

Uma abordagem baseada em LLM para analisar sentimentos dos usuários do Twitter em relação ao impacto da IA no dia a dia considerando a evolução do tempo

An LLM-based approach to analyze Twitter users' sentiments regarding the impact of AI on daily life, considering its evolution over time.

Jacquet Leme   [Universidade Federal | jacquet.leme@estudante.uffs.edu.br]

Denio Duarte  [Universidade Federal | duarte@uffs.edu.br]

 Universidade Federal da Fronteira Sul, Rodovia SC 484, km 02, Bairro Fronteira Sul, Chapecó, SC CEP 89815-899
Tel. (49) 2049-2600 CNPJ 11.234.780/0007-46, Brasil.

Resumo. O crescimento das ferramentas de Inteligência Artificial generativa ampliou as discussões sobre seus impactos e aplicações no cotidiano. Neste contexto, este trabalho apresenta uma análise de sentimentos aplicada a publicações em português do Twitter (X) relacionadas às tecnologias GPT/ChatGPT, Gemini e Copilot. O estudo analisa períodos associados aos lançamentos iniciais e às versões mais recentes dessas ferramentas, utilizando um processo composto por coleta de dados, pré-processamento, filtragem semântica e classificação automática de sentimentos com Modelos de Linguagem de Grande Escala (LLMs) em abordagem *zero-shot*. Além da classificação realizada pelo modelo principal, foi utilizada uma estratégia baseada em múltiplos LLMs para auxiliar na validação das classificações produzidas. Os resultados mostram predominância de sentimentos neutros durante grande parte do período analisado, além de diferenças na evolução dos sentimentos positivos e negativos entre as tecnologias avaliadas. De forma geral, os resultados indicam que LLMs podem ser utilizados em tarefas de análise de sentimentos aplicadas a redes sociais, mesmo em cenários envolvendo textos curtos, linguagem informal e ausência de conjuntos previamente rotulados.

Abstract. The growth of generative Artificial Intelligence tools has expanded discussions about their impacts and applications in everyday life. In this context, this work presents a sentiment analysis of Portuguese-language Twitter (X) posts related to GPT/ChatGPT, Gemini, and Copilot technologies. The study analyzes periods associated with the initial releases and the most recent versions of these tools, using a process composed of data collection, preprocessing, semantic filtering, and automatic sentiment classification with Large Language Models (LLMs) in a *zero-shot* approach. In addition to the classification performed by the main model, a strategy based on multiple LLMs was employed to support the validation of the generated classifications. The results show a predominance of neutral sentiments throughout most of the analyzed period, as well as differences in the evolution of positive and negative sentiments among the evaluated technologies. Overall, the results indicate that LLMs can be applied to sentiment analysis tasks on social media, even in scenarios involving short texts, informal language, and the absence of previously labeled datasets.

Palavras-chave: Análise de sentimentos, Modelos de Linguagem de Grande Escala, Twitter (X), Inteligência Artificial Generativa, Classificação de Sentimentos.

Keywords: Sentiment Analysis, Large Language Models, Twitter (X), Generative Artificial Intelligence, Sentiment Classification.

1 Introdução

Nos últimos anos, modelos de linguagem de grande escala (LLMs) passaram a fazer parte de diferentes atividades do cotidiano. Ferramentas como *ChatGPT*, *Gemini* e *Copilot* vêm sendo utilizadas em tarefas relacionadas à escrita, programação, busca por informação e apoio a estudos, ampliando rapidamente sua presença em ambientes profissionais e acadêmicos. O desenvolvimento e as capacidades desses modelos são discutidos em trabalhos como Brown *et al.* [2020], Google DeepMind [2024] e Touvron *et al.* [2024].

Com a popularização dessas tecnologias, também aumentaram as discussões públicas relacionadas aos seus impactos, limitações e possibilidades de uso. Em redes sociais, usuários compartilham opiniões, experiências, críticas e expectativas envolvendo essas ferramentas, produzindo um grande volume de conteúdo textual espontâneo. Nesse contexto, o *Twitter* tornou-se um ambiente relevante para estudos envolvendo comportamento coletivo e análise de sentimentos, principalmente devido à velocidade de circulação das in-

formações e ao caráter público das interações [Kwak *et al.*, 2010; Murthy, 2018].

