INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS – *CAMPUS* BAMBUÍ
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

Emanuel Elias Ferreira

ANÁLISE DO IMPACTO DE NOTÍCIAS NA PREVISÃO DO ÍNDICE BOVESPA

Bambuí - MG
2023

EMANUEL ELIAS FERREIRA

ANÁLISE DO IMPACTO DE NOTÍCIAS NA PREVISÃO DO ÍNDICE BOVESPA

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí para obtenção do grau de Bacharelado em Engenharia de Computação.

Orientador: Prof. Dr. Marcos Roberto Ribeiro

Bambuí - MG
2023

Catalogação na Fonte Biblioteca IFMG - *Campus* Bambuí

F383a Ferreira, Emanuel Elias.
Análise do impacto de notícias na previsão do Índice Bovespa. / Emanuel Elias Ferreira. – 2023
74 f. : il. ; color.

Orientador: Dr. Marcos Roberto Ribeiro.
Trabalho de Conclusão de Curso (graduação) - Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí, MG, Curso Bacharelado em Engenharia de Computação, 2023.

1. Análise de sentimentos. 2. Ibovespa. 3. Mercado de ações. I. Ribeiro, Marcos Roberto. II. Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí, MG. III. Título.

CDD 332.642

Elaborada por Douglas Bernardes de Castro - CRB-6/2802

Emanuel Elias Ferreira

ANÁLISE DO IMPACTO DE NOTÍCIAS NA PREVISÃO DO ÍNDICE BOVESPA

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí para obtenção do grau de Bacharelado em Engenharia de Computação.

Aprovado em 09 de agosto de 2023 pela banca examinadora:

Prof. Dr. Marcos Roberto Ribeiro – IFMG – *Campus* Bambuí – (Orientador)

Prof. Dr. Ciniro Aparecido Leite Nametala – Instituto Federal de Minas Gerais

Prof. Me. Cláudio Ribeiro de Sousa – Instituto Federal de Minas Gerais

Documento assinado eletronicamente por **Marcos Roberto Ribeiro**, **Professor**, em 09/08/2023, às 14:52, conforme Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ciniro Aparecido Leite Nametala**, **Professor**, em 09/08/2023, às 14:52, conforme Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Claudio Ribeiro de Sousa**, **Professor EBTT**, em 09/08/2023, às 14:52, conforme Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade do documento pode ser conferida no site https://sei.ifmg.edu.br/consultadocs informando o código verificador **1632824** e o código CRC **BECFED8C**.

Bambuí - MG
2023

Este trabalho é dedicado à minha família, em especial, à minha mãe, Marina, e ao meu irmão, Lucas. Este trabalho também é dedicado a todas as pessoas que buscam conhecimento, não para benefício próprio, mas sim com o objetivo de melhorar o mundo.

AGRADECIMENTOS

Ao meu orientador, professor Doutor Marcos Roberto Ribeiro, pelos conselhos, paciência e incentivo durante a realização deste trabalho, além de me inspirar tanto profissionalmente quanto na minha vida pessoal.

Aos meus amigos, por tornarem minha vida mais agradável nos momentos longe de minha família.

Aos meus professores do IFMG, pelos ensinamentos passados no decorrer da minha graduação.

"Life is not a game of luck. If you want to win, work hard."
(Sora)

RESUMO

Atualmente, o principal canal de informação dos brasileiros são as redes sociais. Em uma sociedade na qual os cidadãos possuem fácil acesso à informação, a propagação de notícias ocorre rapidamente. Visto que, através de redes sociais as informações podem ser facilmente disseminadas. Este trabalho, se propõe à auxiliar na previsão de preço do IBOV, considerando os efeitos de informações veiculadas na internet. A fim de contribuir para a compreensão de seus impactos e, como resultado, minimizar os prejuízos. Em virtude que é possível, estabelecer uma relação, entre o conteúdo das notícias e *tweets* com as oscilações do mercado financeiro. Os modelos preditivos desenvolvidos foram obtidos com os métodos da classe Box-Jenkins, e enriquecidos com o valor de sentimento de notícias e *tweets*. Os modelos que obtiveram melhores resultados consideravam os sentimentos de *tweets*. De 2018 a 2022, o modelo de previsão diária que considerava toda a série histórica do preço de fechamento do IBOV, obteve melhor desempenho. Considerando apenas os sentimentos de *tweets*, o protótipo foi capaz de realizar previsões com diferença média entre o valor previsto e o valor real de aproximadamente 0,0066 de MAPE.

Palavras-chave: Análise de Sentimentos. Ibovespa. Mercado de Ações. Processamento de Linguagem Natural.

ABSTRACT

Currently, the principal source of information for Brazilians is social networks. In a society where citizens have easy access to information. The spread of news occurs rapidly. Since, through social networks, information can effortlessly propagate. This work proposes to conduct the IBOV forecast, considering the effects of data transmitted on the internet to contribute to understanding its outcomes. And as a result, minimize damage. Because it is possible to relative news and tweet content with oscillations in the financial market. The predictive models developed utilizing techniques derived from the Box-Jenkins class. And enriched with the sentiment values of news and tweets. The models that achieved the best results considered the sentiments of tweets. From 2018 to 2022, the daily forecast model that considered the entire historical series of the closing price of the IBOV, performed better considering only the sentiments of tweets. The prototype is capable of forecasting with an average difference between the predicted value and the actual value of approximately 0.0066 of MAPE.

Keywords: Ibovespa. Stock Market. Sentiment Analysis. Natural Language Processing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT - Associação Brasileira de Normas Técnicas

ACM - Association for Computing Machinery

ARIMA - Autoregressive Integrated Moving Average

AS - Análise de Sentimentos

B3 - Bolsa de Valores Oficial do Brasil (Antiga BM&FBOVESPA)

BM&F Bovespa - Bolsa de Valores, Mercadorias e Futuros

BOVA11 - Fundo de investimentos do índice Bovespa

BOVESPA - Bolsa de Valores de São Paulo

BVSP - Código de identificação do índice Bovespa na B3

CETIP - Central de Custódia e de Liquidação Financeira de Títulos

CNN - Convolutional Neural Network

ETF - Exchange Traded Funds

IBOV - Índice Bovespa

IBOVESPA - Índice Bovespa

IEEE - Institute of Electrical and Electronics Engineers

IFMG - Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais

INPI - Instituto Nacional de Propriedade Industrial

LeIA - Analisador Léxico de Inferência Aplicada

LSTM - Long Short-Term Memory

MAE - Mean Absolute Error

MAPE - Mean Absolute Percentage Error

MLP - Multilayer Perceptron Neural Network

MSE - Mean Squared Error

NLP - Natural Language Processing

RBM - Restricted Boltzmann Machines

RNN - Recurrent Neural Network

RMSE - Root Mean Squared Error

SVM - Support Vector Machine

TCC - Trabalho de Conclusão de Curso

VADER - Analisador de sentimentos, também conhecido por VaderSentiment

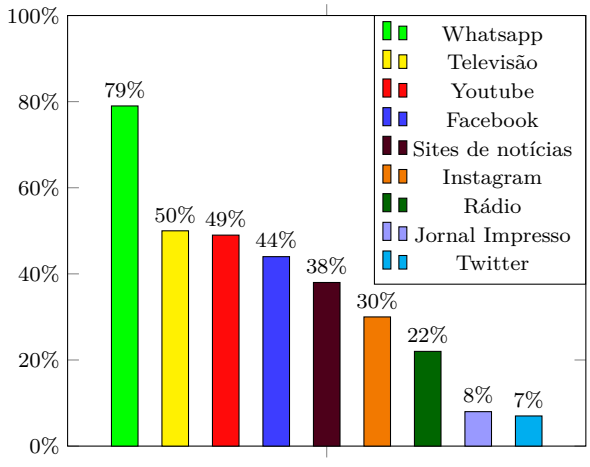
SUMÁRIO

1 INTRODUÇÃO

A partir da década de 1990, a Inteligência Artificial se desenvolveu consideravelmente, em especial, nas áreas de Processamento de Linguagem Natural, de Aprendizado de Máquina e de Representação de Conhecimento (RUIZ; FOGUEM; GRABOT, 2014; SUN; LUO; CHEN, 2017; GÜNDÜZ *et al.* 2019). Tal desenvolvimento permitiu a solução de problemas consideravelmente complexos, como a detecção de notícias falsas (SILVA *et al.* 2020; MORENO; BRESSAN, 2019; MONTEIRO *et al.* 2018; SHU *et al.* 2017) e a predição de impactos causados por redes sociais (CAROSIA; COELHO; SILVA, 2019; MEDEIROS; BORGES, 2019; MACHADO; PEREIRA, 2018; ARAÚJO; MARINHO, 2018).

As movimentações de impacto significativo no valor de mercado das empresas dependem de notícias e eventos externos (MALKIEL, 2003; FAKHRY, 2016). Portanto, é interessante que o investidor tenha conhecimento a respeito da repercussão de notícias relacionadas a empresas e segmentos de mercado de seu interesse. Contudo, devido ao grande volume de notícias na Internet, existe uma dificuldade no trabalho dos investidores ao analisar como as notícias afetam o mercado (CHAN, 2003; LI *et al.* 2014). Assim, é importante que essas pessoas tenham acesso a ferramentas capazes de auxiliar na previsão do movimentos dos preços de ações, levando em consideração a análise de notícias e seus impactos no mercado financeiro. Atualmente, a principal fonte de informação dos brasileiros são as redes sociais (DATASENADO, 2019). Por meio da Figura 1, é possível observar as principais fontes de informação dos brasileiros.

Figura 1 – Gráfico sobre os principais canais de informação dos brasileiros.



Fonte: Adaptado de DataSenado (2019).

É importante destacar que os brasileiros utilizam, principalmente, os meios digitais para se informar. De acordo com Araújo e Marinho (2018) e Carosia, Coelho e Silva (2021), as informações veiculadas nos *sites* de notícias e redes sociais podem afetar o mercado financeiro.

O Índice Bovespa (Ibovespa), cujo código de identificação é *BVSP*, popularmente conhecido por IBOV, representa uma carteira teórica com os principais ativos das empresas mais impactantes do mercado financeiro do Brasil (B3, 2023a). Segundo B3 (2023a), o Ibovespa é o indicador de desempenho mais utilizado para monitorar o mercado de ações brasileiro. Por consequência, é relevante analisar e prever o impacto de influências externas (notícias) sobre o valor desse índice.

1.1 Objetivo Geral

O objetivo principal do presente trabalho foi analisar o impacto de notícias e postagens de redes sociais na previsão do valor do IBOV.

1.1.1 Objetivos Específicos

Para atingir o objetivo principal, os seguintes objetivos específicos foram estabelecidos:

- Construir uma base de dados com histórico do Ibovespa, notícias e *tweets* relacionados;
- Aplicar a análise de sentimentos sobre as bases de dados com o intuito de classificar as notícias como positivas, negativas ou neutras;
- Implementar um método de previsão do Ibovespa considerando o sentimento das notícias.

1.2 Justificativa

Apesar da existência de diversos trabalhos nessa área, como a previsão sobre as influências de mídias sociais e notícias, segundo Carosia, Coelho e Silva (2019, 2020, 2021), a literatura brasileira carece de trabalhos em língua portuguesa. Inclusive, a maioria dos trabalhos correlatos, que empregam a análise de sentimentos com o intuito de analisar ou prever variações na bolsa de valores brasileira, foram publicados em língua inglesa (CAROSIA; COELHO; SILVA, 2020; MACHADO; PEREIRA, 2018; MEDEIROS; BORGES, 2019; ARAÚJO; MARINHO, 2018; JUNIOR; SALOMON; OLIVEIRA PAMPLONA *et al.* 2014; SILVA, 2021).

Na literatura internacional, existem diversos estudos sobre a previsão do mercado financeiro considerando notícias (ELSHENDY *et al.* 2018; SUDHAKAR; NAGANJANEYULU, 2020; SHARMA *et al.* 2020; FORECASTING. . . , 2016). Do ponto de vista científico, é interessante desenvolver e utilizar técnicas computacionais, em especial, a análise de sentimentos, para determinar como as notícias podem impactar na previsão do Ibovespa. Além disso, essas técnicas podem ser aplicadas em outras esferas importantes, como política e segurança digital. Do ponto de vista socioeconômico, uma previsão do

impacto das notícias no Ibovespa pode auxiliar investidores do mercado de ações a evitarem prejuízos que afetem a economia e, por consequência, a sociedade.

Atualmente, não há patentes no Instituto Nacional de Propriedade Industrial (INPI) relacionadas à abordagem adotada neste trabalho. Existem algumas patentes encontradas no INPI que tratam de programas de computador que auxiliam na previsão e análise do mercado financeiro. Contudo, nenhuma dessas patentes aborda um modelo preditivo que utilize técnicas de previsão de séries temporais em conjunto com a análise de sentimentos. Logo, existe a possibilidade de criação de um produto passível de ser patenteado, com grande potencial de aplicação prática.

1.3 Resultados Esperados

Após a conclusão do presente trabalho, diversos resultados interessantes podem ser alcançados. Os algoritmos desenvolvidos para obtenção das bases de dados podem ser aplicados para desenvolvimento de outras bases de dados não necessariamente relacionadas ao mercado financeiro. As bases de dados obtidas com a conclusão deste trabalho podem ser utilizadas por outras pessoas e ferramentas que visem aprimorar análises relacionadas ao Ibovespa. Essas bases de dados também podem ser úteis para treino e teste de classificadores de sentimentos em língua portuguesa, com foco no mercado financeiro. O mesmo método desenvolvido pode ser aplicado em outros índices e ativos do mercado financeiro. A partir deste trabalho, vislumbra-se que seja possível desenvolver ferramentas com métodos de previsão para apoiar os investidores em suas tomadas de decisão. Espera-se que seja possível aplicar o mesmo método de previsão em outras esferas da sociedade, como política, segurança pública, *marketing*, mercado de trabalho,*etc.*

1.4 Organização da Monografia

Esta monografia está organizada como se segue: no Capítulo 2, são abordados os principais trabalhos correlatos e fundamentos teóricos. Em seguida, o Capítulo 3 apresenta os materiais e métodos utilizados, além de discutir sobre a classificação da pesquisa realizada. No Capítulo 4, são descritos o desenvolvimento do trabalho e os experimentos conduzidos. Por fim, o Capítulo 5 apresenta as considerações finais.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta a fundamentação teórica a respeito do tema estudado e também trabalhos correlatos. Inicialmente, a Seção 2.1 expõe os conceitos de Recuperação de Informação. Na Seção 2.2, são apresentados conceitos sobre a Bolsa de Valores brasileira e o índice Bovespa. Em seguida, a Seção 2.3 aborda os principais métodos de previsão e análise do mercado financeiro. Posteriormente, a Análise de Sentimentos é explicada na Seção 2.4. Logo após, a Seção 2.5 trata da previsão de séries temporais. Por fim, um comparativo sobre os principais trabalhos correlatos é apresentado na Seção 2.6.

2.1 Recuperação de informação

A Recuperação de Informação é uma extensa área da Computação que se concentra na extração, organização e visualização de dados. Baeza-Yates (2013) define a recuperação de informação como:

> A recuperação de informação trata da representação, armazenamento, organização e acesso a itens de informação, como documentos, páginas Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia, etc. A representação e a organização dos itens de informação devem fornecer aos usuários facilidade de acesso às informações de seu interesse.

O autor afirma que houve um desenvolvimento da área além das expectativas iniciais. Na atualidade, a Recuperação de Informação também inclui modelagem, classificação de textos, visualização de dados, arquitetura de sistemas, interface homem-máquina, filtragem e linguagens.

Para satisfazer as necessidades dos usuários, um sistema de recuperação de informação deve interpretar os dados coletados nos documentos e classificá-los de acordo com o grau de relevância da consulta solicitada. Tal classificação envolve as análises léxica, sintática e semântica do texto (e, em alguns casos, a análise de sentimentos). Então, de acordo com Baeza-Yates (2013), o objetivo principal da aplicação de técnicas de recuperação de informação é recuperar todas as informações relevantes à necessidade do usuário. A recuperação de informação é uma área fundamental para o presente trabalho, visto que, por meio de seus conceitos, foi possível desenvolver as bases de dados relacionadas ao IBOV, bem como seu histórico na Bolsa de Valores brasileira.

2.2 Bolsa de Valores brasileira

De acordo com Ramos (2023), a Bolsa de Valores brasileira é popularmente conhecida por B3, com base no nome da instituição Brasil, Bolsa, Balcão (B3). A B3 é uma empresa de capital aberto sediada em São Paulo. Foi declarada aberta em 2017, após a fusão das empresas Bolsa de Valores, Mercadorias e Futuros de São Paulo (BM&F

Bovespa) e a Central de Custódia e de Liquidação Financeira de Títulos (CETIP) (B3, 2023b; RAMOS, 2023).

A BM&F Bovespa era uma instituição resultante da junção de duas outras empresas em 2008: a Bolsa de Valores de São Paulo (Bovespa) e a BM&F (B3, 2023b; RAMOS, 2023). Na época, a CETIP e a BM&F Bovespa eram consideradas os principais agentes do mercado financeiro e de capitais. Essas instituições eram responsáveis por todos os processos realizados, desde negociações, registros, ofertas, liquidações, operacionalizações de ativos e demais operações financeiras do mercado (B3, 2023b; RAMOS, 2023).

No contexto do mercado de ações, denomina-se índice a representação quantificável de uma carteira hipotética de papéis listados na Bolsa de Valores. Essa representação é calculada conforme o desempenho desses papéis (B3, 2023a). De acordo com B3 (2023a), o índice mais popular na B3 é o IBOV, que foi idealizado em 1968 e, ao longo de décadas, consolidou-se como referência para os investidores do mercado de capitais brasileiro. De forma geral, o IBOV é o principal índice de desempenho da B3, representando o desempenho médio das empresas de maior importância listadas na B3. Reavaliado a cada quatro meses, esse índice é resultado de uma carteira teórica de ativos (ações e *units*) composta de companhias listadas na B3 que atendem a certos critérios (B3, 2023a). O IBOV corresponde a cerca de 80% do número de negócios e do volume financeiro da Bolsa de Valores brasileira (B3, 2023a).

Segundo B3 (2023a), os critérios de seleção das companhias que compõem o IBOV são:

> Estar entre os ativos que representem 85% em ordem decrescente de Índice de Negociabilidade (IN); 95% de presença em pregão; 0,1% do volume financeiro no mercado à vista (lote-padrão) e não ser *penny stock*.

Basicamente, são selecionadas as empresas consideradas mais relevantes e negociadas na Bolsa de Valores brasileira. As ações do tipo *penny stock* são desconsideradas devido ao seu baixo valor de mercado (comumente próximo ou menor que 1 real) (B3, 2023a).

Segundo B3 (2023a), o Ibovespa é o principal indicador de desempenho das ações negociadas na B3, e se consolidou como referência para os investidores do mercado de capitais brasileiro. De modo geral, o índice Bovespa representa, de forma quantitativa, a fotografia do desempenho das ações mais relevantes na B3 (B3, 2023a). Por exemplo, quando noticiado na mídia que a Bolsa fechou em queda, significa que o valor do IBOV diminuiu em relação ao valor anterior em determinado período. Por consequência, é possível concluir que o desempenho de empresas que compõem a carteira teórica do IBOV também diminuiu. Todavia, nesse mesmo período, pode haver empresas que compõem o IBOV e tiveram melhor desempenho. Entretanto, a maioria das empresas teve pior desempenho.

2.3 Análises do mercado de ações

A maioria dos trabalhos correlatos que utilizam técnicas de processamento de textos para aprimorar métodos de análise ou previsão de índices do mercado financeiro recorrem à *Hipótese do Mercado Eficiente* para fundamentar seu desenvolvimento, visto que essa hipótese considera que eventos e fatores externos podem influenciar no mercado de ações (CAROSIA; COELHO; SILVA, 2020, 2019; BOLLEN; MAO; ZENG, 2011; MONTEIRO; FERREIRA, 2021; SILVA, 2021; RUNDO *et al.* 2019; SHARMA *et al.* 2020; JUNIOR; SALOMON; OLIVEIRA PAMPLONA *et al.* 2014).

Segundo Malkiel e Fama (1970), a *hipótese de mercado eficiente* parte do pressuposto de que altos retornos no mercado financeiro não podem ser obtidos por análise do histórico dos preços das ações. Todavia, os preços das ações no mercado podem ser, pelo menos, parcialmente previsíveis. Isso acontece porque as movimentações de impacto significativo no valor de mercado das empresas dependem também de eventos externos (MALKIEL, 2003). Esses eventos externos, por exemplo, podem ser notícias, posicionamento de chefes de Estado na mídia ou redes sociais, guerras, conflitos e tensões geopolíticas, desastres ambientais, *etc.* Além disso, outro fator influenciador na movimentação de ações é a influência das emoções humanas em relação às tomadas de decisões (DE LONG *et al.* 1990).

