UNISINOS

Programa de Pós-Graduação em

Computação Aplicada

Doutorado Acadêmico

Michele Jackeline Andressa Rosa

Um Modelo Preditivo com Base na Integração de Dados Numéricos e Textuais: Um estudo de Caso no Mercado Acionário Brasileiro

São Leopoldo, 2025

Michele Jackeline Andressa Rosa

UM MODELO PREDITIVO COM BASE NA INTEGRAÇÃO DE DADOS NUMÉRICOS E TEXTUAIS:

Um estudo de caso no mercado acionário brasileiro

Tese apresentada como requisito parcial para a obtenção do título de Doutor em Computação Aplicada, pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos – UNISINOS.

Orientador: Prof. Dr. Sandro José Rigo

Coorientador: Prof. Dr. Jorge Luis Victória Barbosa

São Leopoldo, 2025

R788m Rosa, Michele Jackeline Andressa.
Um modelo preditivo com base na integração de dados numéricos e textuais: um estudo de caso no mercado acionário brasileiro / Michele Jackeline Andressa Rosa – 2025.
152 f.: il. color; 30 cm.

Tese (doutorado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, 2025.
"Orientador: Prof. Dr. Sandro José Rigo; Coorientador: Prof. Dr. Jorge Luis Victória Barbosa."

1. Mercado financeiro. 2. Bolsa de valores. 3. Aprendizagem profundo. 4. Processamento de linguagem natural (Computação). 5. Dados heterogêneos. I. Título.

CDU 004:336.76

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Bruna Sant'Anna – CRB 10/2360)

RESUMO

A análise dos movimentos e preços do mercado brasileiro de ações tem sido amplamente estudada, com um crescimento recente no uso de Inteligência Artificial para essa finalidade. Tradicionalmente, as abordagens preditivas baseiam-se em dados históricos numéricos, com ênfase na análise gráfica. No entanto, essas técnicas não exploraram plenamente o potencial dos dados fundamentalistas, extraídos de relatórios técnicos e balanços contábeis, nem aproveitaram uma grande quantidade de informações em tempo real disponibilizadas por mídias sociais e portais de notícias. Este estudo teve como objetivo identificar a abordagem mais eficaz para aumentar a precisão das previsões de preços de ações por meio da integração de dados numéricos e dados textuais, aplicados a um conjunto de ativos do mercado acionários brasileiro. Diferentes técnicas e modelos de aprendizado profundo foram empregados, e a análise da literatura evidenciou lacunas na integração de dados heterogêneos. Para suprir essas limitações, propôs-se uma abordagem que combina dados numéricos e textuais, avaliando os impactos dessa integração na previsão de preços e movimentos de ações. Os dados textuais incluem informações contábeis, postagens no X (antigo *Twitter*), notícias financeiras e econômicas publicadas na *web*. Os dados numéricos consistem em séries históricas de preços e volume das ações, variáveis macroeconômicas, além do índice de buscas do Google trends. O modelo proposto permite avaliar avanços no tratamento e integração de dados numéricos e textuais, tendo em vista a identificação de movimentos e preços de ações no mercado brasileiro. Foram realizados estudos para explorar o comportamento dos dados numéricos e textuais. Também foram realizados experimentos implementando a abordagem proposta, que permitiram observar um ganho percentual na predição quando comparados com a análise apenas numérica. Os resultados revelaram que a inclusão de *tweets*, notícias (*Google News*) e indicadores técnicos, juntamente com dados de preços e volume das ações, melhoraram a precisão das correções. Comparando os modelos testados, o LSTM apresentou melhor desempenho do que o DNN. Os valores de RMSE coletados foram: PETR4 (0,0114; 0,0111; 0,0210), VALE3 (0,0106; 0,0128; 0,0452), BBDC4 (0,0119; 0,0112; 0,0234) e ITUB4 (0,0117; 0,0119). Conclui-se que a integração de dados heterogêneos pode melhorar significativamente a previsão de preços de ações, contribuindo para o desenvolvimento de estratégias mais eficazes no mercado financeiro.

Palavras-Chave: Mercado financeiro, Aprendizagem profundo, Processamento de linguagem natural, Bolsa de valores (B3), Dados heterogêneos.

ABSTRACT

The analysis of movements and prices in the Brazilian stock market has been widely studied, with a recent increase in the use of Artificial Intelligence for this purpose. Traditionally, predictive approaches rely on historical numerical data, with an emphasis on graphical analysis. However, these techniques have not fully explored the potential of fundamental data extracted from technical reports and financial statements, nor have they taken advantage of the vast amount of real-time information available through social media and news portals. This study aimed to identify the most effective approach to improving the accuracy of stock price predictions by integrating numerical and textual data, applied to a set of assets in the Brazilian stock market. Various deep learning techniques and models were employed, and the literature review revealed gaps in integrating heterogeneous data. To address these limitations, an approach was proposed that combines numerical and textual data, assessing the impact of this integration on stock price and movement predictions. The textual data includes financial statement information, posts on X (formerly Twitter), and financial and economic news published online. The numerical data consists of historical stock price and volume series, macroeconomic variables, and the Google Trends search index. The proposed model allows for an evaluation of advancements in the processing and integration of numerical and textual data to identify stock price movements in the Brazilian market. Studies were conducted to explore the behavior of numerical and textual data. Additionally, experiments implementing the proposed approach demonstrated a percentage gain in prediction accuracy compared to purely numerical analysis. The results revealed that the inclusion of tweets, news (Google News), and technical indicators, along with stock price and volume data, improved forecasting accuracy. When comparing the tested models, the LSTM outperformed the DNN. The collected RMSE values were: PETR4 (0.0114; 0.0111; 0.0210), VALE3 (0.0106; 0.0128; 0.0452), BBDC4 (0.0119; 0.0112; 0.0234), and ITUB4 (0.0117; 0.0119). It is concluded that the integration of heterogeneous data can significantly enhance stock price predictions, contributing to the development of more effective strategies in the financial market.

Keywords: *Financial market, Deep learning, Natural language processing, Stock exchange (B3), Heterogeneous data.*

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O mercado de capitais desempenha um papel fundamental para desenvolvimento econômico de um país e oportuniza a participação coletiva nos resultados da economia. Isso ocorre principalmente no mercado acionário, no qual as empresas de capital aberto colocam à disposição suas ações, possibilitado a negociação de compra e vendas deste ativos financeiros (ALZAZAH; CHENG, 2021; LEMOS, 2022).

Um dos fatores fundamentais para se obter ganhos com ações refere-se ao momento certo de entrada e saída do mercado, ou seja, é necessário antecipar-se aos movimentos do mercado financeiro. Para isso, é preciso realizar uma análise do cenário econômico, do mercado financeiro e do desempenho das companhias listadas na bolsa de valores, tanto da empresa em que se pretende investir quanto de suas concorrentes. As decisões tomadas pelos investidores no mercado financeiro não são feitas em um ambiente de total certeza quanto aos seus resultados. Considerando que essas decisões são fundamentalmente voltadas para o futuro, é imprescindível introduzir a variável incerteza como um dos aspectos mais significativos no estudo das operações do mercado financeiro (ASSAF NETO, 2021).

O risco na tomada de decisão de compras de ativos no mercado acionário pode ser caracterizado como uma variável de incerteza relacionada ao futuro. Em geral este risco é representado por uma medida estatística que reflete a dispersão dos resultados em relação ao valor médio esperado. As tomadas de decisão dos investidores baseiam-se nos cenários econômicos, tanto sistemáticos como não sistemáticos. O risco sistemático está relacionado a eventos de natureza política, econômica e social. Por outro lado, o risco não sistemático é específico do próprio ativo (ASSAF NETO, 2021; PINHEIRO, 2019).

Os investidores utilizam uma variedade de técnicas e modelos tradicionais de previsão para tentar antecipar o comportamento do mercado e, assim, evitar grandes perdas, especialmente em momentos de alta incerteza. Essas previsões são feitas usando métodos estatísticos e econométricos, que ajudam a identificar tendências e padrões nos dados. Entre as principais abordagens utilizadas pelos investidores estão a análise fundamentalista e a análise técnica, também conhecida como grafista.

A análise fundamentalista foca na avaliação do valor intrínseco de uma empresa. Para isso, utiliza-se de informações contábeis e financeiras, como os balanços, demonstrativos de resultados e fluxo de caixa, além de dados econômicos gerais que afetam o setor da empresa. O objetivo é analisar tantos fatores microeconômicos (referente a empresa) quanto

macroeconômico (que afetam o mercado em geral), para projetar o desempenho futuro da empresa e determinar se preço atual das ações reflete o seu valor real (LEMOS, 2022).

Já a análise técnica, ou grafista, adota uma abordagem diferente, baseada na interpretação de dados históricos de preços e volumes de negociação. Eles utilizam gráficos e indicadores para identificar padrões que sugerem o comportamento futuro dos preços, com foco em antecipar movimentos, sem considerar os fatores fundamentais da empresa (LEMOS, 2022).

Prever os movimentos do mercado de ações é uma tarefa desafiadora, dada a multiplicidade de fatores que influenciam as cotações das ações. Fatores macroeconômicos, como taxas de juros, inflação, desemprego, câmbio, política e crescimento econômico, contribuem para a instabilidade e volatilidade do mercado, tornando a previsão uma tarefa complexa (ALZAZAH; CHENG, 2021). Os desafios na previsão do mercado de ações estão associados com a alta volatilidade e a não linearidade destes dados (SILVA, 2016).

Para enfrentar esses desafios, as abordagens de previsão de tendências futuras dos preços estão se voltando para a análise dos movimentos e flutuações do mercado de ações, utilizando recursos de Inteligência Artificial e aprendizado de máquina (ALBAOOTH, 2023; CAROSIA, 2023; CASTRO; REYES; LANDAZÁBAL, 2020; DU; HAO; LI, 2022; KACZOROWSKI et al., 2021; KORABLYOV et al., 2023; RUKE et al., 2024; SISMANOGLU et al., 2019; WERNER; BISOGNIN; ARAUJO, 2020).

Sismanoglu et al. (2019) tiveram como objetivo prever o valor futuro de um ativo específico, utilizando informações do mercado acionário e uma abordagem baseada em *Deep Learning*, com foco no comportamento das ações da IBM para treinamento e construção da base de conhecimento. Bou-Hamad e Jamali (2020) destacam o uso promissor da mineração de dados para a previsão de séries temporais financeiras, analisando as taxas de câmbio do dólar australiano (AUD) e do franco suíço (CHF) em relação ao dólar americano. Por sua vez, Werner, Bisognin e Araujo (2020) apresentam uma aplicação de técnicas de previsão sobre o volume de ações negociadas da empresa Petrobras. Já Castro, Reyes e Landazábal (2020) investigaram o desempenho da Bolsa de Valores Mexicana após a crise de 2008, aplicando mineração de dados e modelos tradicionais de aprendizado de máquina, utilizando dados macroeconômicos para prever o Índice da Bolsa Mexicana.

Tradicionalmente, essas abordagens baseiam-se em dados numéricos extraídos dos valores das ações. No entanto, existe uma oportunidade significativa de expandir esse conjunto de dados para incluir informações provenientes da internet, como dados de bancos abertos e publicações em tempo real nas mídias sociais. Essas fontes de dados estão crescendo

exponencialmente e podem enriquecer as previsões (MAQBOOL et al., 2023; NTI; ADEKOYA; WEYORI, 2020b, 2021; VARGAS et al., 2022).

Este trabalho objetiva explorar a ampliação do campo de atuação das técnicas tradicionais, baseadas na análise grafista, com a combinação dos dados da abordagem fundamentalista. Estes normalmente são conhecidos a partir de relatórios técnicos e demonstrações contábeis, mas estão sendo complementados com a grande quantidade de notícias e comentários disponibilizados nas redes sociais e redes mundiais de notícias online (SHI et al., 2019).

Trabalhos anteriores exploraram a integração do movimento diário dos preços das ações utilizando redes neurais profundas, alimentadas com dados de coleções de notícias financeiras e informações numéricas sobre a movimentação dos preços (DING et al., 2014, 2015; NTI; ADEKOYA; WEYORI, 2020b; PENG; JIANG, 2016; VARGAS et al., 2022). No entanto, as pesquisas analisadas apresentaram pouca diversidade nas fontes adicionais de informação e não abordaram em profundidade os fatores relacionados aos ativos avaliados. Esses aspectos foram identificados como oportunidades para aprofundar a pesquisa nessa área.

A partir da análise do estado da arte nessa área, identifica-se uma lacuna significativa: a ausência de estudos que realizem uma análise abrangente das atuais fontes de dados textuais e do seu potencial para apoiar a previsão de movimentos e valores de ações, considerando a integração de várias fontes de dados. Essa fragilidade é particularmente evidente no mercado de ações brasileiro e em documentos em língua portuguesa, que ainda carecem de investigações mais detalhadas com esse foco. Além disso, os estudos existentes não desenvolvem modelos capazes de analisar de maneira eficaz a correlação entre notícias de diferentes mídias e as oscilações nos preços das ações.

A abordagem proposta considera tanto a dinâmica dos principais documentos tradicionalmente utilizados nesta área quanto a das novas fontes em rápida consolidação. Historicamente, a análise fundamentalista no mercado financeiro envolve o estudo detalhado de balanços contábeis e de notícias financeiras sobre empresas e tendências econômicas. Contudo, a literatura destaca limitações nessa abordagem, apontando a demora na produção, divulgação e acesso a esses materiais. O processo de coleta, edição e publicação pode ser demorado, o que torna esses dados inadequados para suportar análises de alta frequência, como as diárias, que dependem de dados numéricos em tempo real, conforme discutido na abordagem grafista (ARAÚJO; FERNANDES, 2021).

Entretanto, a literatura descreve uma dinâmica diferenciada emergente com a

consolidação das redes sociais. Nesse contexto, as redes sociais funcionam como um mecanismo de produção e divulgação de notícias, percepções e opiniões com uma dinâmica oposta à tradicional. As plataformas permitem a geração e disseminação de informações em tempo real, em uma escala global e por um número massivo de usuários. Estudos indicam que, já há algum tempo, as redes sociais operam em uma escala de milhões de usuários e geração de mensagens diariamente (PINHEIRO, 2019). De forma semelhante, as agências de notícias contemporâneas adaptaram suas dinâmicas ao contexto de acesso imediato à informação, replicando o comportamento das redes sociais ao oferecer notícias e opiniões em tempo real, frequentemente acompanhadas de conteúdo multimídia como texto, áudio e imagens (ROSENTHAL; FARRA; NAKOV, 2019).

Este trabalho considera a possibilidade de que a divulgação em escala global e em tempo real de notícias, percepções e opiniões possa contribuir para a predição de movimentos de ações, quando integrada a dados grafistas. A proposta de tese sugere o estudo dessas fontes textuais com o objetivo de verificar seu potencial no apoio à predição de movimentos no mercado acionário brasileiro. A partir dessa análise, propõe-se uma abordagem que integre esses dados textuais com dados numéricos, buscando avaliar possíveis ganhos em tarefas de previsão de movimentos e valores de ações.

Diferentemente de estudos anteriores que utilizam abordagens semelhantes, esta proposta se destaca pela ampla variedade de fontes de dados e pela aplicação de técnicas avançadas de aprendizado profundo, permitindo uma avaliação abrangente da pertinência dessa integração de dados na predição de ações. O modelo proposto visa explorar avanços na combinação de dados numéricos e textuais, com o objetivo de identificar padrões e movimentos de preços no mercado brasileiro.

Dessa forma, o trabalho apresenta uma análise da previsão de comportamento de ativos específicos do mercado acionário brasileiro, por meio da fusão de séries históricas e dados textuais diversificados, utilizando técnicas de *Deep Learning* e Processamento de Linguagem Natural. Com isso, busca-se contribuir para o desenvolvimento de abordagens eficazes na geração de previsões, auxiliando na tomada de decisões mais assertivas, com a minimização de riscos e incertezas ao incorporar dados adicionais sobre o mercado e as bolsas de valores.

1.1 Motivação

A teoria clássica da economia pressupõe que os agentes econômicos são racionais e qualquer desvio dessa regra é considerado uma anomalia. No entanto, a economia

comportamental contesta essa teoria, sugerindo que a racionalidade dos indivíduos é limitada e que existem várias outras variáveis que influenciam suas decisões econômicas. Nesse processo decisório, diversos fatores são considerados, tais como aspectos emocionais, sociais, econômicos, cognitivos, culturais, entre outros (FERREIRA, 2008; STATMAN, 1995; THALER, 2016).

Para compreender o processo decisório dos agentes, os pesquisadores Kahneman e Tversky (1979); Tversky e Kahneman (1974); Twersky (1972), publicaram diversos trabalhos sobre o comportamento e o processo de decisão. Os autores apresentaram problemas a diversos grupos de pessoas, que eram levadas a tomar decisões considerando o ganho, a perda e o risco envolvido no processo decisório.

Nos últimos anos, o número de investidores pessoa física na Bolsa de Valores brasileira (B3) apresentou grande crescimento. Em 2002, existiam 85,2 mil investidores e desde então vem ocorrendo um aumento contínuo. Em 2010, este número de investidores chegou a 610,9 mil. Já em 2018, o número atingiu 813,9 mil, com uma variação de 31,21% em relação ao ano anterior. Em 2020, o número de investidores quase dobrou em relação a 2019, passando de 1,66 milhões para 3,2 milhões, um aumento de 94%. Em 2021, o crescimento foi de 53,8%, com 4,97 milhões investidores. Nos anos seguintes, os crescimentos foram menores, com 5,84 milhões em 2022 e 5,73 milhões em 2023, uma redução -1,91%. No primeiro semestre de 2024, a Bolsa de Valores já contava com cerca de 6 milhões de investidores. Esse crescimento reflete uma maior conscientização financeira e interesse em diversificar investimentos (B3, 2020, 2022, 2024a).

Recentemente, observa-se um cenário de grandes incertezas devido à pandemia provocada pelo coronavírus (Covid-19), que impactou o mundo inteiro, gerando uma crise sem precedentes para a sociedade e o mercado financeiro. Adicionalmente, várias mudanças no comportamento da sociedade brasileira em relação ao consumo e hábitos financeiros foram observadas. As pessoas passaram a dar mais atenção à vida financeira, tanto aquelas que tinha uma reserva de emergência quanto as que nunca tiveram. A crise motivou muitos a planejar melhor suas finanças, buscando segurança e diversificação de investimentos.

Houve também um aumento na busca por informações sobre finanças e investimentos, impulsionado pela queda da taxa básica de juros (Selic), que incentivou os investidores a buscarem alternativas mais rentáveis do que a renda fixa. Em 2018, a Selic estava em 6,4% no primeiro trimestre, mas em 2019 o Comitê de Política Monetária (Copom) do Banco Central realizou quatro reduções, levando a taxa a 4,5% ao ano. Em 5 de agosto de 2020, a Selic foi

cortada para 2% ao ano, o menor nível da história até então, mantendo-se assim até o início de 2021. Paralelamente, observou-se um aumento do número de investidores pessoas físicas no mercado acionário, mesmo durante um período de sucessivos cortes na taxa Selic (BACEN, 2022).

Desta forma, os principais fatores que influenciaram o crescimento da renda variável foram a redução da taxa Selic, os reflexos da crise causada pela Pandemia, que gerou um cenário de grandes incertezas e riscos no mundo inteiro, além da popularização de plataformas de investimentos digitais, entre outros fatores.

Devido ao cenário de riscos e incertezas na economia, o Bacen realizou vários ajustes, aumentando a taxa de juros desde março de 2021, atingindo 10,75% em fevereiro de 2022. A taxa continuou subindo e atingiu 13,75% em agosto de 2022, com o objetivo de desestimular o consumo e tentar controlar a inflação, devido aos impactos econômicos da pandemia do COVID-19 e às consequências da guerra na Ucrânia. Em 2023, a Selic se manteve em 13,75 por um período, mas, como a inflação mostrou sinais de desaceleração, o Banco Central começou a realizar cortes. Em 2024, a taxa foi ajustada para 10,75% ao ano, visando equilibrar o crescimento econômico e o controle da inflação em um cenário econômico global incerto (BACEN, 2022, 2024).

Portanto, com o início da pandemia de COVID-19, o número de investidores na bolsa de valores aumentou significativamente. Entretanto, muitas pessoas ainda têm certo receio do risco envolvido nestas operações. Bertão (2021) esclarece que muitas pessoas alegam não investir por falta de dinheiro, enquanto outras têm medo de perder os valores que investiram. A volatilidade, os riscos e as incertezas do mercado ainda assustam alguns investidores.

Os investidores sempre procuram tomar as melhores decisões sobre quais ativos comprar ou vender, analisando os movimentos dos mercados financeiros para fazer uma previsão precisa e obter os melhores retornos. Isso inclui identificar os momentos certos para entrar e sair da bolsa e evitar grandes perdas, considerando os riscos e incertezas do mercado financeiro (ASSAF NETO, 2021). Ao mesmo tempo, as técnicas usadas para a tomada de decisão estão incorporando gradativamente o uso de recursos automatizados para a análise de movimentos e tendências. A análise técnica, baseada em gráficos por meio de séries históricas de preços das ações e indicadores, está sendo combinada com técnicas de aprendizagem de máquina para promover melhorias nas previsões de mercado.

No entanto, as técnicas de aprendizagem de máquina atuais ainda utilizam de maneira incipiente os dados não numéricos, como aqueles encontrados em notícias, balanços contábeis

e redes sociais. Dado que a forma de produção e divulgação de notícias mudou drasticamente nos últimos anos, este trabalho busca avaliar o potencial dessas fontes de dados como elementos de apoio em previsão de preços, integrando-as com dados numéricos para aprimorar a precisão e a robustez dos modelos preditivos, oferecendo uma visão mais abrangente da dinâmica de mercado para investidores e pesquisadores.

1.2 Questão de Pesquisa

Diante do exposto, a questão central deste trabalho é: Quais são os elementos fundamentais para desenvolvimento de um modelo preditivo capaz de gerar ganhos de assertividade com a integração de conjuntos de dados numéricos e textuais na previsão de movimentos no mercado de ações Brasileiro?

1.3 Objetivos

Com base no estudo realizado, o objetivo deste trabalho é desenvolver um modelo preditivo de preços de ações utilizando uma combinação de conjuntos de dados, com ênfase na implementação de técnicas de Processamento de Linguagem Natural e de Aprendizagem de Máquina e Aprendizado Profundo, com foco na otimização de métodos de pré-processamento, seleção de recursos e redução de dimensionalidade, visando aprimorar a precisão no desempenho das previsões. A seguir são descritos os objetivos gerais e específicos.

1.3.1 Objetivo geral

Identificar a abordagem mais eficaz para aumentar a precisão das previsões de preços de ações por meio da integração de dados numéricos e dados textuais, aplicados a um conjunto de ativos do mercado acionários brasileiro, através de diferentes técnicas e modelos de aprendizado profundo.

1.3.2 Objetivo específicos

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

1. Descrever as principais teorias, conceitos e trabalhos relacionados ao mercado de capitais, assim como os critérios de análise de ações e análise de sentimentos, com foco especial na aplicação de inteligência artificial e processamento de linguagem natural;
2. Coletar e categorizar diferentes tipos de dados, tanto numéricos quanto textuais, relevantes para a previsão de movimentos no mercado de ações determina uma abordagem ampla e integrada. Isso inclui séries históricas de preços e volume de ações, indicadores técnicos, dados macroeconômicos, sentimentos extraídos de notícias e redes sociais, além de indicadores fundamentalistas;
3. Propor uma abordagem detalhada para a integração de dados textuais e numéricos na previsão de movimentos no mercado acionário, utilizando técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado Profundo;
4. Executar experimentos com o modelo preditivo, utilizando diferentes cenários e dados de teste, para validar a eficácia do modelo proposto na previsão de movimentos de ações;
5. Analisar os resultados dos experimentos, avaliando a precisão das previsões e o impacto da integração de dados textuais na melhoria das previsões de movimentos no mercado de ações;

1.4 Estrutura da Tese

Esta proposta de tese está organizada em seis capítulos. O Capítulo 2 apresenta a fundamentação teórica sobre temas importantes para o desenvolvimento do trabalho. O Capítulo 3 descreve os trabalhos relacionados que foram estudados para mapear o estado da arte e identificar lacunas. O Capítulo 4 apresenta a proposta de abordagem para suportar os objetivos do trabalho. O Capítulo 5 expõe os resultados de experimentos realizados. Por fim, o Capítulo 6 descreve as considerações finais, as contribuições, limitações observadas e trabalhos futuros sugeridos.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão apresentadas as teorias e os conceitos que fundamentam o tema da pesquisa, sendo estes estruturados em cinco seções. A primeira seção aborda os conceitos de hipótese de mercado eficiente e finanças comportamentais. A segunda seção apresenta as principais característica do mercado acionário. A terceira seção mostra os critérios de análise de ações. A quarta seção expõe os conceitos sobre fontes de informações textuais, tipos de dados financeiros e análise de sentimentos. Na quinta seção são apresentadas as teorias e os conceitos de inteligência artificial, aprendizado de máquina, aprendizado profundo e processamento de linguagem natural.

2.1 Hipótese de Mercado Eficiente (HME)

Segundo Fama, (1965, 1970), o autor defende que o mercado de ações é imprevisível, ou seja, mercado é eficiente à informação, onde os preços das ações refletem todas as informações relevantes disponíveis e seu ajuste à novas informações é instantâneo.

Para que os mercados sejam eficientes do ponto de vista econômico, devem ser considerados quatro requisitos relacionados as características de mercados para que sejam mais ou menos eficientes: serem competitivos, transparentes, líquidos, e o tamanho deve possibilitar custos de transação razoavelmente baixos (PINHEIRO, 2019).

Para Casilda, Lamothe e Monjas (1997), existem várias discussões sobre os conceitos de eficiência, cuja principal divergência está nos diferentes níveis de informação que os preços de mercado devem refletir para alcançar a eficiência.

Especialistas do mercado financeiro classificam a eficiência em três níveis: hipótese fraca de eficiência, na qual os preços atuais devem refletir toda a informação histórica de preços e volume; hipótese média ou semiforte, em que os preços devem refletir toda a informação disponível publicamente; e hipótese forte, que exige que os preços reflitam toda a informação disponível sobre mercado, tanto pública como privada (PINHEIRO, 2019). Portanto, existe uma chance de prever o mercado de ações quando se possui conhecimento e compreensão dos dados históricos de ações (análise técnica) e de dados econômico-financeiros (análise fundamentalista), o que pode levar a uma previsão bem-sucedida do preço futuro das ações da empresa.

Portanto, a HME fornece uma base teórica para a eficiência e a racionalidade dos mercados. As finanças comportamentais, no entanto, surgiram para desafiar essa ideia,

destacando que a irracionalidade dos investidores pode levar a ineficiência. Assim, torna-se necessário compreender o comportamento do mercado e desenvolver modelos de previsão de preços que considerem tanto aspectos racionais quanto comportamentais, como proposto nesta tese. Na próxima subseção, são apresentadas as teorias e conceitos das finanças comportamentais.

2.1.1 Finanças Comportamentais

A teoria tradicional das finanças foi construída a partir da abordagem microeconômica neoclássica, em que o modelo essencial é a racionalidade dos agentes econômicos. Nesse modelo, os agentes possuem preferências estáveis e coerentes, levando à hipótese de mercados eficientes, que afirmam que os preços refletem de maneira eficiente todas as informações disponíveis (FAMA, 1991; MALKIEL; FAMA, 1970; MILANEZ, 2003).

As finanças comportamentais surgiram nos anos 50 e início da década de 60, como forma de avaliar as decisões relacionadas aos investimentos ao considerar o comportamento do investidor. Elas incorporam os sentimentos e motivações internas dos investidores. Conceitos provenientes da economia, psicologia cognitiva e finanças contribuem para as finanças comportamentais, construindo modelos do comportamento dos agentes levam em conta a ideia de que os investidores estão sujeitos a vieses comportamentais e heurísticas (ANACHE, 2008).

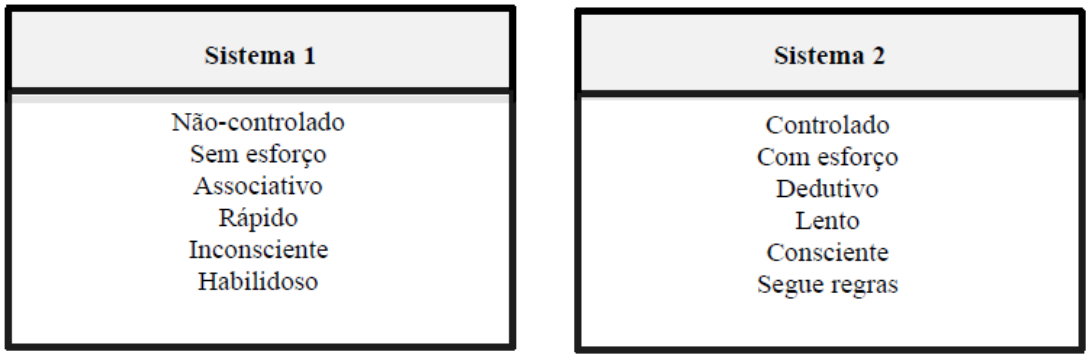
As finanças comportamentais têm como objetivo a compreender as ilusões cognitivas que levam os indivíduos a cometer erros sistemáticos, partindo do pressuposto que os indivíduos possuem racionalidade limitada e, portanto, são propensos aos efeitos dessas ilusões cognitivas (ROGERS; FAVATO; SECURATO, 2008).

As anomalias do comportamento dos indivíduos na tomada de decisão foram alvo de diversos estudos. Tversky & Kahneman (1974), sugeriram uma abordagem para lidar com essas anomalias, conhecida como eliminação por aspectos. Esse método consiste na escolha de um aspecto específico para eliminar opções, reduzindo assim as alternativas disponíveis até restar apenas uma opção.

Para Kahneman (2012), a mente opera com dois sistemas de raciocínio distintos: o sistema um e o sistema dois. O primeiro sistema funciona de maneira automática, sem esforço e de forma instintiva, já o segundo demanda concentração, sendo mais intencional e consciente. Esse segundo sistema é utilizado para atividades complexos, como cálculos complexos. Esse

sistema dual ajuda a explicar os padrões e as falhas nas escolhas cotidianas das pessoas. Os principais recursos de cada sistema estão apresentados na Figura 1.

Figura 1: Dois Sistemas Cognitivos.



Fonte: Adaptado de Tversky e Khaneman (1974).

Thaler e Sunstein (2009), classificam o sistema 1 como automático, rápido e instintivo, sendo associadosos pelos neurocientistas às partes mais antigas do cérebro. O sistema 2, por sua vez, pode ser descrito como reflexivo, intencional e consciente, ou seja, mais racional. No entanto, ambos os sistemas estão sujeitos a falhas cognitivas, o que reflete a racionalidade limitada do ser humano.

Richard H. Thaler, premiado com o Prêmio Nobel de Economia em 2017, por suas pesquisas sobre os mecanismos psicológicos e sociais na tomada de decisão dos investidores, demonstrou a irracionalidade dos agentes econômicos. Em suas pesquisas, ele mostrou como as escolhas se baseiam em critérios individuais, comprovando a ideia de que os traços humanos afetam significativamente o mercado financeiro (THALER, 2016).

Os vieses comportamentais ou erros sistêmicos são divididos em dois tipos. Os vieses cognitivos representam tendências de agir ou pensar de formas pré-definidas, levando a erros sistemáticos nos modelos de bom julgamento e racionalidade. Já os vieses emocionais referem-se às tomadas de decisões baseadas em sentimentos (YOSHINAGA; RAMALHO, 2014).

Segundo Yoshinaga e Ramalho (2014), alguns vieses que se manifestam com mais frequência no mercado financeiro são:

- Viés do arrependimento: a emoção associada a um acontecimento do passado, onde não se tomou a melhor decisão, tendendo a influenciar as decisões futuras.
- Viés do conservadorismo: a tendência do indivíduo de demorar a atualizar suas crenças ou valores diante de mudanças duradoras.
- Viés do excesso de confiança: a crença exagerada nas próprias habilidades, levando o indivíduo a desconsiderar os riscos.

- Viés de confirmação: os investidores procuram informações que confirmem suas crenças e ignoram ou desvalorizam informações que as contradigam.
- Viés de aversão à perda: os investidores têm uma tendência maior a evitar perdas do que a obter ganhos.
- Viés de retrocesso: os investidores acreditam que eventos passados são preditores diretos de eventos futuros, ignorando a independência dos eventos.

Já as Heurísticas, palavra derivada do grego *"Heureka"*, que significa "descobrir" ou "inventar", são atalhos que a mente utiliza para tornar o processo de tomada de decisão menos complexo e mais eficiente. No entanto, o uso de heurísticas pode levar a erros sistemáticos. Entre as muitas existentes, três são as mais importantes e frequentemente utilizadas em situação de incerteza: representatividade, ancoragem e disponibilidade Tversky e Kaneman (1974):

- Representatividade: é um atalho mental que se baseia em algo com características comuns;
- Ancoragem: ocorre quando o indivíduo se ampara em um valor de referência para direcionar suas ações, mesmo que essa referência não seja relevante para a decisão atual;
- Disponibilidade: refere-se à avaliação da probabilidade ou frequência de um evento com base na facilidade de trazer à memória eventos semelhantes.

A Figura 2 apresenta os principais vieses das heurísticas de representatividade, disponibilidade e ajuste (ou ancoragem).

Figura 2: Heurísticas e Vieses.

REPRESENTATIVIDADE	DISPONIBILIDADE	AJUSTE OU ANCORAGEM
Insensibilidade à probabilidade a priori dos resultados	Viés devido à recuperabilidade das ocorrências	Viés do ajuste insuficiente
Insensibilidade ao tamanho da Amostra	Viés devido à efetividade de um ajuste de busca	Viés na avaliação de eventos conjuntivo e disjuntivo
Concepções errôneas da possibilidade	Viés da imaginabilidade	Viés na avaliação das distribuições de probabilidade subjetiva
Insensibilidade à previsibilidade	Viés de correlação ilusória	
Ilusão da validade		
Concepções errôneas da regressão		

Fonte: Adaptado de Tversky e Khaneman (1974).

Os vieses apresentados podem ser classificados como falhas. Dessa forma, a partir de crenças (heurísticas), formam-se espécies de caminho para facilitar as decisões econômicas dos indivíduos (FERREIRA, 2008).

Segundo Bazerman e Moore, (2012), as heurísticas podem ser úteis, no entanto, seus vieses podem provocar graves erros na percepção de padrões, pois os indivíduos podem tomar suas experiências como verdades, ignorando a realidade se baseando-se apenas em crenças individuais ou coletivas.

Kahneman, (2012), nos explica que as heurísticas se relacionam com nosso sistema dual de raciocínio, já que o sistema dois opera com base em eventos já vivenciados, ou seja, fundamentados em dados da memória. Dessa forma, os indivíduos se tornam suscetíveis aos vieses das heurísticas.

Segundo Sbicca (2014), estudos sobre heurísticas e vieses ajudam a explicar as decisões cotidianas da vida econômica, como uso de cartão de crédito, decisões financeiras, contratações de seguros, compras e vendas de ações etc. Assim, as decisões econômicas são, em parte, influenciadas pela percepção individual das pessoas e nem sempre são totalmente racionais, pois distorções cognitivas interferem na tomada de decisão.

A teoria do prospecto, desenvolvida por Kahneman e Tversky, aponta que os indivíduos, de maneira geral, reagem com mais intensidade às perdas do que aos ganhos, o que pode levá-los a rejeitar situações favoráveis e preferir a inércia para evitar riscos. Dessa forma, podemos dizer que as pessoas tendem a ter aversão à perda e ao risco (PAIVA, 2013). Na próxima seção são apresentados os conceitos e caraterísticas de Mercado de capitais.

2.2 Mercado de Capitais

Mercado de capitais é constituído pelo conjunto de empresas que fazem negociação de títulos e valores mobiliário, conectando investidores (agentes superavitários) a aqueles que necessitam de recursos de longo prazo, com as empresas (agentes deficitários). Esse mercado tem como objetivo as suprir as necessidades de investimentos dos agentes econômicos, por meio de diversas modalidades de financiamentos a médio e longo prazos, voltadas para capital de giro e capital fixo. Sendo constituído, principalmente, por instituições financeiras não bancárias (ASSAF NETO, 2021; PINHEIRO, 2019).

Os principais títulos e valores mobiliários negociados no mercado de capitais são as ações, opções sobre ações, *depositary receipts*, *brazilian depositary receipts*, debêntures, letras

de câmbio, certificados/recibos de depósitos bancários (CDB/RDB), caderneta de poupança, letras hipotecárias, letras imobiliárias, *warrants*, títulos conversíveis e letras financeiras (ASSAF NETO, 2021). A subseção a seguir apresenta as principais características do mercado acionário.

2.2.1 Mercado de Ações

No mercado acionário, ocorrem as negociações de ações de sociedades anônimas, empresas de capital aberto. No Brasil, essas empresas são regulamentadas pela Lei nº 6.404, de 15 de dezembro de 1976, conhecida como lei das S.A., com as alterações da Lei nº 9.457, de maio de 1997. Essa legislação apresenta duas formas de sociedade por ações: as de capital aberto, que possuem títulos negociáveis em bolsa de valores, e as de capital fechado, cujos títulos não são negociáveis em bolsa (PINHEIRO, 2019).

No Brasil, a principal bolsa é a B3 S.A, que atua tanto no ambiente de bolsa quanto no mercado de balcão. Em 2017, aconteceu a fusão entre a BM&F Bovespa - Bolsa de Valores, Mercadorias e Futuros com a Cetip, empresa atuante no mercado de balcão organizado. O mercado brasileiro de valores mobiliários é regulamentado pela Comissão de Valores Mobiliários (CVM), bem como pela Conselho Monetário Nacional (CMN) e pelo Banco Central do Brasil (Bacen) (ASSAF NETO, 2021; B3, 2024b).

A B3 é uma sociedade de capital aberto que emite somente ações ordinárias, negociadas no Novo Mercado, e integra os índices Ibovespa, IBrX-50, IBrx e Itag, e entre outros. O Novo Mercado das empresas de capital aberto adota melhores práticas de governança corporativas, com inovações em produtos e tecnologias, sendo B3 uma das maiores bolsas em valor de mercado e em participação global nos setores de bolsas (B3, 2024b).

As ações são títulos que representam uma fração do capital social de uma empresa de capital aberto. O acionista é coproprietário da empresa, com direitos e obrigações proporcionais à sua participação no capital social. Os investidores emitem uma ordem de compra ou venda de determinada ação, que são negociadas nas bolsas de valores por meio de uma corretora, responsável por executar a ordem recebida no pregão da bolsa (ASSAF NETO, 2021; PINHEIRO, 2019).

As ações representam frações do capital social de uma empresa e são classificadas em dois tipos básicos: ordinárias e as preferenciais. As ações ordinárias, além de garantirem a participação nos lucros, conferem aos seus proprietários o direito ao voto, permitindo que assim

influenciem nas decisões da empresa. As ações preferenciais, por outro lado, geralmente não concedem direito a voto, mas oferecem prioridade no recebimento de dividendos e, em alguns casos, percentuais mais elevados da parcela de lucro (ASSAF NETO, 2021).

O mercado acionário é dividido em dois segmentos: o mercado primário e o mercado secundário. No mercado primário, ocorre a emissão primária das ações, em que o dinheiro dos investidores é direcionado diretamente ao caixa da empresa. No mercado secundário, as negociações acontecem entre os próprios investidores nas bolsas de valores, sem que haja qualquer impacto financeiro direto no caixa da empresa (PINHEIRO, 2019).

As aquisições de ações têm sempre um caráter estratégico, e o preço que um comprador está disposto a pagar por um ativo da empresa está associado às expectativas futuras sobre o desempenho dessas ações. Esse valor depende, em grande parte, da capacidade de inovação do comprador em fazer a empresa adquirida gerar resultados acima dos obtidos na condição atual. Assim, os compradores que enxergam na empresa um alto potencial para geração de riqueza, além dos níveis atuais, estarão dispostos a pagar um preço mais elevado pelo seu controle acionário (SILVA, 2016).

