

Preços do Café e Narrativas Midiáticas: uma análise de sentimentos no mercado brasileiro entre 2014 a 2025

Coffee Prices and Media Narratives: a sentiment analysis in the Brazilian market from 2014 to 2025

Milena Magalhães Oliveira

Embrapa Agricultura Digital, Campinas, SP, Brasil

Vitor Manoel Tomé do Nascimento

Universidade de São Paulo, Piracicaba, SP, Brasil

SOBER GT13. Estudos Interdisciplinares no rural

Resumo: Este estudo examina a relação entre a volatilidade das cotações do café e o sentimento presente em reportagens publicadas entre 2014 e 2025. O corpus analisado compreende 8.926 notícias coletadas nos portais Canal Rural e CaféPoint por meio de wescrapping e padronização textual. A metodologia combinou mineração de texto e processamento de linguagem natural, com análise de sentimentos pelo método VADER adaptado ao português e reconhecimento de entidades nomeadas para mapeamento dos principais atores do setor. Os resultados apontam fraca co-movimentação entre intensidade da cobertura e polaridade média, correlação negativa moderada entre sentimento e preços, e elevada associação entre as cotações do arábica e da robusta — sugestiva de fatores estruturais comuns na formação de preços. O trabalho reconhece limitações relativas à granularidade temporal anual, ao viés de corpus e à ausência de variáveis de controle fundamentais. Ainda assim, oferece evidências descritivas sobre o papel informacional da mídia no mercado cafeeiro e indica direções para investigações futuras, como análises em frequências temporais mais elevadas, aprimoramento dos modelos de sentimento para o português e incorporação de variáveis climáticas e de fundamentos de mercado.

Palavras-chave: café, análise de sentimento, mineração de texto.

Abstract: This study examines the relationship between coffee price volatility and the sentiment expressed in news articles published between 2014 and 2025. The corpus comprises 8,926 articles collected from the Canal Rural and CaféPoint portals through web scraping and text standardization. The methodology combined text mining and natural language processing, using a Portuguese-adapted VADER sentiment analysis and named entity recognition to map key market actors. Results indicate weak co-movement between coverage intensity and average polarity, a modest negative correlation between sentiment and prices, and a strong association between arabica and robusta quotations — suggesting common structural drivers in price formation. The study acknowledges limitations related to annual temporal granularity, corpus bias, and the absence of fundamental control variables. Nevertheless, it provides descriptive evidence on the informational role of media in the coffee market and outlines directions for future research, including higher-frequency analyses, improved sentiment models for Portuguese, and the incorporation of climatic and market fundamental variables.

Keywords: coffee, sentiment analysis, text mining.

1. Introdução

O Brasil é o maior produtor e exportador mundial de café, commodity que historicamente ocupa papel central na economia nacional, tanto pela geração de divisas quanto pela criação de empregos ao longo de uma cadeia produtiva ampla e diversificada (Prado Júnior, 2000; Furtado, 2005). Além de sua dimensão econômica, o café integra a identidade cultural do país.

Os preços dessa commodity são estruturalmente voláteis. Fatores como choques macroeconômicos, variações cambiais e desequilíbrios entre oferta e demanda nos mercados globais exercem pressão contínua sobre as cotações (Gilbert e Morgan, 2010; Balcombe, 2009). No Brasil, eventos climáticos extremos tais como geadas, secas e irregularidades no regime de chuvas nas principais regiões produtoras representam um vetor adicional de instabilidade, com impacto direto sobre o volume das safras e, conseqüentemente, sobre os preços internacionais (Camargo, 2009; Ripol, 2016). A demanda global relativamente estável, combinada a uma oferta agrícola rígida no curto prazo, amplifica esse desequilíbrio e condiciona as estratégias de produtores, exportadores e investidores do setor (Gomes et al., 2020; Tavares et al., 2018).

Nesse ambiente de incerteza, a mídia exerce influência relevante sobre a formação de expectativas dos agentes econômicos. O modo como o noticiário enquadra as oscilações do mercado pode amplificar percepções, induzir vieses cognitivos e, em alguma medida, retroalimentar a própria volatilidade que descreve (Tetlock, 2007; Shiller, 2019). Compreender essa dinâmica é, portanto, fundamental para interpretar os mecanismos de transmissão entre informação, narrativa e precificação de commodities.

Este artigo investiga a relação entre a volatilidade das cotações do café e o sentimento presente nas reportagens publicadas nos portais Canal Rural e CaféPoint entre 2014 e 2025. A questão central que orienta a pesquisa é: a cobertura jornalística adota um tom predominantemente mais negativo em períodos de queda nos preços, e essa reação é proporcionalmente mais intensa do que a positividade observada em momentos de alta?

Para respondê-la, foram aplicadas técnicas de Processamento de Linguagem Natural (NLP), com análise de sentimentos baseada no método léxico VADER, adaptado para o português, e Reconhecimento de Entidades Nomeadas (NER), permitindo identificar tanto o tom das narrativas quanto os principais atores mencionados na cobertura do setor.

O artigo está organizado em cinco seções. A seção 2 apresenta a revisão da literatura, abordando a volatilidade do mercado cafeeiro, a relação entre mídia e expectativas econômicas,

e os fundamentos metodológicos da análise de sentimentos. A seção 3 descreve os procedimentos metodológicos e o tratamento dos dados. A seção 4 discute os resultados obtidos. Por fim, as considerações finais sintetizam as principais conclusões e apontam direções para pesquisas futuras.

2. Revisão da Literatura

2.1 O Mercado Cafeeiro Brasileiro e a Volatilidade de Preços

O café constitui um dos pilares históricos da formação econômica do Brasil desde o século XIX, quando sua produção e exportação passaram a desempenhar papel estruturante no desenvolvimento nacional (Prado Júnior, 2000; Furtado, 2005). Atualmente, o país mantém a liderança mundial na produção e exportação da commodity, com reflexos expressivos na balança comercial e na geração de emprego e renda em diversas regiões.

A volatilidade dos preços do café é fenômeno estrutural, resultante da inelasticidade de curto prazo da oferta e da demanda típica de commodities agrícolas (Gilbert e Morgan, 2010; Balcombe, 2009). No Brasil, eventos climáticos extremos em regiões como o Cerrado Mineiro, o Sul de Minas e a Mogiana geram choques de oferta com repercussão imediata nas cotações internacionais (Camargo, 2009; Ripol, 2016). O descompasso entre essa oferta rígida e uma demanda global relativamente estável cria condições propícias à formação de picos de preço, situação que tende a se agravar com o avanço das mudanças climáticas (Gomes et al., 2020; Tavares et al., 2018).

2.2 Mídia, Sentimento e Formação de Expectativas Econômicas

A relação entre cobertura midiática e comportamento dos mercados financeiros e de commodities tem sido objeto crescente de investigação na literatura de finanças comportamentais e economia da informação. O trabalho seminal de Tetlock (2007) demonstrou empiricamente que o tom do noticiário financeiro, medido por análise de sentimentos, possui capacidade preditiva sobre o comportamento do mercado acionário norte-americano, sugerindo que variáveis qualitativas derivadas da mídia carregam informações relevantes que não estão completamente incorporadas aos preços.

Shiller (2017, 2019) amplia essa perspectiva ao desenvolver o conceito de narrativas econômicas, argumentando que histórias e discursos disseminados pela mídia e pela opinião pública influenciam decisões econômicas de forma sistemática, podendo acelerar ou atenuar ciclos de expansão e retração. Para Shiller, as narrativas não são meros reflexos passivos da realidade econômica, mas agentes ativos na construção das expectativas dos agentes, com potencial de retroalimentar a própria volatilidade que descrevem.