Atualmente, os métodos mais utilizados nas análises e previsões do mercado financeiro são baseados na combinação de técnicas de aprendizagem de máquina com análises do mercado (ATSALAKIS; VALAVANIS, 2009; SHAH; ISAH; ZULKERNINE, 2019). No atual estado da arte, as abordagens que demonstraram melhores resultados consideram eventos externos. Para o desenvolvimento dessas análises, são aplicadas combinações de diferentes técnicas de Inteligência Artificial, modelos estatísticos, análises de mercado e a análise de sentimentos (SHAH; ISAH; ZULKERNINE, 2019; SILVA, 2021).

As análises do mercado financeiro que fundamentam o desenvolvimento deste trabalho são a análise técnica, a análise de séries temporais e a hipótese do mercado eficiente. Em virtude de estudar as influências das notícias em relação ao histórico do IBOV, faz-se necessário aplicar a análise técnica com o intuito de realizar análises gráficas em relação a diversos intervalos de tempo (LEMOS, 2017; MARCA; ANTUNES, 2017). A análise de séries temporais é empregada justamente para examinar o histórico do IBOV, a fim de prevê-lo com base em seu comportamento passado, sendo essa análise fundamentada pela Estatística (HYNDMAN; ATHANASOPOULOS, 2018). Por fim, a hipótese do mercado eficiente é empregada, visto que, no modelo preditivo, também são consideradas as influências de fatores e eventos externos à análise; nesse caso, as influências de notícias.

2.4 Análise de sentimentos

A Análise de Sentimentos (AS) é uma subárea do Processamento de Linguagem Natural, podendo ser combinada com técnicas de aprendizagem de máquina para analisar quantidades abundantes de dados textuais de modo a classificá-los com relação à subjetividade das informações (CAMBRIA *et al.* 2017). Basicamente, a análise de sentimentos pode usar a abordagem léxica baseada em *corpus* ou em textos avaliativos.

A abordagem léxica é embasada na utilização de um analisador léxico que compara os caracteres de entrada com caracteres presentes em dicionários preexistentes. A entrada é classificada conforme o valor previamente atribuído a determinadas sequências de caracteres presentes nesses dicionários. Essa técnica é a mais comumente utilizada na literatura devido à sua simplicidade de implementação (MEDHAT; HASSAN; KORASHY, 2014; PRABOWO; THELWALL, 2009).

A abordagem baseada em *corpus* é semelhante à léxica, mas utiliza dicionários com conjuntos de textos subjetivos para avaliação. Assim, são empregadas várias sequências de palavras para a entrada do analisador (FELDMAN, 2013; LIU, B. *et al.* 2010). Por exemplo, o termo "queda", isoladamente, possui sentimento negativo; todavia, em expressões como "queda no prejuízo", indica um sentimento positivo. Caso fosse empregada apenas a abordagem léxica na análise dessa sentença, a classificação dessa expressão tenderia a ser errônea, visto que a base de dados de um analisador léxico contém as palavras de forma isolada. Em contrapartida, a abordagem baseada em *corpus*, por exemplo, pode classificar expressões do tipo "queda" e "prejuízo" como sentimento positivo.

A abordagem fundamentada em textos avaliativos utiliza-se de técnicas de aprendizado de máquina para analisar as sentenças semanticamente (LIN; HE, 2009; LIU, 2015). Para a execução desta técnica, é necessária uma coleção de frases com sentimentos classificados. Nesta coleção, cada expressão é previamente avaliada manualmente e rotulada em termos de sentimento. Um algoritmo de aprendizado de máquina é treinado com as expressões rotuladas para reconhecer o sentimento presente em novos textos. Em geral, a classificação não é ótima quando o algoritmo treinado sobre um domínio de texto tenta classificar textos de outro domínio. Por exemplo, quando o algoritmo é treinado com textos de redes sociais e aplicado em textos de notícias. Todavia, para textos de mesmo domínio, essa abordagem de análise de sentimentos é a mais eficiente (CAMBRIA *et al.* 2017; LIU, 2015).

A abordagem de textos avaliativos não é simples de ser implementada, visto que um grande esforço deve ser conduzido para classificar coleções de expressões de sentimentos, previamente às etapas de treino e teste de algoritmos de aprendizagem de máquina. Além disso, seu desempenho é limitado, de acordo com o domínio das expressões de sentimentos presentes nas bases de dados de treino e teste (CAMBRIA *et al.* 2017; LIU, 2015).

2.5 Previsão de séries temporais

De acordo com Morettin e Toloi (2022), denomina-se série temporal uma lista ordenada de observações de um fenômeno particular no decorrer do tempo; por exemplo, os preços das ações de uma determinada empresa coletados ao longo de um intervalo de tempo (semanas, dias, minutos). A ordem dessas observações importa, e, por consequência, é necessário aplicar técnicas específicas de processamento de dados para que essa estrutura temporal não seja comprometida (MORETTIN; TOLOI, 2022). É importante ressaltar que essas variáveis podem ser quantitativas ou qualitativas, sendo que, no exemplo anterior, trata-se de uma variável quantitativa contínua. Exemplos de séries temporais qualitativas podem ser as palavras em um determinada notícia ou uma série de *tweets*.

A tarefa de previsão de séries temporais, popularmente conhecida como *forecasting*, refere-se à técnica de empregar dados observados de uma série em questão, a fim de estimar valores ainda desconhecidos dessa mesma série ou de outras séries relacionadas (MORETTIN; TOLOI, 2022). As técnicas de previsão de séries temporais são fundamentais para auxiliar análises e tomadas de decisões. A abordagem adotada para a realização deste trabalho fundamentou-se em técnicas de econometria combinadas com Análise de Sentimentos.

O modelo autorregressivo integrado de médias móveis (ARIMA) é uma classe de modelos estatísticos (Box-Jenkins), comumente aplicado para analisar e prever dados de séries temporais. ARIMA é um acrônimo que significa *Autoregressive Integrated Moving Average* Morettin e Toloi (2022) e Zhang (2003). Denominam-se as siglas dos parâmetros desse modelo devido à atribuição das letras iniciais das palavras de língua inglesa que nomeiam cada método embutido na classe do ARIMA, sendo essas técnicas do modelo: autorregressor, integrador e médias móveis. Por meio da combinação dessas técnicas, é possível obter modelos variados, desde a aplicação de apenas uma delas até a combinação de todas.

Por exemplo, caso a série temporal analisada já seja estacionária, não é necessário realizar as diferenciações. Isto é, caso a série temporal possua as propriedades estatísticas (média, variância, autocorrelação, etc.) constantes ao longo do tempo, não é necessário transformá-la em estacionária através do cálculo entre a diferença absoluta ou percentual entre uma observação e outra (MORETTIN; TOLOI, 2022; ZHANG, 2003). Portanto, pode ser aplicado um modelo de Médias Móveis Autorregressivas em séries que já são estacionárias. Caso seja necessário tornar a série temporal estacionária, é preciso incluir o método de diferenciação ao modelo. Caso a série temporal apresente repetição do padrão de forma sequencial em determinado período, o modelo mais adequado seria incrementado com a sazonalidade (SARIMA).

Por fim, caso exista alguma variável externa ao modelo, que influencie na série temporal de modo a diminuir as métricas de erros, um modelo acrescido de uma variável exógena (SARIMAX) atenderia melhor (HO; XIE, 1998; ZHANG, 2003). Tal modelo

atende explicitamente a um conjunto de estruturas padrão em dados de séries temporais e fornece um método simples, mas poderoso, para realizar previsões de séries temporais (HO; XIE, 1998; ZHANG, 2003; MORETTIN; TOLOI, 2022).

A sigla SARIMAX é descritiva e captura os principais aspectos do próprio modelo, e a letra S representa a sazonalidade, isto é, essa técnica deve ser empregada caso haja repetição de padrões da série temporal. A sigla AR denota a autorregressão, ou seja, o modelo utiliza a relação dependente entre uma observação e um determinado número de observações anteriores. A letra I descreve o integrador, o uso de diferenciação entre as observações brutas, com o intuito de obter uma série temporal estacionária. Já a sigla MA demonstra as médias móveis, ou seja, refere-se a um modelo que utiliza a dependência entre uma certa observação e o erro residual de um modelo de médias móveis aplicado nas observações defasadas. Portanto, trata-se de uma média aplicada a uma determinada sequência numérica cronológica, em um intervalo de tempo móvel. A letra X exprime a variável exógena, isto é, refere-se à variável externa à sequência de valores da série temporal analisada.

Cada uma dessas siglas que compõem as técnicas empregadas no acrônimo ARIMA é categorizada de forma explícita no modelo como um parâmetro. Uma notação padrão também é empregada na classe de modelos ARIMA(p, d, q), nos quais os parâmetros são substituídos por valores inteiros para indicar explicitamente o modelo ARIMA aplicado.

Segundo Box *et al.* (2015) e Pankratz (2009), os parâmetros do modelo ARIMA são definidos da seguinte forma: a letra p remete-se ao parâmetro AR, indicando o número de observações atrasadas incluídas no modelo, também chamado de ordem de atraso. A letra d refere-se ao parâmetro I, apresentando, dessa forma, o número de vezes que as observações brutas são diferenciadas, também chamado de grau de diferenciação. Por fim, a letra q relaciona-se ao parâmetro MA, expressando o tamanho da janela de média móvel empregada, também denominada ordem da média móvel.

Os parâmetros do modelo ARIMA, que incluem a sazonalidade (SARIMA), são representados por (P, D, Q, S), sendo que são específicos à sazonalidade. As letras P, D, Q representam os parâmetros equivalentes a p, d, q, referentes à série temporal contida no período intrasazonal, e a letra S remete à frequência de observações intrasazonal. Portanto, o modelo completo pode ser representado na seguinte nomenclatura SARIMA (p, d, q)[P, D, Q, S].

Por meio da Equação (1), é possível observar a notação matemática do modelo ARIMAX, da classe Box-Jenkins. Como mencionado, o modelo ARIMAX pode conter todos os parâmetros do modelo ARIMA, com a inclusão de variável exógena (X) (ELSHENDY *et al.* 2018; HYNDMAN; ATHANASOPOULOS, 2018).

$$Y_{t+1} = \sum_{i=0}^{P} \boldsymbol{\beta}_i Y_{t-i} + \sum_{j=1}^{J} \Theta_j X_{t,j} + \sum_{k=1}^{K} \Phi_k Z_{t,k} + \boldsymbol{\varepsilon}_{t+1} \tag{1}$$

Os termos da Equação (1) foram contextualizados em relação ao escopo do presente

trabalho. Y_t significa o preço de fechamento do IBOV em determinado dia (t). Portanto, Y_{t+1} remete ao preço de fechamento do IBOV no próximo dia em relação ao dia t. Por consequência, Y_{t-1} significa o preço de fechamento do IBOV no dia anterior ao dia t. $X_{t,j}$ relaciona-se à média do valor resultante da análise de sentimentos relacionada às notícias publicadas no dia t. $Z_{t,k}$ refere-se a outros possíveis valores numéricos ligados ao IBOV no dia t; por exemplo, se fosse considerado também o valor de abertura do IBOV, isto é, em casos em que fosse aplicada uma análise multivariada. ε_{t+1} são os ruídos aleatórios, também conhecidos como *white noise*, que significa ruídos brancos, em português, ou seja, são os padrões aleatórios da série temporal. $\boldsymbol{\beta}_i$, Θ_j e Φ_k são os coeficientes de erro relacionados ao emprego das respectivas técnicas: autorregressão, integração e médias móveis.

2.6 Estado-da-arte

Segundo Silva (2021), as técnicas que obtiveram melhores resultados de previsão do fundo de investimento vinculado ao IBOV (BOVA11) foram as baseadas no modelo ARIMA, em especial, as implementações que combinavam a variável resultante da análise de sentimentos com a variável resultante da utilização de indicadores técnicos e o histórico de preços do BOVA11. Em seu trabalho, Silva (2021) desenvolveu um sistema automatizado de negociação de ações usando aprendizado profundo por reforço. O modelo preditivo era composto por dois módulos: um considerava apenas as variáveis do índice BOVA11 e indicadores técnicos, e, em contrapartida, o outro módulo apenas classificava o sentimento dos títulos das notícias vinculadas ao Ibovespa. Em seguida, as saídas desses dois módulos eram combinadas, de modo a alimentar as entradas dos modelos de compra e venda dos ativos. Os modelos de compra e venda que consideravam tanto os indicadores técnicos quanto os sentimentos das notícias obtiveram melhores resultados.

Por meio da Tabela 1, é possível observar os modelos que obtiveram melhores resultados do módulo sem o atributo de sentimentos.

Tabela 1 – Top 9 Melhores Resultados - Previsão BOVA11 - Algoritmos sem análise de sentimentos

Grupo do Modelo	Modelo	Melhores Parâmetros	Divisão Treino/Teste	IT	MAE	MSE
Econometria	SARIMAX	p:1,d:0,q:1,P:4,D:0,Q:4,S:3	Divisão Comum	**X**	0.489	0.657
Econometria	SARIMAX	p:1,d:0,q:1,P:4,D:0,Q:1,S:2	Bloco	**X**	0.490	0.657
Aprendizado de Máquina	SVR_Multi	C:10, Ep:0.001, K:linear	Bloco		0.524	0.735
Aprendizado de Máquina	SVR_Multi	C:10, Ep:0.001, K:linear	Divisão Comum		0.524	0.735
Econometria	SARIMAX	p:1,d:0,q:1,P:4,D:0,Q:4,S:3	Divisão Comum		0.525	0.693
Econometria	SARIMAX	p:2,d:0,q:2,P:3,D:1,Q:3,S:3	Bloco		0.876	1.438
Aprendizado de Máquina	SVR_Multi	C:10, Ep:0.01, K:linear	Bloco	**X**	0.604	1.149
Aprendizado de Máquina	SVR_Multi	C:10, Ep:0.01, K:linear	Divisão Comum	**X**	0.604	1.149
Aprendizado Profundo	LSTM_Multi	Ba:32, Nn:1000, Ph:3, E:30	Divisão Comum	**X**	1.069	1.771

Fonte: Adaptado de Silva (2021).

As colunas da Tabela 1 representam, respectivamente, a qual grupo o modelo

pertence, o nome do modelo, os melhores valores para os parâmetros de cada modelo, o tipo de técnica empregada para dividir o conjunto de dados, se foram utilizados indicadores técnicos, e os valores do erro médio absoluto (MAE) e erro quadrático médio (MSE). As siglas mencionadas na Tabela 1: IT, Ep, K, Ba, Nn, Ph e E significam, respectivamente, indicadores técnicos, taxa de aprendizagem (*epsilon*), *kernel*, *batch size*, número de neurônios, comprimento da janela de intervalo e número de épocas (*epochs*).

Para a divisão do conjunto de dados em treino e teste, é possível dividir a base de dados da forma comum, popularmente conhecida por *split*. Isto é, a base de dados é dividida em treino e teste apenas uma única vez. Em contrapartida, a divisão do conjunto de treino e teste em blocos ocorre mais de uma vez, ou seja, a série temporal original é fatiada em diversas séries, em intervalos menores, formando várias séries temporais menores, cada uma com seu respectivo conjunto de treino e validação.

Costa *et al.* (2018), além de considerar indicadores técnicos em seu trabalho, extraíram sentimentos de artigos de notícias e *tweets*. Foi identificada uma correlação entre as variações do preço das ações das companhias *Apple*, *JPMorgan Chase*, *Exxon Mobil* e *Boeing*.É importante mencionar que o foco do seu trabalho foi no retorno financeiro mediante as variações intradiárias de menor intervalo de tempo. Em seu estudo, os autores focaram nas variações de curto intervalo de tempo (0-5 min, 5-10 min, 10-30 min, 30-60 min e 2-6 h). De acordo com Costa *et al.* (2018), a modelagem ARIMA mostrou que apenas a análise dos preços defasados e das volatilidades defasadas contém pouco poder de previsão consistente sob a *Hipótese do Mercado Eficiente*. Os autores observaram que os modelos ARIMA que consideravam as variáveis externas ao mercado de ações (ARIMAX) eram mais capazes de captar as tendências reais dos preços das ações estudadas, entretanto, ainda longe do ideal.

O trabalho de Sudhakar e Naganjaneyulu (2020) teve foco no mercado de ações indiano. Os autores aplicaram o algoritmo LASSO para extrair as variáveis mais relevantes dos índices SENSEX, NIFTY 50, das ações ICS, HDFC, INFY e das notícias coletadas. Posteriormente, Sudhakar e Naganjaneyulu (2020) aplicaram o modelo ARIMAX, sendo que a variável exógena foi uma combinação de variáveis. Essa combinação foi realizada a partir das técnicas de processamento de linguagem natural e do resultado do algoritmo LASSO. De acordo com Sudhakar e Naganjaneyulu (2020), o modelo preditivo obteve melhores resultados após a adição dos sentimentos extraídos das notícias.

Em Elshendy *et al.* (2018), foram coletados dados de quatro plataformas de mídia diferentes (Twitter, Wikipedia, Google Trends e GDELT), com o objetivo de desenvolver um modelo preditivo do preço do petróleo bruto WTI, incrementado com variáveis externas ao mercado de ações. Segundo os autores, os modelos que consideravam apenas os dados de uma plataforma foram menos precisos. Elshendy *et al.* (2018) observaram o modelo ARIMAX que obteve melhores resultados considerava a combinação das análises dos dados do Twitter, Wikipedia e GDELT como variável exógena. Também

segundo os autores, os sentimentos provenientes dos *tweets* foram mais relevantes para o modelo obter melhores previsões, visto que os modelos que não consideravam os sentimentos de *tweets* apresentaram desempenho pior.

Bollen, Mao e Zeng (2011) utilizou a análise de sentimentos em *tweets* para incrementar o modelo preditivo do índice *Dow Jones Industrial Average* (DJIA) (em português, Média Industrial Dow Jones) da Bolsa de Valores dos Estados Unidos. Esse índice é popularmente conhecido como índice Dow Jones. Segundo Bollen, Mao e Zeng (2011), a adição dessa variável externa ao mercado financeiro aperfeiçoou os resultados do modelo, reduzindo o MAPE (erro percentual absoluto médio) em mais de 6%.

Conforme Pagolu *et al.* (2016), foi identificada uma correlação entre as tendências de altas e quedas no preço de ação da Microsoft na Bolsa dos Estados Unidos (*Dow Jones*) com as tendências dos sentimentos e opiniões extraídos dos *tweets* coletados.

De acordo com Machado e Pereira (2018), a utilização do algoritmo Máquina Restrita de Boltzmann (RBM), combinado com a análise de sentimentos de notícias e *tweets*, coletados na plataforma *Bloomberg*, gerou percentuais de retornos financeiros líquidos aplicáveis como formas de investimentos em relação à utilização dos principais indicadores técnicos aplicados do mercado financeiro.

Segundo Silva (2021), Carosia, Coelho e Silva (2020) e Machado e Pereira (2018), os modelos baseados em aprendizagem profunda, sendo os mais utilizados rede neural Perceptron Multicamadas (MLP), RBM, Redes Neurais Recorrentes (RNN), rede neural Memória de Longo e Curto Prazo (LSTM) e alguns trabalhos baseados em aprendizagem supervisionada, no caso do algoritmo de Máquinas de Vetores de Suporte (SVM), tiveram também bons resultados. Assim, no presente momento, até onde foi pesquisado, não há um modelo que obteve desempenho superior, de forma unânime, em todos os trabalhos correlatos, visto que a literatura ainda carece de trabalhos com metodologias similares.

Em virtude de o presente estudo utilizar os modelos da classe Box-Jenkin e o IBOV, os principais trabalhos tomados como base para o desenvolvimento deste foram Elshendy *et al.* (2018), Sudhakar e Naganjaneyulu (2020), Sharma *et al.* (2020), Silva (2021), Medeiros e Borges (2019), Machado e Pereira (2018), Araújo e Marinho (2018), Monteiro e Ferreira (2021) e Junior, Salomon, Oliveira Pamplona *et al.* (2014). No presente trabalho, foi adotado o modelo econométrico ARIMAX, levando em consideração trabalhos correlatos que apresentaram resultados promissores Silva (2021), Sudhakar e Naganjaneyulu (2020) e Elshendy *et al.* (2018).

Embora o desempenho dos modelos de aprendizagem profunda, em geral, apresentem bons resultados, existem modelos mais simples, que atingem resultados superiores em um menor tempo de execução. Por exemplo, os resultados atingidos pelos algoritmos baseados no modelo ARIMA para previsão diária de índices do mercado financeiro superam os resultados de modelos mais complexos, como apresentado por Silva (2021) e Sudhakar e Naganjaneyulu (2020).

Com o monitoramento e armazenamento do histórico de sentimentos, as técnicas de processamento permitem realizar previsões sobre os futuros preços das ações, incrementado com a previsão da variável exógena de sentimentos, sem que a ordem dos dados seja comprometida. Desse modo, é possível a obtenção de um modelo capaz de realizar previsões com menores valores de métricas de erro, isto é, modelos com melhores resultados de previsões.

3 MATERIAIS E MÉTODOS

No capítulo atual, são abordados os aspectos metodológicos utilizados para o desenvolvimento do presente trabalho. A Seção 3.1 apresenta as classificações da pesquisa. Na Seção 3.2, são explicadas as ferramentas e tecnologias que permitiram o desenvolvimento deste trabalho. Em seguida, a Seção 3.3 aborda os métodos empregados no desenvolvimento do presente estudo.