As vantagens para os investidores ao adquirir ações podem ser definidas de acordo com Assaf Neto, (2021):

a) Dividendos – representam uma parte dos lucros da empresa apurados em determinado exercício e repassados aos acionistas em dinheiro. No Brasil, foi introduzido o pagamento de juros sobre o capital próprio como forma adicional de remuneração aos acionistas. Esse pagamento é facultativo, mas, quando realizado, é considerado parte dos dividendos;
b) Bonificações – ocorrem quando há aumento do capital social da empresa mediante incorporação de reservas patrimoniais. Esse aumentando resulta na emissão de novas ações, que são distribuídas gratuitamente aos acionistas, de forma proporcional à participação no capital;
c) Direito de subscrição – concede aos acionistas atuais o direito de adquirir novas ações, em qualquer aumento de capital, proporcional à quantidade de ações que já possuem. A subscrição não é obrigatória, mas pode ser negociada como um direito;
d) Valorização – representa o ganho de capital que um acionista pode obter a partir da valorização de suas ações no mercado. Esse ganho depende do preço compra, da quantidade de ações emitidas e de fatores econômicos e do desempenho econômico e

financeiro da empresa. Na subseção a seguir, serão apresentadas os riscos, retorno e mercado.

2.2.2 Riscos, Retorno e Mercado

O mercado acionário apresenta maior potencial de rentabilidade em comparação com outros investimentos, porém com um elevado grau de risco, devido às oscilações das cotações do valor de mercado das ações.

O preço de mercado de um ativo é o valor efetivo de negociação das ações na bolsa de valores, definido a partir das percepções dos investidores e das estimativas o desempenho da empresa e da economia. Assim, os preços das ações são formados pela interação da oferta e da demanda de mercado, ou seja, pela quantidade de investidores interessados em comprar ou vender a determinado preço. Os ganhos de capital oferecidos aos acionistas estão sujeitos ao comportamento incerto dos preços de mercados das ações (SINATORA, 2016).

Para obter ganhos com ações, é importante o *timing* (momento oportuno) de entrada e saída do mercado, pois investir em ações envolve assumir certo grau de risco. A compensação desse risco deve ser refletida na remuneração oferecida pelo papel, que será mais elevada quanto maior for o risco. Assim, o diferencial de um bom investidor está em saber antecipar os movimentos do mercado. Para isso, é necessário acompanhar as análises de cenário econômicos, do mercado financeiro e do desempenho da empresa e seus concorrentes (ASSAF NETO, 2021; PINHEIRO, 2019).

Os riscos de um investimento em ações podem ser identificados como: risco da empresa que capta os recursos e risco do mercado. O primeiro está relacionado as decisões financeiras, levando em conta a atratividade do negócio e sua capacidade financeira. O risco de mercado está vinculado às variações imprevistas no comportamento do mercado, principalmente devido a mudanças na economia. O risco da empresa pode ser subdividido em risco econômico e risco financeiro (ASSAF NETO, 2021).

O risco econômico é uma característica do mercado em que a empresa atua, sendo influenciado por fatores como aumento da concorrência, evolução tecnológica, altas taxas de juros e qualidade dos produtos, entres outros fatores. Já o risco financeiro está relacionado ao endividamento da empresa, ou seja, à sua capacidade de horar seus compromissos financeiros. O comportamento desses dois elementos de risco afeta o risco total da empresa e o valor de mercado de suas ações. Deve haver um equilíbrio na relação risco/retorno esperado do

investimento em ações, conseguindo a máxima parcela de rentabilidade associada a um nível de risco que solicite o maior valor de mercado das ações (ASSAF NETO, 2021).

Portanto, o risco pode ser vinculado com a probabilidade. Existe risco sempre que a probabilidade de um determinado evento ocorrer é elevada. Os dois principais tipos de risco são o sistemático e o não sistemático. O risco sistemático é inerente a todos ativos do mercado e é determinado por eventos de natureza política, econômica e social, afetando a economia de um modo geral. Já o risco não sistemático é intrínseco a cada investimento, estando ligado ao desempenho específico das empresas (SINATORA, 2016).

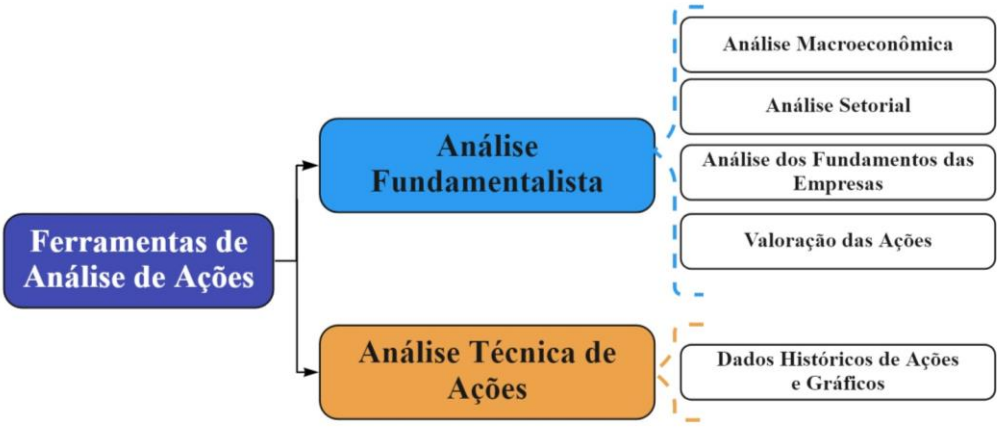
Para mitigar os riscos não sistemáticos no mercado acionário, é importante diversificar da carteira de investimentos, o que consiste em comprar ações de empresas de setores distintos. Dessa forma, uma queda de um setor pode ser compensada por outro, já que um problema setorial não necessariamente afeta os demais. Além disso, a estatística é uma ferramenta valiosa para prever e avaliar riscos. O modelo de *Markowitz*, por exemplo, utiliza a média aritmética, média ponderada e outras medidas estatísticas de dispersão e risco, como desvio padrão, a variância, covariância e correlação (SINATORA, 2016). Na seção a seguir apresenta as descrições e técnicas das ferramentas tradicionais de análise de ações para avaliar o comportamento e as flutuações do mercado de ações.

2.3 Critério de Análise de Ações

Os modelos de avaliação do comportamento futuro dos ativos financeiros são essenciais para projetar previsões sobre as variações de alta e baixa de preço no mercado. Como mencionado incialmente, são adotados dois critérios de análise para investimento em ações: análise técnica e análise fundamentalista, como mostra a Figura 3.

A Figura 3 apresenta as principais ferramentas de análise de ações. A análise fundamentalista baseia-se em fundamentos econômico-financeiros das empresas, incluindo demonstrações contábeis, análise dos fundamentos empresariais e valoração das ações. Além disso, considera o ambiente macroeconômico, o mercado e o setor em que a empresa atua, para determinar seu desempenho no mercado de ações, utilizando dados numéricos e textuais. A análise técnica, por sua vez, concentra-se em dados históricos das ações, observando os movimentos e flutuações dos preços das ações com o auxílio de gráficos, trabalhando predominantemente com dados numéricos. O Quadro 1 apresenta uma comparação das escolas de análise de ações.

Figura 3: Ferramentas de Análise de Ações.



Fonte: Elaborado pela autora.

A principal diferença entre as duas escolas reside em suas metodologias. A escola fundamentalista é embasada em método científico e geralmente seguida por analistas acadêmicos. Já a escola técnica, ou grafista, é composta por agentes que atuam diretamente no mercado e buscam análises rápidas para não perder o *timing* de negociação (KORBES e COSTA JR., 2003; PINHEIRO, 2019).

Apesar das suas características distintas, principalmente em relação às metodologias de análise, ambos os métodos são fundamentais para análise do mercado acionário. São complementares e fornecem instrumentos valiosos para a tomada de decisão.

Quadro 1: Comparação entre as escolas de análise de ações.

Itens	Fundamentalista	Técnica
Idade	Mais nova.	Mais antiga.
Origem	Acadêmica.	Profissional.
Usuário	Administradores de fundos e investidores no longo prazo.	Especulador.
Questionamento	Por quê?	Quando?
Decisão de Investimento	Baseada nos fundamentos da empresa.	Baseada em gráficos de movimentação de ativos.
Hipóteses Básicas	Existe um valor real ou intrínseco para cada ação que está diretamente correlacionado com o desempenho da empresa.	Os preços das ações se movimentam em tendências e existe uma dependência significativa entre as oscilações dos preços que se sucedem.

Objetivos	O objetivo da análise fundamentalista é determinar o real valor de uma ação, calculado com base em receita, lucro, patrimônio, valor presente líquido dos fluxos de caixa futuros.	O objetivo da análise técnica é determinar a tendência de evolução das cotações no curto prazo, a fim de se aproveitar das rápidas oscilações para auferir ganhos de capital (vender as ações por um preço superior ao da compra).

Fonte: Pinheiro (2019, p.458).

Desta forma, a análise fundamentalista mostra diversas alternativas disponíveis no mercado, indicando quais ativos apresentam expectativas de investimentos rentáveis, ou seja, "o que comprar ou vender". Já a análise técnica, ou grafista, aponta o momento mais adequando para execução do investimento, ou seja, "quando comprar ou vender" (KORBES e COSTA JR., 2003). A subseção a seguir apresenta os conceitos e embasamentos da análise técnica e fundamentalista, assim como os principais indicadores utilizados em ambas para prever o movimento do mercado acionário.

2.3.1 Análise Técnica

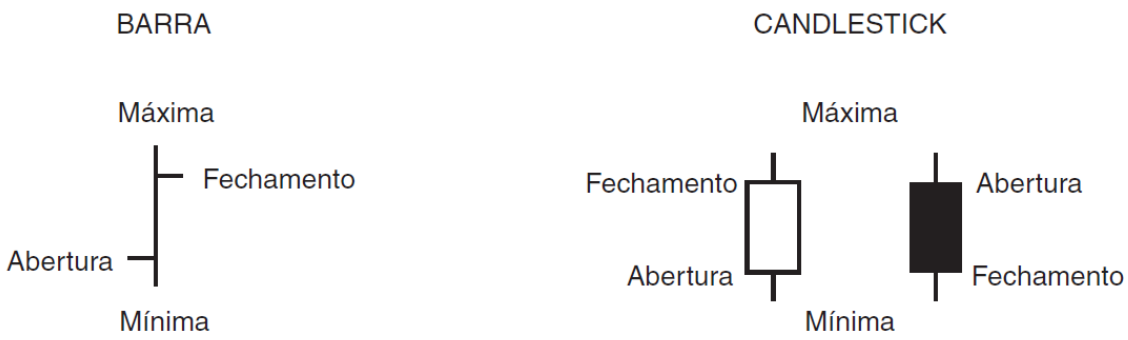
A análise técnica, também conhecida como análise grafista, realiza projeções futuras sobre o comportamento das ações considerando o desempenho do passado do mercado, sendo considerados os parâmetros de oferta e procura desses ativos e a evolução de suas cotações (ASSAF NETO, 2021). Segundo Kobori, (2011) e Lemos, (2022), as premissas básicas que fundamentais que sustentam a análise técnica são:

- O preço do mercado é determinando pela interação das forças da demanda e oferta;
- Demanda e oferta são impulsionadas por fatores racionais e irracionais;
- Os preços se movem em tendências, com objetivo de identificar o movimento dos preços no gráfico, buscando captar tendências desde seus estágios iniciais para aproveitamento em prol do lucro. Dessa forma, uma tendência em andamento tem maior probabilidade de continuar do que de se reverter;
- Os preços tendem a se movimentar em uma direção até que alterações nos fundamentos alterem a sua trajetória;
- Teoria de *Dow* – alguns padrões observados no passado tendem a se repetir no futuro, pois os preços refletem variáveis psicológicas inerentes ao

comportamento humano. Os padrões comportamentais mudam sutilmente ou permanecem estáticos ao longo do tempo.

A análise técnica utiliza gráficos como principais ferramentas recomendações. Para isso, as cotações são divididas em períodos predeterminados, como dias, semanas, meses, ou até intervalos de um ou dois minutos. Após a definição do período, o gráfico é plotado, sendo os mais comuns: gráfico de linha, gráfico de volume, gráfico de barras, gráfico ponto-figura e gráfico *candlestick*, utilizados para as prever tendências futuras de preço (MARTINS, 2010; PINHEIRO, 2019). A Figura 4 apresenta exemplos de gráficos de barras e *candlestick.* No gráfico de barras, cada ponto corresponde a um período de negociação, exibindo o preço de abertura à esquerdo, o fechamento à direito, e os preços máximo e mínimo de forma vertical.

Figura 4: Alguns tipos de gráficos.



Fonte: Martins (2010).

Como mostrado na Figura 4, o gráfico de *Candlestick* (ou candelabro), apresenta uma serie “velas” que indicam a variação de preços no período. Caso o fechamento seja superior à abertura, as velas ficam verdes ou brancos; caso seja inferior, ficam vermelhas ou pretos. A seguir, são comentados os indicadores técnicos, que refletem de forma mais determinística os movimentos do mercado. Eles são divididos em duas categorias principais: indicadores de tendência (média móvel etc.), e osciladores (índice de força relativa, estocástico etc.).

2.3.1.1 Indicadores Análise Técnica

O analista técnico examina as flutuações dos preços das ações por meio da análise de gráficos que retratam os preços históricos de mercado e indicadores. Diversos estudos utilizam indicadores técnicos no pré-processamento dos dados históricos, como Carosia, (2023); Coelho (2020); Nabipour et al. (2020); Salvi, Souza e Branco Neto (2020); Santos (2020); Pauli, Kleina e Bonat (2020); Vasco (2020); Nti, Adekoya e Weyori (2020b).

Os indicadores da análise técnica buscam determinar os movimentos do mercado para identificar o momento mais oportuno para a compra e venda de ações. Eles dividem-se em indicadores de tendência e osciladores que são: Média Móvel Simples (SMA); Média Móvel Exponencial (EMA); Regras de convergência/divergência de média móvel (MACD); Índice de Força Relativa (RSI); Volume no Balanço (OBV); Estocástico %D (%D); Estocástico %K (%K), os quais serão explicados a seguir de acordo com (PINHEIRO, 2019).

Dentro dos indicadores de tendência temos as médias móveis que são utilizadas para identificar flutuações no mercado de ações. A Média Móvel Simples (SMA) (*Simple Moving Average)* é uma média aritmética: soma-se as cotações de um determinado período e divide-se pelo número dos períodos considerados, verificando-se, assim, a média dessas cotações no período estudado. Sua representação está na equação (01):

$$\frac{SMA(n) \;=\; valor(1) \;+\; valor(2) + \ldots + valor(n)}{n} \qquad (01)$$

A Média Móvel Exponencial (EMA) (*Exponential Moving Average*) é um cálculo mais complexo que combina as vantagens das médias simples e ponderada, dando prioridade às cotações mais recentes em relação às mais antigas. Sua fórmula é apresentada na equação (02).

$$EMAx \;=\; EM(x-1) \;+\; K\,x\,\{Fech(x) - EM(x-1)\} \qquad (02)$$

As regras de convergência/divergência de média móvel (MACD) (*Moving Average Convergence/Divergence*), permitem calcular uma média móvel por pontos e sinais de cruzamento e divergência, prevendo assim tendências de baixa e alta, o que pode ajudar a melhorar as operações da empresa. Sua fórmula está na equação (03).

$$Linha\ MACD \;=\; MME\ de\ 12\ períodos - MME\ de\ 26\ períodos \qquad (03)$$

O Índice de Força Relativa (RSI) (*relative strength index*) facilita a identificação dos melhores pontos de venda ou compra de ações. A equação (04) apresenta a fórmula para calcular esse indicador de oscilação.

$$Rsi = 100 - \left\{ \frac{100}{1 + \frac{AU}{AD}} \right\} \quad (04)$$

Na equação 04, considerar que: AU (*average up*) é a Média dos incrementos de preços de fechamento em relação ao pregão anterior; AD (*average dow*) é a Média dos decréscimos dos preços de fechamento. Quando o índice RSI é superior a 70, o ativo está "supercomprado" (*overbought*), indicando momento de venda; já valores de RSI abaixo de 30 indicam uma condição de "supervendida (*oversold*), sugerido oportunidade de compra.

O Volume no Balanço (OBV) calcula o saldo acumulado do volume de compras e vendas no mercado de ações. Há três fórmulas para calcular o volume no balanço, considerando o preço de fechamento da ação no dia atual.

1. Se o preço fechamento do dia atual for maior que o do dia anterior tem-se a seguinte equação (05):

$$\text{OBV atual} = \text{OBV anterior} + \text{volume de hoje} \quad (05)$$

2. Se o preço de fechamento do dia atual for inferior ao preço de fechamento anterior, tem-se a equação (06).

$$\text{OBV atual} = \text{OBV anterior} - \text{volume de hoje} \quad (06)$$

3. E se o preço de fechamento do dia atual for igual ao preço de fechamento do dia anterior, utiliza-se a equação (07).

$$\text{OBV atual} = \text{OBV anterior} \quad (07)$$

O indicador Estocástico %K (*Stochastic*) mostra a relação do preço de fechamento em comparação à faixa formada entre os preços mínimos e máximos de um período, como descrito na equação (08).

$$\%K = \frac{(U - B)}{(A - B)} \; X\, 100 \quad (08)$$

Na equação 8, U representa a última cotação; A indica a cotação mais alta do período; e B refere-se à cotação mais baixa do período. O índice %K varia entre 0 e 100; valor de %K > 80 indicam sinais de venda, enquanto %K < 20 sinais de compra.

O Estocástico %D é a média móvel simples da curva %K, normalmente calculada com um período de três dias, como indicado na equação (09).

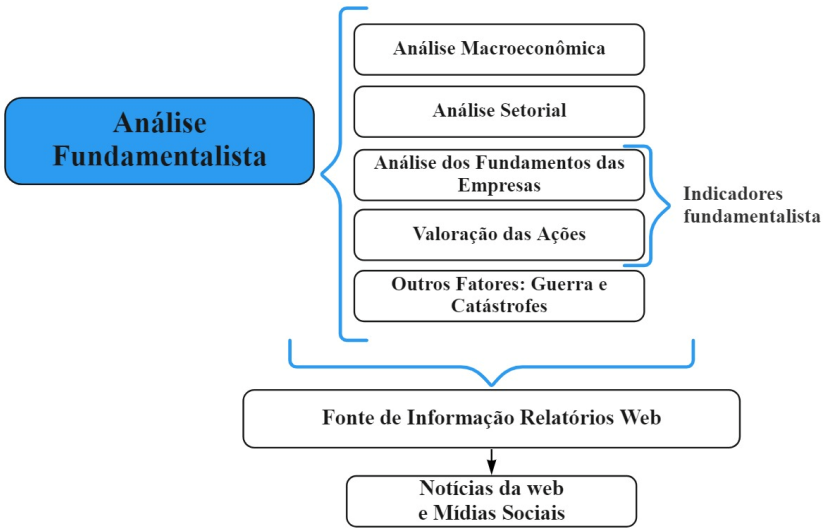
$$\%D = MMA\ (\%K, X) \tag{09}$$

Na subseção a seguir, apresentam-se os conceitos e embasamentos da análise fundamentalista, e os principais indicadores utilizados pelos analistas fundamentalistas.

2.3.2 Análise Fundamentalista

A análise fundamentalista examina os resultados financeiros e econômicos das empresas emissoras para avaliar o valor intrínseco de suas ações. Este processo envolve a análise de variáveis internas, como eficiência operacional, e externas, como condições econômicas gerais, que influenciam o desempenho da empresa e o valor de mercado de suas ações (ASSAF NETO, 2021). Na Figura 5 apresenta-se a descrição da análise fundamentalista.

Figura 5: Descrição da Análise Fundamentalista



Fonte: Elaborado pela autora.

Para realizar essa avaliação, são utilizadas informações financeiras detalhadas, como balanços e demonstrações de resultados. A Figura 5, que ilustra a descrição da análise fundamentalista, demostra como a abordagem integra análises macroeconômicas e setoriais, além da avaliação dos fundamentos das empresas e dos fatores de risco, para uma valoração precisa das ações. Os fatores de risco que afetam a valorização, como catástrofes naturais, crises

e mudanças políticas, entre outros. Essas informações são geralmente acessíveis em relatórios financeiros disponíveis nos sites das empresas e em fontes de notícias econômicas.

Para tomar decisões de investimento, os investidores consultam diversas fontes de informações, incluindo notícias, mídias sociais e análise de relatório. Esse acesso à informação, no entanto, é desigual, o que gera assimetria de informação entre os investidores. Mesmo assim, eles tomam decisões baseadas em seus próprios julgamentos, até que novos fatores os levem a rever suas posições (LEMOS, 2022).

Atualmente, as pessoas têm fácil acesso à tecnologia e os meios de comunicação para troca de informações; contudo, devido à assimetria de informações, é difícil que todos este a par das mesmas notícias ao mesmo tempo. Assim, raramente os investidores chegaram às mesmas conclusões sobre os efeitos dessas informações no valor de uma ação ou ativo financeiro. Portanto, eles fazem escolhas financeiras e se comprometem com elas até que algum novo fator os leve a mudar de opinião, alterando sua participação acionária, seja vendendo ações de uma empresa ou adquirindo de outras (LEMOS, 2022).

No processo de tomada decisão sobre compra ou venda de uma ação, a análise fundamentalista utiliza os métodos *top down* (de cima para baixo) e *bottom up* (de baixo para cima). Esses métodos diferem, basicamente, na classificação da importância dos fatos que afetam o valor das empresas. Na análise *top down*, discorrem que o que move a bolsa no longo prazo são as principais variáveis macroeconômicas. Ou seja, parte-se de um contexto global (análise macroeconômica, setorial e fundamentos das empresas) para formular uma conclusão sobre qual empresa recomendar. A análise *bottom up*, por sua vez foca no comportamento e desempenho de cada empresa para apresentar melhores oportunidades de investimento. Dessa forma, ambos os métodos são fundamentais para comprovar as conclusões (PINHEIRO, 2019; PÓVOA, 2012). A seguir, podem ser observados os impactos da análise macroeconômica e setorial no comportamento do mercado de capitais.

2.3.2.1 Análise Macroeconômica e Setorial

Nesta subseção, abordaremos os diversos fatores relacionados à análise macroeconômica e setorial que podem interferir nas negociações de mercado de uma empresa. Para analisar o desenvolvimento da economia de um país, são mensurados os indicadores econômicos das atividades econômicas gerais de um dado sistema econômico, utilizando-se das

principais variáveis macroeconômicas para verificar o comportamento da economia e de seus agentes.

Para Pinheiro, (2019), a bolsa de valores reflete, de algumas formas, o desempenho da economia na qual está inserida, sendo, portanto, fundamental a análise macroeconomia. Quando a economia está em um cenário positivo, as empresas tendem a apresentar um bom desempenho, o que leva à valorização de suas ações e, consequentemente, à alta da bolsa. Esse ponto de vista é relevante também para administradores de carteiras e analistas generalistas. Assim, a macroeconomia constitui o ambiente em que as empresas operam, e saber antecipar seus movimentos pode proporcionar grandes oportunidades de ganhos nos investimentos em ações.

No Quadro 2, observam-se as variáveis macroeconômicas que influenciam o comportamento do mercado de capitais, representadas por indicadores econômicos como: crescimento econômico, produção industrial, lucros empresariais, oferta de moeda, taxa de juros, taxa de câmbio, gasto públicos, inflação e desemprego.

Quadro 2: Exemplos de impactos de variáveis macroeconômicas nas ações.

Variáveis econômicas	Impacto no mercado de ações
Crescimento econômico	• Seu impacto é positivo e bom para o mercado, já que está correlacionado positivamente com a valorização das ações. • Os ciclos bursáteis tendem a acompanhar os ciclos econômicos. • A internacionalização das empresas cotadas atenua a conexão economia nacional – bolsa nacional.
Produção industrial	Aumentos contínuos são um sinal de força, o que é bom para o mercado.
Lucros empresariais	Grandes lucros são bons para o mercado.
Oferta de moeda	O crescimento moderado pode ter um impacto positivo na economia e no mercado. Mas o rápido crescimento é inflacionário e, portanto, prejudicial para o mercado de ações.
Taxa de juros	É influenciada pela política monetária. As taxas de juros determinam o movimento das cotações da bolsa, isto é, quando sobem, as cotações tendem a cair, e vice-versa. Representa um depressor, já que as taxas crescentes tendem a ter efeito negativo no mercado de ações. Apesar de a maioria dos analistas considerar a taxa de juros como um dos mais significativos determinantes de tendência primária do mercado de ações, deve-se levar em consideração que existem outras variáveis que também podem determinar o comportamento do mercado acionário, e esse *gap* resulta no prêmio de risco que o mercado oferece.
Taxa de câmbio	As cotações de uma bolsa valem em outros países o valor que resultar do câmbio e, por isso, valorizam ou desvalorizam-se com as respectivas moedas. Além da valorização ou desvalorização inerente a um mercado, deve-se considerar o risco cambial dos mercados em que são feitos os investimentos.
Gasto público	Em princípio, quanto maior for o déficit público, mais o Estado terá de recorrer ao financiamento. Com o aumento de procura, as taxas de juros da dívida pública colocada no país tendem a subir; em consequência, as demais taxas também tendem a subir. Superávits são bons para as taxas de juros e os preços das ações. Já os déficits podem causar inflação.

Inflação	O ambiente inflacionário, no curto prazo, não tem, por si só, impacto nas cotações. Porém, a influência indireta negativa nas taxas de juros torna o ambiente negativo para os mercados de ações. Em detrimento dos preços das ações, a inflação mais alta leva a taxas de juros mais altas e menores multiplicadores preço/lucro e geralmente torna as ações menos atrativas.
Desemprego	É um depressor, já que um aumento do desemprego significa que os negócios estão começando a desacelerar.

Fonte: Pinheiro (2019, p.470).

Desta forma, é importante destacar que as variáveis macroeconômicas estão inter-relacionadas, não sendo possível isolar seus efeitos sobre a bolsa. Deve-se sempre se considerar a conjuntura econômica internacional, pois, com processo de globalização, as economias estão cada vezes mais conectadas, e as bolsas muitas vezes apresentam comportamentos similares (PINHEIRO, 2019).

As ações de políticas macroeconômicas afetam tanto pessoas quanto empresas. As decisões do FED (Federal Reserve – banco central norte-americano), influenciam o mundo todo, inclusive as ações dos demais bancos centrais, como o brasileiro. Por exemplo, se o FED reduz a taxa de juros, investidores americanos e estrangeiros que investem nos EUA podem retirar seus recursos do país em busca de melhores taxas em outros países, como o Brasil. Isso aumenta a entrada de dólares, o que, por sua vez, impacta a economia brasileira ao influenciar variáveis como, taxa de câmbio, inflação, taxa de juros, balança comercial, entre outros (KOBORI, 2011).

Após a análise macroeconômica, são avaliadas as implicações dessa análise para os setores específicos de cada empresa. A análise setorial abrange os agrupamentos setoriais e investiga a posição competitiva de um setor em relação a outros, considerando aspectos como: “regulamentação e aspectos legais; ciclos de vida do setor; estrutura da oferta e exposição à concorrência estrangeira; sensibilidade à evolução da economia: setores cíclicos, acíclicos e contracíclicos; exposição a oscilações de preços; e tendências a curto e médio prazo”, além de fatores gerais do setor, oportunidade e riscos (PINHEIRO, 2019, p. 477).

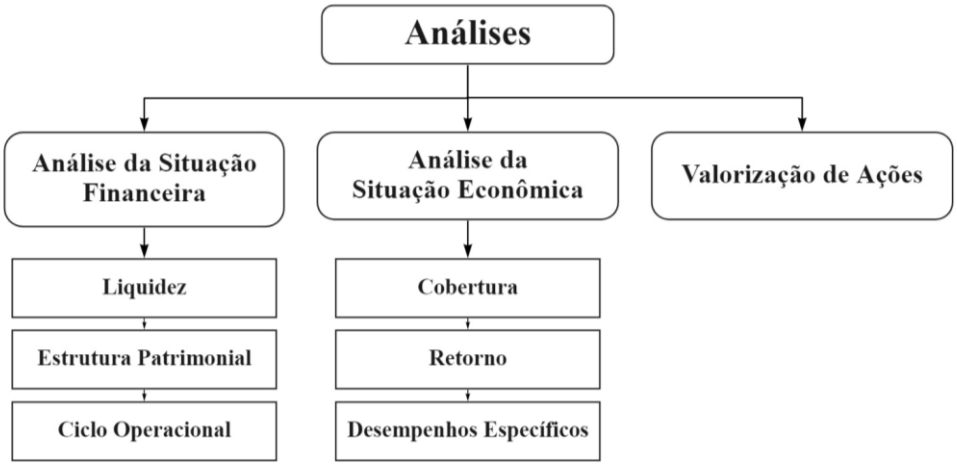
Um setor é um grupo de empresas com estruturas e produções semelhantes ou que oferecem bens e serviços que competem pelas preferencias dos consumidores. A análise do ambiente setorial avalia o estágio de crescimento do setor, se é um setor maduro, em declínio ou em expansão. A rentabilidades das empresas depende dessas características. Por exemplo, o setor da indústria de fumo já atingiu sua maturidade e está em declínio, com expectativa de redução de consumidores devido a intensas campanhas de desestímulo promovidas pela sociedade. Em contrapartida, a indústria de energia renovável está em forte expansão e deverá

crescer continuar crescendo nos próximos anos (KOBORI, 2011). A seguir, observam-se os principais indicadores utilizados pela análise fundamentalista para examinar o comportamento das empresas listadas nas bolsas de valores.

2.3.2.2 Indicadores Fundamentalista

Os indicadores da análise fundamentalista relacionados ao desempenho da uma empresa e valoração das ações baseiam-se em três princípios, conforme apresentado na Figura 6. O primeiro princípio é a análise da situação financeira, o segundo trata da análise da situação econômica da empresa, e o terceiro foca na valorização das ações. Os dois primeiros princípios são avaliados por meio de uma análise dos dados financeiros acumulados ao longo do tempo de operação da empresa, permitindo comparações entre dados de balanços de períodos anteriores para averiguar a saúde financeira da empresa. Já na valorização das ações, a abordagem consiste na comparação entre o valor calculado para a empresa e seu valor de mercado.

Figura 6: Indicadores da Análise Fundamentalista



Fonte: Adaptado, Pinheiro, 2019.

A análise da situação financeira de uma empresa busca avaliar sua capacidade de liquidar compromissos tanto no presente quanto no futuro, considerando a relação entre informações patrimoniais que evidenciam sua liquidez, estrutura patrimonial e ciclo operacional da empresa (DEBASTIANI e RUSSO, 2008; PINHEIRO, 2019).

A análise de liquidez examina as condições financeiras da empresa para honrar seus compromissos financeiros no curto e longo prazo, buscando um equilíbrio entre os ativos e passivos para evitar inadimplência. Essa análise é dividida entre os seguintes níveis: liquidez

global (mediata e estimada), liquidez de curto prazo (corrente, seca e imediata) e análise do capital circulante líquido (volume de recursos permanentes aplicados a curto prazo e liquidez dos recursos permanentes próprios) (PINHEIRO, 2019).

A análise da estrutura patrimonial objetiva identificar a evolução percentual dos diversos itens que compõem os ativos e passivos em relação ao total do ativo. Dessa forma, observa-se quantos pontos percentuais esses itens representam no valor total do ativo e como variam ao longo do tempo. Esse estudo ode ser dividido em duas etapas: a primeira é a análise da estrutura de capitais (participação do exigível a curto prazo e longo prazo, participação do exigível total e composição do passível exigível, imobilização de recursos permanentes, imobilização de recursos próprios e de terceiros a longo prazo) (DEBASTIANI e RUSSO, 2008).

A segunda etapa é a análise da estrutura de imobilização, que avalia a independência e dependência financeira dos recursos investidos, além de examinar o ativo imobilizado (giro do imobilizado, vida útil, nível de automatização, produção por imobilizado, grau de comercialização e retorno da produção). Já a análise do ciclo operacional da empresa identifica a velocidade do giro de caixa em cenários de dificuldade financeira, dividida entre o ciclo econômico (tempo entre a aquisição de matérias-primas e a vendas de produtos) e ciclo financeiro (período entre o desembolso para compras e o recebimento das vendas) (PINHEIRO, 2019). O quadro 3 destaca os principais indicadores usados para avaliar a lucratividade da empresa.

Quadro 3: Indicadores de valorização de ações.

Abreviação	Fórmula
D.Y	Dividendos pagos no período / Preço ação
P/L	Preço atual / Lucro por ação
P/VPA	Preço atual / Valor patrimonial por ação (VPA)
VPA	Patrimônio líquido / Nº de ações
P/Ativo	Preço atual / Ativos
LPA	Lucro Líquido / Nº de ações
P/SR	Preço atual / Receita líquida por ação
P/CAP.GIRO	Preço atual / (Ativo circulante – Passivo Circulante)

Fonte: Pinheiro (2019).

Os indicadores fundamentalistas que analisam a situação econômica da empresa têm como objetivo, mensurar, avaliar e interpretar os lucros ou prejuízos gerados, comparando-os com os recursos aplicados. Para isso, são calculados índices de valorização, que estão detalhados no quadro 3.

O *Dividend Yield* (D.Y), que apresenta a relação entre os rendimentos obtidos de um ativo financeiro e seu preço de mercado. O Preço/Lucro (P/L) é amplamente utilizado pelos investidores para analisar as ações, pois mostra a relação entre os preços pagos por uma ação e o lucro que ela oferece (ASSAF NETO, 2021; PINHEIRO, 2019).

O Preço/Valor Patrimonial por ação (P/VPA) mostra a relação entre o preço de mercado da empresa e seu valor patrimonial, enquanto o Valor Patrimonial por Ação (VPA) apresenta a relação entre o preço da ação e o seu patrimônio líquido. O Preço da Ação /Total de Ativos (P/Ativo) mede a avaliação de uma empresa no mercado em relação ao total de seus ativos. O Lucro por Ação (LPA) indica quanto a empresa gera de lucro em relação ao número de ações em circulação. Já o Preço Atual/Receita (P/SR) compara o preço atual da ação à receita líquida por ação, e o Preço/Capital de Giro (P/CAP.GIRO) relaciona o preço da ação ao capital de giro da empresa (DEBASTIANI e RUSSO, 2008; PINHEIRO, 2019).

Além das análises mencionadas, a análise fundamentalista também considera o impacto de outros fatores de riscos na economia de um país, como catástrofe naturais, guerras e acidentes. Esses riscos sistemáticos e não sistemáticos podem afetar a valorização das ações de uma ou mais empresas.

Segundo Debastiani e Russo (2008), é crucial atentar para os riscos decorrentes de medidas governamentais, indicadores macroeconômicos com inflação e taxa de juros, além de manobras políticas ou comerciais. Contudo, muitas vezes esquecemos que há uma classe de riscos devastadores para economia de uma empresa ou país, que não apresentam sinais claros de informação que possam indicar probabilidade ou impacto no mercado. Exemplos incluem catástrofes naturais, acidentes, atentados terroristas, conflitos armados e epidemias, cuja ocorrência varia em frequência e intensidade.

Esses riscos podem ser classificados como riscos sistemáticos ou individuais, dependendo de sua natureza e extensão. Um exemplo é a epidemia da “vaca loca” que surgiu na década de 1990 no rebanho bovino da Inglaterra, levando os consumidores a deixarem de comprar carne europeia. Por outro lado, o mercado brasileiro se beneficiou deste evento, em um aumento da demanda por carne (DEBASTIANI e RUSSO, 2008).

Portanto, a análise macroeconômica, setorial e dos fundamentos de uma determinada empresa auxilia os analistas fundamentalistas a determinarem um “valor justo” para ela. Com os avanços das informações ao longo do tempo e a propagação na internet, o acesso às informações tornou-se uma *commodity*. Atualmente, a capacidade de processamento da informação possui um valor inestimável, influenciando as decisões e distanciando o futuro do

passado a cada nova informação e dado (PÓVOA, 2012). A próxima seção discutirá as diferentes fontes de informações textuais que afetam o movimento do preço das ações.

2.4 Fontes de Informações Textuais

Nesta seção, são relacionados e discutidos os diferentes tipos de informações textuais disponíveis atualmente. Esta análise é importante para a proposta desta tese, pois há um aumento significativo na quantidade e diversidade de material textual na Web que pode estar associada a aspectos do movimento do preço das ações e das tomadas de decisões dos investidores.

A análise fundamentalista baseia-se em fontes de dados numéricos e textuais, como já descrito. Os analistas coletam informações sobre a empresa, para isso, devem conhecer as variáveis relacionadas ao seu desenvolvimento da empresa. As principais fontes de pesquisa utilizadas, segundo Pinheiro (2019, p.460) são:

a) Entrevistas pessoais com dirigentes das empresas, representantes de associações de classe (patronal e empregatícia), clientes e público interno.
b) Leitura de relatórios macroeconômicos; relatórios setoriais; relatórios de administração; demonstrações contábeis publicadas; informações em *real time* (CNN, *broadcasting* etc.); artigos publicados na imprensa especializada; e publicações especiais.
c) Base de dados disponíveis na internet, bolsa de valores e corretoras de valores, e outros.

Estas principais fontes apresentadas são extraídas de relatórios econômico-financeiro da empresa divulgado trimestralmente, semestralmente ou anualmente pelas empresas listadas na Bolsa de Valores, que apresenta as demonstrações financeiras da empresa. Através dos valores apresentados nessas demonstrações, são calculados os indicadores da análise fundamentalista sobre o desempenho da empresa. Esses relatórios são disponibilizados na internet, nos *sites* das empresas e pela bolsa de valores, entre outras instituições relacionadas.

Os relatórios macroeconômicos contêm dados numéricos dos indicadores econômicos para verificar o comportamento da economia do país e são apresentados no formato textual através de notícias, enquanto os relatórios setoriais trazem tanto dados numéricos quanto textuais. Todas essas informações são disponíveis na internet, nos *sites* de instituições oficiais. Atualmente, muitos *sites* especializados em economia e análise de mercado, administrados por profissionais adequados, também se tornaram fontes ricas de notícias de opiniões diversas sobre o mesmo tema. Com o avanço das mídias sociais, muitos desses profissionais também utilizam

essas plataformas para disseminar informações. No entanto, os investidores devem sempre confrontar essas opiniões com outros pontos de vista e fatores discutidos em seções anteriores para uma melhor tomadas de decisões. Assim, com o aumento do volume de informações na web, diversos estudos começaram a extrair indicadores (sentimentos e eventos) da internet para facilitar e até mesmo aprimora as previsões de movimentos de ações.

A Tabela 1 mostra as mídias sociais mais usadas no mundo e no Brasil em fevereiro 2022 e a Figura 7 indica os principais motores de busca.

Tabela 1: Mídias Sociais mais usadas no Mundo e no Brasil em fevereiro 2022.

Mídias Sociais	**N°. Usuários**	
	Mundo	**Brasil**
Facebook	2.910	148
YouTube	2.562	138
WhatsApp	2.000	120
Instagram	1.478	119,5
TikTok	1.000	74,07
Facebook Messenger	988	65,5
LinkedIn	744	52
Pinterest	444	27
Twitter	436	19,05
Snapchat	557	8,5

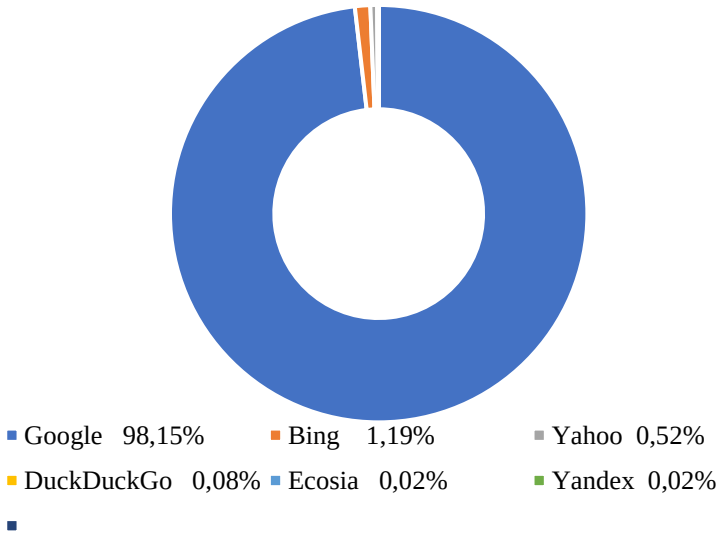
Fonte: Statista (2022).