No campo da economia comportamental, essa abordagem dialoga diretamente com a teoria dos vieses cognitivos e verifica a percepção de risco e a disposição para agir dos agentes econômicos (Kahneman, 2011, apud Shiller, 2019). Essa assimetria é relevante no contexto das commodities agrícolas, cujos mercados envolvem agentes com diferentes níveis de acesso à informação e distintas capacidades de processá-la.

2.3 Análise de Sentimentos e Processamento de Linguagem Natural em Contextos Econômicos

A análise de sentimentos, ramo do Processamento de Linguagem Natural (NLP¹), consiste na identificação e classificação automática do tom emocional ou avaliativo de textos, categorizando-os como positivos, negativos ou neutros. Dentre as abordagens metodológicas disponíveis, os métodos baseados em léxico, sendo essas que associam palavras e expressões a valores de valência afetiva e se destacam pela interpretabilidade e pela aplicabilidade a grandes volumes de texto (Tetlock, 2007).

O método Valence Aware Dictionary and sEntiment Reasoner - VADER, originalmente desenvolvido para textos em inglês por Hutto e Gilbert (2014), tem sido adaptado e aplicado a em outras línguas, incluindo o português, em estudos que buscam capturar o sentimento expresso em textos jornalísticos e em mídias sociais. A aplicação ao contexto econômico permite operacionalizar de forma sistemática e reproduzível a dimensão qualitativa do discurso midiático, convertendo-a em variáveis passíveis de análise quantitativa.

De maneira complementar, o Reconhecimento de Entidades Nomeadas NER² constitui técnica de mineração de texto que possibilita identificar automaticamente menções a atores específicos, sejam eles empresas, governos, instituições financeiras, organizações internacionais, mapeando sua frequência e seu contexto de aparição. A combinação de análise de sentimentos com NER permite não apenas mensurar o tom geral da cobertura midiática, mas também compreender quais atores são associados a narrativas positivas ou negativas, enriquecendo a interpretação dos resultados.

3. Metodologia

Este estudo adotou uma abordagem quantitativa e computacional para investigar a relação entre a volatilidade dos preços do café e o sentimento associado às notícias publicadas sobre o setor. Para tanto, foram utilizadas técnicas de Processamento de Linguagem Natural

¹ Natural Language Processing

² Named Entity Recognition

[NLP], mineração de dados textuais e análise de séries temporais, permitindo a extração e interpretação automatizada de grandes volumes de dados econômicos e noticiosos.

A metodologia foi estruturada em três principais etapas:

- (i) Coleta de Dados;
- (ii) Tratamento dos dados de texto;
- (iii) Análise de Sentimentos nas Notícias e

A obtenção dos dados foi conduzida por meio de métodos automatizados de web scraping e APIs de acesso a bases econômicas e textuais. A coleta foi dividida em duas principais fontes: notícias sobre café e dados econômicos do preço do café. O conjunto de textos foi extraído de veículos de comunicação especializados em economia e no setor agrícola. Entre os principais repositórios de informações considerados os portais agrícolas especializados Canal Rural e CaféPoint.

A coleta de notícias foi efetuada utilizando Web Scraping (BeautifulSoup e Scrapy). Web scraping é o processo de automatizar a coleta de dados de páginas web. Ele envolve o uso de scripts que simulam a navegação de um navegador, extraindo conteúdo e dados de sites que normalmente seriam visualizados por usuários humanos (Mitchel, 2018). Os textos foram armazenados em um banco de dados estruturado e submetidos a pré-processamento para remoção de ruído textual, eliminação de duplicatas e normalização de caracteres.

Após a coleta, os textos foram submetidos a uma análise de sentimentos automatizada, baseada em técnicas de Processamento de Linguagem Natural [NLP] e aprendizado de máquina, com o objetivo de classificar o sentimento das notícias e identificar se apresentam viés positivo, negativo ou neutro em relação ao mercado de café. O processamento de linguagem natural é uma área da ciência da computação e da inteligência artificial que utiliza aprendizado de máquina para que os computadores possam entender textos e palavras faladas da mesma forma que humanos (Chowdhary, 2020; Stryker & Holdsworth, (s.d.)). Esse processo utilizou abordagens léxicas e baseadas em modelos de aprendizado profundo para garantir maior precisão na identificação dos sentimentos expressos nos textos.

3.1 Tratamento dos dados

Antes da aplicação da ferramenta de análise de sentimentos, os dados textuais passaram por um processo de pré-processamento e normalização, etapa necessária para assegurar a consistência das análises subsequentes. A base de dados foi composta por 8.926 notícias coletadas nas fontes : Canal Rural (seção dedicada ao segmento de café) e Café Point, no período de 2014 a 2025, o que possibilitou a construção de uma perspectiva longitudinal das

narrativas midiáticas sobre o setor cafeeiro. Cada notícia foi extraída com informações estruturadas, incluindo data de publicação, título, conteúdo integral e URL de origem, de modo a garantir rastreabilidade e reprodutibilidade.

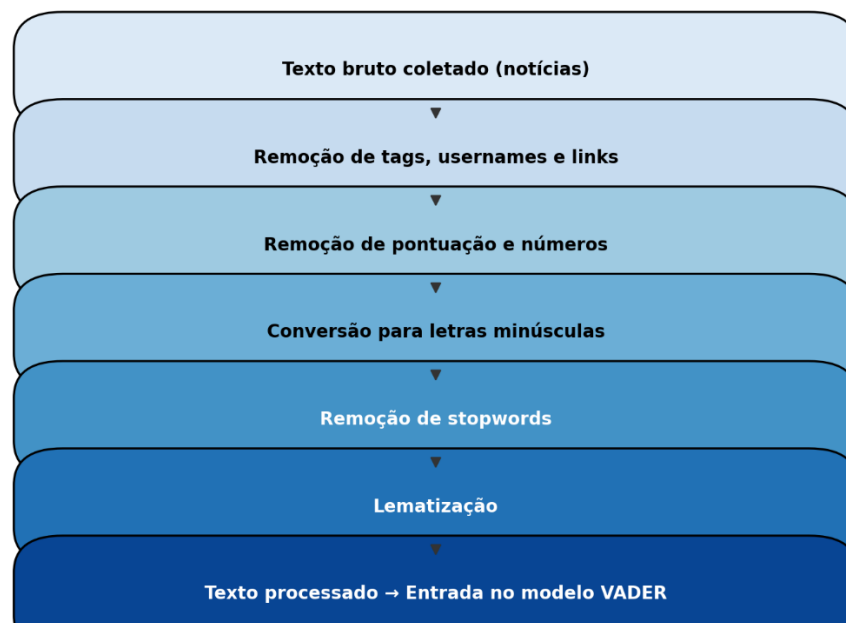
O tratamento inicial dos textos iniciou-se com a conversão integral para letras minúsculas, a remoção de acentos e diacríticos por meio da normalização Unicode e a exclusão de URLs, números, caracteres especiais, quebras de linha e repetições consecutivas de palavras. Essa etapa reduziu os ruídos característicos do texto jornalístico e promoveu a uniformização da estrutura textual. Ademais, implementou-se um dicionário de substituições para contrações informais da língua portuguesa, como “tá” transformado em “está” e “pra” em “para”, a fim de aumentar a precisão da lematização e da análise de sentimentos.