3.1 Classificações do trabalho

O trabalho desenvolvido pode ser classificado, quanto à abordagem, como uma pesquisa quantitativa, que, de acordo com Gerhardt e Silveira (2009), utiliza-se de procedimentos estruturados para coletar dados e analisá-los. Para a análise desses dados, pode-se recorrer a modelos estatísticos descrevendo o motivo da ocorrência de determinado fenômeno e a relação de sua influência através das variáveis.

Em relação à natureza, este trabalho pode ser classificado como uma pesquisa aplicada, por colaborar para uma possível solução de um problema específico e de interesse socioeconômico. Ademais, o trabalho objetiva gerar conhecimentos para aplicação prática.

Quanto aos objetivos, este trabalho pode ser definido como uma pesquisa exploratória, ao visar construir hipóteses, além de utilização de levantamento bibliográfico, de análise e compreensão de exemplos que podem auxiliar na resolução dos problemas encontrados.

Segundo Gerhardt e Silveira (2009), a pesquisa exploratória traz uma maior familiaridade aos pesquisadores sobre o problema pesquisado, e tal fato acontece devido à imersão do pesquisador na área de estudo. Este trabalho também possui essa característica provinda da exploração de técnicas de Processamento de Linguagem Natural, em especial, a Análise de Sentimentos.

No aspecto procedimental, este estudo pode ser definido como uma pesquisa experimental, pois visa desenvolver uma ferramenta para previsão do IBOV considerando os impactos de notícias. A pesquisa foi realizada com base no modelo de previsão criado e testes analíticos conduzidos (GERHARDT; SILVEIRA, 2009).

Com base na abordagem de Wazlawick (2009), a respeito da metodologia de pesquisa para a Ciência da Computação, este trabalho pode ser classificado como a apresentação de um produto. Essa classificação ocorre porque o presente trabalho utiliza a combinação de técnicas de estatística e análise de sentimentos, com o intuito de obter um produto que contribua com a sociedade, a fim de evitar prejuízos no mercado financeiro.

3.2 Materiais e tecnologias

As tecnologias empregadas para o desenvolvimento deste trabalho, agrupadas por cada tarefa, podem ser observadas na Tabela 2. Por meio da Tabela 3, é possível observar a configuração da máquina utilizada tanto para o desenvolvimento quanto para a realização dos testes.

Tabela 3 – Descrição de recursos da máquina de trabalho

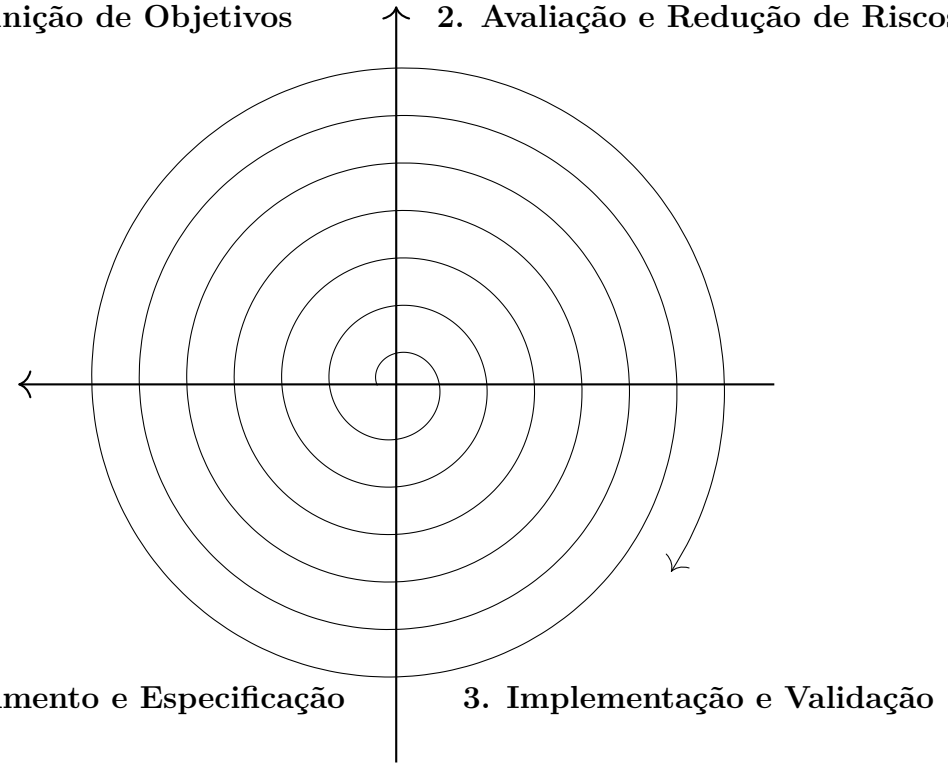
Recurso	**Descrição**
Sistema operacional	Linux Mint 21.1
Processador	Intel i7 7700K 4,2 GHz 8Mb de cache
Placa de vídeo externa	NVIDIA RTX 2060 OC Super Galax 8GB
Memória primária	32 GB HyperX DDR4 2666MHze
Memória secundária	SSD XPG 1TB, leitura de 7400MB/s e gravação de 5500MB/s
Placa-mãe	Gigabyte GA-B250M-GAMING 3

Fonte: Elaborado pelo Autor (2023).

3.3 Métodos e procedimentos

Em relação ao desenvolvimento de software, foi aplicado o modelo espiral. Por se tratar de um modelo baseado na prototipação cíclica e evolutiva, demonstra-se mais adaptável às mudanças, em comparação com outros modelos (BOEHM, 1988), como pode ser observado na Figura 2.

Figura 2 – Modelo Espiral de desenvolvimento de *software*



Fonte: Adaptado de BOEHM, 1988.

A etapa inicial do presente trabalho envolveu uma extensa revisão bibliográfica relacionada aos métodos de predição das influências de notícias sobre o mercado de ações, com foco no mercado brasileiro. Foi pertinente, nessa análise, considerar a disponibilidade dos dados a serem utilizados, a possibilidade de reprodução e acurácia dos métodos. Concomitantemente à etapa de análise dos métodos de previsão, foram utilizadas ferramentas de recuperação de informação para desenvolver e atualizar as bases de dados, sendo que a que contém o histórico do IBOV foi extraída por meio da biblioteca Y!Finance.

As notícias vinculadas ao IBOV foram extraídas do site Money Times[1], através de um *script* escrito em Python utilizando o *framework Selenium WebDriver*. As principais justificativas para escolha da utilização do site Money Times como fonte de origem de dados foram em virtude da possibilidade de extrair notícias escritas por diversos autores de instituições distintas e da viabilidade de filtragem de notícias relacionadas a determinadas palavras-chave. Para a filtragem das notícias nesse site, foi selecionado, no menu de notícias, o elemento Ibovespa. Outros sites também foram analisados, como Investing Brasil[2]; porém, além de possuírem um menor volume de notícias vinculadas ao IBOV, as notícias disponibilizadas eram apenas aquelas mais recentes, inviabilizando sua utilização em um estudo de maior período, conforme conduzido neste trabalho.

Com o propósito de realizar uma análise mais rica, foram considerados também

[1] https://www.moneytimes.com.br/tag/ibovespa/
[2] https://br.investing.com/indices/bovespa-news

os *tweets* públicos em língua portuguesa que continham as palavras-chave IBOV ou Ibovespa, visto que, em trabalhos correlatos, alguns autores observaram que os modelos preditivos que consideravam os *tweets* obtinham melhores resultados do que os que consideravam notícias (CAROSIA; COELHO; SILVA, 2019). A maioria dos trabalhos correlatos que consideram notícias analisam apenas o título destas (NEMES; KISS, 2021; LIU, Y. *et al.* 2020; JARIWALA; AGARWAL; JADHAV, 2020; ONCHAROEN; VATEEKUL, 2018; NIZER; NIEVOLA, 2012). Por esse motivo, a fim de contribuir com a literatura, a base de notícias foi dividida em duas bases de dados: uma delas contendo apenas o título e outra com todo o corpo das notícias. Todas as bases de dados foram armazenadas em arquivos de texto.

Após a etapa anterior, as bases de dados textuais foram submetidas ao processo de análise de sentimentos com o intuito de extrair o sentimento predominante em cada texto. Para isso, foi utilizado o analisador Léxico de Inferência Aplicada (LeIA), que é uma ferramenta baseada no *vaderSentiment*[3]. Por esse motivo, não foi necessário empregar nenhuma técnica de pré-processamento textual, visto que o *vaderSentiment* já contempla *scripts* para tratamento dos textos (GILBERT, 2014). No presente trabalho, a única métrica resultante da aplicação do *LeIA* considerada foi o *compound*, que é o valor tratado resultante da somatória dos valores predominantes obtidos com a análise de sentimentos.

Para preenchimento dos valores nas bases de dados referentes aos dias em que não existiam notícias/cotações, foi utilizada a técnica de interpolação do método *time*, disponível na biblioteca *pandas*. Após a aplicação da análise de sentimentos, foi preciso correlacioná-la com o valor de fechamento do IBOV do mesmo dia. Em seguida, foi desenvolvido um método preditivo, com base nos modelos da classe ARIMA. Como a série temporal do IBOV não é estacionária, foi necessário aplicar uma diferenciação para obtenção de série estacionária, em virtude de ser um pré-requisito para aplicação dos métodos econométricos da classe Box-Jenkins. Após a execução dessa análise estatística, foi implementado um algoritmo de *grid search* para obtenção dos melhores parâmetros, considerando os menores valores das métricas de erro: Erro Quadrático Médio (MSE), Erro Médio Absoluto (MAE) e Raiz do Erro Quadrático Médio (RMSE). Também foi utilizada a biblioteca *pmdarima* para executar testes de sazonalidade e hiperparametrização.

[3] https://vadersentiment.readthedocs.io/en/latest/

Tabela 2 – Descrição de tarefas e tecnologias empregadas

Tarefas	**Descrição**
Desenvolvimento do trabalho	Foi empregada a linguagem de programação *Python*, versão 3.10.6, disponível em https://www.python.org/
Extração de notícias	Utilizou-se o *framework Selenium WebDriver*, versão 4.7.2, disponível em: https://www.selenium.dev/documentation/webdriver/
Extração de *tweets*	Foi aplicado o *scraper* de redes sociais *snscrape*, versão 0.4.3.20220106, disponível em https://github.com/JustAnotherArchivist/snscrape
Extração do histórico do IBOV	Foi utilizada a ferramenta *Y!Finance*, versão 0.2.9, disponível em https://github.com/ranaroussi/yfinance
Manipulação das bases de dados	Empregou-se a biblioteca *pandas*, versão 1.5.2, disponível em https://pandas.pydata.org/
Análise de sentimentos	Recorreu-se ao analisador léxico *LeIA*, versão 0.0.1, disponível em https://github.com/rafjaa/LeIA
Testes estatísticos e aplicação do modelo preditivo	Foi empregada a biblioteca *statsmodels*, versão 0.13.5, disponível em https://www.statsmodels.org
Otimizações de parâmetros e testes de sazonalidades	Empregou-se a biblioteca *pmdarima*, versão 2.0.2, disponível em https://alkaline-ml.com/pmdarima/
Avaliação dos modelos	Aplicou-se a biblioteca *scikit-learn*, versão 1.1.3, disponível em https://scikit-learn.org

Fonte: Elaborado pelo Autor (2023).

4 DESENVOLVIMENTO

Este capítulo aborda o desenvolvimento do presente trabalho. Na Seção 4.1, é realizada uma análise a respeito dos principais métodos preditivos. Posteriormente, a Seção 4.2 apresenta o desenvolvimento das bases de dados de notícias. Em seguida, a Seção 4.3 apresenta as manipulações e tratamentos nas bases de dados. A Seção 4.4 apresenta a aplicação da técnica de Análise de Sentimentos. Aborda-se o desenvolvimento dos modelos preditivos na Seção 4.5,e a Seção 4.6 expõe e discute os experimentos conduzidos. Por fim, a Seção **??** descreve os repositórios deste trabalho.

4.1 Análise dos métodos de predição

A primeira etapa do desenvolvimento deste trabalho envolveu a análise dos métodos de predição relacionados ao mercado financeiro, com foco nos trabalhos referentes ao mercado brasileiro e também naqueles que consideravam eventos externos ao mercado. O resultado dessa análise definiu a abordagem do trabalho, utilizando técnicas econométricas da classe Box-Jenkins combinadas com a Análise de Sentimentos (SILVA, 2021; SUDHAKAR; NAGANJANEYULU, 2020; ELSHENDY *et al.* 2018).

4.2 Construção das bases de dados

Durante a etapa de desenvolvimento das bases de dados de notícias, foram extraídas do site Money Times as notícias vinculadas ao IBOV, através de *script* escrito em Python utilizando o *framework* Selenium WebDriver, para emular o navegador Firefox de forma a realizar as interações no *site* Money Times automatizadamente e recuperar as informações relevantes (títulos das notícias, textos, nomes dos autores, data de publicação das notícias *etc*). As notícias extraídas foram armazenadas localmente em arquivos de texto.

A fim de conduzir um estudo sobre o impacto da análise de sentimentos considerando apenas o título ou o corpo da notícia, foram desenvolvidas duas bases de dados: uma considerando o título da notícia, e a outra, o corpo. Essa análise é relevante, visto que a maioria dos trabalhos correlatos utilizam apenas o título das notícias (NEMES; KISS, 2021; LIU, Y. *et al.* 2020; JARIWALA; AGARWAL; JADHAV, 2020; ONCHAROEN; VATEEKUL, 2018; NIZER; NIEVOLA, 2012). No total, foram extraídas 3.742 notícias referentes ao período de 11 de fevereiro de 2016 até 03 janeiro de 2023. Em média, os títulos contêm 12 palavras, e o corpo das notícias, 436.

Para o desenvolvimento da base de dados de *tweets*, foram extraídos somente os *tweets* públicos em língua portuguesa que continham as palavras-chave IBOV ou IBovespa. No total, foram extraídos 914.277 *tweets* referentes ao período de 09 de novembro de 2007 a 03 de janeiro de 2023. Para a extração dos *tweets*, foi aplicado o *scraper* de redes sociais

snscrape, visto que essa ferramenta não necessita que uma conta no Twitter seja criada para realizar a extração de *tweets* públicos.

Para o desenvolvimento da base de dados que contém o histórico do IBOV, foi aplicada a biblioteca Y!Finance, utilizando o código identificador "^BVSP". Esse código identificador é denominado *ticker*. Esses dados foram manipulados e armazenados na estrutura de dados *dataframe*, disponibilizada através da biblioteca pandas. Posteriormente, o histórico do IBOV foi armazenado localmente em arquivo de texto. Para construção da base de dados, foi considerado apenas o preço de fechamento. O período da base de dados obtida é referente a 27 de abril de 1993 a 02 de fevereiro de 2023, sendo que, ao todo, foram armazenadas 7.369 cotações.

Por meio da tabela 4, é possível observar algumas informações sobre os *datasets* com o período dos dados presentes nas bases de dados, quantidade de registros e alocação de memória para armazenamento delas.

Tabela 4 – Especificações das bases de dados

Base de dados	**Período**	**Quantidade de registros**	**Tamanho**
Histórico do IBOV	27 de abr. de 1993 a 02 de fev. 2023	7.369	385,5 KB
Notícias	11 de fev. de 2016 a 03 jan. de 2023	3.742	11,3 MB
Tweets	09 de nov. de 2007 a 03 de jan. de 2023	914.277	188,1 MB

Fonte: Elaborado pelo Autor (2023).

4.3 Manipulações das bases de dados

Devido à instabilidade de conexão com o servidor do site Money Times, as bases de notícias foram gravadas em diversos arquivos de texto. Essas bases de dados foram concatenadas em um único arquivo, e os registros duplicados foram excluídos. As bases de dados foram tratadas de modo a se obter apenas as colunas de interesse. Essas colunas foram, respectivamente, o preço de fechamento do IBOV, a data do fechamento, a data de publicação das notícias e *tweets*, o texto dos *tweets*, o corpo e título das notícias.

Foram realizados alguns testes a fim de descobrir qual seria o corte temporal mais adequado que englobaria as três bases de dados em um mesmo período, com quantidades de registros relevantes sem que existissem grandes ausências de dados entre semanas. Por meio dessas análises, foi possível determinar o recorte temporal de 02 de janeiro de 2018 a 31 de dezembro de 2022. O método aplicado para preenchimento das lacunas na base do histórico do IBOV foi o método de interpolação *time* disponível na biblioteca pandas.

4.4 Aplicação da Análise de Sentimentos

Posteriormente às etapas de pré-processamento das bases de dados, aplicou-se a técnica de análise de sentimentos com a finalidade de extrair o sentimento predominante

em cada notícia, *tweet* e título de notícia. Para isso, foi utilizado o analisador Léxico de Inferência Aplicada (LeIA), que, por sua vez, não necessita empregar nenhuma técnica de pré-processamento textual, visto que o LeIA já contempla rotinas para tratamento dos textos, embora o LeIA classifique os textos com pontuações para positivas, negativas, neutras e *compound*. Neste trabalho, considerou-se apenas a pontuação *compound*, em português, composta, visto que essa pontuação considera todas as outras, ou seja, são consideradas as pontuações de classificação negativa, positiva e neutra, sendo que, no cálculo de *compound*, as pontuações negativas e positivas são amplificadas (GILBERT, 2014).

Em seguida, para preenchimento do *compound* referente aos dias que não existem notícias, utilizou-se a técnica de interpolação método do *time*, disponível na biblioteca *pandas*. Após o preenchimento dessas lacunas, empregou-se uma média simples para agregar os valores resultantes da análise de sentimentos de forma diária, a fim de obter um único valor *compound* por dia, para tratar os casos em que há mais de uma notícia ou *tweet* em um determinado dia.

4.5 Desenvolvimento dos modelos preditivos

Antes de iniciar o desenvolvimento do modelo preditivo, foi importante realizar análises estatísticas, visando abstrair as características da série temporal do valor de fechamento do IBOV. No Apêndice A, é possível observar a decomposição dessa série temporal utilizando o módulo *seasonal_decompose* nativo da biblioteca *statsmodels*. Como se pode observar, a série temporal do IBOV possui tendências de crescimento e queda não homogêneas, compreendidas no período de 2018 a 2022. Para aplicação dos modelos econométricos da classe Box-Jenkins, faz-se necessário que a série temporal a ser analisada seja estacionária. Deste modo, realizou-se o teste de hipótese aplicando o teste Dickey-Fuller aumentado, também por meio da biblioteca *statsmodels*, com o intuito de verificar, em conjunto com a análise de tendências da série temporal, se, de fato, a série é estacionária.

Portanto, para a hipótese nula ser rejeitada, o valor-p deve ser abaixo de 0.05. Por meio da aplicação do teste Dickey-Fuller aumentado, em conjunto com as análises gráficas de autocorrelação e autocorrelação parcial, foi possível concluir que a série do IBOV não é estacionária, pois a hipótese nula não foi rejeitada. Portanto, após a primeira diferenciação, os testes de estacionariedade foram conduzidos novamente. Verificou-se que a série se tornou estacionária após a primeira diferenciação, porque a hipótese nula foi rejeitada, e, além disso, os gráficos de autocorrelação e autocorrelação parcial da série temporal do IBOV, após a primeira diferenciação, tendem a zero. Logo, conclui-se que a série temporal do IBOV, após a primeira diferenciação, tornou-se estacionária e está pronta para aplicação dos métodos da classe Box-Jenkins.

Para o desenvolvimento dos métodos preditivos, fez-se necessária a importação de algumas bibliotecas específicas de *Python*: *math*, *pandas*, *scikit-learn*, *statsmodels* e

matplotlib. A *math* foi empregada para realização de operações matemáticas; *pandas*, para a manipulação dos dados na forma de *dataframes*; e *scikit-learn*, para avaliação dos modelos através das métricas MSE, MAE, RMSE e MAPE. Utilizou-se a biblioteca *statsmodels* para aplicação dos modelos da classe Box-Jenkins, realização dos testes e análises estatísticas descritas anteriormente. Por fim, empregou-se a biblioteca *matplotlib* para a plotagem dos resultados em forma gráfica.

A fim de realizar a hiperparametrização para obtenção dos melhores modelos, implementou-se o algoritmo *grid search*, por meio da biblioteca *itertools*, para emprego do método *product*, com o propósito de realizar as combinações entre os parâmetros p, d e q. Em conjunto, utilizou-se a biblioteca *pandas*, para manipulação e armazenamento das informações na forma de *dataframes*, aplicou-se também recursos nativos da linguagem *Python* como laços de repetições e a estrutura de dados *list*. Durante a aplicação da hiperparametrização, foram utilizados apenas 80% dos dados da série temporal, isto é, empregou-se o conjunto de treino.