De acordo com os dados da Statista (2022), a rede social líder é o Facebook, com 2.910 bilhões de usuários no mundo e 148 milhões no Brasil, sendo este o país da América Latina com maior presença na rede social, caracterizado pela interação e expansão de contatos. Em segundo lugar está o YouTube, com 2.562 bilhões de usuários no mundo e 138 milhões no Brasil, utilizada principalmente para compartilhamento de vídeos. O WhatsApp, uma rede social de mensagens instantânea e chamadas de voz, é a terceira mais utilizada no mundo, com 2.000 bilhões de usuários, e conta com 120 milhões de usuários no Brasil. Em seguida, vêm as redes sociais Instagram, TikTok, Facebook Messenger, LinkedIn; Pinterest, Twitter e Snapchat, conforme mostrado na Tabela 2.

A Figura 7 mostra os motores de busca mais utilizados no Brasil em fevereiro de 2022, segundo dados Statcounter Globalstats (2022). O Google lidera com 98,15%, participação entre os buscadores mais utilizados no Brasil. Em segundo lugar está o Bing, com 1,19% da preferência dos usuários, seguido de Yahoo, com 0,52%, DuckDuckGo, com 0,8%, e Ecosia e Yandex, com 0,02%.

Segundo a DataReportal (2022), em janeiro de 2022 havia 165,3 milhões de usuários na *internet* no Brasil, com uma taxa de penetração de 77% da população total no início de 2022. Esse número representa um aumento de 5,3 milhões em relação a 2021. Considerando a população brasileira de 214,7 milhões, estima-se que 49,37 milhões de pessoas não utilizavam a internet no início de 2022, indicando que cerca de 23% da população estava *offline*.

Figura 7: Motores de Busca mais usados no Brasil em fevereiro 2022.



Fonte: Statcounter Globalstats (2022).

Nos anos de 2020 e 2021, com a pandemia de Covid-19, percebeu-se uma mudança nos hábitos digitais das pessoas em todo o mundo, o que contribuiu para aumento no consumo de informações em diversos formatos. No Brasil, os usuários passam, em média, 3 horas e 42 minutos por dia conectados às redes sociais. A subseção a seguir expõem os tipos essenciais de dados financeiro.

2.4.1 Tipos de dados Financeiros

De acordo com De Prado (2018), há vários tipos de dados financeiros utilizados no mercado para análise, previsão e tomada de decisões. Esses dados podem ser classificados em quatros tipos essenciais, conforme apresentado no Quadro 4:

- Dados Fundamentais: Os dados fundamentais incluem principalmente informações contábeis divulgadas trimestralmente, oferecendo *insights* sobre a saúde financeira e o desempenho das empresas, como receita, lucro, fluxo de caixa e balanço patrimonial.

Também fazem parte dessa categoria variáveis macroeconômicas, como o produto interno bruto (PIB), taxas de juros, inflação e outros indicadores de mercado.

- Dados de Mercado: Os dados de mercado abrangem todas as atividades de negociação realizadas em uma bolsa ou local de negociação. Incluem informações como preços negociados, volume de negócios, dividendos ou gratificações pagas aos acionistas, contratos em aberto que mostram os interesses dos participantes, ordens de compra e venda, cancelamento de ordens e o lado agressor, ou seja, qual parte da negociação executou a transação.

Quadro 4: Os quatro tipos essenciais de dados financeiros.

Dados Fundamentais	Dados de Mercados	Dados Alternativos	Análises
• Ativos • Passivos • Vendas • Custos/lucros • Variável macro	• Precos/ rendimento / volatilidade implícita • volume • dividendo/cupoes • interesse aberto • ordens de compra ou venda/cancelamentos de ordens • lado agressor	• Imagens de satélite/CCTV • Pesquisas no Google • Twitter/chats • Metadados	• Recomendações de analistas • Classificações de crédito • Expectativas de lucros • Sentimento de notícias

Fonte: De Prado (2018, p.24).

- Dados Alternativos: Dados alternativos referem-se a informações provenientes de fontes não tradicionais, como mídias sociais, notícias, pesquisas na web, imagens de satélite, sistemas de videovigilância, tráfego de voos e contêineres, entre outros. Esses dados são considerados informações primárias, pois ainda não foram incorporados por outras fontes de dados. Por exemplo, no caso de uma empresa de petróleo, antes que esta relate seus ganhos e o mercado reaja positivamente, é possível capturar sinais antecipados de atividade, como o movimento de petroleiros, perfuradores e tráfego de oleodutos, sinais que podem surgir meses antes de serem refletidos nos dados financeiros oficiais. Segundo Jansen (2020), os dados alternativos podem ser classificados pela sua origem:

  - Indivíduos: dados de redes sociais, avaliações de produtos e pesquisas na web.

  - Empresas: dados de transações com cartões de crédito, atividades da cadeia de suprimentos e seus intermediários.

  - Sensores: dados capturados pela atividade econômica, como imagens de satélite e câmeras de segurança.

- Análises: As análises são de dados derivados dos dados fundamentais, de mercado, alternativos e de suas combinações. Empresas de pesquisa e bancos de investimento costumam vender relatórios baseados em análises detalhadas de modelos de negócios, atividades, concorrência e perspectivas das empresas, com recomendações de compra ou venda. Além disso, essas análises podem incluir estatísticas extraídas de dados alternativos, como sentimentos em notícias e redes sociais.

Este estudo tem como objetivo utilizar várias fontes de dados para prever o preço das ações, avaliando as contribuições de cada variável no desempenho do modelo e identificando as que apresentam melhor precisão. Para isso, será realizada a combinação dos quatro tipos de dados classificados por De Prado (2018). A subseção a seguir apresenta os conceitos de análise de sentimentos.

2.4.2 Análise de Sentimentos

A análise de sentimentos é um método utilizado para mensurar as atitudes, opiniões e emoções manifestadas por alguém sobre um determinado assunto. Neste caso é identificado o sentimento através dos textos, sendo detectados aspectos textuais que expressam um sentimento positivo, negativo ou neutro. O sentimento pode ser geral ou sobre um assunto específico, tal como um indivíduo, um produto ou um evento (CHEN et al., 2018; ROSENTHAL, FARRA e NAKOV, 2019).

A análise de sentimento utiliza-se das técnicas de mineração de texto, processamento de linguagem natural e técnicas computacionais para extrair automaticamente sentimentos de um texto. Seu principal foco é a classificação da polaridade de determinado texto em nível de sentença ou nível de classe, verificando se este reflete uma visão positiva, negativa ou neutra. No trabalho de previsão do mercado de ações, geralmente são usadas duas fontes de textos que são importantes, sendo estas as mídias sociais (usando dados principalmente do *Twitter*) e os artigos de notícias financeiras online (ALZAZAH e CHENG, 2021).

Apesar dos avanços dos usos de dados textuais, ainda é desafiador lidar automaticamente com dados não estruturados e extrair deles atributos úteis não é comum. Assim, pesquisadores observaram a importância da mineração de opiniões sobre o preço do mercado de ações, pois identificaram uma forte relação entre as notícias sobre uma empresa e as flutuações futura nos preços das ações (ARAÚJO e FERNANDES, 2021; JOSHI, BHARATHI e RAO, 2016).

Muitos pesquisadores estudaram os movimentos do mercado de ações usando como fonte de dados do Twitter, por ser uma fonte de dados significativa e que se tornou muito popular entre os usuários na internet como uma ferramenta de comunicação. O Twitter foi criado em 2006 e é uma mídia social em formato de *microblog* que permite ao usuário enviar e receber atualizações pessoais de seguidores e outros contatos. As atualizações no perfil são exibidas em tempo real. Inicialmente seu foco é permitir compartilhamentos de ações pessoais, mas atualmente também é usado para discussões no âmbito profissional, questionamentos de assuntos da atualidade, marketing, entre outros (PAK e PAROUBEK, 2010; ZENHA, 2018).

Segundo Zenha, (2018), as principais razões para utilização do Twitter são associadas com o fato de que esta ferramenta é utilizada por diferentes usuários para expressar suas opiniões sobre diferentes assuntos. Desta forma, ele contém grande quantidade de textos de opinião que cresce a cada dia. Os usuários são diversos e variam desde celebridades, presidente da república, representantes de empresas, entre outros usuários. Existe usuários de diferentes países, permitindo coletar dados de diferentes linguagens.

Embora cada tweet seja restrito a 280 caracteres, acredita-se que a informação possa refletir na precisão o humor dos usuários. Os textos extraídos são comumente tratados como problema de classificação (supervisionado) e são categorizados. Os estudos utilizam da API do Twitter e combinando os *cashtags* da empresa no conteúdo do tweet. O *Cashtag* é uma nova forma de compartilhamento de informações financeiras nas mídias sociais desenvolvida pelo Twitter e outros provedores, junto com o código de cotações de ações da empresa são prefixados com um cifrão para compor o *cashtag*, por exemplo, Apple=$AAPL, Google=$GOOG (SHI et al., 2019).

As notícias financeiras são consideradas uma fonte de dado mais consistente e confiável, pois a análise de notícias financeiras pode ajudar a prever os movimentos do mercado de ações, por exercer uma forte influência nas flutuações deste mercado. Os avanços das técnicas de Processamento de linguagem natural permitem a melhoria da capacidade de extrair eventos de artigos de notícias. No entanto existem algumas limitações na previsão, pois os eventos podem representar apenas uma pequena parte da volatilidade das ações, assim sendo insuficientes para estimar o mercado de ações sozinho.

A próxima seção apresenta as técnicas computacionais utilizadas na previsão do mercado de ações, inteligência artificial, aprendizado de máquina, aprendizado profundo e processamento de linguagem natural.

2.5 Inteligência Artificial

Para Rezende (2005), inteligência artificial (IA) simula as funções desempenhadas pelo ser humano usado o conhecimento e raciocínio, quando são habilitadas para o computador as executar, traduzidas em habilidade e capacidade de solucionar problemas complexos. Ou seja, é um processo de desenvolvimento e realizações de tarefas desempenhado por uma máquina. Portanto a IA simula o comportamento inteligência humano, através da utilização de algoritmos definidos por especialistas, capazes de identificar um problema, ou uma tarefa a ser realizada, analisar dados e até determinar tomada de decisões (LOBO, 2018).

Os sistemas computadorizados de suporte a tomada de decisões já existem há décadas. Estes sistemas estão aumentando sua capacidade nos últimos anos, com aumento da velocidade de processamento e de armazenamento de informações dos computadores, mas também com o acesso a um grande volume de dados, melhores algoritmos, adquirindo assim a capacidade de solucionar problemas mais complexos, sem orientação constante de um usuário e a capacidade de melhorar o desempenho aprendendo com novas experiências (LOBO, 2018). O Quadro 5 apresenta algumas definições de IA, organizadas em quatros categorias.

Quadro 5: Algumas definições de Inteligência Artificial.

Pensando como um humano	**Pensando racionalmente**
"O novo e interessante esforço para fazer os computadores pensarem (...) *máquinas com mentes*, no sentido total e literal." (Haugeland, 1985) "[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman,1978)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak e McDermott, 1985) "O estudo das computações que tornam possível perceber, raciocinar e agir." (Winston, 1992)
Agindo como seres humanos	**Agindo racionalmente**
"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas." (Kurzweil, 1990) "O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas." (Rich and Knight, 1991)	"Inteligência Computacional é o estudo do projeto de agentes inteligentes." (Poole *et al*., 1998) "AI... está relacionada a um desempenho inteligente de artefatos." (Nilsson, 1998)

Fonte: Russel e Norvig (2013, p. 25).

A IA também pode ser conhecida pelos campos de pesquisa com: aprendizado de máquina, ciência de dados e aprendizado profundo que são subcampos da IA, como pode ser observado nas subseções a seguir algumas técnicas tradicionais de Aprendizado de Máquina.

2.5.1 Aprendizado de Máquina

O termo Aprendizado de Máquina descreve sistema que melhoram seu desempenho em uma determinada tarefa com base em maior experiência ou uso de volume de dados, ou seja, o

agente melhora seu desempenho em tarefas futuras com base em experiência e observações do mundo (RUSSELL e NORVIG, 2013).

Os três principais tipos de aprendizados são: aprendizado supervisionado; aprendizado não supervisionado e aprendizado por reforço. No aprendizado supervisionado, o agente aprende uma função capaz de mapear entradas para saídas a partir de um conjunto de exemplos de entradas e saídas (RUSSELL e NORVIG, 2013).

No aprendizado não supervisionando, o agente busca aprender padrões em conjuntos de dados não rotulados, ou seja, padrões de entradas sem que haja o fornecimento dos valores que se deseja obter na saída, ou seja, valores desconhecidos, normalmente usados para gerar agrupamento. O aprendizado por reforço busca aprender um comportamento através de recompensas eventuais (recebidas ao interagir com ambiente) para atingir determinado objetivo, ou seja, tentativa e erro (RUSSELL e NORVIG, 2013).

As técnicas são escolhidas de acordo com as tarefas a serem desempenhadas. Podem ser utilizadas de maneira individual ou combinadas. São várias as técnicas que podem ser aplicadas no segmento financeiro. A seguir são comentadas técnicas associadas com este contexto.

A Rede Neural Artifical (RNAs) é um sistema de computação que baseada sua formação em uma estrutura de neurônio. Uma RNA é composta por três componentes: (i) camada de entrada, que recebe os dados brutos de entrada; (ii) uma ou mais camadas ocultas; (iii) camada de saída, que indicam a previsão final do sistema (PETERSON e RÖGNVALDSSON, 1991). A equação (10) mostra a implementação vetorizada de um neurônio.

$$Y = f(wk + b) \tag{10}$$

Na equação 10 o valor w representa um parâmetro – valores de peso da conexão do neurônio da camada anterior; b é valor de polarização e f é a função de ativação linear. Os neurônios são utilizados em camadas como uma unidade de ativação. Assim, essas unidades de ativação levam os dados da camada anterior, para serem processados e transmitidos os dados processados o para os neurônios da própria camada, onde uma unidade de ativação de uma camada está conectada a todas as outras unidades na próxima camada, com valores diferentes para cada conexão. Os valores de peso serão otimizados usando algoritmos de propagação direta e posterior (PETERSON e RÖGNVALDSSON, 1991).

As RNAs implementam técnicas de aprendizado de máquina. A camada oculta da rede neural captura os padrões de dados, e característica para estabelecer uma dinâmica complexa

relacionada a entrada e saída de variáveis não linear (RAY, KHANDELWAL e BARANIDHARAN, 2018).

De acordo com Gujarati e Porter (2011), a análise de regressão refere-se ao estudo da dependência de uma variável, a variável dependente em relação a uma ou mais variáveis, as variáveis explanatórias visam estimar e/ou prever o valor médio (da população) da primeira, em termos dos valores conhecidos ou fixados (em amostragens repetidas) das segundas.

Assim, o resultado deve ser bom para tornar pequenos os erros da previsão, ou seja, diferença entre o valor real e o valor projetado pelo algoritmo. De acordo com Sartoris (2003), suponhamos, ainda, que X é a variável independente e Y é a variável dependente, isto é, Y que é afetado por X e não o contrário. Uma função afim (linear), cada Y pode ser escrito em função de cada X da seguinte forma:

$$Y = \alpha + \beta X \quad + \varepsilon \tag{11}$$

sendo $\alpha + \beta X$ a equação da reta, e ε 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 predizer as saídas, trabalhando apenas com valores numéricos e os pesos são atualizados conforme a função que minimiza o erro.

Florestas aleatórias - *Random Forests* é um algoritmo supervisionado proposto por Breiman (2001), constrói dezenas de árvores combinadas para predizer o melhor resultado, pode ser utilizado para classificação ou regressão.

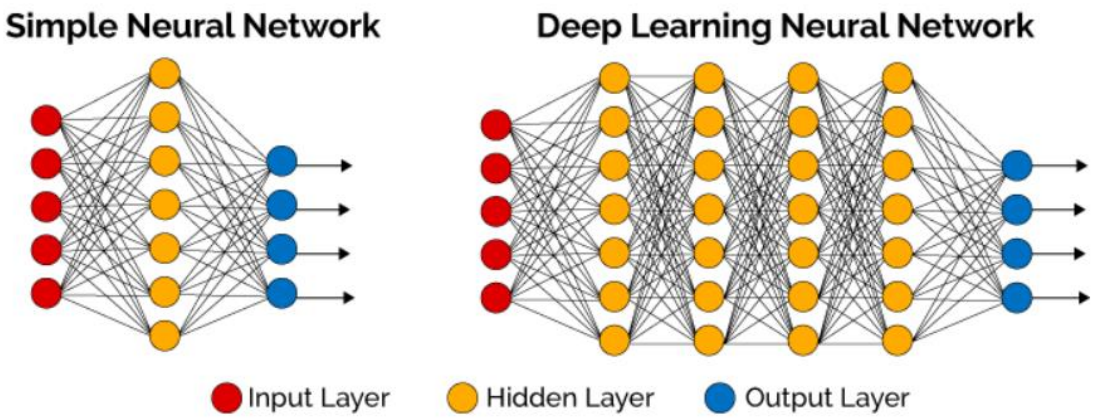
As seleções dos atributos são aleatórias, ao invés da seleção a partir do cálculo de impureza, e resolve o problema de *overfitting* da árvore de decisão. Suas vantagens são de maior robustez, menor propenso a sofre *overfitting* em comparação com uma única árvore de decisão, permite a descoberta de conhecimento e poucos parâmetros para ajustes. Suas desvantagens são por exigir um maior poder de processamento e pode ser lento processo de classificação de novas amostras (BREIMAN, 2001; RUSSELL e NORVIG, 2013).

2.5.2 Aprendizado Profundo

Para Kim, (2016), aprendizado profundo (*deep learning*) é uma configuração de aprendizado de máquina que permite os computadores aprenderem com suas experiências

anteriores e compreender de forma geral com uma hierarquia de conceitos. Assim, por extrair informações passadas, não é preciso fornecer todo o conhecimento ao modelo. Uma rede de aprendizado profundo se descobre no número de camadas ocultas que elas possuem, diferente da rede neural como pode ser observado na Figura 8 a seguir.

Figura 8: Diagrama de aprendizado profundo.



Fonte: Vázquez (2017).

As redes neurais recorrentes (RNN) são estruturadas por neurônios cujas saídas realimentam a si própria e os outros neurônios de maneira a formar ciclos, por conter *loops*, deixa que ele tenha memória e explore a capacidade de armazenamento de informações temporais e sinais sequenciais. Uma deficiência enfrentada pelas redes neurais recorrentes é desaparecimento do gradiente, as redes com muitas camadas perdem a informação do gradiente. Para solucionar este problema foi desenvolvida a estrutura de *Long Short Term Mermory* LSTM, foi a inserção de uma memória à NN para manter seu status durante vários passos de tempo (KIM, 2016).

O LSTM foi criado na década de 1990 para resolver problema de gradiente de desaparecimento que faz com que o aprendizado do modelo se torne muito lento ou pare. LSTM é modelo de rede neural recorrente eficiente que pode manter internamente a memória da entrada, sendo uma solução para os problemas que envolver dados sequenciais de série temporal. Assim, tem um memórias mais longas e podem aprender com as entradas que estão separadas uma das outra por muito tempo (YADAV, JHA e SHARAN, 2020).

2.5.3 Processamento de Linguagem Natural

Nos últimos anos com avanços tecnológico, o uso de computadores poderoso com a capacidade de manipulação e processamento de grande quantidade de dados, vem aumentando o uso de processamento de linguagem natural usando métodos de redes neurais profundas e aprendizado de máquina para relacionar os movimentos dos preços das ações e os sentimentos do mercado, através da integração de dados estruturado e não estruturado para melhor predição do modelo.

Processamento de Linguagem Natural (PLN) é uma área da computação que é utilizada para tratamento de informações digitais expressa em texto extraído representações e significado mais completo de textos livres em linguagem natural para processamento de dados textuais, é uma subárea da inteligência artificial e da linguística computacional (INDURKHYA; DAMERAU, 2010).

Segundo Gupta, (2014), Perera e Nand, (2017), é uma interação entre máquina e seres humanos, dentro da área de processamento de linguagem natural a uma área de estudo de geração de linguagem natural, que é responsável em produzir textos de linguagem natural com ótimo nível de entendimento e precisão de informações armazenadas em sistemas computacionais.

A análise linguística e o processamento de linguagem natural pode ser classificada em: pré-processamento (conversão do arquivo para formato de texto e segmentação do texto em unidade lexicais e sentenças) definindo os delimitadores textuais como espaços em brancos e vírgulas, assim, a divisão em unidades lexicais resultam os *tokens* do texto (tokenização) e em seguida é reconhecer as orações presente no textos; análise léxica é realizada a decomposição das unidades lexicais determinada no pré-processamento; análise sintática (organização da ordem e estrutura do texto); análise semântica (significado); e análise pragmática (análise do discurso). Está divisão é fundamental para gerenciamento mais adequado do processamento de analisar um texto (INDURKHYA; DAMERAU, 2010). No próximo capítulo mostra uma discussão dos trabalhos relacionados de previsão do mercado de ações.

3 TRABALHOS RELACIONADOS

Atualmente, diversos estudos exploram técnicas de IA para previsão do mercado acionário, embora ainda existam desafios a serem investigados. Este capítulo apresenta uma revisão de pesquisas que aplicam abordagens de análise técnica e fundamentalista na previsão do mercado de ações, visando compreender o estado atual da arte e identificar possíveis direcionamentos futuros.

Os estudos foram coletados por meio de uma revisão de literatura, com buscas realizadas nas bases de dados Portal de Periódicos das Capes (IEEE *Xplore*), *Scielo*, *Google* Acadêmico e *ScienceDirect*. Essas bases foram configuradas para selecionar textos dos seguintes tipos: artigos publicados em periódicos, artigos de anais e conferências, livros, dissertação de mestrado e teses de doutorado, além de trabalhos acadêmicos adicionais não publicados e relatórios. A seleção compreendeu publicações no período de 2017 a 2024, em qualquer idioma.

Na presente pesquisa, foram selecionadas publicações que continham as seguintes palavras-chave: “previsão de movimentos no mercado de ações”, “técnicas de Inteligência Artificial no mercado acionário”; “previsão de ações”. Os resultados foram classificados em três categorias: (i) análise técnica com dados numéricos, (ii) análise fundamentalista com dados textuais, (iii) combinação dos de dados numéricos e textuais. A seleção dos estudos para análise considerou critérios como a natureza dos conjuntos de dados, o tamanho do conjunto de dados, os algoritmos de aprendizado de máquina, aprendizado profundo e processamento de linguagem natural empregados, além das métricas de precisão adotadas nos estudos analisados.

Este capítulo está organizado em três seções. A primeira seção expõe a análise de estudos que utilizam exclusivamente dados numéricos para previsão do mercado de ações por meio de abordagens técnicas, aplicando algoritmos de aprendizado de máquina e aprendizado profundo. A segunda seção explora estudos focados na análise fundamentalista com dados não numéricos, incluindo combinações de dados e o uso de algoritmos de processamento de linguagem natural, além de técnicas de aprendizado de máquina e aprendizado profundo. Por fim, a terceira seção fornece uma análise comparativa e crítica dos estudos revisados.

3.1 Trabalhos de Análise Técnica

Esta seção apresenta os trabalhos relacionados de análise técnica, que realizam a previsão do comportamento futuro dos preços das ações estudando as tendências futuras através de séries históricas do mercado de ações, da natureza de dados estruturados ou quantitativos. A maioria dos métodos de previsão do mercado de ações geralmente utiliza da análise técnica,

tais como os estudos: Nelson (2017); Sismanoglu et al. (2019); Bou-Hamad e Jamali (2020); Castro, Reyes e Landazábal, (2020); Werner, Bisognin e Araujo (2020), Liu e Long (2020)

O Estudo de Nelson (2017) teve como objetivo estudar as técnicas de inteligência computacional para compreender e prever tendências de preços na Bolsa de valores brasileira. Foram utilizados dados de cinco ativos listados, incluindo preços de abertura, fechamento, máximo, mínimo, volume e 138 indicadores técnicos gerados pela TA-Lib. Esses dados foram analisados por meio de redes recorrentes LSTM e comparados com MLP e *Random Forest*. Observou-se que o modelo LSTM proposto apresentou desempenho superior aos *baselines,* com algumas exceções. Os resultados obtidos indicam uma acurácia média de até 55,9% ao prever se o preço dos ativos selecionados irá subir ou não no futuro imediato.

Sismanoglu et al. (2019), investigaram a possibilidade de prever o valor futuro de uma ação com base em informações históricas do mercado financeiro, usando uma abordagem de *Deep Learning*. Os dados incluíram ações da IBM, e informações diárias de preço e volume de todas as ações negociadas nas Bolsas NYSE, NASDAQ e NYSE MKT, no período de 2 de janeiro de 1968 a 9 de abril de 2018, totalizando 12.648 observações, analisadas com o modelo LSTM. O teste de previsão de 300 dias foi realizado com treinamento nos dados históricos e verificação da precisão através da raiz do erro quadrático médio (RMSE), comparando os valores reais das ações com os valores previstos. Os resultados mostraram que o modelo LSTM produziu previsões satisfatórias para ações especificas, capturando dinâmicas de mercado e produzindo previsões eficientes, com RMSE de 0.04.

Bou-Hamad e Jamali, (2020) investigaram a capacidade preditiva, estática e dinâmica, de rede neurais artificiais e florestas aleatórias para séries temporais financeiras, utilizando de uma simulação de dados com um modelo Autoregressivo (AR) cuja variação condicional segue um processo Heteroscedasticidade Condicional Autorregressiva Generalizada (GARCH). A previsão fora da amostra foi avaliada com um período de validação de 252 observações, equivalente a um ano de negociação diárias. Os dados utilizados foram Taxas de Câmbio do Dólar Australiano (AUD) e do Franco Suíço (CHF) em relação ao dólar americano no período de 16 de outubro de 1989 a 28 de março de 2019 (BOU-HAMAD e JAMALI, 2020). Os resultados mostraram que as técnicas de mineração de dados superaram os modelos AR e GARCH, especialmente sob um esquema de previsão dinâmica para séries com persistentes moderada a alta. Quanto maior o grau de persistência, mais evidente foi o desempenho superior das técnicas de mineração de dados, que produziram previsões mais precisas da dinâmica de retorno em comparação o modelo verdadeiro (BOU-HAMAD e JAMALI, 2020).

No trabalho de Castro, Reyes e Landazábal, (2020), os autores aplicaram as técnicas de aprendizado de máquina para previsão do desempenho da Bolsa de Valores do México (PQI) após a crise de 2008. O objetivo foi explorar os modelos tradicionais de aprendizado de máquina, como *perceptron* multicamada, processos Gaussianos, tabela de dados, árvore de decisão, máquina de vetores de suporte, avaliando seus desempenhos em comparação com a regressão linear (LR), comumente utilizada, e redes neurais (RN), amplamente utilizadas em finanças. O modelo foi construído usando variáveis macroeconômicas relevantes para o Índice de Preços e Cotações (PQI): *Index Dow* Jones Média Industrial (DJIA), Índice Nacional de Preços Consumidor (INPC), Índice de Preços ao Consumidor (IPC), Reservas Internacionais (RI), Rendimentos em Tesouraria Certificados (CETES28), o peso-dólar mexicano Taxa de Câmbio (USDMX), Agregado Monetário 1 (M1) e Risco Soberano do México (SDRM). O conjunto de dados é composto de preços mensais a partir de janeiro de 2009 a dezembro de 2016. Foram realizados dois casos de teste para cada modelo, utilizando a média absoluta percentual de erro (MAPE) para compará-los, onde foi realizada a previsão de um a cinco meses à frente. No primeiro caso, os processos Gaussianos apresentaram erros superiores a 4,0%, enquanto a regressão linear teve desempenho inferior. Em contrapartida, as árvores de decisão mostraram o melhor desempenho, com erros inferiores de 1,0%. No segundo caso, as porcentagens de erro foram inferiores a 2,0% para Tabela de Dados, enquanto a regressão linear teve as piores taxas, acima de 7,0%; com os processos gaussianos apresentando resultados semelhantes (CASTRO, REYES e LANDAZÁBAL, 2020).

Conforme Werner, Bisognin e Araujo (2020), diversas técnicas foram avaliadas para prever o volume diários das ações da Petrobras (PETR3) na B3, a bolsa de valores brasileira. Os dados utilizados foram o volume de negociações entre 04 de janeiro de 2010 a 18 de setembro de 2018. Utilizou-se o modelo *AutoRegressive Fractionally Integrated Moving Average* (ARFIMA), aliado à rede neural recorrente (RNN), para combinações de três abordagens de previsão: média aritmética, variância mínima e regressão. As métricas de acurácia aplicadas para análise da tomada de decisão incluíram RMSE, MAPE e U de *Theil*. A análise foi dividida em dois períodos. No primeiro, de 4 de janeiro de 2010 a 20 de setembro de 2016, a combinação de previsão por regressão linear apresentou melhor desempenho nas métricas RMSE e U de *Theil*, enquanto a média aritmética teve o menor MAPE. O segundo período, de 21 de setembro de 2016 a 18 de setembro de 2018, foi usado para validação dos modelos, e a combinação de previsões com média aritmética obteve as melhores medidas de acurácia (WERNER, BISOGNIN e ARAUJO, 2020).

Liu e Long (2020), desenvolveram um *framework* para prever o valor de fechamento diário de ações, estruturado em três fases: pré-processamento, predição e pós-processamento. No pré-processamento, técnica *empirical wavelet transform* (EWT) é utilizada para decompor a série temporal em subcamadas, facilitando a análise. Em seguida, um modelo LSTM com *dropout* é treinado para cada subcamada, com seus hiperparâmetros otimizados pelo particle swarm optimization (PSO). No pós-processamento, o método *outlier robust extreme learning* (ORELM) corrige o erro de previsão com vase nos resultados anteriores de cada subcamada, ajustando a previsão final.

O modelo foi testado para prever o fechamento diário usando dados dos 10 dias anteriores para prever o décimo primeiro dia, sendo aplicado a três conjuntos de dados: S&P 500 de 17 de dezembro de 2010 a 17 de janeiro de 2013 (totalizando 541 amostras); CMSB (China *Minsheng Bank*) de 18 dezembro de 2013 a 18 de janeiro de 2016 (totalizando 528 amostras) e o índice da *Dow Jones* (DJI) de 17 de dezembro de 2014 a 17 de janeiro 2017 (totalizando 543 amostras), com as últimas 30 amostras de cada conjunto reservadas para testar o modelo treinado. As métricas de avaliação do desempenho do modelo foram: erro médio absoluto (MAE), média percentual absoluta do erro (MAPE), raiz do erro quadrático médio (RMSE) e desvio padrão do erro (SDE). demonstrando que a abordagem híbrida proposta, combinando EWT, LSTM e ORELM, superou os modelos tradicionais (LIU e LONG, 2020).

Sob outra perspectiva, Wang et al. (2021) realizaram um estudo inovador ao desenvolver um modelo de aprendizado profundo baseado em computação de reservatório, utilizando redes aleatórias, para avaliar o desempenho preditivo em previsões de ações no horizonte de um dia. Os resultados indicaram que a abordagem proposta superou outras técnicas populares de aprendizado profundo, como LSTM, RNN e EMD2FNN apresentando o menor RMSE nas previsões realizadas.

Em Goswami et al. (2022), utilizou-se o conjunto de dados da NSE (NIFTY 50) para previsão do preço das ações: LT, CIPLA, HCL Tech, HINDUSTANUNILVR (HINDLEVER) e TATASTEEL (TISCO), representando setores distintos. Para isso, utilizaram o modelo LSTM proposto, calculados para dois cenários de treinamento (20 épocas e 10 épocas). Os resultados mostraram que, com 20 épocas, o modelo obteve um desempenho satisfatório para a maioria das ações. Observou-se que o desempenho do modelo variou conforme o número de épocas, evidenciando que a quantidade de épocas é um fator relevante no treinamento de modelos de previsão.

A tabela 2 apresenta uma comparação dos trabalhos relacionados, destacando o conjunto de dados relatado, o tipo de período considerado, algoritmos empregados, quantidade de observações na amostra, métricas empregadas para avaliação e os resultados obtidos.

Tabela 2: Comparação dos trabalhos sobre abordagens técnicas (uso de dados numéricos).

Trabalho	Conjunto de dados	Período	Algoritmos - Modelo	Observações/amostra	Métricas	Resultados
Nelson (2017)	Preço - abertura, fechamento, máximo, mínimo, volume e 138 indicadores técnicos de cinco ativo listado na Bolsa brasileira.	Diários (15 min) – 2014	Redes neurais - LSTM e comparação com: MLP, *Random Forest*, e o pseudoaleatório	1.920	Acurácia, Precisão Revocação, Medida F1	LSTM - 55,9%
Sismanoglu et al (2019)	Preços e Volumes IBM - Ações EUA na NYSE, NASDAQ e NYSE MKT	Diários – 2 jan. 1968 a 9 abr. 2018	Redes neurais recorrentes (RNN) – LSTM	12.648	RMSE	0,04
Bou-Hamad e Jamali (2020)	Taxas de câmbio do dólar australiano (AUD); Franco Suíço (CHF) em relação ao dólar americano	Diários – 16 out. 1989 e 28 mar. 2019	AR (1) GARCH (1,1) Artificial Neural Networks (ANNs) *Random forests (RF)*	1.000 2.000 5.000 10.000	RMSE MAE	
Castro, Reyes e Landazábal (2020)	Índice de Preços e Cotações (PQI): (DJIA), (INPC), (IPC), (RI), (CETES28), (USDMX), (M1) e (SDRM)	Mensais – jan. 2009 a dez. 2016	Multilayer Perceptron Gaussian Processes Data Table Decision Tree Support Vector Machine	-	MAPE	Decision Tree 0.14
Werner, Bisognin e Araujo (2020)	Volume de ações negociadas da Petrobras (PETR3.SA)	Diários – 04 jan. 2010 a 18 set. 2018.	ARFIMA (p,d,q) e Redes Neurais – LSTM Combinações: Variância Mínima, Regressão Linear e Média aritmética.	2157	RMSE MAPE U de Theil	
Liu e Long (2020)	Preços de fechamento diários: S&P 500, CMSB e DJI	Diários - 17 dez 2010 17 jan. 2017	EWT-dpLSTM-PSO-ORELM	1.612	MAPE, MAE, RMSE e SDE	
Wang et al. (2021)	Preços de fechamento S&P 500, DJI, SSE, NYSE, N225, FTSE, NASDAQ	4 jan 2010 e 31 dez. 2018	Reservoir computing (RC) model - LSTM, RNN, and EMD2FNN	2.272	RMSE, MAE, MAPE, $R2$	
Goswami et al. (2022)	Preços de fechamento diários: cinco ações do NSE - NIFTY 50		LSTM	5.205	RMSE e MAPE	Tata Steel 11,95, 0,025
Carosia (2023)	Preço - abertura, fechamento, máximo, mínimo, volume e 6 indicadores técnicos, Ibovespa (B3)	Diários 01 jan. 2015 a 23 mar. 2020	Lin. Reg. SVR (Linear, Pol, RBF) XG Boost Rand. Forest MLP		MSE	Lin. Reg. 0,0002
Kakde e Dale (2024)	Preços de fechamento diários: cinco ações do NSE - NIFTY 50	Diários - 01 de jan. 2014 a 01 fev. 2024	CNN, RNN, LSTM e Bi-LSTM		RMSE e MAPE	MAPE - Sun Pharma 0.010 Bi-LSTM

Fonte: Elaborado pela autora.

Carosia (2023), investigou as técnicas de machine learning para previsão de perdas no mercado de ações brasileiro durante a pandemia da Covid-19, aonde foram estudados os algoritmos: Regressão Linear, Máquinas de Vetores de Suporte, *Random Forest*, *XGBoost*, *Multilayer Perceptron* e um *Ensemble*, com conjunto de dados de séries de preços do índice Ibovespa (B3), realizando experimentos com os indicadores técnicos. Os resultados demostraram que alguns modelos de machine learning até podem proporcionar pequenos lucros durante a pandemia do Covid-19. O modelo de regressão linear mostrou menor medida de erro para conjunto de dados de validação. Porém este modelo não apresentou resultados satisfatórios na simulação de investimento. Já o modelo de MLP e a técnica ensemble tiveram o segundo e o terceiro menores valores de erro e, na simulação de investimento, tiveram resultados significativos em comparação ao modelo regressão linear.

Kakde e Dale (2024), o modelo proposto de Bi-LSTM alcançou um RMSE de 6,6 para a ação da TATA e um MAPE de 0,010 para a ação da Sun Pharma. O estudo demonstra que o Bi-LSTM é um algoritmo eficaz de aprendizagem profunda para prever o preço de fechamento futuro de ação, superando o desempenho em comparação com a CNN, RNN e LSTM. Foram consideradas ações de 5 empresas: Tata Steel, Tata Motors, Sun Pharmacy, Infosys e HDFC Bank, da *National Stock Exchange* (NIFTY 50).

Diversos estudos utilizam, além das séries históricas de preços das ações, indicadores de análise técnica no pré-processamento dos dados. Indicadores como Média Móvel Simples (SMA), Média Móvel Exponencial (EMA), Média Móvel de Convergência/Divergência (MACD), Índice de Força Relativa (RSI), Volume no Balanço (OBV), Estocásticos %D e %K, Índice Acumulativo (AR) e Relação de Volume (VR) são frequentemente incorporados. Essas métricas oferecem informações sobre tendências, força do preço, volumes, e momentos de sobrecompra ou sobrevenda, enriquecendo o modelo preditivo.

Estudos como os de Nelson (2017); Nti, Adekoya e Weyori (2020c); Coelho (2020); Nabipour et al. (2020); Salvi, Souza e Branco Neto (2020); Santos (2020); Pauli, Kleina e Bonat (2020); Vasco (2020); Nti, Adekoya e Weyori (2020b); Carosia (2023) e Mannam (2023), entre outros, exploram previsões do mercado de ações com base em preços e indicadores técnicos, buscando otimizar a precisão dos modelos ao incluir esses indicadores. Na seção a seguir, serão apresentados os trabalhos relacionados à análise fundamentalista e à combinação de dados para previsão no mercado de ações.

3.2 Trabalhos de Análise Fundamentalista e Integração dos dados

Esta seção descreve os estudos relacionados à análise fundamentalista, que utiliza fundamentos econômico-financeiros para determinação do valor da empresa e avaliar seu desempenho por meio de demonstrações contábeis. Além disso, também considera informações de fontes como notícias financeiras, fórum de discussões, sentimentos dos usuários em mídia sociais e indicadores macroeconômicos. Geralmente esses dados fundamentais, alternativos e de análise são não estruturados (textuais). São exemplos de estudos que integram dados de análise fundamentalista e técnica (abrangendo tanto textos quanto números) os trabalhos de Juwono et al. (2024); Nti, Adekoya; Weyori (2020b, 2020c, 2021); Shi et al. (2019); Vargas et al. (2022); Zhang; Yang; Zhou (2021); Zhang et al. (2018a, 2018b, 2019).

Zhang et al. (2018a, 2018b, 2019) apresentou três estudos sobre previsão do mercado acionário. No estudo de Zhang et al. (2018a), este apresentou a previsão de ações por meio de várias fontes de aprendizagem e várias instâncias (modelo *Multi-source Multiple Instance*), utilizando dados históricos, dados de sentimentos de mídias sociais e informações de notícias na Web, no período de 2015 a 2016 do mercado China A-share. O estudo registrou um aumento na precisão das múltiplas fontes de dados em comparação com diferentes fontes.

No estudo de Zhang et al. (2018b) este utilizou de fontes de dados de eventos das notícias da Web e os sentimentos dos usuários das mídias sociais, além de dados de preços e indicadores fundamentalistas, mostrando seus impactos conjuntos nos movimentos dos preços das ações por meio de uma matriz acoplada e uma estrutura de fatoração de tensores do mercado de ações China A-share e ações Hong Kong no ano de 2015. O estudo alcançou uma precisão de predição entre 55% e 62,5%, além disso, os experimentos mostraram uma alta associação entre movimento de preço das ações e sentimentos dos usuários.

Por fim, no seu estudo de Zhang et al. (2019) propôs um modelo de *Markov* oculto acoplado estendido, através de dados de notícias web e séries históricas de ações da China A-share no período de 1 de janeiro de 2016 a 31 de dezembro de 2016. Um total de 21.728 artigos de notícias, incluindo títulos e tempo de publicação em 2016, foram coletados da Wind. Esta combinação de notícias e dados históricos mostrou a melhor precisão do modelo proposto.

A Tabela 3 apresenta uma comparação entre esses e outros estudos, destacando aspectos como as fontes de dados, os conjuntos de dados utilizado, os períodos analisados, os algoritmos aplicados e as métricas adotadas para avaliação.

Tabela 3 – Trabalhos sobre abordagens híbridas (integrando dados numéricos e textuais).