A filtragem de termos irrelevantes foi conduzida mediante a utilização da lista padrão de “stopwords” disponibilizada pelo spaCy para português, complementada por uma lista personalizada criado pela autora com palavras de elevada frequência, mas baixo valor semântico no contexto da pesquisa, como “café”, “preço”, “milhões” e “ontem”. Essa alteração reduziu a probabilidade de enviesamento decorrente de tokens repetitivos e semanticamente pouco informativos. Após a normalização, aplicou-se prioritariamente a lematização via spaCy, que reduziu cada termo à sua forma canônica. Paralelamente, utilizou-se o stemmer RSLP do NLTK em análises complementares, com o propósito de verificar a robustez dos resultados frente a diferentes estratégias de redução morfológica.

Os textos resultantes foram tokenizados e reconstruídos em cadeias limpas, que serviram de entrada para o algoritmo de análise de sentimentos. A análise de sentimentos foi realizada por meio de duas abordagens complementares. O modelo VADER [Valence Aware Dictionary and Sentiment Reasoner] é amplamente utilizado na literatura para análise de sentimentos e a mineração de palavras-chaves com reconhecimento de entidades. O VADER é um modelo baseado em léxicos otimizado para a análise de linguagem informal, amplamente empregado na avaliação de sentimentos em redes sociais e notícias financeiras. Destaca-se por sua capacidade de atribuir polaridade e intensidade emocional às expressões textuais, proporcionando uma classificação eficiente e interpretável dos dados textuais (Rosenberg et al, 2023; Suardi, Rasel, Liu, 2022; Hutto, Gilbert, 2014). O índice “compound”, fornecido pelo algoritmo, foi utilizado como métrica de referência para a classificação das notícias, atribuindo rótulos de sentimento negativo para valores inferiores a zero, positivo para valores superiores a zero e neutro para valores iguais a zero.

A segunda abordagem empregou técnicas de mineração de texto, combinando mineração de palavras-chave e Reconhecimento de Entidades Nomeadas [Named Entity Recognition – NER]. A extração de termos relevantes associados a sentimentos positivos e negativos foi realizada por meio do TF-IDF [Term Frequency - Inverse Document Frequency]. Esse método estatístico pondera a importância de palavras em um corpus textual. Paralelamente, o NER permitiu a identificação e categorização de entidades como empresas, órgãos governamentais, instituições financeiras e agentes do setor cafeeiro mencionados nas notícias. Ademais, foram utilizadas visualizações gráficas para identificar padrões sazonais e possíveis tendências na cobertura midiática e nos preços do café, principais palavras associadas e as frequências relacionadas (WordCloud). A Figura 1 sintetiza esse processo:

Figura 1 - Pré-processamento dos dados de texto



Fonte: Dados originais do trabalho.

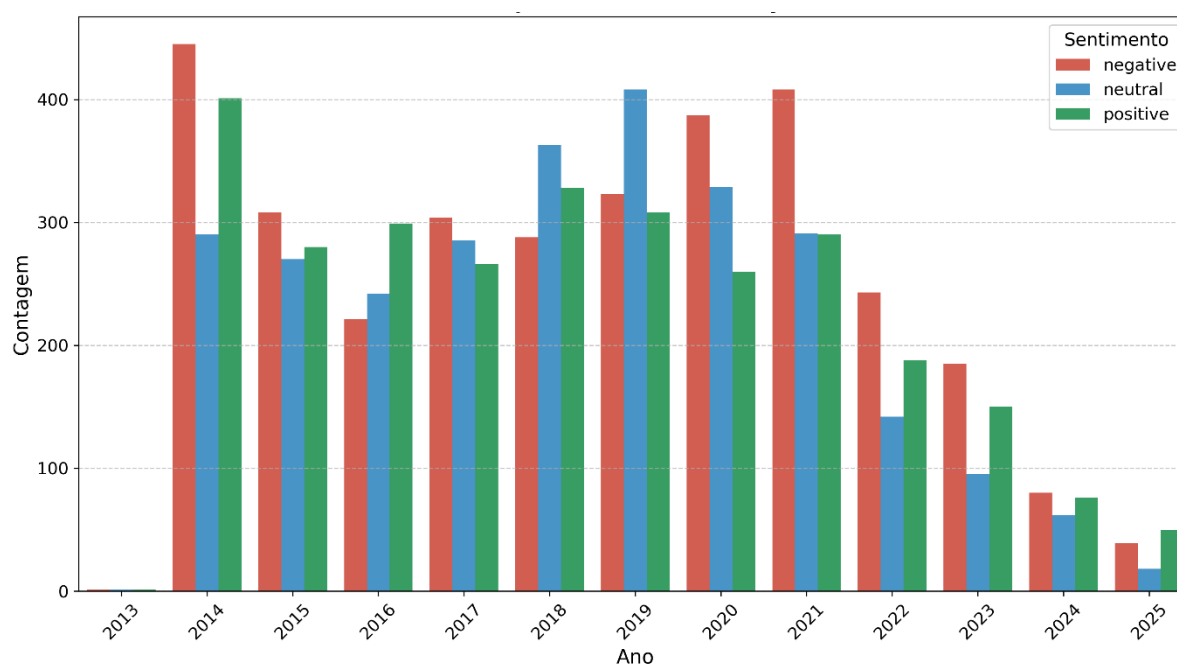
Em suma, ressalta-se que neste estudo a análise de dados foi conduzida utilizando a linguagem Python na plataforma Google Colab, que possibilitou processamento em nuvem e integração eficiente de pacotes especializados. Para o pré-processamento textual empregaram-se as bibliotecas re, nltk, spacy, gensim e unicodedata. A etapa de análise de sentimento utilizou a aplicação da biblioteca nltk.sentiment.vader e da biblioteca leIA, adequada ao idioma português. No tratamento de stopwords e léxico foram aplicados nltk.corpus.stopwords e sklearn.feature_extraction.text. Para representação e análise estatística e econométrica dos

dados foram utilizadas scikit-learn, pandas, numpy e scipy. A visualização foi estruturada por meio de matplotlib, seaborn, wordcloud e plotly, permitindo desde gráficos básicos até representações interativas e de redes. Como suporte adicional, o fluxo de trabalho incorporou tqdm, joblib e pickle para otimização e armazenamento de modelos, além de torch e tensorflow para o suporte a arquiteturas de aprendizado profundo.

4. Resultados e Discussão

A análise da distribuição temporal dos sentimentos expressos revela variações significativas ao longo dos anos, destacando tendências relevantes para a compreensão do contexto estudado. Observa-se através da Figura 2, que os sentimentos negativos predominam de forma consistente, especialmente nos anos de 2014, 2015, 2021 e 2022, indicando uma percepção desfavorável recorrentes durante este período. Em contraste, os sentimentos neutros e positivos apresentam flutuações mais equilibradas, com picos de neutralidade em 2019 e aumento progressivo da positividade a partir de 2018 até 2021. A diminuição gradual das contagens totais a partir de 2022 sugere uma redução na produção ou relevância do conteúdo relacionados ao café, podendo refletir mudanças no interesse ou na dinâmica do fenômeno estudado. A distribuição decrescente e o equilíbrio relativo entre os três tipos de sentimento nos últimos anos indicam uma possível estabilização das percepções, apontando para uma maturação ou mudança qualitativa no ambiente analisado.