Ao mesmo tempo, avanços recentes em processamento de linguagem natural ampliaram a capacidade de análise automática de grandes volumes de texto. A arquitetura *Transformer*, proposta por Vaswani *et al.* [2017], possibilitou o desenvolvimento de modelos capazes de lidar com tarefas como classificação textual, interpretação semântica e análise de sentimentos em diferentes contextos. Trabalhos como Jurafsky and Martin [2023] e Touvron *et al.* [2024] mostram que esses modelos conseguem interpretar textos curtos e informais com resultados consistentes em diferentes aplicações.

Apesar do crescimento de pesquisas envolvendo análise de sentimentos em redes sociais, ainda são menos frequentes estudos voltados especificamente para a compreensão da evolução temporal das percepções relacionadas às tecnologias de IA generativa. Além disso, trabalhos como os de Ahnert *et al.* [2024]; Bhargava *et al.* [2025]; Mun and

Kim [2025] concentram-se em eventos específicos ou utilizam conjuntos previamente rotulados, o que pode limitar análises em cenários mais dinâmicos.

Neste contexto, este trabalho apresenta uma análise de sentimentos aplicada a publicações em português do *Twitter* relacionadas às tecnologias GPT/ChatGPT, Gemini e Copilot. O estudo considera períodos associados aos lançamentos iniciais e às versões mais recentes dessas ferramentas, buscando observar como os sentimentos relacionados a essas tecnologias variaram ao longo do tempo.

Para isso, foi desenvolvido um processo envolvendo coleta de dados, pré-processamento, filtragem semântica e classificação automática de sentimentos utilizando LLM em abordagem *zero-shot*. Além da classificação realizada pelo modelo principal, também foi utilizada uma estratégia baseada em múltiplos modelos de linguagem para auxiliar na validação das classificações produzidas.

Como principais contribuições, este trabalho apresenta: (i) uma análise temporal de sentimentos relacionados a ferramentas de IA generativa em publicações em português; (ii) uma abordagem baseada em múltiplos modelos de linguagem para validação das classificações; e (iii) uma análise comparativa do comportamento dos sentimentos ao longo do tempo para diferentes tecnologias de IA.

O restante deste trabalho está organizado da seguinte forma: a próxima seção apresenta o referencial teórico, a Seção 3 discorre sobre alguns trabalhos relacionados ao tema aqui proposto; as Seções 4 e 5 apresentam a metodologia e a discussão final da proposta e, finalmente, a Seção 6 conclui e apresenta algumas direções futuras deste trabalho.

2 Referencial Teórico

Esta seção apresenta alguns conceitos importantes para o entendimento deste trabalho.

2.1 Twitter (X)

O Twitter, que passou a ser denominado oficialmente X em 2023, é amplamente utilizado em estudos acadêmicos devido ao seu caráter público e à velocidade das interações. A plataforma permite que usuários expressem opiniões em tempo real, gerando registros espontâneos de percepções sobre temas do cotidiano, como, por exemplo, a Inteligência Artificial [Kwak et al., 2010; Murthy, 2018].

Para organizar esse volume de dados, o uso de *hashtags* é fundamental. Elas funcionam como marcadores que agrupam conteúdos em torno de tópicos específicos, facilitando a identificação de discussões relevantes [Bruns and Burgess, 2015]. No contexto deste estudo, termos como #IA, #ChatGPT e #InteligênciaArtificial permitem selecionar publicações diretamente relacionadas ao tema analisado.

Além do conteúdo textual, a plataforma disponibiliza informações estruturadas, como data de publicação e métricas de interação, que auxiliam na análise de padrões e na observação da evolução das opiniões ao longo do tempo [Marwick, 2013].

Dessa forma, a escolha pelo Twitter se justifica pela disponibilidade de dados, diversidade de opiniões e possibilidade de acompanhamento contínuo do comportamento dos usuários, sendo utilizado como base para a análise de sentimentos proposta neste trabalho.

2.2 Modelos de Linguagem de Grande Escala

Os LLMs representam um avanço significativo no campo do Processamento de Linguagem Natural, permitindo a análise e geração de textos de forma contextualizada [LeCun et al., 2015; Brown et al., 2020]. Esses modelos são baseados em arquiteturas neurais profundas, especialmente o *Transformer*, que possibilitam o processamento eficiente de grandes volumes de dados textuais [Vaswani et al., 2017].

