UNIVERSIDADE FEDERAL DA PARAÍBA CENTRO DE CIÊNCIAS HUMANAS, SOCIAIS E AGRÁRIAS DEPARTAMENTO DE CIÊNCIAS SOCIAIS APLICADAS CURSO DE BACHARELADO EM ADMINISTRAÇÃO

HAROLDO RAIR MELO DOS SANTOS

Previsão do **Retorno Acionário** do Mercado Brasileiro com o uso de **dados textuais, notícias** especializadas do G1 e técnicas de **Aprendizado de Máquina** Supervisionado.

Bananeiras

- PB 2023

Catalogação na publicação
Seção de Catalogação e Classificação

S237p Santos, Haroldo Rair Melo Dos.
Previsão do retorno acionário do mercado brasileiro com o uso de dados textuais, notícias especializadas do G1 e técnicas de Aprendizado de Máquina Supervisionado. / Haroldo Rair Melo Dos Santos. - Bananeiras, 2023.
27 f. : il.

Orientação: Gustavo Correia Xavier Xavier.
TCC (Graduação) - UFPB/CCHSA.

1. MACHINE LEARNING. 2. Previsão. 3. Sentimentos. I. Xavier, Gustavo Correia Xavier. II. Título.

UFPB/CCHSA-BANANEIRAS CDU 658 (042)

Elaborado por BRUNA ISABELLE MEDEIROS DE MORAIS - CRB-15/813

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS HUMANAS, SOCIAIS E AGRÁRIAS
CURSO DE ADMINISTRAÇÃO

ATA DE AVALIAÇÃO DO TCC2

Modalidade: ARTIGO CIENTÍFICO

Aos décimo terceiro dia de novembro do ano de dois mil e vinte e três (13/11/2023), na presença dos professores GUSTAVO CORREIA XAVIER E DANILO RAIMUNDO DE ARRUDA apresentou-se a defesa do **Artigo Científico** do (a) estudante HAROLDO RAIR MELO DOS SANTOS, intitulado Previsão do Retorno Acionário do Mercado Brasileiro com o uso de dados textuais, notícias especializadas do G1 e técnicas de Aprendizado de Máquina Supervisionado, o qual obteve aprovação com nota final 10,00 (dez), conforme o resultado das notas dadas pelos professores, abaixo descritas:

Observação: atribuir notas de 0 a 10 em cada critério.

CRITÉRIOS DE AVALIAÇÃO DO ARTIGO CIENTÍFICO	Avaliador 1	Avaliador 2	Avaliador 3
Introdução: apresentação, justificativa, o problema e os objetivos dapesquisa e estrutura geral do trabalho.	10	10	
Referencial teórico: apresentação da literatura relevante sobre o assunto.	10	10	
Método: apresentação das principais decisões e procedimentos do trabalho de campo, com definição coerente com a opção de pesquisa definida (entre qualitativa e quantitativa).	10	10	
Resultados: apresentação dos resultados do trabalho empírico, juntamente com a discussão dos resultados à luz da construção teórica.	10	10	
Considerações finais: apresentação do fechamento da pesquisa, comretomada dos objetivos e sua análise, assim como as implicações teóricas e práticas da pesquisa e as recomendações de estudos futuros.	10	10	
Referências bibliográficas: apresentação somente dos itens de bibli- ografia efetivamente citados no texto.	10	10	
Apresentação física do trabalho: coerência com as normas	10	10	
Apresentação pública do trabalho	10	10	
Total			
Média	10	10	

Observações da Banca:

Bananeiras, 17 de novembro de 2023

gov.br Documento assinado digitalmente
GUSTAVO CORREIA XAVIER
Data: 17/11/2023 18:36:19-0300
Verifique em https://validar.iti.gov.br

Professor(a) Orientador (a)

Danilo Raimundo de Arruda

Avaliador (b)

SANTOS, H. R. M. **PREVISÃO DO RETORNO ACIONÁRIO DO MERCADO BRASILEIRO COM O USO DE DADOS TEXTUAIS, NOTÍCIAS ESPECIALIZADAS DO G1 E TÉCNICAS DE APRENDIZADO DE MÁQUINA SUPERVISIONADO.** 2023. 26 f. Trabalho de Conclusão de Curso (Graduação em Administração) – Universidade Federal da Paraíba, Bananeiras, 2023.

RESUMO

O objetivo deste estudo foi investigar e prever o retorno acionário do mercado brasileiro medido através dos índices da Bovespa, durante o período de dezembro de 2020 até maio de 2023. Utilizou-se para tal, o uso de dados textuais, das notícias especializadas de economia e política do portal G1, que foram analisados por meio de ferramentas de aprendizado de Máquina Supervisionado. Este estudo se concentra na análise da relação entre o sentimento expresso em notícias financeiras e as flutuações nos preços do mercado acionário brasileiro, tendo como base em uma revisão de literatura que abordou os princípios do mercado eficiente e examinou o comportamento e sentimento dos investidores, além de mostrar modelos de predições e ferramentas de *Web Scraping.* Os dados textuais relacionados a notícias financeiras ao conduzir o estudo e na coleta de informações, resultam em uma amostra de 17.999 notícias, os resultados destacam a predominância de médias negativas em certos períodos, muitas vezes relacionadas a eventos econômicos e políticos significativos, além disso, a análise explorou a evolução dos sentimentos ao longo do tempo, revelando flutuações notáveis e tendências anuais. Este estudo reforça a importância da análise de sentimentos em notícias financeiras como um recurso valioso para prever as tendências do mercado de ações no Brasil e destaca o potencial da combinação de técnicas de *Machine Learning* com análise de texto.

Palavras-Chave: Aprendizado de Máquina; Previsão; Sentimentos

Sumário

1.INTRODUÇÃO.

Nos últimos anos, tem havido um reconhecimento crescente da relevância das previsões de variáveis macroeconômicas, que se tornaram ferramentas amplamente empregadas na antecipação do desempenho empresarial. Essas previsões permitem aos gestores tomar decisões mais informadas e estratégicas, tendo em vista as flutuações macroeconômicas que podem impactar seus negócios. Segundo Godeiro (2018), a previsão de variáveis econômicas é um dos maiores desafios para os economistas, principalmente nos últimos anos, com o surgimento de novas técnicas e dados. Atualmente, temos a influência do *Big Data*, onde a cada minuto são produzidos milhões de informações, sejam elas via publicações em redes sociais, *scanner* data (dados dos consumidores ao efetuarem suas compras), entre outros.

Dentro do contexto literário, essa fonte de informação é referida como dados não estruturados. A sua utilização é caracterizada por meio de uma variedade de informações, originando em um grande armazenamento de elementos, chamado de *Big Data* (SILVA, 2018). Sendo os principais exemplos no mercado a previsão da qual pode ser usada como uma ferramenta de análise preditiva, identificando padrões históricos, probabilidades e a observação de valores de tendência, existem inúmeros campos que lidam com uma imensa quantidade de dados e informações dispersas. A aplicação eficiente de técnicas de *Big Data* permite identificar a lógica subjacente a esse caos, eliminando o ruído informacional desses conjuntos de dados e, assim, encontrando respostas para os diversos dilemas existentes.

Comumente, para realizar essas análises, empregam-se técnicas aprimoradas pelo aprendizado de máquina (*Machine Learning*). Trata-se da capacidade que um equipamento construído pelo homem tem de analisar dados para automatizar a criação de modelos analíticos. É, portanto, uma vertente da inteligência artificial, um conceito mais amplo, que diz respeito à capacidade que uma máquina tem de tomar decisões a partir de um raciocínio que lembra o pensamento humano (FIA, 2021)

No *Machine Learning*, o material de estudo das máquinas são os dados, quanto mais dados alimentam os sistemas, mais perguntas serão feitas, e mais respostas surgirão para solucionar problemas. É por isso que o *Machine Learning* alcança seu pleno potencial com o *Big Data*, o armazenamento e processamento de volumes gigantescos de dados (NEIL PATEL, 2019).