Figura 2 - Distribuição de sentimentos por ano

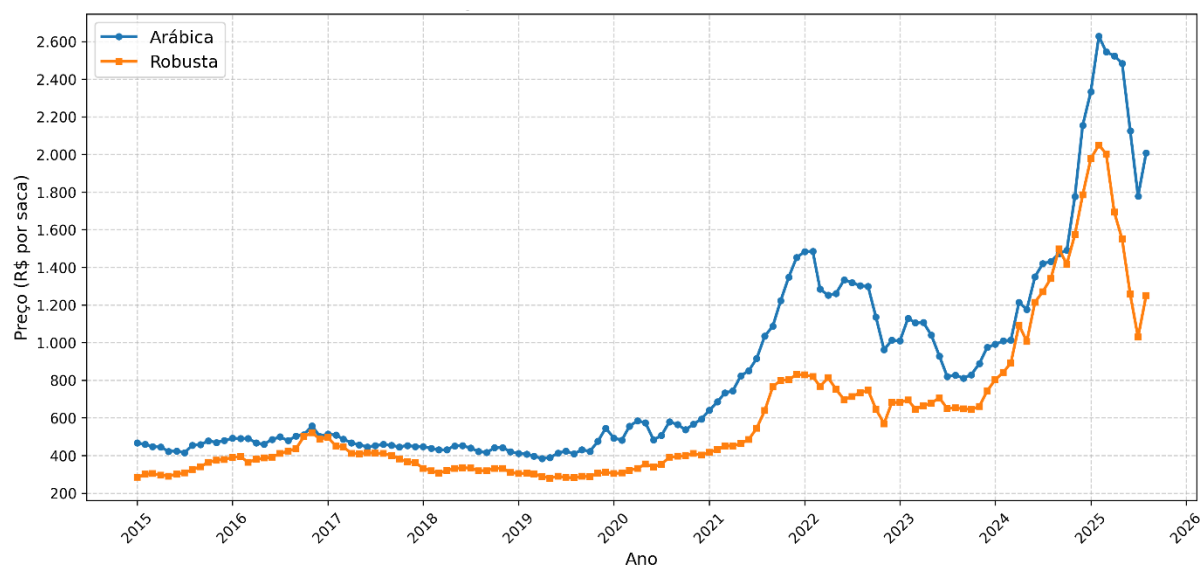


Fonte: Resultados originais da pesquisa.

A predominância de notícias com conotação negativa ao longo do período analisado pode estar relacionada a fatores que impactam diretamente a oferta e demanda do mercado cafeeiro, como eventos climáticos adversos, problemas logísticos ou crises econômicas. Tais elementos, frequentemente reportados com tom negativo, tendem a gerar expectativas de redução na produção ou aumento nos custos de produção, o que pode pressionar o preço do café para cima.

A Figura 3 apresenta gráfico que ilustra a trajetória dos preços do café arábica e robusta no mercado brasileiro entre 2015 e 2025, evidenciando períodos de forte valorização, sobretudo a partir de 2020. A elevação mais acentuada nos preços do arábica coincide com eventos climáticos adversos, como geadas e secas em regiões produtoras, que reduziram a oferta e foram amplamente noticiados com viés negativo. Essa dinâmica reforça a relação discutida no parágrafo anterior, em que a cobertura midiática de eventos desfavoráveis tende a amplificar percepções de escassez e custos crescentes, retroalimentando a volatilidade do mercado. A trajetória do robusta, embora menos volátil, também apresentou elevações significativas no mesmo período, sugerindo que choques de oferta e expectativas negativas não se restringiram a uma única variedade, mas afetaram de forma sistêmica o setor cafeeiro brasileiro.

Figura 3 - Preço do café arábica e robusta em reais

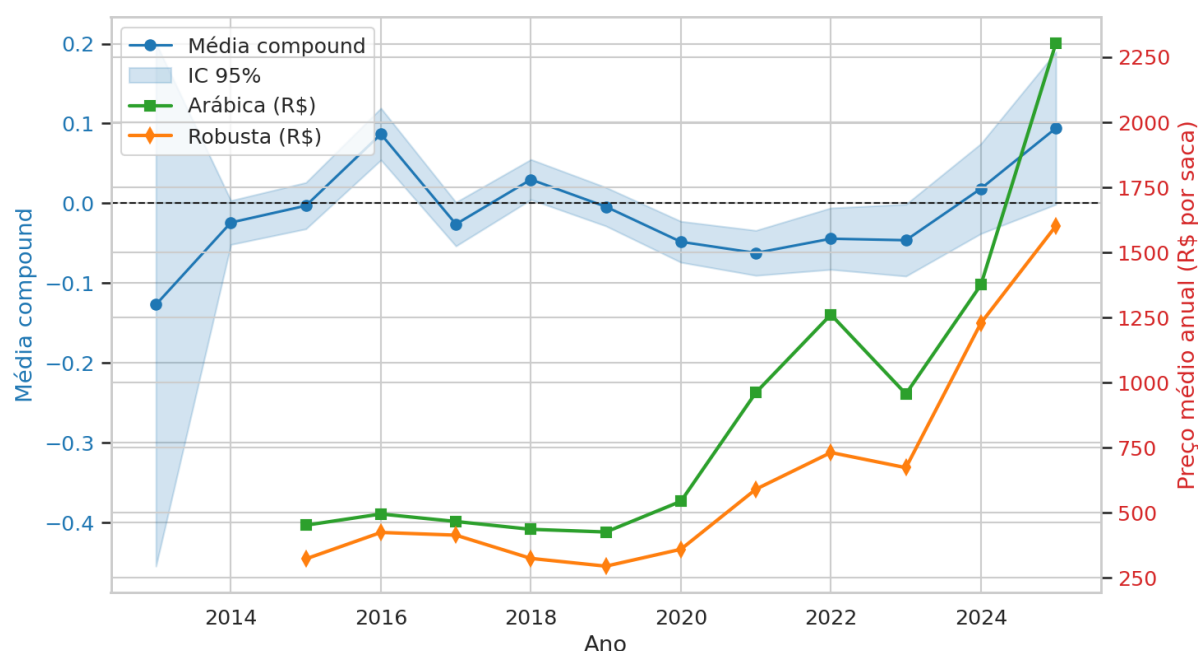


Fonte: Elaborado pela autora a partir de dados CEPEA/Esalq.

A Figura 4 aprofunda a análise ao relacionar a trajetória do sentimento médio captado nas notícias com os preços anuais médios do café arábica e robusta. Nota-se que o período entre 2016 e 2020, caracterizado por predominância de sentimentos negativos, correspondeu a níveis relativamente mais baixos de preços, ao passo que a recuperação gradual do indicador de

sentimento a partir de 2021 ocorreu em paralelo à forte valorização das cotações, especialmente do arábica. Essa correlação sugere que a cobertura midiática não se limita a espelhar choques de oferta e demanda, mas pode atuar como variável explicativa complementar na formação de expectativas de mercado, reforçando tendências já presentes e contribuindo para a amplificação de movimentos altistas ou baixistas no setor cafeeiro.

