UNIVERSIDADE DO VALE DO RIO DOS SINOS
UNIDADE ACADÊMICA DE GRADUAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

CARLOS LEANDRO SILVA MACHADO

PROJEÇÃO DE PREÇOS NO MERCADO DE AÇÕES BRASILEIRO CRUZANDO DADOS TEXTUAIS E HISTÓRICOS DE VALORES COM MODELOS DE APRENDIZADO PROFUNDO

São Leopoldo
2022

CARLOS LEANDRO SILVA MACHADO

PROJEÇÃO DE PREÇOS NO MERCADO DE AÇÕES BRASILEIRO CRUZANDO DADOS TEXTUAIS E HISTÓRICOS DE VALORES COM MODELOS DE APRENDIZADO PROFUNDO

Artigo apresentado como requisito parcial para obtenção do título de Bacharel em Ciência da Computação, pelo Curso de Ciência da Computação da Universidade do Vale do Rio dos Sinos (UNISINOS)

Orientador(a): Prof. Dr. Sandro José Rigo

São Leopoldo
2022

PROJEÇÃO DE PREÇOS NO MERCADO DE AÇÕES BRASILEIRO CRUZANDO DADOS TEXTUAIS E HISTÓRICOS DE VALORES COM MODELOS DE APRENDIZADO PROFUNDO

Carlos Leandro Silva Machado[1]
Sandro José Rigo[2]
Michele Rosa[3]

Resumo: O grande crescimento do mercado de ações brasileiro nos últimos anos e o aumento na quantidade de investidores em ações mais arriscadas gerou a demanda de ferramentas de apoio à predições. Entretanto os métodos tradicionais de apoio à predição automática utilizam exclusivamente as bases de valores numéricos sobre os preços de ações. Com o objetivo de avaliar novas formas de predição e análise do mercado financeiro com melhores taxas de acerto das predições, desenvolvemos neste trabalho processos para unir dados do histórico de valores de ações com dados textuais extraídos de mensagens do *Twitter*, mensagens provindas de mídias focadas no mercado financeiro e notícias relacionadas coletadas do *Google News* sobre os ativos das empresas. No processo foram utilizados técnicas de aprendizagem de máquina focada em redes neurais artificiais combinadas com técnicas de processamento de linguagem natural, mais especificamente ligadas a área de análise de sentimentos, efetuando assim a união dos dados para o treinamento de três modelos de redes neurais, sendo um modelo estruturado como uma rede neural profunda composta por camadas do tipo *Dense*, um modelo de rede neural estruturado em *LSTM* com retroalimentação e outro modelo *LSTM* sem retroalimentação. Ao final pode-se constatar que o modelo utilizando apenas rede neural profunda obteve melhores resultados nas comparações utilizando precisão direcional como métrica para cada ativo testado. Também verificou-se que os experimentos contendo dados numéricos e dados textuais apresentam melhor precisão do que os experimentos contendo apenas dados numéricos.

Palavras-chave: Rede Neural Artificial. Análise de Sentimentos. Mercado Financeiro. Redes Sociais. Processamento de Linguagem Natura.

1 INTRODUÇÃO

Com a ascensão do mercado de ações nos últimos anos e mesmo havendo uma elevação dos juros houve um aumento na quantidade de investidores em ações mais arriscadas (MENDES, 2022) nos últimos anos, e para esse cenário tornou-se necessário criar novas formas de auxiliar antigos e novos investidores, principalmente pelo fato de muitos novos investidores que saíram de investimentos mais conservadores e seguros migraram para o mercado de ações com pouco ou nenhum conhecimento referente ao assunto.

Dessa forma, surgiram muitos projetos para predição de dados utilizando apenas informações extraídas de históricos de negociações de bases de dados diversas, como por exemplo

[1]Graduando em de Ciência da Computação pela Unisinos. Email: carlossmachado@edu.unisinos.br
[2]Orientador, professor da Unisinos, Doutor em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (2008). Email: rigo@unisinos.br
[3]Co-Orientadora, professor da UNEMAT, Mestre em Economia (2008). Email: michele.rosa@unemat.br

Yahoo! Finance (YAHOOFINANCE, 1997), Google Finance (GOOGLEFINANCE, 2006), Motley Fool (MOTLEYFOOL, 1993), entre outros. Este tipo de base de dados é usado em trabalhos que abordam o tema de predição de séries históricas do mercado de ações brasileiro com usos de aprendizado de máquina (STEFFENON, 2021). Nos métodos tradicionais de predição utilizando dados numéricos, normalmente é feito a extração em seguida os dados são normalizados e combinados com cálculos de média móvel para enfim serem utilizados nos modelos de redes neurais artificiais (IBM, 2021a). Além deste tipo de abordagem, trabalhos como o de Isaac Nti (NTI; ADEKOYA; WEYORI, 2021) apresentam um foco na fusão de dados obtidos de múltiplas fontes para predição de ações de mercado utilizando redes neurais artificiais com uma técnica que combina redes convolucionais com LSTM.

O objetivo desse trabalho é atuar em uma lacuna de pesquisa com foco na avaliação da integração de diferentes fontes de dados, tais como notícias coletadas de portais, mensagens textuais extraídas de redes sociais e séries históricas do mercado de ações brasileiras. Consideramos assim que essas integrações podem apresentar um ganho na melhoria das predições.

Para testar novas formas de predição do mercado de ações foram efetuadas análises com a unificação de dados textuais extraídos de mensagens do Twitter diretamente de mídias focadas no mercado financeiro como Investing, Money Times, InfoMoney e Valor Econômico, além de notícias relacionadas, coletadas do Google News sobre o ativo e a empresa em questão.

Este trabalho foi organizado da forma descrita a seguir. Na seção 2 detalhamos as fundamentações teóricas, demonstrando os recursos e ferramentas utilizadas e conhecimentos adquiridos sobre o mercado de ações, Twitter, Google News, redes neurais artificiais, métricas, bibliotecas e ferramentas utilizadas no projeto. Seguindo, na seção 3 foram destacados trabalhos relacionados que serviram de base para este artigo. Na seção 4 detalhamos a metodologia do trabalho demonstrando o funcionamento da análise de sentimentos utilizada para os dados obtidos nas redes sociais, o cruzamento de informações e as arquiteturas de redes neurais artificiais utilizadas. A seção 5 apresenta os detalhes dos experimentos e sua interpretação. E por fim, na seção 6 são destacadas as conclusões sobre o projeto e possibilidades para o futuro.

2 REFERENCIAL TEÓRICO

Nessa sessão são comentados detalhes sobre o mercado de ações, sobre fontes alternativas de informações textuais sobre movimentos de ações, bem como aspectos das técnicas de Inteligência Artificial utilizadas.

2.1 Mercado de ações brasileiro

Como base de dados para a verificação por valores, vamos utilizar dados obtidos das ações PETR4 (Petróleo Brasileiro SA), VALE3 (Vale S.A.), BBDC4 (Banco Bradesco SA) e ITUB4 (Itaú Unibanco). Como referência nas ações brasileiras temos a Brasil Bolsa Balcão (B3 S.A.) onde apenas ela atua como bolsa de valores, mercadorias e futuros no Brasil, a qual surgiu em 2017, quando a Bolsa de Valores, Mercadorias e Futuros de São Paulo (BM&F-BOVESPA) se uniu com a Central de custódia e de liquidação Financeira de Títulos(CETIP). Iremos utilizar neste trabalho dados obtidos de negociações do mercado de ações e dados textuais coletados de mensagens no Twitter para cruzar informações e obter um modelo adaptativo utilizando redes neurais, uma das bases para a pesquisa foi a tese defendida pela Mestre Michele Rosa (ROSA, 2022) onde foi avaliada a possibilidade de integração de dados estruturados e não estruturados com o objetivo de identificar movimentações e preços do mercado brasileiro de ações.

2.2 Histórico de ações

Vamos utilizar nesse trabalho dados de históricos de ações brasileiras que foram extraídos diretamente do Yahoo! Finance a partir de uma data de início e uma data final das cotações. Um acionista pode ser definido como um indivíduo que adquiri ações de uma determinada empresa, tornando-se um acionista, ele passa a ter direito na participação dos lucros quando a empresa se dispõe para a distribuição. Uma ação é caracterizada como a menor parte do capital social de uma empresa ou sociedade anônima.

2.3 Fontes de informação textual sobre movimentos no mercado de ações

O Google News é uma ferramenta disponibilizada pelo Google que efetua uma compilação de notícias agregadas em um fluxo contínuo de artigos dispostos a partir de editores e revistas.

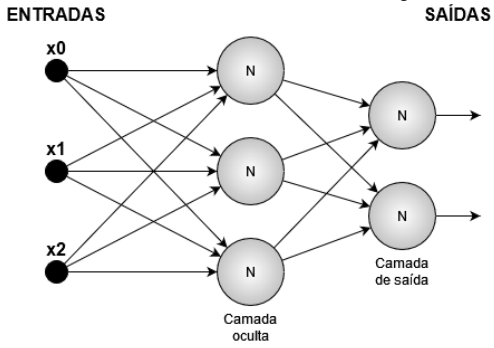
O Twitter é uma rede social para comunicação em tempo real criada em 2006 por Jack Dorsey, Evan Williams, Biz Stone e Noah Glass. Os tweets podem ser definidos como mensagens publicadas na plataforma virtual Twitter, sendo que essas mensagens podem conter imagens, vídeos ou dados textuais com um limite de 280 caracteres. Em alguns tipos de tweets, podem ser indicadas as menções destacando o nome de um usuário da plataforma, o que é caracterizado pelo símbolo "@"antes no nome. Os tweets podem conter respostas de outros usuários.

2.4 Redes neurais artificiais

Podemos dizer nos tópicos de inteligência artificial (ORACLE, 2020) que um neurônio artificial de uma visão computacional realiza o processamento para gerar uma saída tendo uma ou várias entradas. O neurônio pode ser considerado uma estrutura lógica e matemática com o objetivo de simular a estrutura, ações e funções de um neurônio biológico.