Estudos	Fonte de Dados[1]	Conjunto de dados	Modelo	Métricas
Zhang et al. (2018a)	PV + MS + NW	2015 a 2016 China - Shanghai Composite Index	Multi-source Multiple Instance (M-MI); SVM; TeSIA; Nmil; O-MI; WoR-MI; WoH-MI	F1-score e Accuracy (ACC)
Zhang et al. (2018b)	PV + IF + MS + NW	2015 e 2015 China A-share - ações de Hong Kong	SVM; PCA+SVM; TeSIA; CMT- Z − X; CMT- Z; CMT	Accuracy (ACC) e Correlação de Matthews (MCC)
Zhang et al. (2019)	PV + NW	2016 a 2016 - China -100 ações do Índice de Ações Chinês (CSI) 100	SVM; TeSIA; CMT; ECHMM-NE; ECHMM-NC; ECHMM	Accuracy (ACC) e Coeficiente de Correlação de Matthews (MCC)
Shi et al. (2019)	PV + X + NW	2006 a 2015 EUA - ações do S&P 500	DeepClue, CNN e LSTM	Accuracy (ACC)
Nti, Adekoya e Weyori (2020b)	PV + GT + X + FD + NW	2010 a set. 2019 Gana - três (3) empresas GSE	ANN Multi-Layer Perceptron (MLP)	Specitivity; Sensitivity; Accuracy; RMSE; MAPE
Nti, Adekoya e Weyori (2020c)	PV + IT	2017 a jan. 2020 Gana - duas empresas dos setores bancário e petrolífero	Ensemble GASVM; (SVM); (NN); (DT); (RF); (GA) ESVM; GASVM	Accuracy; AUC; Precision; RMSE; MAE; SD
Nti, Adekoya e Weyori (2021)	PV + GT + VM+NW+FD+X	2017 a Jan. 2020 Gana -	(CNN e LSTM empilhadas), IKN-ConvLSTM	Accuracy; Specificity; F - score; Sensitivity
Zhang Q (2021)	PV + IT+NW	2017 a Jul. 2020 China -BGI Genomics	LSTM	Accuracy; Precision; Recall; F-measure; MAE; MAPE; RMSE; MD; PnL/MD
Vargas et al. (2022)	PV + X + VM	2013 a 2017 - Brasil - empresa Vale, JHSF e Usiminas	LSTM	RMSE; U de Theil
Juwono et al. (2024)	PV + IT + VM + IT	2000 a 2023 - Ações indonésias 8 - ASII; BBCA; BBRI; BMRI; HITS; INDX; SMMT e TLKM	LSTM e GRU	RMSE; MAPE; AcMAPE

Fonte: Elaborado pela autora.

1 Variáveis Macroeconômicas (VM), Informações Contábeis (IC), Preços e Volume (PV); Google Trends (GT); X/Twitter (X), Notícias da Web (NW), Fóruns de Discussão (FD), Indicadores Técnicos (IT), Análise de Sentimentos (AS), Indicadores Fundamentalista (IF), Mídias Sociais (MS).

Shi et al. (2019) apresentou o *DeepClue*, um sistema para interpretar visualmente os fatores aprendidos por modelos de aprendizagem profunda baseados em texto, facilitando a previsão de preços de ações. A abordagem envolveu três etapas principais: (1) construção de uma rede neural profunda para extrair fatores relevantes e permitir interpretações úteis fora do modelo; (2) organização e exibição desses fatores em uma interface de visualização interativa e hierárquica e (3) a avaliação do sistema com dois estudos de caso, utilizando notícias financeiras e tweets relacionados a empresas.

Os dados analisados incluíram séries históricas do mercado norte-americano (S&P 500, 2006-2015, coletadas do Yahoo *Finance*), 341.310 notícias financeiras extraídas de fontes como *Reuters* e *Bloomberg*, e 6.869.771 *tweets* coletados via API do *Twitter* (abril a novembro de 2015). As informações textuais foram vinculadas às empresas por meio de palavras-chave (ex.: “Apple”, “AAPL”, etc.) e *cashtags* (ex.: “$AAPL”) (SHI et al., 2019).

Foram avaliados três modelos preditivos: DeepClue, CNN e LSTM, comparados com previsões humanas realizadas por especialistas. O desempenho foi medido pela acurácia (ACC), destacando-se a LSTM para a *General Motors* com 63% na fase de desenvolvimento e o DeepClue para o *Bank of America* com 60% na fase de teste. No geral, o DeepClue demonstrou precisão comparável ou superior aos outros modelos, reforçando sua eficácia na previsão de movimentos de preços de ações (SHI et al., 2019).

Nti, Adekoya e Weyori (2020b, 2020c, 2021) apresentaram três estudos sobre a previsão de ações. No primeiro estudo, Nti, Adekoya e Weyori (2020b) analisaram a associação entre os sentimentos dos usuários e a previsibilidade do movimento futuro dos preços das 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.

No segundo estudo, Nti, Adekoya e Weyori (2020c) apresentam um novo classificador de ensemble homogêneo chamado GASVM, baseado em suporte máquina vetorial (SVM) aprimorada com Algoritmo Genético (GA). Essa abordagem utiliza GA para a seleção de atributos e a otimização dos parâmetros do kernel SVM, com objetivo de prever o mercado de ações de Gana. O experimento foi realizado com dados de ações ao longo de onze (11) anos e

os resultados demonstraram que o modelo proposto (GASVM) superou algoritmos clássicos de aprendizado de máquina, como Árvore de Decisão (DT), *Random Forest* (RF) e Neural Network (NN), na previsão de um movimento de preço das ações com uma antecedência de 10 dias.

Por fim, o terceiro estudo Nti, Adekoya e Weyori, (2021) propuseram uma nova estrutura para previsão do mercado de ações, baseada na fusão de múltiplas fontes de dados e utilizando uma arquitetura híbrida de redes neurais, composta por *Convolution Neural Networks* (CNN) e *Long Short-Term Memory* (LSTM), chamada *IKN-ConvLSTM*. Enquanto outros estudos costumam explorar uma ou duas fontes de dados para realizar previsões, o estudo de Nti, Adekoya e Weyori (2021) integrou seis fontes de informações relacionada ao mercado de ações de fonte de informações heterogêneas. 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 (NTI; ADEKOYA; WEYORI, 2021).

Os dados qualitativos consistiram em *tweets*, notícias financeiras da web e discussões de fóruns. Os *tweets* foram coletados via API *Twitter Tweepy*, utilizando o cifrão ($) como filtro, resultando em 1.101 tweets relacionados. As notícias financeiras foram extraídas de três sites populares em Gana (ghanaweb.com, myjoyonline.com e graphic.com.gh), por meio da *API BeautifulSoup*, totalizando 251 artigos. Por fim, as discussões de fóruns foram coletadas do site sikasem.org e analisadas utilizando ferramentas de análise de sentimentos. A integração dessas seis fontes de dados resultou em uma boa precisão de previsão para o mercado de ações de Gana, utilizando o modelo IKN-ConvLSTM (NTI; ADEKOYA; WEYORI, 2021).

Zhang; Yang; Zhou (2021) apresentam uma estrutura analítica integrada que utiliza múltiplas fontes de dados, incluindo dados de negociação diária, notícias online, indicadores técnicos e características tempo-frequência derivadas dos preços de fechamento. O estudo fornece uma demonstração prática de como combinar e explorar essas informações para prever os preços das ações da BGI Genomics. Para isso, os autores implementaram uma rede de memória longo e curto prazo (LSTM) aprimorada com um mecanismo de atenção, capaz de identificar dependências temporais de longo alcance e destacar, de forma adaptativa, os principais recursos informativos.

Os autores também compararam o desempenho do modelo proposto 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 (ZHANG; YANG; ZHOU, 2021).

Em Vargas et al. (2022) o objetivo foi prever os preços das ações e integrar a análise de sentimentos das postagens do *Twitter*, através de uma Rede Neural de Memória de Longo e Curta Prazo (LSTM). Os resultados demostram que a análise de sentimento ajudou a alcançar uma raiz do erro quadrado médio (RMSE) de 0,021 para a empresa Vale com a fusão do conjunto de dados de cotações de preços e análise de sentimentos.

Juwono et al. (2024) realizaram um estudo 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, apresentada uma comparação e análise crítica.

3.3 Comparação e Análise Crítica

Com o avanço da tecnologia e o crescente volume de dados provenientes da web e de mídias sociais, métodos mais avançados, como aprendizado profundo e processamento de linguagem natural (PLN), têm sido amplamente aplicados na previsão do mercado de ações. Além disso, diversos estudos têm investigado a integração de dados textuais e numéricos, aliada a técnicas sofisticadas para melhorar a precisão das previsões, conforme destacado por autores como Juwono et al. (2024); Nti, Adekoya; Weyori (2020b, 2020c, 2021); Shi et al. (2019);

Vargas et al. (2022); Zhang; Yang; Zhou (2021); Zhang et al. (2018a, 2018b, 2019) e Jin et al. (2017).

Os autores Nti, Adekoya e Weyori, (2020a), analisaram 122 estudos sobre previsão de ações, identificando que 66% desses estudos utilizam análise técnica, baseada no histórico de preço de ações, enquanto 23% se concentram na análise fundamentalista, que considera fatores como notícias da web, sentimentos de mídias social e variáveis macroeconômicas. Apenas 11% dos estudos analisados aplicam uma abordagem baseada na combinação de fontes de dados e métodos. Além disso, 89,38% dos trabalhos utilizam uma única fonte de dados, enquanto 8,2% combinam duas fontes, e apenas 2,46% integram três fontes, evidenciando a escassez de pesquisas que explorem a combinação de mais de três tipos de fontes de dados numéricos e textuais para previsões de ações.

Com base nos dados brutos de entrada e nos recursos extraídos, observa-se que as combinações mais utilizadas de recursos de entrada incluem preços históricos e indicadores técnicos, que são os mais comumente empregados. Esses são seguidos por dados textuais e macroeconômicos. Essa preferência pode ser atribuída à maior facilidade de acesso e processamento dos dados de mercado em comparação com outros tipos de informações (JIANG, 2021).

Entretanto, os modelos de aprendizagem profunda desenvolvidos nesses estudos dificilmente são diretamente aplicáveis pelos usuários finais em sua forma original. Na literatura, identificam-se dois principais grupos de usuários: os *Traders* de ações, que podem ser investidores independentes ou gestores de fundos públicos e privados, e os analistas de mercado que desenvolvem modelos preditivos para auxiliar os *Traders* (SHI et al., 2019).

A maioria dos modelos de previsão adota uma abordagem de aprendizado supervisionado, onde o conjunto de treinamento é utilizado para ajustar os parâmetros do modelo e o conjunto de teste é reservado para avaliação. Apenas uma pequena parcela dos estudos utiliza aprendizado semi-supervisionado, aplicado quando os rótulos não estão disponíveis na etapa de extração de recursos. Os modelos de previsão podem ser classificados em três categorias principais: modelos padrão e suas variantes, modelos híbridos e outros modelos (JIANG, 2021).

No contexto da ciência da computação, diversos estudos vêm explorando o comportamento mercado financeiro por meio de novos modelos preditivos baseados em Inteligência Artificial (IA) usando técnicas de Aprendizado de Máquina (ML) e aprendizado aprofundo (DL). Segundo Alzazah e Cheng, (2021), os algoritmos mais utilizados incluem

SVM; LSTM; CNN e RNN. As métricas de avaliação mais frequentes na literatura são o MSE, RMSE e MAPE Nti, Adekoya e Weyori, (2020a).

A partir da revisão de literatura apresentada, foram identificadas as seguintes lacunas e oportunidades para futuras pesquisas:

a) Desenvolvimento de estudos que utilizem a integração de quatro ou mais fontes de dados para melhorar a precisão na previsão de ações, explorando combinações mais abrangentes e complexas;
b) Foco na previsão de preços futuros das ações, considerando que a maioria dos estudos existentes se limita a prever apenas os movimentos de alta ou baixa, sem fornecer valores absolutos de preços. Essa limitação reduz a aplicabilidade prática dos modelos para investidores e gestores de carteiras;
c) Investigações que integrem dados de indicadores fundamentados na análise fundamentalista, como desempenho financeiro calculado a partir das demonstrações contábeis das empresas listadas na bolsa de valores, para enriquecer os modelos preditivos;
d) Estudos que combinem variáveis macroeconômicas, como políticas monetárias, taxas de juros e eventos globais, com dados textuais, como sentimentos dos usuários em mídias sociais e notícias financeiras, a fim de captar as interações entre esses fatores e oferecer maior contexto às previsões;
e) Estudos que explorem notícias e mídias sociais brasileiras, levando em conta o potencial das notícias financeiras locais e de plataformas populares, como o *Twitter*. A análise de sentimentos com foco no idioma e no contexto brasileiro ainda carece de modelos específicos capazes de capturar as nuances culturais e linguísticas locais.
f) Aplicações de métodos de fusão de dados no mercado de ações brasileiro, que é pouco explorado em comparação aos mercados asiáticos, europeu e americano. Há uma escassez significativa de pesquisas específicas para o contexto brasileiro, considerando as características econômicas, políticas e culturais que influenciam o mercado acionário local.

Essas lacunas evidenciam a necessidade de expandir o escopo das pesquisas, aprimorando metodologias e explorando novos contextos e fontes de dados para prever o comportamento do mercado de ações. Modelos que combinem técnicas de ciências de dados, economia comportamental e finanças para interpretar melhor as dinâmicas do mercado brasileiro ainda são poucos, deixando de explorar todo o potencial de *insights* interdisciplinares.

Dessa forma, este estudo diferencia-se por seu foco na compreensão da dinâmica dos movimentos de preços das principais ações da Bolsa de Valores B3, analisando o impacto da integração de dados numéricos e textuais na previsão dos preços dos ativos. Além de avaliar a influência dessas variáveis no comportamento do mercado, a pesquisa investiga a capacidade de generalização do modelo proposto para outros mercados, contribuindo para o avanço das técnicas de modelagem preditiva e para uma melhor compreensão dos fatores que afetam a volatilidade e a precificação de ativos financeiros.

4 MATERIAIS E MÉTODOS

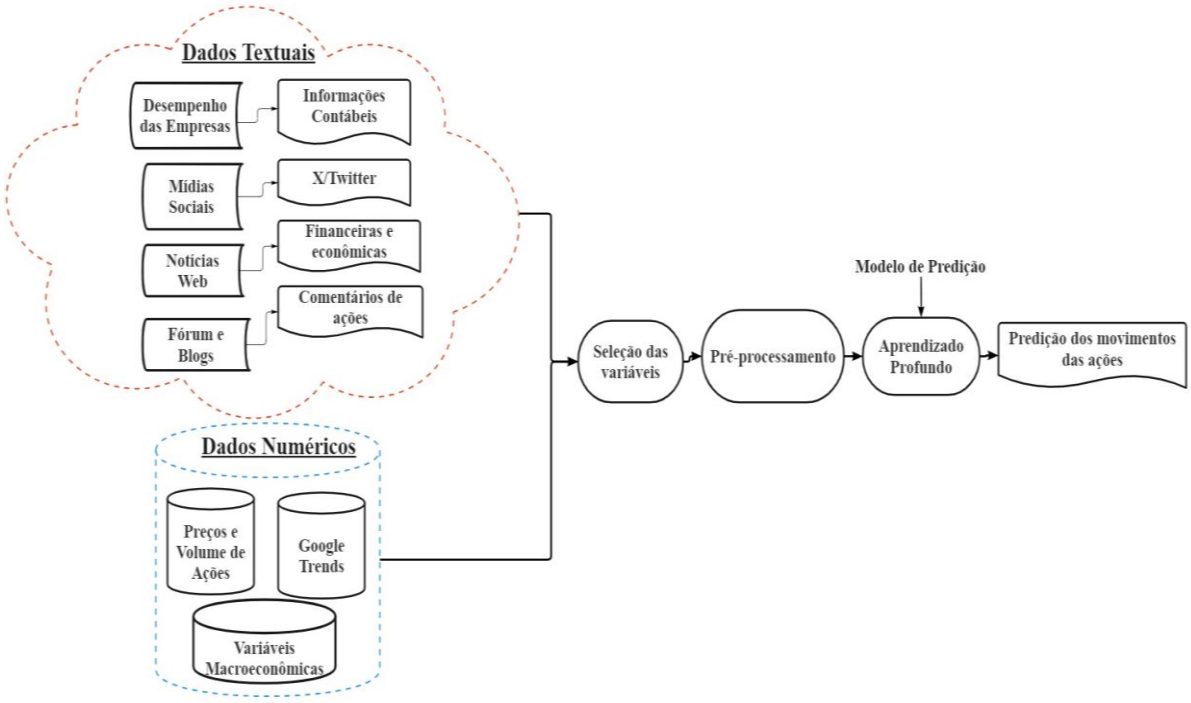
Este capítulo está estruturado em três seções. A primeira seção expõe a abordagem geral da proposta desta pesquisa. A segunda seção apresenta a descrição dos dados previstos para estudo e a última seção apresenta os procedimentos previstos para tratar dos aspectos de avaliação do modelo.

4.1 Visão geral da abordagem proposta

Nesta seção são apresentados os problemas abordados nesta tese, justamente com a abordagem geral proposto nesta pesquisa, conforme resumido na Figura 9.

A Figura 9 apresenta a descrição geral da proposta da pesquisa, com foco na avaliação os efeitos da integração de dados textuais e numéricos para predição do movimento do mercado de ações, considerando os ativos listados na Bolsa de Valores Brasileira (B3). A Figura 9 ilustra as diferentes fontes de dados (numéricas e textuais), a sua integração e o fluxo geral de processamento previsto.

Figura 9: Abordagem Geral Proposto.



Fonte: Elaborado pela autora.

Para a implementação desta abordagem, será realizada inicialmente a coleta de dados provenientes de fontes heterogêneas. Esses dados serão classificados em dados textuais e numéricos.

Os dados textuais incluem informações contábeis, postagens no X (antigo *Twitter),* notícias financeiras e econômicas publicadas na web, fóruns e *blogs* com comentários de usuários sobre ações. Os dados numéricos consistem em séries históricas de preços e volume das ações, variáveis macroeconômicas, além do índice de buscas do *Google trends.*

Após a coleta, a abordagem segue etapas estruturadas, que incluem: a seleção de variáveis, pré-processamento, treinamento, avaliação e ajustes do modelo de predição utilizado para o mercado de ações.

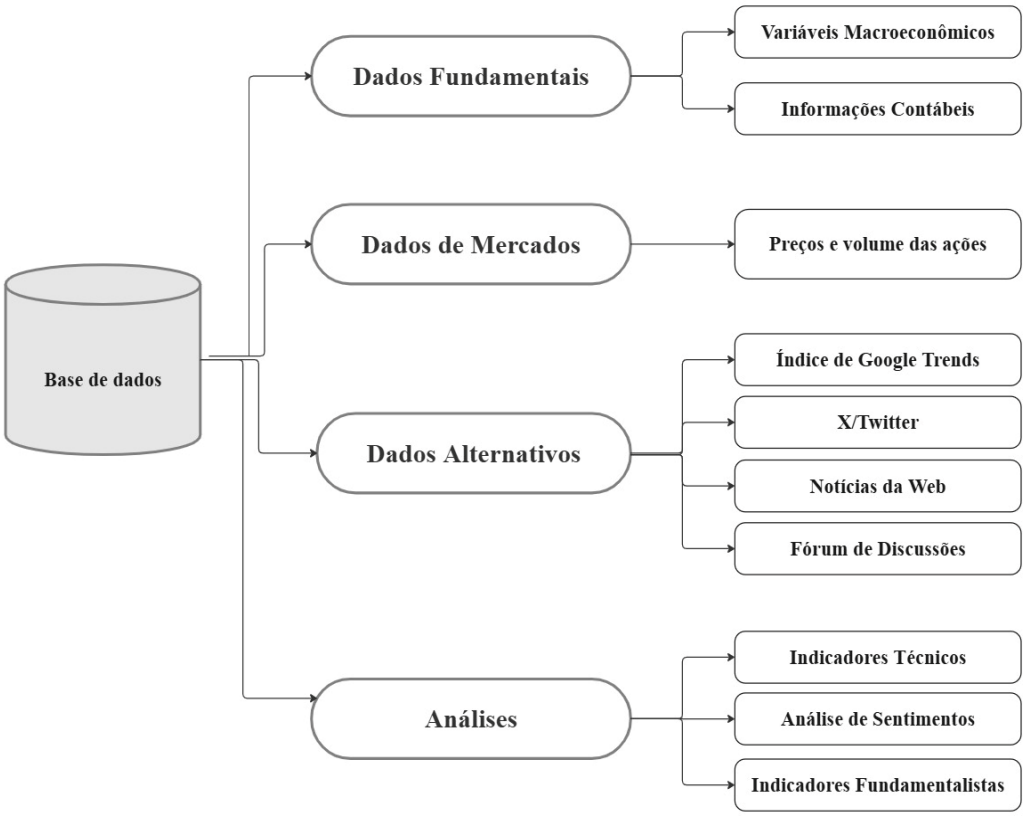
Na seção a seguir, será apresentada a descrição detalhada dos dados que compõem a abordagem proposta.

4.2 Descrição dos Dados

A Figura 10 apresenta os conjuntos de dados que serão utilizados na abordagem proposta neste estudo. Quatro bases de dados serão empregadas: dados fundamentais (variáveis macroeconômicas e informações contábeis), dados de mercado (preços e volumes das ações), dados alternativos (índice do Google Trends, X/*Twitter*, notícias da web e fóruns de discussão) e análises (indicadores técnicos, sentimento das notícias e indicadores fundamentalistas) (DE PRADO, 2018).

Os dados de mercados são amplamente utilizados no setor financeiro, sendo geralmente apresentados em formato numéricos. Neste estudo, foram coletados os dados históricos de preços e volume das ações de cinco empresas listadas na Bolsa de Valores Brasileira (B3). Estes conjuntos de dados foram coletados no *site Yahoo* Finanças (https://br.financas.yahoo.com/) para o período de janeiro de 2000 a dezembro de 2022. Os dados incluem as cotações das séries temporais de volume, fechamento, abertura, mínimo e máximo dos ativos.

Figura 10: Fontes de Dados da abordagem Proposta.



Fonte: Elaborado pela autora.

A Tabela 4 apresenta um exemplo desses dados, mostrando parte da série histórica das ações da Vale (VALE3).

Tabela 4: Exemplo de dados históricos das Ações da Vale (VALE3).

Data	Abertura	Máximo	Mínimo	Fechamento	Volume
28 de fev. de 2022	17,6	18,5	17,6	18,49	35.617.700
25 de fev. de 2022	16,73	17,9	16,72	17,82	42.396.700
24 de fev. de 2022	16,44	17,17	16,29	17,07	53.693.900
23 de fev. de 2022	17,12	17,32	16,99	17,27	30.476.300
22 de fev. de 2022	16,92	17,26	16,89	17,18	34.551.700
18 de fev. de 2022	16,87	16,87	16,61	16,64	26.247.500
17 de fev. de 2022	17,08	17,13	16,46	16,6	38.657.400
16 de fev. de 2022	17,34	17,51	17,25	17,44	20.295.800
15 de fev. de 2022	17,24	17,27	16,83	17,14	35.562.100
14 de fev. de 2022	17,6	17,61	17,36	17,49	24.138.500

Fonte: Elaborado pela autora.

As empresas selecionadas para a realização do modelo proposto são: Vale (VALE3); Petrobras (PETR4); Itaú Unibanco (ITUB4); Bradesco (BBDC4) e B3 (B3SA3). A escolha foi baseada no fato de serem ativos que compõe o índice Ibovespa e estarem entre as maiores empresas negociadas na B3. Conforme apresentado na Tabela 5, estão descritos os códigos dos ativos, nome das empresas, a quantidade teórica dos ativos e a participação no índice Ibovespa. A Vale representa a maior participação no índice, com 16,9395%, seguida pela Petrobras, com 6,411%, Itaú Unibanco 5,774%, Bradesco 3,926%, e B3 3,926%, de acordo com os dados de março de 2022 (B3, 2022).

Os dados de análises foram processados a partir dos preços e volumes para o cálculo dos indicadores de análise técnica, tais como: Média Móvel Simples (SMA), Média Móvel Exponencial (EMA), Regras de Convergência/Divergência de Média Móvel (MACD), Índice de Força Relativa (RSI), Bandas Bollinger (BB), Índice de Fluxo de Dinheiro (MFI) e Balanço de Volume (OBV). A escolha desses indicadores foi embasada em estudos realizados por Coelho (2020); Nabipour et al. (2020); Salvi, Souza e Branco Neto (2020); Santos (2020); Pauli, Kleina e Bonat (2020); Vasco (2020); Nti, Adekoya e Weyori (2020b) e Carosia (2023), Juwono et al. (2024)

Tabela 5: Os 5 ativos com o maior volume e participação na B3.

Código	Ação	Qtde. Teórica	Part. (%)
VALE3	Vale	3.843.570.705	16,939
PETR4	Petrobras	4.566.442.248	6,411
ITUB4	Itaú Unibanco	4.780.002.924	5,774
BBDC4	Bradesco	4.691.427.537	4,546
B3SA3	B3	6.065.856.318	3,926

Fonte: Elaborado pela autora.

A figura 11 apresenta o gráfico de *candlestick* e índice de Regras de convergência/divergência de média móvel (MACD) das ações da empresa Vale (VALE3) referentes aos meses de outubro de 2021 a março de 22, onde podemos observar os movimentos e flutuações do preço, através da análise técnica, elaborado no *site* da ADVFN.

Figura 11: Exemplo de gráfico de *candlestick* e índice de (MACD).

Fonte: Elaborado pela autora.

Com relação aos dados fundamentais, foram escolhidas as variáveis macroeconômicas que podem influenciar o desempenho da Bolsa de Valores (B3) com maior intensidade. A Tabela 6 apresenta exemplos de variáveis macroeconômicas consideradas neste estudo.

Tabela 6: Exemplo de variáveis macroeconômicas.

Variáveis	**2006**	**2007**	**2008**
Nível de preços (inflação)	3,1	4,5	5,9
PIB (US$ bilhões)	1.088,50	1.33,6	1.575,20
População (milhões)	186,8	189,3	191,9
Investimento direto estrangeiro (%)	18,8	34,6	45,1
Exportações globais (US$ bilhões)	137,8	160,6	197,9
Importações globais (US$ bilhões)	91,4	120,6	173,2
Desemprego (IBGE média anual)	10	9,3	7,9
Taxa de juros Selic (FDP)	13,25	11,25	13,75
TJLP (taxa de juros de longo prazo)	7,86	6,42	6,28
Reservas internacionais (US$ bilhões)	85,8	180,3	206,8
Taxa de câmbio R$/US$	2,14	1,77	2,39

Fonte: Pinheiro (2008, p.412).

Os dados das variáveis macroeconômicas representam fatores que influenciam a economia como um todo, impactando diretamente no mercado financeiro. Para este estudo, foi

fundamental a análise de agregados macroeconômicos como inflação, taxa de juros e câmbio entre outros, coletados no site Banco Central Brasil (https://www.bcb.gov.br/). Dessa forma, a análise fundamentalista utiliza tanto o cenário macroeconômico quanto o microeconômico.

Além das variáveis macroeconômicas, os dados fundamentais são, em grande parte, baseados em informações contábeis divulgadas trimestralmente pelas empresas listadas na Bolsa de Valores (B3). Esses dados refletem o desempenho financeiro das empresas, incluindo informações sobre receita, lucro, ativo, passivo, fluxo de caixa e balanço patrimonial, geralmente disponibilizados em relatórios no formato de dados textuais.

Para a obtenção desses dados, serão extraídos os dados textuais dos Balanços Patrimoniais e dos Demonstrativos de Resultados de cada empresa analisada, disponibilizados pela Comissão de Valores Mobiliários (CVM) (http://dados.cvm.gov.br). Para a análise das informações contábeis, serão realizados os cálculos dos indicadores fundamentalistas, conforme apresentado na Figura 12.

A partir dessas informações contábeis, são processados os dados de análises, gerando os indicadores fundamentalista amplamente utilizados na escola fundamentalista. Esses indicadores incluem: índice de preço/Lucro (P/L), preço/valor patrimonial (P/VPA), preço/venda (P/S), indicadores de valor de mercado; Ebitda (LAJIDA); *dividend yield* e retorno sobre patrimônio líquido (ROE) entre outros.

Figura 12: Exemplo dos Indicadores Fundamentalistas.

Indicadores fundamentalistas			
P/L	3,96	LPA	24,25
P/VP	2,50	VPA	38,49
P/EBIT	2,77	Marg. Bruta	60,0%
PSR	1,64	Marg. EBIT	59,2%
P/Ativos	0,96	Marg. Líquida	41,3%
P/Cap. Giro	12,01	EBIT / Ativo	34,8%
P/Ativ Circ Liq	-2,71	ROIC	42,0%
Div. Yield	15,4%	ROE	63,0%
EV / EBITDA	2,58	Liquidez Corr	1,47
EV / EBIT	2,83	Div Br/ Patrim	0,40
Cres. Rec (5a)	27,5%	Giro Ativos	0,59

Dados Balanço Patrimonial			
Ativo	499.128.000.000	Dív. Bruta	76.909.000.000
Disponibilidades	66.437.000.000	Dív. Líquida	10.472.000.000
Ativo Circulante	124.800.000.000	Patrim. Líq	192.403.000.000

Dados demonstrativos de resultados			
Últimos 12 meses		Últimos 3 meses	
Receita Líquida	293.524.000.000	Receita Líquida	70.115.300.000
EBIT	173.656.000.000	EBIT	38.641.800.000
Lucro Líquido	121.228.000.000	Lucro Líquido	30.365.800.000

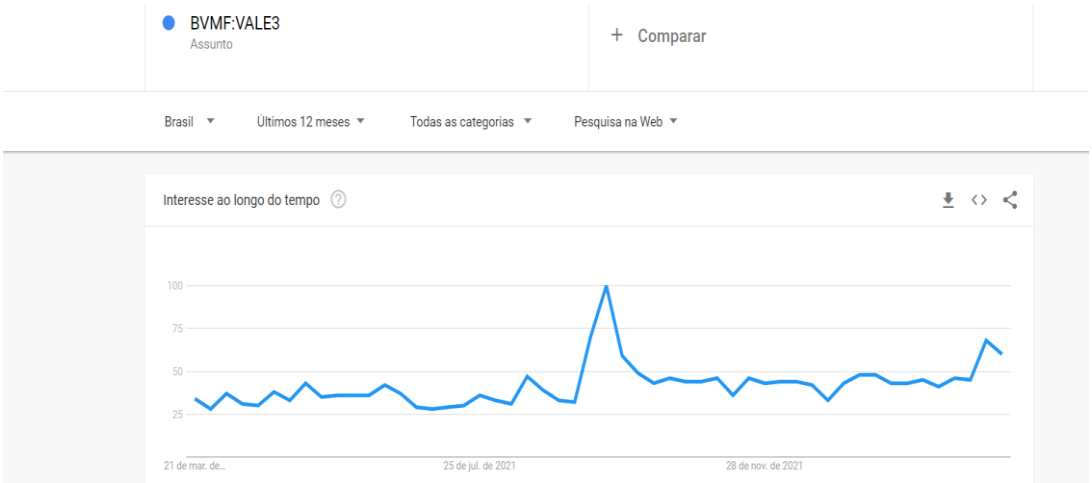
Fonte: Fundamentus (2022).

Os dados alternativos do mercado financeiros provêm de fontes não tradicionais (textuais), tais como: índice do *Google Trends*, X/*Twitter*, notícias da web e fóruns de discussão.

Em 2006, o Google lançou a ferramenta "*Google Trends"*, que exibe a quantidade de vezes que um termo é pesquisado. Os dados fornecidos por essa plataforma podem ser extremamente úteis e são considerados uma fonte de pesquisa confiável, ajudando na análise de sentimento do mercado ao fornecer informações sobre o interesse de pesquisa por termos específicos (GOOGLE TRENDS, 2024). As pessoas estão cada vez mais conectadas à internet. Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), em 2019 cerca de 82,7% dos domicílios brasileiros tinham acesso à internet (IBGE, 2021). Segundo *Statcounter GlobalStats, em (2024)* o Google é um dos principais buscadores, com uma participação de 94,59% no Brasil e 91,05% globalmente.

Esse índice é determinado pelo volume de buscas sobre assuntos específicos relacionados às ações das empresas analisadas, geralmente em uma escala de 0 a 100, onde 100 representa maior volume de buscas em um determinado dia e 0, o menor volume. Esse serviço é oferecido por uma ferramenta do Google. No *site*, os dados são apresentados por meio de um gráfico, mas existe a opção de *download* em arquivo CSV, permitindo que sejam agrupados por categorias e por localização geográfico, como ilustrado na Figura 13. Os índices são calculados com base na frequência de buscas que incluem uma palavra ou um conjunto de palavras, denominados "termos de Pesquisa". Por exemplo, ao pesquisar a palavra "BVMF:VALE3" no *Google Trends*, retorna um índice que quantifica as buscas desse termo em um período.

Figura 13: Exemplo do *Google Trends*.



Fonte: *Google Trends* (2024).

A frequência semanal do índice VHP do *Google Trends* inicia-se no domingo e termina no sábado. Contudo, para a análise, serão considerados apenas os dias úteis, de segunda-feira a sexta-feira, durante o funcionamento da Bolsa de Valores.

A análise proposta não considerará a limitação geográfica, devido à relevância dos investidores estrangeiros na Bolsa brasileira. Serão coletadas duas variáveis do *Google Trends*: i) o nome da empresa; e ii) o *ticker* da empresa, para a construção do modelo proposto. Essas variáveis servem com *Proxies* da demanda por informações, sendo mais simples de acessar e refletindo as ações naturais dos investidores, sem qualquer intervenção.

Com o *Google Trends* é possível criar variáveis de sentimento sobre temas e empresas, considerando o pessimismo ou otimismo dos investidores, uma vez que essas informações são geradas espontaneamente pelos indivíduos. Diversos autores utilizaram *Google Trends* para analisar os movimentos das ações, como Correia (2021); Guzella et al., (2023); Latoeiro (2012); Mourão (2018); Viana (2017); Viola (2022).

A maioria dos investidores não age racionalmente, pois são influenciados por suas emoções. O sentimento do investidor pode explicar o retorno das ações e os termos de notícias negativas são particularmente eficazes para medir esse sentimento (BAKER; WURGLER, 2007; BARBERIS; SHLEIFER; VISHNY, 1998; TETLOCK, 2005).

O conjunto de dados textuais do X/*Twitter* foi obtido por meio da leitura das postagens (*tweets)* que contêm o código dos ativos ou o nome das empresas selecionadas. Para este estudo, os *tweets* serão extraídos diretamente do X/*Twitter* utilizando a API do *Twitter*, acessada por meio da biblioteca *Tweepy* e *stocktwits* (https://developer.x.com/en/docs/twitter-api/).

Essa API permite capturar postagens que mencionem *cashtags* (indicadas pelo cifrão $) e tópicos de interesse relacionados (hashtags, #), como $MGLU, $WEGE, que são amplamente utilizados em estudos acadêmicos para análise de sentimentos e impacto no mercado financeiro (NTI, ADEKOYA e WEYORI, 2021; SHI et al., 2019).

Por meio dessa coleta, será possível extrair opinião e emoções expressas pelos usuários em relação aos ativos analisados, proporcionando *insights* sobre os sentimentos predominantes no mercado. Essas informações são valiosas para identificar padrões comportamentais e tendências que podem influenciar os preços das ações

O conjunto de dados de notícias financeiras da Web, relacionada às empresas analisadas, foi extraído de fonte confiáveis, incluindo os sites *Infomoney*, Valor Econômico, Google News e ADVFN. A análise foi focada nos títulos das notícias, pois eles geralmente resumem os principais eventos e assuntos abordados, facilitando a identificação de eventos relevantes.

Esses títulos foram processados para extrair informações que possam impactar o comportamento do mercado financeiro, conforme apresentado na Figura 14. Esse método permite capturar rapidamente os tópicos mais importantes, proporcionando uma visão eficiente e direcionada das notícias que influenciam os preços das ações das empresas analisadas.

Figura 14: Exemplo de dados de notícias da Web.

Notícias Vale S/A

Data	Hora	Fonte	Notícias sobre as ações da VALE ON
15/03/2022	14:27	ADVFN N…	Vale: mineração em terras indígenas só pode ser realizada..
11/03/2022	09:52	ADVFN N…	Vale é condenada a pagar indenização por danos morais pelo..
07/03/2022	22:26	ADVFN N…	Vale: chuva causa suspensão temporária de trens na estrada..
02/03/2022	14:24	ADVFN N…	Vale faz acordo com governo de Minas Gerais e MP após perder..
25/02/2022	14:03	ADVFN N…	Vale (VALE3): lucro de US$ 5,427 bilhões no 4T21, alta de..
25/02/2022	09:57	ADVFN N…	Embraer eleva valor máximo em oferta para recompra de bonds..
25/02/2022	00:40	ADVFN N…	Vale irá pagar montante total de US$ 3,5 bilhões em..
23/02/2022	10:17	ADVFN N…	Vale e grupo chinês Valin assinam memorando para projeto de..

Fonte: ADVFN (2022).

Da mesma forma, foi realizado o estudo utilizando notícias econômicas da web, a partir das quais serão extraídos os eventos mencionados nos títulos relacionados a fatores externos que impactam a economia local e mundial. Como exemplo de fatores externos que impactam diretamente o mercado financeiro, podem ser citados a pandemia de Covid-19 e a guerra entre Rússia e Ucrânia. Esses eventos são classificados na literatura como risco sistemático e têm sido poucos explorados em modelos de predição. Além disso, para os fóruns de discussões, foi utilizado um analisador de sentimentos, que permitiu captar os sentimentos coletivos expressos nas mensagens desses espaços, proporcionando uma compreensão mais ampla do comportamento dos investidores. A Figura 15 mostra o exemplo dessas discussões sobre ações nos fóruns.

Figura 15: Exemplo de Fórum de discussões.

Forum Vale S/A

Data	Hora	Comentários	Discussão sobre as ações da VALE ON
09/03/2022	09:46	512	VALE3
26/02/2022	18:28	4	Notícia: Vale (vale-n1) - Pagamento Complementar De Remuneracao Das Debentures
10/01/2022	11:30	563	O QUE É A VALE E QTO VALE ?????????????
22/11/2021	05:55	153	VALE 3 NÃO TEM PRA NIMGUEM
22/11/2021	05:55	88	A VALE JA INDICA REVERSÃO ?
29/03/2021	06:37	14	petr4 R$15.10..VPA R$ 20,07....VALE3 R$ 37,00..VPA 26,98
23/03/2021	09:14	12	Notícia: Novo presidente da Vale fala de seus projetos para a empresa

Fonte: ADVFN (2022).

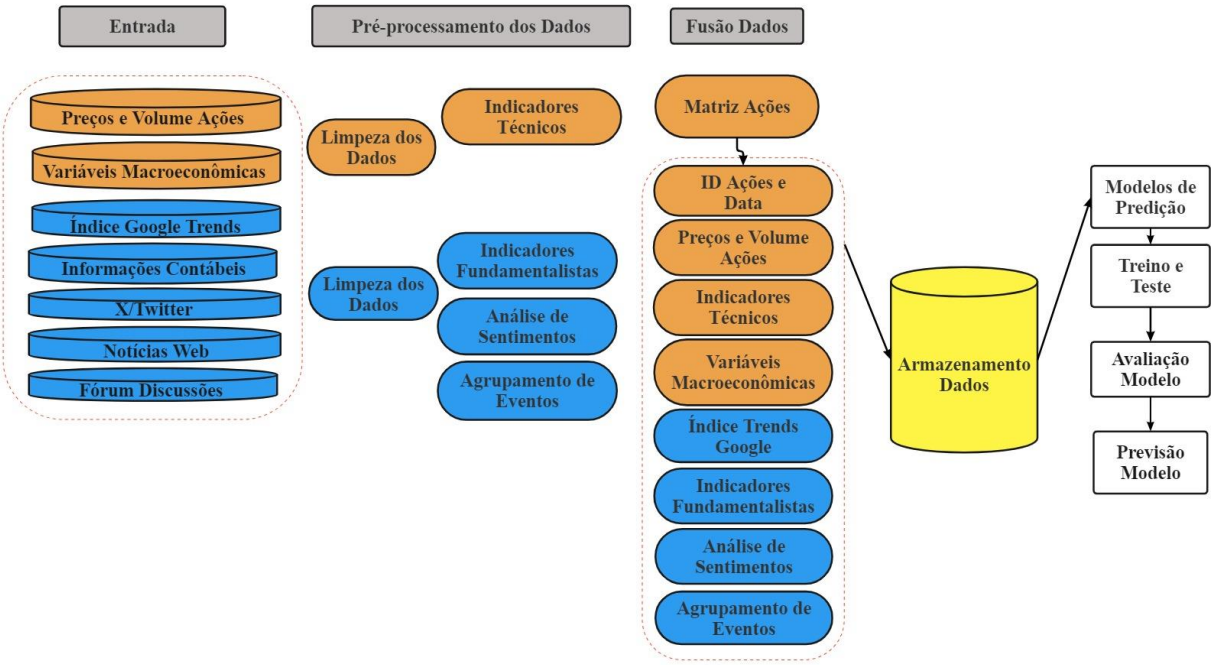
Todos os conjuntos de dados não estruturados (textuais), incluindo postagens X/*Twitter (tweets)*, notícias da *web* (com análise de sentimento baseada nos títulos) e mensagens de fóruns de discussão (com análise do conteúdo das mensagens), foram submetidos a processos de tokenização, segmentação e normalização. Esses procedimentos são fundamentais para a análise de sentimentos, garantindo que as informações relevantes sejam extraídas de forma estruturada e eficiente.