Recentemente, alguns pesquisadores tem se dedicado a capturar o sentimento por meio de *Big data*, com o objetivo de avaliar o impacto desse sentimento sobre o mercado. Silva (2017) buscou identificar como o sentimento extraído de notícias tinha efeitos sobre o preço, que representa o retorno do mercado brasileiro, e a volatilidade, que representa o risco. A oscilação dos preços das ações pode ser influenciada por uma série de fatores, incluindo informações divulgadas na mídia, notícias que envolvem eventos como mudanças na política econômica, a divulgação de resultados financeiros de empresas, escândalos corporativos, entre outros, podem ter um impacto significativo no mercado acionário. Por exemplo, se uma empresa divulga um resultado financeiro melhor do que o esperado, suas ações podem valorizar-se, da mesma forma, se uma notícia negativa sobre uma empresa é divulgada na mídia, suas ações podem sofrer uma desvalorização.

Além disso, autores como Fama (1970) e Roberts (1967), procuraram identificar o que afeta os preços na bolsa de valores, e ainda buscaram detectar mudanças em seus padrões, diante disso, provar esse processo da estruturação dos preços, não é uma tarefa simples. A economia, a política e outros eventos macroeconômicos também podem afetar o comportamento do mercado acionário, por isso, muitos estudiosos acreditam que acompanhar as notícias e informações divulgadas na mídia é importante para compreender a dinâmica do mercado acionário e identificar possíveis oportunidades de investimento. O estudo do efeito do sentimento textual das notícias nos mercados financeiros é uma área de interesse crescente, que busca entender melhor como as notícias e outras informações divulgadas na mídia podem afetar o comportamento dos investidores e, consequentemente, o desempenho do mercado.

Nesse sentido, este estudo pretende investigar se o tom emocional das notícias influencia o comportamento dos investidores e, consequentemente, o desempenho do mercado de ações, com técnicas de Aprendizado de Máquinas. A partir dessa análise, os resultados podem contribuir para a compreensão dos fatores que afetam o mercado acionário e, assim, auxiliar investidores na tomada de decisões. Os resultados indicaram que o sentimento textual das notícias dos cadernos de economia e política do G1 tem relação significativa com o retorno acionário do mercado brasileiro. A relação entre o tom das notícias e o desempenho do índice bovespa é um tópico amplamente discutido no campo da economia comportamental e da análise financeira, vários estudos e pesquisadores exploraram essa relação e apresentaram análises sobre o impacto das notícias com tom negativo e positivo no mercado financeiro. (SHILLER, 2000; STATMAN, 2010; LO ANDREW, 2017). Diante disso, espera-se que

notícias com tom negativo estão associadas a quedas no índice Bovespa, enquanto notícias com tom positivo estão associadas a altas no índice, além disso, as notícias com tom mais intenso, tanto positivo quanto negativo, apresentaram uma relação mais forte com o desempenho do mercado.

Esses resultados sugerem que as notícias especializadas em economia e política podem ter um papel importante na formação das expectativas dos investidores e, consequentemente, na volatilidade do mercado acionário brasileiro. Portanto, é importante que investidores e analistas considerem o efeito do sentimento textual das notícias ao tomarem suas decisões de investimento. O presente estudo está organizado em seções que proporcionam uma visão abrangente do tema. Iniciando por introduzir a problemática, justificativa e objetivos, delineando a motivação da pesquisa, prosseguindo para aprofundar-se na fundamentação teórica, oferecendo uma base sólida para compreensão do contexto. Dando seguimento são detalhados os aspectos metodológicos, apresentando o modelo e os dados utilizados. E por fim a transição para revelar resultados, acompanhados de análises, enriquecendo a compreensão. O trabalho encerra-se, destacando as referências bibliográficas.

2. REFERENCIAL TEÓRICO

2.1 Hipótese do Mercado Eficiente HME

Um dos tópicos mais importantes dentro da teoria de finanças, é a HME (Hipótese do Mercado Eficiente). A eficiência de mercado é um conceito fundamental na literatura financeira, uma vez que a sua função exerce um papel decisivo na concepção de modelos financeiros e na realização de estudos voltados ao mercado de capitais. Esse estudo foi idealizado e proposto por Fama (1970), trazendo à tona que a precificação dos ativos financeiros espelha integralmente todo o conteúdo de informações que são divulgadas.

Fama (1970) destaca ainda que existem três níveis de eficiência: fraca, semi-forte e forte. No mercado com eficiência fraca, todas as informações públicas históricas estão refletidas nos preços dos ativos financeiros. No mercado com eficiência semi-forte, além das informações públicas históricas, novas informações são rapidamente incorporadas nos preços dos ativos. Já no mercado com eficiência forte, os preços dos ativos refletem tanto as informações públicas quanto as informações ocultas ou privilegiadas. Esses três níveis de

eficiência propostos por Fama (1970) são amplamente discutidos na literatura financeira e são fundamentais para o desenvolvimento de modelos de previsão e estratégias de investimento.

Diante desse cenário, existem autores que buscam explicar em outras vertentes o mercado de capitais.Kahneman e Tversky (1979) vão contra a HME quando se fala sobre a racionalidade dos investidores, os autores defendem que os investidores são irracionais e que suas emoções afetam e interferem a sua tomada de decisão, além de mostrarem a ideia que os seres humanos possuem racionalidade limitada, e que tais fatores comportamentais e psicológicos na tomada de decisão podem influenciar o retorno das ações.

Outros fatores que afetam o mercado e podem servir como formas para se prever são as notícias transmitidas pelas mídias, a ascensão das mídias sociais criou oportunidades para investidores compartilharem suas opiniões e visões de mercado sobre empresas de capital aberto, o que se tornou um campo promissor para capturar o sentimento do investidor por meio de suas publicações e relacioná-lo com o retorno das ações.

2.2 Previsão de retorno acionário

A previsão do retorno acionário tem sido objeto de estudo de muitos pesquisadores, e o uso de dados textuais tem se mostrado promissor para melhorar a precisão dos modelos de previsão. Segundo Ribeiro e Lima (2020), as notícias financeiras veiculadas pela imprensa são uma importante fonte de informações para prever o comportamento do mercado acionário. Além disso, o uso de técnicas de *Machine Learning* supervisionado, como redes neurais artificiais e árvores de decisão, tem se mostrado eficaz na análise e interpretação desses dados textuais.

No contexto brasileiro, Souza e Silva (2019) realizaram um estudo sobre o uso de dados textuais para prever o retorno de ações na Bolsa de Valores de São Paulo (B3). Os autores utilizaram dados do Twitter para identificar o sentimento do mercado em relação a determinadas empresas e, em seguida, aplicaram técnicas de análise de sentimentos e regressão para prever o retorno das ações. Os resultados indicaram que a inclusão de informações textuais melhorou significativamente a precisão do modelo de previsão.

Outro estudo relevante é o de Morais et al. (2017), que utilizaram dados textuais de notícias para prever o retorno de ações na B3. Os autores aplicaram técnicas de mineração de

textos e *Machine Learning* para extrair informações relevantes das notícias e, em seguida, construíram um modelo de previsão baseado em regressão linear múltipla. Os resultados indicaram que a inclusão de informações textuais melhorou a precisão do modelo de previsão em comparação com modelos que não incluíam essas informações.

Diante disso, é possível verificar que o uso de dados textuais e técnicas *Machine Learning* pode ser uma abordagem promissora para a previsão do retorno acionário no mercado brasileiro.

2.3 Sentimento Textual

Godeiro (2018) tem adotado o índice de sentimento textual como um indicador preditivo inovador. Esses índices são baseados em uma abordagem linguística que examina as palavras encontradas em uma fonte de informação específica, e por meio de técnicas especializadas, avalia subjetivamente o conteúdo textual. Essa abordagem permite capturar nuances emocionais e subjetivas presentes nos textos, fornecendo uma dimensão adicional para a análise financeira.

Godeiro (2018) apresenta uma abordagem inovadora para a extração das informações mais relevantes e preditivas das minutas do FED, ao invés de se basear em um dicionário com um conjunto fixo de palavras, propõe-se a construção de um dicionário adaptável, capaz de se atualizar ao longo do tempo. Esse processo é realizado utilizando técnicas de *Machine Learning*, que permitem identificar as palavras mais preditivas dentro de uma determinada minuta. Essas palavras selecionadas são então utilizadas para gerar novos preditores, aumentando assim a capacidade de previsão do modelo.