As redes neurais alimentadas com múltiplas camadas são um tipo de rede neural mais comumente utilizado, ela possui retroalimentação, onde efetua a ligação da camada de entrada com camadas ocultas e alimentando as mesmas camadas a partir de suas saídas, a fim de melhorar os resultados. A figura 1 ilustra este conceito. Esse tipo de rede é muito útil para grandes volumes de dados.

Figura 1 – Exemplo de rede neural com alimentação de múltiplas camadas



Fonte: Elaborado pelo autor.

A arquitetura LSTM é baseada na rede neural recorrente (RNN) (IBM, 2021b) e é usada principalmente para processamento, classificação e predição de séries temporais. A LSTM é estruturada em cadeia, contendo quatro redes neurais e células que funcionam como blocos de memória. Os dados são retidos pelas células e o gerenciamento de memória é efetuado por portões, sendo eles o Forget Gate que faz a remoção das informações que já não são mais úteis, o Input Gate que efetua a entrada das informações utilizáveis para a célula e o Output Gate, que possui como objetivo a extração das informações uteis do estado atual da célula e retornar para a saída.

A arquitetura DNN (JOHNSON, 2020) pode ser definida como uma rede neural artificial composta por múltiplas camadas entre as camadas de entrada e camadas de saída gerando níveis de complexidade. Devido a grande complexidade computacional, muitos dispositivos com recursos limitados como *smartphones* e *IoTs*[4] foram limitados na utilização de redes neurais profundas. Por conta disso são efetuadas buscas por melhorias na eficiência, mantendo a precisão das DNNs.

[4]Internet das Coisas

2.5 Métricas

Aqui veremos algumas das métricas utilizadas durante os testes dos modelos.

O erro médio quadrado (MSE) É geralmente utilizado para a verificação da precisão dos modelos dando um maior peso aos erros com mais proporção, pois ao ser calculado, cada erro é elevado individualmente ao quadrado e em seguida é efetuado um cálculo da média dos erros quadráticos.

$$MSE = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{1}$$

A raiz quadrada do erro médio (RMSE) é frequentemente utilizada, pois possui maior sensibilidade aos erros maiores devido ao processo quadrático efetuado. Ela é calculada fazendo a raiz quadrada do MSE.

$$RMSE = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

O MAE ou erro médio absoluto é uma medida utilizada em séries temporais para medir a distância do valor verdadeiro. Ela é calculada com base na média dos erros absolutos utilizando o módulo de cada erro.

$$MAE = \sum_{i=1}^{n} |x_i - y_i| \tag{3}$$

O Erro Percentual Absoluto também chamado de MAPE, é um método utilizado para calcular a precisão de uma determinada previsão. Podemos expressar o cálculo sendo *Xi* como o valor real e *Yi* como o valor previsto. Primeiramente é calculado a diferença entre os dois valores e dividido por *Xi*. Para cada ponto previsto no tempo é feita a soma do valor absoluto do cálculo efetuado e dividido pela quantidade ajustada (n).

$$MAPE = \sum_{i=1}^{n} \frac{|x_i - y_i|}{x_i} \tag{4}$$

A média móvel simples conhecida pela sigla MMS é definida a partir do somatório de um conjunto de valores numéricos que por fim é dividido pela quantidade de elementos do conjunto.

$$MMS = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5}$$

O MME ou média móvel exponencial possui como característica associar um peso maior ao preço mais recente na média com o objetivo de efetuar um rápido acompanhamento das mudanças do preço de uma ação.

$$MME = MME_{d-1} + \frac{2}{n+1} \times \left(P_d - MME_{d-1}\right) \tag{6}$$

O IFR é um indicador criado por J. Welles Wilder, que possui como objetivo informar sobre variações dos preços e a velocidade com que essas variações ocorrem no mercado financeiro. O IFR possui uma variação fixa que varia entre 0 e 100, onde tradicionalmente é considerado superavaliado quando acima de 70 e subavaliado quando abaixo de 30. O FR é definido como as médias de ganhos divididas pelas médias de perdas em um período determinado pelos dias de cotações. Geralmente, é levado em conta um período de 14 dias, pelo padrão sugerido por Wilder.

$$IFR = 100 - (\frac{100}{1+FR}) \tag{7}$$

Na precisão direcional são efetuados cálculos para definir a taxa de acerto direcional com base no conjunto de preços reais e no conjunto dos preços previstos. O cálculo é efetuado através de um algoritmo que percorre a lista de preços previstos e preenche uma lista de direções com valores zero ou um, quando o valor real atual e o valor previsto atual forem maior ou menor que o respectivo valor anterior será adicionado o valor 1 na lista de direções, caso contrário será adicionado o valor 0. Após isso, será feito uma soma de todos os itens da lista de direções e dividido pela quantidade total, retornando a taxa média direcional.

2.6 Bibliotecas e ferramentas utilizadas

Neste estudo, foi utilizado a biblioteca *Numpy*[5] que é focada para a linguagem de programação *Python*[6], cujo objetivo é suportar o processamento de grandes quantidades vetoriais e matriciais além de possuir uma vasta coleção de funções matemáticas. *Tweepy*[7], onde trata-se de uma biblioteca para Python de código aberto que providencia um caminho para os desenvolvedores se comunicarem com a API do Twitter lidando com vários processos de baixo nível como requisições HTTP, autenticações, serialização, etc. *Scikit Learn*[8] para processamento de linguagens naturais, surgiu em 2007 dentro do programa *Google Summer of Code* como um projeto de David Cournapeau. No mesmo ano Matthieu Brucher continuou o projeto como parte de sua tese de doutorado. Em fevereiro de 2010, pesquisadores do INRIA (Institut National de Recherche en Informatique et en Automatique) lançaram a primeira versão. O *Scikit-Learn* predispõe funções importantes para projetos de aprendizado de máquina como *datasets*, dados pré-processados e modelos de *machine learning*. Para análises de sentimentos foi utilizado a biblioteca Leia (ALMEIDA, 2018) disponível no *github*[9] que é baseada no *Vader Sentiment* (GILBERT, 2014)[10] a qual possuí como foco a linguagem portuguesa. Essa biblioteca é utilizada para classificar dados e efetuar a polarização de textos retornando uma análise dos senti-

[5]https://numpy.org/
[6]https://www.python.org/
[7]https://www.tweepy.org/
[8]https://scikit-learn.org/stable/
[9]https://github.com/rafjaa/LeIA
[10]https://github.com/cjhutto/vaderSentiment

mentos em relação as palavras contidas em um dicionário pré-alocado. *Keras*, ou *Tensorflow Keras* [11] é uma *API* do *Tensorflow* utilizada para a criação e treinamento de modelos de aprendizado profundo e a biblioteca Pandas que é uma biblioteca gratuita disponibilizada no site oficial [12] para *Python* utilizada para manipulação e análise de dados.

3 TRABALHOS RELACIONADOS

Para o desenvolvimento desse artigo foi realizada a leitura de trabalhos que abordam questões ligadas a predição de dados através de séries históricas, principalmente artigos onde foram feitos experimentos com a unificação de dados textuais cruzados com dados numéricos. Nessa sessão são destacados alguns textos analisados sobre o tema.

No artigo de Reni A. Steffanon (STEFFENON, 2021) o autor aborda o tema de predição de séries históricas do mercado de ações brasileiro propondo três modelos de redes neurais artificiais para calcular os valores futuros dos históricos de preços. Para este estudo o autor realizou a extração de dados do *Yahoo! Finance* para as ações Banco Inter S.A. (BIDI4), Bco Bradesco S.A. (BBDC4), Cielo S.A. (CIEL3), Magazine Luiza S.A. (MGLU3), Petróleo Brasileiro S.A. Petrobras (PETR4), Vale S.A. (VALE3) e Via S.A (VIIA3). Foram utilizado para cada ativo o preço de abertura (Open), preço máximo (High), preço mínimo (Low), preço de fechamento (Close) e volume negociado (Volume). Para os modelos foram utilizadas camadas do tipo LSTM.

Um dos textos utilizados como base para o artigo, foi o trabalho de Xi Zhang (ZHANG et al., 2017) que detalhou sobre a melhoria das predições do mercado de ações através da fusão de informações heterogêneas, ou seja, informações obtidas de diferentes fontes. Foram extraídos dados de redes sociais e notícias de páginas na internet com o objetivo de investigar os impactos nas movimentações dos preços das ações utilizando uma matriz acoplada e uma estrutura de fatoração de tensores. Um tensor é primeiramente construído para unir dados heterogêneos e obter as relações entre os eventos ocorridos e os sentimentos dos investidores. Devido à dispersão do tensor, foram criadas duas matrizes auxiliares para auxiliar na decomposição tensorial, a matriz de características quantitativas das ações e a matriz de correlação das ações. Como as ações que são altamente correlacionadas entre si tendem a serem afetadas pelo mesmo evento, ao invés de realizar a tarefa de previsão das ações separadamente e independentemente, será feito a previsão de várias.

Também foi utilizada como referência para este estudo a obra de Nti, Adekoya e Weyori (NTI; ADEKOYA; WEYORI, 2021) com foco para fusão de informações de múltiplas fontes para predição de ações de mercado utilizando redes neurais artificiais. É proposto nesse projeto uma arquitetura baseada nas redes neurais convolucionais (CNN) e LSTM que faz a união de ambas as arquiteturas para utilizar na predição de fontes quantitativas e qualitativas, essa

[11] https://keras.io/
[12] https://pandas.pydata.org/

arquitetura hibrida é chamada de *IKN-ConvLSTM*. No processo do *IKN-ConvLSTM* são realizadas as entradas dos dados de ações, termos de buscas mais populares no *Google*, variáveis macroeconômicas, notícias, discussões em fóruns e mensagens do Twitter. Em seguida, é feito o processamento desses dados efetuando a limpeza das informações, agrupamento dos eventos em relação as notícias e análise de sentimento dos textos extraídos dos fóruns e *tweets*. Após isso, os dados das ações são armazenados de forma matricial e em seguida são unidos com os índices do *Google Trends*, dados macroeconômicos, as categorias das notícias e os sentimentos polarizados. Posteriormente, as informações são armazenadas em um banco de dados que transfere as informações para uma rede CNN onde após efetuar a execução os dados são separados em 25% para teste e 75% para treino e, por fim, é feito o treinamento do modelo utilizando *LSTM* e retornando a avaliação dos resultados.