Dessa forma, os dados de análise, conforme ilustrado na Figura 10, incluem três categorias principais: dados de indicadores técnicos (extraídos dos preços e volumes das ações), dados de análise de sentimentos (extraídos de postagens X/*Twitter* e notícias *web* e mensagens de fóruns) e dados de indicadores fundamentalistas (extraídos das informações contábeis). Esses grupos de informações são integrados para proporcionar uma visão abrangente e detalhada do comportamento do mercado, contribuindo para a construção da abordagem proposta.

4.3 Abordagem de integração de dados e predição

Esta seção apresenta a abordagem de integração das setes fontes de dados que serão utilizadas na abordagem proposta, como pode ser observado na Figura 16.

Figura 16: Abordagem Proposta da integração de dados para predição.



Fonte: Elaborado pela autora.

Na Figura 16 apresenta a abordagem proposta de integração de dados para análise e previsão de preços de ações. A abordagem abrange desde a coleta de dados brutos até a aplicação de modelos de aprendizado de máquina, demonstrando a complexidade de integração de fontes de dados estruturados (numéricos) e não numéricos (textuais). Assim, objetivou-se construir um modelo preditivo robusto que combine dados heterogêneos, utilizando técnicas de inteligência artificial, aprendizado profundo e processamento de linguagem natural (PLN), para capturar movimentos dos preços que modelos tradicionais baseados apenas em séries históricas podem não identificar.

Os dados de entrada são coletados de fontes diversas, que capturam diferentes dimensões relevantes para o comportamento do mercado financeiro.

Os dados numéricos (estruturados) incluem preços e volumes de ações, variáveis macroeconômicas, além do índice *Google Trends*. As séries temporais de preços de abertura, fechamento, máximos, mínimos e volumes de negociação são dados de mercados fundamentais para construção de indicadores técnicos e para modelagem estatística baseada em séries temporais. Os dados macroeconômicos são dados fundamentais, representados por séries temporais contínuas, as quais apresentam os fatores externos, como taxa de juros, câmbio, PIB e índice de inflação, que impactam significativamente os preços das ações devido à influência no cenário econômico geral.

O índice *Google Trends*, consiste em dados alternativos derivados de séries temporais de interesse público por busca de palavras-chaves ou tópicos ao longo do tempo. Estes dados são usados como *proxies* para medir a atenção pública, o comportamento coletivo e o interesse em eventos ou ativos financeiros específicos.

Os dados textuais (não estruturados) incluem informações contábeis (dados fundamentais), postagens X/*Twitter*, notícias web e fóruns de discussão (dados alternativos). As informações contábeis correspondem a dados fundamentais como balanços patrimoniais, demonstrações de resultados e fluxo de caixa, que apresentam detalhes sobre passivos, ativos, receitas, custos, endividamento, patrimônio, entre outros. Esses dados, disponibilizados em relatórios geralmente no formato PDF, são tabulados, categorizados e numericamente organizados. Através destas informações são extraídos os indicadores fundamentalistas, os quais permitem avaliar o desempenho das empresas.

As postagens X/*Twitter*, notícias web e fóruns de discussão, consistem de dados provenientes de mídias sociais, portais de notícias e plataformas de debate. Estes dados

capturam opiniões, reações em tempo real e discussões contextuais sobre eventos relevantes ao mercado.

O pré-processamento dos dados envolve a organização, transformação e limpeza, com objetivo de facilitar sua fusão em uma matriz unificada e eliminar ruídos. Na limpeza dos dados de preços e indicadores técnicos, são realizadas a normalização, o tratamento de valores ausentes e o cálculo de indicadores técnicos, como médias móveis e índice de força relativa, entre outros.

No caso dos dados textuais, o processo de limpeza inclui a aplicação de técnicas de processamento de linguagem natural (PLN), como remoção de *stopwords*, *stemming*, lematização e limpeza semântica, para unificar *tokens* textuais. Técnicas avançadas de PLN são utilizadas para a extração de entidades nomeadas e a detecção de tópicos relevantes, além da análise de polaridade de sentimentos (positivo, negativo, neutro).

A partir dos dados contábeis, são extraídas informações financeiras e calculados indicadores fundamentalistas, como preço/Lucro (P/L), preço/valor patrimonial (P/VPA), entre outros.

Após o pré-processamento, os dados são integrados em uma matriz de recurso para análise conjunta, utilizando dimensões como ID ação e data, para garantir a indexação temporal e o alinhamento de todas as variáveis. Na etapa de modelagem e previsão, os dados integrados são utilizados para construir previsões robustas, combinando dados históricos de preços com diferentes fontes de dados (ex.: macroeconômicos, textuais e fundamentalistas). Posteriormente, é realizada uma análise considerando todas as variáveis, com o objetivo de compreender o comportamento de cada fonte de dados e suas contribuições para a previsão de preços no mercado acionário.

São empregados modelos supervisionados, como Redes Neurais Recorrentes (RNNs), *Long Short-Tem Memory* (LSTM) e modelos híbridos que combinam redes neurais convolucionais para capturar padrões complexos e melhorar a precisão das previsões.

Por fim, é realizada a fusão das diferentes fontes de dados para definir o modelo de predição mais adequado. O processo inclui o treinamento do modelo, testes e a avaliação de sua eficácia, garantindo previsões consistente e alinhadas ao comportamento do mercado acionário.

4.4 Métricas de avaliação previstas

A análise de desempenho de modelos de previsão frequentemente utiliza métricas tradicionais, conforme descrito na literatura. Essas métricas permitem avaliar a precisão e eficácia dos modelos em prever valores, como utilizadas nos estudos (CAROSIA, 2023; GOSWAMI et al., 2022; JUWONO et al., 2024; KAKDE; DALE, 2024; ZHANG; YANG; ZHOU, 2021). A seguir, destacam-se algumas das métricas mais utilizadas nestes contextos.

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 (12),

$$RMSE = \sqrt{\frac{1}{n}\sum(previsto_i - real_i)^2} \tag{12}$$

Na equação 12, 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 Erro Quadrático Médio (*Mean Squared Error* – MSE), é amplamente utilizado para medir o desempenho de modelos preditivos. Ele representa a soma dos quadrados das diferenças entre os valores reais e previstos, normalizada pelo número de observações. Sua equação é descrita em (13) e os componentes da equação são os mesmos do RMSE.

$$MSE = \frac{1}{n}\sum(previsto_i - real_i)^2 \tag{13}$$

O Erro Absoluto Médio (*Mean Absolute Error* – MAE) mede o desvio médio absoluto entre os valores previsto e os reais. É uma métrica simples e intuitiva, calculada como apresenta a equação (14):

$$MAE = \frac{1}{n}\sum|previsto_i - real_i| \tag{14}$$

Enquanto o RMSE é mais sensível a erros maiores devido à elevação ao quadrado, o

MAE fornece uma visão linear das discrepâncias.

O Erro Percentual médio absoluto (*Mean Absolute Percentage Error* – MAPE) mede o erro percentual médio entre os valores previstos e os reais. É uma métrica útil para avaliar a precisão relativa de previsões, especialmente quando os valores têm escalas diferentes.

$$MAPE = \frac{1}{n}\sum \left|\frac{real_i - previsto_i}{real_i}\right| x100 \quad (15)$$

Valores menores de MAPE indicam melhor desempenho do modelo, com 0% sendo o ideal (sem erro).

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.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(real_i - previsto_i)^2}{\sum_{i=1}^{N}(real_i - \overline{real})^2} \quad (16)$$

No próximo capítulo são descritos os experimentos realizados com o modelo proposto.

5 EXPERIMENTOS

Este capítulo apresenta os experimentos realizados para identificar o comportamento do mercado de ações e sua relação com os dados textuais adicionais explorados neste trabalho. Os experimentos foram organizados em 3 categorias principais.

A primeira categoria utiliza a abordagem padrão de predição de movimentos do mercado acionário, desenvolvida com base em dados numéricos. Para isso, foram empregados algoritmos de aprendizagem de máquina e aprendizado profundo.

A segunda categoria de experimentos tem como objetivo analisar a relação entre as variáveis macroeconômicas, microeconômicas e volume de buscas no *Google Trends* com o comportamento do mercado de ações. Esses experimentos visam compreender como essas variáveis podem influenciar o sentimento dos investidores e, consequentemente, afetar os movimentos do mercado.

Por fim, a terceira categoria demonstra a integração dos dados numéricos e textuais para a realização de predição, com uso de algoritmos de aprendizagem de profundo.

O capítulo está estruturado em cinco seções. A primeira seção descreve os experimentos relacionados ao uso de dados numéricos na previsão de movimentos do mercado com recursos de aprendizagem de máquina. A segunda seção aborda o uso de técnicas de aprendizado profundo. A terceira e quarta seções detalham os experimentos que avaliam a correlação entre variáveis adicionais e os movimentos do mercado acionário. Por último, a quinta seção apresenta um experimento que integra dados numéricos e textuais para predição.

5.1 Experimento 1 – Modelo de Aprendizado de Máquina

Nesta seção será apresentado o experimento 1 proposto, que teve como objetivo identificar as melhores técnicas de Inteligência Artificial e Aprendizado de Máquina e sua possibilidade de apoio na identificação do comportamento e movimento do mercado acionário. A Figura 17 expõe o modelo geral proposto para o experimento 1.

Figura 17: modelo geral adotada para previsão do Experimento 1.

Fonte: Elaborado pela autora.

Para analisar este movimento dos preços, foi selecionado o ativo da empresa Petróleo Brasileira S.A. (PETR4.SA) listado na Bolsa de Valores brasileira – B3. A empresa Petrobras (PETR4.SA), foi o ativo escolhido para realização deste experimento por ser um ativo que compõe o índice Ibovespa e é uma das maiores empresas negociadas na B3, com um valor de mercado de R$ 269.040.164,00 bilhões em maio de 2020, segundo a Economatica. Foram utilizadas 2.571 observações diárias regulamente disponíveis das séries temporais de quantidades negociadas, quantidade de títulos, volume $, fechamento, abertura, mínimo, máximo, e média, do período de 01 de janeiro de 2010 a 26 de maio de 2020.

5.1.1 Descrição dos Dados

Os dados foram coletados na Plataforma Economatica [2]. A Economatica é uma plataforma de análise de dados de informações financeiras sobre o mercado latino-americano. É uma referência no desenvolvimento de sistemas para análise de investimento, tendo sido fundada em 1986. Seu foco é a coleta e o gerenciamento de bases de dados de altíssima confiabilidade, bem como o desenvolvimento contínuo de ferramentas de análise de alta performance. As informações são constantemente atualizadas com dados mais recentes do mercado financeiro e permitem ao usuário manipular um grande volume de informações, criar insights, fazer simulações avançadas e analisar e comparar ativos de maneira fácil e eficiente (ECONOMATICA, 2020).

Os dados utilizados na pesquisa são informações de preço (abertura, fechamento, máximo, mínimo) de um determinado ativo (PETR4). Este conjunto também é representado como *candlestick*, que indica a variação dos preços de um determinado ativo em uma unidade de tempo (dia, hora, minuto, mês etc.). Ou seja, descreve o valor do preço no início do período, juntamente com o valor no final do período, além dos valores máximo e mínimo que o preço atingiu dentro do intervalo. O Quadro 6 apresenta as descrições do conjunto de dados utilizados no experimento.

Quadro 6: Descrição do Conjunto de Dados Utilizados no Experimento 1.

Entradas/Saídas	Variáveis	Descrição
Feature	Preço - Abertura	Primeira cotação, na abertura de negócios de um dia de negociação.
Target	Preço – Fechamento	Última cotação, no encerramento do mercado.
Feature	Preço – Máxima	Maior cotação do dia.
Feature	Preço- Média	Cotação média do dia.
Feature	Preço - Mínima	Menor cotação do dia.
Feature	Q Negs	Quantidade de ativos negociados no dia.
Feature	Q Títs	Quantidade de títulos dia
Feature	Volume $	Volume das negociações do dia em $
Feature	Retorno	Retorno financeiro

Fonte: Elaborado pela autora.

A pesquisa utilizou de dados de preços, quantidade e volume para previsão do preço de fechamento com a técnica de aprendizado de máquina, procurando aprender e evidenciar

2https://economatica.com/

padrões e tendências no comportamento do preço do ativo. O experimento foi realizado por meio da coleta de dados, preparação de dados, escolha do modelo, treinamento, avaliação, aprimoramento dos paramentos e predição, de modo a determinar a capacidade dos modelos de aprendizagem de máquinas para solucionar os problemas.

No pré-processamento foi realizada a organização e o tratamento dos dados para preparação dos dados para uso com os algoritmos da seguinte etapa, sendo responsável para consolidação das informações relevantes para o algoritmo.

Na base de dados da plataforma do Economatica, na opção de cotações, foi selecionado o ativo "PETR4". Os parâmetros utilizados na extração das cotações foram: escala de data em dias; faixa de data 01/01/2010 a 26/05/2020; não mostrar dias em branco (sábado, domingo e feriado) em que não há funcionamento da Bolsa; preços ajustados por proventos inclusive dividendos; preço em moeda original. Os dados foram extraídos da plataforma Economatica para uma planilha do Excel, como mostra a Tabela 7.

Tabela 7: Conjunto de Dados do Experimento 1.

Data	Q Títs	Volume$	Q Negs	Retorno	Fechamento	Abertura	Mínimo	Máximo	Médio
04/01/2010	13303600,00	493660216,00	13531,00	0,00	28,46	28,18	28,08	28,46	28,30
05/01/2010	21396400,00	794327759,00	22782,00	1,00	28,22	28,51	28,07	28,55	28,31
06/01/2010	18720600,00	697345692,00	18647,00	0,00	28,60	28,07	28,07	28,60	28,41
07/01/2010	10964600,00	408386356,00	12720,00	0,00	28,33	28,43	28,27	28,56	28,41
08/01/2010	14624200,00	542061948,00	14192,00	0,00	28,18	28,34	28,11	28,52	28,27
11/01/2010	15317700,00	566849171,00	16438,00	0,00	28,09	28,38	27,93	28,48	28,23
12/01/2010	14886200,00	540049755,00	18226,00	0,00	27,73	27,91	27,49	27,98	27,67
13/01/2010	23228200,00	840384872,00	23390,00	0,00	27,69	27,88	27,33	28,02	27,59
14/01/2010	20073400,00	721136496,00	21130,00	1,00	27,21	27,67	27,12	27,81	27,40
15/01/2010	21169000,00	756590226,00	19681,00	1,00	27,27	27,13	27,08	27,53	27,26
18/01/2010	23168400,00	840761716,00	17270,00	0,00	27,88	27,38	27,28	27,97	27,68
19/01/2010	16895900,00	615281388,00	14555,00	0,00	27,75	27,80	27,64	27,95	27,78
20/01/2010	18757500,00	668141190,00	17869,00	0,00	27,04	27,53	26,89	27,57	27,17
21/01/2010	24553000,00	852386876,00	24874,00	1,00	26,18	27,14	26,15	27,23	26,48
22/01/2010	19025100,00	656497120,00	17734,00	0,00	26,50	26,02	25,98	26,54	26,32
26/01/2010	20281200,00	689112251,00	18713,00	1,00	25,86	26,15	25,66	26,19	25,92
27/01/2010	19908600,00	675391070,00	24639,00	1,00	26,02	25,86	25,56	26,15	25,87
28/01/2010	13819200,00	474578084,00	12952,00	0,00	26,40	26,39	25,83	26,46	26,19
29/01/2010	16106700,00	557368783,00	15696,00	1,00	26,06	26,53	25,88	26,73	26,39
01/02/2010	15869300,00	543143906,00	13168,00	0,00	26,16	26,15	25,86	26,31	26,11

Q Tits – Quantidade em unidade; Volume ($) em reais; Q Negs – quantidade negociada do dia; retorno; fechamento, abetura, mínimo, máximo e média são valores do preço em reais ($). O Retorno é indicador financeiro de retorno positivo ou negativo das ações em que:1 – preço da ação subiu, 0 – preço das ações caiu dia.

Fonte: Elaborado pela autora.

Para uma análise financeira do preço diário do ativo, foi calculado o retorno diário do ativo, para poder validar o modelo do ponto de vista financeiro, de acordo com equação (17):

$$Retorno = fechamento_{t+1} - fechamento_t \quad (17)$$

Onde $fechamento_{t+1}$ é preço de fechamento do dia anterior em relação ao $fechamento_t$ do dia. A entrada do dado de retorno, foi atribuído uma classe binária, sendo que valor de "1" indica que preço subiu, e "0" que não subiu, teve uma variação negativa no preço do dia em relação ao dia anterior. A tabela 7 apresenta 20 observações do conjunto de dados utilizados na pesquisa.

O arquivo gerado foi utilizado, posteriormente, na ferramenta de análise Orange 3.23.1, que consiste em uma estrutura de mineração de dados baseada em componentes, disponibilizando métodos de aprendizado de máquina de código aberto, visualização de dados e análise de dados. Antes de alimentar os dados para realização das técnicas de aprendizado de máquina, foi realizado um processo de normalização para facilitar a convergência no treinamento e generalização na previsão. Para os dados de preço (fechamento, mínimo, máximo, abertura), quantidade e volume, os atributos de entradas foram normalizados em valores dentro do intervalo de 0 e 1. Como os valores absolutos podem variar bastante ao longo prazo, são tratados pela variação de preço pelo tempo, que são padrões mais prováveis de se repetir.

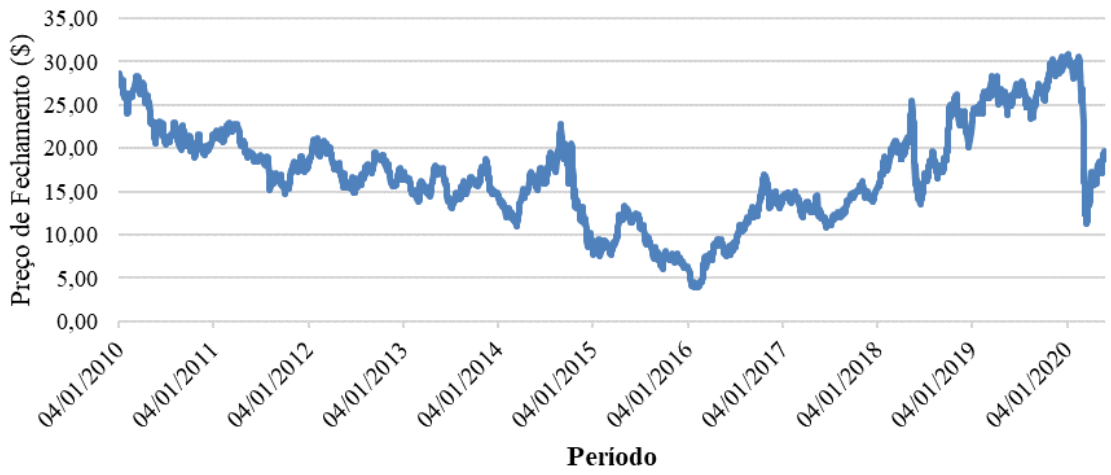
Com processamento e tratamento da base de dados do ativo, geramos um conjunto de dados, dividido em conjunto de treino e teste para avaliar as técnicas de aprendizado de máquina utilizadas. Os modelos escolhidos foram: o algoritmo de regressão linear, florestas aleatórias e rede neural que são os tradicionais utilizados para séries históricas do mercado financeiro e apresentam um ótimo desempenho na previsão. Além disso, foram feitas comparações de desempenho com outros algoritmos, o de Máquina de Vetores de Suporte - SVM, e k-Vizinhos Mais Próximos – kNN.

Para analisar o desempenho de previsão das séries temporais, são utilizadas métricas para analisar as diferenças entre a série de valores previstos e a série de valores reais. Algumas métricas de avaliação foram escolhidas para avaliar quantitativamente o experimento: foram utilizadas as métricas: Erro Médio Quadrático – MSE, Raiz do Erro Quadrático Médio - RMSE, também foi avaliado pela métricas Erro Médio Absoluto - -MAE e Coeficiente de Determinação - R2. Na próxima subseção são descritos os resultados do modelo proposto.

5.1.2 Análise dos Resultados

Esta subseção apresenta os resultados obtidos utilizando o modelo descrito na subseção anterior. As séries temporais financeiras possuem comportamentos conhecidos, como a não estacionariedade dos preços dos ativos (presença de raiz unitária), alta volatilidade e acentuada não linearidade. Esses comportamentos podem ser observados na Figura 18, que mostra a série histórica de preço, em reais, das ações preferenciais da Petrobras (PETR4) listada na B3.

Figura 18: Séries Temporais do Preço de fechamento da PETR4 em reais ($).



Fonte: Elaborado pela autora.

A figura 16 mostra o comportamento do preço de fechamento do ativo, do período de 2010 a 2020, utilizado no experimento. Como observado, o preço do ativo teve grandes oscilações no período analisado. Em 2016 apresentou a menor queda do valor da série, devido a fatores financeiros, onde a empresa teve uma redução no plano de investimentos para os próximos períodos, além dos fatores econômicos globais.

Nos últimos meses de 2020, principalmente no mês de março, as Bolsas de Valores sofreram vários *Circuit breacker* interrompendo as negociações na Bolsa, devido à grande queda no principal Índice acionário do País. O mercado acionário de todo o mundo vem sofrendo bastante quedas, por conta da Pandemia do Covid-19 que se alastrou pelos Países, uma crise sanitária que está impactado fortemente as economias e comportamento das pessoas.

Mostra-se assim que mercado acionário é totalmente incerto, pois os preços das ações têm flutuações diariamente, por conta de inúmeros fatores que influenciam. Desta forma, os investidores buscam sempre ter retorno nos seus investimentos, sendo que para isso precisam de ferramentas para analisar o comportamento do preço e a tendência de vários ativos. Neste experimento 1 foram utilizados algoritmos de aprendizado de máquina que vem sendo usado

para construir modelos de previsão e podem prever os preços das ações e tendência do mercado com boa precisão. A Tabela 8 apresenta as métricas de desempenho que foram utilizadas para avaliar a precisão dos modelos utilizados.

Tabela 8: Métrica de Desempenho do Experimento 1.

Modelo	MSE	RMSE	MAE	R^2
Regressão Linear	0.022	0.148	0.105	0.999
Florestas Aleatórias	0.033	0.183	0.135	0.999
Rede Neural	0.086	0.294	0.207	0.997
Máquina de Vetores de Suporte– SVM	0.503	0.709	0.486	0.985
k-Vizinhos Mais Próximos – kNN	3.126	1.768	1.131	0.907

Fonte: Elaborado pela autora.

Neste experimento, escolheu-se como saída a variável fechamento a ser obtida pelas técnicas: regressão linear, florestas aleatórias, e rede neural. Também foi realizado o experimento com os modelos de máquinas de vetores de suporte e k- vizinhos mais próximos para comparação com os modelos escolhidos.

De acordo com a definição de proporção do conjunto dos valores de entradas, o preço de abertura, máximo, mínimo, média, quantidade negociada, quantidade de títulos, volume de negociação e retorno do ativo. Foram feitas a escolha de alguns parâmetros na definição do modelo que mais se aproxime da resposta ideal. Nesta abordagem foram experimentados algoritmos, sendo conduzidos vários testes para verificar qual apresentava melhor desempenho.

Os parâmetros utilizados para obter os resultados da medida de desempenho foram: *cross validation* de 20 *folders*. O RMSE para regressão linear apresentou o melhor desempenho como o menor erro, de 0.148, um erro médio absoluto de 0.105, e um coeficiente determinação de 0,99 percentual. Além do modelo de regressão linear, os modelos de florestas aleatórias e redes neurais tiveram ótimos resultados, como pode observado na tabela 8. Para floresta aleatória RMSE foi de 0.182 e um MAE de 0.134 tem uma ótima performance de precisão de previsão do preço das ações.

O modelo RNA não tem nenhuma forma funcional, sendo melhor na captura de relação não linear entre os dados e na previsão da saída. Como o mercado de ações prevê também muita não linearidade, RNA fornece um ótimo resultado também para modelo com RMSE de 0.271 e um erro médio absoluto de 0.193.

Os modelos de máquina de vetores de suporte e k-vizinhos mais próximos foram os que tiveram um erro maior que os demais modelos apresentados, com um RMSE de 0.733 e um

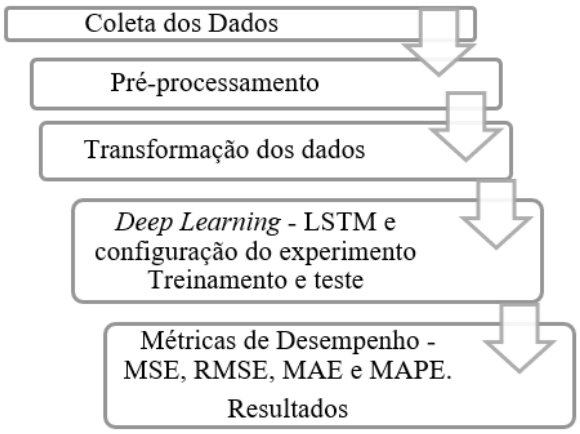
MAE 0.505 de erro absoluto do algoritmo SVM, maior RMSE foi de 1.761 do kNN.

Os modelos mostraram uma ótima precisão da previsão das tendências das cotações do ativo ao longo do tempo, pois são modelos tradicionais de aprendizado de máquina, que são muito utilizados no mercado financeiro, tal como observado no capítulo 3. Cada modelo apresenta suas vantagens e desvantagens sobre os fatores de avaliação e os conjuntos de dados usados para os experimentos, alguns modelos funcionam melhor com determinados tipos de dados históricos, sendo este um aspecto sem ampliado em trabalhos futuros.

5.2 Experimento 2 – Modelo de Aprendizado Profundo

Nesta seção será apresentado o experimento 2, cujo objetivo foi identificar o comportamento e o movimento das cotações do ativo da Petrobras (PETR4.SA) por meio de técnicas de *Deep Learning*, utilizando o modelo de *Long Short Term Memory* (LSTM). A Figura 19 expõe modelo geral proposta para o experimento 2.

Figura 19: Modelo geral para previsão do Experimento 2.



Fonte: Elaborado pela autora.

Os dados foram coletados na Plataforma Economatica. Foram utilizadas 6.269 observações diária, das séries temporais de volume das negociações do dia em reais ($) e preço de fechamento que é última cotação no encerramento do mercado, no período de 01 de janeiro de 1995 a 31 de maio de 2020, o início da série foi escolhido ano de 1995 por causa do Plano Real que foi iniciado em 27 de fevereiro de 1994. A empresa Petróleo Brasileira S.A. - Petrobras (PETR4.SA), foi o ativo escolhido para realização do experimento, por ser um ativo que compõe o índice Ibovespa e é uma das maiores empresas negociadas na B3, com um valor de mercado de R$ 269.040.164,00 bilhões em maio de 2020 segundo Economatica.

A técnica empregada para previsão das ações da PETR4 foi *Deep Learning*, com modelo de *Long Short Term Memory* (LSTM), procurando aprender e evidenciar padrões e tendências no comportamento futuro do preço do ativo.

5.2.1 Descrição dos Dados

Desta forma, a implementação do experimento 2 deste estudo foi realizada por meio da coleta de dados, preparação de dados, escolha do modelo, treinamento, avaliação, aprimoramento dos paramentos e predição, determinar a capacidade do modelo de aprendizagem profunda, para solucionar os problemas.

Os dados foram extraídos da plataforma Economatica para uma planilha do Excel, das cotações das séries temporais de quantidades negociadas, quantidade de títulos, volume $, fechamento, abertura, mínimo, máximo e média, do ativo “PETR4” do período de 02/01/1986 a 26/05/2020. Para realização da implementação foram somente utilizados os dados de volume$ e preço de fechamento do período de 01/01/1995 a 26/05/2020.

No pré-processamento foi realizada a exclusão dos valores faltantes em branco (sábado, domingo e feriado) que não à funcionamento da Bolsa, preços ajustados por proventos inclusive dividendos e preço em moeda original. Os valores foram colocados e decimais e data no formato de data.

Antes de alimentar os dados para realização das técnicas de aprendizado profundo, foi realizado um processo de *MinMaxScaler* para facilitar a convergência no treinamento e generalização na previsão. Para os dados de preço de fechamento e volume $, os atributos de entradas reescalados de forma independente de cada coluna os valores dentro da escala de 0 e 1. Pois os valores absolutos podem variar bastante ao longo prazo, são trocados pela variação de preço pelo tempo, que são padrões mais prováveis de se repetir. Uma proporção de 80 por cento para treinamento e 20 para teste foi usada para a criação do modelo.

Algumas métricas de avaliação foram escolhidas para avaliar quantitativamente o método proposto LSTM. Foram utilizadas as métricas: Erro Médio Quadrático – MSE, Raiz Quadrada do Erro Quadrático Médio – RMSE. Também foi avaliado pela métricas Erro Médio Absoluto - -MAE e Erro Percentual Absoluto Médio - MAPE.

A linguagem de programação utilizada foi Python 3, onde o desenvolvimento foi realizado na ferramenta do Google *Colaboratory* (Google Colab) que permite a criação e execução de códigos em nuvem. Um ambiente virtual foi criado para executar o experimento

para este estudo. Neste ambiente virtual, os seguintes pacotes foram instalados: *Tensorflow* 1.13.1; *Keras* 2.2.4-tf; *pandas* 0.24.2; *Sklearn* 0.21.1; *Numpy* 1.16.3 e *Matplotlib* 3.0.3. A subseção a seguir apresenta as análises dos resultados.

5.2.2 Análise dos Resultados

Esta subseção apresenta os resultados obtidos com o modelo discutido na subseção anterior. As séries temporais financeiras possuem características bem conhecidas, como a não estacionariedade (raiz unitária), alta volatilidade e comportamento não linear. O mercado acionário, em sua essência, é incerto, com flutuações diárias nos preços das ações influenciadas por uma vasta gama de fatores. Diante dessa incerteza, os investidores buscam maximizar o retorno de seus investimentos, recorrendo a ferramentas que auxiliem na análise do comportamento e das tendências de diversos ativos. Neste estudo, utilizou-se o algoritmo de aprendizado profundo *Long Short Term Memory* (LSTM), amplamente empregado para construir modelos de previsão que fornecem estimativas precisas dos preços das ações e tendências de mercado. A Tabela 9 apresenta as descrições do experimento 2.

Tabela 9: Descrição do Experimento 2.

Model – Sequential _ 5		
Layee (type)	**Output shape**	**Param #**
Lstm_5 (LSTM)	(None, 16)	1216
Dense_4 (Dense)	(None, 1)	17

Fonte: Elaborado pela autora.

Os dados foram divididos em: dados de treinamento e dados de teste, resultando 1216 dias para teste, utilizado 5 dias antes para previsão do sexto dia, ou seja, tamanho da janela que se desloca à medida que os dados são lidos, com número de 16 unidades LSTM de neurônios.

As medidas de desempenho de previsão de série temporal fornecem informações essenciais sobre a capacidade do modelo de previsão que esperava fazer as previsões. Como mostra a tabela 10 e figura 20 os resultados das métricas utilizadas no experimento 2.

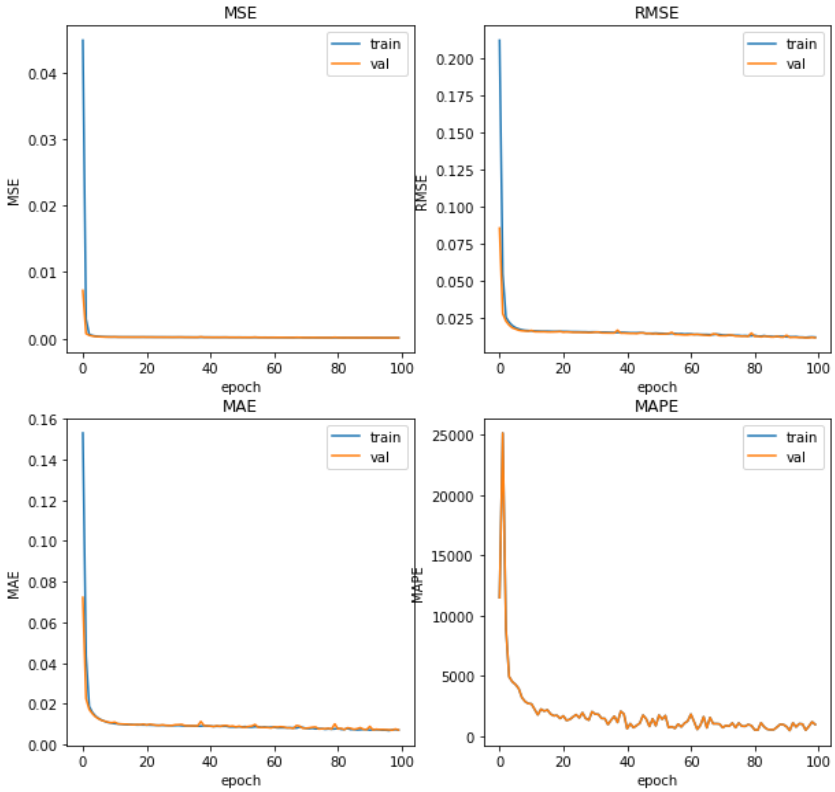
Tabela 10: Métrica de desempenho do Experimento 2.

Métricas	Teste	Validação	Treinamento
Men Squared Error - MSE	0.000176	0.000125	0.000135
Root Mean Squared Error - RMSE	0.013275	0.011201	0.011607
Mean Absolute Error -MAE	0.007329	0.007249	0.007181
Mean Absolute Percentage Error - MAPE	672.771918	5.151606	940.111877

Fonte: Elaborado pela autora.

Para treinamento do modelo foram utilizados 10 por cento da amostra de treinamento para validação do conjunto de dados, considerando o número de epochs de 100 e *batch size* de 64. Os resultados deste processo são mostrados na Tabela 10.

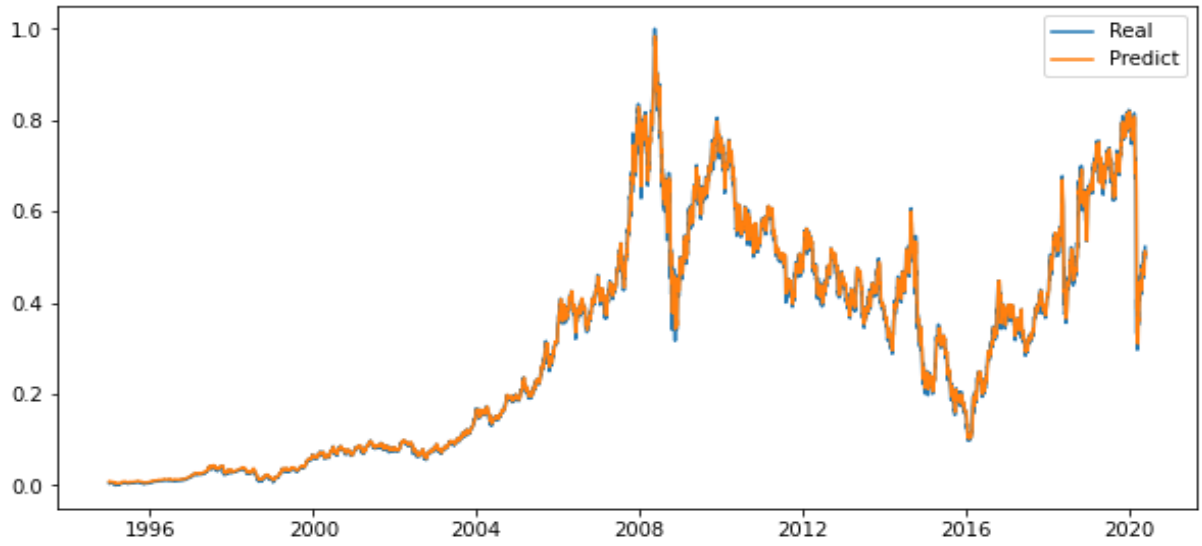
Figura 20: Métrica de desempenho do Experimento 2



Fonte: Elaborado pela autora.

O MSE apresentou o melhor desempenho como o menor erro, de 0.000176, uma raiz quadrada do erro quadrático Médio de 0.013275, erro médio absoluto de 0.007329, e um erro percentual absoluto médio de 672.771918.

Figura 21: Séries de preço de fechamento da previsão e preço real.



Fonte: Elaborado pela autora.

A Figura 21 mostra o comportamento das séries de preço das ações PETR4 preço de fechamento real e preço previsto pelo modelo proposto. Os valores dos preços apresentados na figura estão na escala de 0 e 1, mas a mesma pode ser redimensionada para a escala original.

Cada modelo apresenta suas vantagens e desvantagens sobre os fatores de avaliação e os conjuntos de dados usados para os experimentos, alguns modelos funcionam melhor com determinados tipos de dados históricos ou outras fonte de dados.

5.3 Experimento 3 – Modelo de Regressão Linear

Esta seção apresenta o experimento 3, que possui como objetivo identificar os principais fatores que influenciaram o comportamento do preço das ações da Petrobras no período de 2006 a 2022.

A Petrobras tem uma importância significativa para o Brasil, especialmente por sua contribuição à economia nacional. Sua relevância não se limita apenas ao seu porte; a empresa é uma das maiores do setor de energia e petróleo na América Latina. Como uma sociedade anônima de capital aberto, a Petrobras negocia suas ações na bolsa de valores. A companhia oferece dois tipos principais de ações: ordinárias (ON) e preferenciais (PN). As ações preferenciais (PETR4) se destacam entre os ativos de maior volume de negociação, além de terem uma expressiva participação no índice Ibovespa. Em 9 de setembro de 2022, as ações PETR4 representavam 7,012% do índice, sendo a segunda maior participação (B3, 2022b).

5.3.1 Descrição dos Dados

Para a realização deste experimento, foram utilizados dados secundários provenientes de diversas fontes, incluindo o Instituto de Pesquisa Econômica Aplicada (IPEADATA), *Yahoo* Finanças, Brasil, Bolsa e Balcão (B3), *Investing*, *Google Trends* e a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). Os dados incluem a série histórica de preços de fechamento das ações (PETR4) da Petrobras, índice Ibovespa, taxa de juros Selic, taxa de câmbio (dólar), EMBI+Risco-Brasil, valor internacional do barril de petróleo Brent, índice do *Google Trends* e a produção de petróleo pela Petrobras. O período analisado abrange de janeiro de 2006 a julho de 2022. A escolha do período e das variáveis foi justificada pela disponibilidade de dados adequados para a pesquisa. O Quadro 7 apresenta a descrição das variáveis utilizadas no experimento 3.

Quadro 7: Descrição das variáveis do modelo.

Siglas	Variáveis	Fonte
PETR4	Preço das Ações preferenciais da Petrobras	*Yahoo* Finanças
IBOV	Índice da Bolsa de Valores	B3
SELIC	Taxa de Juros - Selic	Ipeadata
DÓLAR	Taxa de Câmbio - Dólar	*Yahoo* Finanças
EMBI+	EMBI+Risco-Brasil	Ipeadata
BRENT	Preço do barril de Petróleo - Brent	*Investing*
IGT	Índice do *Google Trends*	*Google trends*
PROD	Produção de Petróleo	ANP

Fonte: Elaborado pela autora.