Figura 4 - Sentimento (compound) vs preço médio do café arábica e robusta em reais



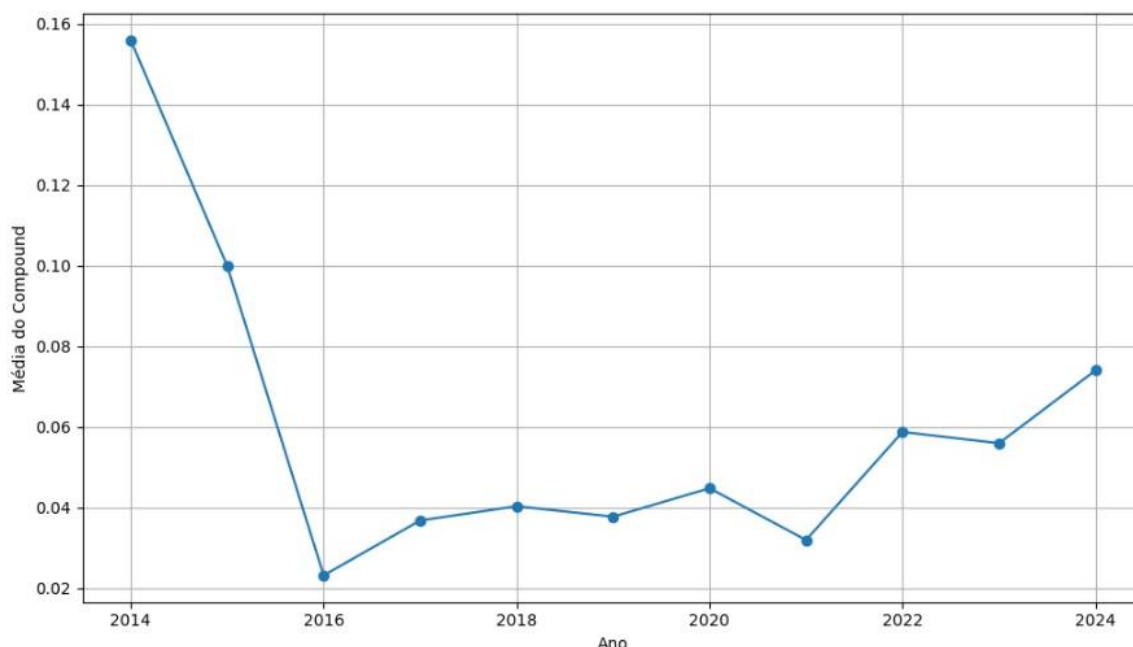
Fonte: Elaborado pela autora a partir de dados CEPEA/Esalq e dados originais da pesquisa.

A Figura 5 apresenta a evolução temporal da média do score compound em notícias relacionadas ao café, no período de 2014 a 2025. O score compound é uma métrica composta que reflete a polaridade geral do sentimento expresso, com valores positivos indicando uma tendência mais positiva, valores negativos indicando tendência negativa, e valores próximos a zero sinalizando neutralidade.

Observa-se uma queda acentuada no score compound (composto) entre 2014 e 2016, passando de aproximadamente 0,16 para pouco acima de 0,02. Esse declínio sugere uma redução significativa no teor positivo das notícias ao longo desses anos iniciais. Entre 2016 e 2021, a pontuação composta se mantém relativamente estável, com leves oscilações em torno de valores baixos, indicando uma predominância de notícias com sentimento mais neutro ou ligeiramente positivo. A partir de 2021, nota-se uma tendência crescente na pontuação composta, alcançando valores próximos a 0,075 em 2025, o que indica uma melhora no tom geral das notícias, com maior presença de conteúdo positivo. Essa evolução pode refletir

mudanças no cenário econômico, político ou social que influenciaram a percepção midiática e o discurso público sobre o tema.

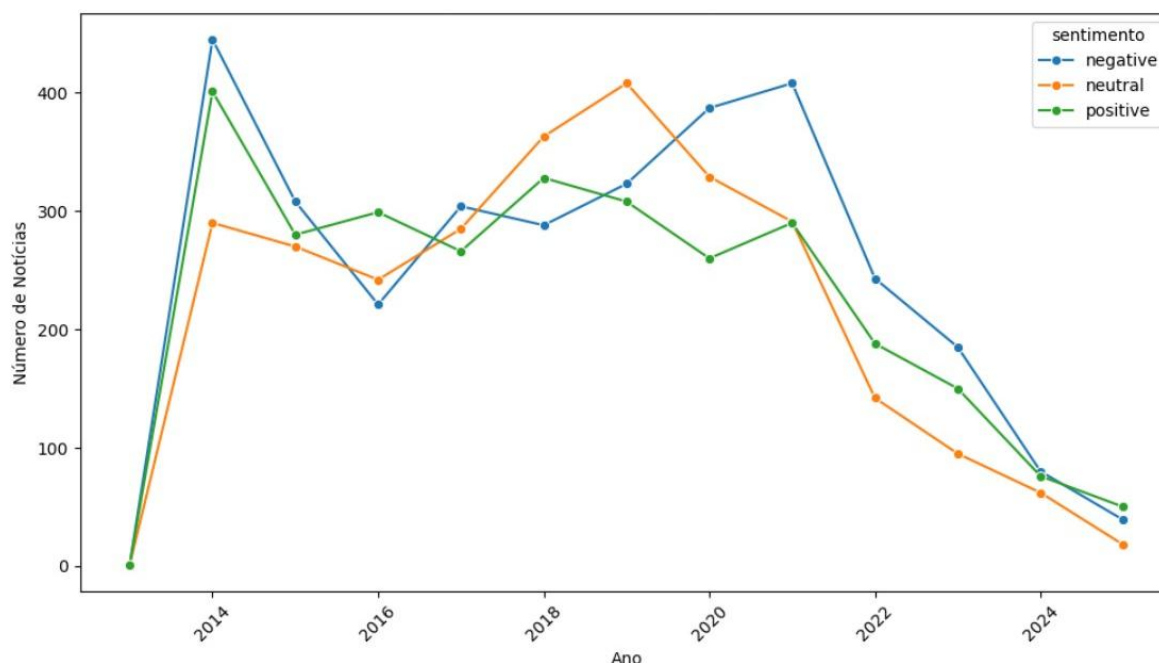
Figura 5 - Média do score compound por ano



Fonte: Resultados originais da pesquisa.

Em seguida, a Figura 6 ilustra a evolução anual do volume absoluto de notícias classificadas em três categorias de sentimento: negativo, neutro e positivo, no mesmo período. Inicialmente, entre 2013 e 2014, há um aumento significativo no volume de notícias em todas as categorias, indicando um maior interesse e cobertura midiática. A partir de 2014 até aproximadamente 2021, observa-se uma predominância das notícias com sentimento negativo, que em alguns anos, como 2019 e 2020, atingem picos superiores a 400 notícias. O volume de notícias neutras e positivas, apesar de considerável, permanece abaixo do volume negativo nesse intervalo. A partir de 2021, todas as categorias apresentam uma tendência de declínio, com destaque para a queda mais acentuada nas notícias negativas e neutras. No entanto, a redução das notícias positivas é menos pronunciada, o que pode estar alinhado com o aumento observado na Figura 3 na pontuação composta, reforçando a percepção de melhora no sentimento geral das notícias, apesar do menor volume total de cobertura.