Diferentemente de abordagens tradicionais baseadas em regras, os LLMs aprendem padrões diretamente a partir de dados, o que permite capturar relações semânticas complexas e lidar com ambiguidades da linguagem natural [Goodfellow et al., 2016; Jurafsky and Martin, 2023]. Essa capacidade tem impulsionado o uso desses modelos em diversas tarefas, incluindo classificação de textos, sumarização e análise de sentimentos.

2.2.1 Arquitetura e Representação Semântica

Um dos elementos centrais dos LLMs é o uso de representações vetoriais, conhecidas como *embeddings*, que permitem mapear palavras em espaços semânticos contínuos [Mikolov et al., 2013]. Essa abordagem possibilita que termos com significados semelhantes sejam posicionados próximos entre si, facilitando a identificação de padrões linguísticos.

Modelos mais recentes incorporam representações contextuais, nas quais o significado das palavras depende do contexto em que são utilizadas [Jurafsky and Martin, 2023; Raschka, 2024]. Essa característica é particularmente importante na análise de sentimentos, pois permite interpretar nuances como ambiguidade, variação de sentido e dependência do contexto discursivo.

A arquitetura *Transformer*, introduzida por Vaswani et al. [2017], utiliza mecanismos de atenção para identificar relações entre diferentes partes do texto, permitindo capturar dependências de longo alcance e melhorar a compreensão semântica.

2.2.2 Análise de Sentimentos

A análise de sentimentos (*Sentiment Analysis*) é uma área do Processamento de Linguagem Natural voltada à identificação automática de opiniões, emoções e polaridades presentes em textos [Zhang et al., 2024]. De forma geral, essa tarefa busca classificar conteúdos em categorias como positivo, negativo ou neutro, permitindo compreender como determinados temas são percebidos por indivíduos ou grupos.

A Tabela 1 apresenta exemplos de publicações classificadas durante o processo de análise de sentimentos realizado neste trabalho. Observa-se que, além de opiniões explícitas, os modelos também foram capazes de interpretar construções irônicas e contextuais presentes em postagens de redes sociais. No segundo exemplo, por exemplo, a publicação utiliza ironia ao aparentar elogiar a ferramenta de IA, enquanto o contexto expressa insatisfação, sendo corretamente classificada como negativa.

Com o crescimento das redes sociais, a análise de sentimentos passou a ser amplamente utilizada na investigação de opiniões públicas em ambientes digitais. Plataformas como o Twitter possibilitam o acesso a grandes volumes de textos produzidos espontaneamente pelos usuários, favorecendo es-

Tabela 1. Exemplos de publicações e classificações de sentimento

Publicação	Sentimento
O ChatGPT me ajudou muito a organizar meus estudos hoje.	Positivo
Nossa, maravilhoso esse Copilot apagando metade do meu código sozinho. Tecnologia incrível mesmo.	Negativo
Testei a nova versão do Gemini hoje, mas ainda preciso usar mais para formar uma opinião.	Neutro

tudos relacionados a comportamento, percepção social e monitoramento de tendências [Cambria *et al.*, 2023].

Abordagens tradicionais de análise de sentimentos eram frequentemente baseadas em métodos supervisionados e conjuntos de dados previamente anotados. Entretanto, o avanço dos LLMs permitiu a realização dessa tarefa de maneira mais flexível e contextualizada, reduzindo a dependência de treinamento específico [Liu *et al.*, 2023].

Além da classificação de polaridade, modelos contemporâneos são capazes de lidar com características complexas da linguagem natural, como ambiguidades, ironias e variações contextuais, aspectos recorrentes em textos de redes sociais [Simmering *et al.*, 2023; Rathje *et al.*, 2024]. Essa capacidade amplia o potencial de aplicação dos LLMs em estudos que envolvem grandes volumes de dados textuais e análises temporais de opinião pública.

No contexto deste trabalho, a análise de sentimentos é utilizada para investigar as percepções expressas por usuários do Twitter sobre ferramentas de Inteligência Artificial, considerando diferentes períodos relacionados ao lançamento e à evolução desses modelos.

2.2.3 Modelos Contemporâneos e Aplicação

Diversos modelos foram desenvolvidos com base nessa arquitetura, tais como GPT, Gemini e LLaMA. Embora diferem em escala, janela de contexto e estratégias de treinamento, esses modelos compartilham a capacidade de interpretar linguagem natural de maneira eficiente [Brown *et al.*, 2020; Google DeepMind, 2024; Touvron *et al.*, 2024].