Os dados históricos de preços das ações e do dólar foram coletados no Yahoo Finanças (2022), considerando apenas os dias úteis de operação da bolsa de valores, de segunda a sexta-feira, para a análise. Para a análise, foram utilizados dados diários da série histórica de fechamento dos preços das ações PETR4, que correspondem às ações preferenciais da Petrobras. Essas ações oferecem aos investidores prioridade no recebimento de dividendos, mas não concedem direito a voto nas decisões corporativas (ASSAF NETO, 2021).

Além disso, a taxa de câmbio - dólar, por ser a moeda mais relevante globalmente e amplamente utilizada como reserva por diversas economias, desempenha um papel significativo no mercado internacional, especialmente porque muitas *commodities* são negociadas em dólar.

O Ibovespa, índice que mede o desempenho das ações negociadas na B3, engloba as empresas mais relevantes do mercado de capitais brasileiro. Foram utilizados dados diários da série histórica de fechamento do Ibovespa, coletados no site da B3 (B3, 2022b).

Os dados diários da taxa de juros Selic e do EMBI+(*Emerging Markets Bond Index Plus*) Risco-Brasil, foram extraídos do IPEADATA (2022). A taxa Selic, que é a taxa básica de juros da economia brasileira, serve como referência para as demais taxas de juros no país. O EMBI+Risco-Brasil, indicador que avalia a capacidade de países emergentes de honrar suas dívidas e atrair investimentos, ou seja, medir o desempenho dos títulos da dívida dos países emergentes em relação aos títulos do Tesouro dos Estados Unidos, é desenvolvido pelo banco J.P. Morgan.

O preço do barril de petróleo Brent, um importante indicador no mercado de petróleo, teve seus dados diários obtidos no site Investing (2022). A série de preços foi deflacionada pelo Índice de Preços ao Consumidor Amplo (IPCA).

A produção de petróleo refere-se ao conjunto de operações coordenadas para a extração de petróleo ou gás natural e seu preparo para o fluxo. Os dados da série histórica de produção de petróleo foram obtidos na Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP, 2022).

Em 2006, o Google lançou a ferramenta "*Google Trends*", que exibe a quantidade de vezes que um termo é pesquisado. Os dados fornecidos por essa plataforma podem ser extremamente úteis e são considerados uma fonte de pesquisa confiável, ajudando na análise de sentimento do mercado ao fornecer informações sobre o interesse de pesquisa por termos específicos (GOOGLE TRENDS, 2024).

A ferramenta permite agregar dados por categorias ou regiões geográficas e apresenta os resultados em uma escala de 0 a 100, em que 100 representa o maior volume de buscas em um determinado dia. Para este estudo, foi analisado o termo "Petrobras", com dados restritos ao Brasil, no período de janeiro de 2006 a julho de 2022, com o objetivo de investigar como o interesse popular influencia a oscilação do preço das ações da Petrobras. Esses dados foram coletados diretamente no site do *Google Trends*. Na subseção seguinte, serão detalhadas as características do coeficiente de correlação de Pearson e do modelo de regressão linear aplicado na pesquisa.

5.3.2 Modelo de Coeficiente de Correlação Linear e Regressão Linear

Para analisar a correlação entre as variáveis descritas anteriormente, utilizou-se o coeficiente de correlação de *Pearson*, descrita pela equação (18).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (18)$$

onde:

r = Correlação; xi = Valor variável x no instante i; yi = Valor variável y no instante i; $\bar{x}$ = Média da variável x; $\bar{y}$ = Média da variável yi.

"O coeficiente varia entre -1 e 1. O sinal indica a direção da correlação (negativa ou positiva), enquanto o valor indica a magnitude. Quanto mais próximo de 1, mais forte é o nível de associação linear entre as variáveis. Quanto mais próximo de zero, menor é a associação" (PARANHOS et al., 2014, p. 69).

Segundo Wooldridge (2023, p. 2), "a econometria é baseada no desenvolvimento de métodos estatísticos para estimar relações econômicas, testar teorias, avaliar e implementar políticas de governo e de negócios". Para Gujarati (2019) de modo geral, o modelo de regressão linear pode ser interpretado como uma relação entre a variável dependente (Y) e a variável independente (X), onde Y é afetada por X, e não o contrário.

A equação de regressão linear pode ser representada da seguinte forma (equação 19):

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \varepsilon \quad (19)$$

Onde:

Y - é a variável dependente;

β_0 – é o intercepto;

β_1 – é o coeficiente da variável X;

X - é variável independente; e

ε – é o termo de erro da equação.

De acordo com Hill; Judge e Griffiths (2010), quando se introduz mais de uma variável independente em um modelo econômico, ou seja, quando existem várias variáveis (X) no

modelo econométrico, este é caracterizado como um modelo de regressão linear múltipla. Esse tipo de modelo estuda a relação entre uma variável dependente (Y) e várias variáveis independentes (X).

A equação de regressão linear múltipla pode ser expressa pela equação 20.

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + ... + \beta_n X_{tn} + \varepsilon_t \quad (20)$$

Em que:

Y_t é a variável dependente; X_{tn} as variáveis independentes, e ε_té o termo de erro da equação. O intercepto (constante linear) é representado por β_1, enquanto os coeficientes angulares $\beta_2; \beta_n$ determinam a inclinação da reta de regressão. Na subseção seguinte, será apresentado o modelo econométrico de regressão linear múltipla que possibilitará os resultados do experimento.

5.3.3 Modelo Empírico de Regressão Linear Múltipla

A pesquisa foi conduzida com base em um modelo econométrico de regressão linear múltipla. Nesse modelo, a variável dependente (Y) é o preço das ações preferenciais da Petrobras (PETR4), e as variáveis independentes são: Ibovespa, taxa de juros - Selic, taxa de câmbio - dólar, EMBI+Risco-Brasil, cotação internacional do petróleo Brent, produção de petróleo pela Petrobras e o índice do *Google Trends*.

O modelo tem como objetivo verificar a relação das variáveis macroeconômicas e micro com a variável dependente, o preço das ações preferenciais da Petrobras (PETR4).

O modelo pode ser representado pela equação 21:

$$PETR4_t = \beta_1 + \beta_2 IBOV_t + \beta_3 SELIC_t + \beta_4 DÓLAR_t + \beta_5 EMBI +_t + \beta_6 BRENT_t + \beta_7 IGT_t + \beta_8 PROD_t + \varepsilon_t \quad (21)$$

Em que:

PETR4 = Preço das ações preferenciais da Petrobras;

IBOV = Índice da Bolsa de Valores de São Paulo;

SELIC = Taxa de Juros;

DÓLAR = Taxa de Câmbio;

EMBI+ = EMBI+ (*Emerging Markets Bond Index Plus*) Risco-Brasil;

BRENT = Preço do barril de petróleo Brent;

IGT = Índice do *Google Trends*;

PROD = Produção de petróleo pela Petrobras;

t = = período em meses; 2006 a julho de 2022;

β_1 = intercepto constante;

$\beta_2, \ldots \beta_8$ = coeficiente da angulares;

ε = é o termo de erro da equação.

Assim, com base no modelo econométrico de regressão linear múltipla, objetiva-se identificar as variáveis diretamente relacionadas ao preço das ações preferenciais da Petrobras.

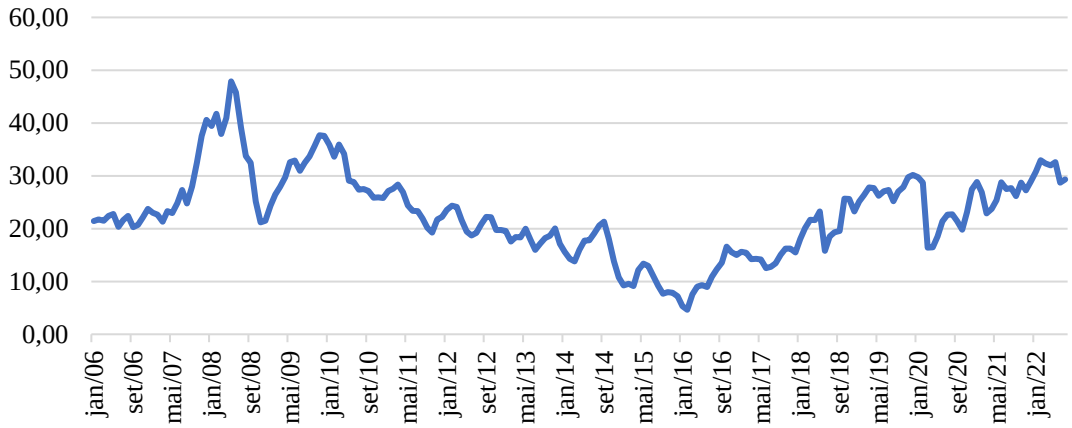
Portanto, a partir das séries históricas diárias, os dados foram tratados e transformados, em séries mensais, e realizada a correlação linear de Pearson com as variáveis estudadas. Em seguida, foi realizada estimação do modelo de regressão linear múltipla, no qual todas as variáveis foram normalizadas em relação às séries originais, por apresentarem escalas com valores diferentes, foi necessário ajustar a escala de valores para cada série dentro do intervalo de escala de 1 e -1. Esse procedimento teve como objetivo facilitar a comparação entre variáveis com magnitudes muitos diferentes, garantindo que uma variável com um valor numericamente maior não domine o modelo. Dessa forma, obteve-se um modelo mais robusto e interpretações mais confiáveis. A regressão linear e a verificação dos resultados foram realizadas utilizando o *software* R.

5.3.3 Análise dos Resultados

Nesta subseção, apresentamos os resultados e as discussões do experimento 3. Diversos fatores podem influenciar a variação dos preços das ações no mercado acionário. A seguir, são apresentadas as variáveis que podem afetam a oscilação dos preços das ações da Petrobras, incluindo: índice Ibovespa (B3), taxa de juros (Selic), taxa de câmbio (dólar), EMBI+Risco-Brasil, preço internacional do barril de petróleo (Brent) e o índice do *Google Trends* (IGT).

Observando o comportamento das séries históricas do preço de fechamento das ações PETR4, percebe-se que ocorreram diversas oscilações ao longo do período estudado. A Figura 22 apresenta esses dados, mostrando as variações no preço de fechamento das ações PETR4, em reais, no período de 2006 a julho de 2022.

Figura 22: Séries históricas do preço de fechamento da PETR4 em reais ($) de 2006 a julho de 2022.



Fonte: *Yahoo* Finanças (2022).

Na Figura 22 observa-se grandes variações no preço das ações PETR4 da Petrobras ao logo do período analisado. Em outubro de 2007, o preço subiu rapidamente, fechando a média do mês em R$ 32,46, um aumento de 16% em comparação ao mês anterior. Em maio de 2008, o ativo alcançou seu valor máximo histórico, fechando a R$ 47,88, com uma variação de 17% em relação a abril, durante o boom das *commodities* e o aquecimento econômico global.

No entanto, a crise do subprime, iniciada em 2007 nos Estados Unidos e intensificada em 2008, impactou severamente os mercados financeiros globais. A desvalorização das *commodities* contribuiu para a queda das ações da Petrobas, que em outubro de 2008 registrarm uma queda de 22,6%, fechando o ano com uma média de R$ 21,51.

A partir de 2014, a Operação Lava Jato trouxe grande instabilidade à Petrobras. Para combater à corrupção e lavagem de dinheiro, porém os maiores impactos surgiram nos anos seguintes, em dezembro de 2015, as ações caíram para R$ 7,19, representando uma queda de 8,6% em comparação ao mês anterior. Em fevereiro de 2016, o preço atingiu seu menor valor, R$ 4,64, devido à redução dos investimentos da Petrobras e à crise econômica de 2016. O *Impeachment* da presidente do Brasil, Dilma Rousseff, também contribuiu para instabilidade política e afetou negativamente o valor das ações (TORGA et al., 2021).

Em junho de 2018, as ações sofreram uma nova queda, fechando em R$ 15,82, uma desvalorização de 32,1% em relação a mês anterior, influenciada pela greve dos caminhoneiros, que gerou incertezas para a Petrobras e seus investidores (DAL PUPPO, 2020).

Em março de 2020, a pandemia de Covid-19 causou uma queda de 42,9% nas ações, com os preços em R$ 16,40. A crise sanitária e econômica resultante da pandemia afetou não apenas a Petrobras, mas também a economia global, levando o Estado a adotar medidas

financeiras emergenciais para conter os impactos econômicos e sociais (PAIVA; PAIVA, 2021).

Para identificar as principais variáveis que afetam o comportamento do preço das ações da Petrobras (PETR4), foi utilizada uma série histórica mensal com 199 observações, de 01 de janeiro de 2006 a 1 de julho de 2022. Para avaliar o comportamento da PETR4 (preço de fechamento das ações, variável dependente), foram utilizadas como variáveis independentes: índice Ibovespa (IBOV); taxa de juros (Selic), taxa de câmbio (dólar); EMBI+ (risco-Brasil); preço do barril de petróleo (Brent); índice *Google Trends* (IGT) e produção de petróleo (PROD).

A Tabela 11 apresenta a matriz de correlação entre as variáveis, permitindo observar a dinâmica entre elas e identificar potenciais de relações para o modelo de regressão linear múltipla. O preço de fechamento das ações (PETR4) mostra uma correlação positiva relativamente forte com o índice Ibovespa (IBOV), com coeficiente de 0,67. Esse resultado é esperado, pois empresas de grande capitalização, como a Petrobras, tendem a acompanhar o movimento geral do mercado de ações no Brasil. Além disso, a (PETR4) representa 7,012% do índice em 9 de setembro de 2022, o que intensifica essa relação (B3, 2022b).

Tabela 11: Correlação entre as variáveis 2006 a 2022.

Variáveis	PETR4	IBOV	SELIC	DÓLAR	EMBI+	BRENT	IGT	PROD
PETR4	1,00							
IBOV	**0,67**	1,00						
SELIC	-0,36	-0,75	1,00					
DÓLAR	0,25	**0,75**	-0,53	1,00				
EMBI+	-0,18	0,01	0,11	0,53	1,00			
BRENT	0,14	-0,43	0,30	-0,80	-0,64	1,00		
IGT	0,09	-0,32	0,37	-0,40	-0,17	0,41	1,00	
PROD	0,21	**0,78**	-0,56	**0,87**	0,36	-0,75	-0,52	1,00

Fonte: Elaborado pela autora.

A taxa de juros (SELIC) obteve uma correlação negativa fraca de (-0,36) em relação ao preço da ação da Petrobras (PETR4). Esse comportamento sugere que taxa de juros mais altas tornam os investimentos de renda fixa mais atrativos, incentivando investidores a migrarem para alternativas menos voláteis, o que reduz a demanda por ações e pressiona seus preços para baixo. Dessa forma, uma taxa de juros elevada tende a desestimular os investimentos no mercado acionário.

A taxa de câmbio o dólar e a produção de petróleo (PROD) mostram uma correlação positiva fraca com o preço das ações, de 0,25 e 0,21, respectivamente. No caso do dólar, a Petrobras se beneficia de sua valorização, pois é uma grande exportadora de petróleo e outras *commodities*, que são cotadas em dólar. A alta do dólar ajuda a sustentar o valor das ações da empresa, já que seus produtos ganham valor em termos de receita.

O índice EMBI+ risco-Brasil apresenta uma correlação negativa muito fraca (-0,18) com o preço das ações (PETR4), indicando que o aumento do risco país tem uma influência limitada sobre o desempenho da Petrobras no mercado acionário, mas ainda assim tende a impactá-lo negativamente.

A correlação entre o preço das ações da Petrobras e preço do barril de petróleo (BRENT) é positiva e muito fraca, com coeficiente de 0,14. Isso sugere que, emborra um aumento no preço do barril de petróleo (BRENT) possa levar a uma alta nas ações da Petrobras, essa relação é relativamente fraca e não representa o único fator para o valor das ações.

Por fim, o volume de pesquisas no *Google Trends* (IGT) pelo nome das empresas tem uma correlação muito fraca 0,09 com o preço de fechamento das ações. Embora haja uma leve tendência de o interesse por buscas relacionadas à Petrobras acompanhe um aumento modesto no preço das ações, essa relação é bastante sutil, indicando que outros fatores exercem maior influência sobre o valor das ações da empresa.

A Tabela 11 mostra que algumas variáveis independentes apresentaram correlações positivas e fortes, como dólar e produção de petróleo (PROD) de 0,87, além do índice Ibovespa (IBOV) e a produção de petróleo (PROD), com correlação de 0,78. Essas relações podem indicar possíveis problemas de multicolinearidade, afetando assim a estabilidade dos coeficientes de regressão.

O (IBOV) está correlacionado com (DÓLAR) 0,75 e com a (PROD) 0,78, mostrando que esses fatores afetam significativamente o índice da Bolsa de Valores. Observa-se também a relevância das variáveis (IBOV) e (SELIC), destacando a importância dos fatores macroeconômicos na precificação de ações no Brasil. Os efeitos do dólar e da produção de petróleo (PROD) refletem a dependência de setores exportadores e *commodities*.

Foi estimando um modelo utilizado a linearização das variáveis, porém os resultados não foram tão expressivos. Em contrapartida, o modelo com as variáveis normalizadas apresentou um desempenho superior.

Vários testes foram realizados para avaliar o desempenho das variáveis que influenciam o comportamento das ações da Petrobras. Observou-se a presença de multicolinearidade no

modelo proposto, que utiliza séries normalizada e mensais, com as variáveis: índice Ibovespa (IBOV), taxa de câmbio (DÓLAR) e produção de petróleo (PROD). Os valores de VIF para essas variáveis foram 7,197 (IBOV); 9,495 e 6,868 (PROD), indicando uma multicolinearidade significativa. A análise do Fator de Inflação da Variância (VIF) evidenciou fortes relações entre as variáveis independentes, comprometendo a estabilidade e a interpretação dos coeficientes no modelo de regressão.

Variáveis independente com VIF acima de 5, indica uma multicolinearidade elevada. Para solucionar este problema e evitar que a precisão do modelo seja afetada, aplicou-se a técnica de remoção das variáveis com alta colinearidade. Isso resultou em um modelo de regressão linear múltipla mais estável e confiável, com coeficientes menos suscetíveis a variações, como apresentado no resultado do modelo proposto na Tabela 12.

Tabela 12: Resultado da estimação do modelo de regressão linear múltipla.

	Coeficiente	**Erro Padrão**	**razão-t**	**p-valor**	
Const	-0.24678	0.08918	-2.767	0.00621	**
Selic	-0.23415	0.04390	-5.333	2.71e-07	***
EMBI+	0.36417	0.06461	5.636	6.16e-08	***
BRENT_norm	0.28216	0.05765	4.894	2.09e-06	***
IGT_norm	0.06402	0.04637	1.381	0.16898	ns
R^2	0.7997				
R^2 AJUSTADO	0.7924				

***, **, * destaca os resultados estatisticamente significativos ao nível de 0,01, 0,05 e 0,10, respectivamente.

Fonte: Elaborado pela autora.

A estimação da regressão revelou que o R^2 do modelo é equivalente a 0,7997, o que indica que 80% das variações da variável dependente (Y) podem ser explicadas pelas variáveis independentes (X), enquanto os 20% restantes são atribuídos ao termo de erro. Na análise, foram considerados níveis de significância de 5% para as variáveis explicativas, assegurando que apenas variáveis estatisticamente relevante sejam incluídas no modelo.

A partir da Tabela 12, observa-se que as variáveis, taxa de juros (SELIC), EMBI+ (risco-Brasil) e preço do barril de petróleo (BRENT) apresentaram relação positiva e coeficientes significativos com a variável dependente, com p-valores menores que 0,0001.

A variável taxa de juros (SELIC) apresentou uma relação significativa e negativa com o preço das ações PETR4. O p-valor de 2.71e-07 indica a significância estatística para esse

coeficiente. Isso significa que, para um aumento de 1% na taxa de juros (SELIC), o preço das ações PETR4 tende a diminuir aproximadamente 0,23%. Esse resultado está em linha com a teoria financeira, pois taxas de juros mais altas tornam os investimentos em renda fixa mais atraentes, reduzindo a demanda por ações e pressionando seus preços para baixo.

Paraboni et al. (2016) um aumento da taxa de juros fará o governo elevar o custo do crédito, dessa forma torna-se mais caro realizar empréstimos ou fazer financiamentos de bens ou serviços. Como consequência, isto irá gerar uma redução do consumo e do investimento.

Do mesmo modo, Righi et al. (2012) relatam que ao aumentar a taxa de juros, o governo elevará os preços dos títulos públicos para rendimentos mais altos, competindo com os rendimentos do mercado financeiro, além de reduzir o consumo e os investimentos desincentivando aplicações no mercado de capitais.

O índice de EMBI+ (risco-Brasil) mostrou uma relação positiva e significativa com o preço das ações (PETR4), com um p-valor de 6.16e-08. Isso significa que um aumento de 1% no risco-país está associado a um crescimento de aproximadamente 0,36% no preço das ações da Petrobras. Segundo Lopes (2019) esse índice é uma referência importante para investidores estrangeiros, que historicamente movimentam volumes substanciais na bolsa de valores brasileira. Esse comportamento ajuda a explicar a relação positiva entre o preço das ações da PETR4 e o risco-país.

O valor internacional do barril de petróleo Brent apontou uma relação positiva e significativa com o preço da (PETR4) com p-valor de 2.09e-06, em outras palavras, se houver um aumento de US$ 1, no preço do barril de petróleo Brent, o preço das ações (PETR4) irá aumentar cerca de R$ 0,28. No estudo de Nunes et al. (2018), os resultados demonstraram que a variável preço do barril de petróleo Brent apresentou relação positiva e significativa com a variação dos preços das ações (PETR4).

O índice do *Google Trends* (IGT) apresentou um coeficiente de 0,06402 e um p-valor de 0,16898, indicando que o volume de buscas pelo termo “Petrobras” não apresenta uma relação significativa com o preço das ações. No estudo de Caraça (2019), foram realizadas diversas análises, e embora os resultados também não tenham sido significativos, o autor sugere que o comportamento das ações pode influenciar positivamente o volume de buscas no *Google Trends*. Isso significa que uma valorização no preço das ações poderia levar a um aumento das buscas por parte dos investidores.

Assim, com base nos resultados do modelo proposto, observa-se que variáveis macroeconômicas e setoriais, como taxa de juros (Selic), EMBI+ (risco-Brasil) e preço do barril

de petróleo (Brent), têm um impacto significativo sobre o comportamento das ações da Petrobras. Embora o índice do *Google Trends* (IGT), que poderia captar o interesse público, não tenha mostrado impacto estaticamente relevante neste modelo, ele ainda pode ter utilidade em análises complementares.

5.4 Experimento 4 – Modelo de Coeficiente de Correlação de Pearson

Esta seção apresenta o experimento 4. Neste contexto o objetivo do experimento foi identificar o efeito das pesquisas na internet via *Google Trends* nas decisões dos investidores no período de 2019 a setembro de 2022, bem como seus impactos sobre o preço, retorno e volume dos ativos, PETR4, VALE3 e ITUB4. Foram observadas as tendências ao longo do tempo das empresas que apresentam o maior peso na composição da carteira teórica do Ibovespa, o índice da B3. A empresa Vale representa a maior participação no índice no índice Ibovespa, com 13,73%, seguida pela Petrobras, com 7,76%, e o Itaú Unibanco, com 7,14%, conforme dados de janeiro de 2024 (B3, 2024c). Este estudo é importante como um experimento de metodologia para evidenciar a correlação entre elementos textuais e a variação do preço das ações.

Foram coletados dados de séries histórica do fechamento ajustado do preço das ações e volume de negociações diárias em R$ de 02 de janeiro de 2019 a 30 de setembro de 2022, através do Yahoo Finanças. Este período abrangeu uma evolução significativa no número de investidores na Bolsa de Valores – B3 e os impactos da pandemia de COVID-19 na economia global e nas Bolsas de Valores.

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 volume de pesquisa no Google foi coletado através da ferramenta *Google Trends* para compreender a demanda dos investidores e analistas sobre o comportamento do mercado e suas decisões de compra e venda de ações. Os termos de pesquisa utilizados foram os nomes das empresas selecionadas (VALE S.A, Petrobras e Itaú Unibanco) e os *ticker* de cada empresa (códigos das ações “VALE3, PETR4, ITUB4”). O volume de buscas foi observado numa escala de 0 a 100, onde 100 representa pico de volume de buscas em um determinado dia e 0 é o menor número de buscas. Para normalização das variáveis, foi criada uma escala de 0 a 1, pois os dados não gerados em números absolutos.

O formato de arquivo de texto específico permite o salvamento de dados em um formato estruturado de tabela, medindo índices pela frequência de busca dos termos. A pesquisa pelo termo “BVMF:VALE3” no *Google Trends*, retornando um índice das buscas em determinado período. A frequência semanal do *Google Trends* começa no domingo e termina no sábado, mas apenas dias úteis foram considerados para a análise, com limitação geográfica ao território do Brasil. Essas variáveis servem como *Proxies* de demanda por informações, refletindo ações naturais dos investidores. Os dados foram tabulados no Excel e tratados conforme o modelo descrito a seguir.

5.4.1 Modelo de coeficiente de correlação Pearson, Retorno e Volatilidade

A covariância indica que duas variáveis variam juntas, enquanto a correlação mensura a direção e o grau de variância entre duas variáveis (SARTORIS, 2003). Para analisar a correlação entre as variáveis descritas anteriormente, utilizou-se o coeficiente de correlação de Pearson, descrita pela fórmula 22.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{22}$$

onde:

r = Correlação;

xi = Valor variável x no instante i;

yi = Valor variável y no instante i;

$\bar{x}$ = Média da variável xi;

$\bar{y}$ = Média da variável yi.

O coeficiente de relação varia de -1 a 1, onde o sinal indica uma relação positiva ou negativa e o valor absoluto mostra o grau de força na relação entre as variáveis. Um coeficiente de 0 significa que não há relação linear entre as variáveis (FIGUEIREDO FILHO; SILVA JÚNIOR, 2009).

Para uma análise financeira diária do ativo, o retorno diário foi calculado conforme a seguinte equação. Os retornos percentuais foram obtidos com a fórmula 23:

(23)

$$R_i t = \left(\left(\frac{p_t}{p_{t-1}} \right) - 1 \right) \cdot 100$$

onde:

R = Retorno;

i = Ativo;

t = Tempo;

P= Preço.

Foi calculada também a volatilidade das ações, sendo ela diária, anual, mensal, trimestral e semestral, com a seguinte fórmula. A volatilidade diária é calculada com o desvio padrão com a seguinte fórmula (24):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (Rai - Ra)}{n - 1}} \quad (24)$$

onde:

σ = Desvio padrão;

Rai= Retorno diário no instante i;

Ra =Média aritmética dos retornos diários;

N= Número de observações.

A volatilidade anual, mensal, trimestral e semestral, foram calculadas a partir da volatilidade diárias usando as seguintes fórmulas:

A volatilidade anual:

$$Volatilidade\ Anual = Volatilidade\ diária * \sqrt{252} \quad (25)$$

A volatilidade mensal:

$$Volatilidade\ Mensal = Volatilidade\ diária * \sqrt{21} \quad (26)$$

A volatilidade trimestral:

$$Volatilidade\ Trimestral = Volatilidade\ diária * \sqrt{63} \quad (27)$$

A volatilidade Semestral:

$$Volatilidade\ semestral = Volatilidade\ diária * \sqrt{126} \quad (28)$$

Antes de calcular o coeficiente de correlação de *Pearson*, foi realizado um processo de normalização para garantir que os dados fossem tratados de maneira consistente e eficaz. Isso é necessário porque os valores absolutos podem variar significativamente ao longo do tempo. Os dados de preço de fechamento ajustado e volume de pesquisas na *web* (*Google Trends*)

foram todos normalizados para facilitar a análise. A subseção a seguir apresenta as análises e discussões dos resultados.

5.4.2 Análise e discussões dos Resultados

A Tabela 13 apresenta as correlações entre o preço de fechamento, retorno, volume, e o volume de buscas no *Google Trends*, para identificar como o volume de buscas pode influenciar ou estar associado com o comportamento desses ativos no mercado. A correlação das empresas Vale S.A, Itaú Unibanco e Petrobras entre as variáveis do modelo, com base em 2.802 observações referentes à 3 ações e 934 dias de pregão. A tabela mostra a matriz do coeficiente de correlação de *Pearson* entre as ações para preço de fechamento, volume de transações das ações, retorno e volume de buscas no Google Trends pelo nome da empresa – GT e volume de buscas pelas *Tickers* (código das ações na Bolsa de Valores – B3).

Tabela 13: Correlação entre as variáveis do modelo.

Empresa Vale S.A.					
Variáveis	**Preço**	**GT**	**Ticker**	**Volume**	**Retorno**
Preço	1				
GT	**0,195699495**	1			
Ticker	0,093624675	0,469255	1		
Volume	0,080887569	0,055216	0,074973	1	
Retorno	-0,04116874	-0,0183	0,008709	0,02624	1
Empresa Itaú Unibanco					
Variáveis	**Preço**	**GT**	**Ticker**	**Volume**	**Retorno**
Preço	1				
GT	**0,253988998**	1			
Ticker	-0,029010173	0,274799	1		
Volume	-0,387217461	-0,09416	0,241462	1	
Retorno	-0,049785084	0,014943	0,02148	-0,02498	1
Empresa Petrobras					
Variáveis	**Preço**	**GT**	**Ticker**	**Volume**	**Retorno**
Preço	1				
GT	**-0,38939344**	1			
Ticker	-0,172558997	0,282424	1		
Volume	-0,02622076	0,116185	0,248704	1	
Retorno	-0,081258846	0,061349	0,04067	0,019913	1

Fonte: Elaborado pela autora.

Na Tabela 13, podemos observar que volume de pesquisas no *Google Trends* pelo nome das empresas tem uma correlação maior com a variável preço de fechamento, em todas as empresas analisadas no modelo. A empresa Vale S.A apresentou uma correlação de 19%, o

banco Itaú Unibanco apresentou uma correlação de 25% (positiva e fraca), e a Petrobras apresentou uma correlação inversa de -38% (negativa e fraca). Em relação ao preço de fechamento das ações e as *tickers*, os valores para estas três empresas foram respectivamente de 9%, -2% e -17% configurando uma correlação muito fraca.

Preis et al. (2013) identificaram padrões que podem ser analisados e interpretados como sinais de alerta do movimento das ações, pois a *Google Trends* não indica somente o atual estado do mercado, mas também as condições futuras. Quando ocorrem quedas significantes no preço, há uma maior preocupação do investidor e, portanto, um maior número de pesquisas.

Consequentemente, quando uma quantidade grande de investidores pesquisar sobre determinado ativo (empresa) ou *ticker*, pode ocorrer impacto no preço e volume de negociação no mercado. No entanto, esse impacto pode não ser significativo, pois para isso é necessário que sejam efetuadas milhares de negociações para acarretar pequenas alterações no preço (PEREIRA; ROSA; BENDER FILHO, 2020).

A variável retorno apresentou uma correlação muito baixa com todas as demais variáveis no modelo proposto para todas as empresas. O volume de negociação, por sua vez, mostrou uma correlação fraca com a variável *Ticker* da ITUB4 e PETR4 com 24% de correlação. A empresa Vale S.A apresentou uma correlação positiva muito baixa, de 7%. Em relação à variável GT, a Petrobras teve uma correlação de 11% com o volume, conforme observado em estudos como Barber & Odean (2008), Guzella et al. (2023), e Paiva (2020). Além disso, o volume apresentou uma correlação inversa fraca de -38% sobre o preço de fechamento na matriz de correlação da empresa Itaú Unibanco, indicado que um volume alto está associado a preços baixos.

As pesquisas no *Google Trends* podem não estar diretamente relacionadas aos retornos das ações, indicando uma ausência de correlação significativa entre elas, o que sugere uma neutralidade das empresas analisadas em relação a esse fator. No entanto, o aumento no volume de pesquisas no Google parece estar associado a um aumento nas negociações e na liquidez das ações, o que pode estar temporariamente vinculado a retornos futuros mais altos (PAIVA, 2020).

A análise da correlação entre o volume de buscas pelo nome da empresa (GT) e pelas buscas pelas *tickers* mostrou uma correlação significativa entre si. Para a empresa Vale, a correlação foi moderada e positiva, com 46%. Já para Itaú Unibanco e Petrobras, a correlação foi fraca e positiva, com 27% e 28%, respectivamente. Esses resultados indicam que, em alguns

casos, os *insights* sobre percepção e o interesse dos investidores e do público em geral podem não estar completamente alinhados.

Durante o período analisado da empresa Vale S.A, observou-se que o número de pesquisas aumentava em períodos de queda nas ações, provavelmente devido à insegurança dos investidores. Um exemplo marcante foi o aumento nas buscas sobre as ações após a tragédia do rompimento da barragem de Brumadinho, em Minas Gerais, em 2019. Nesse evento as ações da Vale S.A caíram 24,52%, e o número de pesquisas tanto pelo nome da empresa quanto pela *ticker* aumentou drasticamente após as notícias e a queda das ações, resultando em uma perda de 70 bilhões de valor de mercado para a empresa.

Esse período também foi marcado pela suspensão do pagamento de dividendos e juros sobre capital próprio, aumentando o pessimismo entre os investidores. Esse cenário pode ter levado ao viés de arrependimento, especialmente porque um desastre similar já havia ocorrido anteriormente em Mariana, Minas Gerais, em 2015. Além disso, as inúmeras acusações de crimes ambientais contra a empresa reforçam o viés de confiança, onde o investidor acredita em suas próprias habilidades de estimativa, mesmo diante de evidências contrárias.

Em 2020, no auge da pandemia, as ações do Itaú Unibanco caíram 35%, resultando em uma perda de cerca de R$ 130 milhões em valor de mercado, período em que muitos investidores migraram para bancos digitais. No entanto, em 2022, as ações do Itaú Unibanco voltaram a ser uma das mais recomendadas pelas corretoras, devido ao bom desempenho impulsionado por fatores externos, tais como a inflação.

Segundo Barber & Odean (2008), a atenção dos investidores individuais é um recurso escasso, já que a vasta quantidade de informações disponíveis dificulta a busca por dados relevantes. Essa limitação se torna ainda mais evidente, pois a maioria desses investidores são compradores, geralmente orientados pela atenção. Eles tendem a ser atraídos por notícias, retornos anormais, ou volume elevados de negociações em determinados ativos. Isso os leva ao viés da representatividade, em que acabam deletando informações cruciais e tomam atalhos mentais, associando informações de forma simplificada, o que pode resultar em decisões subótimas.

Em 2014, a Petrobras foi envolvida em um dos maiores escândalos de corrupção do mundo, revelado pela Operação Lava Jato. Esse evento resultou em um prejuízo de mais de R$ 25 bilhões, causando desinvestimento e descapitalização da empresa, o que impactou drasticamente suas finanças e operações. Mesmo com sinais de recuperação começando a aparecer em 2019, o valor das ações da Petrobras continuou sendo negociado em baixa.

Segundo Paraboni et al. (2016) o cenário econômico de um país é crucial para a tomada de decisão dos investidores. Esse ambiente pode influenciar o nível de conforto dos investidores em relação aos seus investimentos, levando-os ao viés da disponibilidade. Esse viés ocorre quando o investidor dá mais peso às informações que são mais facilmente acessíveis, mesmos que essas informações não reflitam necessariamente a realidade. Isso pode resultar em distorções na avaliação dos investimentos.

Em março de 2020, as ações da Petrobras sofreram uma queda de cerca de 57%, devido a fatores como a pandemia do COVID-19, que afetou a economia global e reduziu a demanda por petróleo, e as disputas entre a Rússia e a Arábia Saudita, que levaram a uma guerra de preços no mercado de petróleo, resultando em uma queda nos preços a níveis não vistos nos últimos 30 anos. Durante esse período, houve um aumento nas pesquisas no Google sobre a empresa e a *ticker*, refletindo um otimismo crescente sobre a possível recuperação da Petrobras.

Em 2021, a Petrobras recuperou cerca de R$5,3 bilhões desviados pela rede de corrupção, segundo o Ministério Público Federal (MPF). As ações da empresa passaram por diversos momentos de queda e volatilidade devido à pandemia e aos escândalos de corrupção mencionados anteriormente. Os dados indicam que, para haver um impacto significativo das pesquisas do *Google Trends* nos preços das ações, é necessário que ocorram muitas pesquisas. De acordo com Latoeiro (2012), essas pesquisas tendem a ter um impacto negativo nos preços das ações.

Desta forma, a influência do volume de pesquisa no Google afeta as decisões dos investidores, tanto profissionais que utilizam de outras ferramentas avançadas quanto investidores individuais iniciantes. Além do Google Trends, outras ferramentas são usadas para identificar os sentimentos dos investidores e observar o efeito no movimento do mercado acionário, como notícias da web e das mídias sociais. Vários fatores afetam o movimento dos preços no mercado acionário, incluindo fatores econômicos, políticos, corporativos e psicossociais.

A principal vantagem de utilizar pesquisas na Web é a geração espontânea de dados pelos investidores. Combinando esses dados com outros índices, pode-se obter novas percepções e um melhor entendimento do complexo comportamento coletivo dos investidores, que nem sempre é descrita pela literatura (MOURÃO, 2018).

5.4.3 Avaliação empírica dos atributos de retorno e de volatilidade

Para compreender o comportamento dos investidores durante o período analisado, foi realizada uma comparação do retorno e da volatilidade dos ativos selecionados. Este estudo é fundamental para entender a relação risco-retorno dos investimentos, ajudando os investidores a avaliarem o desempenho dos ativos e a tomar decisões mais alinhadas com seus objetivos e tolerância ao risco. A Tabela 14 apresenta o retorno das ações da Vale, Itaú Unibanco e Petrobras.

Tabela 14: Retorno das ações Vale, Petrobras e Itaú Unibanco no período de 2019 a 2022.

RETORNO	2019	2020	2021	2022*
VALE3	0,06%	0,29%	0,02%	0,01%
PETR4	0,12%	0,08%	0,12%	0,26%
ITUB4	0,03%	-0,02%	-0,11%	0,18%

Fonte: Elaborado pela autora. *Dados coletados até 30/09/2022.

Como já mencionado, em 2019, a Vale S.A, apresentou um retorno baixo de 0,06%, principalmente devido à tragédia da barragem de Brumadinho, que resultou na morte de 270 pessoas. Esse desastre causou destruição de comunidades, devastação ambiental e impactos socioeconômicos severos, levando a uma perda de cerca de R$ 70 bilhões em valor de mercado. Foi a maior perda já registrada por uma empresa brasileira em um único dia, encerrando o ano com um desempenho negativo.

Em 2020, apesar da pandemia de COVID-19, a Vale S.A registrou o maior retorno entre todos os anos analisados, alcançando 0,29%. O aumento do preço do minério de ferro, a demanda aquecida da China e a produção a baixos custos contribuíram para que a empresa amenizasse os efeitos do desastre de Brumadinho. As ações da Vale acumularam uma alta 64%, tornando 2020 um ano de recuperação para a empresa, que alcançou um lucro líquido de mais de R$24 bilhões.

Em 2021, a Vale S.A registrou um lucro recorde de mais de R$ 120 bilhões. No entanto, o retorno das suas ações foi baixo, devido à queda nos preços do minério de ferro e à estabilidade das ações. As projeções indicam que os preços do minério de ferro continuariam em queda, refletindo um cenário econômico global desfavorável que impactou a demanda por materiais básicos, contribuindo para o baixo retorno das ações da Vale.

A Petrobras apresentou um retorno das ações relativamente estável em 2019 a 2021 com valores de 0,12%, 0,08%, e 0,12%. Em 2019, a empresa registrou um lucro de R$ 40 bilhões, um crescimento de mais de 55% em relação ao ano anterior. Esse aumento significativo no

lucro foi impulsionado, em grande parte, pelo desinvestimento da empresa, que se desfez de diversos ativos, incluindo a venda da BR Distribuidora e da TAG Investimentos.

Em 2022, as ações da Petrobras valorizaram mais de 80%, impulsionadas pelo aumento no preço do combustível e do barril de petróleo, com os preços subindo de R$ 18,61 em junho de 2021, para R$ 34,15 em julho de 2022, resultando em um retorno de 0,26%. O retorno elevado da Petrobras também foi influenciado pela alta do dólar, que favoreceu a empresa, além do aumento no consumo de combustível com o retorno à rotina pós-pandemia e os efeitos da guerra entre Ucrânia e Rússia.