Com base neste estudo inicial foram levantadas possibilidades de atuar no experimento de integração de dados para predição de movimento de ações no mercado brasileiro.

4 MATERIAIS E MÉTODOS

Esta seção descreve a metodologia geral utilizada para o desenvolvimento do trabalho. São descritas as fontes de dados, os processos de aquisição e preparação, a abordagem de integração, além dos modelos descritos para os experimentos de predição.

A seguir são descritos inicialmente, na subseção 4.1, os procedimentos para a aquisição dos dados, incluindo as séries históricas e os processos de extração dos textos, preparação dos dados, polarização dos sentimentos, ou seja, a identificação das frases e palavras contendo sentimentos negativos, positivos e neutros. São demonstradas as polarizações resultantes para as notícias do Google News e mensagens do Twitter. Na subseção 4.2 é descrito o cruzamento das informações através do processo de unificação dos dados a serem mesclados e da etapa de composição das informações obtidas de séries históricas extraídas do Yahoo! Finance, as mensagens já polarizadas do Twitter e os textos coletados do Google News igualmente polarizados.

Na subseção 4.3 são evidenciadas as abordagens para estruturação de três modelos de redes neurais artificiais arquitetados e parametrizados demonstrando suas camadas e valores. A partir das origens de dados descritas, seu processamento e da abordagem de integração, foram realizados experimentos para avaliação dos resultados com cada modelo de arquitetura definido.

4.1 Aquisição e processamento dos dados

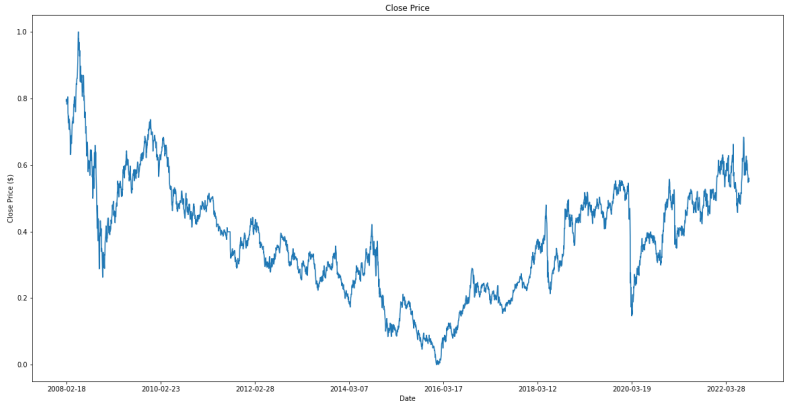
A aquisição dos dados ocorre de forma singular para cada tipo de informação a ser importada, diferenciando a forma de extração das séries históricas, notícias do Google News e postagens do Twitter. A seguir são descritos estes processos.

4.1.1 Séries Históricas

Para a extração dos dados de séries históricas foi utilizada a base de dados do *Yahoo Finances*. Seguindo métricas para filtrar as informações obtidas, utilizamos para o processo as bibliotecas na linguagem de programação *Python*: *yfinance* e *pandas_datareader*, ambas as bibliotecas permitiram implementar a extração dos dados históricos das ações PETR4, VALE3, ITUB4 e BBDC4 entre 01/01/2008 e 19/09/2022 que contabilizaram um total de 3653 registros para cada ativo nesse período.

PETR4 é definido como o código para as ações da Petrobras, cujo destaque se da na produção de petróleo e derivados. A PETR4 pode ser definida como ação preferencial, a ação preferencial é definida por dar o direito ao recebimento dos dividendos aos acionistas com maiores valores do que as ações ordinárias, mas sem disponibilizar o direito ao voto nas assembleias gerais.

Figura 2 – Gráfico mostrando os dados históricos extraídos da Petrobras entre 01/01/2008 e 19/09/2022



Fonte: Elaborado pelo autor.

VALE3 é o código das ações da empresa Vale do Rio Doce que atualmente é uma das maiores mineradoras do mundo, a empresa possui atividades da extração e mineração de ferro, níquel, carvão, manganês, entre outros minerais. Grande parte da produção da Vale do Rio Doce é destinada para exportação. A maior parte dos investidores trabalham com operações mais conservadoras ou de longo prazo utilizando análises fundamentalistas como base. A VALE3 utiliza como base ações ordinárias, as ações ordinárias dão direito a uma parte dos dividendos, mas também permitem o voto nas assembleias gerais da empresa.

A ação BBDC4 representa o Banco Bradesco e é considerada uma das maiores companhias financeiras do Brasil.

ITUB4 é o código da empresa Itaú Unibanco Holding S.A. A empresa é considerada uma das maiores no ramo financeiro do Brasil e conta com mais de 5.142.042.020 ações gerenciadas.

Para padronizar os dados obtidos das séries históricas, foi efetuado um processo de normalização do banco de dados nas colunas '*high*','*low*','*open*','*close*' e '*volume*'. O processo consiste em padronizar os valores de cada coluna variando entre no mínimo zero e máximo um, efetuando a divisão de cada valor pelo valor máximo da coluna em questão.

4.1.2 Dados textuais

Para os dados qualitativos foram utilizados processos de limpeza e filtragem dos textos, além de tratativas específicas para postagens do Twitter e para as notícias do *Google News*. As figuras 3 e 4 apresentam exemplos de dados textuais extraídos. Foram utilizados um total de 2870 registros do Google News e 2749 registros do Twitter para PETR4, um total de 1834 registros do Google News e 1916 registros do Twitter para VALE3, um total de 1783 registros do Google News e 1586 registros do Twitter para BBDC4 e um total de 1844 registros do Google News e 1854 registros do Twitter para ITUB4.

Para os dados de notícias foi efetuado a extração diretamente do portal de notícias *Google News* utilizando a biblioteca *pygooglenews* em um período de tempo predeterminado, no caso dos testes desse trabalho foram efetuados no período entre 01/01/2008 e 19/09/2022. Seguindo uma ideia semelhante ao proposto no artigo de Joshi Kalyani (JOSHI; N.; RAO, 2016) onde as notícias foram coletadas e agregadas as séries históricas de forma que os dados já estavam estruturados e processados para serem classificados. Além disso, 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, unificando as análises obtidas, executando a biblioteca *pygooglenews* em divisões de um mês no período determinado pelo método. Em seguida, foram capturados os dados de cada dia e 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, teremos como resultado um *dataset* com os títulos das notícias unificadas para cada dia.

Figura 3 – Exemplos de textos extraídos do ativo PETR4 pelo Google News

Data	Texto extraído
15/07/2010	Brasil começa a produção comercial do 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	ConJur - Petrobras é condenada a indenizar família por acidente com explosivo - Consultor Jurídico
30/08/2010	MP investigará venda de ativos da Petrobras à Braskem - VEJA
15/09/2010	MP vai investigar se policiais de SP quebraram sigilos para Petrobras - Globo.com

Fonte: Elaborado pelo autor.

Os dados do Twitter foram extraídos com base em um período de tempo entre 01/01/2008 e 19/09/2022, eliminando feriados, finais de semana e mantendo apenas dias úteis. O artigo de James Briggs (BRIGGS, 2020) sobre análise de sentimentos para predição dos preços de ações em *Python*, ajudou a efetuar a extração dos dados para o *Twitter*, principalmente na utilização da *API*. A extração das postagens foi provenientes de quatro fontes de notícias do mercado financeiro sendo elas, *Investing*, *Money Times*, *InfoMoney* e Valor Econômico. Como parâmetro

de pesquisa foi utilizada uma combinação dos nomes do ativo com o nome da empresa que está sendo efetuada a extração. Após o processo de extração, foram gerados quatro conjuntos de dados e em seguida esses conjuntos de dados foram unificados em apenas um conjunto para que seja feita a análise de sentimentos posteriormente.

Figura 4 – Exemplos de textos extraídos do ativo PETR4 pelo Twitter

Data	Texto extraído
23/10/2009	petrobras, tam e net captam us$ 5,6 bilhões http://tinyurl.com/yzcpce6
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 http://tinyurl.com/y8hagtr

Fonte: Elaborado pelo autor.

4.1.3 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 foram baseadas nos processos de vetorização e indexação (SCHUMAKER; CHEN, 2006), onde as palavras foram transformadas em *tokens* como será visto posteriormente. Primeiramente, foram removidos textos duplicados, em seguida os textos foram convertidos para letras minúsculas garantindo uma padronização, após isso, removemos pontuações, números e símbolos que não seriam utilizáveis mantendo apenas espaços em branco e alguns caracteres especiais que estão ligados à língua portuguesa (áéíóúãõàèìòùçâêîôû). Logo após, foi efetuado a remoção dos nomes dos ativos utilizados nos testes (petrobras, vale, itaú, bradesco), foram removidos também endereços "*http*"e eliminação de espaços em branco desnecessários. O próximo passo foi efetuar a tokenização dos textos separando cada palavra e armazenando em uma lista, seguindo o processo realizamos uma remoção de todas as palavras vazias (*stop words*) da lista de *tokens* gerada anteriormente. Neste processo, foi utilizado um dicionário de *stop words* contido na biblioteca "*spacy*". Após isso, foi efetuado uma análise em cada palavra restante da lista para verificar se ela consta em um banco de dados de palavras preestabelecido. Caso a palavra não exista no banco de dados, ela será removida da lista de *tokens*. Por fim, utilizamos um processo de lematização das palavras a fim de deflexionar os *tokens* e retornar as palavras a sua forma base. Após os processos, um novo *dataframe* é criado com os dados textuais já limpos e unificados pela data.