Por meio da aplicação de técnicas de *Machine Learning*, é possível utilizar algoritmos para identificar as raízes de palavras ou termos mais preditivos presentes em uma determinada notícia. Esses conteúdos identificados como mais relevantes podem ser utilizados para derivar novos preditores que podem melhorar a acurácia na previsão do prêmio de risco, ao incorporar esses novos preditores, observa-se um aprimoramento estatisticamente significativo na capacidade de previsão. Essa abordagem permite explorar de forma mais precisa e eficiente as informações contidas nas notícias, contribuindo para aprimorar as estratégias de previsão e tomada de decisão no mercado financeiro.

2.4 Modelos de Predição

Os modelos de predição Ridge e Lasso são técnicas populares na área de *Machine Learning* e análise de dados. Eles são utilizados para lidar com problemas de multicolinearidade e seleção de variáveis, melhorando a precisão das previsões.

Um dos estudos importantes sobre o modelo de predição Ridge é o de Hoerl e Kennard (1970). Eles introduziram o método Ridge, também conhecido como regressão Ridge, como uma técnica para lidar com problemas de multicolinearidade. Eles utilizaram um termo de regularização na função objetivo, que adiciona uma penalidade às estimativas dos coeficientes de regressão, o que permite reduzir a variância dos estimadores, tornando-os menos sensíveis às perturbações nos dados. Outro estudo relevante sobre o modelo de predição Ridge é o de Friedman et al. (2001). Nesse estudo, intitulado "*Elements of Statistical Learning*", os autores apresentam uma abordagem abrangente dos métodos estatísticos e de *Machine Learning*, incluindo a regressão Ridge.

No caso do modelo de predição Lasso, um dos estudos fundamentais é o de Tibshirani (1996). O autor introduz o método Lasso como uma técnica para realizar regressão com seleção automática de variáveis, é uma extensão do modelo Ridge que além de lidar com a multicolinearidade, também realiza a seleção automática de variáveis. Ele utiliza um termo de regularização que promove a diminuição dos coeficientes de algumas variáveis para zero, eliminando-as do modelo.

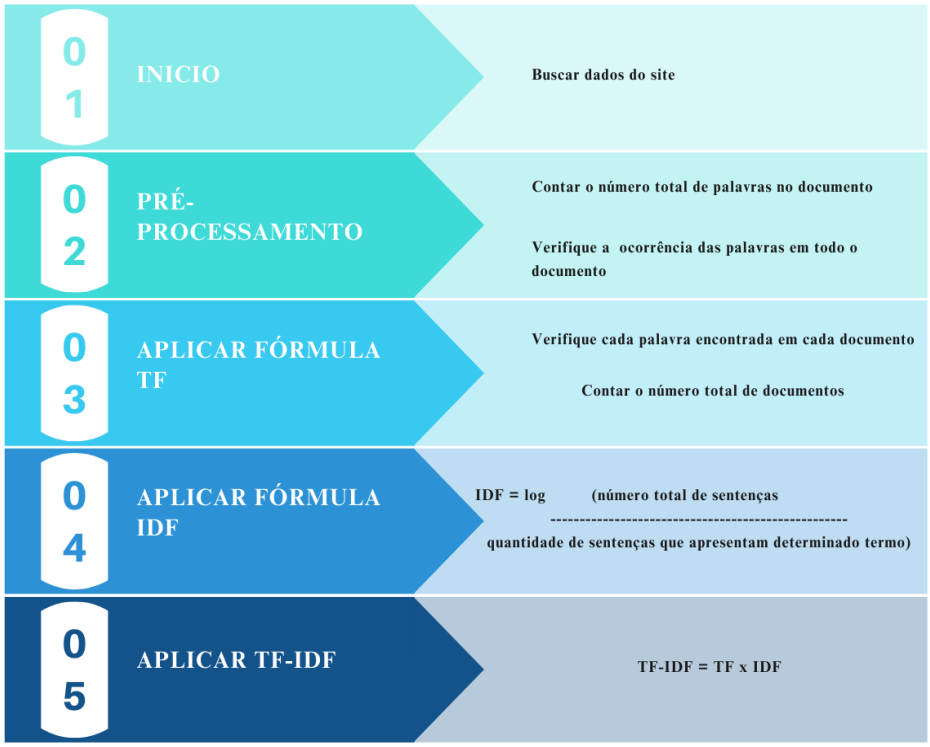
Um estudo mais recente que discute a aplicação do modelo de predição Lasso é o de Zou e Hastie (2005). Os autores apresentaram uma extensão desse modelo chamada de *Elastic Net*, que combina a penalidade do Lasso com a penalidade de Ridge. Essa abordagem foi desenvolvida para superar algumas limitações do Lasso, especialmente quando há alta correlação entre as variáveis preditoras, enquanto o Lasso tende a selecionar apenas uma das variáveis altamente correlacionadas e eliminar as outras, o *Elastic Net* permite que um grupo de variáveis correlacionadas seja selecionado ao mesmo tempo.

2.5 TF-IDF

A aplicação da ferramenta TF-IDF (*Term Frequency — Inverse Data Frequency*), se dá com o foco de analisar, mensurar e categorizar documentos. Onde o TF: mede a frequência de uma palavra em cada documento, sendo a proporção de suas ocorrências em relação ao total

de palavras, à medida que a contagem da palavra cresce no documento, o valor do TF também aumenta. Já o IDF: é utilizado para determinar a ponderação de termos menos comuns em toda a coleção de documentos. Termos de baixa frequência no corpus recebem uma pontuação elevada de IDF (TRIPATHI, 2018).

Figura 1: Processo TF-IDF (passo a passo)



Fonte: Autoria própria.

A extração dos documentos é feita utilizando *Web Scraping*, e após essa extração ser feita a aplicação do TF-IDF é usada.

2.6 Web Scraping

Os *web scrapers* são ferramentas poderosas para coletar e processar grandes volumes de dados, proporcionando uma abordagem eficiente e escalável. Ao contrário da abordagem tradicional de navegar e acessar uma página de cada vez, os scrapers têm a capacidade de visualizar e acessar milhares de páginas simultaneamente. Essa capacidade de coletar dados em escala permite uma análise abrangente e uma visão mais completa do conteúdo disponível na *web*. Com o uso de *web scrapers*, é possível extrair informações valiosas de uma ampla variedade de fontes e aproveitar ao máximo o potencial dos dados *online*. Segundo Mitchell

(2015), o *Web Scraping* é a prática de coletar dados de forma automatizada, por meio da programação que torna possível desenvolver programas que requisitam dados a servidores *web* e os analisam para extrair as informações desejadas.

> O Web Scraping "É essencialmente extrair e reunir conjuntos de dados da web (o que pode ser considerado Big Data em alguns casos), dados esses que são a pedra angular do Big Data Analytics, Machine Learning e Inteligência Artificial. Esses dados podem ser usados em projetos de Data Science para resolver problemas de negócio específicos e ajudar os tomadores de decisão, podendo trazer vantagem competitiva." (DATA SCIENCE ACADEMY, 2018).

Além de suas capacidades limitadas, os dispositivos de busca tradicionais não são capazes de oferecer acesso a uma ampla gama de informações disponíveis na *web*. Enquanto os sites de busca fornecem resultados baseados em anúncios e *sites* populares, essas informações representam apenas uma fração do vasto conteúdo disponível. Por outro lado, os *scrapers* possuem a habilidade de extrair informações de fontes variadas e menos conhecidas, permitindo um acesso mais abrangente e detalhado aos dados disponíveis *online*, como explica Mitchell (2015), que segundo ele, um web scraper bem desenvolvido pode colocar em um gráfico o custo de um voo para Boston ao longo do tempo para uma variedade de sites e informar qual vai ser o melhor momento para comprar uma passagem.

