Investigando a Associação de Canais Financeiros do YouTube na Composição de Carteiras: Um Estudo com Análise de Sentimentos

Bruno de Oliveira Silva[1], **Renato Miranda Filho**[1], **Bruno Nonato Gomes**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais
Campus Sabará (IFMG)
CEP 34.590-390 – Sabará – Minas Gerais – Brasil

brunooliveirasilva77@gmail.com

{renato.miranda, bruno.nonato}@ifmg.edu.br

Abstract. *YouTube has become an increasingly relevant source of information for investors, expanding the influence of financial content creators on investment-related perceptions and decisions. In this context, this study investigates potential relationships between the sentiments expressed by financial influencers in YouTube videos, the composition of the Carteira Valor, and the behavior of stocks traded on the Brazilian stock exchange. Natural Language Processing (NLP) techniques were applied to subtitle data extracted from videos of two major financial education channels. A sentiment classification model based on the Random Forest algorithm was developed, together with data balancing techniques. In the best experimental scenario, using two classes, the model achieved an accuracy of 81%. The results indicate that, although no significant patterns were found between the extracted sentiments and the composition of the Carteira Valor, there is evidence of an association between stock mentions in the videos and short-term price variations of the respective assets.*

Resumo. *O YouTube tem se consolidado como uma importante fonte de informação para investidores, ampliando a influência de criadores de conteúdo financeiro sobre as decisões de investimento. Nesse contexto, este trabalho teve como objetivo investigar possíveis relações dos sentimentos expressos por influenciadores financeiros em vídeos da plataforma com as recomendações da Carteira Valor e com o comportamento de ações negociadas na bolsa de valores brasileira. Para isso, foram aplicadas técnicas de Processamento de Linguagem Natural (PLN) e um modelo de classificação de sentimentos baseado no algoritmo Random Forest, juntamente com técnicas de balanceamento da base de dados. No melhor cenário experimental, utilizando duas classes, o modelo obteve acurácia de 81%. Embora os resultados não mostrem a existência de padrões significativos entre os sentimentos identificados e a composição da Carteira Valor, observou-se uma associação entre as citações e a variação dos preços dos ativos.*

1. Introdução

O interesse por finanças pessoais e investimentos cresceu significativamente nos últimos anos, especialmente com a popularização das redes sociais. Plataformas como o YouTube

passaram a concentrar grande parte desse conteúdo, permitindo que influenciadores compartilhem suas experiências, estratégias e opiniões com um público cada vez maior. Nesse contexto, esses criadores passaram a exercer influência relevante na forma como muitas pessoas entendem e tomam decisões sobre investimentos [FIn 2025].

Com o avanço do consumo de conteúdo digital, o YouTube tornou-se uma das principais plataformas de disseminação de conteúdo educativo financeiro no Brasil. Entre os diversos criadores de conteúdo que atuam nesse segmento, dois influenciadores se destacam tanto pelo alcance quanto pela credibilidade de seus respectivos canais: Thiago Nigro, do canal "O Primo Rico"[1] , com mais de 7 milhões de inscritos, e Bruno Perini, do canal "Bruno Perini - Você MAIS Rico"[2], com mais de 2 milhões de inscritos, dados obtidos em janeiro de 2025 na plataforma do Youtube. A escolha desses influenciadores para este estudo deve-se ao fato de que ambos realizam e compartilham publicamente seus investimentos, o que oferece maior transparência aos seguidores. Além disso, eles figuram entre os 10 principais influenciadores com maior engajamento no mercado financeiro [FIn 2025]. Neste trabalho analisa-se as legendas dos vídeos desses dois canais a fim de identificar as ações mais mencionadas e os sentimentos atribuídos a elas, como positivos, negativos ou neutros.

A análise de sentimentos é uma abordagem utilizada no Processamento de Linguagem Natural (PLN) que busca identificar o tom de uma mensagem, classificando-a, por exemplo, como positiva, negativa ou neutra [IBM 2023]. No contexto deste trabalho, essa técnica é aplicada às legendas de vídeos, permitindo avaliar como determinadas ações são mencionadas pelos influenciadores ao longo do tempo. Essa abordagem considera o contexto das palavras empregadas, possibilitando a classificação de sentenças como positivas, negativas ou neutras. Por meio de algoritmos de PLN, é possível extrair *insights* relevantes sobre percepções e comportamentos relacionados a um determinado tema. Por exemplo, a frase "Mesmo após a recente queda das ações, continuo acreditando que o Banco do Brasil permanece como uma excelente oportunidade de investimento" apresenta sentimento positivo. Já a frase "A Petrobras registrou lucro recorde neste trimestre, mas ainda considero a ação uma opção arriscada para investir" expressa um sentimento negativo, apesar da presença de uma informação favorável. Por sua vez, a frase "O Bradesco divulgará seu balanço financeiro na próxima semana" representa um sentimento neutro, por apenas informar um fato, sem manifestar uma opinião positiva ou negativa.

A Carteira Valor[3] é composta por indicações de 16 instituições financeiras, que consolidam mensalmente as 10 ações mais votadas entre os analistas das entidades participantes. Essa carteira tem como objetivo refletir o consenso do mercado quanto aos ativos com maior potencial no período considerado, sendo amplamente utilizada como referência para análises comparativas de desempenho e de expectativas do mercado financeiro brasileiro[Valor Econômico 2025].

A bolsa de valores é o ambiente onde ocorre a negociação de ativos financeiros, como ações e fundos de investimento. No Brasil, esse papel é desempenhado pela B3, que centraliza essas operações e garante a estrutura necessária para que compradores e vendedores realizem transações de forma segura e organizada. As variações nos preços

[1] `https://www.youtube.com/@primorico`
[2] `https://www.youtube.com/@brunoperini`
[3] `https://infograficos.valor.globo.com/carteira-valor/`

desses ativos refletem diversos fatores, incluindo expectativas do mercado e informações disponíveis no momento[B3 2025].

Para a primeira etapa do trabalho, que consiste em criar um modelo classificador de sentimentos, foram realizados os processos de extração, processamento e análise manual das legendas. Apesar das limitações iniciais da base de dados, como o desequilíbrio entre as classes de sentimentos, a aplicação de técnicas de balanceamento aprimorou o desempenho do modelo. Nas análises iniciais, observa-se que o influenciador Thiago Nigro, do canal "O Primo Rico", mencionou com maior frequência a ação do Banco do Brasil, destacando-a de forma recorrente e, em muitos casos, com sentimento positivo. Por outro lado, Bruno Perini, do canal "Você MAIS Rico", teve como ação mais citada o Bradesco, apresentando um tom predominantemente neutro ao longo dos episódios. Esses padrões indicam que, mesmo com as limitações da base de dados, é possível identificar direcionamentos distintos na abordagem de cada influenciador.

O modelo desenvolvido foi posteriormente aplicado a novos conteúdos, o que permitiu analisar possíveis relações entre os sentimentos identificados nas falas dos influenciadores, a composição da Carteira Valor e o comportamento das ações negociadas na B3.

O objetivo deste trabalho é analisar as legendas de vídeos publicados por influenciadores financeiros no YouTube, comparadas às recomendações da Carteira Valor e aos dados do mercado acionário brasileiro, a fim de investigar possíveis relações entre os sentimentos expressos pelos influenciadores, a composição da Carteira Valor e o comportamento dos ativos negociados na bolsa de valores. O estudo busca contribuir para uma compreensão mais aprofundada do papel informacional desses conteúdos no contexto do mercado financeiro brasileiro. Ressalta-se, entretanto, que este trabalho não tem como objetivo prever o preço de ações.

A estrutura deste trabalho está organizada da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve a metodologia aplicada, incluindo a extração dos dados, o pré-processamento das legendas , a implementação do modelo de classificação *Random Forest Classifier* e as métricas utilizadas para avaliar o modelo; a Seção 4 dedica-se à apresentação e discussão dos resultados, incluindo a avaliação do desempenho do classificador, a análise das *playlists* "Rumo ao Bilhão" e "Viver de Ações" com base nas classificações manuais, a investigação da relação entre a *playlist* "Rumo ao Bilhão" e a Carteira Valor e, por fim, a extração de conhecimento a partir das classificações realizadas pelo modelo; e a Seção 5 apresenta as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Diversos estudos têm explorado o uso de Processamento de Linguagem Natural (PLN), técnicas de aprendizado de máquina e estratégias de comunicação digital para a análise de sentimentos e previsão de comportamentos em diferentes domínios.