4.1.4 Análise de sentimentos

Neste trabalho, vamos efetuar uma análise de sentimentos utilizando dados extraídos do Twitter e de notícias do *Google News*. Os dados foram lidos em um período predeterminado e assim como no trabalho sobre *Deep Clue* na sessão sobre aprendizagem de redes sociais (SHI et

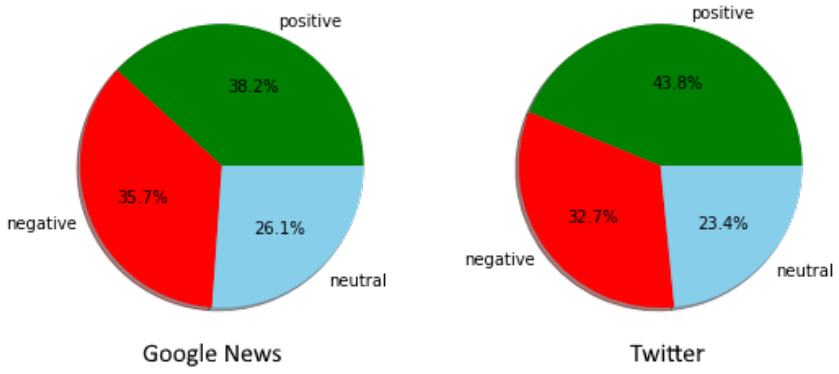
al., 2019), os dados para cada mensagem foram armazenados como uma cadeia de caracteres. Em seguida, as mensagens ou notícias foram unidas para cada dia e retornaram uma análise sentimental positiva ou negativa. Vamos utilizar palavras e frases com determinados pesos para definir a positividade e negatividade da análise.

Tabela 1 – Exemplos de textos extraídos do ativo PETR4 pelo Google News com os respectivos sentimentos analisados

Sentimento	Texto	positive	negative	neutral
Positivo	Brasil tem sete entre 500 maiores empresas do mundo, aponta 'Fortune'	0.495	0	0.505
Negativo	ConJur - Petrobras é condenada a indenizar família por acidente com explosivo	0	0.5	.5
Neutro	O que foi 2015 para os brasileiros segundo as pesquisas no Google	0	0	1

Após a preparação dos dados, foi realizada a análise dos sentimentos. Para isso, vamos utilizar a biblioteca *Leia* (ALMEIDA, 2018) que é uma variação da biblioteca *Vader Sentiments* (GILBERT, 2014) onde é utilizado um dicionário focado para a língua portuguesa. Primeiramente para aplicarmos os processos de polarização dos sentimentos, vamos iterar sobre os dados obtidos após a preparação, unificando os textos para cada dia especificamente, após isso vamos executar um método para classificar a polarização dos sentimentos (*polarity_scores*). Os sentimentos são classificados em três categorias, sendo elas sentimentos positivos, negativos e neutro. Valores relativos percentuais entre 0 e 1 e outra métrica de composição dos sentimentos chamada *compound* que é definida como a soma dos valores da polarização, onde no final terá como resultante um valor normalizado entre –1 e +1, sendo –1 um resultado extremamente negativo e +1 um resultado extremamente positivo. Dessa forma conseguimos analisar o sentimento geral em relação ao dia corrente. Algo que pode ser destacado são os experimentos realizados no trabalho de Faten Alzazah (ALZAZAH; CHENG, 2020) onde os resultados de notícias negativas tiveram mais influência no mercado de ações dos que notícias positivas. A tabela 1 apresenta exemplos de textos analisados.

Figura 5 – Porcentagem de sentimentos extraídos do Google News e Twitter da Petrobras entre período de 2008 e 2022.



Fonte: Elaborado pelo autor.

4.2 Integração de informações

Para tentar melhorar a predição, utilizamos alguns métodos para concatenar os dados textuais extraídos do Twitter e notícias do *Google News* com valores numéricos dos históricos de ações especialmente extraídas do *Yahoo Finances* em determinados períodos temporais. Nesta seção são descritas as abordagens utilizadas para integrar os conjuntos multimodais utilizados, ou seja, valores numéricos e dados textuais.

4.2.1 Unificação dos dados

Para a unificação dos dados foram integrados três conjuntos de dados sendo eles, os dados históricos normalizados, os dados polarizados pelas análises de sentimentos de notícias do *Google News* e de os dados polarizados das postagens provenientes do Twitter.

Inicialmente filtramos os dados para utilizar apenas informações pertinentes ao projeto, sendo que para o histórico de preços foram mantidos apenas os dados "*date*", "*close*", "*open*", "*high*", "*low*"e "*volume*". Já para os sentimentos analisados do *Twitter* e *Google News* utilizamos as seguintes colunas de dados "*date*", "*compound*", "*negative*", "*neutral*", "*positive*".

A próxima etapa será renomear as colunas de alguns dados para diferenciar as informações do *Twitter* e do *Google News*. As colunas a serem renomeadas são, "*compound*", "*negative*", "*neutral*", "*positive*"onde ficaram consecutivamente com os nomes "*compound_gn*", "*negative_gn*", "*neutral_gn*", "*positive_gn*"para o *Google News* e "*compound_tw*", "*negative_tw*", "*neutral_tw*", "*positive_tw*"para o *Twitter*.

Na etapa seguinte, todos os conjuntos de dados já filtrados e renomeados foram unidos aos dados dos históricos de preços utilizando a data como valor de referência resultando em um conjunto de dados com os seguintes campos '*date*','*close*','*compound_gn*', '*negative_gn*', '*neutral_gn*', '*positive_gn*', '*compound_tw*', '*negative_tw*', '*neutral_tw*', '*positive_tw*', '*open*', '*high*', '*low*', '*volume*'.

Para evitar dias com valores do tipo *NaN*, ou seja, sem valores numéricos devido a não haver postagens ou notícias no dia em questão, utilizamos um método para preencher os dados de forma a evitar as polarizações sem valores. Caso o primeiro dia não possua valores, este será definido como neutro colocando o valor zero para as colunas "*compound*", "*positive*"e "*negative*"e o valor um para a coluna "*neutral*". Para os dias posteriores, caso não haja valores polarizados, será feito um processo de redução em 50% para o próximo dia em relação ao anterior. Neste caso os dados do dia atual das colunas de sentimentos com valor *NaN* foram alterados, de forma que a coluna "*compound*", "*positive*"e "*negative*"receba o valor do dia anterior dividido por dois. Para garantir o equilíbrio a coluna "neutral"irá receber a soma das diferenças entre o dia anterior e o dia atual das colunas "*positive*"e "*negative*"após a divisão.

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.

As tabelas seguintes exemplificam o processo.

Tabela 2 – Exemplo utilizando dias sem postagens

Data	Fechamento	Compound Twitter
2022-05-20	0.025	0.153
2022-05-21	0.05	NaN
2022-05-22	0.033	NaN
2022-05-23	0.04	-1.0
2022-05-24	0.045	NaN

Tabela 3 – Exemplo utilizando método de preenchimento com 50% do valor anterior.

Data	Fechamento	Compound Twitter
2022-05-20	0.025	0.153
2022-05-21	0.05	0.0765
2022-05-22	0.033	0.03825
2022-05-23	0.04	-1.0
2022-05-24	0.045	-0.5

Após esta etapa, a coluna referente as datas foram definidas como a coluna de índices do conjunto. Em seguida, vamos copiar o conjunto já unificado para outra variável e remover os 30 primeiros dias para utilizarmos no cálculo de IFR.

4.2.2 Composição dos dados para treinamento

Para a composição dos dados de treinamento dos modelos utilizamos uma série de procedimentos para preparar todas as informações necessárias para os modelos. Inicialmente foi criada uma variável *data_to_use* que registra a quantidade de dados a serem usados do *dataframe* principal, sendo que para este trabalho será utilizado 100% do conjunto de dados. Em seguida foi criada a variável *train_end*, que armazena a quantidade de dados a serem usados no treinamento. Foram utilizados para os modelos 80% dos dados para treino e 20% para teste. Após os processos iniciais de definição das variáveis, foi feita a coleta dos dados de cada coluna os quais foram armazenados em variáveis, a fim de serem usadas no treinamento posteriormente.

Logo após esta etapa, foram estabelecidas mais três variáveis *close_price_shifted*, *compound_gn_shifted* e *compound_tw_shifted*, sendo que essas variáveis vão armazenar dados referentes com um dia de antecedência. Foram usadas a variável *close_price_shifted* para os dados das colunas de fechamento e as variáveis *compound_gn_shifted* e *compound_tw_shifted* para sentimentos polarizados. Em seguida, foi definido um novo parâmetro para treinamento de modo a calcular o índice de força relativa também conhecido como IFR.

Na tabela 4 pode ser visto um exemplo de composição de dados, onde na coluna *compound_gn* estão os números referentes a soma dos valores das polarizações do Google News, negative_gn, neutral_gn e positive_gn contém os percentuais de negatividade, neutralidade e positividade dos textos do Google News e negative_tw, neutral_tw e positive_tw contém os percentuais de negatividade, neutralidade e positividade dos textos do Twitter.

Tabela 4 – Exemplo do processamento, filtragem e unificação dos dados do Twitter e Google News para a ação PETR4.SA entre 01/01/2008 e 19/09/2022

compound_gn	negative_gn	neutral_gn	positive_gn	compound_tw	negative_tw	neutral_tw	positive_tw
0.4019	0.000	0.828	0.172	0.6908	0.110	0.668	0.222
0.1531	0.076	0.827	0.097	0.1779	0.000	0.876	0.124
0.4939	0.063	0.810	0.126	0.0000	0.000	1.000	0.000
0.0000	0.000	1.000	0.000	0.4588	0.000	0.826	0.174
-0.5859	0.096	0.881	0.023	0.9201	0.000	0.880	0.120
0.2263	0.000	0.863	0.137	0.3612	0.000	0.839	0.161
-0.5267	0.074	0.895	0.031	0.2023	0.068	0.833	0.098
-0.1027	0.069	0.931	0.000	-0.5423	0.209	0.791	0.000
0.0000	0.000	1.000	0.000	0.7096	0.000	0.836	0.164
0.8481	0.055	0.796	0.149	0.3612	0.000	0.848	0.152

Na tabela 5 pode ser visto um exemplo de composição completa de dados.