O tempo de execução do *grid search* é extenso, visto que, ao total, seriam variados 7 parâmetros, e cada parâmetro pode ser variado entre diversos intervalos, fazendo com que a execução do algoritmo, dependendo do tamanho dos intervalos, possa demorar longos prazos, como dias, semanas e meses. A princípio, adotou-se a variação apenas dos parâmetros p, d e q. Os parâmetros p e q foram variados entre os valores inteiros 0 e 6. Apesar de os testes de transformação da série temporal em estacionária terem sido suficientes para determinar que uma diferenciação era suficiente para transformar a série em estacionária, a título experimental, o parâmetro d foi variado com os número um e dois, pois, em trabalhos correlatos, como de Silva (2021), o parâmetro d também foi variado. Entretanto, conforme os testes conduzidos de estacionariedade, em conjunto com hiperparametrização dos parâmetros, através do *grid search*, o melhor parâmetro obtido para d é igual a um, visto que não é adequado realizar diferenciações além da necessidade, já que a série temporal pode perder suas características e padrões, resultando em uma previsão de baixa confiabilidade.

O algoritmo *grid search* foi empregado com o objetivo de selecionar e obter os melhores parâmetros para os algoritmos da classe Box-Jenkins. Foram considerados os menores valores das métricas de erro Erro Quadrático Médio (MSE), Erro Médio Absoluto (MAE) e Raiz do Erro Quadrático Médio (RMSE). Ou seja, foram realizadas três aplicações desse algoritmo para cada base de dados. Após a obtenção dos melhores parâmetros p, d e q, para cada caso de base de dados diferente, para verificar se os modelos também deveriam considerar a sazonalidade, foram selecionados os três modelos p,d e q com menor MSE. Estes parâmetros p, d e q foram fixados, e os parâmetros P, D, Q e S foram variados com outras aplicações do *grid search*, considerando o valor máximo de P, D e Q como seis, e valor-limite de S como quatro. A tabela com os resultados dos testes do *grid search* pode ser observada no Apêndice B. Esses mesmos valores-limite foram empregados nos

testes conduzidos com a seleção de parâmetros disponível pela biblioteca *pmdarima*. As plotagens e métricas de erros dos modelos desenvolvidos com os parâmetros selecionados pelo módulo autoarima, nativo da biblioteca *pmdarima*, podem ser observadas no Apêndice C. Apesar de todas essas tentativas, os modelos obtidos com menores valores para as métricas de erro não consideravam a sazonalidade, ou seja, em geral, os melhores modelos resultaram nos valores zerados para os parâmetros P, D, Q e S.

Por meio da Tabela 5, é possível observar os melhores modelos obtidos com a aplicação do algoritmo *Grid Search*, em conjunto com a descrição de seus respectivos parâmetros e erros. Para mais informações, consultar o Apêndice B, que contempla os resultados da aplicação do *Grid Search*.

Tabela 5 – Descrição de melhores modelos obtidos e seus respectivos erros referentes ao período de 2018 a 2022.

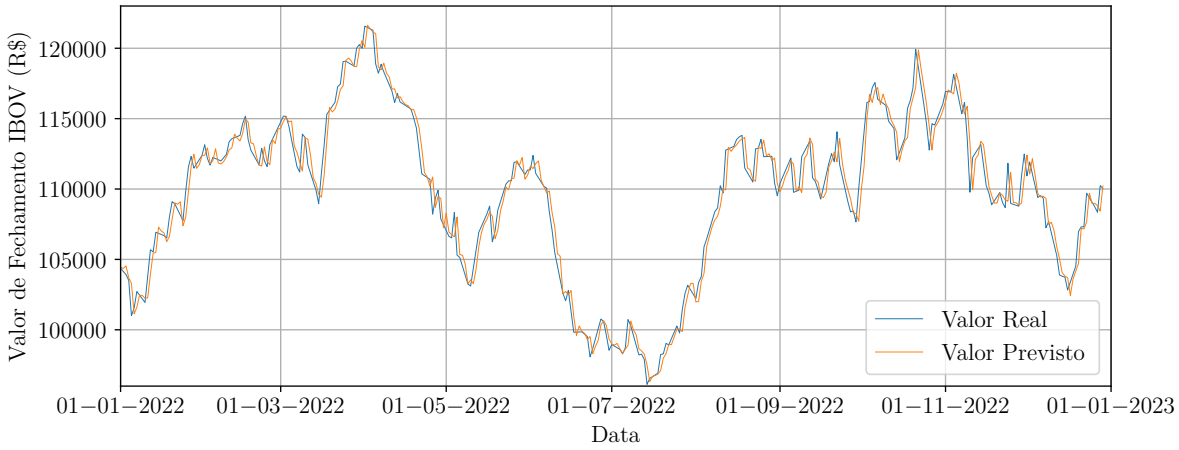
Variável Exógena	**MAE**	**MSE**	**RMSE**	**pdq**	**PDQS**
Sentimentos de *tweets*	751,2256	1.240.347,5251	1.113,7089	(4, 1, 4)	(0, 0, 0, 0)
Sentimentos de títulos das notícias e *tweets*	759,9004	1.296.753,7399	1.138,7510	(3, 1, 5)	(0, 0, 0, 0)
Sentimentos de notícias e *tweets*	765,3313[1]	1.322.861,8229*	1.150,1573*	(2, 1, 3)	(2, 0, 0, 3)[1], (5, 0, 0, 3)*
Sentimentos de títulos das notícias	773,3067	1.336.791,4418	1.156,1970	(5, 1, 5)	(0, 0, 0, 0)
Sentimentos de notícias	774,9664	1.335.876,6019	1.155,8013	(5, 1, 4)	(0, 0, 0, 0)
Sem variável exógena	781,1319	1.351.618,3070	1.162,5912	(5, 1, 4)	(0, 0, 0, 0)

Fonte: Elaborado pelo Autor (2023).

4.6 Experimentos

Esta seção apresenta os melhores resultados. Para mais informações sobre os experimentos, consultar o Apêndice D. Esse Apêndice contempla outras variações, além de apresentar as plotagens gráficas de cada previsão. No decorrer do desenvolvimento dos experimentos, também se considerou a análise no período pós-pandemia, referente a 01 de janeiro de 2021 até 31 de dezembro de 2022. A realização dos experimentos seguiu a metodologia empregada nos trabalhos de Silva (2021), Sudhakar e Naganjaneyulu (2020) e Elshendy *et al.* (2018).

Figura 3 – Previsão diária do IBOV com sentimentos de *tweets* referente ao período de 2018 a 2022.



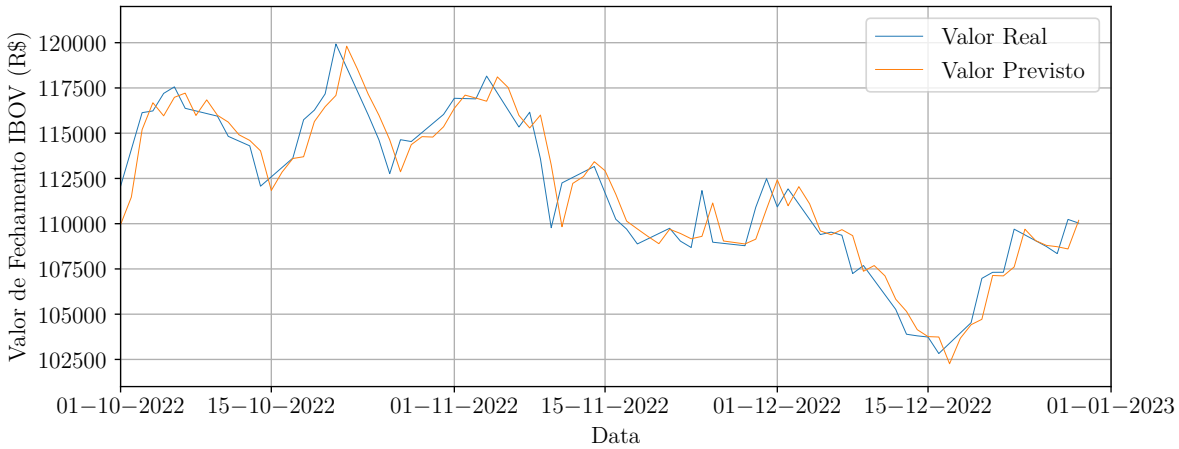
Fonte: Elaborado pelo Autor (2023).

Tabela 6 – Descrição de modelos e respectivos erros de previsão diária do IBOV referente ao período de 2018 a 2022.

Variável Exógena	**MAPE**	**MAE**	**MSE**	**RMSE**	**pdq**	**PDQS**
Sentimentos de *tweets*	0,0066	736,1539	997.422,6312	998,7104	(4, 1, 4)	(0, 0, 0, 0)
Sentimentos de títulos de notícias e *tweets*	0,0068	751,7421	1.046.133,0050	1.022,8064	(3, 1, 5)	(0, 0, 0, 0)
Sentimentos de notícias e *tweets*	0,0069	760,4713	1.066.094,4456	1.032,5184	(2, 1, 3)	(5, 0, 0, 3)
Sentimentos de notícias	0,0069	765,3690	1.084.079,0252	1.041,1911	(5, 1, 4)	(0, 0, 0, 0)
Sentimentos de títulos das notícias	0,0069	767,3590	1.095.810,8827	1.046,8098	(5, 1, 5)	(0, 0, 0, 0)
Sem variável exógena	0,0070	777,1252	1.125.359,7878	1.060,8297	(5, 1, 4)	(0, 0, 0, 0)

Fonte: Elaborado pelo Autor (2023).

Figura 4 – Previsão diária do IBOV com sentimentos de *tweets* referente ao período 2021 a 2022.



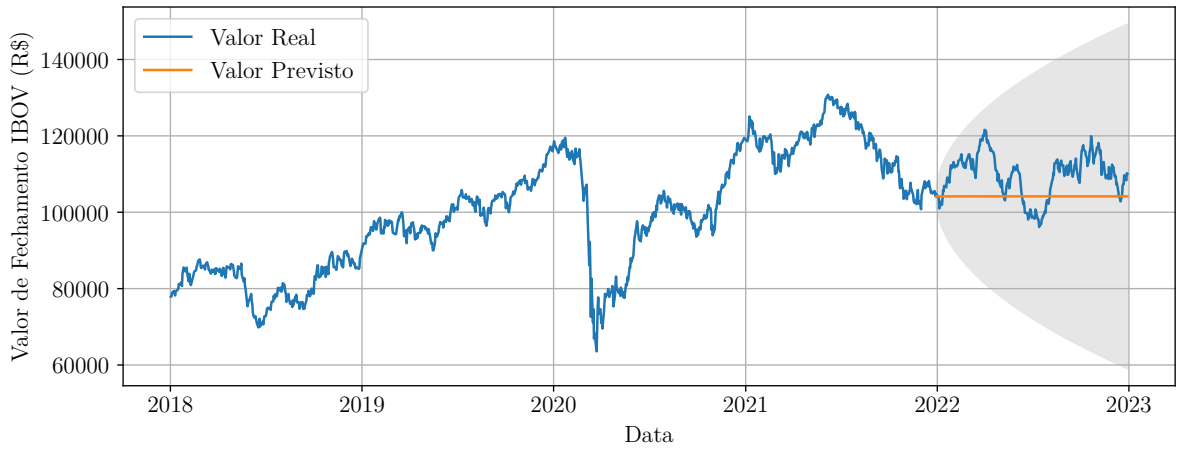
Fonte: Elaborado pelo Autor (2023).

Tabela 7 – Descrição de modelos e respectivos erros de previsão diária do IBOV referente ao período 2021 a 2022

Variável Exógena	MAPE	MAE	MSE	RMSE	PDQ	PDQS
Sentimentos de *tweets*	0,0080	899,5446	1.419.333,1539	1.191,3576	(4, 1, 4)	(0, 0, 0, 0)
Sentimentos de títulos das notícias	0,0083	937,7044	1.536.731,5183	1.239,6497	(5, 1, 5)	(0, 0, 0, 0)
Sentimentos de notícias	0,0084	949,6574	1.521.684,3713	1.233,5657	(5, 1, 4)	(0, 0, 0, 0)
Sem variável exógena	0,0085	958,1709	1.609.636,3738	1.268,7144	(5, 1, 4)	(0, 0, 0, 0)

Fonte: Elaborado pelo Autor (2023).

Figura 5 – Previsão de 365 dias do IBOV referente ao período de 2018 a 2022.



Fonte: Elaborado pelo Autor (2023).

Tabela 8 – Descrição de modelos e resultados de previsões longas do IBOV referente ao período 2018 a 2022

Período Previsto	Variável Exógena	MAPE	MAE	MSE	RMSE	PDQ	PDQS
4 dias	Sentimentos de *tweets*	0,0067	744,1509	646.745,3560	804,2047	(4, 1, 4)	(0, 0, 0, 0)
7 dias	Sentimentos de *tweets*	0,0175	1.919,5656	4.037.199,3982	2.009,2783	(4, 1, 4)	(0, 0, 0, 0)
15 dias	Sentimentos de *tweets*	0,0303	3.300,1361	15.889.195,5894	3.986,1253	(4, 1, 4)	(0, 0, 0, 0)
30 dias	Sentimentos de *tweets*	0,0318	3.370,7848	17.181.567,4878	4.145,0654	(4, 1, 4)	(0, 0, 0, 0)
90 dias	Sentimentos de *tweets*	0,0321	3.623,7733	18.844.550,6158	4.341,0310	(4, 1, 4)	(0, 0, 0, 0)
180 dias	Sentimentos de *tweets*	0,0951	10.651,0124	140.958.132,2214	11.872,5790	(4, 1, 4)	(0, 0, 0, 0)
365 dias	Sentimentos de *tweets*	0,0616	6.906,0238	62.595.787,2155	7.911,7499	(4, 1, 4)	(0, 0, 0, 0)

Fonte: Elaborado pelo Autor (2023).

Por meio das informações apresentadas neste capítulo, é possível observar que, em todos os casos, as variáveis externas à série histórica do IBOV aprimoraram os modelos, diminuindo os valores das métricas de erros MSE, RMSE MAE e MAPE das previsões. Logo, é possível concluir que os modelos que incluíram sentimentos alcançaram melhores previsões. Na Tabela 5, é possível observar que os modelos que incluíram apenas os sentimentos de *tweets* obtiveram melhores resultados do que os que consideraram os sentimentos de notícias. Com as informações apresentadas na Tabela 8, é possível notar que o modelo incrementado com sentimentos de *tweets* foi capaz de prever o valor de

fechamento do IBOV em até 4 dias com 0,0067 de MAPE. por meio da Figura 6, é possível observar que, em termos de MAPE, as previsões não obtiveram resultados distantes. Por meio das informações apresentadas nas Tabelas 6 e 7, pode-se concluir que os modelos que levaram em consideração um maior período da série histórica do IBOV obtiveram melhores resultados. Outro ponto a ressaltar, com relação à ausência dos parâmetros de sazonalidade é que, como a série histórica do IBOV não evidencia relações de sazonalidade, as previsões longas não obtiveram resultados satisfatórios, como se pode verificar nas informações abordadas nos Apêndices A, D. Portanto, é possível concluir que os modelos da classe Box-Jenkins não são os mais adequados para efetuar previsões longas da série temporal do preço do IBOV.

5 CONCLUSÃO

As mídias sociais e sites de notícias são as principais fontes de informação dos brasileiros (DATASENADO, 2019; IPSOS, 2018). Devido ao grande volume de informações veiculadas na internet, é uma tarefa inviável para os investidores brasileiros acompanharem diariamente a maioria dessas informações relacionadas ao mercado financeiro. Além disso, muitos investidores pessoas físicas não vivem de investimentos e possuem ocupação profissional (TEIXEIRA; LOPES; MEURER, 2021), o que dificultaria ainda mais o acompanhamento diário dessas informações. Por esse motivo, o desenvolvimento de ferramentas capazes de realizar análises como a apresentada neste trabalho é necessário para facilitar a atuação de investidores, além de permitir a identificação de janelas de oportunidades e auxiliá-los a evitar prejuízos, caso os modelos desenvolvidos identifiquem quedas em observações futuras do valor do IBOV. Embora essas ferramentas não sejam capazes de varrer todas as páginas web, as informações das páginas mais relevantes podem ser obtidas, processadas e incluídas em modelos preditivos.

Conforme evidenciado no decorrer deste trabalho, a inclusão de sentimentos de notícias e *tweets*, de fato, aprimorou as previsões no escopo dos testes realizados. Também foi possível observar que a análise de sentimentos nos títulos de notícias não obteve resultados distantes da análise de sentimentos que incluiu todo o corpo das notícias. Em alguns casos, as previsões que consideraram a análise de sentimentos em todo o corpo das notícias obtiveram resultados até piores. Outro ponto a ressaltar são as previsões com os sentimentos de *tweets* que obtiveram melhores resultados que as previsões que consideraram os sentimentos de notícias. Identificou-se que os valores obtidos com a análise de sentimentos dos *tweets* possuíam menores valores em relação aos valores absolutos alcançados com a análise de sentimentos das notícias. Portanto, supõe-se que essa seja a causa para as previsões com sentimento dos *tweets* terem obtido melhores resultados em relação às previsões que consideraram os sentimentos das notícias. Destaca-se, também, que as previsões das séries temporais que consideravam um período histórico mais longo atingiram melhores resultados. O modelo que considerou os sentimentos de *tweets* foi capaz de prever em até 4 dias o valor de fechamento do IBOV com 0,0067. Outro produto deste trabalho são as bases de dados desenvolvidas e os *scripts* para mineração dos dados, visto que, com pequenos ajustes, esses *scripts* podem ser aplicados em outros trabalhos, até mesmo em trabalhos de escopos diferentes.

5.1 Trabalhos futuros

Com a conclusão deste estudo, novas propostas podem ser aplicadas a trabalhos futuros com os produtos e métodos desenvolvidos neste. Uma das propostas possíveis é a aplicação dessa ferramenta em outros índices e ativos do mercado de ações brasileiro. Outra alternativa seria aplicar o mesmo método de previsão em outras esferas da sociedade,

como política, segurança pública, *marketing*, mercado de trabalho,*etc.*

Outra proposta seria implementar outros métodos preditivos, para que a previsão longa alcance melhores resultados. Isso porque, devido ao fato de a série histórica do IBOV não conter relações de sazonalidade, os modelos da classe Box-Jenkins ficam limitados para previsões de longos períodos, uma vez que os modelos dessa classe são dependentes da sazonalidade para executar esse tipo de previsão.

REFERÊNCIAS

ARAÚJO, J. G. de; MARINHO, L. B. Using online economic news to predict trends in brazilian stock market sectors. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, XXIV., Salvador. **Proceedings**. New York: Association for Computing Machinery (ACM), 2018. p. 37–44.

ATSALAKIS, G. S.; VALAVANIS, K. P. Surveying stock market forecasting techniques - Part II: Soft computing methods. **Expert Systems with Applications**, Pergamon Press, Rockville, v. 36, n. 3, p. 5932–5941, abr. 2009.

BAEZA-YATES, R. **Recuperação de Informação**: Conceitos e Tecnologia das Máquinas de Busca. Porto Alegre: Bookman, 2013.

BOEHM, B. W. A spiral model of software development and enhancement. **Computer**, IEEE, New York, v. 21, n. 5, p. 61–72, 1988.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of computational science**, Elsevier, Amsterdam, v. 2, n. 1, p. 1–8, 2011.

BOX, G. E. *et al.* **Time series analysis: forecasting and control**. Hoboken: John Wiley & Sons, 2015.

BRASIL, BOLSA, BALCÃO (B3). **Índice Bovespa**. 2023. Disponível em: https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm. Acesso em: 9 jan. 2023.

_____. **Uma das principais empresas de infraestrutura de mercado financeiro do mundo**. 2023. Disponível em: https://www.b3.com.br/pt_br/b3/institucional/quem-somos/. Acesso em: 9 jan. 2023.

CAMBRIA, E. *et al.* **A practical guide to sentiment analysis**. Cham: Springer, 2017.

CAROSIA, A. E. O.; COELHO, G. P.; SILVA, A. E. A. da. Analyzing the Brazilian Financial Market through Portuguese Sentiment Analysis in Social Media. **Applied Artificial Intelligence**, Taylor & Francis, Oxford, v. 34, n. 1, p. 1–19, 2020.

_____. Investment strategies applied to the Brazilian stock market: a methodology based on sentiment analysis with deep learning. **Expert Systems with Applications**, Elsevier, Shreveport, v. 184, p. 115470, 2021.

_____. The Influence of Tweets and News on the Brazilian Stock Market through Sentiment Analysis. In: BRAZILLIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, XXV., Rio de Janeiro. **Proceedings**. New York: Association for Computing Machinery, 2019. p. 385–392.

CHAN, W. S. Stock price reaction to news and no-news: drift and reversal after headlines. **Journal of Financial Economics**, Elsevier, New York, v. 70, n. 2, p. 223–260, 2003.

COSTA, M. *et al.* **Investigation of sentiment importance on intraday stock returns**. 2018. Master Thesis (Degree in Data Science) – Graduate School of Economics (GSE), Barcelona, 2018.

DATASENADO. **Redes sociais, notícias falsas e privacidade de dados na internet**. 2019. 113 p. Disponível em: https://www12.senado.leg.br/institucional/ouvidoria/publicacoes-ouvidoria/redes-sociais-noticias-falsas-e-privacidade-de-dados-na-internet. Acesso em: 15 fev. 2020.