O estudo de [Cunto 2021] contribui para a compreensão do impacto dos influenciadores digitais no campo da educação financeira, ao analisar as estratégias de comunicação adotadas pelo canal "O Primo Rico" no YouTube. A pesquisa focou nos três vídeos mais visualizados entre 2019 e 2020, aplicando a técnica de análise de conteúdo para identificar a presença de gatilhos mentais, como curiosidade, prova social, autoridade, entre outros. Os resultados mostraram que os gatilhos mais recorrentes foram curiosidade,

histórias e prova social, frequentemente associados a estratégias de marketing digital e testemunhos. Esses recursos comunicacionais são eficazes para gerar engajamento e criar uma relação de confiança com o público, o que reforça a relevância de investigar esse canal neste trabalho.

O estudo feito por [Sinoara et al. 2021] apresenta uma revisão das etapas e desafios da Mineração de Textos com foco na semântica, destacando que, para extrair conhecimento de dados textuais em linguagem natural, é necessário aplicar transformações que resultem em representações estruturadas. Foram descritas as fases de pré-processamento (remoção de *stopwords*, radicalização e lematização para lidar com a alta dimensionalidade e esparsidade dos dados textuais), extração de padrões (uso de algoritmos de aprendizado de máquina) e pós-processamento (utilização do conhecimento). O estudo conclui que ainda não existe uma solução definitiva para a compreensão semântica automatizada, o que exige decisões criteriosas do analista quanto à abordagem a ser utilizada.

O trabalho de [Pinto 2024] teve como objetivo classificar as avaliações de usuários da plataforma Steam — um serviço online voltado à compra, download e gerenciamento de jogos digitais para computador — em sentimentos positivos, negativos ou neutros no contexto das experiências com os jogos. Para isso, foi coletado um grande volume de dados de comentários, seguido de um rigoroso pré-processamento textual, que incluiu remoção de *stopwords*, lematização, *stemming*, normalização, tokenização e vetorização. O modelo *Random Forest* foi implementado com a biblioteca Scikit-learn em Python, e técnicas de *oversampling* e *undersampling* foram aplicadas para lidar com o desbalanceamento das classes. Como resultado, o modelo obteve alto desempenho, com *recall* de 0,91, precisão de 0,89, *F1-Score* de 0,88 e acurácia de 0,89, demonstrando eficácia na identificação de padrões de opinião e aspectos como jogabilidade, gráficos e enredo nas avaliações de jogos.

O estudo apresentado por [Silva et al. 2024] analisou diferentes algoritmos de Inteligência Artificial aplicados à tarefa de análise de sentimentos em publicações do Twitter (atualmente X) durante o período das eleições presidenciais de 2022 no Brasil. A pesquisa teve como foco as opiniões manifestadas sobre as candidaturas presidenciais, utilizando uma base composta por 1.000 *tweets* rotulados com o auxílio da API do ChatGPT para o treinamento dos modelos e outros 300 *tweets* rotulados manualmente para teste. Foram avaliados diversos algoritmos, incluindo *Random Forest, Multilayer Perceptron, Support Vector Machine (SVM), K-Nearest Neighbors (KNN)* e o modelo pré-treinado em português BERTimbau. Os melhores resultados foram obtidos com a combinação entre BERTimbau Large Embedding e *Random Forest*, com uma acurácia de 68%. O estudo demonstrou ainda que, mesmo com o uso de rotulagem automatizada, os resultados obtidos foram satisfatórios; no entanto, ainda há espaço para aprimoramentos, especialmente no que se refere ao balanceamento da base de dados. Esses achados contribuem para a discussão sobre o uso de modelos pré-treinados e classificadores tradicionais em domínios distintos, além de reforçar a relevância do *Random Forest* em contextos de análise de sentimentos.

O trabalho de [Santos et al. 2023] foca na análise de sentimentos em textos financeiros em português. O modelo foi desenvolvido a partir do BERTimbau, com *fine-tuning* aplicado em dados do domínio financeiro. Técnicas como Gradual *Unfreezing* foram empregadas para preservar o conhecimento previamente adquirido, e a métrica de per-

plexidade [Chen et al. 1998] — que avalia o quão bem o modelo consegue prever uma palavra dado o contexto — foi utilizada para a avaliação. O modelo final, SentFinBERT-PT-BR, obteve resultados expressivos, com acurácia de 0,76 e *F1-Score* de 0,73, além de apresentar bom desempenho qualitativo na identificação de sentimentos correlacionados a eventos do mercado de ações.

O trabalho [Piloneto et al. 2020] teve como objetivo prever variações no mercado acionário brasileiro com base na análise dos sentimentos de notícias extraídas da web. Para isso, foi implementado um processo de *web scraping* com as bibliotecas *Selenium e BeautifulSoup*, que coletou textos de notícias relacionadas a ações específicas. Em seguida, os textos passaram por pré-processamento com o NLTK, incluindo tokenização, remoção de *stopwords* e atribuição de pesos às palavras mais relevantes. A análise de sentimentos foi realizada com uma rede neural treinada com o modelo *Bag of Words*, em que dados foram normalizados, vetorizados e embaralhados antes do treinamento com 1000 épocas. Para validar a eficiência do modelo, foram analisadas as variações das ações nos dias subsequentes à coleta das notícias. Observou-se que, nas ocasiões em que o algoritmo identificou predominância de notícias positivas, as ações tendiam a apresentar valorização. Por outro lado, quando predominavam notícias negativas, verificou-se uma queda nos preços das ações. Esses resultados sugerem uma correlação direta entre o sentimento extraído das notícias e o comportamento do mercado acionário.

Dessa forma, os trabalhos revisados fornecem uma base sólida para o desenvolvimento da presente pesquisa, tanto no aspecto metodológico quanto na escolha de ferramentas, algoritmos e abordagens de análise. Enquanto estudos como os de [Sinoara et al. 2021], [Pinto 2024] e [Silva et al. 2024] contribuíram para a compreensão das etapas e desafios do processamento de linguagem natural e da análise de sentimentos, pesquisas como as de [Santos et al. 2023], [Piloneto et al. 2020] e [Cunto 2021] evidenciaram, respectivamente, a eficácia de modelos especializados no contexto financeiro e o papel estratégico da comunicação digital na construção de influência.

O presente trabalho se diferencia ao aplicar essas técnicas a legendas automáticas de vídeos de influenciadores financeiros no *YouTube*, que, apesar de apresentarem pequenas imprecisões de transcrição, mostraram-se adequadas para as análises realizadas, explorando não apenas a acurácia dos modelos de classificação, mas também a possível relação entre o conteúdo desses canais, a composição da Carteira Valor e o comportamento das ações no mercado brasileiro. Assim, busca-se contribuir com uma abordagem inovadora e complementar aos estudos existentes nas áreas de ciência de dados, finanças e comunicação digital.

3. Metodologia

A presente seção descreve detalhadamente os procedimentos adotados para atingir os objetivos deste trabalho. A metodologia foi estruturada em etapas sequenciais que abrangem desde a extração das legendas dos vídeos dos canais selecionados até as métricas utilizadas para avaliar o modelo. A Figura 1 ilustra o fluxograma das fases do processo metodológico, que serão aprofundadas nas subseções seguintes.

3.1. Extração dos Dados

Todo o processo de extração e análise das legendas dos vídeos foi realizado em Python, considerada uma das linguagens mais populares e eficazes na área de ciência de dados

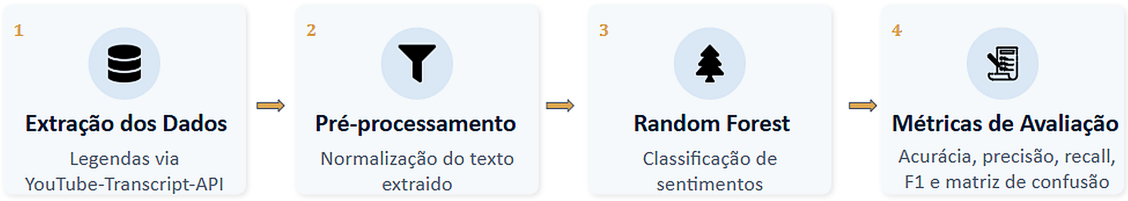


Figura 1. Fluxograma das etapas metodológicas do estudo

[Rice University 2023]. O período de coleta de dados para o treinamento do modelo compreendeu os meses de agosto de 2023 a agosto de 2024, abrangendo os episódios 63 a 75 da *playlist* "Rumo ao Bilhão" e os episódios 07 a 20 da *playlist* "Viver de Ações". Esse intervalo proporcionou um recorte temporal consistente para a fase de treinamento, uma vez que ambos os canais publicaram vídeos regularmente em suas respectivas *playlists* ao longo desse período.