3. METODOLOGIA

A base deste estudo consistiu na previsão do retorno acionário e analisar a relação entre o sentimento das notícias financeiras e o uso de dados textuais com a utilização de técnicas de *Machine Learning*. Para isso, foram utilizados dados do mercado nacional e notícias relacionadas a economia e política extraídas do portal de notícias do G1, os dados coletados estão entre o período de dezembro de 2020 à maio de 2023. A fim de compreender o comportamento textual das notícias, foram aplicadas técnicas estatísticas utilizando o ambiente de programação do R Studio e o dicionário sentiLex_PT02 de SILVA (2010), usado através do pacote LexinconPT do R.

Além disso, foi realizado um pré-processamento das notícias utilizando o pacote tidytext do R, removendo palavras como artigos, preposições e pontuações que não contribuem para a análise em questão. Para isso, foi utilizada a biblioteca BeautifulSoup do Python, que permitiu realizar o parse dos *sites*. A pesquisa foi conduzida realizando a análise dos

documentos por meio do processamento das *tags* HTML. Essa abordagem possibilitou a obtenção de informações relevantes contidas nas publicações, que serviram como base para a análise posterior.

Com a ideia de analisar e mostrar a previsão do retorno acionário do mercado brasileiro com o uso de dados textuais de notícias especializadas do G1 e técnicas de *Machine Learning* supervisionadas, todo o processo e instrumentos metodológicos foram divididos em várias etapas, que são elas:

Coleta de dados: Inicialmente, foram coletadas as notícias especializadas pelo portal G1 relacionadas a economia e política. Essa etapa utilizou técnicas de *web scraping*.

Pré-processamento dos dados textuais: Em seguida, foi utilizado o pacote *tidytext* do R. Os dados textuais são submetidos a técnicas de pré-processamento, como remoção de *stopwords*, lematização ou *stemming*, remoção de caracteres especiais e normalização do texto. Isso visa reduzir o ruído e padronizar o formato dos dados textuais.

Extração de características: Com os dados pré-processados, é realizada a extração de características (*features*) relevantes dos textos. Inicialmente foi aplicado o TF-IDF (*Term Frequency-Inverse Document Frequency*), um método que atribui um peso para cada palavra de acordo com a frequência. Dessa forma, palavras muito frequentes ou muito raras têm uma menor relevância. Adicionalmente, o *score* final do TF-IDF foi multiplicado pela polaridade de cada palavra conforme o dicionário sentiLex_PT02 (-1 para palavras negativas e +1 para os demais casos). O resultado da multiplicação foi uma medida de sentimento para cada termo.

Estimativa das medidas de Sentimento: Como forma de reduzir o número de características extraídas no passo anterior, foi estimado duas medidas de sentimento, SentNews e SentNewsIBOV. A primeira medida (SentNews) é a média da pontuação de cada palavra contida nas notícias do dia. A segunda usa o modelo *XGBoost*, uma técnica de machine learning, para prever o retorno acionário a partir das características (*features*) extraídas no item anterior. Dessa forma, as palavras menos relevantes para previsão tendem a ser descartadas ao longo do tempo, ficando assim apenas os termos mais importantes. Após a estimação do modelo, é possível analisar as palavras mais relevantes bem como utilizar o valor previsto pelo o modelo como um sentimento textual (SentNewsIBOV) que é baseado na relação entre as palavras e o retorno acionário até o dia anterior.

Construção dos modelos de Machine Learning: Após a criação das medidas de sentimento textual que resume as *features* baseadas em palavras, é construído um modelo de previsão no qual a média histórica do retorno, o SentNews e o SentNewsIBOV são utilizados para prever o retorno acionário do dia seguinte (um passo à frente) com o uso de diversos modelos de *Machine Learning*, incluindo algoritmos como Regressão Linear, Árvores de Decisão, *Random Forest, Gradiante Boost Machine (GBM) e Extreme Gradient Boost Machine (XGBoost).*

Treinamento e validação do modelo: Todos os modelos são treinados utilizando um conjunto de dados históricos contendo informações do retorno acionário. Nos dados de treinamento foi feita validação do modelo utilizando a técnica de validação cruzada como forma de evitar o *overfitting*, e por fim realizado o teste para avaliar a sua capacidade de generalização.

Avaliação do desempenho: Para avaliação do desempenho do modelo foi calculado métricas apropriadas para problemas de regressão, como erro médio absoluto (MAE), erro quadrático médio (RMSE), Percentual Médio de Erro Absoluto (MAPE) ou coeficiente de determinação (R^2). Essas métricas permitem avaliar quão bem o modelo é capaz de prever o retorno acionário com base nos dados textuais. A maior parte da análise se concentrou no RMSE, mas as outras medidas corroboram a conclusão do RMSE.

Análise e interpretação dos resultados: Por fim, os resultados obtidos foram analisados e interpretados para entender o impacto das características textuais na previsão do retorno acionário. Adicionalmente para entender se a técnica utilizada é capaz de superar o *benchmarking* que se trata de um modelo simples (*naive*) baseado apenas na média histórica. Essa análise pode auxiliar na identificação de padrões e insights sobre como as notícias especializadas do G1 influenciam o mercado financeiro.

4 ANÁLISE DOS RESULTADOS E DISCUSSÃO

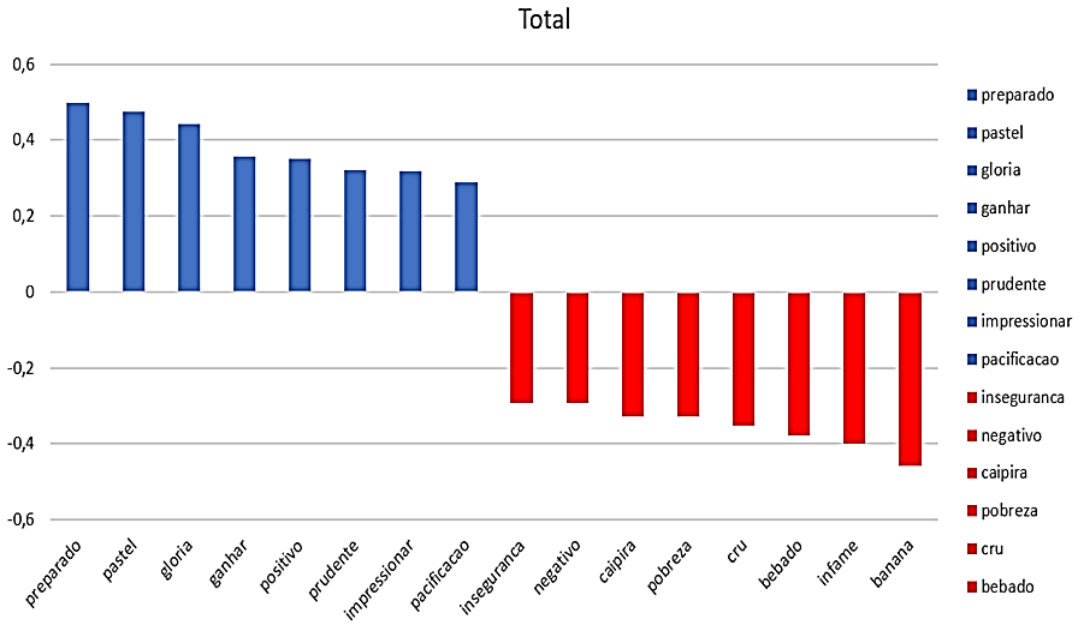
Os resultados apresentados neste estudo foram cuidadosamente obtidos por meio da análise dos sentimentos expressos nas notícias veiculadas pela imprensa especializada. Esta análise se concentrou de forma particular no desempenho e nas consequências refletidas no índice Ibovespa, abordando os sentimentos relacionados à conjuntura econômica e ao ambiente político nacional.

4.1 O tom e a importância das palavras

Essa subseção analisa o sentimento das palavras calculado a partir do peso TF-IDF multiplicado pela polaridade de cada palavra no Dicionário Lexicon PT-BR. Em outras palavras trata-se de uma análise do tom (positivo ou negativo) e da importância de cada palavra (peso TF-IDF). A exposição a seguir apresenta uma seleção representativa dos termos abordados nas notícias financeiras coletadas entre dezembro de 2020 e maio de 2023, destacando as palavras respectivamente de maior e menor relevância. Esse gráfico proporciona uma síntese visual dos termos associados a comportamentos pessimistas e otimistas, oferecendo uma análise quantitativa da importância de cada palavra ou termo.