Figura 6 - Evolução do volume de notícia por sentimento

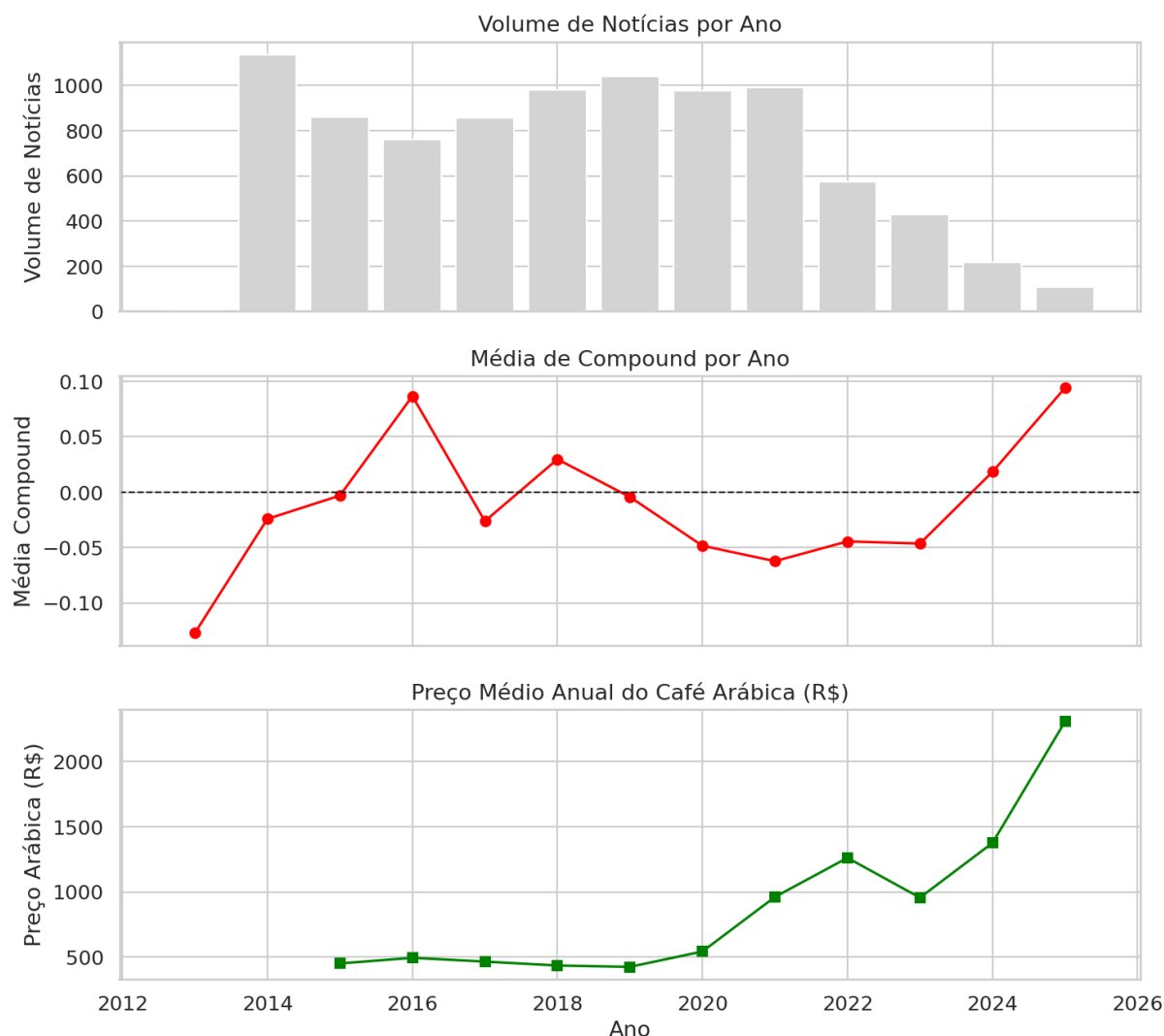


Fonte: Resultados originais da pesquisa.

Observa-se que o volume anual de notícias atingiu seu ápice em 2014, com aproximadamente 1100 registros, seguido por uma queda significativa até 2016 e relativa estabilização entre 2017 e 2019. A partir de 2019, o volume evidencia uma tendência de declínio contínuo, projetando-se um volume inferior a 200 notícias em 2024. Tal comportamento pode refletir uma diminuição do interesse midiático ou da relevância do tema na agenda pública durante os anos recentes.

A Figura 7 combina dados sobre análise de sentimento e o volume de notícias. A pontuação composta, apresentou comportamento inverso ao volume em alguns períodos, evidenciando a independência entre quantidade e qualidade da cobertura. Iniciou-se com valores negativos em 2013, atingindo pico positivo em 2016. A recuperação projetada do sentimento para 2024-2025, apesar do baixo volume de notícias, sugere predominância de conteúdos favoráveis em uma cobertura reduzida.

Figura 7 - Análise temporal do sentimento, preços e volume de notícias sobre café



Fonte: Resultados originais da pesquisa.

Esses resultados indicam que o monitoramento conjunto do volume e do sentimento das notícias oferece uma maior compreensão sobre a percepção pública e o interesse midiático no setor cafeeiro. Nos próximos passos essa análise quantitativa será complementada com estudos qualitativos para identificar eventos e narrativas específicas que influenciam tais tendências.

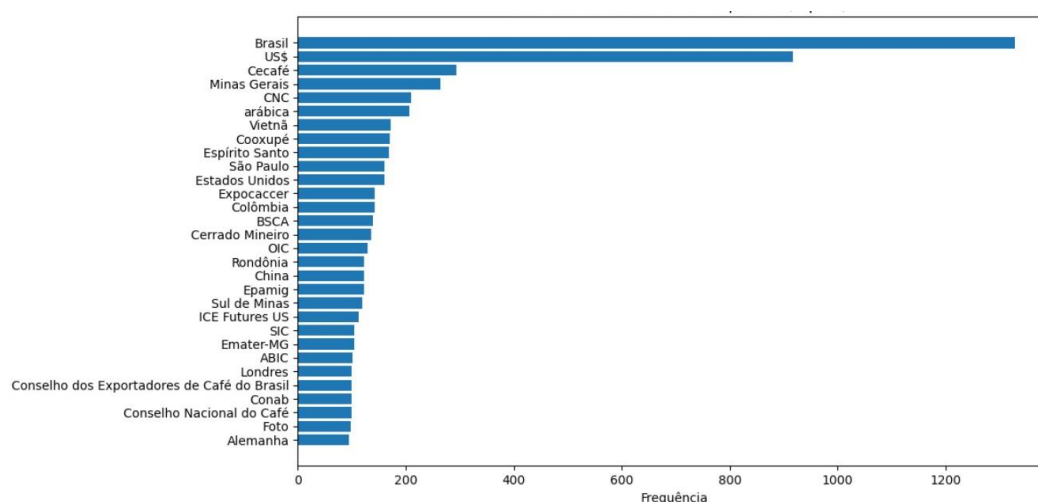
A Figura 7 sintetiza três séries anuais: volume de notícias, polaridade média (compound) e preço médio do arábica. O padrão dominante é a fraca co-movimentação entre volume e polaridade. A polaridade inicia negativa em 2013, alcança valor positivo elevado em 2016 e volta a terreno negativo entre 2020 e 2022, com recuperação em 2024–2025 e novo pico próximo de 0,10. O volume cai de forma monotônica após 2019 e atinge patamar reduzido em 2024–2025. Mantém-se, portanto, evidência de desacoplamento: anos com cobertura intensa

não exibem, necessariamente, tom favorável, e anos recentes de baixa cobertura concentram menções positivas.

A inferência estatística fica limitada por agregação anual e amostra curta. Para pesquisas futuras, este estudo apresenta direcionamentos de reamostrar para frequência mensal e estimar correlações móveis; estimar VAR com choques identificados de “volume” e “polaridade”, controlando preços CEPEA/OIC e variáveis de oferta/demanda (USDA, clima); decompor por tópicos e medir sentimento por tópico para dissociar produção, comércio e finanças; realizar testes de robustez com classificadores em português (VADER-PT, ajustado ao domínio) e dicionários financeiros; e por fim, mapear eventos (choques climáticos e logísticos) em desenho de event study para quantificar direção temporal, persistência e magnitude das relações observadas.

A Figura 8 apresenta um gráfico de barras horizontais com as 30 entidades nomeadas mais frequentes extraídas do conjunto de notícias analisadas sobre o setor cafeeiro. A frequência indica o número de vezes que cada entidade foi mencionada no corpus textual, refletindo sua relevância e presença no discurso midiático.

Figura 8 - Entidades nomeadas mais frequentes



Fonte: Resultados originais da pesquisa.