Tabela 5 – Exemplo do processamento dos dados e unificação dos dados do Twitter e Google News para dados extraídos da ação PETR4.SA após processamentos e shifts

date	close_price	close_price _shifted	compound_gn	compound _gn_shifted	compound_tw	compound _tw_shifted	volume	open_price	high	low
2021-01-06	30.100000	31.000000	0.7906	0.7096	0.6369	0.0258	96562500	30.160000	30.900000	30.049999
2021-01-07	31.000000	31.120001	0.7096	0.4404	0.0258	0.9062	56171300	30.340000	31.150000	30.340000
2021-01-08	31.120001	30.860001	0.4404	0.0000	0.9062	0.1531	67136300	31.459999	31.760000	30.350000
2021-01-11	30.860001	30.629999	0.0000	0.0000	0.1531	0.6249	48744700	30.610001	31.059999	30.400000
2021-01-12	30.629999	29.150000	0.0000	0.6369	0.6249	0.9744	65691900	31.120001	31.559999	30.629999
2021-01-13	29.150000	29.450001	0.6369	-0.1280	0.9744	0.9300	93826600	30.680000	30.860001	29.000000
2021-01-14	29.450001	28.120001	-0.1280	0.0000	0.9300	0.7184	50745400	29.170000	29.670000	28.719999
2021-01-15	28.120001	28.690001	0.0000	0.4588	0.7184	0.5859	80673300	29.049999	29.080000	28.030001
2021-01-19	28.690001	28.209999	0.4588	0.4215	0.5859	0.9716	61656000	28.480000	28.860001	27.639999
2021-01-20	28.209999	27.549999	0.4215	0.6808	0.9716	0.7003	60306200	28.950001	29.120001	28.110001

4.2.3 Calculando a Média Móvel

De forma optativa, podemos escolher entre utilizar a média móvel simples (MMS) ou média móvel exponencial (MME) para as predições. Em ambos os casos foram utilizados dois cálculos de média móvel, sendo eles para um período de 30 dias e 5 dias. Após efetuar os cálculos foram geradas duas colunas para o conjunto de dados de treinamento, sendo uma para a média móvel no período de 5 dias e outra para o período de 30 dias. Por fim, concatenamos os dados em um único conjunto para ser utilizado no treinamento.

4.3 Parametrização e Arquitetura dos modelos

Utilizamos três modelos de redes neurais, sendo primeiramente utilizado um modelo composto por 2 camadas que foi proposto na pesquisa da autora Michele Rosa (ROSA, 2020). A primeira camada é uma LSTM com 16 neurônios e como camada de saída um Dense com 1 neurônio, é possível visualizar a estrutura na tabela 6.

O segundo modelo foi extraído do artigo proposto por Reni Steffenon (STEFFENON, 2021) sendo estruturado por 4 camadas, com a primeira camada sendo uma LSTM de 200 neurônios a segunda camada LSTM de 300 neurônios, a terceira camada uma LSTM de 400 neurônios e tendo como saída uma Dense de 1 neurônio, as camadas LSTM deste modelo possuem a ativação como uma tangente hiperbólica (tanh) e como ativação recorrente a função *sigmoid*, as duas últimas camadas LSTM possuem um dropout de 3%, a estrutura do modelo pode ser vista na tabela 7.

O terceiro modelo segue a composição de redes neurais profundas e foi proposto por Yang Lyla (LYLA, 2020). O modelo é composto por 3 camadas, sendo a camada de entrada uma Dense de 32 neurônios e ativação do tipo *relu*, novamente na segunda camada temos uma Dense de 8 neurônios com ativação *relu* e por fim uma saída Dense de um neurônio, podemos visualizar a estrutura na tabela 8.

Tabela 6 – Primeiro modelo LSTM proposto

Ordem	Modelo	Nº neurônios	Ativação	Ativação Recorrente	Dropout	Feedback
1	LSTM	16			20%	não
2	Dense	1				não

Fonte: Elaborado pelo autor.

Tabela 7 – Segundo modelo LSTM proposto

Ordem	Modelo	Nº neurônios	Ativação	Ativação Recorrente	Dropout	Feedback
1	LSTM	200	tanh	sigmoid	3%	sim
2	LSTM	300	tanh	sigmoid	3%	sim
3	LSTM	400	tanh	sigmoid	3%	não
4	Dense	1				não

Fonte: Elaborado pelo autor.

Tabela 8 – Terceiro modelo seguindo a arquitetura DNN

Ordem	Modelo	Nº neurônios	Ativação	Ativação Recorrente	Dropout	Feedback
1	Dense	32	relu			não
2	Dense	8	relu			não
3	Dense	1				não

Fonte: Elaborado pelo autor.

5 RESULTADOS

Para a obtenção dos resultados foi efetuada uma análise das combinações de dados para testar cada modelo. Em seguida, foi efetuado a aplicação de todos os modelos combinando os dados do Google News, Twitter, cálculo do índice de força relativa (IFR) e média móvel exponencial (MME).

Portanto, nesta seção são comparados os modelos LSTM sem retroalimentação, LSTM com retroalimentação e o modelo estruturado apenas com rede neural profunda. Os modelos foram testados utilizando combinações diversas de parâmetros para o ativo PETR4, sendo que as análises podem ser conferidas nas figuras da subseção 5.1. Em seguida, nas subseções 5.2, 5.3, 5.4 e 5.5, foram efetuados testes utilizando três modelos de redes neurais artificiais sendo eles, um modelo estruturado como uma rede neural profunda composta por camadas do tipo *Dense*, um modelo de rede neural estruturado em *LSTM* com retroalimentação e outro modelo *LSTM* sem retroalimentação, a estruturação dos modelos pode ser vista na subseção 4.3, esses modelos foram utilizados para treinamento dos ativos PETR4 (Petróleo Brasileiro SA), VALE3 (Vale S.A.), BBDC4 (Banco Bradesco SA) e ITUB4 (Itaú Unibanco) no período entre 01/01/2008 e 19/09/2022, exibindo os resultados obtidos através de gráficos e demonstrando as métricas do erro-médio, erro médio absoluto e precisão direcional.

5.1 Análise de combinações de dados

Para confirmarmos a hipótese de que o conjunto ampliado de dados garante melhores resultados na predição, realizamos experimentos comparando as diferentes combinações disponíveis. Abaixo, são descritos estes experimentos.

Para os testes foram utilizadas combinações que incluem as configurações a seguir. A primeira (referida como "sem nenhum parâmetro"se refere aos testes usando apenas os dados numéricos de compra e venda. Os demais são configurações entre as possibilidades existentes.

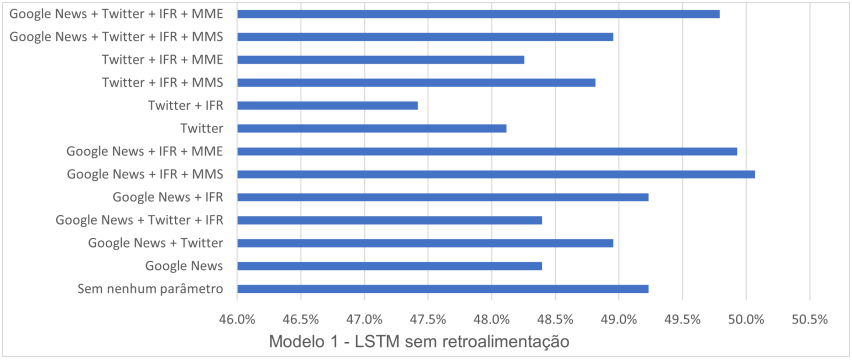
- Sem nenhum parâmetro
- Apenas Google News
- Apenas Twitter
- Google News + Twitter
- Google News + Twitter + cálculo de IFR
- Google News + cálculo de IFR
- Google News + cálculo de IFR + MMS
- Google News + cálculo de IFR + MME
- Twitter + cálculo de IFR
- Twitter + cálculo de IFR + MMS
- Twitter + cálculo de IFR + MME
- Google News + Twitter + cálculo de IFR + MMS
- Google News + Twitter + cálculo de IFR + MME

As séries históricas dos preços de ações estiveram presentes em todos os parâmetros dos testes.

Foram efetuados testes utilizando os modelos de LSTM sem retroalimentação (Tabela 6), LSTM com retroalimentação (Tabela 7) e o modelo de redes neurais profunda (Tabela 8) sendo utilizados múltiplos parâmetros de entrada conforme as figuras 6, 7 e 8 para verificar a diferença nos resultados.

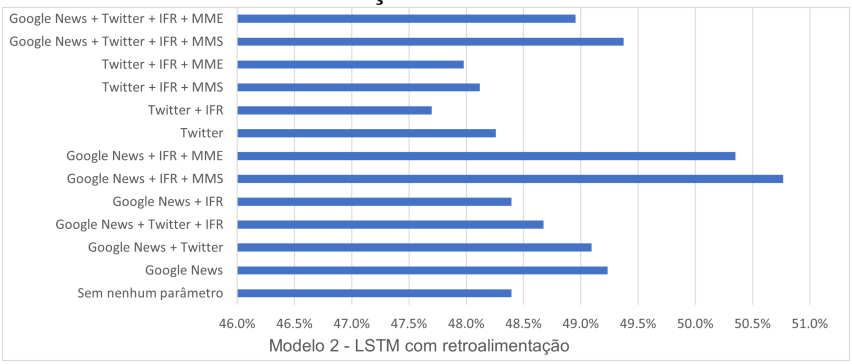
Conforme as figuras 6, 7 e 8 ficou evidenciado que os melhores resultados foram obtidos com a junção de notícias do Google News e com cálculo de média móvel das séries históricas, mas ao incluir os dados das análises de sentimentos do Twitter ocorreram quedas nas precisões. Isso ocorre devido a sujeira nas mensagens e imprecisões das postagens em comparação as notícias devido ao número de dados obtidos.