A análise dos dados coletados, como ilustrado no Gráfico 1, revela a proeminência de termos que desempenham um papel significativo na medição dos sentimentos, com destaque para palavras como “preparado”, “ganhar”, “positivo” e “impressionar”, que denotam conotações positivas. É destaque que as palavras com sentimento textual positivo têm o percentual mais alto se comparados com as médias negativas.

Gráfico 1 – Relevância das palavras fundamentada nas notícias extraídas de economia e política doportal G1.



Fonte: Elaboração própria.

Focando nos termos negativos, pode-se observar as palavras que mais impactam diretamente com a sua relevância, “infame”, “pobreza” e “insegurança”, tais termos tem uma

implicação negativa mensuradas nos sentimentos. Vale salientar que os termos como “banana” que se destaca com alta nas palavras negativas, “cru”, “pastel” e “glória” (esta última que por sua vez apareceu três vezes, dentre os cinco termos mais frequentes positivos entre a coleta), podem ser justificadas por repetirem muito Portanto é natural que palavras desse tipo sempre apareçam durante a coleta e análise de dados.

Comparado a estudos anteriores, Oliveira (2022) observa uma frequência mais elevada de palavras ao longo do tempo, com um aumento notável na quantidade de termos em períodos mais recentes. Isso pode ser atribuído a várias razões, como a utilização de um método de *Web Scraping* mais atualizado, capaz de capturar dados mais recentes. Além disso, à medida que o tempo avança, mais termos e informações são gerados diariamente, e portais de notícias como o G1 expandem seu portfólio de artigos e notícias. Esse aumento no volume de dados pode ser percebido de maneira mais acentuada no contexto deste estudo, e, ao analisar um período mais longo, entre 5 a 10 anos, é provável que seja mais fácil destacar essa tendência.

4.2 Sentimento Textual das Notícias (SentNews)

Na tabela 1 apresenta os dados de notícias obtidas durante o período de dezembro de 2020, até dezembro de 2021, mostrando a média dos índices de sentimentos tanto positivos quanto negativos. Ao examinar os dados coletados, notou-se que o sentimento textual proveniente das fontes de mídia revelou uma média negativa durante o final 2020 de (-0,115), vale ressaltar que a média dos sentimentos de dezembro de 2020 é válida apenas para os últimos dias do mês, não cobrindo período inteiro como um todo. E partindo agora para o ano de 2021 observa-se que a média negativa fica entre os meses de março com (-3,313), com o maior pico negativo, e abril com o resultado de (-1,127), enquanto os demais meses do ano ficam com uma média positiva.

Tabela 1 – Média dos sentimentos, fundamentada nas notícias extraídas de economia e política do portal G1.

Datas	Sentimentos
dez/2020	-0,1154
jan/2021	0,1834
fev/2021	0,2223
mar/2021	-3,3130
abr/2021	-1,1279
mai/2021	0,2226
jun/2021	1,9637

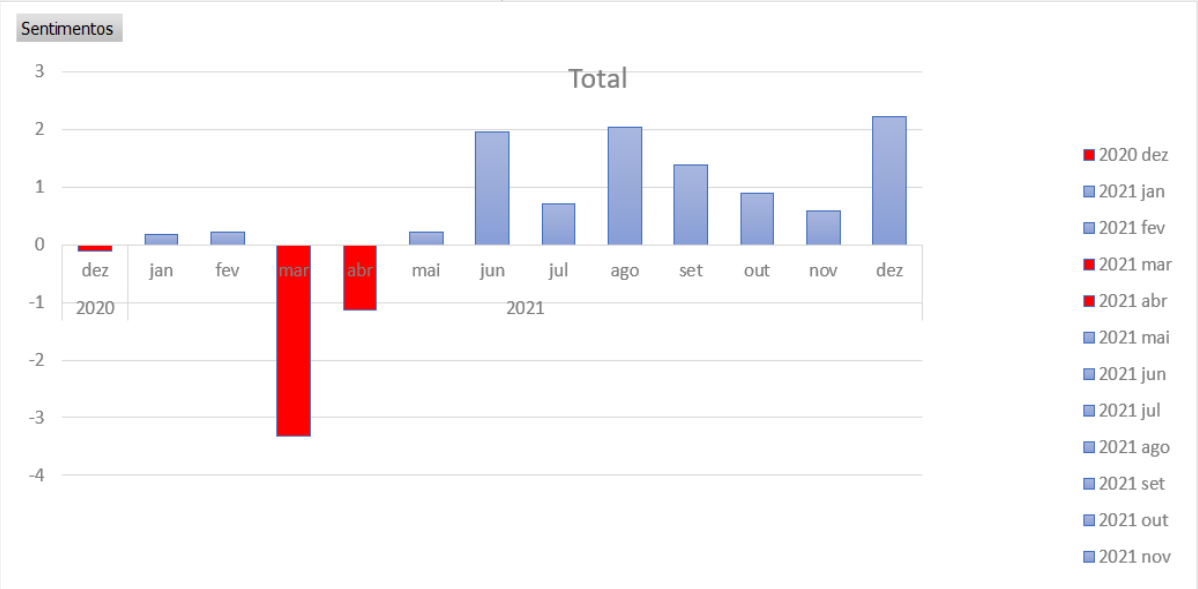
jul/2021	0,7080
ago/2021	2,0398
set/2021	1,3932
out/2021	0,9008
nov/2021	0,5946
dez/2021	2,2182

Fonte: Elaboração própria.

É pertinente salientar que os pontos altos, caracterizados pelas médias mais positivas, se destacam nos meses de junho, registrando (1,963), seguido por agosto, com (2,039), e atingindo o pico mais significativo em dezembro, alcançando (2,218), como demonstrado na Tabela 1. Diante desses resultados, torna-se evidente que o ano de 2021 manteve consistentemente um patamar de sentimentos positivos, com base nas médias extraídas das fontes de mídia.

Para uma compreensão mais nítida, o Gráfico 2 visualiza a média de sentimentos ao longo dos períodos mencionados anteriormente. Ao analisar as flutuações do índice de sentimento e seu comportamento, observa-se que as variações de natureza pessimista nas notícias são particularmente acentuadas, com maior destaque para o período relacionado aos meses de março e abril. Durante esse mesmo intervalo, é possível destacar eventos de impacto mundial, como a escalada da pandemia de COVID-19, com março e abril representando os meses mais críticos e com maior número de fatalidades no país.

Gráfico 2 – Média dos sentimentos, fundamentada nas notícias extraídas de economia e política do portal G1.



Fonte: Elaboração própria.

No que tange ao período impactado pela pandemia, é pertinente considerar que o otimismo nos sentimentos, que se manifestou a partir do mês de maio, coincide com a distribuição das doses das vacinas aos estados brasileiros. Posteriormente, nos meses subsequentes, a taxa de mortalidade apresentou um declínio significativo como resultado direto desse processo, uma vez que até dezembro, aproximadamente 90% da população já havia recebido a primeira dose da vacina.

Ao examinarmos o panorama abrangente, notamos que, apesar de um ano caracterizado por uma predominância de médias de sentimentos positivos e uma relativa estabilidade a partir da segunda metade do ano, as médias negativas continuam a se destacar, mantendo o ponto mais alto do ano.

Na Tabela 2, que engloba o ano de 2022, observa-se um cenário em que, apesar de começar o ano com uma média de sentimentos positiva em janeiro, registrando (1,506), somente em agosto a média retorna a um valor positivo, atingindo (0,532). Isso contrasta com o ano anterior, onde as médias positivas eram mais frequentes.