Para a obtenção das legendas dos vídeos das *playlists* Rumo ao Bilhão do Thiago Nigro e "Viver de Ações" do Bruno Perini, utilizou-se a biblioteca *YouTubeTranscriptApi* [PyPI 2024], que permite a extração das transcrições diretamente da plataforma do YouTube. Ao todo, foram extraídos 538 KB de dados textuais da *playlist* Rumo ao Bilhão, contendo 68.103 palavras, e 1.096 KB da *playlist* "Viver de Ações", contendo 136.602 palavras, os quais serviram de insumo para o treinamento e teste do modelo.

Além disso, buscou-se ampliar a base de dados de treinamento por meio de métodos artificiais. Para isso, utilizou-se o ChatGPT para a geração de novas frases, a partir de *prompt* que solicitava a criação de textos semelhantes às falas utilizadas pelos influenciadores no YouTube. Também foram utilizadas as legendas da *playlist* Rumo ao Bilhão, publicadas antes de agosto de 2023, esses trechos foram submetido ao ChatGPT para a classificação dos sentimentos. Adicionalmente, foi empregada a técnica de *oversampling SMOTE (Synthetic Minority Over-sampling Technique)*, que gera artificialmente novas amostras da classe minoritária por meio da interpolação entre exemplos existentes, contribuindo para reduzir o desbalanceamento entre as classes sem simplesmente replicar os dados originais.

Após a conclusão do treinamento do modelo, foram extraídos 345 KB de dados textuais da *playlist* "Rumo ao Bilhão", contendo 66.115 palavras, referentes ao período de fevereiro a maio de 2025. Não foram utilizados dados da *playlist* "Viver de Ações" nessa etapa, em razão do Bruno Perini ter encerrado a publicação de novos vídeos. A base extraída foi utilizada na etapa de extração de conhecimento, em que os sentimentos classificados pelo modelo foram comparados com as ações presentes na Carteira Valor e com o comportamento dos respectivos preços desses ativos na bolsa de valores (B3).

3.2. Pré-processamento dos Dados

As legendas extraídas dos vídeos foram inicialmente obtidas em um formato estruturado contendo informações como texto (*text*), tempo de início (*start*) e duração (*duration*). No entanto, para os fins deste trabalho, apenas o conteúdo textual foi considerado, descartando-se os metadados relacionados ao tempo. Na Figura 2 consta um exemplo de como estão formatados os dados obtidos por meio da biblioteca *YouTubeTranscriptApi*.

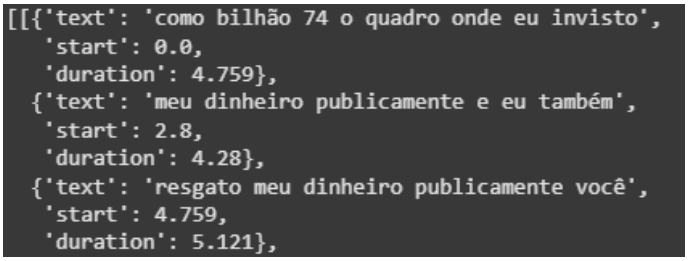


Figura 2. Formato das legendas

Após a seleção da base textual para treinamento do modelo, foi implementado uma função para uniformizar os dados extraídos das legendas dos vídeos. Para isso, foi implementada uma função responsável por concatenar os trechos de legenda de cada vídeo, converter todos os caracteres para letras minúsculas e remover as acentuações por meio da normalização Unicode da linguagem Python. Essa etapa teve como objetivo padronizar os dados, facilitando as fases subsequentes de análise e o treinamento dos modelos de aprendizado de máquina.

Em seguida, foi selecionado um conjunto de empresas com forte relevância no mercado financeiro brasileiro, abrangendo setores como energia, finanças, varejo, saúde, infraestrutura e consumo. Entre elas estão: “Petrobras”, “Cemig”, “Itaú”, “Bradesco”, “Magazine Luiza”, “Via Varejo”, “CCR”, “Raia Drogasil”, “Equatorial”, “Hapvida”, “Ambev”, “EcoRodovias”, “B3”, “Copasa”, “Klabin”, “Banco do Brasil”, “Banco ABC”, “Sanepar” e “Transmissão Paulista” [INVESTIDOR10 2025]. A seleção dessas companhias teve como objetivo garantir diversidade setorial e representatividade de ativos amplamente negociados na bolsa brasileira, muitos dos quais compõem o índice Ibovespa.

Após a etapa de seleção das empresas, foi utilizada a biblioteca nativa *re* da linguagem Python para localizar as menções às empresas nas legendas dos vídeos. Por meio de Expressões Regulares (Regular Expressions – Regex), foram identificados os padrões correspondentes aos nomes das empresas, possibilitando a extração das frases de interesse para a análise de sentimentos. Com o objetivo de preservar o contexto das menções, cada frase extraída foi composta por cinco palavras anteriores e dez posteriores ao nome da empresa, permitindo uma análise mais precisa dos sentimentos expressos. Como exemplo, para a empresa Bradesco, foi extraído o trecho “a projeção de lucro de **Bradesco** para 2023 ela fica em torno de 20 bilhões e”.

Durante a etapa de análise, foi necessário realizar um ajuste manual na nomenclatura para garantir a detecção correta das palavras-chave nos textos. Por exemplo, o termo **“Semig”** foi adotado em substituição a **“Cemig”**, com o objetivo de padronizar a nomenclatura conforme as legendas extraídas dos vídeos. Adicionalmente, a empresa “Vale” foi excluída do conjunto analisado devido à ambiguidade do termo, que poderia gerar ruído na análise de sentimentos por possuir múltiplos significados no idioma português. Não foi identificada outra empresa cuja nomenclatura pudesse ocasionar esse tipo de ruído.

Na etapa de análise de sentimentos, foram aplicadas técnicas de Processamento de Linguagem Natural (PLN) utilizando o modelo SentFinBERT-PT-BR, proposto por [Santos et al. 2023], com o objetivo de identificar o sentimento predominante das legendas — positivo, negativo ou neutro. No entanto, durante os testes, observou-se que o

modelo apresentou dificuldades para classificar corretamente diversas frases presentes no conjunto de dados. Como exemplo, a frase “e muito bizarro nunca o banco do brasil teve um lucro tao alto”, retirada de um vídeo da *playlist* Rumo ao Bilhão, foi erroneamente interpretada como negativa.

Diante dessas limitações, optou-se por realizar a rotulagem manual de 526 frases provenientes dos dados coletados, conduzida exclusivamente pelo autor deste trabalho. Embora não tenha havido a participação de outros avaliadores para validação dos rótulos, buscou-se manter critérios consistentes durante todo o processo de anotação, considerando o contexto das frases e o sentimento expresso em relação aos ativos mencionados. Posteriormente as frases serviram como base para o treinamento do algoritmo *Random Forest Classifier*. A escolha desse modelo foi motivada pelos resultados positivos obtidos no estudo de [Pinto 2024], em que o algoritmo demonstrou eficácia na classificação de sentimentos (negativo, neutro e positivo) no contexto de avaliações de jogos online, alcançando acurácia de 89% e F1-Score de 88%.

3.3. *Random Forest Classifier*

O algoritmo *Random Forest Classifier* (RFC) é um método de aprendizado supervisionado baseado em um conjunto de árvores de decisão, onde cada árvore é construída a partir de uma amostra do conjunto de dados de treinamento obtida por meio da técnica de *bootstrapped* (amostragem com reposição). Durante a construção de cada árvore, aplica-se um processo de partição recursiva, no qual os dados são divididos sucessivamente a partir do nó raiz até atingir condições de parada predefinidas. A capacidade preditiva do modelo decorre da agregação dos resultados de múltiplas árvores individuais — classificadores considerados fracos isoladamente, mas que, em conjunto, formam um modelo robusto. O desempenho do RFC é particularmente eficaz quando há baixa correlação entre as árvores, o que favorece uma diversidade estrutural que reduz o risco de *overfitting*, que acontece quando o modelo se ajusta excessivamente aos dados de treino e não generaliza bem [Hu and Szymczak 2023].

Inicialmente, foram realizados testes com diferentes algoritmos de classificação, incluindo o *Support Vector Machine* (SVM), o *Extreme Gradient Boosting* (XGBoost) e o *Random Forest Classifier* (RFC). Entre os modelos avaliados, o RFC apresentou o melhor desempenho na classificação dos sentimentos, motivo pelo qual foi selecionado para os experimentos subsequentes. A escolha desse algoritmo também foi fundamentada nos resultados obtidos por [Pinto 2024] e [Silva et al. 2024], que demonstraram sua eficácia em tarefas de classificação de sentimentos.