Destacam-se como principais entidades:

- Brasil, com frequência superior a 1300 menções, evidenciando o protagonismo do país na cobertura midiática sobre café, justificável por ser um dos maiores produtores e exportadores mundiais.

- US\$, segunda entidade mais citada, indicando forte ênfase nas discussões relacionadas a preços, valores e câmbio, aspectos importantes para o setor.

• Instituições e regiões produtoras como Cecafé [Conselho dos Exportadores de Café do Brasil], Minas Gerais, CNC [Confederação Nacional do Comércio], e localidades como Espírito Santo, São Paulo, e Rondônia, que aparecem entre as primeiras posições, reforçando o foco geográfico e institucional da cobertura.

• Tipos e variedades de café, como arábica, aparecem com destaque, o que é coerente com a predominância do café arábica no mercado brasileiro.

• Entidades internacionais, como Estados Unidos, Colômbia, China, e Alemanha, indicam a importância dos mercados consumidores e concorrentes globais nas notícias.

• Organizações e eventos relevantes para o setor, como OIC [Organização Internacional do Café], Expocaccer [principal feira cafeeira da América Latina], e Conab [Companhia Nacional de Abastecimento], também são mencionados, destacando o papel institucional e comercial na dinâmica do mercado.

A presença de termos técnicos e regionais sugere que as notícias abordam tanto aspectos econômicos quanto produtivos, geográficos e institucionais, fornecendo uma visão abrangente do setor.

Por fim, a Figura 9 expõe a matriz de correlações de Pearson anuais. A matriz apresentada foi estimada pelo coeficiente de correlação de Pearson, escolha alinhada ao objetivo de avaliar associações lineares entre variáveis contínuas tais como preços, volume de notícias e escores de sentimento. Pearson é apropriado quando o interesse recai sobre a covariação média entre séries já transformadas em escalas comparáveis e com distribuição próxima ao contínuo, como ocorre após a agregação anual do compound e do volume, além das médias de preços. A métrica também facilita a leitura e a comparação entre pares de variáveis ao produzir valores simétricos e diretamente interpretáveis.

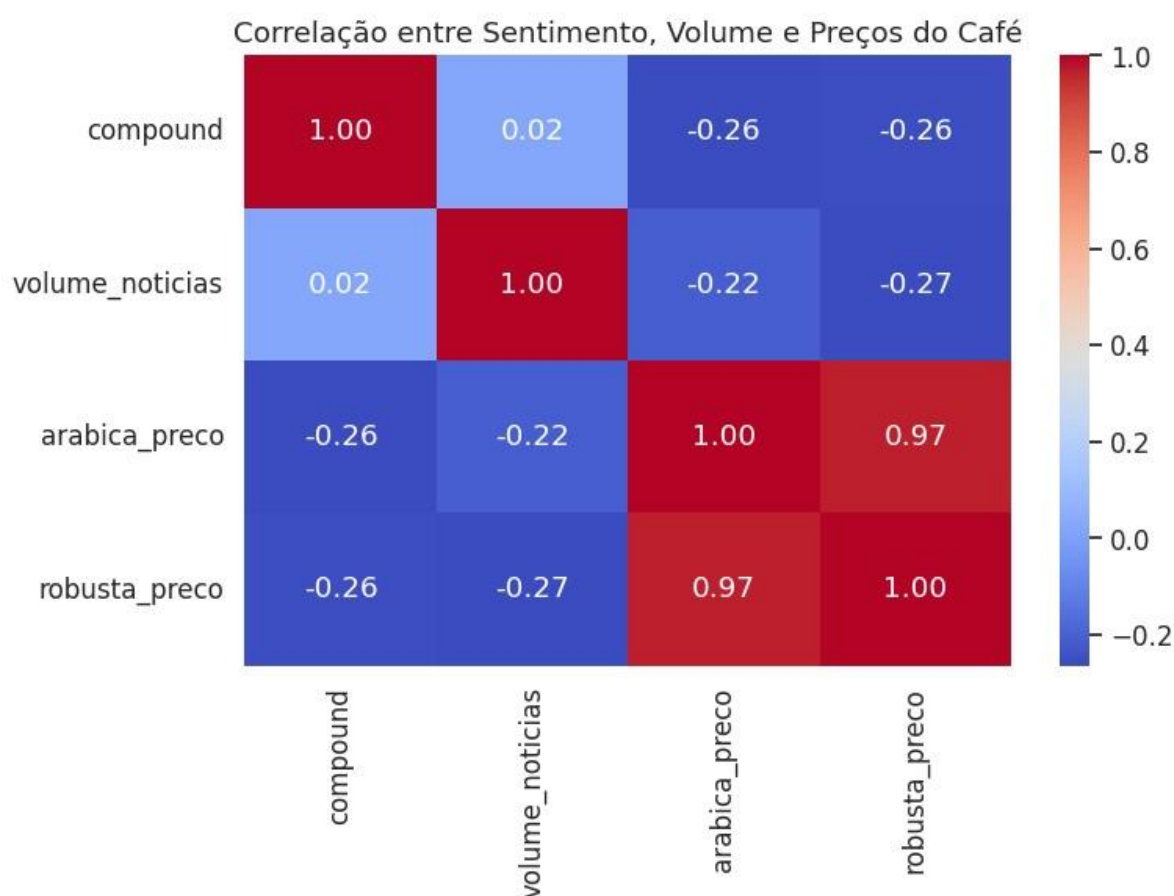
Alternativas como Spearman ou Kendall, baseadas em postos, seriam úteis diante de distribuições altamente assimétricas ou relações monotônicas não lineares. No entanto, a agregação temporal reduz grande parte da assimetria presente na frequência diária de notícias e nas cotações, o que atenua ganhos adicionais dessas medidas. Além disso, a análise aqui proposta busca detectar covariações lineares simples, e não dependências monotônicas mais gerais. Métodos mais complexos, como correlação policórica ou medidas baseadas em informação mútua, exigiriam amostras maiores para produzir estimativas estáveis, o que não se aplica ao conjunto anual disponível.

A opção por Pearson, portanto, atende ao objetivo descritivo do estudo e preserva interpretabilidade, sem sugerir qualquer inferência causal. A amostra curta limita potencia

estatística e pode mascarar relações defasadas, motivo pelo qual os coeficientes devem ser entendidos apenas como uma leitura preliminar da estrutura linear de associações entre sentimento, cobertura jornalística e preços do café.

A relação entre tom médio das notícias (compound) e volume é praticamente nula (0,02), indicando que intensidade da cobertura não acompanha o sentimento. O sentimento apresenta correlação negativa e fraca com os preços do arábica e do robusta (-0,26 em ambos), sinal típico de períodos em que choques de oferta e risco elevam cotações enquanto a narrativa jornalística enfatiza perdas e incertezas. O volume de notícias também se associa negativamente aos preços (-0,22 e -0,27), compatível com janelas de estresse de mercado. Arábica e robusta exibem correlação muito alta (0,97), refletindo integração e fatores comuns de formação de preços. A interpretação permanece descritiva: correlações não implicam causalidade e a agregação anual, com amostra curta, reduz potência estatística e pode ocultar defasagens relevantes.

Figura 9 - Correlação entre sentimento, volume e preços do café



Fonte: Resultados originais da pesquisa.