DE LONG, J. B. *et al.* Noise Trader Risk in Financial Markets. **Journal of Political Economy**, Chicago, v. 98, n. 4, p. 703–738, 1990.

ELSHENDY, M. *et al.* Using four different online media sources to forecast the crude oil price. **Journal of Information Science**, SAGE Publications, London, v. 44, n. 3, p. 408–421, 2018.

FAKHRY, B. A literature review of the efficient market hypothesis. **Turkish Economic Review**, Istanbul, v. 3, n. 3, p. 431–442, 2016.

FELDMAN, R. Techniques and applications for sentiment analysis. **Communications of the ACM**, Association for Computing Machinery (ACM), New York, v. 56, n. 4, p. 82–89, 2013.

FORECASTING Oil Price Trends with Sentiment of Online News Articles. **Procedia Computer Science**, v. 91, p. 1081–1087, 2016.

GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de pesquisa**. Porto Alegre: Plageder, 2009.

GILBERT, C. H. E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: VIII INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA (ICWSM). **Proceedings**. Michigan, 2014.

GÜNDÜZ, D. *et al.* Machine Learning in the Air. **Journal on Selected Areas in Communications (JSAC)**, IEEE, New York, v. 37, n. 10, p. 2184–2199, 2019.

HO, S. L.; XIE, M. The use of ARIMA models for reliability forecasting and analysis. **Computers & industrial engineering**, Elsevier, Amsterdam, v. 35, n. 1-2, p. 213–216, 1998.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. Melbourne: OTexts, 2018.

IPSOS. **Fake news, filter bubbles, post-truth and trust**. Paris, 2018.

JARIWALA, G.; AGARWAL, H.; JADHAV, V. Sentimental analysis of news headlines for stock market. In: IEEE INTERNATIONAL CONFERENCE FOR INNOVATION IN TECHNOLOGY (INOCON). **Proceedings**. 2020. IEEE, p. 1–5.

JUNIOR, P. R.; SALOMON, F. L. R.; OLIVEIRA PAMPLONA, E. de *et al.* ARIMA: An applied time series forecasting model for the Bovespa stock index. **Journal of Applied Mathematics**, Scientific Research Publishing, London, v. 5, n. 21, p. 3383, 2014.

LEMOS, F. A. C. D. A. **Análise técnica dos mercados financeiros**. São Paulo: Saraiva Educação SA, 2017.

LI, X. *et al.* News impact on stock price return via sentiment analysis. **Knowledge-Based Systems**, Elsevier, Iwate, v. 69, p. 14–23, 2014.

LIN, C.; HE, Y. Joint sentiment/topic model for sentiment analysis. In: CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, XVIII., Hong Kong. **Proceedings**. New York: ACM, 2009. p. 375–384.

LIU, B. **Sentiment analysis**: Mining opinions, sentiments, and emotions. Cambridge: Cambridge University Press, 2015.

LIU, B. *et al.* Sentiment analysis and subjectivity. In: HANDBOOK of natural language processing. Oxford: Chapman & Hall/CRC, 2010. v. 2. p. 627–666.

LIU, Y. *et al.* Machine learning for predicting stock market movement using news headlines. In: IEEE GREEN ENERGY AND SMART SYSTEMS CONFERENCE (IGESSC). **Proceedings**. Long Beach, 2020. IEEE, p. 1–6.

MACHADO, E. J.; PEREIRA, A. C. M. Proposal and Implementation of Machine Learning Models for Stock Markets Using Web Data. In: BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, XXIV., Salvador. **Proceedings**. New York: Association for Computing Machinery, 2018. p. 61–64.

MALKIEL, B. G.; FAMA, E. F. EFFICIENT CAPITAL MARKETS: A REVIEW OF THEORY AND EMPIRICAL WORK. **The Journal of Finance**, Massachusetts, v. 25, n. 2, p. 383–417, 1970.

MALKIEL, G. B. The Efficient Market Hypothesis and Its Critics. **Journal of Economic Perspectives**, Jersey, v. 17, n. 1, p. 59–82, 2003.

MARCA, E. C.; ANTUNES, A. G. Mercado de ações e a análise técnica como principal ferramenta dos investidores. **Unoesc & Ciência-ACSA**, Joaçaba, v. 8, n. 1, p. 59–66, 2017.

MEDEIROS, M.; BORGES, V. Tweet Sentiment Analysis Regarding the Brazilian Stock Market. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 8., Belém. **Anais**. Porto Alegre: SBC, 2019. p. 71–82.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams engineering journal**, Elsevier, Cairo, v. 5, n. 4, p. 1093–1113, 2014.

MONTEIRO, F. G.; FERREIRA, D. R. How Much Does Stock Prediction Improve with Sentiment Analysis? In: MINING DATA FOR FINANCIAL APPLICATIONS (MIDAS): ECML PKDD WORKSHOP, V., Ghent. **Proceedings**. Bangladesh, 2021. Springer, p. 16–31.

MONTEIRO, R. A. *et al.* Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE (PROPOR), XIV., Canela. **Proceedings**. Cham: Springer International Publishing, 2018. p. 324–334.

MORENO, J.; BRESSAN, G. FACTCK.BR: A New Dataset to Study Fake News. In: BRAZILLIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, XXV., Rio de Janeiro. **Proceedings**. New York: Association for Computing Machinery (ACM), 2019. p. 525–527.

MORETTIN, P. A.; TOLOI, C. M. d. C. **Análise de séries temporais**. São Paulo: Blucher, 2022.

NEMES, L.; KISS, A. Prediction of stock values changes using sentiment analysis of stock news headlines. **Journal of Information and Telecommunication (JIT)**, Taylor & Francis, Oxford, v. 5, p. 375–394, 2021.

NIZER, P.; NIEVOLA, J. C. Predicting published news effect in the Brazilian stock market. **Expert Systems with Applications**, Elsevier, College Park, v. 39, n. 12, p. 10674–10680, 2012.

ONCHAROEN, P.; VATEEKUL, P. Deep learning for stock market prediction using event embedding and technical indicators. In: INTERNATIONAL CONFERENCE ON ADVANCED INFORMATICS: CONCEPT THEORY AND APPLICATIONS (ICAICTA), V., Krabi. **Proceedings**. New York, 2018. IEEE, p. 19–24.

PAGOLU, V. S. *et al.* Sentiment analysis of Twitter data for predicting stock market movements. In: INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING, COMMUNICATION, POWER AND EMBEDDED SYSTEM (SCOPES) Paralakhemundi. **Proceedings**. New York, 2016. IEEE, p. 1345–1350.

PANKRATZ, A. **Forecasting with univariate Box-Jenkins models: Concepts and cases**. Hoboken: John Wiley & Sons, 2009.

PRABOWO, R.; THELWALL, M. Sentiment analysis: A combined approach. **Journal of Informetrics**, Elsevier, Oxford, v. 3, n. 2, p. 143–157, 2009.

RAMOS, F. **O que é B3 e como funciona a bolsa de valores brasileira**. 2023. Disponível em: https://www.serasa.com.br/blog/o-que-e-b3-e-como-funciona-a-bolsa-de-valores-brasileira/. Acesso em: 9 jan. 2023.

RUIZ, P. P.; FOGUEM, B. K.; GRABOT, B. Generating knowledge in maintenance from Experience Feedback. **Knowledge-Based Systems**, Elsevier, Iwate, v. 68, p. 4–20, 2014.

RUNDO, F. *et al.* Machine learning for quantitative finance applications: A survey. **Applied Sciences**, MDPI, Basel, v. 9, n. 24, p. 5574, 2019.

SHAH, D.; ISAH, H.; ZULKERNINE, F. Stock market analysis: A review and taxonomy of prediction techniques. **International Journal of Financial Studies**, MDPI, Basel, v. 7, n. 2, p. 26, 2019.

SHARMA, A. *et al.* Use of LSTM and ARIMAX algorithms to analyze impact of sentiment analysis in stock market prediction. In: INTELLIGENT DATA COMMUNICATION TECHNOLOGIES AND INTERNET OF THING (ICICI) Coimbatore. **Proceedings**. New York, 2020. Springer, p. 377–394.

SHU, K. *et al.* Fake News Detection on Social Media: A Data Mining Perspective. **ACM SIGKDD Explorations Newsletter**, New York, v. 19, n. 1, 2017.

SILVA, R. M. *et al.* Towards automatically filtering fake news in Portuguese. **Expert Systems with Applications**, Shreveport, v. 146, 2020.

SILVA, R. F. d. **Automated stock trading system using deep reinforcement learning and price and sentiment prediction modules.** 2021. Universidade de São Paulo, São Paulo.

SUDHAKAR, K.; NAGANJANEYULU, S. Stock Price Prediction based on Finance Related News using NLP, LASSO and ARIMAX. **Journal on Software Engineering (JSE)**, iManager Publications, Nagercoil, v. 14, n. 4, p. 11, 2020.

SUN, S.; LUO, C.; CHEN, J. A review of natural language processing techniques for opinion mining systems. **Information Fusion**, Granada, v. 36, p. 10–25, 2017.

TEIXEIRA, K. A.; LOPES, I. F.; MEURER, A. M. Perfil de Investidor de Acordo com a Autoeficácia Financeira e Características Sociodemográficas: Evidências em Estudantes de Graduação em Ciências Contábeis. In: XVIII. CONGRESSO USP de Iniciação Científica em Contabilidade. 2021.

WAZLAWICK, R. S. **Metodologia de pesquisa para ciência da computação**. Rio de Janeiro: Elsevier - Campus, 2009.

ZHANG, G. P. Time series forecasting using a hybrid ARIMA and neural network model. **Neurocomputing**, Elsevier, Toulouse, v. 50, p. 159–175, 2003.

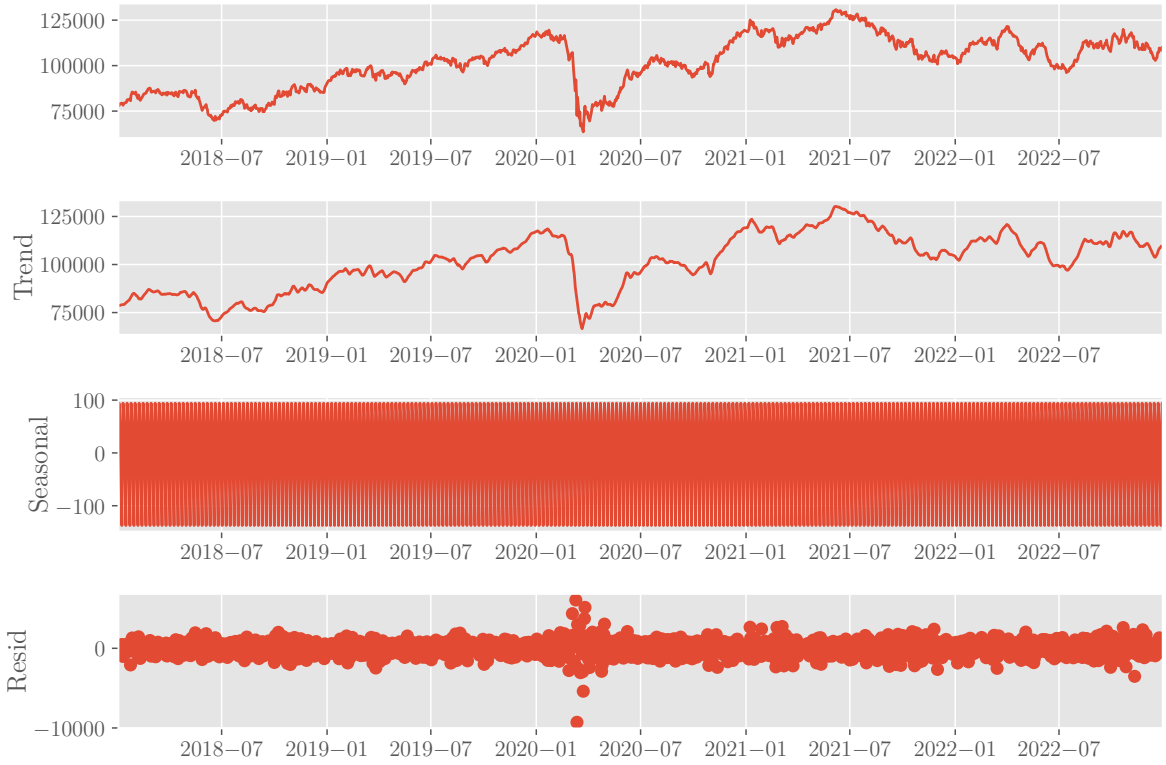
APÊNDICES

APÊNDICE A – TESTES ESTATÍSTICOS

Este apêndice apresenta os resultados de experimentos e testes estatísticos descritos no capítulo de desenvolvimento. Inicialmente, são apresentados as decomposições das séries temporais utilizadas no decorrer deste trabalho. Pode-se, observar nessas decomposições a série temporal original, a tendência da série (*trend*), a sazonalidade (*seasonal*) e os resíduos/ruídos (*resid*). Em seguida, são avaliadas as autocorrelações da série temporal do IBOV levando em consideração o resultado do teste de Dickey-Fuller aumentado (ADF), com intuito de determinar o melhor parâmetro de diferenciação para tornar a série temporal em uma série estacionária, em virtude, de ser um dos pré-requisitos para aplicações dos métodos econométricos da classe Box-Jenkins. Posteriormente, após a identificação do melhor parâmetro de diferenciação, aplicou-se o algoritmo de *grid search* para obtenção dos melhores parâmetros, levando em consideração a transformação da série para estacionária.

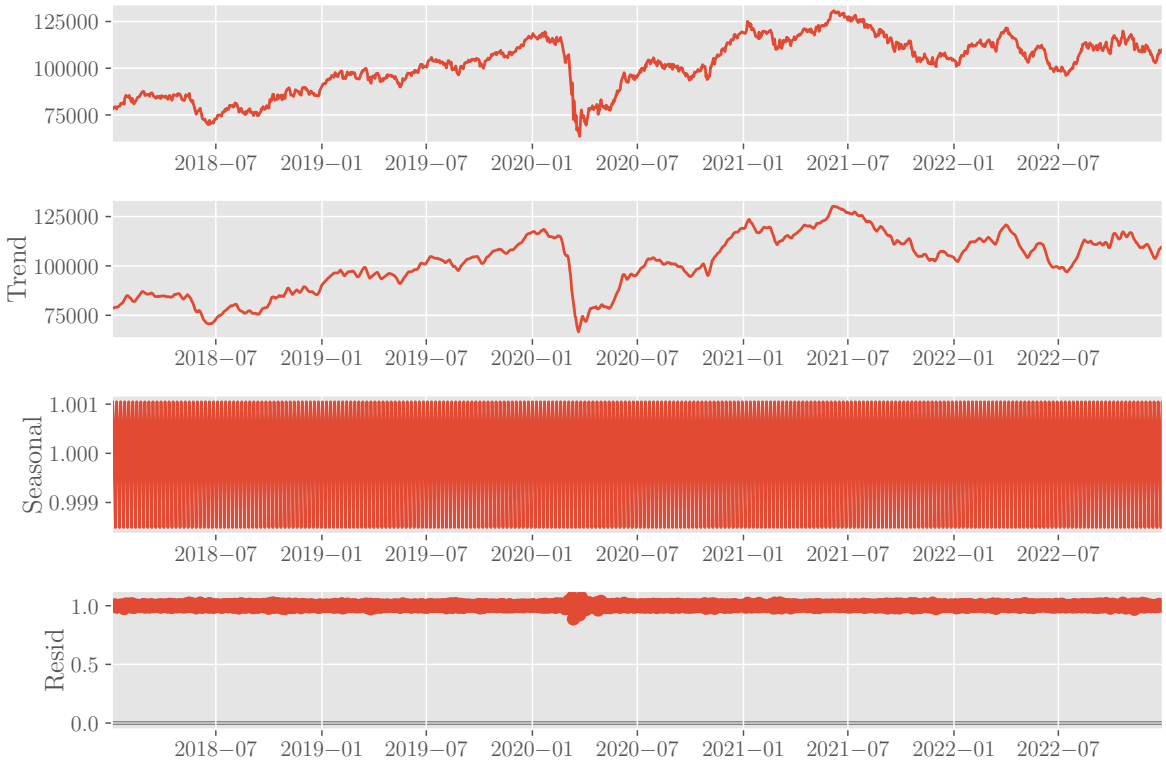
Análise da série temporal de preço de fechamento do IBOV.

Decomposição aditiva IBOV.



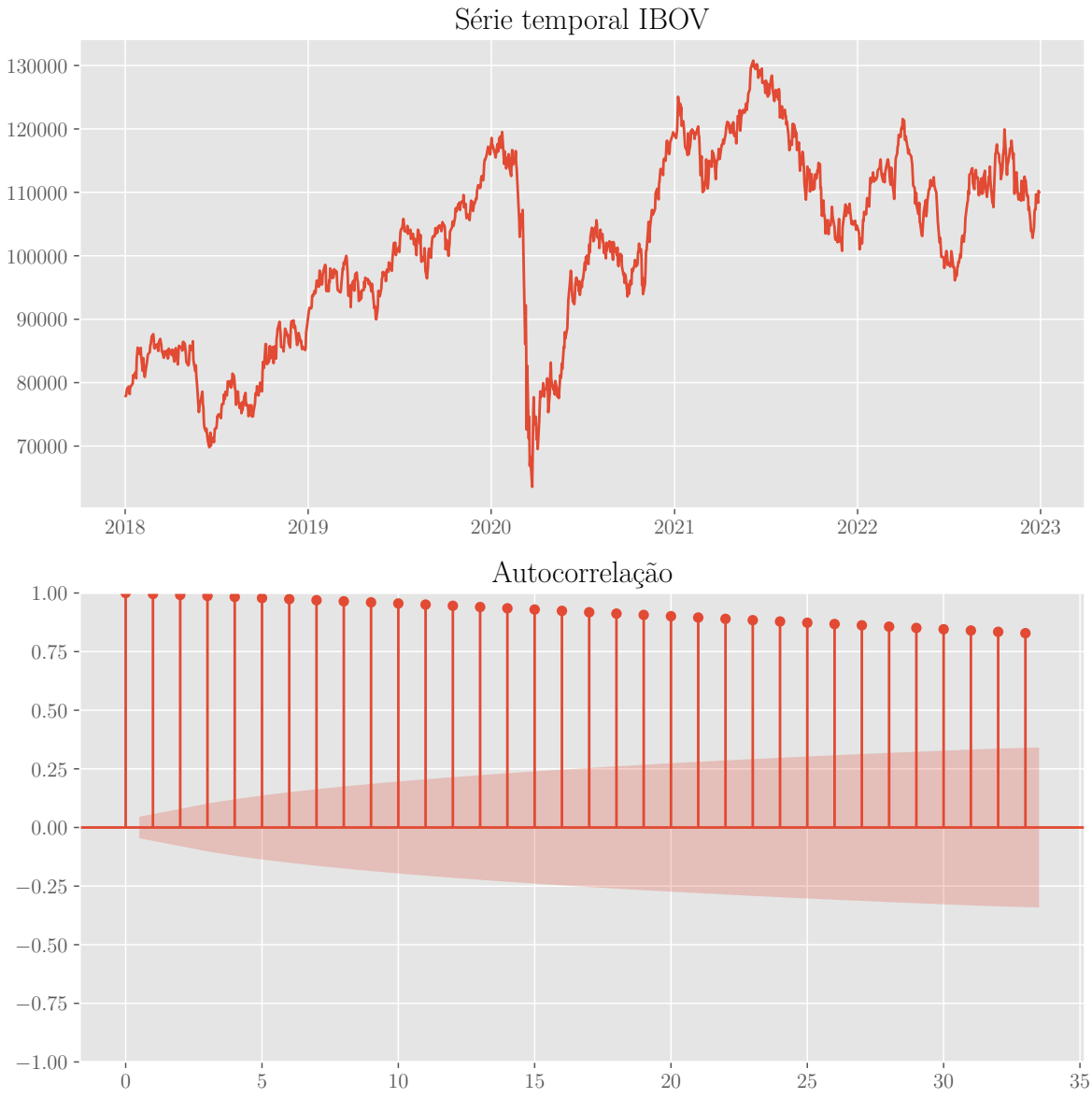
Fonte: Elaborado pelo Autor (2023).

Decomposição multiplifcativa IBOV.