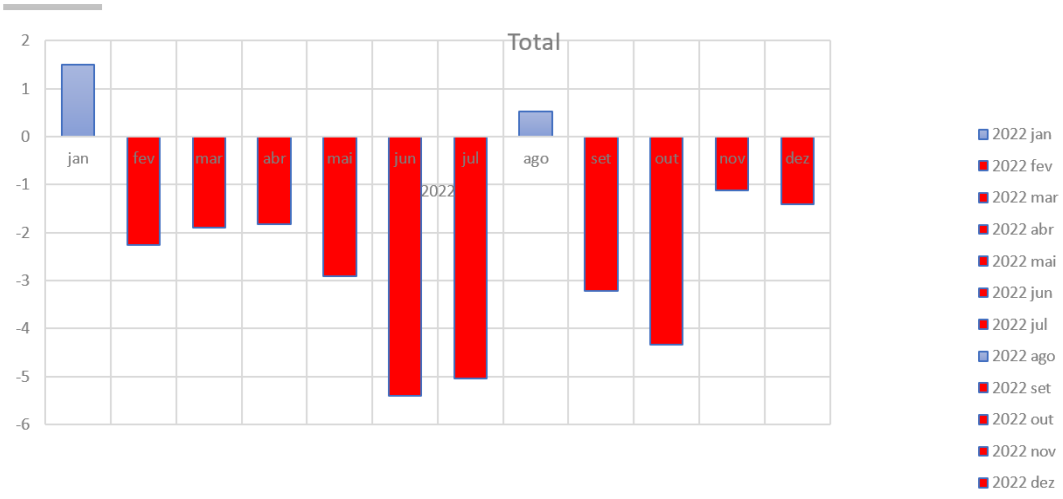
Tabela 2 – Média dos sentimentos, fundamentada nas notícias extraídas de economia e política do portal G1.

Datas	Sentimentos
jan/2022	1,5062
fev/2022	-2,2560
mar/2022	-1,8938
abr/2022	-1,8248
mai/2022	-2,9020
jun/2022	-5,3937
jul/2022	-5,0343
ago/2022	0,5322
set/2022	-3,2171
out/2022	-4,3425
nov/2022	-1,1193
dez/2022	-1,4026

Fonte: Elaboração própria.

No contexto do Gráfico 3, referente ao ano de 2022, merece destaque a presença dos picos mais acentuados de sentimentos negativos nos meses de junho, com (-5,393), e julho, com (-5,034), representando os pontos mais elevados de negatividade ao longo do ano. Ao examinar os sentimentos evidenciados, é possível notar que os comportamentos pessimistas demonstram uma maior frequência na ocorrência de picos elevados em comparação aos comportamentos de sentimentos otimistas.

Gráfico 3 – Média de Sentimentos, fundamentada nas notícias extraídas de economia e política do portal G1.



Fonte: Elaboração própria.

Portanto, ao examinar o cenário de 2022, torna-se evidente que as médias negativas prevaleceram durante o ano. Mesmo com a retomada das atividades presenciais para a maioria das pessoas, esse período foi caracterizado por significativos impactos, incluindo conflitos entre Ucrânia e Rússia, devastação na Floresta Amazônica e eventos climáticos catastróficos. Além disso, o ano foi marcado por uma eleição presidencial, fatores que influenciaram de forma abrangente a política e a economia, gerando uma variedade de sentimentos.

Em relação a 2023, conforme os dados apresentados na Tabela 3 e de acordo com as informações coletadas das notícias, vale ressaltar que a extração de dados se estendeu apenas até o mês de maio, coincidindo com o início deste estudo. Diante dos meses analisados, é notório que mais um ano se destaca pela predominância de médias negativas de sentimentos, com a ausência de sequer uma média positiva.

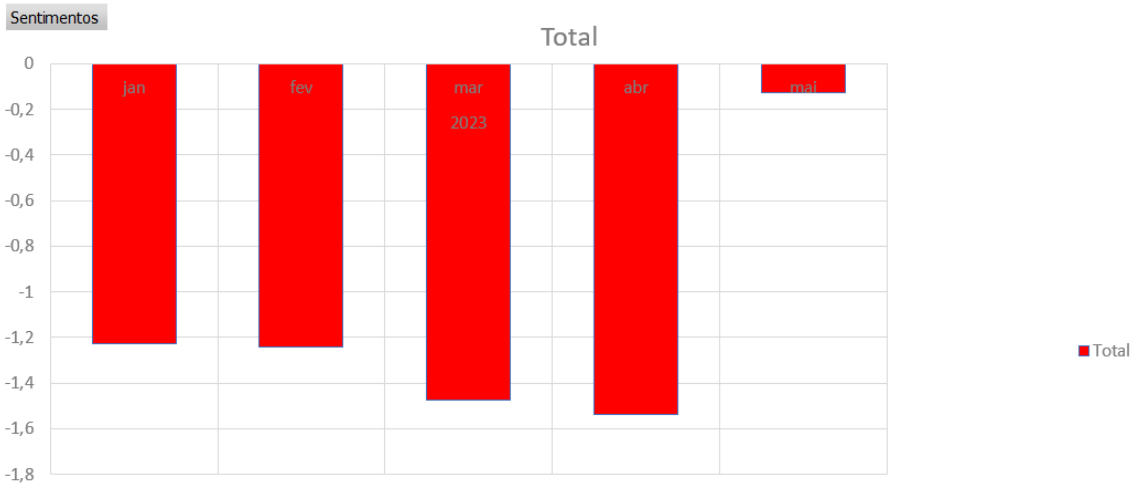
Tabela 3 – Média dos sentimentos, fundamentada nas notícias extraídas de economia e política do portal G1.

Datas	Sentimentos
jan/2023	-1,2281
fev/2023	-1,2432
mar/2023	-1,4749
abr/2023	-1,5371
mai/2023	-0,1263

Fonte: Elaboração própria.

Ainda com base na tabela anterior observa-se que as médias dos sentimentos negativos oscilaram de forma mínima e se mantiveram mais constantes, e isso pode ser observado de forma mais clara no gráfico 4, então tem-se o destaque para abril com a média de (-1,53), se mantendo assim como o maior pico negativo, e o mínimo sendo o mês seguinte, maio com (-0.126).

Gráfico 4 – Média dos sentimentos, fundamentada nas notícias extraídas de economia e política do portal G1.



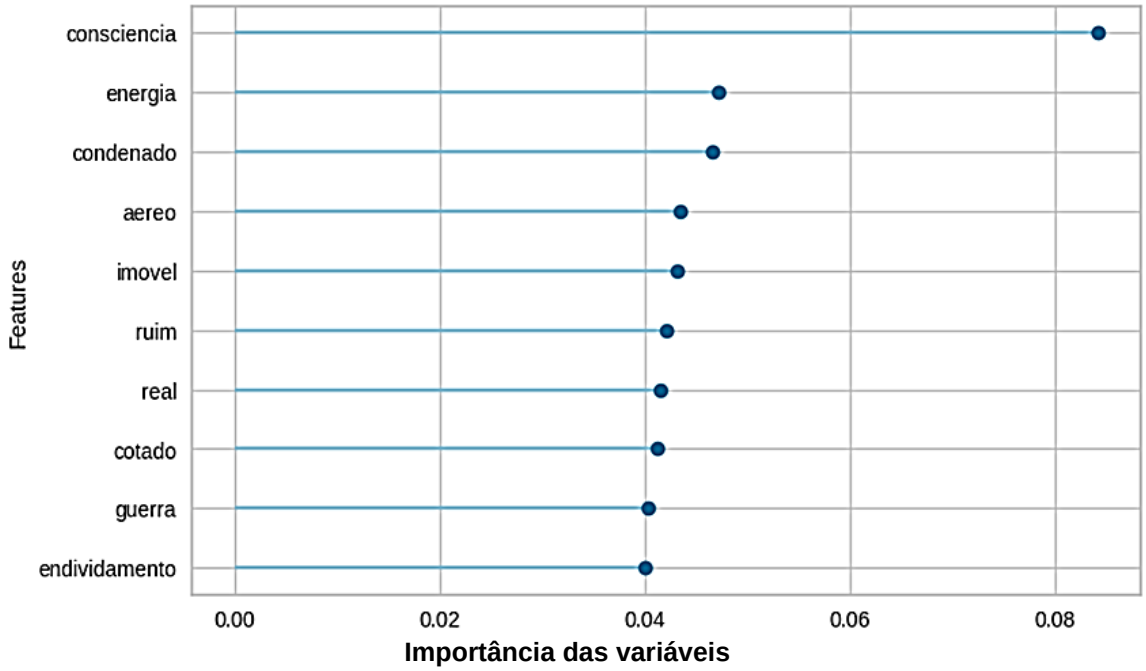
Fonte: Elaboração própria.