Para a implementação do modelo, utilizou-se a biblioteca *scikit-learn*. Inicialmente, os textos foram convertidos em vetores numéricos por meio da técnica TF-IDF (*Term Frequency-Inverse Document Frequency*), de modo a representar cada frase pelas características mais relevantes do conjunto de dados. Em seguida, foram realizados testes com diferentes quantidades de árvores de decisão, avaliando valores inferiores e superiores, até que o modelo com 95 árvores apresentasse o melhor desempenho experimental. Além disso, foram definidos 42 estados aleatórios responsáveis por garantir a reprodutibilidade dos resultados.

O modelo foi treinado com uma base de dados construída a partir das legendas extraídas dos vídeos analisados. A rotulagem inicial dos sentimentos, classificados em

positivo, negativo e neutro, foi realizada manualmente pelo autor deste trabalho, totalizando 293 frases neutras, 156 positivas e 77 negativas. Essa abordagem supervisionada teve como objetivo fornecer ao algoritmo um conjunto de dados confiável para a etapa de treinamento.

3.4. Fundamentação Teórica

A avaliação de desempenho de um modelo de aprendizado supervisionado é uma etapa essencial para verificar sua eficácia e capacidade de generalização. No contexto da análise de sentimentos, onde os dados são rotulados em diferentes classes (positivo, negativo e neutro), é fundamental adotar métricas que vão além do simples número de acertos. Para isso, utilizam-se indicadores como acurácia, precisão, *recall* e *F1-Score*, que fornecem uma visão mais detalhada sobre o comportamento do classificador em diferentes situações. Esses indicadores baseiam-se na matriz de confusão, uma ferramenta que resume as previsões do modelo em relação aos valores reais, permitindo uma análise mais crítica e aprofundada dos resultados.

A matriz de confusão permite avaliar a qualidade das previsões de um classificador. Os elementos da diagonal principal representam as classificações corretas, ou seja, os casos em que o rótulo previsto corresponde ao rótulo verdadeiro, sendo compostos pelos verdadeiros positivos (VP) e verdadeiros negativos (VN). Já os elementos fora da diagonal representam erros de classificação, correspondentes aos falsos positivos (FP) e falsos negativos (FN), conforme ilustrado na Tabela 1 [Pedregosa et al. 2011].

	Predito Positivo	**Predito Negativo**
Real Positivo	VP	FN
Real Negativo	FP	VN

Tabela 1. Matriz de Confusão

A acurácia é uma métrica que indica a proporção de previsões corretas do classificador em relação ao total de amostras avaliadas. Em outras palavras, mede o quão frequentemente o modelo acerta, considerando tanto as classificações positivas quanto as negativas corretas. Sua fórmula é expressa da seguinte forma [Junior 2023]:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

A precisão é uma métrica que avalia a proporção de acertos entre as amostras classificadas como positivas pelo modelo. Em outras palavras, indica o quanto das previsões positivas do classificador efetivamente pertencem à classe positiva. A métrica é calculada por meio da seguinte expressão:

$$\text{Precisão} = \frac{VP}{VP + FP}$$

A *recall* (revocação ou sensibilidade) é uma métrica que avalia a capacidade do modelo em identificar corretamente os casos positivos. Ou seja, mede a proporção de

amostras positivas que foram corretamente classificadas como tais. Essa métrica é especialmente relevante em contextos em que é mais importante não deixar de identificar positivos reais, mesmo que isso implique alguns falsos positivos. A fórmula é expressa:

$$\text{Recall} = \frac{VP}{VP + FN}$$

A medida F1 (F1-Score) é a média harmônica entre a precisão e a *recall*, sendo especialmente útil em contextos onde é necessário equilibrar os erros do tipo falso positivo (FP) e falso negativo (FN). Diferentemente da média aritmética, a média harmônica penaliza valores muito discrepantes entre as duas métricas, oferecendo uma visão mais balanceada do desempenho do modelo. A fórmula que define essa métrica é:

$$\text{F1} = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

4. Discussão e Análise dos Resultados

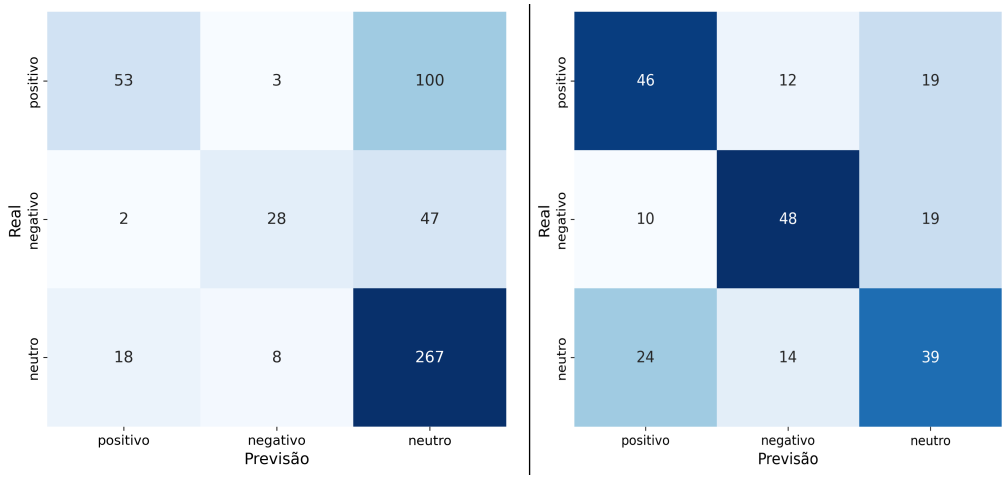
Nesta seção são apresentados e discutidos os principais resultados obtidos a partir da aplicação das técnicas descritas na metodologia. A análise dos resultados está dividida em cinco partes principais: a avaliação do modelo construído pelo algoritmo *Random Forest Classifier*; a interpretação das menções e sentimentos extraídos das *playlists* "Rumo ao Bilhão" e "Viver de Ações", utilizadas para o treinamento do modelo; a análise da relação entre a *playlist* "Rumo ao Bilhão" e a Carteira Valor; e a extração do conhecimento aplicado nos valores das ações.

4.1. Análise do *Random Forest Classifier*

Os testes iniciais, realizados a partir da base de dados coletada e rotulada manualmente — composta por 293 frases neutras, 156 positivas e 77 negativas — apresentaram acurácia de 66%. Na primeira parte da Figura 3 (a), encontra-se a matriz de confusão obtida na classificação inicial. Observa-se que a classe neutra apresentou aproximadamente 91% de acertos, enquanto a classe positiva alcançou cerca de 34% de classificações corretas e a classe negativa obteve aproximadamente 36% de acertos. Esses resultados evidenciam a dificuldade do modelo em generalizar adequadamente as categorias positiva e negativa. Além disso, o elevado desempenho observado na classificação da classe neutra pode estar relacionado ao desbalanceamento da base de dados, uma vez que essa categoria possui uma quantidade significativamente maior de amostras em comparação às demais classes.

Para mitigar esse problema, foi aplicado o balanceamento por *undersampling*, que consiste na seleção aleatória de amostras das classes majoritárias, descartando os exemplos excedentes com o objetivo de equilibrar a distribuição das classes em relação à classe minoritária, conforme discutido por [Pinto 2024]. Além disso, foi implementada uma validação cruzada em cinco folds (5-fold cross-validation), na qual a base de dados foi particionada em cinco subconjuntos. Em cada iteração, um subconjunto foi utilizado para teste e os quatro restantes para treinamento, garantindo que todas as amostras fossem avaliadas sem que houvesse sobreposição entre os conjuntos de treinamento e teste.

Após o balanceamento, cada classe passou a conter 77 frases, conforme apresentado na Figura 3(b). Observou-se uma melhora significativa no desempenho das classes



(a) Matriz de confusão base desbalanceada

(b) Matriz de confusão base balanceada

Figura 3. Matrizes de confusão dos experimentos realizados

positiva e negativa. A categoria positiva, que inicialmente apresentava baixo desempenho, passou a atingir aproximadamente 60% de acertos, representando um aumento de cerca de 26 pontos percentuais em relação ao experimento anterior. De forma semelhante, a classe negativa alcançou aproximadamente 62% de acertos, também registrando um ganho de cerca de 26 pontos percentuais. Em contrapartida, a classe neutra apresentou aproximadamente 51% de classificações corretas, correspondendo a uma redução de cerca de 40 pontos percentuais em comparação com a base desbalanceada.

Adicionalmente, observou-se que o algoritmo apresentou dificuldade em diferenciar adequadamente as classes neutra e positiva, classificando incorretamente 24 frases neutras como positivas, o que corresponde a aproximadamente 31% das amostras dessa classe. De modo geral, os resultados indicam que o balanceamento dos dados contribuiu para uma representação mais equilibrada dos sentimentos durante o processo de aprendizado do modelo.