No contexto da análise de sentimentos, esses modelos podem ser utilizados na classificação automática de textos em categorias como positivo, negativo ou neutro. Essa capacidade é especialmente útil em redes sociais, em que o grande volume de publicações dificulta análises realizadas manualmente.

Além disso, os LLMs conseguem interpretar padrões linguísticos mesmo em cenários com poucos dados anotados, favorecendo aplicações exploratórias e estudos baseados em grandes volumes de texto.

2.2.4 Engenharia de Prompt

A engenharia de *prompt* consiste na formulação de instruções que orientam o comportamento dos modelos em tarefas específicas. Técnicas como *zero-shot prompting*, em que o modelo realiza a tarefa sem exemplos prévios, e *few-shot prompting*, em que poucos exemplos são fornecidos como referência, permitem aplicar os LLMs diretamente a novos contextos, reduzindo a necessidade de dados anotados [Raschka, 2024].

No contexto deste trabalho, foi utilizada principalmente a abordagem *zero-shot prompting*, na qual os modelos receberam instruções específicas para classificar automaticamente as publicações em positivo, negativo ou neutro. O *prompt* foi estruturado de forma objetiva, definindo claramente o papel do modelo, a tarefa esperada e o formato da resposta, buscando reduzir ambiguidades durante o processo de classificação.

A seguir, apresenta-se um exemplo simplificado do *prompt* utilizado:

Você é um especialista em análise de sentimentos aplicado a publicações do Twitter (X).

Sua tarefa é analisar o texto fornecido e classificá-lo em apenas uma das seguintes categorias:

- POSITIVO
- NEGATIVO
- NEUTRO

Considere o contexto da frase, incluindo possíveis ironias, sarcasmos e opiniões implícitas. Responda apenas com uma única palavra: a categoria escolhida. Não inclua pontuação, explicações ou texto adicional.

Texto: "[INSERIR TEXTO]"

Resposta:

No caso da análise de sentimentos, essa abordagem possibilita a classificação de textos de forma direta, sem a necessidade de treinamento adicional, mantendo consistência e escalabilidade nos resultados [Rathje *et al.*, 2024; Simmering *et al.*, 2023].

De modo geral, o uso de LLMs neste trabalho contribui para a análise de grandes volumes de textos e textos publicados no Twitter, permitindo observar padrões e variações nos sentimentos expressos pelos usuários ao longo do tempo.

3 Trabalhos Relacionados

A análise de sentimentos em redes sociais tem recebido atenção de maneira crescente nos últimos anos, especialmente com o avanço dos LLMs. Baseados principalmente na arquitetura *Transformer* [Vaswani *et al.*, 2017], esses modelos apresentam grande capacidade de compreensão contextual e vêm sendo utilizados em tarefas como classificação de sentimentos, geração de texto e análise de discurso [Brown *et al.*, 2020; OpenAI, 2023].

Em redes sociais, o grande volume de textos publicados diariamente, aliado à capacidade dos LLMs de interpretar contexto e linguagem natural, tem contribuído para estudos voltados à identificação de padrões emocionais e mudanças de percepção ao longo do tempo [Liu *et al.*, 2023; Cambria *et al.*, 2023].

Em um estudo recente, Ahnert *et al.* [2024] investigam a modelagem temporal de emoções em redes sociais utilizando adaptações aplicadas ao modelo LLaMA. Os autores demonstram que técnicas de ajuste leve permitem acompanhar variações emocionais ao longo do tempo sem a necessidade de reentrenamento completo do modelo.

Outra abordagem é apresentada por Wang and Chandra [2024], que analisa a evolução de sentimentos em um contexto social específico por meio de modelos supervisionados

baseados em BERT. Os resultados indicam que mudanças no contexto social influenciam diretamente os sentimentos expressos nas publicações, reforçando a importância de análises temporais em redes sociais.

Já Bhargava *et al.* [2025] exploram o impacto de características linguísticas, como sarcasmo e emojis, no desempenho de modelos de linguagem. Utilizando modelos como LLaMA e Mistral, os autores mostram que, apesar dos avanços recentes, ainda existem dificuldades na interpretação de nuances presentes em textos informais.