Figura 6 – Gráfico mostrando as precisões de cada combinação de parâmetros no ativo PETR4 usando o modelo LSTM sem retroalimentação



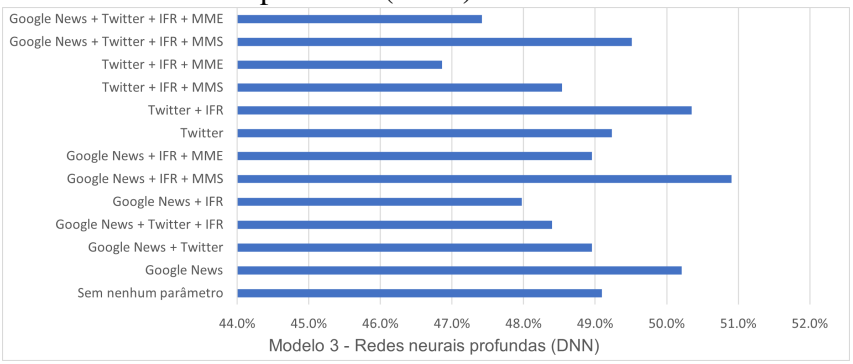
Fonte: Elaborado pelo autor.

Figura 7 – Gráfico mostrando as precisões de cada combinação de parâmetros no ativo PETR4 usando o modelo LSTM com retroalimentação



Fonte: Elaborado pelo autor.

Figura 8 – Gráfico mostrando as precisões de cada combinação de parâmetros no ativo PETR4 usando o modelo de redes neurais profunda (DNN)



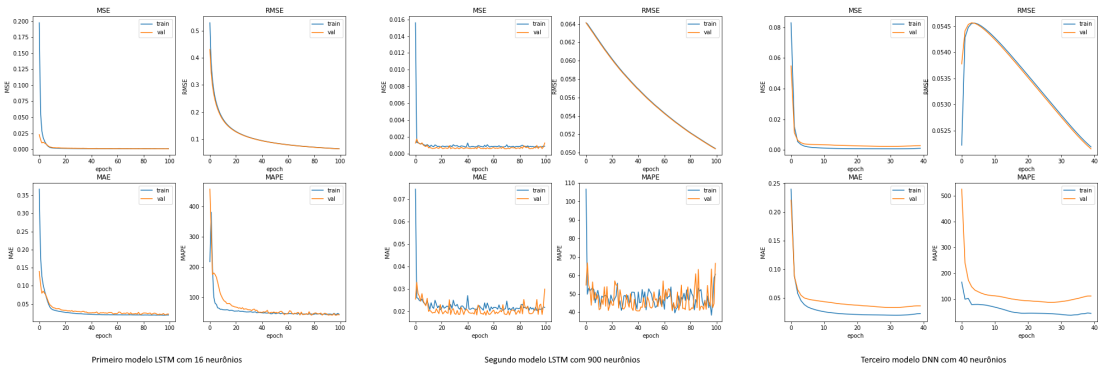
Fonte: Elaborado pelo autor.

5.2 PETR4

Após aplicar o primeiro modelo, obtivemos uma precisão de 49,23%, além disso a diferença prevista foi de 0,5531. Esse valor foi obtido através do cálculo da raiz quadrada do erro-médio, o erro médio absoluto foi de 0,014265. Utilizamos dados extraídos do Yahoo Finances, Google News e Twitter da Petrobras entre o dia 01/01/2008 e 19/09/2022.

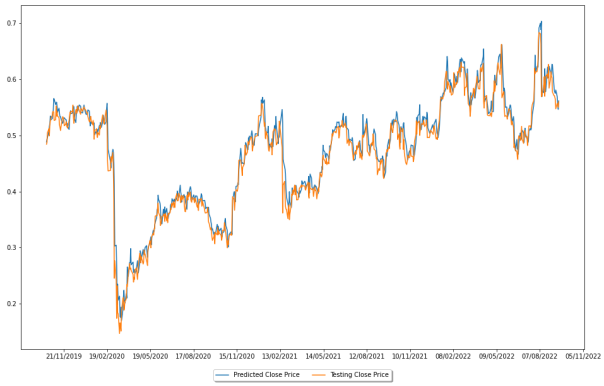
Os dados foram utilizados de acordo com a combinação completa de todos os tipos de dados disponíveis, que apresenta a melhor precisão, conforme as figuras 6, 7, 8.

Figura 9 – Gráficos mostrando as métricas de cada modelo para PETR4



Fonte: Elaborado pelo autor.

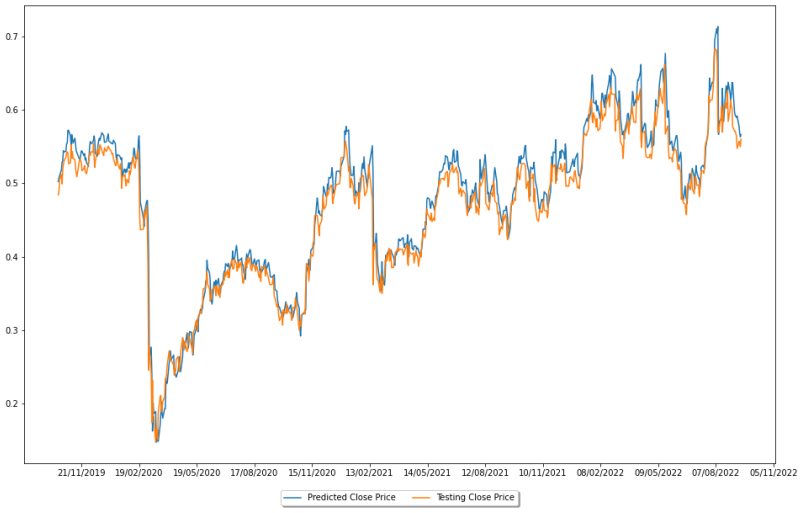
Figura 10 – Gráfico mostrando as predições do primeiro modelo para a PETR4



Fonte: Elaborado pelo autor.

Nesta etapa, aplicamos o segundo modelo projetado por Reni Steffenon (STEFFENON, 2021), onde a precisão foi de 48,54% com a diferença prevista de aproximadamente 0,5565 e o erro médio absoluto foi de 0,017415.

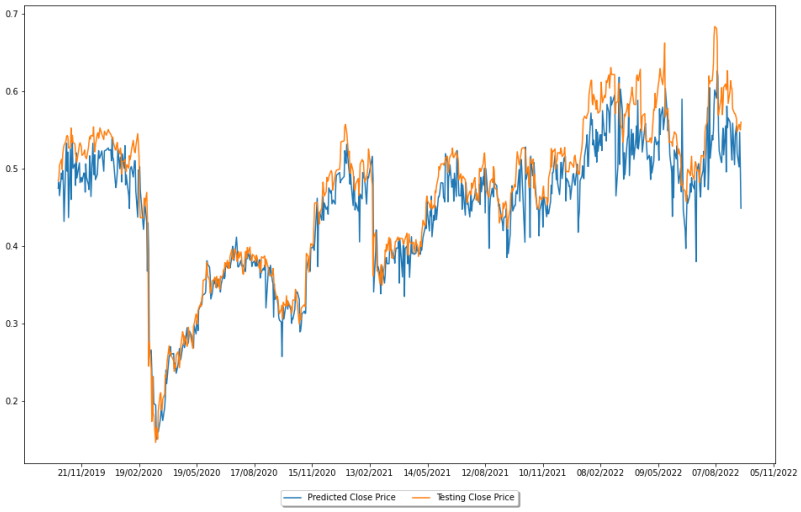
Figura 11 – Gráfico mostrando as predições do segundo modelo para a PETR4



Fonte: Elaborado pelo autor.

Por último, utilizamos um terceiro modelo projetado por Yang Lyla (LYLA, 2020) que utilizou a arquitetura DNN. Nesse teste foi obtido uma previsão de 51,46%, a diferença prevista foi de aproximadamente 0,5261 e o erro médio absoluto foi de 0,027728.

Figura 12 – Gráfico mostrando as predições do terceiro modelo para a PETR4



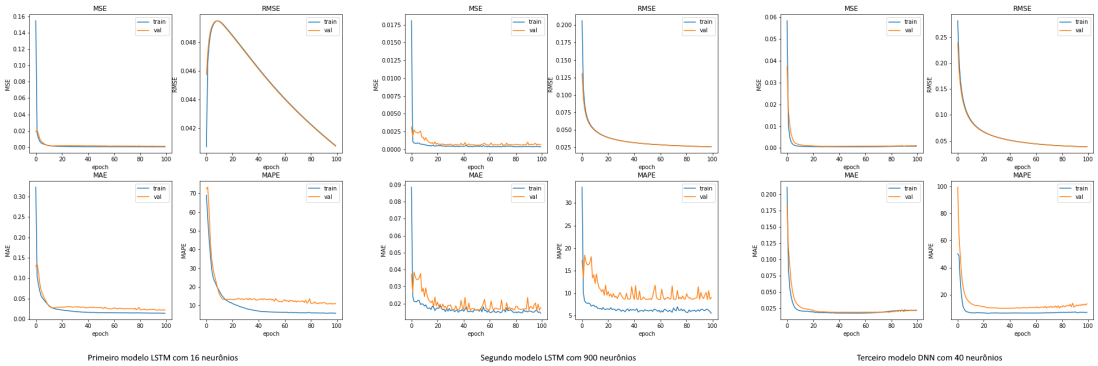
Fonte: Elaborado pelo autor.

5.3 VALE3

Ao utilizar o primeiro modelo, obtivemos um resultado com precisão de 51,04%, a diferença prevista foi de 0,434 onde esse valor foi obtido através do cálculo da raiz quadrada do erro médio e o erro médio absoluto foi de 0,04061. Utilizamos dados extraídos do Yahoo Finances, Google News e Twitter da empresa Vale entre o dia 01/01/2008 e 19/09/2022.

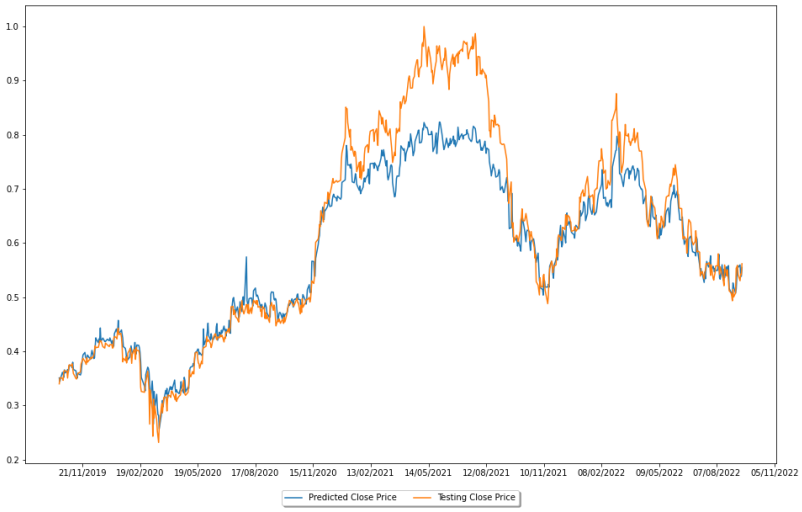
Os dados foram utilizados de acordo com a combinação completa de todos os tipos de dados disponíveis, que apresenta a melhor precisão, conforme as figuras 6, 7, 8.