Classificação	Precisão	Recall	F1
negativo	0,83	0,88	0,85
neutro	0,78	0,68	0,73
positivo	0,53	0,57	0,55

Tabela 2. Relatório de classificação com a base de dados balanceada

A Tabela 2 apresenta os resultados obtidos pelo modelo para as métricas analisadas; foram analisados o *F1-Score*, a precisão e o *recall*. A classe negativa apresentou o melhor desempenho, com *F1-Score* de 85%, indicando um bom equilíbrio entre acertos e a redução de falsos positivos e falsos negativos. A classe neutra obteve um F1 de 73%, demonstrando desempenho razoável, mas ainda com margem de aprimoramento. Já a classe positiva, embora tenha registrado o menor F1-Score (55%), apresentou um avanço de aproximadamente 10 pontos percentuais em relação ao desempenho inicial evidenciado na matriz de confusão (Figura 3 (a)).

A métrica de precisão, que indica a proporção de acertos entre todas as previsões

positivas realizadas, foi mais alta para a classe negativa (83,33%), revelando elevada confiabilidade nas classificações dessa categoria. A classe neutra obteve 78%, enquanto a positiva registrou 53%.

Quanto à *recall*, os resultados também foram mais expressivos na classe negativa, com 88%, confirmando a capacidade do modelo em identificar corretamente grande parte das amostras dessa classe. As classes neutra e positiva, por sua vez, apresentaram revocações de 68% e 57%, respectivamente, o que indica desafios persistentes na detecção precisa desses sentimentos.

A Tabela 3 apresenta os resultados obtidos na comparação do modelo utilizando a base de dados com e sem a classe neutra. Observa-se que o algoritmo apresenta maior dificuldade na classificação da base de dados que contem a classe neutra. Ao analisar a acurácia, verifica-se um ganho de aproximadamente 15% na base manual completa e de 18% na base manual balanceada quando a classe neutra é removida. Esses resultados indicam que a redução da complexidade da tarefa de classificação, passando de três classes para duas, contribuiu para a melhora no desempenho do modelo em ambas as bases de dados. Além disso, a remoção da classe neutra reduziu a ambiguidade entre as categorias, permitindo ao algoritmo identificar de forma mais clara os padrões associados aos sentimentos positivos e negativos.

	Com neutro		Sem neutro	
Base de dados	Acurácia	F1	Acurácia	F1
Rot. manual completa	58%	44%	73%	61%
Rot. manual balanceada	59%	57%	77%	76%

Tabela 3. Desempenho do modelo com e sem a classe neutra

Embora o balanceamento e a remoção da classe neutra tenham melhorado o desempenho, a base de dados ainda carece de ampliação para se tornar mais robusta e confiável. Com esse objetivo, foram adotadas estratégias voltadas ao aumento do conjunto de dados. Uma das abordagens consistiu na geração de frases artificiais por meio do ChatGPT, utilizando o seguinte *prompt*: “Gere um arquivo CSV a partir da análise das legendas do youtube dos seguintes canais: Luan Onofre Economista Sincero, Professor Baroni, Clube do Dividendos e Luiz Vaz Trader. As colunas devem ter: Frase: A frase deve expressar claramente um sentimento (positivo, negativo ou neutro) sobre a ação correspondente; Ação: o nome da ação mencionada na frase; Sentimento: positivo, negativo ou neutro.Requisitos adicionais: O arquivo deve estar balanceado entre as três classes de sentimento (ou seja, com quantidades iguais de frases positivas, negativas e neutras). A expectativa é a geração 300 falas de cada classe (positivo, negativo, neutro)”.

Além disso, realizou-se uma tentativa de ampliar a base de dados por meio da inclusão de novos dados reais extraídos da *playlist* Rumo ao Bilhão, referentes a vídeos publicados antes do período inicialmente considerado para as análises, isto é, antes de agosto de 2023. Esse procedimento resultou na geração de um arquivo .txt com 1483 KB de texto processado. Em seguida, o conteúdo foi submetido ao ChatGPT para classificação, utilizando o seguinte *prompt*: “Gere um arquivo CSV a partir do conteúdo anexo com as seguintes colunas: Frase: uma sentença que contenha 5 palavras antes da ação 10 palavras depois da menção da ação. A frase deve expressar claramente um sentimento

(positivo, negativo ou neutro) em relação à ação correspondente. Ação: o nome da ação mencionada (ex: Banco do Brasil, Bradesco, Cemig etc.). Não incluir a ação Vale. Sentimento: positivo, negativo ou neutro. Requisitos adicionais: Gere o máximo de texto possível, pois o arquivo textual contém muitas falas de ação. por exemplo, o Banco do Brasil tem 142 correspondências". Obteve ao final 505 frases classificadas por esse método.

Ademais, foi aplicada a técnica de *oversampling* por meio do método *SMOTE (Synthetic Minority Over-sampling Technique)* na base de dados rotulada manualmente completa, conforme proposto por [Pinto 2024]. O método foi utilizado para gerar amostras sintéticas das classes minoritárias por meio da interpolação entre instâncias vizinhas no espaço de características, aumentando artificialmente o número de exemplos dessas classes. Dessa forma, buscou-se reduzir o desbalanceamento entre as classes e proporcionar uma distribuição mais equilibrada dos dados, contribuindo para o treinamento de modelos de classificação mais robustos e menos enviesados.

A Tabela 4 apresenta os resultados obtidos a partir das diferentes bases de dados utilizadas no treinamento do modelo. Para fins de avaliação, utilizou-se como conjunto de teste a base rotulada manualmente completa. Foram avaliadas quatro estratégias distintas de construção da base de treinamento.

	Com neutro		Sem neutro	
Base Treinamento	Acurácia	F1	Acurácia	F1
Rot. GPT artificias balanceada	15%	9%	33%	26%
Rot. GPT reais completa	58%	36%	69%	47%
Rot. GPT reais balanceada	61%	45%	70%	55%
Smote	73%	65%	81%	75%

Tabela 4. Experimentos artificias para aumentar a base de dados

A primeira estratégia, denominada Rot. GPT artificiais balanceada, consistiu na geração de frases artificiais pelo ChatGPT, produzindo 300 frases para cada classe de sentimento (positiva, negativa e neutra). Entretanto, essa abordagem apresentou baixo desempenho, alcançando acurácia de 15% e F1 de 9% considerando a presença da classe neutra, e acurácia de 33% e F1 de 26% após sua remoção. Esses resultados sugerem que as frases geradas artificialmente não representaram adequadamente o padrão linguístico presente nos dados reais.

A segunda estratégia, denominada Rot. GPT reais completa, foi construída a partir de frases reais extraídas da *playlist* "Rumo ao Bilhão", classificadas automaticamente pelo ChatGPT. Essa base apresentou distribuição desbalanceada, composta por 400 frases neutras, 81 positivas e 24 negativas. Nessa configuração, o modelo obteve acurácia de 58% e F1 de 36% com três classes, passando para 69% de acurácia e 47% de F1 após a remoção da classe neutra. Observa-se que o desbalanceamento entre as classes influenciou negativamente o desempenho do classificador.

A terceira estratégia, denominada Rot. GPT reais balanceada, foi obtida a partir da aplicação da técnica de *undersampling* sobre a base anterior, reduzindo o número de amostras para 24 frases em cada classe. O balanceamento proporcionou melhora no desempenho, atingindo acurácia de 61% e F1 de 45% considerando as três classes, e acu-

rácia de 70% e F1 de 55% utilizando apenas as classes positiva e negativa. Entretanto, a redução significativa da quantidade de amostras pode limitar a capacidade de generalização do modelo.

A base Rot. GPT reais balanceada por meio da aplicação de *undersampling* sobre a base anterior, resultando em 24 frases por classe. Essa abordagem proporcionou melhora nos resultados, com acurácia de 61% e F1 de 45% considerando todas as classes, e acurácia de 70% e F1 de 55% sem a classe neutra. Contudo, ressalta-se que o balanceamento reduziu significativamente o tamanho da base, o que pode limitar a capacidade de generalização do modelo.

Por fim, a estratégia baseada em *SMOTE* consistiu na geração sintética de novas amostras a partir da base rotulada manualmente, preservando as características das classes originais e equilibrando sua distribuição. Nessa configuração, foram obtidas 293 amostras por classe no cenário com três classes (positiva, negativa e neutro) e 156 amostras por classe no cenário com duas classes (positiva e negativa). Essa abordagem apresentou o melhor desempenho entre todos os experimentos, alcançando acurácia de 73% e F1 de 65% com a classe neutra e, no melhor cenário experimental, 81% de acurácia e 75% de F1 utilizando apenas as classes positiva e negativa. O conjunto de dados foi inicialmente dividido em 80% para treinamento e 20% para teste, sendo o SMOTE aplicado exclusivamente sobre o conjunto de treinamento.