Diante das informações apresentadas, é possível discernir uma tendência desfavorável ao examinar os meses individualmente, culminando em seu ápice em abril. Essa trajetória denota um início de ano marcado por incertezas políticas, que remontam ao ano anterior, juntamente com desafios econômicos, tais como a crescente inflação e os esforços em busca de medidas de estímulo econômico, mas que ainda segue com saldo positivo em relação às médias de sentimentos pessimistas do ano de 2022.

4.3 Sentimento Textual condicional ao IBOVESPA

Nessa subseção é feita uma análise do Sentimento Textual das Notícias Condicional ao Índice Bovespa (SentNewsIBOV). Em outras palavras, analisa uma medida de sentimento que usa técnicas de *machine learning* para aprender ao longo do tempo, quais palavras são mais úteis em explicar o retorno acionário. Para tanto, foi estimado um modelo *XGBoost*, que é um tipo de modelo de árvore de decisão. Nessa etapa foram utilizados apenas os dados textuais para tentar explicar o retorno, mas sem agregação, ou seja, cada palavra foi colocada como uma variável preditora formando um modelo com mais de 1000 preditores (características ou *features)*. Na Figura 2 é possível analisar as palavras mais relevantes na explicação do retorno.

Figura 2 – Importância das características, fundamentada nas notícias extraídas de economia e política do portal G1.



Fonte: Elaboração própria.

Especificamente, a Figura 2 apresenta as 10 palavras mais importantes para prever o índice Bovespa. Observou-se um destaque notável para a palavra "consciência". Essa palavra se sobressai consideravelmente em comparação com as outras palavras extraídas. Logo em seguida, temos as palavras "energia" e "condenado", que aparecem com destaque próximas umas das outras.

O destaque de certas palavras, como "consciência", pode estar relacionado a eventos ou tendências específicas que ocorreram nesse período e que impactaram a previsão do mercado de ações. Além das palavras, outros fatores, como indicadores econômicos, eventos políticos ou eventos de mercado, também podem influenciar a explicação do retorno acionário. Essas variáveis podem ter um impacto mais forte do que palavras específicas.

Ainda entre as palavras mais importantes estão os termos "guerra" e "endividamento", que remetem a temas discutidos recentemente e que provavelmente afetaram o movimento do mercado acionário, respectivamente os efeitos globais da Guerra da Ucrânia e o risco local do endividamento público brasileiro.

Após a estimação do *XGBoost*, foi utilizado o valor previsto pelo modelo como um sentimento textual (SentNewsIBOV) para ser incluído no modelo final de previsão.

4.4 Análise de Previsão do Retorno Acionário

Após estimar as medidas de sentimento, foi feita a análise de um modelo final de previsão. Para isso foi testado vários modelos de *Machine learning* para problemas de regressão, especificamente tentando prever o retorno do Ibovespa do dia seguinte a partir de três variáveis preditoras: Média Histórica dos Retornos, SentNews e SentNewsIBOV. A tabela 4 apresenta métricas de desempenho de diferentes modelos de regressão em um contexto de previsão, com o foco principal na métrica de erro RMSE.

Tabela 4 – Métricas de Previsão, fundamentada nas notícias extraídas de economia e política do portal G1.

Modelo	RMSE	MAE	MSE	RMSLE	MAPE
Modelo Base	0,0127	0,0101	0,0002	0,0118	1,2530
XGBoost Tunado	0,0120	0,0096	0,0001	0,0110	1,4177
Ridge	0,0121	0,0097	0,0001	0,0113	1,3101
Linear Regression	0,0122	0,0098	0,0002	0,0101	2,0994
Random Forest	0,0125	0,0100	0,0002	0,0091	3,8143
Lasso	0,0127	0,0102	0,0002	0,0123	1,0828
Elastic Net	0,0127	0,0102	0,0002	0,0123	1,0828
Gradient Boosting	0,0127	0,0100	0,0002	0,0095	3,5485
XGboost	0,0135	0,0107	0,0002	0,0096	5,2683

Fonte: Elaboração própria.

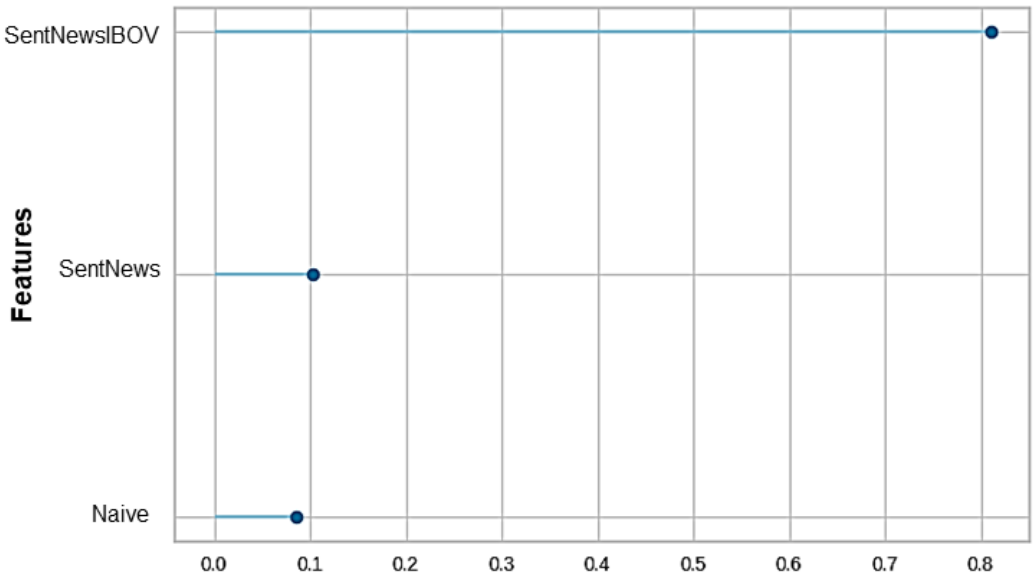
Analisando a tabela 4, entre os modelos listados, o modelo *XGBoost* tunado apresentou os menores valores de MAE (0,0096) e RMSE (0,0120), indicando que ele tem a menor magnitude média dos erros e o menor erro quadrático médio em suas previsões, respectivamente. Esses valores indicam que nosso modelo apresentou um desempenho melhor que o *benchmark* dado pela média histórica do IBO. Além disso, o *XGBoost* Tunado exibiu um MAPE relativamente baixo (1,4177), o que sugere que ele tem um bom desempenho nas previsões em termos percentuais.

Destacando outros modelos, o *ridge* também exibe um RMSE competitivo (0,0121), o que o coloca como o segundo modelo com melhor desempenho, o Lasso juntamente com o *Elastic Net* tem o menor MAPE (1,0288), indicando uma menor variação percentual em relação aos valore reais, indo de oposto ao *XGBoost* com (5,2683). O *Random Forest* tem o menor

RMSLE (0,0091), sugerindo um desempenho superior na previsão de valores em uma escala logarítmica.

O RMSE é uma métrica que mede a magnitude média dos erros nas previsões, no caso do *XGBoost* tunado, seu RMSE mais baixo pode ser atribuído a uma combinação de vários fatores, como por exemplo, a capacidade do modelo *XGBoost* de ajustar-se a padrões complexos nos dados, considerando as características relevantes, o ajuste de parâmetros durante a etapa de sintonia, que otimiza o desempenho do modelo, além da robustez do algoritmo *XGBoost* em lidar com dados. Portanto, esses fatores podem contribuir para um RMSE menor em comparação com outros modelos. A escolha do modelo ideal dependerá dos objetivos específicos do estudo ou projeto, mas na amostra analisada neste estudo o *XGBoost* tunado parece ser uma escolha sólida com base nas métricas apresentadas.

Figura 3 – Resultados das características



Importância das variáveis

Fonte: Elaboração própria.