Figura 13 – Gráficos mostrando as métricas de cada modelo para VALE3



Fonte: Elaborado pelo autor.

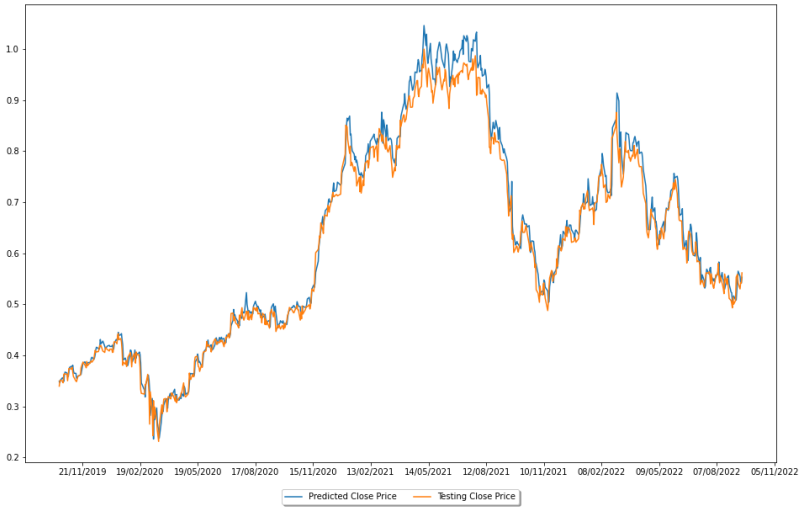
Figura 14 – Gráfico mostrando as predições do primeiro modelo para a VALE3



Fonte: Elaborado pelo autor.

Em seguida, aplicamos o segundo modelo projetado por Reni Steffenon (STEFFENON, 2021), onde foi obtido uma precisão de 49,24% com uma diferença prevista de aproximadamente 0,4426 e o erro médio absoluto de 0,01961.

Figura 15 – Gráfico mostrando as predições do segundo modelo para a VALE3



Fonte: Elaborado pelo autor.

Por último, utilizamos um terceiro modelo projetado por Yang Lyla (LYLA, 2020), tendo como a precisão do resultado 53,19%, uma diferença prevista de aproximadamente 0,427 e o erro médio absoluto de 0,039188.

Figura 16 – Gráfico mostrando as predições do terceiro modelo para a VALE3

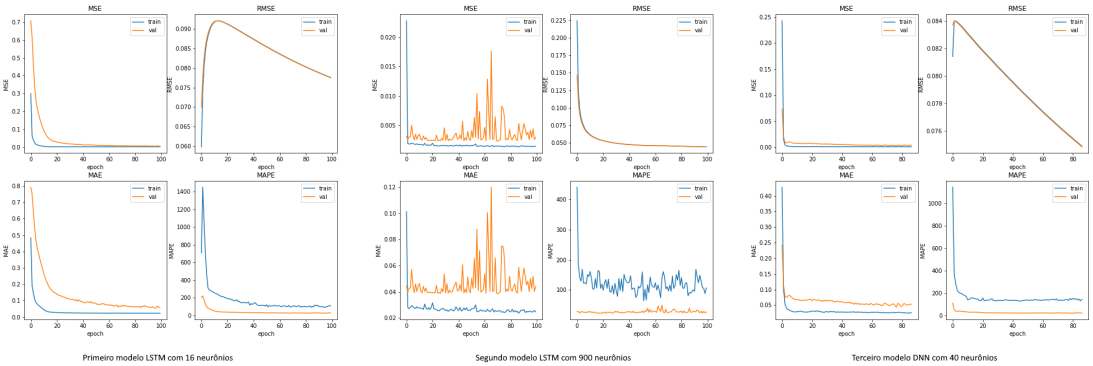
Fonte: Elaborado pelo autor.

5.4 BBDC4

Na utilização do primeiro modelo, obtivemos uma precisão de 50% com uma diferença prevista de 0,4636, esse valor foi obtido através do cálculo da raiz quadrada do erro médio, o erro médio absoluto foi de 0,018296. Utilizamos dados extraídos do Yahoo Finances, Google News e Twitter do banco Bradesco entre o dia 01/01/2008 e 19/09/2022.

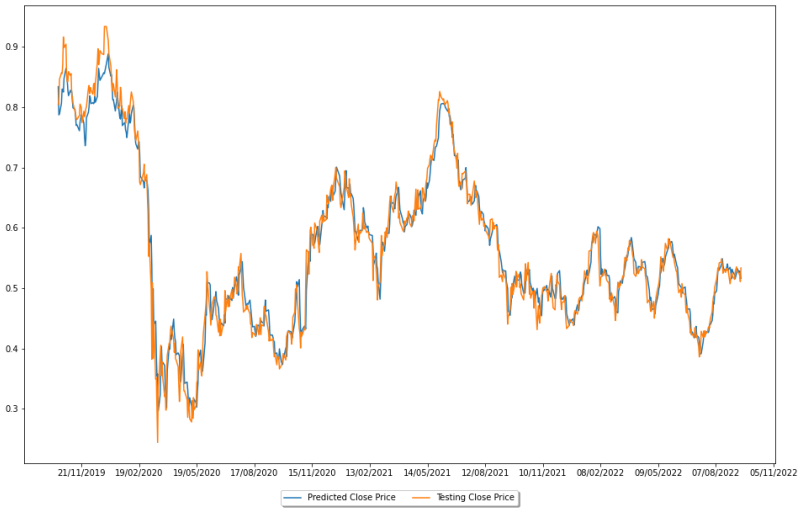
Os dados foram utilizados de acordo com a combinação completa de todos os tipos de dados disponíveis, que apresenta a melhor precisão, conforme as figuras 6, 7, 8.

Figura 17 – Gráficos mostrando as métricas de cada modelo para BBDC4



Fonte: Elaborado pelo autor.

Figura 18 – Gráfico mostrando as predições do primeiro modelo para a BBDC4



Fonte: Elaborado pelo autor.

Em seguida, aplicamos o segundo modelo projetado por Reni Steffenon (STEFFENON, 2021), onde conseguimos uma predição de 49,58% junto com uma diferença prevista de 0,468 e o erro médio absoluto foi de 0,01927.

Figura 19 – Gráfico mostrando as predições do segundo modelo para a BBDC4

Fonte: Elaborado pelo autor.

Por último, utilizamos um terceiro modelo baseado em uma arquitetura DNN projetado por Yang Lyla (LYLA, 2020), onde obteve uma precisão direcional de 48,6% com a diferença prevista de aproximadamente 0,4687 e o erro médio absoluto foi de 0,02049.

Figura 20 – Gráfico mostrando as predições do terceiro modelo para a BBDC4

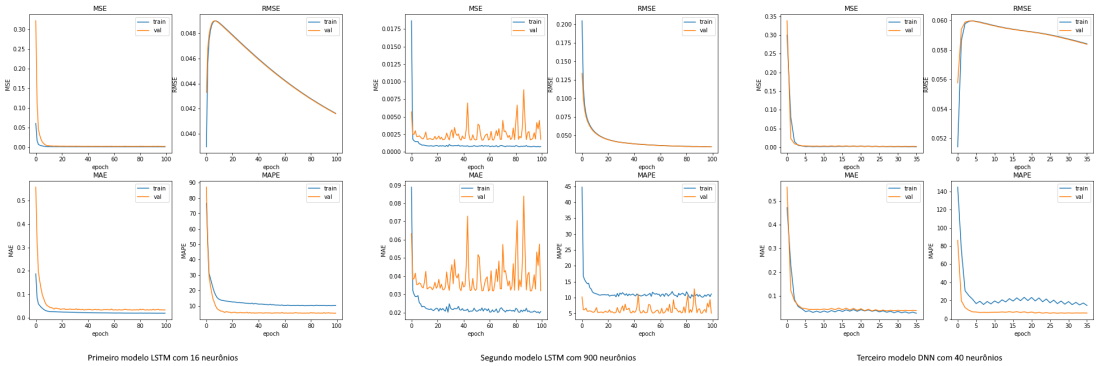
Fonte: Elaborado pelo autor.

5.5 ITUB4

Para o ativo ITUB4 iniciamos usando o primeiro modelo, o qual obtivemos uma precisão de 47,9% com a diferença prevista de 0,405, onde esse valor foi obtido através do cálculo da raiz quadrada do erro médio e o erro médio absoluto foi de 0,01526. Utilizamos dados extraídos do Yahoo Finances, Google News e Twitter do banco Itaú entre o dia 01/01/2008 e 19/09/2022.

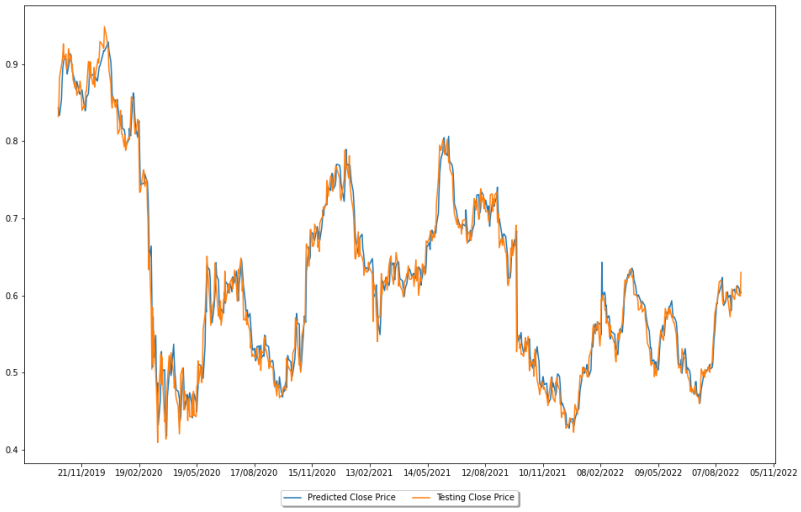
Os dados foram utilizados de acordo com a combinação completa de todos os tipos de dados disponíveis, que apresenta a melhor precisão, conforme as figuras 6, 7, 8.