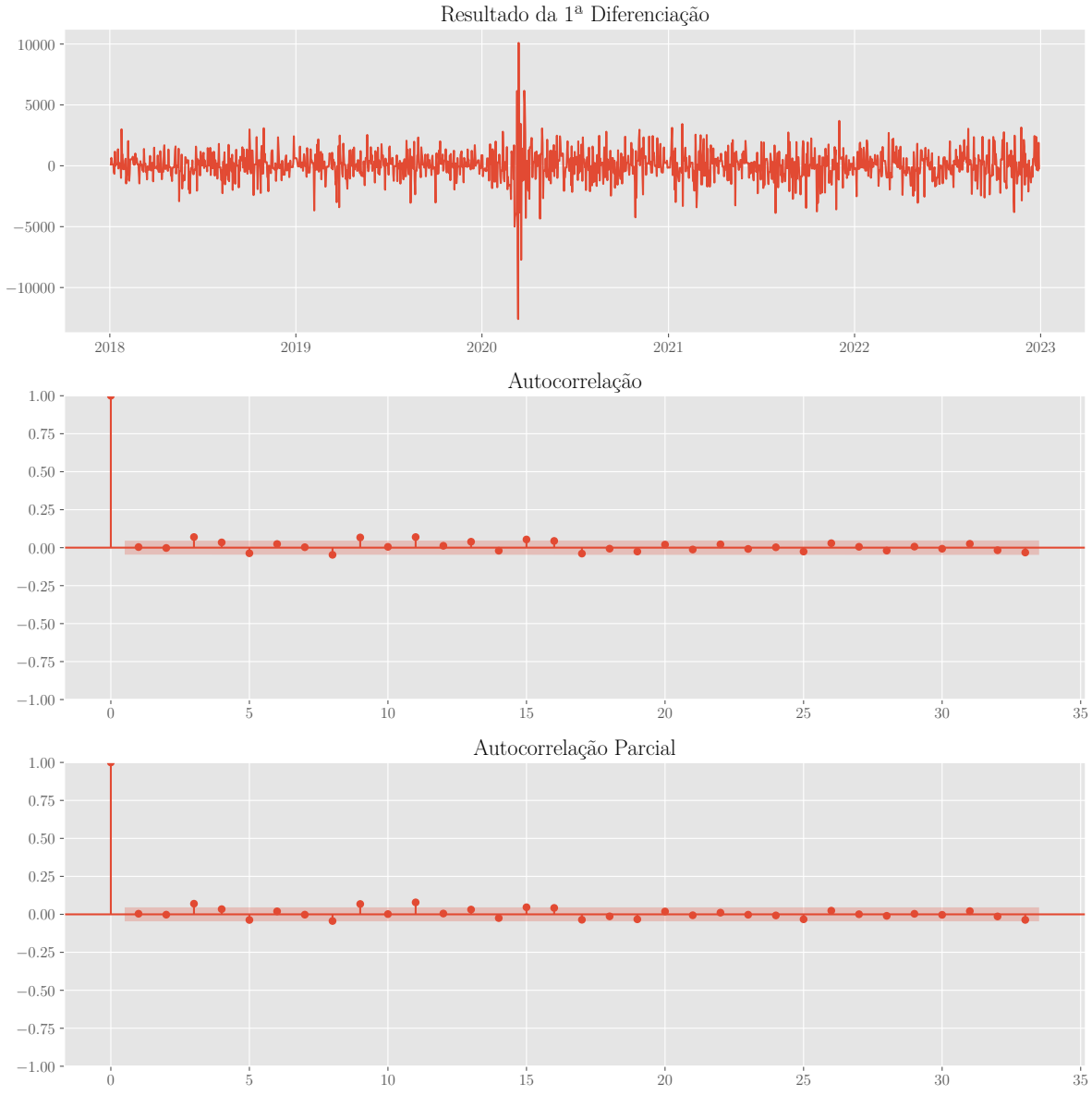
Fonte: Elaborado pelo Autor (2023).

Autocorrelação série original IBOV.



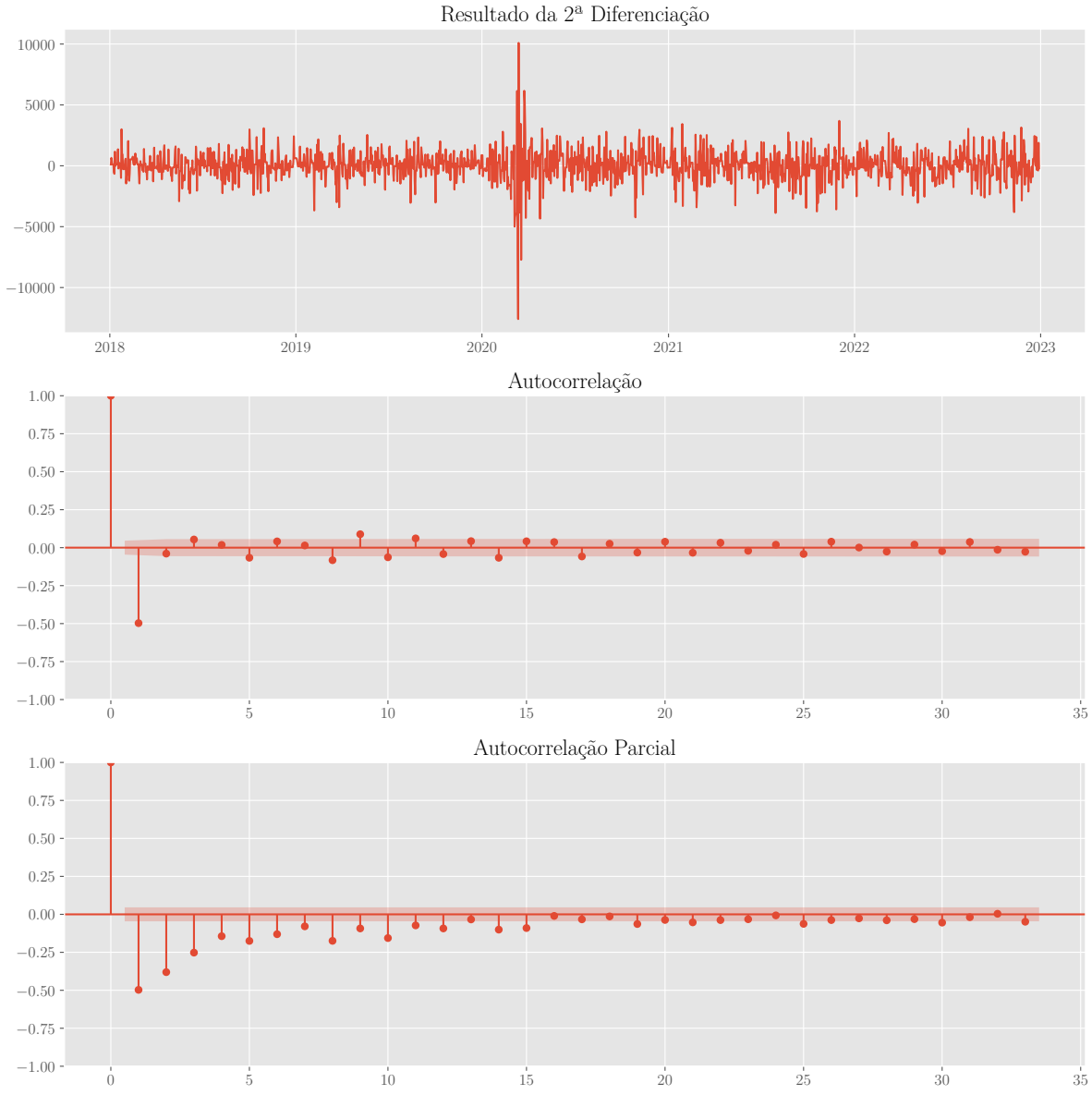
Fonte: Elaborado pelo Autor (2023).

Autocorrelação do IBOV após 1ª diferenciação.



Fonte: Elaborado pelo Autor (2023).

Autocorrelação do IBOV após 2ª diferenciação.



Fonte: Elaborado pelo Autor (2023).

Decomposição aditiva da série temporal de sentimentos das notícias.

Fonte: Elaborado pelo Autor (2023).

Decomposição aditiva da série temporal de sentimentos dos títulos das notícias.

Fonte: Elaborado pelo Autor (2023).

Decomposição aditiva da série temporal de sentimentos dos tweets.

Fonte: Elaborado pelo Autor (2023).

APÊNDICE B – MELHORES PARÂMETROS E MODELOS OBTIDOS

Tabela 9 – Melhores parâmetros obtidos com grid search, desconsiderando a sazonalidade.

Variável Exógena	MAE	pdq	MSE	pdq	RMSE	pdq
Sentimentos de *tweets*	749,0676	(2, 1, 0)	1.240.347,5251	(4, 1, 4)	1.113,7089	(4, 1, 4)
Sentimentos de títulos das notícias e *tweets*	757,7653	(2, 1, 0)	1.296.753,7399	(3, 1, 5)	1.138,7510	(3, 1, 5)
Sentimentos de notícias	769,4781	(2, 1, 0)	1.335.876,6019	(5, 1, 4)	1.155,8013	(5, 1, 4)
Sentimentos de títulos das notícias	770,6798	(2, 1, 1)	1.336.791,4418	(5, 1, 5)	1.156,1970	(5, 1, 5)
Sentimentos de notícias e *tweets*	764,0557	(2, 1, 0)	1.327.272,7887	(2, 1, 3)	1.152,0733	(4, 1, 4)
Sem variável exógena	776,8208	(2, 1, 0)	1.351.618,3070	(5, 1, 4)	1.162,5912	(5, 1, 4)

Fonte: Elaborado pelo Autor (2023).

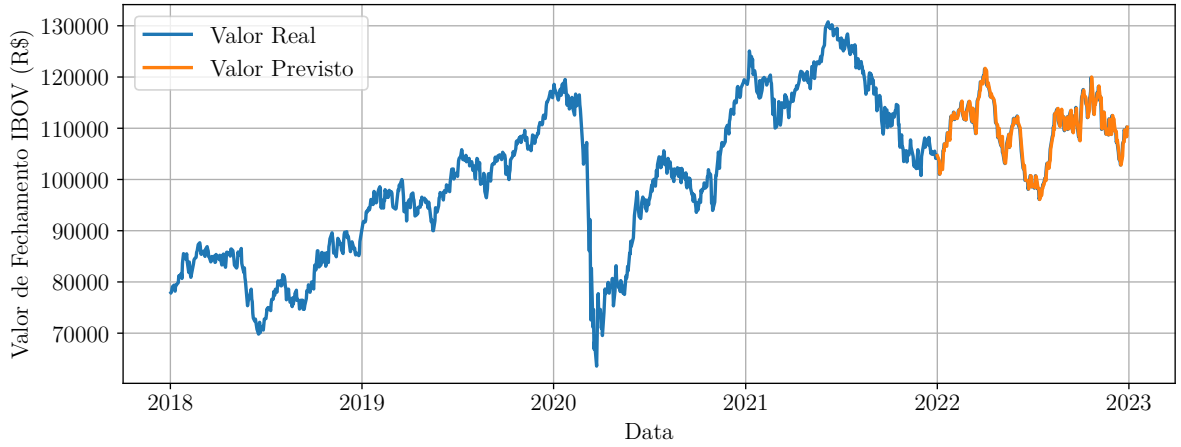
Tabela 10 – Melhores parâmetros obtidos com grid search, considerando a sazonalidade.

Variável Exógena	MAE	MSE	RMSE	pdq	PDQS
Sentimentos de *tweets*	751,2256	1.240.347,5251	1.113,7089	(4, 1, 4)	(0, 0, 0, 0)
Sentimentos de títulos das notícias e *tweets*	759,9004	1.296.753,7399	1.138,7510	(3, 1, 5)	(0, 0, 0, 0)
Sentimentos de notícias e *tweets*	765,3313[1]	1.322.861,8229*	1.150,1573*	(2, 1, 3)	(2, 0, 0, 3)[1], (5, 0, 0, 3)*
Sentimentos de títulos das notícias	773,3067	1.336.791,4418	1.156,1970	(5, 1, 5)	(0, 0, 0, 0)
Sentimentos de notícias	774,9664	1.335.876,6019	1.155,8013	(5, 1, 4)	(0, 0, 0, 0)
Sem variável exógena	781,1319	1.351.618,3070	1.162,5912	(5, 1, 4)	(0, 0, 0, 0)

Fonte: Elaborado pelo Autor (2023).

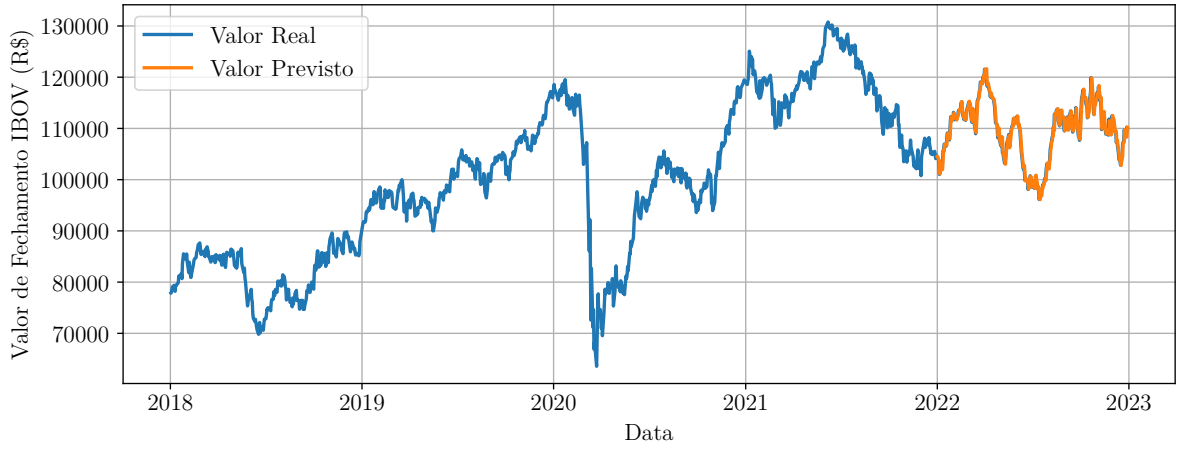
APÊNDICE C – TESTES DE MODELOS PREDITIVOS COM SELEÇÃO DE PARÂMETROS ATRAVÉS DE AUTOARIMA

Previsão do IBOV sem sentimentos.



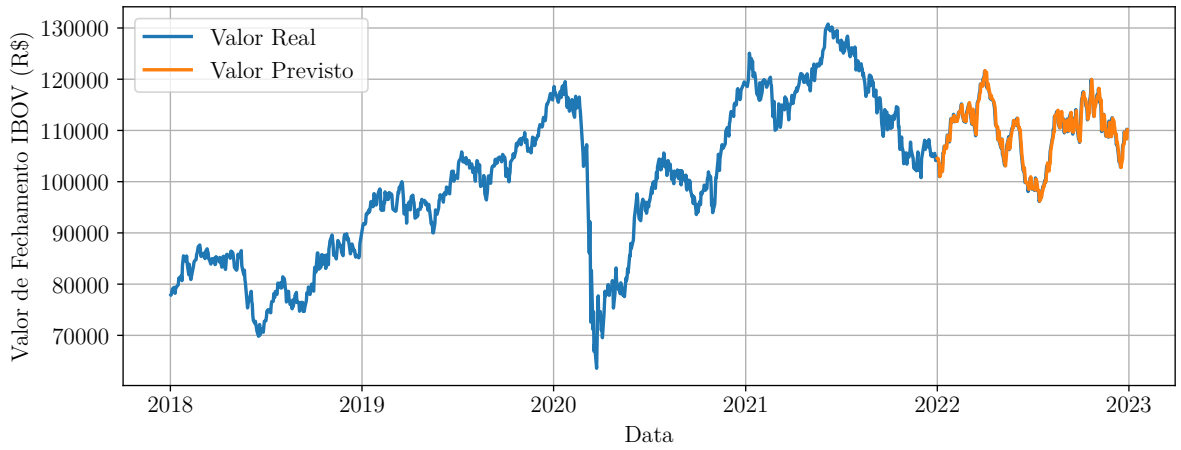
Fonte: Elaborado pelo Autor (2023).

Previsão do IBOV considerando os sentimentos de notícias.



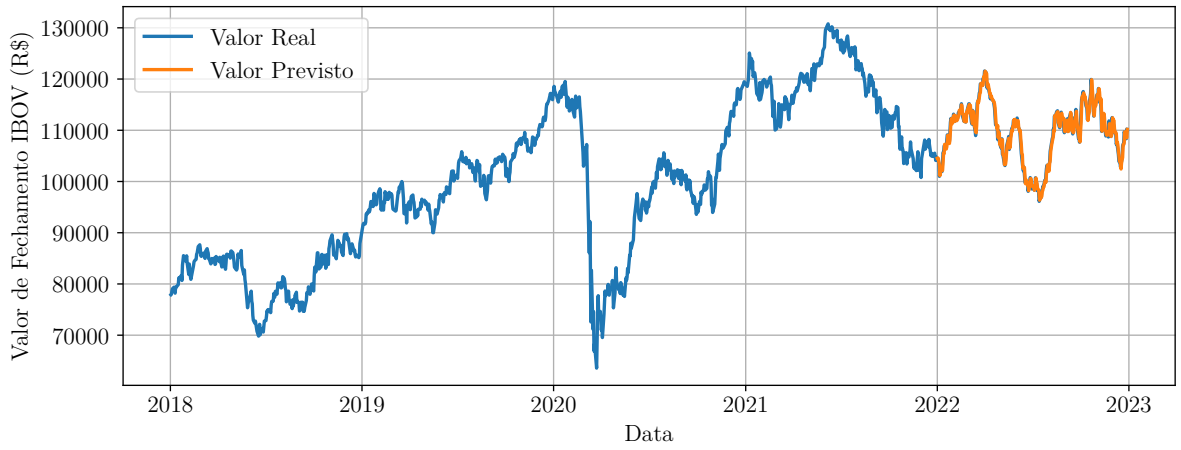
Fonte: Elaborado pelo Autor (2023).

Previsão do IBOV considerando os sentimentos dos títulos das notícias.



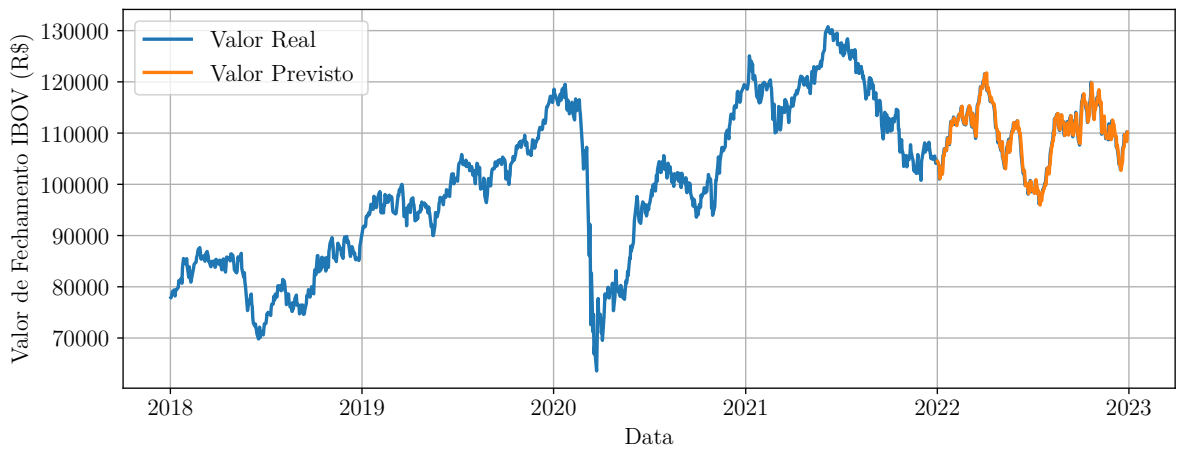
Fonte: Elaborado pelo Autor (2023).

Previsão do IBOV considerando os sentimentos de tweets.



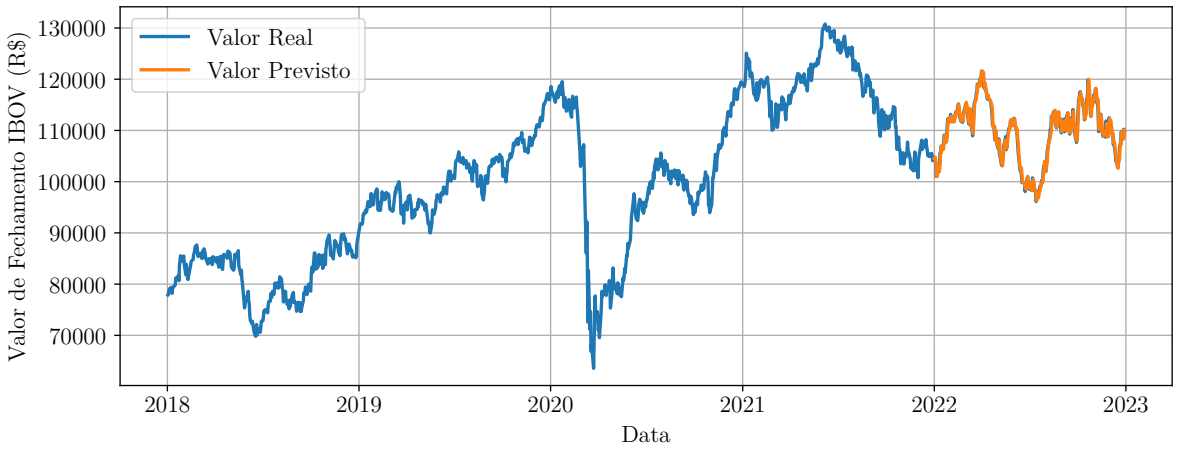
Fonte: Elaborado pelo Autor (2023).