A Figura 3 apresenta quais preditores têm maior contribuição na explicação do modelo *XGBoost* tunado. A variável com sentimento textual condicional ao IBOVESPA (SentNewsIBOV) foi a que mais contribuiu com a explicação do modelo, superando a variável baseada na média histórica (modelo "*naive*" ou modelo base) e até mesmo o sentimento textual agregado do dia SentNews

O *XGBoost* tunado é uma escolha mais avançada e poderosa em comparação. Ele demonstra habilidade para lidar com uma ampla gama de tarefas analíticas com um desempenho e precisão aprimorados. Esse modelo é particularmente adequado para lidar com problemas complexos e dados de alta dimensionalidade, onde a simplicidade do modelo "*naive*" pode ser insuficiente para capturar as relações intrínsecas nos dados. A escolha do modelo ideal dependerá dos objetivos específicos do estudo ou projeto, no entanto, com base nas métricas apresentadas, o *XGBoost* tunado parece ser a melhor escolha para os dados analisados neste estudo.

No contexto da análise métrica realizada, pode-se destacar a importância do uso de técnicas de *Machine Learning* na análise dos cadernos de economia e política em relação aos retornos do índice Bovespa. A aplicação dessas técnicas possibilitou uma avaliação mais aprofundada, permitindo a identificação de padrões, tendências e a comparação de modelos. Isso demonstra como a combinação de *Machine Learning* com a análise de conteúdo e os sentimentos textuais, podem fornecer *insights* valiosos, facilitando a tomada de decisões informadas e refinando a compreensão do comportamento do mercado de ações em relação às notícias econômicas e políticas.

5 CONCLUSÃO

O estudo realizado explorou a relação entre o sentimento expresso nas notícias financeiras e as flutuações nos preços do mercado acionário brasileiro, empregando técnicas de *Machine Learning* supervisionado e análise de texto. Na análise das notícias, foram coletadas ao longo de um período de dezembro de 2020 a maio de 2023, e isso destacou a predominância de médias negativas em vários meses, refletindo eventos econômicos e políticos significativos.

Os resultados também incluíram a avaliação de diferentes modelos de previsão, as métricas desempenham um papel essencial na avaliação de modelos de previsão. No contexto deste estudo, o "XGBoost Tunado" se destacou como o modelo mais preciso, com baixos valores de RMSE e MAE, bem como um MAPE razoavelmente baixo, a relevância dessas métricas reside em sua capacidade de fornecer informações claras e interpretações sobre o desempenho dos modelos, permitindo que os tomadores de decisões escolham o modelo que

melhor se ajusta aos seus objetivos específicos. Portanto, a análise de métricas desempenha um papel crítico na seleção e validação de modelos para previsão e tomada de decisões informadas.

De forma geral, o estudo demonstrou a utilidade das técnicas de *Machine Learning* na análise dos cadernos de economia e política do G1, e como essas técnicas podem proporcionar *insights* valiosos não apenas para os profissionais da área financeira, mas também para pesquisadores, a compreensão do impacto das notícias econômicas e políticas no mercado de ações é fundamental para a tomada de decisões informadas e a melhoria na previsão do comportamento do mercado.

Em seguimento às pesquisas anteriores (GODEIRO; 2018, OLIVEIRA; 2022), foi-se ampliada a amostra de dados, observando um aumento substancial na frequência de termos ao longo do tempo. Dessa forma, pode ser justificado esse fenômeno considerando vários aspectos, dentre as quais a adoção de um método de *Web Scraping* mais atualizado e eficaz, capaz de capturar informações recentes, tendo em vista que ocorre um crescimento exponencial na geração de termos e informações diariamente, acompanhado pelo contínuo enriquecimento do portfólio de artigos e notícias nos portais de notícias, como no caso do estudo o portal G1. A presente tendência de crescimento exponencial no volume de dados, tende a se manifestar de maneira ainda mais proeminente caso seja estendida a pesquisa para um período mais abrangente, compreendendo um intervalo temporal ainda maior.

No entanto, é importante ressaltar que qualquer conclusão ou interpretação dos resultados específicos do estudo deve ser baseada em uma análise mais aprofundada dos dados e da metodologia. Além disso, as condições do mercado financeiro e as influências políticas e econômicas podem mudar ao longo do tempo, tornando necessário um acompanhamento contínuo para manter a relevância das descobertas.

6. REFERÊNCIAS

DATA SCIENCE ACADEMY et al. **Web scraping e web crawling são legais ou ilegais?.** Acesso em 05 jun. 2023. Disponível em https://blog.dsacademy.com.br/web-scraping-e-web-crawling-sao-legais-ou-ilegais/.

FAMA, E. F. (1970). Efficient Capital Markets: **A Review of Theory and Empirical Work.** The Journal of Finance, 25 (2), 383–417. doi: 10.1111/j.1540-6261.1970.tb00518.x.

FIA - Fundação Instituto de Administração**. Machine Learning: o que é e como funciona**. Acesso em 10 maio 2023. Disponível em https://fia.com.br/blog/machine-learning/.

FRIEDMAN, J. H. et al. **The elements of statistical learning: data mining, inference, and prediction**. Springer Science & Business Media, 2001.

GODEIRO, L. L. et al. (2018). **Ensaios sobre modelos de previsão econômica**.

HOERL, A. E.; KENNARD, R. W. **Ridge Regression: Biased Estimation for Nonorthogonal Problems**. Technometrics, 12 (1), 55-67, 1970.

MITCHELL, Ryan**. Web Scraping with Python: Collecting Data from the Modern Web**. 1st. [S.l.]: O'Reilly Media, Inc., 2015. ISBN 1491910291.

MORAIS, F. P. et al. **Análise de sentimentos em notícias para previsão de retornos de ações na Bolsa de Valores de São Paulo**. Revista Brasileira de Gestão e Desenvolvimento Regional, 13 (2), 68-87, 2017.

NEIL PATEL. **Machine Learning: Um Guia Completo para Iniciantes**. Acesso em: 10 maio 2023. Disponível em https://neilpatel.com/br/blog/machine-learning/.

OLIVEIRA, E. E. C. **NOTÍCIAS ESPECIALIZADAS: UMA ANÁLISE DA RELAÇÃO ENTRE O RETORNO ACIONÁRIO DO MERCADO BRASILEIRO E O SENTIMENTO TEXTUAL DOS CADERNOS DE POLÍTICA E ECONOMIA DO PORTAL G1**. 2022. 31 f. Trabalho de Conclusão de Curso (Graduação em Administração) - Universidade Federal da Paraíba, Bananeiras, 2022.

RIBEIRO, V. L.; LIMA, L. V. **Previsão do Retorno Acionário com uso de Notícias Financeiras em Redes Neurais Artificiais.** In: Anais do 6º Congresso Brasileiro de Sistemas Fuzzy, 2020.

SILVA, M. D. d. O. P. d. (2018). **O efeito do sentimento das notícias sobre o comportamento dos preços no mercado acionário brasileiro.**

SILVA, Mário J. et al. **Automatic expansion of a social judgment lexicon for sentiment analysis**. 2010.

SOUZA, D. M. S.; LUCENA, Wenner Glaucio Lopes; QUEIROZ, D. B**. O Efeito do Sentimento do Investidor Expresso via Twitter sobre o Comportamento do Mercado Acionário Brasileiro Durante o Período Eleitoral.** In: Anais do Congresso USP, 2019.

SOUZA, G. M.; SILVA, D. J. **Análise de sentimentos em tweets para previsão de retorno de ações na Bovespa.** Revista de Informática Aplicada, 15 (2), 29-38, 2019.

TIBSHIRANI, R. **Regression Shrinkage and Selection via the Lasso**. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58 (1), 267-288, 1996.

TRIPATHI, Mayank**. How to Process Textual Data Using TF-IDF in Python**. Acesso em 08 de agosto de 2023. Disponível em https://www.freecodecamp.org/news/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94.

TVERSKY, A.; KAHNEMAN, D. **Prospect theory: An analysis of decision under risk. Econometrica**, 47 (2), 263–291, 1979.

ZOU, H.; HASTIE, T. **Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society:** Series B (Statistical Methodology), v. 67, n. 2, p. 301-320, 2005.