4.2. Análise da *playlist* Rumo ao Bilhão via Base Rotulada Manualmente

A Figura 4 apresenta a frequência de menções às ações na *playlist* "Rumo ao Bilhão", considerando os vídeos publicados entre agosto de 2023 e agosto de 2024 e analisados a partir da base de dados rotulada manualmente. Observa-se que a ação do Banco do Brasil foi a mais citada, com aproximadamente 70 menções. Esse resultado indica que o ativo recebeu maior atenção do influenciador durante o período analisado em comparação às demais empresas.

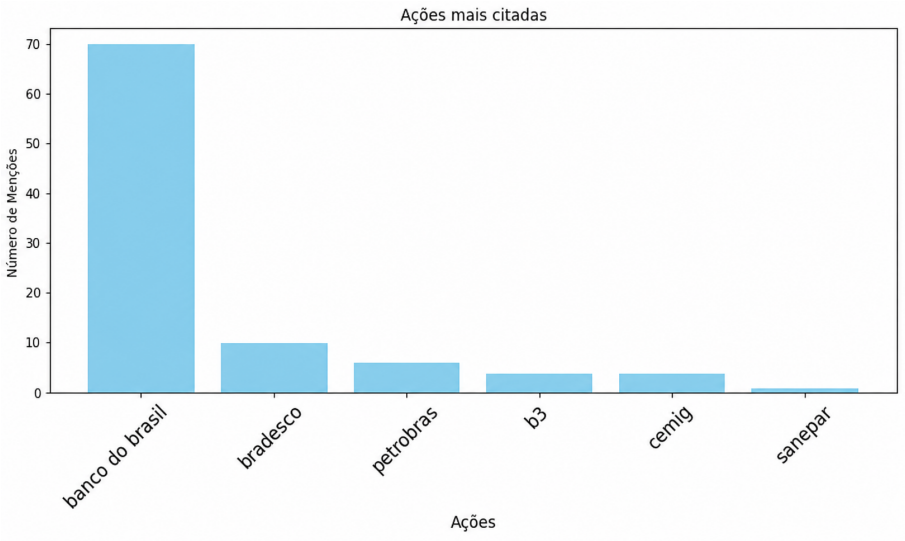


Figura 4. Ações mais citadas Thiago Nigro

Além da frequência de menções, a análise dos sentimentos ao longo do tempo revela variações significativas nas percepções em relação ao Banco do Brasil, conforme

ilustrado na Figura 5. De maneira geral, observa-se um predomínio de sentimentos positivos no período analisado, especialmente entre dezembro de 2023 e agosto de 2024. Uma exceção notável ocorre em junho de 2024, quando há um aumento expressivo de comentários neutros. Também foram identificadas menções com sentimento negativo em determinados momentos, como em agosto e outubro de 2023.

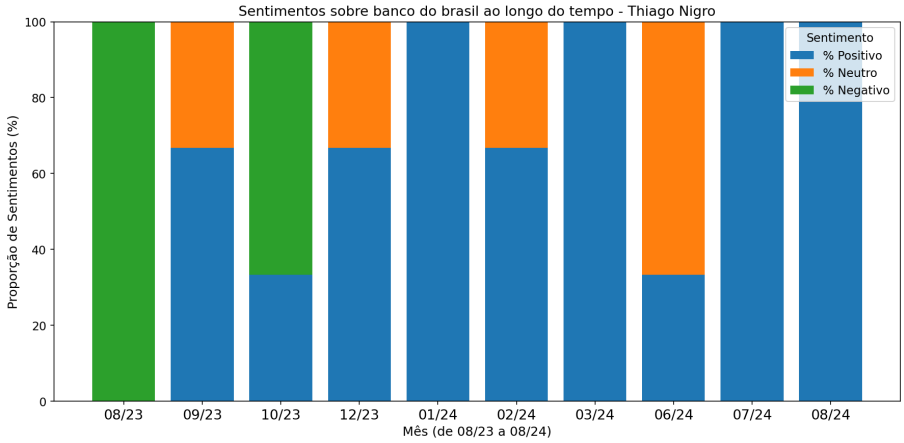
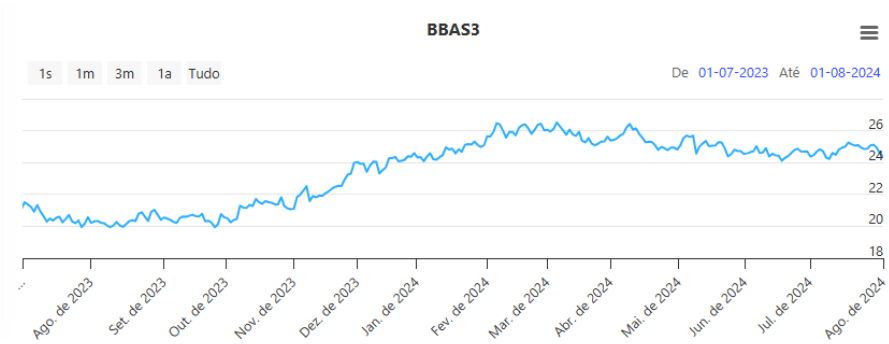


Figura 5. Sentimentos Banco do Brasil

A Figura 6, extraída do portal [InfoMoney25 2025], apresenta a variação do preço da ação do Banco do Brasil (BBAS3) ao longo do período analisado. Entre os meses de agosto e novembro de 2023, o papel foi negociado abaixo de R$22,00. A partir desse ponto, observou-se um movimento de valorização significativa, com o preço alcançando aproximadamente R$26,00. Ao relacionarmos esses dados aos sentimentos expressos nas menções — conforme ilustrado na Figura 5 — é possível perceber uma correlação interessante: durante o período de baixa, o influenciador proferiu comentários majoritariamente negativos sobre a ação, enquanto, no período de alta, suas menções tornaram-se predominantemente positivas.



Fonte: InfoMoney

Figura 6. Ação do Banco do Brasil ao longo do período analisado

4.3. Análise da *playlist* Viver de Ações via Base Rotulada Manualmente

A Figura 7 destaca as duas ações mais mencionadas pelo influenciador Bruno Perini em sua *playlist* "Viver de Ações", B3 e Bradesco. No entanto, após análise mais criteriosa,

observou-se que, em algumas ocasiões, a menção à B3 não se referia diretamente à ação, como no trecho extraído do vídeo 7 da *playlist*: "única bolsa que temos no Brasil é a B3, e aí quando você compra fundo imobiliário e ações talvez...". Diante disso, optou-se por focar a análise na ação do Bradesco, cuja recorrência nas falas do influenciador pode indicar uma ênfase maior em temas relacionados à empresa, seja no contexto de investimentos, estratégias ou conjuntura econômica.

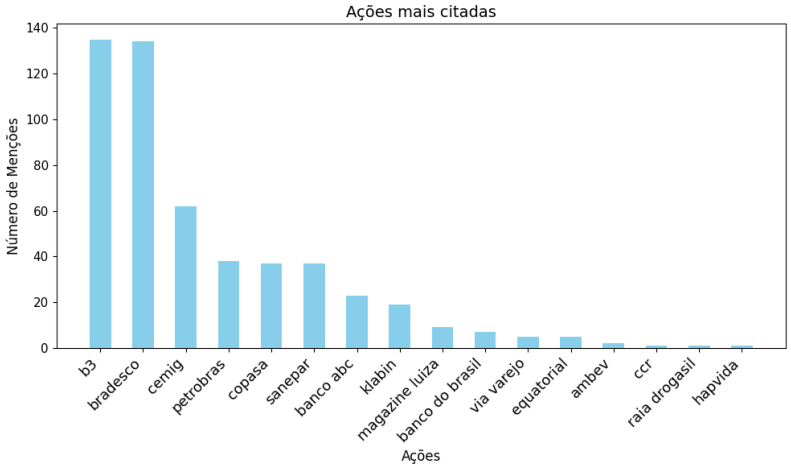


Figura 7. Ações mais citadas Bruno Perine

Ao analisar o comportamento da ação do Bradesco na base de dados rotulada manualmente, conforme ilustrado na Figura 8, observa-se que os sentimentos predominantes foram neutros e positivos. Isso sugere que, de modo geral, as menções ao Bradesco na *playlist* "Viver de Ações" adotaram um tom equilibrado ou favorável. Destaca-se, por exemplo, que, em janeiro de 2024, todas as menções registraram um sentimento positivo. No entanto, entre junho e agosto de 2024, foram identificadas algumas ocorrências de sentimento negativo, indicando possíveis críticas ou cautelas levantadas pelo influenciador nesse período.