O Itaú Unibanco apresentou os menores retornos das três empresas analisadas. Em 2019, as ações tiveram pouca oscilação e o banco registrou um retorno de apenas 0,03%, além do fechamento de 400 agências no Brasil. Em 2020, a pandemia de coronavírus impactou significativamente o banco, com despesas de provisões que reduziram a rentabilidade em mais de 34%. As ações sofreram quedas significativas, resultando em retornos negativos, que continuaram em 2021, refletindo a queda dos grandes bancos na bolsa de valores. No entanto, em 2022, as ações se recuperaram, acumulando ganhos de quase 40%, e apresentando um retorno positivo de 0,18%, com um lucro de 17% no segundo semestre de 2022, fatores que foram cruciais para a recuperação da empresa.

A volatilidade do retorno é um importante indicador de risco, refletindo a incerteza enfrentada pelos investidores. A frequência de pesquisas na internet também pode ser considerada um bom indicador de risco do mercado, pois há uma relação entre aversão a riscos e a busca por informação (LATOEIRO, 2012). A Tabela 15 mostra a volatilidade das empresas analisadas, em diárias, anuais, mensais, trimestrais e semestrais, com dados coletado de 2019 até 30 de setembro de 2022.

Tabela 15: Volatilidade das ações Vale, Petrobras e Itaú Unibanco no período de 2019 a 2022.

VOLATILIDADE	VALE3	PETR4	ITUB4
Diária	2,713%	2,981%	2,206%
Anual*	43,07%	47,32%	35,02%
Mensal	12,43%	13,66%	10,11%
Trimestral	21,53%	23,66%	17,51%
Semestral	30,45%	33,46%	24,76%

Fonte: Elaborado pela autora. *Dados coletados até 30/09/2022.

A volatilidade é uma variável econômica que avalia a estabilidade de uma ação, preço ou índice, indicando a frequência e a intensidade das variações de preços em um determinado período. É possível estimar possíveis comportamentos futuros dos preços, uma vez que a

volatilidade está diretamente relacionada ao risco. Um mercado com alta volatilidade reflete grande instabilidade nos preços, o que pode indicar maior incerteza para os investidores.

Há evidências de que o aumento de pesquisas no Google sobre uma determinada empresa pode ter efeitos positivos na volatilidade de suas ações. O Itaú Unibanco foi a empresa que apresentou a menor volatilidade entre as analisadas, mesmo durante o período da pandemia e o fechamento de diversas agências. O banco enfrentou menos volatilidade em comparação com outras empresas, devido à desvalorização do real e à inflação, resultando em uma volatilidade diária de 2,206% e a anual de 35,02%. Isso se reflete em uma análise robusta na correlação de Pearson, contrastando com a Vale, que apresentou maior volatilidade ao longo do período analisado.

As empresas Petrobras e Vale, ambas produtoras de commodities, têm seus preços definidos internacionalmente e cotados em dólar, o que as torna mais suscetíveis aos impactos do mercado externo e da demanda global. Esses setores tendem a acompanhar os ciclos econômicos e enfrentam incertezas quantos aos preços futuros.

Em 2021, o minério de ferro produzido pela Vale, alcançou máximas históricas devido à demanda aquecida da China e Estados Unidos. Por outro lado, a volatilidade da Petrobras foi principalmente causada pela alta do petróleo. Ambos os setores experimentam volatilidades significativas devido às flutuações nos preços das commodities e aos ciclos econômicos globais.

Para a empresa Vale, os vieses comportamentais observados foi a heurística da representatividade, que está intimamente ligada à heurística da probabilidade. Esse viés ocorre quando os investidores avaliam a probabilidade de um evento acontecer, como a tragédia de Brumadinho, que foi precedido pelo acidente em Mariana. Quando os preços das ações da Vale caíram devido à tragédia, observa-se o efeito manada, onde os investidores seguiram as decisões de outros, ignorando informações relevantes. Esse pessimismo, gerado pelos eventos, impactou significativamente os preços das ações, que só começaram a se recuperar em 2021 (KUTCHUKIAN, 2010; SAMSON, 2015).

Diferente da Vale, o Itaú Unibanco apresentou o viés do conservadorismo, que está mais relacionado à sensibilidade ao risco. Esse viés se manifesta quando os investidores preferem evitar riscos, mantendo suas decisões em linha com informações já conhecidas e minimizando a exposição a novas incertezas. Como resultado, o Itaú Unibanco se manteve com a menor volatilidade entre as empresas pesquisadas, refletindo uma postura mais cautelosa e neutra no mercado (POMPIAN, 2012).

A Petrobras apresentou uma recuperação nos preços das ações após os escândalos de corrupção em que esteve envolvida, o que gerou otimismo entre os investidores. Esse cenário também revelou o viés do excesso de confiança, onde os investidores acreditaram excessivamente em suas habilidades, subestimando os riscos associados à empresa (YOSHINAGA; RAMALHO, 2014).

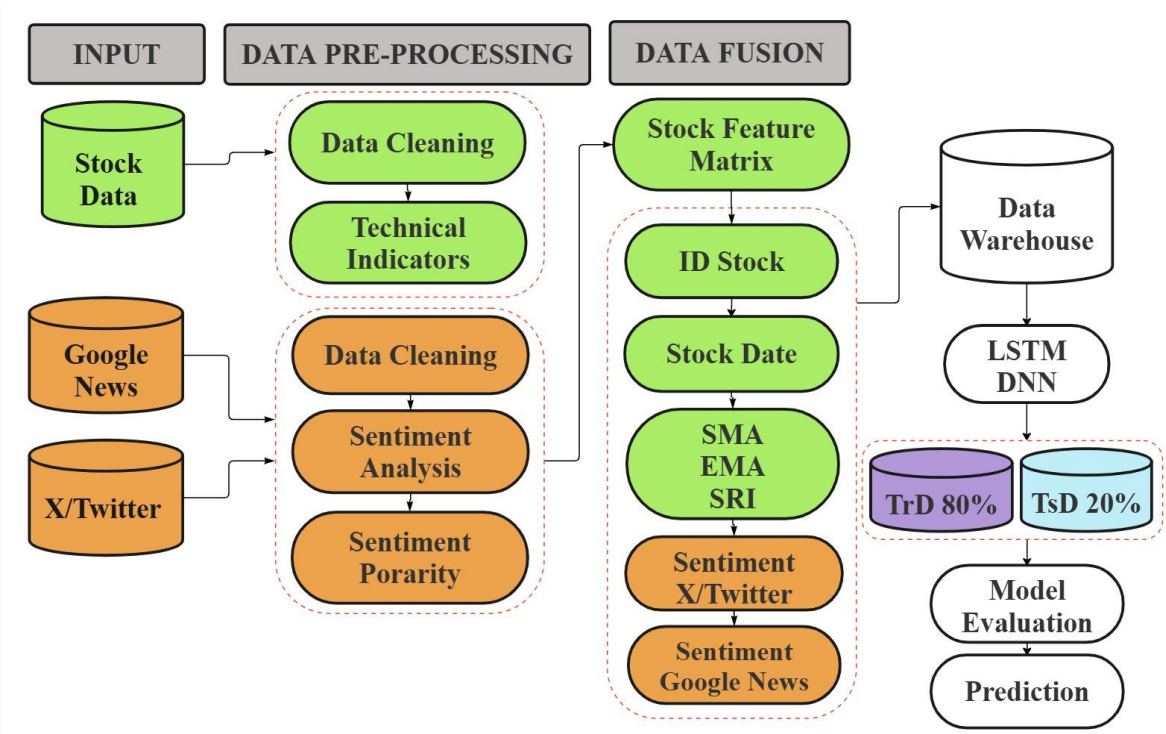
5.5 Experimento 5 – Integração de dados numéricos e textuais

Esta seção apresenta o experimento 5, que teve como objetivo analisar a previsão do comportamento de alguns ativos do mercado de ações brasileiro, por meio da fusão de dados de séries históricas (numéricas) e dados textuais diversificados, com as técnicas de *Deep Learning* e Processamento de Linguagem Natural.

Para analisar esse movimento de preços foram selecionados quatro ativos listados na Bolsa de Valores Brasileira – B3. Também foram explorados dados textuais obtidos de redes sociais *Twitter* (X) e sites de notícias (*Google News*) a partir de técnicas de análise de sentimentos em texto. Desta forma buscamos contribuir para o desenvolvimento de abordagens para gerar previsões de melhor desempenho, apoiando assim a tomada de decisão mais assertiva, através da minimização dos seus riscos e incertezas considerando dados adicionais sobre o mercado e sobre as bolsas de valores.

A Figura 23 ilustra as etapas e elementos envolvidos nesta abordagem. São empregadas bases de dados numéricos com o movimento de ações em conjunto com dados textuais obtidos em duas fontes, que são o *Google News* e o *Twitter*(X). Cada conjunto de dados passa por uma preparação específica, que permite a sua posterior integração em uma base única para utilização no treinamento e avaliação de uma rede neural. Esta rede é utilizada para a predição de valores das ações.

Figura 23: Modelo Geral Proposto do Experimento 5.



Fonte: Elaborado pela autora.

O conjunto de dados utilizado neste estudo inclui, portanto, dados numéricos (séries históricas dos preços das ações) e dados textuais (*Google News* e *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. A seguir são descritos os procedimentos para tratar cada conjunto de dados, seu processamento, integração e os modelos de predição.

5.5.1 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. 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[3]. 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 3.652 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) (NABIPOUR et al., 2020; NTI; ADEKOYA; WEYORI, 2020b).

Estes indicadores são descritos a seguir (PINHEIRO, 2019), sendo a média móvel simples (MMS) descrita na equação (29),

$$\text{MMS} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{29}$$

Onde: 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 (30),

$$MME_d = MME_{d-1} + \frac{2}{n+1} x (P_d - MME_{d-1}) \tag{30}$$

Onde: $MME_{d-1} -$ 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 (31).

$$\text{IFR} = 100 - \left(\frac{100}{1 + FR}\right) \tag{31}$$

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

3https://pypi.org/project/yahoofinancials/

Para padronizar os dados das series históricas, 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 (BISHOP, 2006), descrito na equação (32),

$$\text{normalized\_value} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{32}$$

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

5.5.2 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 2.869 observações do *Google News* e 2.748 observações do *Twitter* para Petrobras (PETR4). Para Vale (VALE3) foram coletadas 1.833 notícias no *Google News* e 1915 mensagens do *Twitter*. Para o Bradesco (BBDC4) foram obtidas 1.782 notícias do *Google News* e 1.585 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 Quadro 8 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[4] no período predeterminado. Seguindo uma abordagem semelhante à proposta (JOSHI; H. N; RAO, 2016), onde as notícias foram coletadas e integradas às séries históricas de modo que os dados já estivessem estruturados e processados para classificação.

4https://pypi.org/project/pygooglenews/

Quadro 8: 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 Inovação 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 à Braskem - Veja
15/09/2010	MP vai investigar se policiais de SP quebraram sigilos para Petrobras – Globo.com
Postagens extraída do X (antigo *Twitter*) *tweets*	
23/10/2009	Petrobras, TAM e NET captam US$ 5,6 bilhões
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 bilhões

Fonte: Elaborado pela autora.

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 utilizado a API fornecida pelo Twitter[5], conforme estudo (BRIGGS, 2020). Foram empregadas quatro fontes de notícias do mercado financeiro sendo elas, InvestingBrasil[6,] Money Times[7], InfoMoney[8] e Valor Econômico[9].

5.5.2 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 (SCHUMAKER; CHEN,

5https://developer.x.com/en/docs/twitter-api/

6https://twitter.com/InvestingBrasil

7https://twitter.com/leiamoneytimes

8https://twitter.com/infomoney

9https://twitter.com/valoreconomico

2009), 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*" (HONNIBAL; MONTANI, 2017). 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.

5.5.3 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 (ALMEIDA, 2018), que é uma variação da biblioteca *Vader Sentiments* (HUTTO; GILBERT, 2014), 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 positiva, negativa e neutra. A Tabela 16 apresenta alguns exemplos de textos analisados e sua classificação.

Tabela 16: Exemplos de Análise de Sentimentos das Notícias.

Data	Sentimentos	Textos	Compound	Positive	Negative	Neutral
18/08/2010	Positive	Brasil tem sete entre 500 maiores empresas do mundo, aponto "Fortune"	0.8689	0.495	0	0.505
07/07/2014	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

Fonte: Elaborado pela autora.

Como exemplifica a Tabela 16, os valores relativos aos percentuais entre 0 e 1 são compostos com uma métrica de composição dos sentimentos chamada *compound*. 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.

5.5.4 Fusão do Conjunto de Dados

Para a fusão dos dados numéricos e textuais foram combinados três diferentes conjuntos de dados. São eles os dados numéricos normalizados das séries históricas de preços das ações, os dados polarizados pelas análises de sentimentos de notícias do *Google News* e os dados polarizados das postagens provenientes do *Twitter*.

Para as séries históricas de preços foram utilizados os dados "data", "fechamento", "abertura", "mínimo", "máximo" e "volume". No caso dos dados do *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 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 17, no qual as ocorrências de valores "NaN" são substituídas de acordo com este método.

Tabela 17: 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	-1.0	-1.0
24/05/2022	0.045	NaN	-0.5

Fonte: Elaborado pela autora.

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.

5.5.5 Composição dos Dados para Treinamento

Para a composição dos dados de treinamento dos modelos utilizou-se uma série de procedimentos para preparar todas as informações necessárias para os modelos. Foram utilizados 80% dos dados para treino do modelo e 20% para seu teste. Logo após esta etapa, foram estabelecidas três variáveis denominadas de *close_price_shifted, compound_gn_shifted e compound_tw_shifted*, utilizadas para armazenar dados referentes a um dia de antecedência, como ilustra a Tabela 18. Foi usada a variável *close_price_shifted* para os dados de valores de fechamento e as variáveis *compound_gn_shifted* e *compound_tw_shifted* foram utilizadas para os sentimentos polarizados.

Tabela 18: Exemplo do Processamento e Unificação dos Dados.

Date	close_price	close_price_shifted	compound_gn	compound_gn_shifted	compound_tw	compound_tw_shifted	volume	open_price	high	low
06/01/2021	30.100	31.000	0.7906	0.7096	0.6369	0.0258	96562500	30.160	30.900	30.049
07/01/2021	31.000	31.120	0.7906	0.4404	0.0258	0.9062	56171300	30.340	31.150	30.340
08/01/2021	31.120	30.860	0.4404	0.0000	0.9062	0.1531	67136300	31.459	31.760	30.350
11/01/2021	30.860	30.629	0.0000	0.0000	0.1531	0.6249	48744700	30.610	31.059	30.400
12/01/2021	30.629	29.150	0.0000	0.6369	0.6249	0.9744	65691900	31.120	31.559	30.629
13/01/2021	29.150	29.450	0.6369	-0.1280	0.9744	0.9300	93826600	30.680	30.860	29.000
14/01/2021	29.450	28.120	-0.1280	0.0000	0.9300	0.7184	50745400	29.170	29.670	28.710
15/01/2021	28.120	28.690	0.0000	0.4588	0.7184	0.5859	80673300	29.049	29.080	28.030
19/01/2021	28.690	28.209	0.4588	0.4215	0.5859	0.9716	61656000	28.480	28.860	27.639
20/01/2021	28.209	27.549	0.4215	0.6808	0.9716	0.7003	60306200	28.950	29.120	28.110

Fonte: Elaborado pela autora.

Os modelos de predição empregados foram: modelos *Long Short-Term Memory* – LSTM sem retroalimentação, LSTM com retroalimentação e modelo de *Deep Neural Network* – DNN.

Foi 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.

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.

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

Portanto, 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 (KIM, 2016). Assim, o estudo propôs uma comparação para prever o comportamento das ações, observando qual modelo se adequa melhor à captura e dinâmica dos ativos selecionados.

As métricas de avaliação escolhidas foram: Erro Médio Quadrático (MSE), Raiz Quadrada do Erro Quadrático Médio (RMSE), Erro Médio Absoluto (MAE) e Acurácia. Essas métricas oferecem uma avaliação equilibrada e detalhada dos modelos de previsão (LIU; LONG, 2020; NTI; ADEKOYA; WEYORI, 2020b, 2020c).

Enquanto o MSE e o RMSE medem erros quadráticos e são sensíveis a grandes desvios, o MAE oferece uma medida robusta dos erros médios, e a acurácia avalia a consistência das previsões.

A abordagem utilizada da Acurácia, foi uma métrica de precisão adaptada para problemas de regressão, complementando métricas tradicionais como MSE, MAE e MAPE.

Seja y_i o valor real e $\hat{y}_i$ o valor previsto, então a previsão é considerada correta se:

$$|y - \hat{y}_i| \leq 0.02 \cdot y_i$$

$$Acurácia = \frac{Número\ de\ Previsões\ Corretas}{Total\ de\ Previsões} \tag{33}$$

Essa métrica fornece uma análise mais abrangente e intuitiva do desempenho do modelo ao considerar previsões “aceitáveis” dentro de uma margem de erro percentual definida. Definimos

um limite de 2% em relação aos valores reais, onde uma previsão é considerada correta se a diferença absoluta entre os valores previstos e reais for menor ou igual a 2% do valor real. Assim, essas métricas permitem uma análise quantitativa abrangente do desempenho dos modelos. Assim, essas métricas permitem uma análise quantitativa abrangente do desempenho dos modelos.

5.5.6 Resultados e Análise da Fusão de Dados

Este experimento inicialmente foi realizado uma análise das fusões das fontes dados para avaliar o resultado de cada uma com os diferentes modelos descritos. Nesta primeira etapa foi utilizado um único ativo para testar os três modelos descritos e as diversas combinações de dados. Para tal foi utilizado o ativo da empresa Petrobras (PETR4).

A Petrobras é maior empresa de petróleo e gás do Brasil, com um impacto significativo na economia do país e no mercado global de energia. Suas ações são altamente sensíveis a mudanças nos preços do petróleo, decisões políticas, e eventos econômicos globais.

Em seguida, foi selecionada a combinação de dados mais promissora de acordo com os resultados deste experimento inicial, sendo esta combinação e o modelo utilizados em experimentos com os demais ativos.

A combinação dos dados com melhor resultado foi a composta por *Google News*, *Twitter*, cálculo do índice de forca relativa (IFR) e média móvel exponencial (MME). Com esta combinação foi realizada a aplicação do experimento a 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, de setores econômicos variados como petróleo, mineração, e bancos, têm relevâncias no mercado da B3 e no índice Ibovespa. Elas são amplamente cobertas por notícias e discussões em mídias sociais, oferecendo uma grande quantidade de dados textuais e numéricos, o que permitiu uma análise abrangente das dinâmicas de mercado e das reações dos setores e eventos econômicos e financeiros.

A primeira combinação, denominada “Sem nenhum parâmetro”, se refere ao uso dos dados numéricos de séries históricas de preços das ações unicamente. As demais configurações envolvem sempre os dados numéricos acrescidos das demais fontes de dados textuais e de indicadores calculados adicionalmente. Foram realizadas combinações utilizando a média móvel simples (MMS) e a média móvel exponencial (MME) para as predições. Foram utilizados para cálculos

das médias um período de 5 dias e de 30 dias para séries históricas do preço de fechamento, além do índice de força relativa (IFR) para 14 dias.

A Tabela 19 exibe os resultados das métricas de desempenho das combinações estudadas e apresenta o quadro completo de combinações utilizadas para os experimentos com os ativos da empresa Petrobras. A acurácia foi utilizada para verificar a proporção de previsões dos modelos para o comportamento dos preços do ativo da Petrobras. O modelo 2 (LSTM) com retroalimentação apresentou o melhor desempenho, com uma acurácia de 53,60% para combinação completo (Google News+Twitter+IFR+MME) com média móvel exponencial, comparado a 52,91% para o modelo 1 e 50% para o modelo 3 (DNN).

Tabela 19: Métricas de desempenho das combinações dos dados da empresa Petrobras.

Modelo 1 - LSTM	**MSE**	**RMSE**	**MAE**	**Accuracy**
Sem nenhum parâmetro	0,00012	0,01090	0,01272	50,28%
Apenas Google News	0,00013	0,01130	0,01239	47,92%
Google News + Twitter	0,00013	0,01120	0,01232	48,89%
Google News + Twitter + IFR	0,00014	0,01160	0,01252	51,66%
Google News + IFR	0,00013	0,01130	0,01252	50,42%
Google News + IFR + MMS	0,00013	0,01120	0,01348	51,52%
Google News + IFR + MME	0,00013	0,01130	0,01356	50,97%
Apenas Twitter	0,00013	0,01150	0,01241	51,11%
Twitter + IFR	0,00014	0,01190	0,01268	51,66%
Twitter + IFR + MMS	0,00014	0,01170	0,01363	52,63%
Twitter + IFR + MME	0,00014	0,01170	0,01284	51,25%
Google News + Twitter + IFR + MMS	0,00013	0,01160	0,01375	52,35%
Google News + Twitter + IFR + MME	0,00014	0,01160	0,01367	52,91%
Modelo 2 - LSTM	**MSE**	**RMSE**	**MAE**	**Accuracy**
Sem nenhum parâmetro	0,00014	0,01170	0,01717	52,49%
Apenas Google News	0,00012	0,01110	0,01242	49,45%
Google News + Twitter	0,00013	0,01130	0,01249	49,31%
Google News + Twitter + IFR	0,00015	0,01210	0,01305	48,61%
Google News + IFR	0,00015	0,01210	0,01324	48,20%
Google News + IFR + MMS	0,00012	0,01090	0,01304	51,39%
Google News + IFR + MME	0,00013	0,01140	0,01316	50,00%
Apenas Twitter	0,00012	0,01100	0,01252	49,45%
Twitter + IFR	0,00013	0,01130	0,01304	50,69%
Twitter + IFR + MMS	0,00013	0,01130	0,01512	52,08%
Twitter + IFR + MME	0,00014	0,01180	0,01679	51,80%
Google News + Twitter + IFR + MMS	0,00015	0,01220	0,01482	51,25%
Google News + Twitter + IFR + MME	0,00015	0,01230	0,01653	53,60%

Modelo 3 - DNN	MSE	RMSE	MAE	Accuracy
Sem nenhum parâmetro	0,00011	0,01060	0,01460	48,75%
Apenas Google News	0,00017	0,01290	0,01496	47,51%
Google News + Twitter	0,00023	0,01500	0,01484	50,97%
Google News + Twitter + IFR	0,00034	0,01850	0,02812	48,61%
Google News + IFR	0,00017	0,01320	0,01695	50,00%
Google News + IFR + MMS	0,00024	0,01560	0,01915	46,81%
Google News + IFR + MME	0,00025	0,01580	0,03923	50,00%
Apenas Twitter	0,00024	0,01560	0,01549	50,83%
Twitter + IFR	0,00026	0,01610	0,01872	46,40%
Twitter + IFR + MMS	0,00012	0,01100	0,01866	48,48%
Twitter + IFR + MME	0,00018	0,01350	0,01599	50,28%
Google News + Twitter + IFR + MMS	0,00050	0,02240	0,02526	47,65%
Google News + Twitter + IFR + MME	0,00023	0,01510	0,01881	50,00%

Fonte: Elaborado pela autora.

Em relação aos erros médios, MSE e RMSE, bem como ao e erros médio absoluto MAE, os modelos de LSTM apresentaram menores erros em relação ao modelo DNN. Podemos observar na Tabela 19 que os resultados demostraram que os sentimentos dos *tweets* e *Google News*, com a combinação dos dados numéricos, apresentaram uma melhora no resultado da previsão das ações. Os *tweets* tiveram um melhor desempenho do que as notícias do *Google News*, como observado no estudo (NTI; ADEKOYA; WEYORI, 2020b), como observado no capítulo 3.

Observa-se que, entre os modelos analisados, o LSTM com retroalimentação oferece uma melhor dinâmica para o conjunto de dados utilizados, capturando dependências de longo prazo em séries temporais. O DNN teve um desempenho significativo, mas enfrentou dificuldades em capturar o comportamento específico dos eventos.

Uma deficiência enfrentada pelas redes neurais recorrentes é desaparecimento do gradiente, as redes com muitas camadas perdem a informação do gradiente. Para solucionar este problema foi desenvolvida a estrutura de *Long Short Term Mermory* LSTM, foi a inserção de uma memória à NN para manter seu *status* durante vários passos de tempo (KIM, 2016).

Observa-se que os resultados dos experimentos com as combinações que acrescentam um maior número de elementos aos dados numéricos originais apresentam melhorias, evidenciando o impacto positivo desta abordagem. Em todos os experimentos a combinação que apresenta o melhor contexto é aquela com *Google News*, *Twitter*, índice de força relativa (IFR) e média móvel exponencial (MME). Desta maneira, foram realizados experimentos com esta combinação para todos os ativos, como pode ser observado na Tabela 20, que apresenta as métricas de avaliação para os três modelos considerando-se a combinação de dados de melhor desempenho.

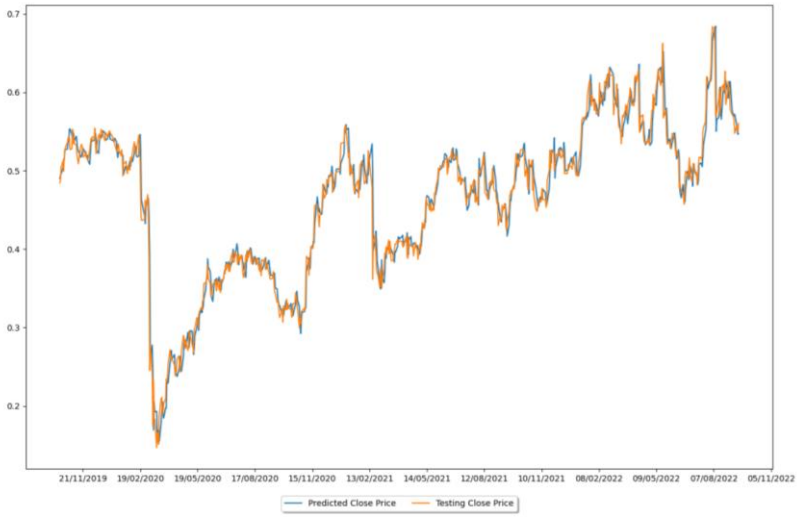
Tabela 20: Métricas de desempenho.

Ações	LSTM 1			LSTM 2			DNN		
	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
PETR4	0,00013	0,0114	0,013137	0,000123	0,0111	0,012187	0,00044	0,021	0,018075
VALE3	0,000113	0,0106		0,000164	0,0128	0,024508	0,002047	0,0452	0,039187
BBC4	0,000142	0,0119	0,017411	0,000125	0,0112	0,027252	0,000548	0,0234	0,023953
ITUB4	0,000137	0,0117	0,016797	0,00014	0,0119	0,016637	0,000323	0,018	0,018996

Fonte: Elaborado pela autora.

A empresa Petrobras, com o ativo PETR4, apresentou o melhor desempenho em suas métricas nos três modelos testados, exibindo os menores erros. Os valores de MSE foram 0,000130, 0,000123 e 0,000440. Como ilustra a Tabela 20, os modelos LSTM 1 e LSTM 2 apresentaram as melhores performances, com evidenciado pelo RMSE: 0,0114; 0,0111 e 0,021. O MAE, uma métrica mais robusta dos erros médios, apresentou valores de 0,013137, 0,012187 e 0,18075. A Figura 24 mostra o comportamento da predição do modelo 2 para o preço de fechamento das ações da PETR4, que apresentou os menores erros.

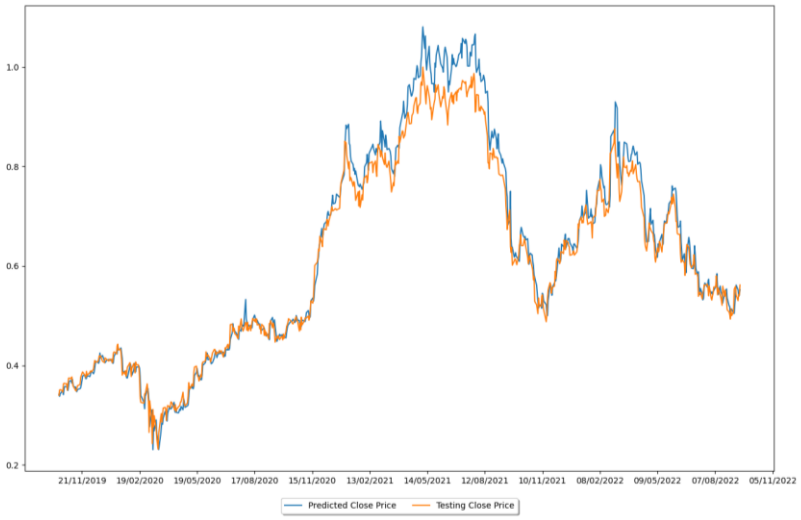
Figura 24: Comportamento das predições do modelo 2 para o preço de fechamento da empresa Petrobras PETR4.



Fonte: Elaborado pela autora.

A figura 25 apresenta o comportamento da predição do modelo 2 para o preço de fechamento das ações da VALE3. A empresa Vale, ativo VALE3, apresentou MSE de 0,000113, 0,000164 e 0,002047 mostrando o desempenho com menor erro no modelo LSTM 1. Os modelos LSTM 1 e LSTM 2 apresentaram as melhores performances. O RMSE: 0,0106; 0,0128 e 0,0452. O MAE, apresentou valores de 0,0024508 para LSTM 2 e 0,039187 para DNN.

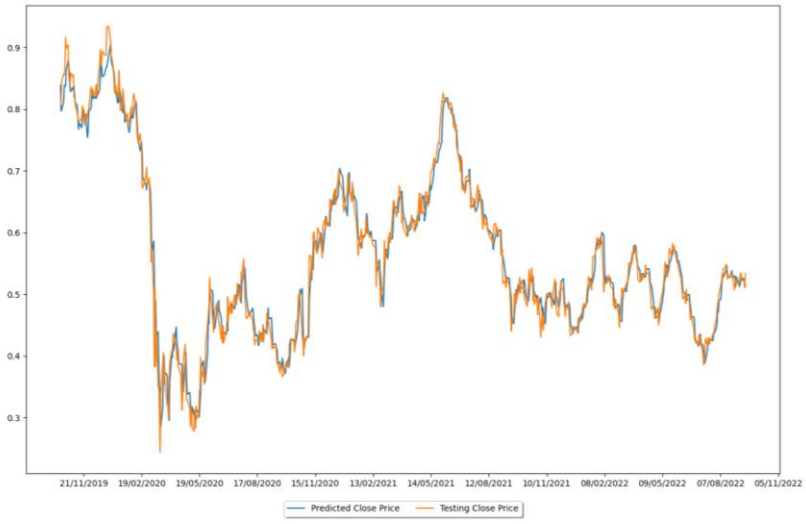
Figura 25: Comportamento das predições do modelo 2 do preço de fechamento da empresa Vale VALE3.



Fonte: Elaborado pela autora.

A empresa Bradesco, ativo BBDC4, nos modelos LSTM 1 e LSTM 2 apresentou MSE de 0,000142 e 0,000125. Mostrou assim o melhor desempenho no modelo LSTM 2. A figura 26 apresenta o comportamento da predição do modelo LSTM 1, o preço de fechamento das ações da BBDC4, modelo que apresentou menores erros no experimento. O RMSE: 0,0119; 0,0112 e 0,0234. O MAE, apresentou valores de 0,017411, 0,027252, e 0,023953.

Figura 26: Comportamento das predições do modelo 1 do preço de fechamento da empresa Bradesco BBDC4.



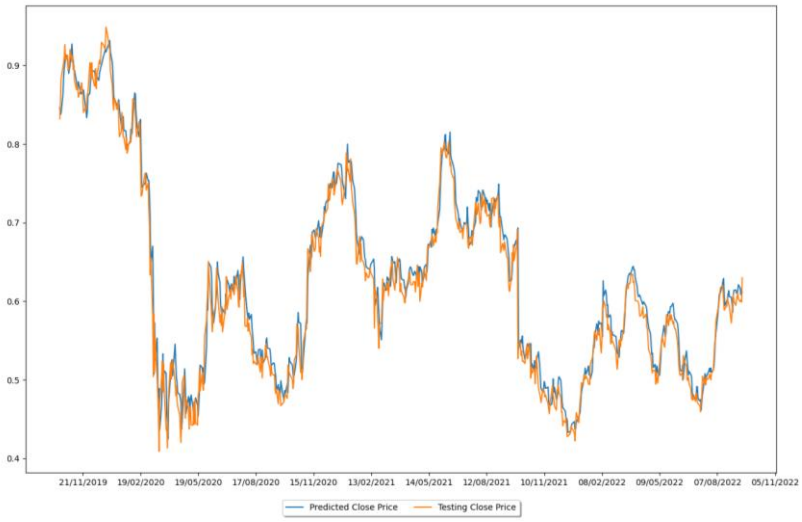
Fonte: Elaborado pela autora.

A empresa Itaú Unibanco ITUB4 teve um bom desempenho nos três modelos, apresentando MSE de 0,000137, 0,000140 e 0,00323, com melhor desempenho no modelo LSTM 1. A figura 27 apresenta o comportamento da predição do modelo LSTM 1 com o preço de fechamento das ações da ITUB4. O RMSE: 0,0117; 0,0119 e 0,018. O MAE, apresentou valores de 0,016797, 0,016637, e 0,018996.

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.

Os resultados mostraram que o modelo com LSTM apresenta mais eficiência na previsão de séries temporais de preço de ações do que o modelo com DNN, que apresentou o pior desempenhos para todos os ativos analisados. Os ativos da PETR4 e ITUB4 apresentaram os melhores desempenhos entre os três modelos avaliados, seguidos pelos ativos BBDC4 e VALE3.

Figura 27: Comportamento das predições do modelo 1 do preço de fechamento da empresa Itaú Unibanco ITUB4.



Fonte: Elaborado pela autora.

A empresa Vale apresentou maior volatilidade no preço de suas ações devido ao desastre da barragem de Mariana -MG (VARGAS et al., 2022). Podemos observar que os ativos enfrentam desafios significativos devido à Hipótese de Eficiência dos Mercados (HME). De acordo com a HME (FAMA, 1965, 1970), 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.

Desta forma, 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 (FAMA, 1991; MALKIEL; FAMA, 1970; STATMAN, 1995; THALER, 2016).

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 menores erros e melhores previsões.

5.6 Análise Crítica dos Resultados

O primeiro experimento teve como objetivo identificar as melhores técnicas de inteligência computacional para entendimento e previsão de tendências de preços da Bolsa de valores brasileira. Foram utilizados dados de preço, volume e quantidade para fazer a previsão por meio de aprendizado de máquina, buscando apreender e evidenciar padrões e tendências no comportamento de preços de ativos. Observou-se o comportamento de um ativo listado na Bolsa de Valores brasileira B3, da empresa Petrobras PETR4. Alguns apresentaram modelos com resultados promissores, como a regressão linear, a floresta aleatória e redes neurais. Os modelos SVM e kNN também demonstraram bom desempenho, com intervalos de erro consistentes entre os modelos, que se mostraram competitivos.

O segundo experimento teve como objetivo prever os preços das ações da Petrobras PETR4 utilizado o modelo de *Long Short Term Memory* (LSTM). O modelo utilizou dados históricos dos cinco dias anteriores para prever o sexto dia. O conjunto de dados históricos desde experimento consistiu em 6.269 observações diárias, uma base de dados mais robusta em comparação ao primeiro, que contou com 2.571 observações diárias. Além disso, o algoritmo utilizado no segundo experimento empregou a técnica de aprendizado profundo.

O terceiro experimento teve como objetivo identificar os efeitos das variáveis macroeconômicas e setoriais no comportamento dos preços das ações da Petrobras (PETR4) ao

longo do período analisado, de 2006 a 2022. A pesquisa revelou que fatores como taxa de juros (Selic), EMBI+ risco-Brasil e preço do barril de petróleo Brent apresentaram relações significativas com as oscilações nos preços das ações, destacando sua relevância na dinâmica do mercado financeiro. Além disso, a análise demostrou que eventos externos e internos, como crises econômicas globais, instabilidades políticas e questões específicas da empresa, desempenham papéis fundamentais na formação dos preços das ações, evidenciando que o mercado acionário brasileiro é altamente sensível a choques econômicos e políticos.

Por outro lado, o índice do *Google Trends*, usado para medir o interesse público, não apresentou significância estatística no modelo proposto. Entretanto, essa variável pode ser considerada em análises complementares, pois tem potencial para capturar aspectos comportamentais dos investidores em cenários específicos.

O quarto experimento teve como objetivo identificar o efeito das pesquisas da *Google Trends* nas decisões dos investidores entre os anos de 2019 a setembro de 2022. Foram avaliados os impactos das buscas pelo nome da empresa e pelas *tickers* sobre o preço, retorno e volume dos ativos listado na Bolsa de Valores Brasileira (B3), especificamente PETR4, VALE3 e ITUB4. Utilizou-se o coeficiente de correlação Linear de Pearson e compararam-se os retornos e a volatilidade.

Os resultados sugerem que esse campo de pesquisa é promissor, indicando que a demanda por informações dos investidores é parcialmente atendida pelo *Google*. Contudo, o impacto das buscas nos preços é pequeno, de curto prazo e momentâneo, aparentemente incapaz de gerar retornos elevados a longo prazo. O volume de buscas no *Google Trends* pelo nome da empresa apresentou baixa correlação com os preços das ações, enquanto as *tickers* tiveram uma correlação ainda menor, não significativo para o movimento do preço, mas com uma baixa correlação com o volume de negociações de alguns ativos. Em relação ao retorno, não houve correlação significativa com nenhuma das variáveis estudadas.

Destaca-se que o volume de pesquisa no *Google Trends* é um dos fatores que pode influenciar o comportamento e o movimento do mercado acionário, juntamente com indicadores de análise de sentimentos, como notícias da web e mídias sociais, além de fatores econômicos, políticos, sociais, corporativos e psicológicos.

Por fim, o quinto experimento buscou estudar o comportamento e a movimentação de alguns ativos do mercado de ações brasileiro por meio da integração de dados numéricos e dados textuais. Realizou-se a análise de sentimentos em textos utilizando técnicas de processamento de linguagem natural, permitindo a integração de dados. Os experimentos

analisaram o comportamento de ativos listados na Bolsa de Valores Brasileira, como PETR4, VALE3, BBDC4 e ITUB4, empregando modelos de *Long Short Term Memory* (LSTM) e *Deep Neural Network* (DNN). Os resultados indicaram ganhos de desempenho com a integração de dados numéricos e textuais, aumentando a acurácia nos experimentos com conjuntos de dados ampliados. Avalia-se que esses resultados podem ser aprimorados com o aumento do volume de dados disponível para treinamento. Especificamente, o modelo LSTM mostrou-se mais eficiente na previsão dos movimentos dos preços das ações em comparação com o modelo com DNN. Para Petrobras (PETR4), os valores de RMSE foram 0,0114; 0,0111 e 0,021. Para a Vale (VALE3), os valores de RMSE foram 0,0106; 0,0128 e 0,0452. O Bradesco (BBDC4) apresentou RMSE de 0,0119; 0,0112 e 0,0234. finalmente, para o Itaú Unibanco (ITUB4), os valores de RMSE foram 0,0117; 0,0119 e 0,018

Os experimentos realizados evidenciam que a aplicação de técnicas de inteligência artificial e aprendizado profundo no mercado financeiro brasileiro apresentam potencial significativo para a previsão de preços e análise de comportamento de ativos. Modelos como LSTM e DNN demonstraram desempenho superior na integração de dados numéricos e textuais, destacando-se frente a técnicas mais tradicionais, como regressão linear e kNN. No entanto, fatores econômicos, políticos e comportamentais, como taxas de juros, preço do petróleo e variáveis de interesse público, mostram-se igualmente relevantes para compreender as oscilações no mercado, indicando que a previsão precisa exigir um modelo que combine aspectos numéricos e textuais.

O tema da pesquisa é de grande relevância para economia global, demandando mais estudos nesta área para melhorar as previsões de preços utilizando conjuntos de dados que relacionem todos os fatores que afetam as tendências de oscilações dos ativos. A aplicação de técnicas de IA, aprendizado de máquina e aprendizado profundo para mercado financeiro possibilita o desenvolvimento de métodos mais eficientes para os investidores alocarem ativos ao longo do tempo, gerando previsões que melhoram o desempenho no mercado e auxiliam na tomada de decisões frente a riscos e incertezas. Apesar disso, a previsão do mercado de ações permanece desafiadora devido à alta volatilidade e à não linearidade dos dados, demandando novas ferramentas para aumentar a previsibilidade do mercado acionário com maior precisão, através do desenvolvimento de modelos híbridos que combinem diferentes abordagens.