Previsão do IBOV considerando sentimentos de notícias e tweets.



Fonte: Elaborado pelo Autor (2023).

Previsão do IBOV considerando sentimentos dos títulos das notícias e tweets.



Fonte: Elaborado pelo Autor (2023).

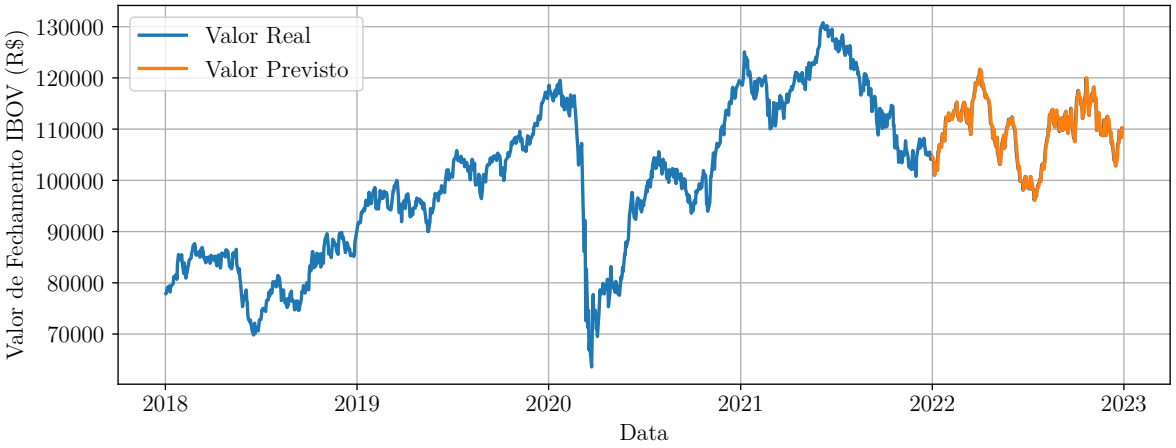
Tabela 11 – Erros das previsões do IBOV com parâmetros selecionados pelo autoarima dos modelos referentes ao período de 2018 a 2022.

Variável Exógena	**Parâmetros**	**MAPE**	**MAE**	**MSE**	**RMSE**
Sentimentos de tweets	(p:0,d:1,q:0)	0.0066	734.2763	998957.7734	999.4787
Sentimentos de títulos de notícias e tweets	(p:2,d:1,q:2)	0.0068	752.6103	1040783.9418	1020.1881
Sentimentos de notícias e tweets	(p:0,d:1,q:0)	0.0069	760.9598	1071833.0114	1035.2936
Sentimentos de títulos de notícias	(p:2,d:1,q:2)	0.0069	766.4535	1089707.3585	1043.8904
Sentimentos de notícias	(p:0,d:1,q:0)	0.0069	766.9113	1091706.0410	1044.8473
Sem variável exógena	(p:2,d:1,q:2)	0.0070	779.0925	1124410.7053	1060.3823

Fonte: Elaborado pelo Autor (2023).

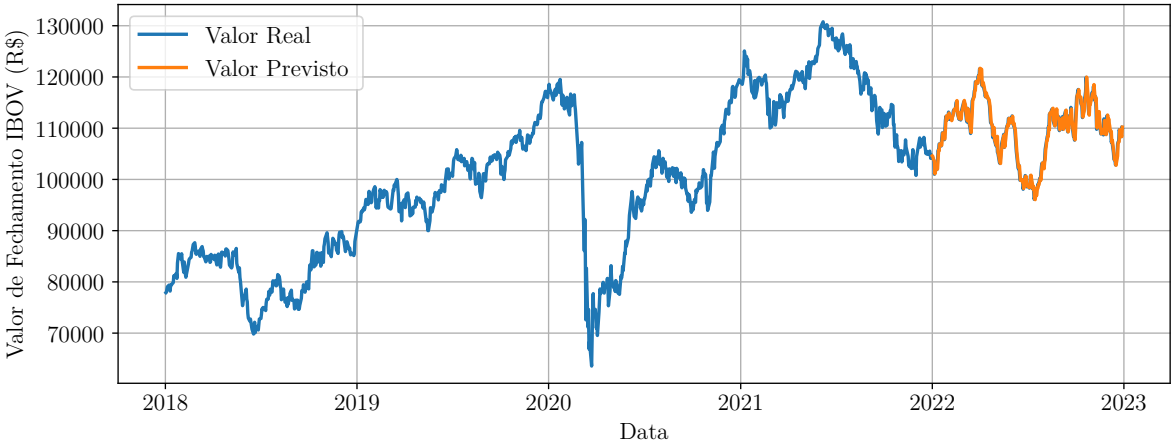
APÊNDICE D – EXPERIMENTOS

Figura 6 – Previsão do IBOV sem sentimentos.



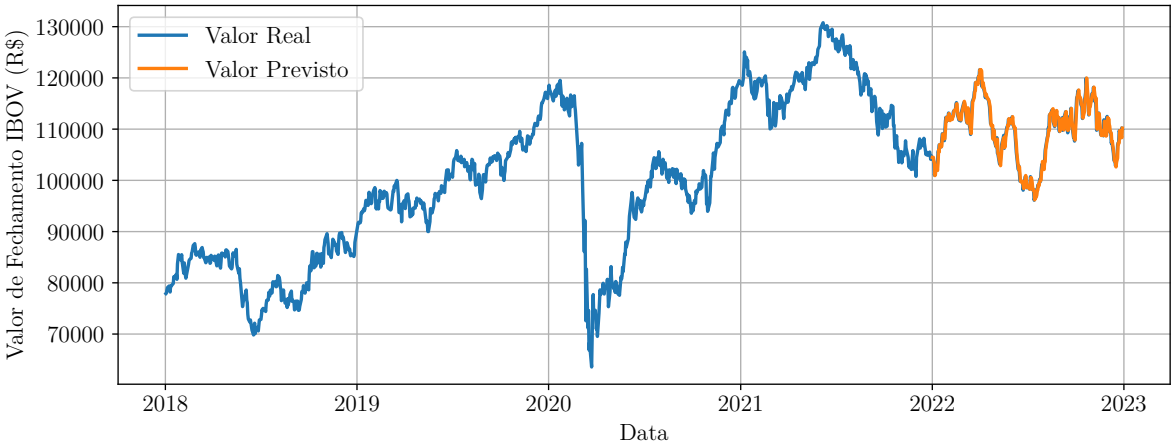
Fonte: Elaborado pelo Autor (2023).

Figura 7 – Previsão do IBOV com sentimentos de notícias.



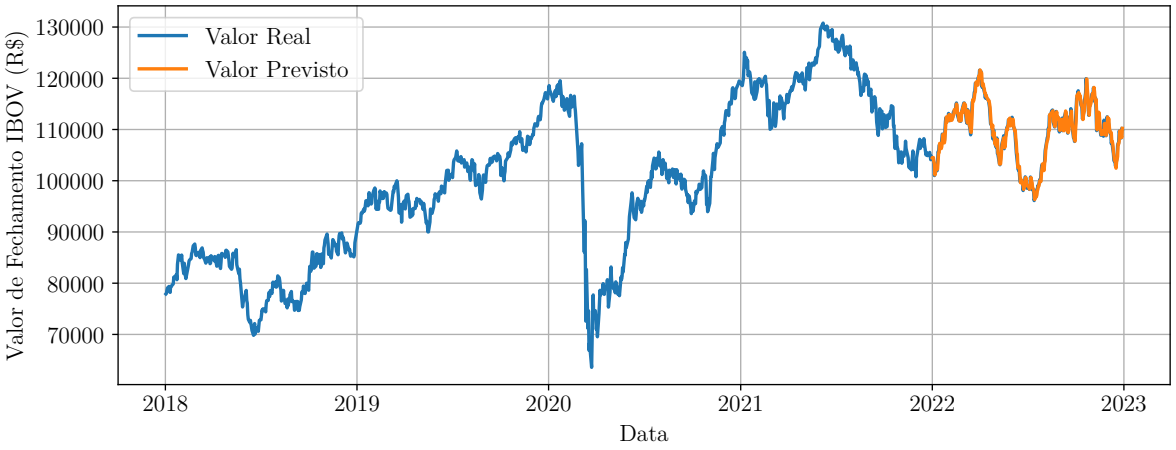
Fonte: Elaborado pelo Autor (2023).

Figura 8 – Previsão do IBOV com sentimentos dos títulos das notícias.



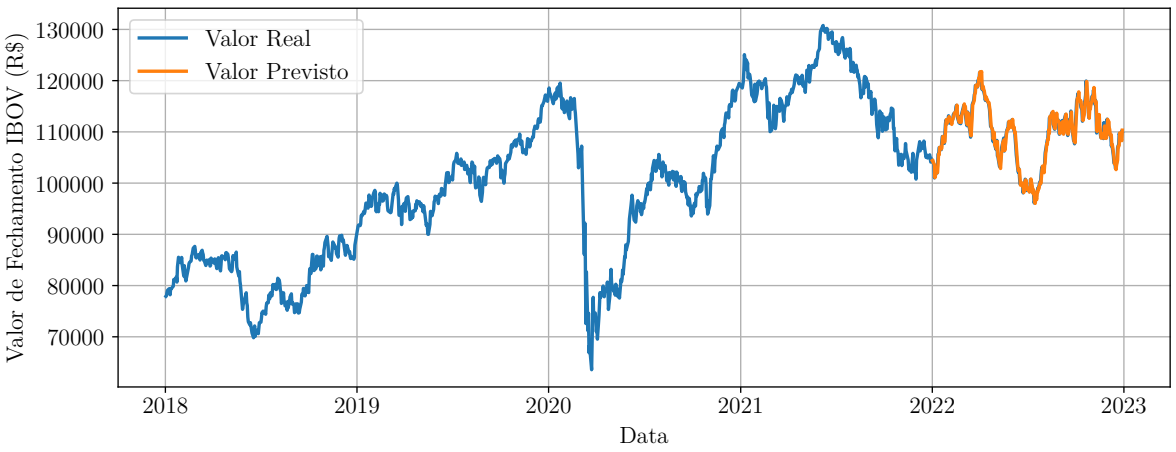
Fonte: Elaborado pelo Autor (2023).

Figura 9 – Previsão do IBOV com sentimentos de *tweets*.



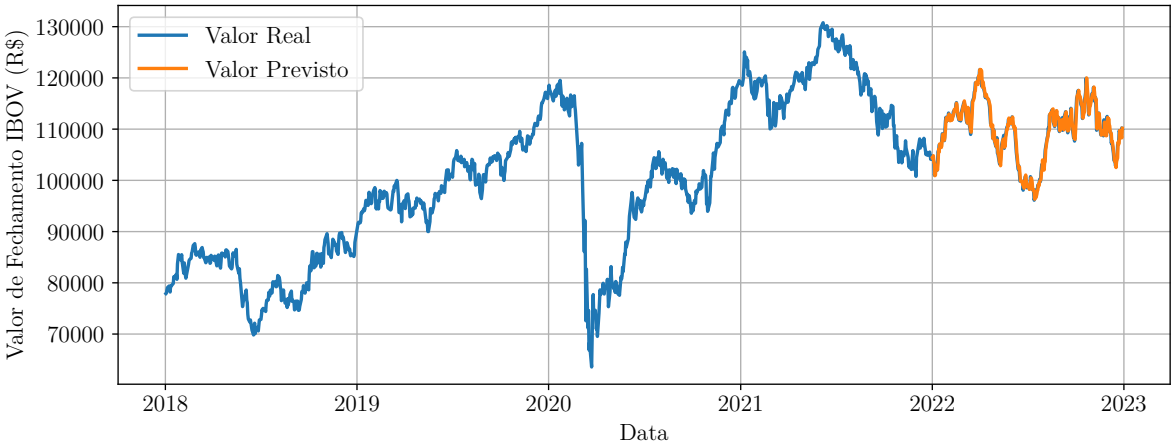
Fonte: Elaborado pelo Autor (2023).

Figura 10 – Previsão do IBOV com sentimentos de notícias e *tweets*.



Fonte: Elaborado pelo Autor (2023).

Figura 11 – Previsão do IBOV com sentimentos dos títulos das notícias e *tweets*.



Fonte: Elaborado pelo Autor (2023).

Tabela 12 – Descrição de modelos e respectivos erros de previsão diária do IBOV referente ao período de 2018 a 2022.

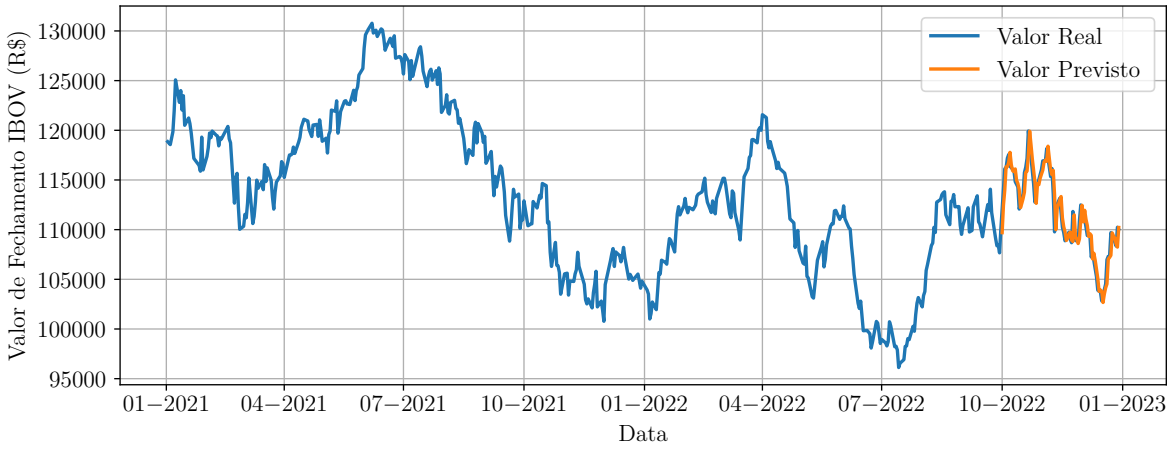
Variável Exógena	**MAPE**	**MAE**	**MSE**	**RMSE**	**pdq**	**PDQS**
Sentimentos de *tweets*	0,0066	736,1539	997.422,6312	998,7104	(4, 1, 4)	(0, 0, 0, 0)
Sentimentos de títulos de notícias e *tweets*	0,0068	751,7421	1.046.133,0050	1.022,8064	(3, 1, 5)	(0, 0, 0, 0)
Sentimentos de notícias e *tweets*	0,0069	760,4713	1.066.094,4456	1.032,5184	(2, 1, 3)	(5, 0, 0, 3)
Sentimentos de notícias	0,0069	765,3690	1.084.079,0252	1.041,1911	(5, 1, 4)	(0, 0, 0, 0)
Sentimentos de títulos das notícias	0,0069	767,3590	1.095.810,8827	1.046,8098	(5, 1, 5)	(0, 0, 0, 0)
Sem variável exógena	0,0070	777,1252	1.125.359,7878	1.060,8297	(5, 1, 4)	(0, 0, 0, 0)

Fonte: Elaborado pelo Autor (2023).

Figura 12 – Previsão do IBOV considerando o histórico de 2021 a 2022 sem sentimentos.

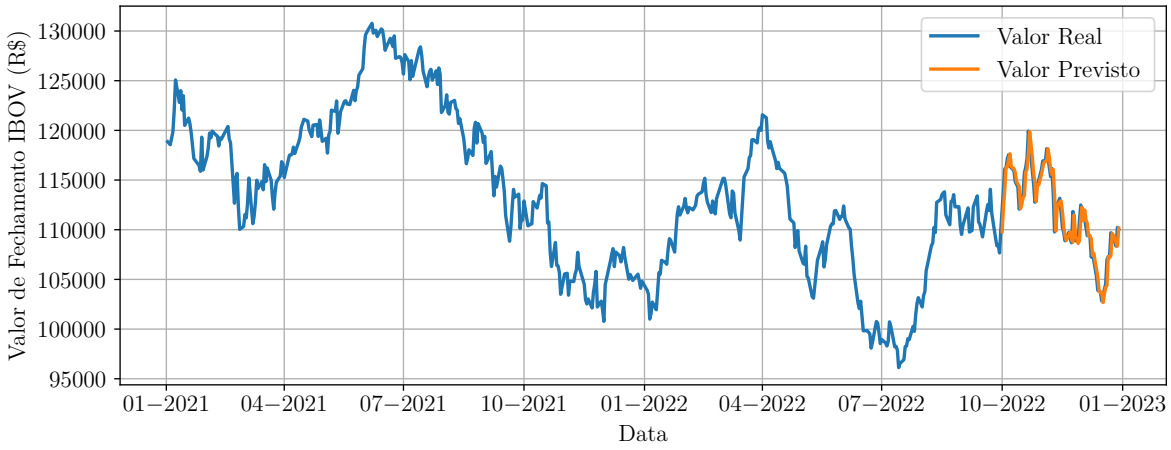
Fonte: Elaborado pelo Autor (2023).

Figura 13 – Previsão do IBOV considerando o histórico de 2021 a 2022 com sentimentos de notícias



Fonte: Elaborado pelo Autor (2023).

Figura 14 – Previsão do IBOV com sentimentos dos títulos das notícias



Fonte: Elaborado pelo Autor (2023).

Figura 15 – Previsão do IBOV com sentimentos de *tweets*

Fonte: Elaborado pelo Autor (2023).

Tabela 13 – Descrição de modelos e respectivos erros de previsão diária do IBOV referente ao período após 2020

Variável Exógena	**MAPE**	**MAE**	**MSE**	**RMSE**	**pdq**	**PDQS**
Sentimentos de *tweets*	0,0080	899,5446	1.419.333,1539	1.191,3576	(4, 1, 4)	(0, 0, 0, 0)
Sentimentos de títulos das notícias	0,0083	937,7044	1.536.731,5183	1.239,6497	(5, 1, 5)	(0, 0, 0, 0)
Sentimentos de notícias	0,0084	949,6574	1.521.684,3713	1.233,5657	(5, 1, 4)	(0, 0, 0, 0)
Sem variável exógena	0,0085	958,1709	1.609.636,3738	1.268,7144	(5, 1, 4)	(0, 0, 0, 0)

Fonte: Elaborado pelo Autor (2023).

Tabela 14 – Descrição de modelo e respectivos erros de previsão de 7 dias do IBOV referente ao período 2018 a 2022

Variável Exógena	**MAPE**	**MAE**	**MSE**	**RMSE**	**pdq**	**PDQS**
Sentimentos de *tweets*	0,0175	1.919,5656	4.037.199,3982	2.009,2783	(4, 1, 4)	(0, 0, 0, 0)

Fonte: Elaborado pelo Autor (2023).

Tabela 15 – Descrição de modelo e respectivos erros de previsão de 15 dias do IBOV referente ao período 2018 a 2022

Variável Exógena	**MAPE**	**MAE**	**MSE**	**RMSE**	**pdq**	**PDQS**
Sentimentos de *tweets*	0,0303	3.300,1361	15.889.195,5894	3.986,1253	(4, 1, 4)	(0, 0, 0, 0)

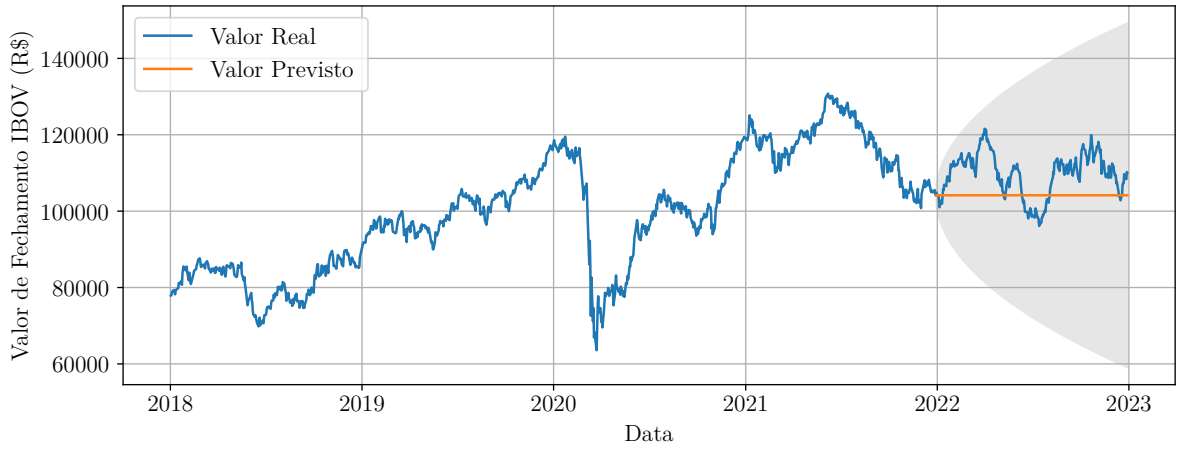
Fonte: Elaborado pelo Autor (2023).