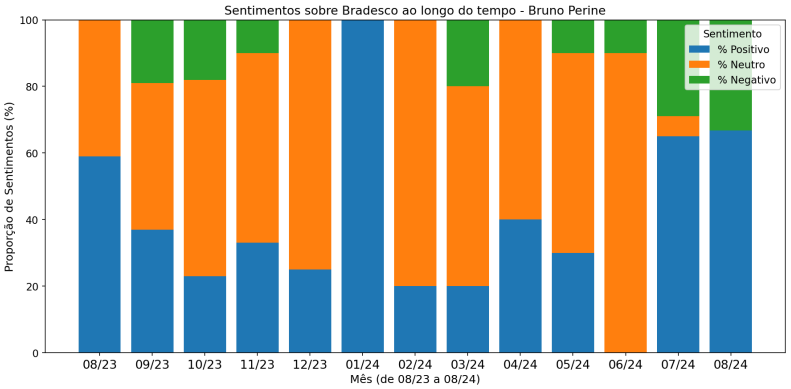
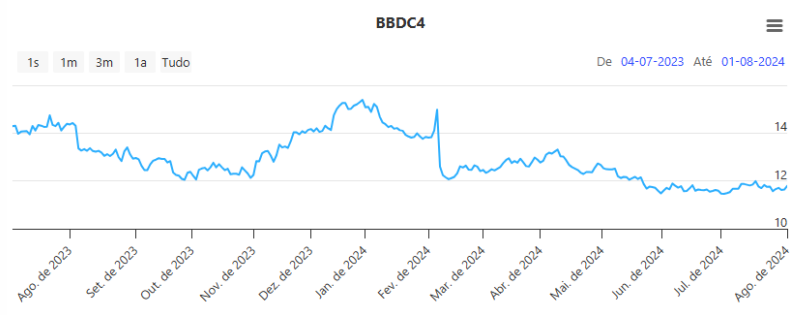


Figura 8. Sentimentos Bradesco

Ao analisar a Figura 9, observa-se um período de valorização da ação entre novembro de 2023 e janeiro de 2024, com os preços se aproximando de R$14,00. No entanto, em fevereiro, ocorreu uma queda acentuada, que se prolongou nos meses seguintes,

atingindo valores abaixo de R$12,00. Ao relacionar esses dados à Figura 8, nota-se que, entre agosto de 2023 e janeiro de 2024, os sentimentos predominantes nas menções foram mais positivos. Após esse período de alta, porém, houve um aumento na ocorrência de sentimentos negativos, possivelmente refletindo a tendência de queda da ação observada posteriormente.



Fonte: InfoMoney

Figura 9. Ação do Bradesco ao longo do período analisado

4.4. Influencia da *playlist* Rumo ao Bilhão na Carteira Valor

Após a definição do modelo de classificação que apresentou o melhor desempenho, treinado com a técnica *SMOTE*, foram extraídos os vídeos da *playlist* Rumo ao Bilhão referentes ao período de fevereiro a maio de 2025, para comparação com os ativos presentes na Carteira Valor nos respectivos meses, conforme apresentado na Tabela 5. A ação Vale foi desconsiderada na análise, pois o termo pode aparecer com frequência em falas que não necessariamente se referem à empresa.

CARTEIRA VALOR			
FEVEREIRO	**MARÇO**	**ABRIL**	**MAIO**
ITAU UNIBANCO PN,	ITAU UNIBANCO PN,	ITAU UNIBANCO PN,	ITAU UNIBANCO PN,
PETROBRAS PN,	SUZANO S.A. ON,	SABESP ON,	BANCO DO BRASIL,
WEG ON,	SABESP ON,	TELEF BRASIL ON,	ELETROBRAS ON,
BB SEGURIDADE ON,	JBS ON,	PETROBRAS PN,	JBS ON,
SABESP ON,	TELEF BRASIL ON,	BANCO DO BRASIL,	PETROBRAS PN,
JBS ON,	CPFL ENERGIA ON,	EMBRAER ON,	EMBRAER ON,
BRADESCO PN,	WEG ON,	ITAUSA PN,	SABESP ON,
SUZANO S.A. ON,	EMBRAER ON,	BB SEGURIDADE ON,	BB SEGURIDADE ON,
EMBRAER ON,	BB SEGURIDADE ON,	PETROBRAS ON,	COPEL PNB,
VALE ON	VALE ON	VALE ON	TELEF BRASIL ON,

Tabela 5. Carteira Valor dos meses de fevereiro a maio de 2025

Após a classificação do modelo, foram extraídas as 4 ações mais comentadas nesse período pelo Thiago Nigro e que participam da composição da Carteira Valor: Banco do Brasil, Sabesp, Petrobras e Bradesco. Conforme a Tabela 6, pode-se notar, nesses meses, houve predominância de comentários positivos do Thiago Nigro em relação ao Banco do Brasil, o que pode indicar uma possível influência na Carteira Valor, uma vez que essa ação apareceu nos meses de abril e maio; já a Sabesp teve citação do Thiago Nigro apenas no mês de maio, sendo que a ação aparece em todos os meses na carteira, não sendo possível identificar uma influência nesse caso; a Petrobras e o Bradesco, devido ao baixo número de citações, tornam inviável notar alguma influência.

Ação	Positvo	Negativo	Publicação	Meses na carteira
Banco do Brasil	23	0	fevereiro	abril e maio
Banco do Brasil	2	0	março	abril e maio
Banco do Brasil	8	2	abril	abril e maio
Banco do Brasil	5	2	maio	abril e maio
Sabesb	6	1	maio	fevereiro, março, abril e maio
Petrobras	1	0	fevereiro	fevereiro, abril e maio
Petrobras	1	0	maio	fevereiro, abril e maio
Bradesco	0	1	maio	fevereiro

Tabela 6. Relação entre o Sentimento das Ações Citadas e a Carteira de Valor

4.5. Análise Comparativa entre Falas do Thiago Nigro e Valores das Ações

Para melhor aproveitamento do conhecimento extraído após a classificação do modelo, foi analisada a influência do Thiago Nigro no valor da ação ao longo do período classificado pelo modelo (fevereiro à maio de 2025). Foram selecionadas as 5 ações mais citadas neste período (Banco do Brasil, Boa Safra, Sabesp, Cemig e XP Malls) e submetidas ao classificador, no total, para extrair o sentimento positivo ou negativo; foram selecionadas 125 frases, conforme a tabela 7.

Ação	Positivo	Negativo
Banco do Brasil	38	4
Boa Safra	24	6
Petrobras	2	0
Sabesp	6	1
Cemig	25	5
XP Malls	13	1

Tabela 7. Sentimentos das ações período de fevereiro a maio de 2025

Foi realizada a extração dos valores das respectivas ações nos meses analisados e desenvolvido um algoritmo para comparar o comportamento dos ativos após a data exata em que foram mencionados por Thiago Nigro. Dessa forma, buscou-se identificar se, após uma citação positiva ou negativa, o ativo apresentou movimento de alta ou de baixa. A Tabela 8 apresenta os indicadores D+1, D+3, D+7, D+15 e D+30, que representam as janelas temporais utilizadas para avaliar o comportamento das ações após as citações. Por exemplo, o indicador D+1 corresponde ao desempenho do ativo um dia após a menção. Assim, supondo que uma ação estivesse cotada a 10,00 reais no dia da citação e, após uma menção positiva realizada por Thiago Nigro, passasse a valer 11,00 reais no dia seguinte, seria registrado um acerto para o indicador D+1, pois a valorização do ativo foi compatível com o sentimento identificado. De forma análoga, quando uma menção negativa é seguida por uma desvalorização do ativo, também é registrado um acerto para a respectiva janela temporal.

Observa-se que os resultados relacionados ao Banco Safra apresentaram os índices mais expressivos, atingindo 100% de acertos nos diferentes períodos analisados. Destaca-se também o desempenho observado para o Banco do Brasil no sétimo dia após a análise, no qual foi registrado um índice de acerto de 75%.

	D+1	D+3	D+7	D+15	D+30
Banco do Brasil	62,50%	57,14%	75,00%	71,43%	12,50%
Boa Safra	100%	100%	100%	100%	100%
Sabesp	0%	0%	0%	100%	0%
Cemig	50%	66,67%	50%	25%	50%
XP Malls	25%	44,44%	44,44%	22,22%	20%

Tabela 8. Média de acertos por ação

A Tabela 9 apresenta a média ponderada de acertos obtida na comparação entre os sentimentos identificados nas citações dos influenciadores e a direção da variação dos preços das ações nas diferentes janelas temporais (D+X), que representam o comportamento dos ativos após a data da citação. Observa-se que a janela D+7 apresentou o melhor desempenho, com média de acertos de 58,33%. Esse resultado sugere que a associação entre os sentimentos identificados e a variação dos preços das ações tende a ser mais evidente aproximadamente uma semana após as citações realizadas pelos influenciadores.