6 CONCLUSÃO

Com o avanço da tecnologia e o crescimento exponencial do volume de informações disponíveis na web, a utilização dessas informações no processo decisório tornou-se essencial para tomadas de decisões no mercado financeiro. Portanto este trabalho buscou identificar e avaliar como publicações em mídias sociais e notícias financeiras podem ser integrados com dados históricos do mercado acionário, ampliando a capacidade preditiva e a compreensão do comportamento das ações.

Dessa forma, o problema de pesquisa foi respondido e os objetivos foram alcançados por meio dos resultados obtidos na análise da integração de dados numéricos e textuais para a previsão de preços de ativos na Bolsa de Valores B3. Os experimentos demonstraram que a inclusão de dados textuais, como notícias e postagens em mídias sociais, melhora a precisão das previsões quando combinados com séries temporais de preços e indicadores técnicos. Além disso, a comparação entre os modelos evidenciou a superioridade do LSTM em relação ao DNN na captura de padrões complexos do mercado. Esses achados reforçam a relevância da abordagem proposta e indicam o potencial da metodologia para aplicações em outros mercados, contribuindo para o avanço das técnicas de modelagem preditiva no setor financeiro.

Cada ativo apresenta um comportamento distinto, influenciado por fatores como volume de notícias, nível de liquidez e sensibilidade a eventos de mercado. Ativos com maior liquidez tendem a ter um fluxo informacional mais intenso, o que pode impactar a previsibilidade dos modelos. Dessa forma, para que a generalização do modelo seja viável em outros ativos e mercados, é essencial que o comportamento dos preços apresente padrões semelhantes. A aplicabilidade da abordagem proposta depende, portanto, da identificação de ativos cujas dinâmicas de precificação compartilhem características estruturais, garantindo a consistência dos resultados preditivos.

A análise de sentimentos expressos nas mídias sociais permite identificar opiniões, emoções e atitudes que podem influenciar o mercado acionário. Essas informações, em conjunto com notícias econômicas e financeiras, impactam nas decisões de compra e venda de ativos, contribuindo para as oscilações nos preços. Embora os investidores tradicionalmente utilizem análises fundamentalistas para avaliar demonstrações financeiras e notícias de

mercado, a exploração desses dados não estruturados (textuais) em modelos preditivos de aprendizado de máquina ainda é limitada.

Este estudo propôs avaliar os impactos da combinação de dados numéricos e textuais na previsão de movimentos do mercado acionário brasileiro. Por meio de técnicas avançadas de processamento de linguagem natural (PLN) e aprendizado profundo (DL), buscamos integrar fontes de dados diversas e testar sua influência sobre a precisão preditiva. A seguir, discutem-se as principais contribuições e limitações do trabalho.

6.1 Contribuições

As principais contribuições desta tese incluem a revisão de estudos com foco na integração de técnicas de aprendizado de máquina e aprendizado profundo com o processamento de linguagem natural para a predição de ações; o estudo de fontes de dados numéricos, como preços e volume das ações, e dados textuais, como *tweets* e notícias da web; além da identificação de lacunas e oportunidades para a integração de dados estruturados (numéricos) e não estruturados (textuais).

Por meio de uma revisão de literatura dos trabalhos relacionados, foram identificados estudos na literatura atuando sobre a previsão do movimento dos preços das ações, com foco em análises baseadas em dados numéricos (análise técnica) e dados textuais (análise fundamentalista). A partir das principais lacunas observadas, foi proposto um modelo para abordar essas questões.

Foi proposto um modelo que combina dados numéricos (estruturados) que incluem preços e volumes de ações, variáveis macroeconômicas, o índice *Google Trend* e os dados textuais (não estruturados), que incluem informações contábeis (dados fundamentais), postagens X/*Twitter*, notícias web e fóruns de discussão (dados alternativos). Destaca-se a ampliação do número de fontes de dados utilizadas, totalizando sete categorias, para melhorar a compreensão das múltiplas influências no mercado acionário.

Nesta tese foram desenvolvidos cinco experimentos com três categorias para abordar as principais lacunas identificadas nos estudos. A primeira categoria de experimentos utiliza algoritmos de aprendizado de máquina e aprendizado profundo para prever movimentos do mercado acionário com base em dados numéricos exclusivamente. A segunda categoria de experimentos analisa a relação entre variáveis econômicas, volume de buscas no *Google Trends* e o comportamento do mercado, avaliando seu impacto no sentimento dos investidores, medido

de acordo com as variações observadas nos preços das ações. Por fim, a terceira categoria integra dados numéricos e textuais para a predição, utilizando algoritmos de aprendizado profundo.

O desenvolvimento dos experimentos teve aplicação de modelo no contexto da Bolsa de Valores Brasileira (B3), considerando as especificidades do mercado local, como maior volatilidade, influência de notícias locais e menor nível de transparência em comparação a mercados maduros. A análise das ações listadas na B3 contribui para o desenvolvimento de modelos preditivos adaptados às características de economias emergentes.

A utilização de algoritmos de aprendizado profundo (*Deep Learning*) e processamento de linguagem natural (PLN) para capturar padrões complexos em dados textuais (não estruturados), como sentimentos e eventos, e correlacioná-los com as oscilações nos preços das ações, expande a capacidade preditiva em comparação as abordagens tradicionais, que se baseiam apenas em dados históricos quantitativos.

Este estudo incorpora diversas fontes de dados nos experimentos para avaliar seu impacto no movimento dos preços em diversos cenários, incluindo variáveis macroeconômicas (taxa de juros, câmbio, preço do petróleo), índice de tendências do Google, indicadores técnicos e eventos textuais. Isso amplia a abrangência do modelo, permitindo a análise dos impactos de múltiplos fatores interdependentes.

A pesquisa destaca a importância dos sentimentos expressos em mídias sociais e notícias da web, demostrando como esses fatores podem complementar as análises técnicas e fundamentalistas. Essa abordagem pode contribuir para uma melhor compreensão do impacto comportamental no mercado financeiro, contribuindo com a literatura de finanças comportamentais.

O experimento descrito no item 5.5 demostra a integração dos dados numéricos e textuais, trazendo contribuições para pré-processamento e fusão desses dados, incluindo técnicas para sincronização temporal, remoção de ruído e agrupamento por eventos e sentimentos, bem como o uso de técnicas de PLN em português. Esse processo pode ser replicado em outros contextos ou mercados.

A abordagem proposta aprimora a capacidade de previsão em mercados caracterizados por incertezas, riscos e oscilações frequentes, como o brasileiro. Isso tem implicações práticas relevantes para investidores, gestores de fundos e pesquisadores que buscam entender melhor as dinâmicas de preços em mercados voláteis, servindo como base para estudos futuros,

promovendo avanços na integração de dados para medir os fatores que impactam as flutuações de preços das ações.

6.2 Limitações do Estudo

Apesar das contribuições, este estudo apresenta algumas limitações. O principal desafio destacado relaciona-se com o pré-processamento dos dados, especialmente na identificação e o agrupamento de variáveis relevantes, além da dificuldade em correlacionar eventos textuais com dados numéricos em linhas temporais distintas. A falta de sincronização temporal entre diferentes fontes de dados pode introduzir ruído nos modelos preditivos.

Outra limitação é a irrelevância de algumas informações extraídas das notícias ou sentimentos, considerando que nem todas impactam significativamente nos preços das ações. Por exemplo, a divulgação de demonstrações financeiras pode influenciar os preços no dia do anúncio, mas não necessariamente nos dias subsequentes.

A qualidade dos dados textuais também é uma preocupação, pois muitas fontes contêm informações redundantes ou irrelevantes para o contexto do mercado acionário, o que pode reduzir a precisão do modelo. Além disso, a interpretação de *nuances* linguísticas em português, como ironia e sarcasmo, representa um desafio adicional para os modelos de PLN. Embora existam poucos modelos de processamento de linguagem natural para o português, especialmente voltado para mercado financeiro, muitas vezes esses não apresentam um bom desempenho na análise de sentimentos. Por isso, muitos estudos optam por bibliotecas em inglês e fazem tradução.

Quanto à aplicabilidade dos resultados, ela pode ser restrita ao mercado acionário brasileiro, que 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 no experimento 5, os comportamentos dos ativos variaram de acordo com a situações que afetaram os setores analisados e os impactos internos e externos da economia.

No estudo, não foram utilizadas as variáveis de indicadores fundamentalistas, extraídas dos balanços, como proposto no modelo, devido ao tempo necessário para o processamento da extração e cálculo desses dados. Embora existam plataformas no Brasil que oferecem essas informações diariamente, elas possuem um alto custo de acesso. Em mercados mais maduros, o acesso a essas informações e ferramentas de processamento de dados é mais facilitado. Além

disso, os dados de fóruns de discussões não foram incluídos devido às dificuldades envolvidas em sua coleta e ao tempo limitado disponível para o desenvolvimento deste experimento com essas variáveis.

6.3 Trabalhos Futuros

Algumas das atividades podem ser ampliadas em trabalhos futuros, bem como algumas linhas de investigação não previstas tomaram forma. A seguir são comentados alguns exemplos de possibilidade de trabalhos futuros.

Um deles é a expansão do modelo para contextos internacionais. Um trabalho futuro promissor poderia ser a adaptação do modelo desenvolvido para mercados acionários de outras economias, incluindo mercados mais maduros e outros emergentes. Essa abordagem permitiria avaliar como as especificidades de diferentes mercados influenciam a eficácia da integração de dados numéricos e textuais. Além disso, poderiam ser incorporados dados adicionais, como indicadores específicos de cada mercado e variáveis macroeconômicas locais, para validar a generalização do modelo.

Outro trabalho futuro de interesse trata do desenvolvimento de modelos de PLN para o português de acordo com o vocabulário financeiro. Isso é importante pois um dos desafios encontrados foi a limitação de ferramentas de PLN específicas para o português voltadas ao mercado financeiro. Estudos futuros poderiam focar na criação de modelos treinados exclusivamente com dados financeiros em português, como relatórios contábeis, notícias econômicas e discussões em fóruns. Essa linha de pesquisa poderia incluir o desenvolvimento de ferramentas para capturar nuances linguísticas, como ironia e sarcasmo, comuns em mídias sociais.

Outro ponto relevante seria a incorporação de dados de fóruns de discussão e outras mídias alternativas. Estudos futuros poderiam investigar métodos de coleta e processamento eficientes para esses dados, avaliando seu impacto na previsão de movimentos do mercado. O uso de algoritmos de aprendizado profundo para identificar padrões nesses conteúdos textuais também seria uma extensão natural da pesquisa.

Estudos futuros poderiam investigar o impacto de eventos específicos, como crises econômicas ou mudanças políticas, utilizando o modelo proposto. Isso permitiria avaliar a

capacidade preditiva do modelo em cenários de alta volatilidade e incerteza, além de explorar a interação entre variáveis macroeconômicas e dados de mídias sociais nesses contextos.

Considerando as limitações de acesso a plataformas de dados no Brasil, estudos futuros poderiam focar no desenvolvimento de ferramentas acessíveis para investidores locais. Essas ferramentas poderiam integrar dados numéricos e textuais em tempo real, auxiliando na tomada de decisão de pequenos e médios investidores.

Essas propostas de trabalhos futuros podem contribuir ainda mais para o avanço da integração de dados estruturados e não estruturados, ampliando a compreensão das dinâmicas de mercado e gerando novas oportunidades para pesquisadores e profissionais do setor financeiro.

6.4 Publicações

Durante a realização desta tese, foram desenvolvidos experimentos (conforme apresentados no capítulo 5), os quais resultaram na elaboração de artigos para publicação. A seguir estão listados os artigos já publicados e os que estão em andamento, aguardando parecer das respectivas revistas:

a) Publicação na revista: *International Journal of Professional Business Review*, ISSN 2525-3654, intitulado “Análise do Efeito das Pesquisas da Internet Via “Google Trends” no Comportamento do Mercado Acionário”. DOI:https://doi.org/10.26668/businessreview/2024.v9i9.4868.
b) Publicação na revista: *International Journal of Professional Business Review*, ISSN 2525-3654, intitulado “Análise dos Determinantes da Oscilação dos Preços das Ações da Petrobras no Período de 2006 a 2022". DOI: https://doi.org/10.26668/businessreview/2024.v9i12.5142.
c) Em andamento: revisões necessárias estão sendo realizadas - IEEE *Latin America Transactions*, intitulado “*Brazilian Stock Market Forecast with Heterogeneous Data Integration for a Set of Stocks*”

Além destes pretende-se publicar um novo experimento em andamento, no qual estão sendo utilizadas fontes de dados além das séries históricas de preços de ações, incluindo sentimentos extraídos de notícias, indicadores técnicos, dados macroeconômicos, índices de

tendências do Google Trends e indicadores fundamentalistas, com foco no aprimoramento dos modelos utilizados.

REFERÊNCIAS

ALBAOOTH, B. **The Role of Artificial Intelligence Prediction in Stock Market Investors' Decisions**. 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). **Anais**...IEEE, 4 dez. 2023.

ALMEIDA, R. J. A. **LeIA - Léxico para Inferência Adaptada**. Disponível em: <https://github.com/rafjaa/LeIA>. Acesso em: 10 ago. 2024.

ALZAZAH, F. S.; CHENG, X. 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.

ANACHE, M. DE C. A. **Finanças comportamentais: uma avaliação crítica da moderna teoria de finanças**. Vitória: Dissertação (Mestrado) Universidade Federal do Espírito Santo, 2008.

ANP. **Agência Nacional do Petróleo, Gás Natural e Biocombustíveis.** Disponível em: <https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-estatisticos/dados-estatisticos>. Acesso em: 26 set. 2022.

ARAÚJO, L. V.; FERNANDES, T. J. B. Análise de Sentimentos de Textos do Twitter sobre Mercado de Ações Brasileiro. **Engenharia e Pesquisa Aplicada, Recife**. 6, p. 18–26, maio 2021.

ASSAF NETO, A. **Mercado Financeiro**. 15. ed. Barueri: SP: Atlas, Instituto Assaf, 2021.

B3. **B3 divulga estudo sobre os 2 milhões de investidores que entraram na bolsa entre 2019 e 2020**. Disponível em: <https://www.b3.com.br/pt_br/notícias/investidores.htm>. Acesso em: 5 mar. 2022.

B3. **B3 atinge 5 milhões de contas de investidores em renda variável em janeiro**. Disponível em: <https://www.b3.com.br/pt_br/noticias/5-milhoes-de-contas-de-investidores.htm>. Acesso em: 5 mar. 2022a.

B3. Índice Bovespa (Ibovespa B3). 2022b. Disponível em: <https://www.b3.com.br>. Acesso em: 5 mar. 2022b.

B3. **Perfil pessoas físicas**. Disponível em: <https://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/perfil-pessoas-fisicas/perfil-pessoa-fisica/>. Acesso em: 3 jul. 2024a.

B3. **A Bolsa do Brasil - B3**. Disponível em: <https://www.b3.com.br/pt_br/>. Acesso em: 17 jun. 2024b.

B3. **Índice Bovespa (Ibovespa B3)**. Disponível em: <https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/indice-ibovespa-ibovespa-composicao-da-carteira.htm>. Acesso em: 9 jan. 2024c.

BACEN. **Banco Central. Taxa Selic**. Disponível em: <https://www.bcb.gov.br/controleinflacao/taxaselic>. Acesso em: 5 mar. 2022.

BACEN. **Banco Central. Taxa Selic**. Disponível em: <https://www.bcb.gov.br/controleinflacao/taxaselic>. Acesso em: 6 jul. 2024.

BAKER, M.; WURGLER, J. Investor Sentiment in the Stock Market. **Journal of Economic Perspectives**, v. 21, n. 2, p. 129–151, 1 abr. 2007.

BARBER, B. M.; ODEAN, T. All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. **Review of Financial Studies**, v. 21, n. 2, p. 785–818, abr. 2008.

BARBERIS, N.; SHLEIFER, A.; VISHNY, R. A model of investor sentiment. **Journal of Financial Economics**, v. 49, n. 3, p. 307–343, set. 1998.

BAZERMAN, M. H.; MOORE, D. A. **Judgment in managerial decision making**. John Wiley & Sons, 2012.

BERTÃO, N. **Rota do dinheiro em 2020: Brasileiro sai de fundos de renda fixa, entra na bolsa e compra novos ativos para a carteira**. Disponível em: <https://valorinveste.globo.com/objetivo/hora-de-investir/noticia/2021/01/26/rota-do-dinheiro-em-2020-brasileiro-sai-de-fundos-de-renda-fixa-entra-na-bolsa-e-compra-novos-ativos-pra-carteira.ghtml>. Acesso em: 6 fev. 2021.

BISHOP, C. M. **Pattern recognition and machine learning**. Springer Science+ Business Media, LLC, 2006.

BOU-HAMAD, I.; JAMALI, I. Forecasting financial time-series using data mining models: A simulation study. **Research in International Business and Finance**, v. 51, p. 101072, Jan. 2020.

BREIMAN, L. Florestas Aleatórias. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BRIGGS, J. **Sentiment Analysis for Stock Price Prediction in Python**. Disponível em: <https://towardsdatascience.com/sentiment-analysis-for-stock-price-prediction-in-python-bed40c65d178>. Acesso em: 10 jun. 2024.

CARAÇA, L. V. N. **Relação entre o desempenho do mercado brasileiro de ações e o Google Trends**. Osasco: Universidade Federal de São Paulo Escola Paulista de Política, Economia e Negócios, 2019.

CAROSIA, A. Using Machine Learning to Prevent Losses in the Brazilian Stock Market During the Covid-19 Pandemic. **IEEE Latin America Transactions**, v. 21, n. 8, p. 867–873, ago. 2023.

CASILDA, B. R.; LAMOTHE, F. Z. P.; MONJAS, B. M. **La banca y los mercados financieros**. Madrid: Alianza, 1997.

CASTRO, M. A.; REYES, R. E.; LANDAZÁBAL, S. N. J. Mexican Stock Exchange Performance after the Crisis of 2008: Application of Data Mining. **Dimensión Empresarial**, v. 18, n. (1), 13 jan. 2020.

CHEN, Y. et al. Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM. **MM '18: Proceedings of the 26th ACM International Conference on Multimedia**, p. 117–125, 2018.

COELHO, F. F. **Machine learning e Análise Técnica como Ferramentas para Construção de Portfólios de Renda Variável no Mercado Brasileiro**. Rio de Janeiro: Fundação Getulio Vargas, 2020.

CORREIA, L. R. **Google Trends e o Desempenho do Mercado Brasileiro de Ações**. Curitiba - PR: Universidade Federal do Paraná, 2021.

DAL PUPPO, G. **Análise de eventos socioeconômicos e políticos que impactaram o preço das ações da Petrobras entre 2016 e 2018**. Pato Branco: Universidade Tecnológica Federal do Paraná, 2020.

DATAREPORTAL. **Digital 2022: Brasil DataReportal – Global Digital Insights**. Disponível em: <https://datareportal.com/reports/digital-2022-brazil?rq=braz>. Acesso em: 19 fev. 2022.

DE PRADO, M. L. **Advances in financial machine learning**. John Wiley & Sons, 2018.

DEBASTIANI, A. C.; RUSSO, A. F. **Avaliando empresas, investindo em ações: a aplicação prática da análise fundamentalista na avaliação de empresas**. São Paulo: Novatec Editora, 2008.

DING, X. et al. **Using Structured Events to Predict Stock Price Movement: An Empirical Investigation**. Proceedings of the 2014 Conference on Empirical Methods in

Natural Language Processing (EMNLP). **Anais**...Stroudsburg, PA, USA: Association for Computational Linguistics, 2014. Disponível em: <http://aclweb.org/anthology/D14-1148>. Acesso em: 3 fev. 2022

DING, X. et al. Deep Learning for Event-Driven Stock Prediction. **24th. International Joint Conferences on Artificial Intelligence Organization - IJCAI**, p. 2327–2333, 2015.

DU, S.; HAO, D.; LI, X. **Research on stock forecasting based on random forest**. 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA). **Anais**...IEEE, 28 out. 2022.

ECONOMATICA. **Uma poderosa plataforma de investimento para estar no topo**. Disponível em: <https://economatica.com/>. Acesso em: 5 mar. 2022.

FAMA, E. F. Random Walks in Stock Market Prices. **Financial Analysts Journal**, p. 55–59, 1965.

FAMA, E. F. Efficient Capital Markets: A Review of Theory and Empirical Work. **The Journal of Finance**, v. 25, n. 2, p. 383, maio 1970.

FAMA, E. F. Efficient Capital Markets: II. **The Journal of Finance**, v. 46, n. 5, p. 1575–1617, 30 dez. 1991.

FERREIRA, V. R. **Psicologia Econômica: Estudo do comportamento econômico e da tomada de decisão**. 1. ed. Rio de Janeiro: Elsevier, 2008.

FIGUEIREDO FILHO, D. B.; SILVA JÚNIOR, J. A. DA. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). **Revista Política Hoje**, v. 18, n. 1, p. 115–146, 2009.

FUNDAMENTUS. Disponível em: <https://www.fundamentus.com.br/>. Acesso em: 5 mar. 2022.

GOOGLE TRENDS. **Google Trends**. Disponível em: <https://trends.google.com>. Acesso em: 13 jul. 2024.

GOSWAMI, S. et al. **Time Series Analysis Using Stacked LSTM Model for Indian Stock Market**. 2022 IEEE IAS Global Conference on Emerging Technologies (GlobConET). **Anais**...IEEE, 20 maio 2022.

GUJARATI, D. **Econometria: princípios, teoria e aplicações práticas**. São Paulo: Saraiva, 2019.

GUJARATI, D. N.; PORTER, D. C. **Econometria Básica**. 5. ed. Rio de Janeiro: Mc Graw-Hill, 2011.

GUPTA, V. A Survey of Natural Language Processing Techniques. **International Journal of Computer Science & Engineering Technology (IJCSET)**, p. 14–16, 2014.

GUZELLA, M.; CASTRO, F. H.; SANTANA, V. DE F. Efeito da atenção do investidor na eficiência do mercado brasileiro de ações. **Revista Contabilidade & Finanças**, v. 34, n. 93, 2023.

HILL, C. R.; JUDGE, G. G.; GRIFFITS, W. E. **Econometria**. São Paulo: Saraiva, 2010.

HONNIBAL, M.; MONTANI, I. **spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing**. Disponível em: <https://spacy.io/>. Acesso em: 10 ago. 2024.

HUTTO, C.; GILBERT, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. **Proceedings of the International AAAI Conference on Web and social media**, v. 8, n. 1, p. 216–225, 16 maio 2014.

IBGE. **PNAD Contínua TIC 2019: internet chega a 82,7% dos domicílios do país**. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-

agencia-de-noticias/releases/30521-pnad-continua-tic-2019-internet-chega-a-82-7-dos-domicilios-do-pais>. Acesso em: 13 jul. 2024.

INDURKHYA, N.; DAMERAU, J. F. **Handbook of Natural Language Processing**. 2. ed. Nova Iorque: Chapman and Hall/CRC, 2010.

INVESTING. **Série histórica do petróleo Brent**. Disponível em: <https://br.investing.com/commodities/brent-oil-historical-data>. Acesso em: 26 set. 2022.

IPEADATA. **Base de dados macroeconômicos, financeiros e regionais do Brasil**. Disponível em: <http://www.ipeadata.gov.br/Default.aspx>. Acesso em: 26 set. 2022.

JANSEN, S. **Machine Learning for Algorithmic Trading: Predictive models to extract signals from market and alternative data for systematic trading strategies with Python**. Packt Publishing Ltd, 2020.

JIANG, W. Applications of deep learning in stock market prediction: Recent progress. **Expert Systems with Applications**, v. 184, p. 115537, dez. 2021.

JIN, F. et al. Tracking Multiple social media for Stock Market Event Prediction. Em: PERNER PETRA (Ed.). **Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2017. Lecture Notes in Computer Science**. Cham: Springer, 2017. v. 10357p. 16–30.

JOSHI, K.; BHARATHI, H. N.; RAO, J. Stock Trend Prediction Using News Sentiment. **International Journal of Computer Science & Information Technology (IJCSIT)**, v. 8, n. 3, 2016.

JOSHI, K.; H. N, B.; RAO, J. Stock Trend Prediction Using News Sentiment Analysis. **International Journal of Computer Science and Information Technology**, v. 8, n. 3, p. 67–76, 30 jun. 2016.

JUWONO, Y. et al. **Comparative Study on Stock Price Forecasting Using Deep Learning Method Based on Combination Dataset**. 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS). **Anais**...IEEE, 21 fev. 2024.

KACZOROWSKI, B. et al. Artificial Intelligence and The Multivariate Approach In Predictive Analysis Of The Small Cap Index Of The Brazilian Stock Exchange. **IEEE Latin America Transactions**, v. 19, n. 11, p. 1924–1932, nov. 2021.

KAHNEMAN, D. **Rápido e devagar: duas formas de pensar**. Rio de Janeiro: Objetiva, 2012.

KAHNEMAN, D.; TVERSKY, A. Prospect Theory: An Analysis of Decision under Risk. **Econometrica**, v. 47, n. 2, p. 263, mar. 1979.

KAKDE, A. K.; DALE, M. P. **Comparative Analysis Of Deep Learning Approaches Used For Stock Price Prediction**. 2024 2nd World Conference on Communication & Computing (WCONF). **Anais**...IEEE, 12 jul. 2024.

KIM, K. G. Book Review: Deep Learning. **Healthcare Informatics Research**, v. 22, n. 4, p. 351, 2016.

KOBORI, J. **Análise Fundamentalista Análise Fundamentalista - Performance Superior e Consistente no Mercado de Ações. Coleção Expo Money**. Rio de Janeiro: Campus, Elsevier, 2011.

KORABLYOV, M. et al. **Hybrid stock analysis model for financial market forecasting**. 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT). **Anais**...IEEE, 19 out. 2023.

KORBES, P. J.; COSTA JR., N. C. A. **Bolsa de Valores e Estratégias de Investimento**. Sinop: Unemat editora, 2003.

KUTCHUKIAN, E. **O efeito manada nos fundos de investimento no Brasil: um teste em finanças comportamentais**. São Paulo: Fundação Getúlio Vargas, 2010.

LATOEIRO, P. J. C. **Pesquisar para decidir: O Google como barómetro da atenção do investidor**. Lisboa: Instituto Universitário de Lisboa, 2012.

LEMOS, F. **Análise técnica dos mercados financeiros: um guia completo e definitivo dos métodos de negociação de ativos**. 3. ed. São Paulo: SaraivaUni, 2022.

LIU, H.; LONG, Z. An improved deep learning model for predicting stock market price time series. **Digital Signal Processing**, v. 102, p. 102741, jul. 2020.

LOBO, L. C. Inteligência artificial, o Futuro da Medicina e a Educação Médica. **Revista Brasileira de Educação Médica**, v. 42, n. 3, p. 3–8, set. 2018.

LOPES, A. DE O. **Determinantes do preço das ações da empresa Petrobras (PETR4), entre 2009 a 2018.** Palmeira das Missões: Universidade Federal de Santa Maria, 2019.

MALKIEL, B. G.; FAMA, E. F. Efficient Capital Markets: A Review of Theory and Empirical Work. **The Journal of Finance**, v. 25, n. 2, p. 383–417, 30 maio 1970.

MANNAM, P. K. **Stock Exchange Market Performance and Data Analysis Techniques by using Machine Learning Approach**. 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT). **Anais**...IEEE, 20 out. 2023.

MAQBOOL, J. et al. Stock Prediction by Integrating Sentiment Scores of Financial News and MLP-Regressor: A Machine Learning Approach. **Procedia Computer Science**, v. 218, p. 1067–1078, 2023.

MARTINS, C. **Os supersinais da análise técnica: guia para investimentos lucrativos na bolsa**. Rio de Janeiro: Elsevier, 2010.

MILANEZ, D. Y. **Finanças comportamentais no Brasil**. São Paulo: Dissertação (Mestrado) – Universidade de São Paulo, 2003.

MOURÃO, A. I. M. **O Google como medida de sentimento nos mercados financeiros**. [s.l.] Dissertação de mestrado, Iscte - Instituto Universitário de Lisboa, 2018.

NABIPOUR, M. et al. Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis. **IEEE Access**, v. 8, p. 150199–150212, 2020.

NELSON, D. M. Q. **Uso de redes neurais recorrentes para previsão de séries temporais financeiras**. Dissertação (mestrado) —Belo Horizonte: Universidade Federal de Minas Gerais, 2017.

NTI, I. K.; ADEKOYA, A. F.; WEYORI, B. A. A systematic review of fundamental and technical analysis of stock market predictions. **Artificial Intelligence Review**, v. 53, n. 4, p. 3007–3057, 1 abr. 2020a.

NTI, I. K.; ADEKOYA, A. F.; WEYORI, B. A. Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence from Ghana. **Applied Computer Systems**, v. 25, n. 1, p. 33–42, 1 maio 2020b.

NTI, I. K.; ADEKOYA, A. F.; WEYORI, B. A. Efficient Stock-Market Prediction Using Ensemble Support Vector Machine. **Open Computer Science**, v. 10, n. 1, p. 153–163, 4 jul. 2020c.

NTI, I. K.; ADEKOYA, A. F.; WEYORI, B. A. A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction. **Journal of Big Data**, v. 8, n. 1, p. 17, 9 dez. 2021.

NUNES, C.; RICKROT, J.; WATANABE, M. Uma pesquisa sobre a variação do preço das ações da Petrobras S.A. **Revista Científica Hermes**, v. 22, p. 606–622, 2018.

PAIVA, C. C. DE; PAIVA, S. C. F. DE. **No Brasil, impacto econômico da pandemia será forte e duradouro**. Disponível em: <https://jornal.unesp.br/2021/07/02/no-brasil-impacto-economico-da-pandemia-sera-forte-e-duradouro/>. Acesso em: 23 nov. 2022.

PAIVA, F. S. **O Processo de decisão sob a perspectiva da economia comportamental e da neurociência**. Lisboa: Dissertação de Mestrado. Instituto Superior de Contabilidade e Administração, 2013.

PAIVA, N. F. S. R. **Pesquisas no Google e a atividade no mercado de ações - Evidência de Portugal**. Coimbra, Portugal: Faculdade de Economia da Universidade de Coimbra, 2020.

PAK, A.; PAROUBEK, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. **In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) , Valletta, Malta. Associação Europeia de Recursos Linguísticos (ELRA).**, 2010.

PARABONI, A. L. et al. Impacto de variáveis macroeconômicas e corrupção na Petrobrás no Sentimento de Mercado. **Revista ESPACIOS**, v. 37 (Nº 31), 2016.

PARANHOS, R. et al. Desvendando os Mistérios do Coeficiente de Correlação de Pearson: o Retorno. **Leviathan (São Paulo)**, n. 8, p. 66, 13 ago. 2014.

PENG, Y.; JIANG, H. **Leverage Financial News to Predict Stock Price Movements Using Word Embeddings and Deep Neural Networks**. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Stroudsburg, PA, USA: Association for Computational Linguistics, 2016. Disponível em: <http://aclweb.org/anthology/N16-1041>. Acesso em: 3 fev. 2022

PEREIRA, M. M. DE; ROSA, T. G. DA; BENDER FILHO, R. Influência do Google Trends em Ações Listadas na Bolsa de Valores Brasileira: Evidências a Partir da Modelagem

PVAR. **REAd. Revista Eletrônica de Administração (Porto Alegre)**, v. 26, n. 3, p. 796–818, dez. 2020.

PERERA, R.; NAND, P. Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature. **Computing and Informatics**, v. 36, n. 1, p. 1–32, 2017.

PETERSON, C.; RÖGNVALDSSON, T. An introduction to artificial neural networks. Em: **CERN Summer School of Computing, CERN Yellow Report 92-02, pp.113-170, 1992.** [s.l.] CERN Yellow Report, 1991. p. 93–23.

PINHEIRO, J. L. **Capital markets**. 9. ed. São Paulo: Atlas, 2019.

POMPIAN, M. M. **Behavioral Finance and Wealth Management: How to Build Investment Strategies That Account for Investor Biases**. 2. ed. [s.l.] Wiley, 2012.

PÓVOA, A. **Valuation: como precificar ações**. Rio de Janeiro: Elsevier, 2012.

PREIS, T.; MOAT, H. S.; STANLEY, H. E. Quantifying Trading Behavior in Financial Markets Using Google Trends. **Scientific Reports**, v. 3, n. 1, p. 1684, 25 abr. 2013.

RAY, R.; KHANDELWAL, P.; BARANIDHARAN, B. **A Survey on Stock Market Prediction using Artificial Intelligence Techniques**. 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT). **Anais**...IEEE, dez. 2018.

REZENDE, O. S. **Sistemas Inteligentes: Fundamentos e Aplicações**. São Paulo: Manole, 2005.

RIGHI, M. B.; SCHLENDER, S. G.; CERETTA, P. S. Análise dos impactos esperados e não-esperados da taxa de juros, câmbio e inflação no mercado brasileiro. **Revista de Administração da UFSM**, v. 5, n. 3, p. 539–548, 16 nov. 2012.

ROGERS, P.; FAVATO, V.; SECURATO, J. R. **Efeito educação financeira no processo de tomada de decisões em investimentos: um estudo a luz das finanças comportamentais**. II Congresso ANPCONT-Associação Nacional dos Programas de Pós-Graduação em Ciências Contábeis. **Anais**...Salvador/BA: 2008.

ROSENTHAL, S.; FARRA, N.; NAKOV, P. **SemEval-2017 Task 4: Sentiment Analysis in Twitter**. 2019. Disponível em: <https://trends24.in/>.

RUKE, A. et al. **Predictive Analysis of Stock Market Trends: A Machine Learning Approach**. 2024 4th International Conference on Data Engineering and Communication Systems (ICDECS). **Anais**...IEEE, 22 mar. 2024.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial. tradução Regina Célia Simille**. Rio de Janeiro: Elsevier, 2013.

SALVI, A. DE A.; SOUZA, W. P. DE; BRANCO NETO, W. C. Hybrid Neural Networks Applied to Brazilian Stock Market. **Revista de Informática Teórica e Aplicada**, v. 27, n. 2, p. 42–65, 27 abr. 2020.

SAMSON, A. Introdução à Economia Comportamental e Experimental. Em: ECONOMIACOMPORTAMENTAL.ORG (Ed.). **Guia de Economia Comportamental e Experimental**. São Paulo: 2015.

SANTOS, G. C. **Algoritmos de Machine Learning para Previsão de Ações da B3**. Uberlândia: Universidade Federal de Uberlândia, 2020

SARTORIS, A. **Estatística e Introdução à econometria**. São Paulo: Saraiva, 2003.

SBICCA, A. Heurísticas no estudo das decisões econômicas: contribuições de Herbert Simon, Daniel Kahneman e Amos Tversky. **Estudos Econômicos (São Paulo)**, v. 44, n. 3, p. 579–603, set. 2014.

SCHUMAKER, R. P.; CHEN, H. Textual analysis of stock market prediction using breaking financial news. **ACM Transactions on Information Systems**, v. 27, n. 2, p. 1–19, 9 fev. 2009.

SHI, L. et al. DeepClue: Visual interpretation of text-based deep stock prediction. **IEEE Transactions on Knowledge and Data Engineering**, v. 31, n. 6, p. 1094–1108, 1 jun. 2019.

SILVA, P. DA J. **Análise financeira das empresas**. 13. ed. São Paulo: Atlas, 2016.

SINATORA, J. R. P. **Mercado de Capitais**. Londrina: Editora e Distribuidora Educacional S.A., 2016.

SISMANOGLU, G. et al. **Deep Learning Based Forecasting in Stock Market with Big Data Analytics**. 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). **Anais**...IEEE, abr. 2019.

STATCOUNTER GLOBALSTATS. **Participação no mercado de mecanismos de busca no Brasil - junho de 2024**. Disponível em: <https://gs.statcounter.com/search-engine-market-share/all/brazil>. Acesso em: 13 jul. 2024.

STATISTA. **Redes sociais mais populares em todo o mundo em janeiro de 2022, classificadas pelo número de usuários ativos mensais**. Disponível em: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. Acesso em: 27 fev. 2022.

STATMAN, M. Behavioral Finance versus Standard Finance. **AIMR Conference Proceedings**, v. 1995, n. 7, p. 14–22, dez. 1995.

TEIXEIRA ZAVADZKI DE PAULI, S.; KLEINA, M.; BONAT, W. H. **Comparing Artificial Neural Network Architectures for Brazilian Stock Market Prediction**. **Annals of Data Science** Springer Science and Business Media Deutschland GmbH, 1 dez. 2020.

TETLOCK, P. C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. **SSRN Electronic Journal**, 2005.

THALER, R. H. Behavioral Economics: Past, Present, and Future. **American Economic Review**, v. 106, n. 7, p. 1577–1600, 1 jul. 2016.

THALER, R. H.; SUNSTEIN, C. R. **Nudge: o empurrão para a escolha certo**. Rio de janeiro: Elsevier, 2009.

TORGA, E. M. M. F. et al. The Effects of Car Wash Operation on The Brazilian Capital Market: The Petrobras Case. **RAM. Revista de Administração Mackenzie**, v. 22, n. 2, 2021.

TVERSKY, A.; KAHNEMAN, D. Judgment under Uncertainty: Heuristics and Biases. **Science**, v. 185, n. 4157, p. 1124–1131, 27 set. 1974.

TWERSKY, I. **A Maimonides Reader**. Behrman House Publishing, 1972.

VARGAS, G. et al. B3 Stock Price Prediction Using LSTM Neural Networks and Sentiment Analysis. **IEEE Latin America Transactions**, v. 20, n. 7, p. 1067–1074, jul. 2022.

VASCO, L. P. **Um Estudo de Redes Neurais Recorrentes no Contexto de Previsões no Mercado Financeiro**. Universidade Federal de São Carlos – UFSCar, 2020.

VÁZQUEZ, F. **Deep Learning made easy with Deep Cognition**. Disponível em: <https://becominghuman.ai/deep-learning-made-easy-with-deep-cognition-403fbe445351>. Acesso em: 10 mar. 2022.

VIANA, L. F. **Google Trends e o Comportamento do Mercado Acionário Brasileiro**. Brasília - DF: Universidade de Brasília, 2017.

VIOLA, R. **Influência do Google Search na previsão do retorno de ações do mercado brasileiro**. São Paulo - SP: Insper, 2022.

WANG, W.-J. et al. Stock market index prediction based on reservoir computing models. **Expert Systems with Applications**, v. 178, p. 115022, set. 2021.

WERNER, L.; BISOGNIN, C.; ARAUJO, C. W. Análise de técnicas de previsão: um estudo de caso para o volume de ações da Petrobras. **Brazilian Journal of Development**, v. 6, n. 1, p. 1103–1115, 2020.

WOOLDRIDGE, J. M. **Introdução à econometria: uma abordagem moderna**. 7. ed. São Paulo: Tradução da 7ª Edição Norte-Americana: Cengage Learning, 2023.

YADAV, A.; JHA, C. K.; SHARAN, A. Optimizing LSTM for time series prediction in Indian stock market. **Procedia Computer Science**, v. 167, p. 2091–2100, 2020.

YAHOO FINANÇAS. **Dados históricos**. Disponível em: <https://br.financas.yahoo.com/quote/PETR4.SA/history?p=PETR4.SA>. Acesso em: 5 mar. 2022.

YOSHINAGA, C. E.; RAMALHO, T. B. Finanças comportamentais no Brasil: uma aplicação da teoria da perspectiva em potenciais investidores. **Revista Brasileira de Gestão De Negócios**, v. 16, p. 594–615, 2014.

ZENHA, L. Redes sociais online: o que são as redes sociais e como se organizam? **Caderno de Educação**, v. 1, n. 49, p. 19–42, 2018.

ZHANG, Q.; YANG, L.; ZHOU, F. Attention enhanced long short-term memory network with multi-source heterogeneous information fusion: An application to BGI Genomics. **Information Sciences**, v. 553, p. 305–330, 1 abr. 2021.

ZHANG, X. et al. Stock Market Prediction via Multi-Source Multiple Instance Learning. **IEEE Access**, v. 6, p. 50720–50728, 2018a.

ZHANG, X. et al. Improving stock market prediction via heterogeneous information fusion. **Knowledge-Based Systems**, v. 143, p. 236–247, 1 mar. 2018b.

ZHANG, X. et al. Enhancing stock market prediction with extended coupled hidden Markov model over multi-sourced data. **Knowledge and Information Systems**, v. 61, n. 2, p. 1071–1090, 17 nov. 2019.