5. Considerações finais

O trabalho teve como objetivos: (i) construir uma base sistematizada de notícias especializadas sobre o mercado de café; (ii) extrair medidas de sentimento e intensidade de cobertura por meio de técnicas de mineração de texto; (iii) examinar se as narrativas jornalísticas apresentam alguma associação com a volatilidade e o comportamento dos preços do arábica e do robusta; e (iv) identificar, via reconhecimento de entidades nomeadas, quais atores são mais recorrentes na cobertura e como essa presença se relaciona ao tom das reportagens. Esses objetivos foram atendidos de forma consistente, permitindo integrar métricas textuais de sentimento e volume de notícias ao comportamento dos preços e gerar uma leitura clara da estrutura de co-movimentações no período analisado.

O estudo integra métricas textuais de sentimento e volume de notícias ao comportamento dos preços do café e indica fraca co-movimentação entre intensidade da cobertura e polaridade média, correlação negativa modesta entre sentimento e preços compatível com narrativas de risco em janelas de restrição de oferta, e associação quase perfeita entre arábica e robusta que remete a fatores comuns de formação de preços; a evidência permanece descritiva e orienta hipóteses testáveis sobre o papel informacional da mídia no mercado cafeeiro.

As limitações deste trabalho decorrem da agregação anual que comprime choques intra- anuais e reduz a capacidade de dedução estatística, da mensuração de sentimento com modelos não treinados no léxico agrofinanceiro e suscetíveis a ironia e negações, da cobertura do corpus com viés de seleção e duplicatas, da endogeneidade entre preços e notícia que impede identificação de direção causal com correlações simples, da omissão de fundamentos de oferta, demanda, finanças e clima capazes de enviesar estimativas e do uso de preços nominais sem deflacionamento nem integração a futuros ICE.

Pesquisas futuras podem avançar em relação a este trabalho ao incluir reamostragem para frequência mensal ou diária com estimação de correlações móveis; identificação de choques de “volume” e “polaridade” em VAR, TVP-VAR ou local projections com instrumentos climáticos (geadas, secas), quebras logísticas e anúncios regulatórios; incorporação sistemática de controles de mercado e fundamentos (CEPEA, OIC, USDA/Conab); outras modelagens para cointegração entre arábica e robusta. No que tange aos aspectos de análise de sentimento, sugere-se que outras pesquisas dediquem-se ao aprimoramento do pipeline textual com deduplicação por similaridade semântica, topic modeling dinâmico (BERTopic/DTM) e sentimento por tópico, além de ajuste fino de modelos

em português como BERTimbau com base anotada por humanos; exercícios preditivos que combinem texto, Google Trends e fundamentos com machine learning regularizado e gradient boosting; padronização de preços em termos reais, harmonização de unidades e disponibilização de script e repositório reprodutível. Esses passos permitem avançar de descrições para identificação causal, quantificar canais de transmissão entre mídia e preços e estruturar métricas operacionais de monitoramento de risco e apoio a decisões no mercado de café.

Referências

BALCOMBE, K. The nature and determinants of volatility in agricultural prices: an empirical study from 1962–2008. In: FAO. *The evolving structure of world agricultural trade*. Rome: FAO, 2009. p. 109–136.

CAMARGO, M. B. P. Mudanças climáticas e a produção de café no Brasil. *Bragantia*, v. 68, n. 1, p. 229–247, 2009. Disponível em: <https://www.scielo.br/j/brag/a/XYKdCC5mVKwx3F5P3m6VrzM/>. Acesso em: 9 jun. 2025.

CHOWDHARY, K. R. Natural Language Processing. In: _____. *Fundamentals of Artificial Intelligence*. New Delhi: Springer, 2020. Disponível em: https://doi.org/10.1007/978-81-322-3972-7_19. Acesso em: 9 jun. 2025.

CNN BRASIL. Preços do café atingem máximas com preocupações por oferta no Brasil. 2024. Disponível em: <https://www.cnnbrasil.com.br/economia/mercado/precos-do-cafe-atigem-maximas-com-preocupacoes-por-oferta-no-brasil/>. Acesso em: 9 jun. 2025.

FURTADO, C. *Formação econômica do Brasil*. 34. ed. São Paulo: Companhia Editora Nacional, 2005.

G1. Café a quase R\$ 50: entenda por que preço deve continuar subindo em 2025. 13 dez. 2024. Disponível em: <https://g1.globo.com/economia/agronegocios/noticia/2024/12/13/cafe-a-quase-r-50-entenda-por-que-preco-deve-continuar-subindo-em-2025.ghtml>. Acesso em: 9 jun. 2025.

GILBERT, C. L.; MORGAN, C. W. Food price volatility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, v. 365, n. 1554, p. 3023–3034, 2010.

GOMES, L. C. et al. Agroforestry systems can mitigate the impacts of climate change on coffee production: a spatially explicit assessment in Brazil. *Agriculture, Ecosystems and Environment*, v. 294, p. 106858, 2020. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167880920300438>. Acesso em: 9 jun. 2025.

HUTTO, C.; GILBERT, E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the International AAAI Conference on Web and Social Media*, v. 8, n. 1, 2014. Disponível em: <http://sentic.net/>. Acesso em: 9 jun. 2025.

MITCHELL, R. *Web scraping with Python: collecting more data from the modern web*. 2. ed. Sebastopol: O'Reilly Media, 2018.

O GLOBO. Preços do café seguem em alta e valor pago ao produtor atinge maior patamar em 28 anos. 3 fev. 2025. Disponível em:

<https://oglobo.globo.com/economia/noticia/2025/02/03/precos-do-cafe-seguem-em-alta-e-valor-pago-ao-produtor-atinge-maior-patamar-em-28-anos.ghtml>. Acesso em: 9 jun. 2025.

PRADO JÚNIOR, C. *Formação do Brasil contemporâneo: colônia*. 5. ed. São Paulo: Brasiliense, 2000.

RIPOL, C. V. *O café no Paraná: do modelo tradicional manual ao adensado mecanizado*. Londrina: Universidade Estadual de Londrina, 2016.

RICHARDSON, L. *Beautiful Soup documentation*. 2007. Disponível em:

https://ucilnica.fri.uni-lj.si/pluginfile.php/217774/mod_resource/content/1/beautiful-soup-4-readthedocs-io-en-latest.pdf. Acesso em: 9 jun. 2025.

ROSENBERG, E. et al. Sentiment analysis on Twitter data towards climate action. *Results in Engineering*, v. 19, p. 101287, 2023. Disponível em: <https://doi.org/10.1016/j.rineng.2023.101287>. Acesso em: 9 jun. 2025.

SHILLER, R. J. *Narrative economics: how stories go viral and drive major economic events*. Princeton: Princeton University Press, 2019.

STRYKER, C.; HOLDSWORTH, J. O que é processamento de linguagem natural (PLN)? IBM Think Brasil, s.d. Disponível em: <https://www.ibm.com/br-pt/think/topics/natural-language-processing>. Acesso em: 9 jun. 2025.

SUARDI, S.; RASEL, A. R.; LIU, B. On the predictive power of tweet sentiments and attention on Bitcoin. *International Review of Economics and Finance*, v. 79, p. 289–301, 2022. Disponível em: <https://doi.org/10.1016/j.iref.2022.02.017>. Acesso em: 9 jun. 2025.

TAVARES, P. S. et al. Climate change impact on the potential yield of Arabica coffee in southeast Brazil. *Regional Environmental Change*, v. 18, p. 873–883, 2018. Disponível em: <https://doi.org/10.1007/s10113-017-1236-z>. Acesso em: 9 jun. 2025.

TETLOCK, P. C. Giving content to investor sentiment: the role of media in the stock market. *The Journal of Finance*, v. 62, n. 3, p. 1139–1168, 2007. Disponível em: <https://doi.org/10.1111/j.1540-6261.2007.01232.x>. Acesso em: 9 jun. 2025.