Geral	D+1	D+3	D+7	D+15	D+30
	47%	54,55%	58,33%	47,83%	28%

Tabela 9. Média ponderada de acertos

A Figura 10 apresenta a distribuição dos percentuais de acerto das ações que obtiveram os resultados mais representativos, destacando Banco do Brasil, Cemig e XP Malls, além da média geral. Observa-se que o Banco do Brasil apresentou os maiores índices de acerto, atingindo aproximadamente 75% na janela D+7 e mantendo desempenho superior à média até D+15. A Cemig apresentou melhor desempenho na janela D+3, enquanto a XP Malls obteve índices inferiores às demais ações em praticamente todas as janelas temporais. Nota-se ainda que, após D+7, a média geral passou a apresentar redução gradual, chegando a 28% em D+30. Esse comportamento sugere que eventuais associações entre as citações dos influenciadores e a variação dos preços dos ativos tendem a ser mais perceptíveis no curto prazo, perdendo força à medida que o intervalo de análise aumenta.

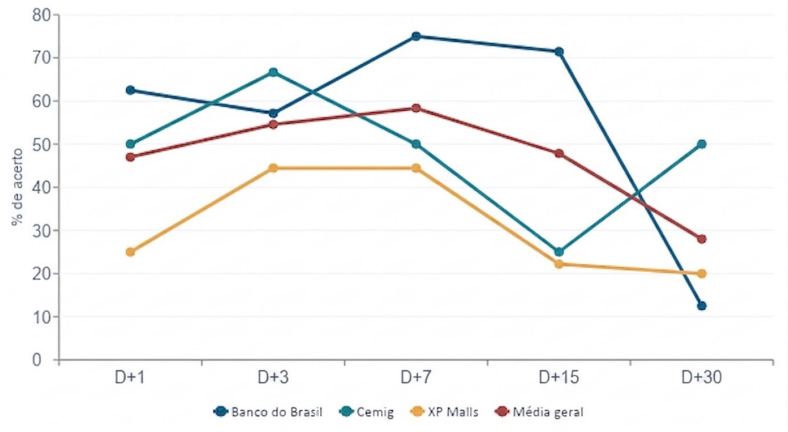


Figura 10. Compilado de acertos

5. Conclusões e Trabalhos Futuros

Este trabalho teve como objetivo principal a extração, análise e processamento de dados textuais provenientes do YouTube, com o intuito de extrair conhecimento a partir de conteúdos produzidos por influenciadores digitais do mercado financeiro e comparar com recomendações da Carteira Valor e com as variações nos valores das ações. Para isso, foram aplicadas técnicas de Processamento de Linguagem Natural (PLN) e métodos de classificação de textos. A metodologia adotada envolveu a coleta dos dados, o pré-processamento textual, a aplicação do classificador *Random Forest* e, posteriormente, a análise dos resultados obtidos a partir dos dados coletados e das classificações.

A partir dos experimentos realizados, foi possível identificar padrões relevantes na polaridade das falas analisadas, bem como avaliar o impacto do tratamento da base de dados no desempenho dos modelos, especialmente com a aplicação de técnicas de balanceamento. Além disso, verificou-se que a análise temporal das menções às ações pode fornecer indícios sobre o comportamento dos ativos no curto prazo, contribuindo para uma melhor compreensão da relação entre o discurso e o mercado.

Entretanto, um dos principais entraves encontrados durante o desenvolvimento do trabalho foi o tamanho reduzido da base de dados utilizada, o que pode limitar a capacidade de generalização dos resultados. Bases de dados maiores tendem a proporcionar modelos mais robustos e com maior capacidade de aprendizado, o que pode impactar positivamente o desempenho dos algoritmos de classificação.

Como sugestão para trabalhos futuros, destaca-se a ampliação da base de dados por meio da inclusão de conteúdos de mais influenciadores digitais. Recomenda-se também a realização de análises mais específicas das frases classificadas, avaliando separadamente as polaridades positiva e negativa ao longo do tempo, o que pode possibilitar uma compreensão mais detalhada da evolução dos sentimentos nos conteúdos analisados. Outra possibilidade consiste em realizar análises segmentadas dos ativos mencionados, avaliando separadamente conteúdos relacionados a fundos imobiliários e ações, o que permite identificar diferenças no comportamento e na polaridade das mensagens associadas a cada tipo de investimento. Além disso, sugere-se a aplicação de testes e cálculos estatísticos para complementar a análise dos resultados, possibilitando uma avaliação mais robusta. Por fim, propõe-se a utilização e a comparação de outros algoritmos de classificação, com o objetivo de identificar possíveis melhorias no desempenho dos modelos e na qualidade dos resultados.

Dessa forma, o presente trabalho demonstra o potencial da aplicação de técnicas de mineração de texto e de aprendizado de máquina na análise de conteúdos digitais, contribuindo para o avanço de estudos que buscam compreender a influência de influenciadores no mercado financeiro.

Referências

[FIn 2025] (2025). Quem mais fala de investimento nas redes sociais. `https://www.anbima.com.br/pt_br/especial/influenciadores-de-investimentos.htm`. Acesso em: 09 junho 2025.

[B3 2025] B3 (2025). `https://www.b3.com.br/pt_br/para-voce`. Acesso em: 11 novembro 2024.

[Chen et al. 1998] Chen, S., Beeferman, D., and Rosenfeld, R. (1998). Evaluation metrics for language models. Master's thesis, School of Computer Science Carnegie Mellon Universit, Pittsburgh, PA.

[Cunto 2021] Cunto, M. F. (2021). A divulgação de conteúdo pelo canal primo rico no youtube: gatilhos mentais na educação financeira. Master's thesis, Universidade Franciscana. Monografia submetida ao curso de Publicidade e Propaganda, UFN.

[Hu and Szymczak 2023] Hu, J. and Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, 24(2):bbad002.

[IBM 2023] IBM (2023). https://www.ibm.com/br-pt/think/topics/sentiment-analysis. Acesso em: 16 novembro 2024.

[InfoMoney25 2025] InfoMoney25 (2025). Cotações b3: Ações. https://www.infomoney.com.br/cotacoes/b3/acao/. Acesso em: 20 junho 2025.

[INVESTIDOR10 2025] INVESTIDOR10 (2025). https://investidor10.com.br/setores/. Acesso em: 20 janeiro 2025.

[Junior 2023] Junior, E. (2023). Principais métricas de avaliação de modelos em machine learning. https://medium.com/data-hackers/principais-m%C3%A9tricas-de-classifica%C3%A7%C3%A3o-de-modelos-em-machine-learning-94eeb4b40ea9. Acesso em: 20 maio 2025.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Piloneto et al. 2020] Piloneto, J., Vieira, M., and Hecht, R. (2020). Análise preditiva do mercado de ações com processamento de linguagem natural. Master's thesis, Centro Universitário Internacional Uninter, Curitiba, Brasil/Paraná. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação), orientado por Luciano Medeiros.

[Pinto 2024] Pinto, G. C. M. (2024). Random forest na análise de sentimentos em jogos digitais. Artigo (Especialização em Data Science & Analytics), orientado por Éldman de Oliveira Nunes.

[PyPI 2024] PyPI (2024). Youtube transcript api. https://pypi.org/project/youtube-transcript-api/. Acesso em: 14 outubro 2024.

[Rice University 2023] Rice University (2023). 12 melhores linguagens de programação para ciência de dados e análise. https://csweb.rice.edu/academics/graduate-programs/online-mds/blog/programming-languages-for-data-science. Acesso em: 10 setembro 2024.

[Santos et al. 2023] Santos, L., Bianchi, R., and Costa, A. (2023). Finbert-pt-br: Análise de sentimentos de textos em português do mercado financeiro. In *Anais do II Brazilian Workshop on Artificial Intelligence in Finance*, pages 144–155, Porto Alegre, RS, Brasil. SBC.

[Silva et al. 2024] Silva, D., Oliveira, A., and Junior, P. A. (2024). Análise de sentimentos de tweets em relação à eleição presidencial de 2022 no brasil. In *Anais da XII Escola*

Regional de Computação do Ceará, Maranhão e Piauí, pages 21–30, Porto Alegre, RS, Brasil. SBC.

[Sinoara et al. 2021] Sinoara, R. A., Marcacini, R. M., and Rezende, S. O. (2021). Mineração de textos e semântica: desafios, abordagens e aplicações. In *Revista de Sistemas de Informação da FSMA*, pages 41–53. FSMA.

[Valor Econômico 2025] Valor Econômico (2025). Carteira valor. `https://infograficos.valor.globo.com/carteira-valor/`. Acesso em: 11 novembro 2024.