Tabela 16 – Descrição de modelo e respectivos erros de previsão de 30 dias do IBOV referente ao período 2018 a 2022

Variável Exógena	**MAPE**	**MAE**	**MSE**	**RMSE**	**pdq**	**PDQS**
Sentimentos de *tweets*	0,0318	3.370,7848	17.181.567,4878	4.145,0654	(4, 1, 4)	(0, 0, 0, 0)

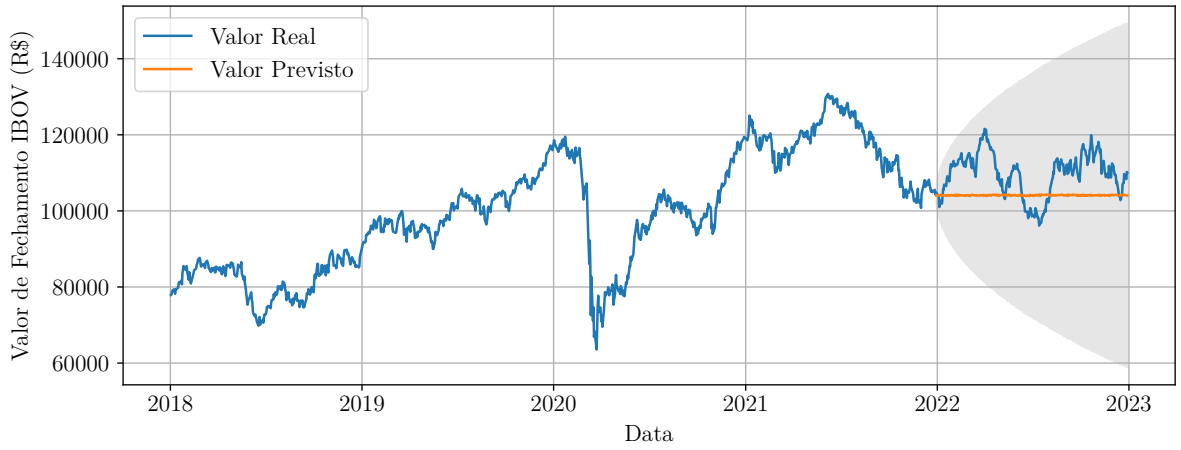
Fonte: Elaborado pelo Autor (2023).

Previsão de 365 dias do IBOV sem sentimentos.



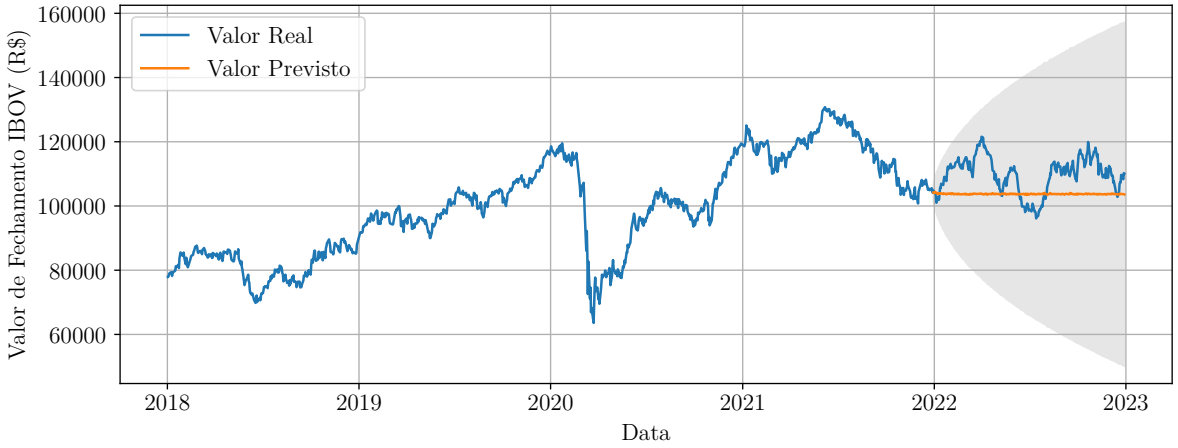
Fonte: Elaborado pelo Autor (2023).

Previsão de 365 dias do IBOV com sentimentos de notícias.



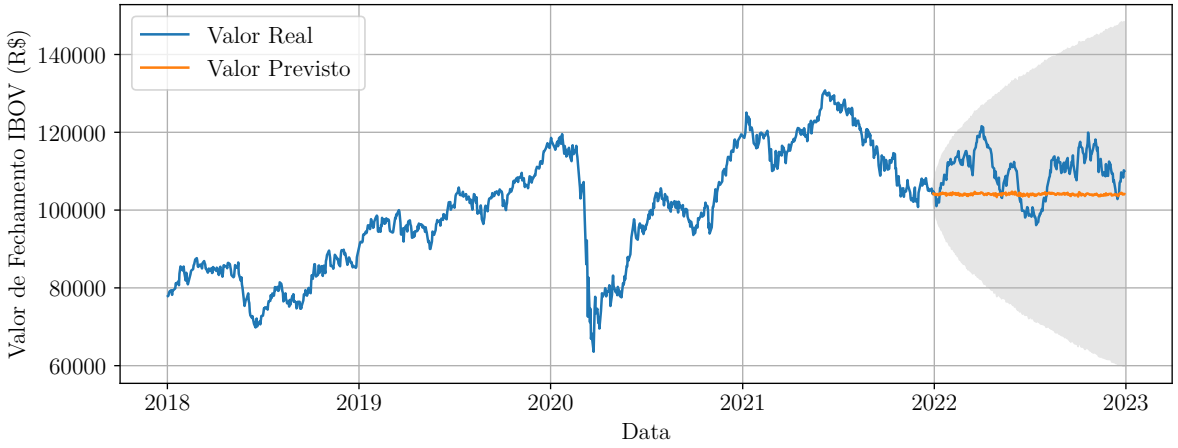
Fonte: Elaborado pelo Autor (2023).

Previsão de 365 dias do IBOV com sentimentos dos títulos das notícias.



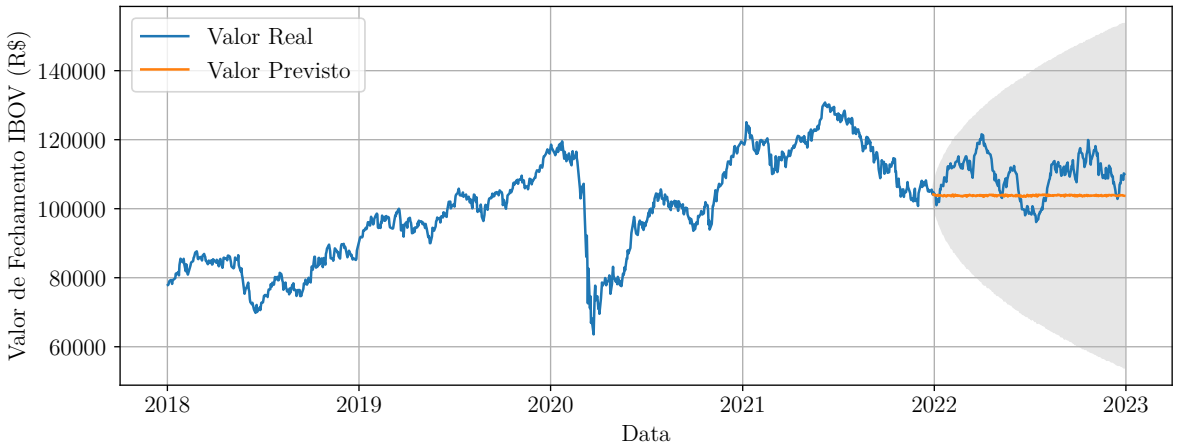
Fonte: Elaborado pelo Autor (2023).

Previsão de 365 dias do IBOV com sentimentos de tweets.



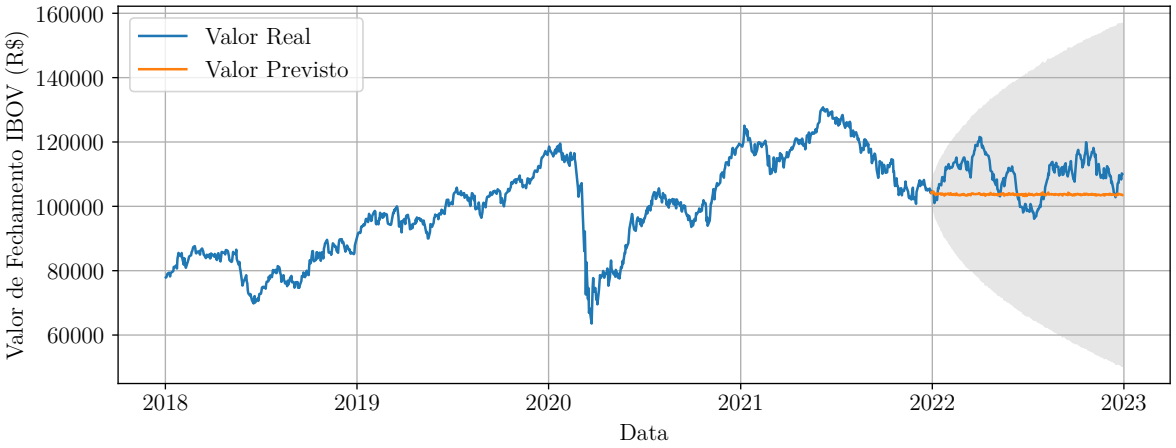
Fonte: Elaborado pelo Autor (2023).

Previsão de 365 dias do IBOV com sentimentos de notícias e tweets.



Fonte: Elaborado pelo Autor (2023).

Previsão de 365 dias do IBOV com sentimentos dos títulos das notícias e tweets.



Fonte: Elaborado pelo Autor (2023).

Tabela 17 – Erros das previsões de 365 dias dos modelos referentes ao período de 2018 a 2022.

Variável Exógena	Parâmetros	MAPE	MAE	MSE	RMSE
Sem variável exógena	(p:3,d:1,q:4; P:0,D:0,Q:0,S:0)	0.0615	6885.8357	62250397.4105	7889.8921
Sentimentos de notícias	(p:2,d:1,q:3; P:0,D:0,Q:0,S:0)	0.0616	6902.4753	62603819.9113	7912.2575
Sentimentos de tweets	(p:4,d:1,q:4; P:0,D:0,Q:0,S:0)	0.0616	6906.0238	62595787.2155	7911.7499
Sentimentos de notícias e tweets	(p:2,d:1,q:3; P:5,D:0,Q:0,S:3)	0.0631	7081.4641	65928746.3572	8119.6518
Sentimentos de títulos das notícias	(p:3,d:1,q:3; P:0,D:0,Q:0,S:0)	0.0638	7152.0223	67108641.2857	8191.9864
Sentimentos de títulos das notícias e tweets	(p:3,d:1,q:5; P:0,D:0,Q:0,S:0)	0.0639	7161.8947	67208553.9905	8198.0823

Fonte: Elaborado pelo Autor (2023).

Previsão longa do IBOV considerando o histórico de 2021 a 2022 sem sentimentos.

Fonte: Elaborado pelo Autor (2023).

Previsão longa do IBOV considerando o histórico de 2021 a 2022 com sentimentos de notícias.

Fonte: Elaborado pelo Autor (2023).

Previsão longa do IBOV com sentimentos dos títulos das notícias.

Fonte: Elaborado pelo Autor (2023).

Previsão longa do IBOV com sentimentos de tweets.

Fonte: Elaborado pelo Autor (2023).

Tabela 18 – Erros das previsões longas dos modelos referentes ao período de 2021 a 2022.

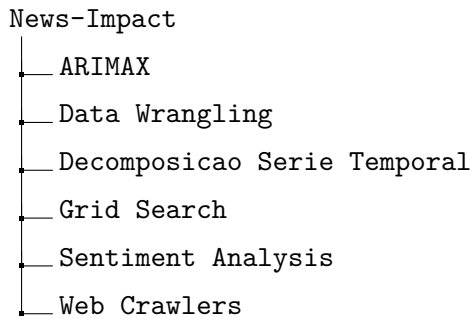
Variável Exógena	**Parâmetros**	**MAPE**	**MAE**	**MSE**	**RMSE**
Sem variável exógena	(p:3,d:1,q:4; P:0,D:0,Q:0,S:0)	0.0447	5068.4282	34140395.7789	5842.9783
Sentimentos de títulos das notícias	(p:3,d:1,q:3; P:0,D:0,Q:0,S:0)	0.0452	5126.5123	34767118.4566	5896.3648
Sentimentos de tweets	(p:4,d:1,q:4; P:0,D:0,Q:0,S:0)	0.0480	5445.1549	38415376.5500	6198.0139
Sentimentos de notícias	(p:2,d:1,q:3; P:0,D:0,Q:0,S:0)	0.0504	5705.8003	41774262.5725	6463.3012

Fonte: Elaborado pelo Autor (2023).

APÊNDICE E – REPOSITÓRIOS

Neste Apêndice, estão descritos os repositórios do presente trabalho, bem como a forma de acessá-los, além de também apresentar, a organização dos mesmos. Os *scripts* utilizados no desenvolvimento deste trabalho, foram armazenados no *GitHub* através do repositório privado *News-Impact*[1]. A Figura 16, apresenta a divisão de diretórios para armazenamento dos arquivos.

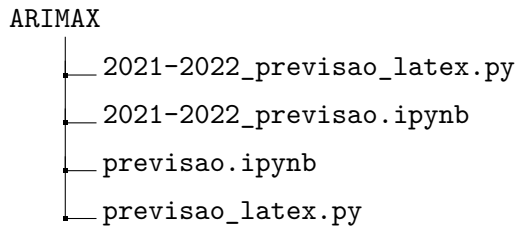
Figura 16 – Árvore de diretórios do repositório



Fonte: Elaborado pelo Autor (2023).

Em seguida, a Figura 17, aborda os *scripts* "2021-2022_previsao.ipynb" e "previsao.ipynb" utilizados para desenvolvimento dos modelos preditivos e análise exploratória, no diretório ARIMAX também são expostos os arquivos "2021-2022_previsao_latex.py' e "previsao_latex.py", empregados na realização de plotagens gráficas apresentadas neste trabalho.

Figura 17 – Diretório com arquivos de desenvolvimento do modelo preditivo.

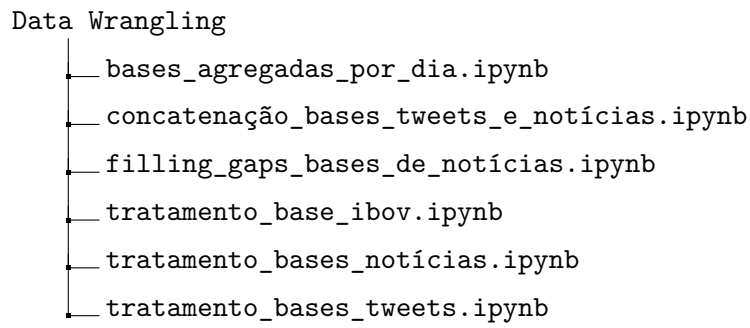


Fonte: Elaborado pelo Autor (2023).

Posteriormente, na Figura 18 são explicitados os notebooks utilizados para empregar os tratamento das bases de dados.

[1] https://github.com/newsimpact/News-Impact/

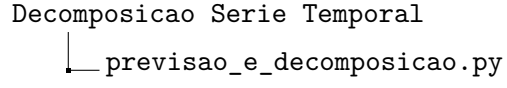
Figura 18 – Diretório com arquivos utilizados para manipulação e tratamento das bases de dados.



Fonte: Elaborado pelo Autor (2023).

A Figura 19, evidencia o arquivo que contém a análise da série temporal do IBOV, em conjunto com testes de correlações, estacionaridade, tendências, sazonalidade e ruídos.

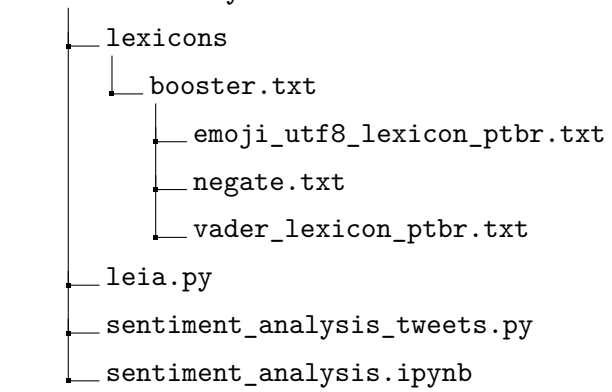
Figura 19 – Diretório com arquivo incluindo análises da série temporal.



Fonte: Elaborado pelo Autor (2023).

Ademais, a Figura 20 apresenta os arquivos "sentiment_analysis_tweets.py" "sentiment_analysis.ipynb" utilizados para emprego da análise de sentimentos, seus respectivos arquivos e pastas de dependência também podem ser observados.

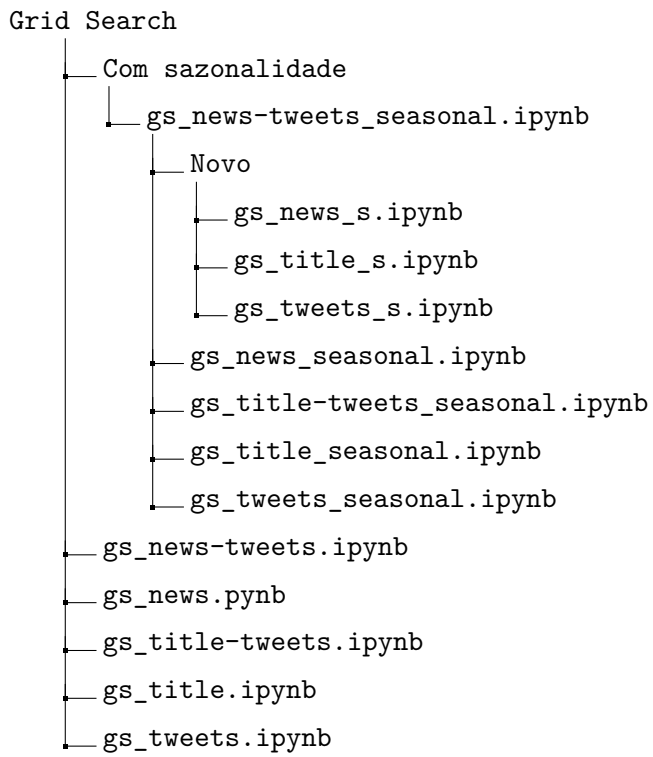
Figura 20 – Diretório com arquivos utilizados na aplicação da análise de sentimentos.



Fonte: Elaborado pelo Autor (2023).

Na Figura 21, é possível observar a organização dos *scripts* empregados nos testes com o algoritmo *Grid Search* para seleção dos melhores modelos.

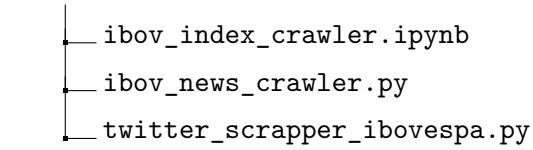
Figura 21 – Diretório com arquivos utilizados na seleção dos melhores modelo com *Grid Search*.



Fonte: Elaborado pelo Autor (2023).

Por fim, a Figura 22 expõe os arquivos responsáveis por extrair as informações utilizadas no desenvolvimento de cada base de dados. As bases de dados de *tweets*[2], notícias[3] e cotações do IBOV[4] foram publicadas de forma privada na *Kaggle*. O acesso às bases pode ser realizado através dos links descritos no rodapé.

Figura 22 – Diretório com arquivos utilizados para desenvolvimento das bases de dados.



Fonte: Elaborado pelo Autor (2023).

[2] https://www.kaggle.com/datasets/emanuelelias/ibovtweets
[3] https://www.kaggle.com/datasets/emanuelelias/ibovnews
[4] https://www.kaggle.com/datasets/emanuelelias/ibov-historical-data-of-the-brazilian-stock-market

ANEXOS

ANEXO A – PÁGINAS NA INTERNET RELACIONADAS ÀS FONTES DE INFORMAÇÕES DOS BRASILEIROS

Senado Federal: Levantamento sobre *fake news*, redes sociais, privacidade e os principais meios de informação para a população do Brasil em 2019. Disponível em: https://www12.senado.leg.br/institucional/ouvidoria/publicacoes-ouvidoria/redes-sociais-noticias-falsas-e-privacidade-de-dados-na-internets.

Digital News Report: Página com informações resumidas sobre o levantamento mais recente de sites de notícias mais influentes no Brasil (2019). Disponível em: http://www.digitalnewsreport.org/survey/2019/brazil-2019/.

Reuters Institute: Site do Instituto Britânico Reuters conveniado com a Universidade de Oxford, responsáveis pelo desenvolvimento da pesquisa sobre notícias digitais. Disponível em: https://reutersinstitute.politics.ox.ac.uk/.

IPSOS GROUP: Site Ipsos Group S.A. é uma empresa de pesquisa de mercado global e consultoria, responsáveis pelo desenvolvimento de pesquisas sobre *fake news* em 2018. Disponível em: https://www.ipsos.com/pt-br/global-advisor-fake-news.