Figura 21 – Gráficos mostrando as métricas de cada modelo para ITUB4



Fonte: Elaborado pelo autor.

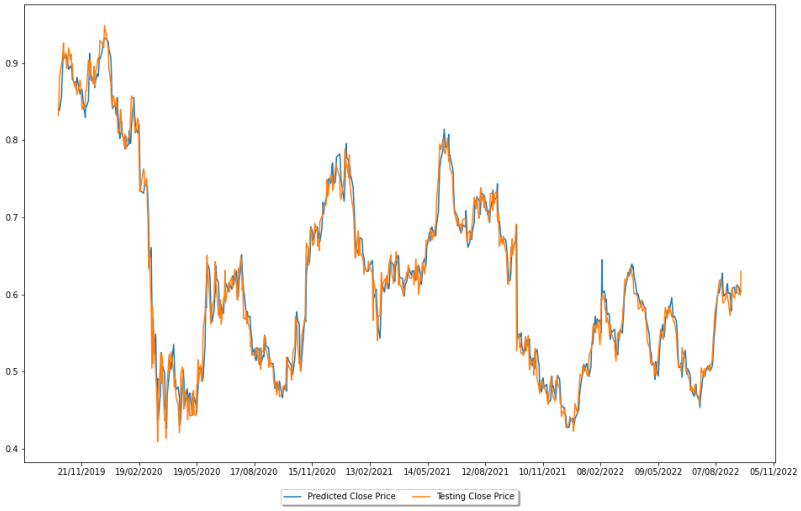
Figura 22 – Gráfico mostrando as predições do primeiro modelo para a ITUB4



Fonte: Elaborado pelo autor.

Em seguida, aplicamos o segundo modelo projetado por Reni Steffenon (STEFFENON, 2021), tendo uma precisão de 47,33% com a diferença prevista foi de aproximadamente 0,4038 e o erro médio absoluto foi de 0,01512.

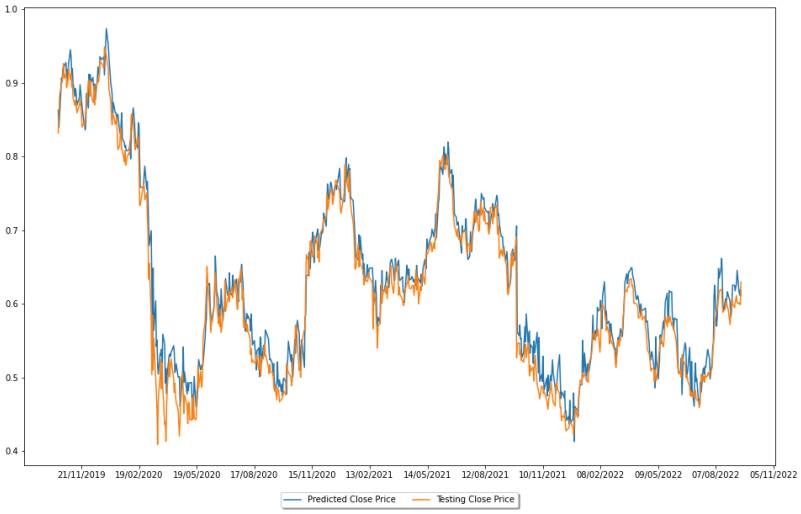
Figura 23 – Gráfico mostrando as predições do segundo modelo para a ITUB4



Fonte: Elaborado pelo autor.

Por último, utilizamos um terceiro modelo projetado por Yang Lyla (LYLA, 2020), com uma precisão direcional de 47,34%, uma diferença prevista de aproximadamente 0,4172 e o erro médio absoluto foi de 0,021086.

Figura 24 – Gráfico mostrando as predições do terceiro modelo para a ITUB4



Fonte: Elaborado pelo autor.

5.6 Sobre os resultados

Para o ativo PETR4, o terceiro modelo obteve uma maior precisão no gráfico final com o RMSE igual a 0,5261 e uma precisão e 51,46% em seguida no ativo VALE3, o segundo modelo também foi o que mais se aproximou na predição com o RMSE igual a 0,427 e uma precisão direcional de 53,19%.

No ativo BBDC4, o primeiro modelo obteve uma precisão com menos oscilações tendendo a uma aproximação dos testes no gráfico final com o RMSE igual a 0,4636 e uma precisão de 50%.

Para o ativo ITUB4 o primeiro modelo foi o que obteve a maior aproximação entre as predições tendo uma precisão direcional de 47,9% e o RMSE de 0,405.

6 CONCLUSÕES E TRABALHOS FUTUROS

O principal objetivo desse estudo foi efetuar a fusão de dados qualitativos como mensagens do Twitter e notícias da web juntamente com dados quantitativos como series temporais do mercado de ações. Para isso, foram utilizadas as técnicas de aprendizagem de máquina LSTM e DNN em três modelos de redes neurais artificiais distintos. Os ativos PETR4 e VALE3 apresentaram bons resultados na aproximação das predições com resultados que variaram de 51% a 53,19% conforme as figuras 12, 14 e 16. Já o ativo BBDC4 apresentou resultados medianos com destaque para a figura 18 a qual obteve uma precisão de 50% tendo utilizado o primeiro modelo de redes neurais. Por fim, o ativo ITUB4 apresentou baixa predição em todos os modelos com

precisões que variaram entre 47% e 48%.

Por meio dos testes efetuados nesse artigo foi possível analisar que a união de dados numéricos com dados textuais pode melhorar a predição, mas os processos de limpeza e preparação dos dados assim como a extração devem ser rigorosos para evitar inconsistência no treinamento dos modelos.

Para possíveis trabalhos futuros podemos ampliar o número de fontes a serem utilizadas na união dos dados além disso utilizar outras técnicas de redes neurais como redes convolucionais (CNN), GRU ou *IKN-ConvLSTM* (NTI; ADEKOYA; WEYORI, 2021).

Referências

ALMEIDA, R. J. A. **LeIA - Léxico para Inferência Adaptada**. [S.l.]: GitHub, 2018. <https://github.com/rafjaa/LeIA>.

ALZAZAH, F.; CHENG, X. Recent Advances in Stock Market Prediction Using Text Mining: A Survey. **Department of Computer Science, Middlesex University**, 2020.

BRIGGS, J. Sentiment Analysis for Stock Price Prediction in Python. **Towards Data Science**, 2020. Disponível em: <https://towardsdatascience.com/sentiment-analysis-for-stock-price-prediction-in-python-bed40c65d178>.

GILBERT, C. H. E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: **Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf**. [S.l.: s.n.], 2014.

GOOGLEFINANCE. **Google Finanças: preços da bolsa de valores, cotações em tempo real e notícias sobre o mercado financeiro**. 2006. Disponível em: <https://www.google.com/finance/>.

IBM. **Neural Networks**. 2021. Disponível em: <https://www.ibm.com/cloud/learn/neural-networks>.

IBM. **Recurrent Neural Networks**. 2021. Disponível em: <https://www.ibm.com/cloud/learn/recurrent-neural-networks>.

JOHNSON, J. What's a deep neural network? deep nets explained. **BMC**, 2020. Disponível em: <https://www.bmc.com/blogs/deep-neural-network/>.

JOSHI, K.; N., B.; RAO, J. Stock Trend Prediction Using News Sentiment Analysis. **Department of Computer Engineering, KJSCE, Mumbai**, 2016.

LYLA, Y. **A Quick Deep Learning Recipe: Time Series Forecasting with Keras in Python**. [S.l.]: Towards Data Science, 2020. <https://towardsdatascience.com/a-quick-deep-learning-recipe-time-series-forecasting-with-keras-in-python-f759923ba64>.

MENDES, D. Mercado reduz, de novo, projeção de inflação para 2022 e 2023. **CNN Brasil**, 2022. Disponível em: <https://www.cnnbrasil.com.br/business/numero-de-investidores-na-bolsa-cresce-15-em-2022-apostando-na-diversificacao/>.

MOTLEYFOOL. **The Motley Fool:Stock Investing and Stock Market Research**. 1993. Disponível em: <https://www.fool.com/>.

NTI, I.; ADEKOYA, A.; WEYORI, B. A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction. **Journal of Big Data**, 2021. Disponível em: <https://doi.org/10.1186/s40537-020-00400-y>.

ORACLE. **Previsão do mercado de ações usando as técnicas de Deep Learning**. 2020. Disponível em: <https://www.oracle.com/br/artificial-intelligence/what-is-ai/>.

ROSA, M. **Previsão do mercado de ações usando as técnicas de Deep Learning**. 2020.

ROSA, M. Projeção dos preços de ações do mercado brasileiro com integração de dados da análise grafistas e fundamentalistas. **Universidade do Vale do Rio dos Sinos**, 2022.

SCHUMAKER, R.; CHEN, H. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System. **The University of Arizona**, 2006.

SHI, L. et al. DeepClue: Visual Interpretation of Text-Based Deep Stock Prediction. **IEEE Transactions On Knowledge And Data Engineering**, v. 31, n. 6, 2019.

STEFFENON, R. Modelo adaptativo para projeção de preços de ações no mercado brasileiro. **Universidade do Vale do Rio dos Sinos**, 2021.

YAHOOFINANCE. **Yahoo Finance - Stock Market Live, Quotes, Business and Finance News**. 1997. Disponível em: <https://finance.yahoo.com/>.

ZHANG, X. et al. Improving stock market prediction via heterogeneous information fusion. **ScienceDirect**, 2017.