De forma semelhante, Mun and Kim [2025] compararam diferentes modelos de linguagem em tarefas de análise de sentimentos aplicadas a dados financeiros. Os resultados mostram que o desempenho dos modelos pode variar de acordo com o contexto analisado e com as características dos dados utilizados.

De forma geral, os trabalhos analisados mostram que os LLMs têm obtido bons resultados em tarefas de análise de sentimentos. No entanto, muitos desses estudos utilizam dados em inglês, focam em contextos específicos ou avaliam apenas um único modelo, o que dificulta a aplicação dos resultados em outros cenários.

Também são poucos os estudos que exploram o uso combinado de diferentes modelos de linguagem em processos de validação ou votação. Considerando essa limitação, este trabalho utiliza diferentes LLMs na análise de sentimentos de publicações do Twitter relacionadas aos modelos de linguagem de grande escala, buscando reduzir inconsistências nas classificações e tornar os resultados mais confiáveis.

Além disso, ainda são poucos os estudos voltados à análise de sentimentos em língua portuguesa, especialmente em pesquisas realizadas ao longo de períodos mais extensos. Nesse contexto, esta pesquisa busca contribuir para essa discussão, observando como as percepções dos usuários variam ao longo do tempo.

4 Metodologia

A metodologia adotada neste trabalho foi organizada com o objetivo de analisar a evolução dos sentimentos expressos por usuários do Twitter (X) em relação a modelos de Inteligência Artificial baseados em LLMs, considerando especificamente publicações relacionadas ao ChatGPT, Gemini e Copilot ao longo do tempo. O processo inclui etapas de coleta de dados, pré-processamento, filtragem semântica, classificação de sentimentos e rotulagem automática utilizando múltiplos modelos de linguagem.

Essa abordagem permite identificar a polaridade das publicações (*i.e.*, positiva, neutra ou negativa) e também comparar diferentes modelos de linguagem na tarefa de classificação de sentimentos. A Figura 1 apresenta o fluxo geral do processo utilizado neste trabalho e serve como base para as etapas descritas nas subseções seguintes.

4.1 Visão Geral do Processo

O processo metodológico adotado neste trabalho segue um fluxo estruturado, desde a coleta dos dados até a análise final dos resultados. Inicialmente, os dados foram coletados a partir do Twitter (X) e da plataforma Nitter, uma interface alternativa utilizada para acessar e visualizar conteúdos públicos da rede social.

Em seguida, os dados passaram por uma etapa de pré-processamento, que incluiu limpeza textual, remoção de duplicidades, filtragem por idioma e verificação de relevância do conteúdo. Após esse processo, os dados foram organizados em formato estruturado para utilização nas etapas seguintes.

A classificação de sentimentos foi realizada utilizando LLMs por meio de uma estratégia de *zero-shot prompting*, mantendo o mesmo padrão de interpretação para todos os textos analisados.

A rotulagem manual dos *tweets* extraídos é muito complexa, tediosa e suscetível a erros. Assim, foi criada uma abordagem automática de rotulagem: três modelos foram escolhidos para formarem um comitê de rotulagem, o rótulo associado ao tweet é aquele com mais votos. Em caso de empate, um humano associava o rótulo manualmente.

Por fim, os dados classificados foram utilizados em análises quantitativas relacionadas à distribuição dos sentimentos e à variação dessas percepções ao longo do tempo.

4.2 Coleta de Dados

A coleta de dados foi realizada a partir do Twitter (X), utilizando também a plataforma Nitter¹ como alternativa de acesso às publicações. A escolha dessas fontes ocorreu devido à disponibilidade de conteúdo público e à possibilidade de obtenção dos dados sem o uso da API oficial da plataforma.

O processo foi conduzido por meio de scripts desenvolvidos especificamente para este trabalho, responsáveis por recuperar publicações a partir de termos de busca previamente definidos. A coleta foi realizada separadamente para cada tecnologia analisada, permitindo associar diretamente os dados obtidos ao respectivo modelo.

Para o modelo GPT, foram utilizados termos como “GPT” e “ChatGPT”. Para o modelo Gemini, foram considerados “Gemini”, “Google Gemini” e “Bard”, uma vez que versões anteriores da tecnologia eram identificadas por essa denominação. Já para o Copilot, foram empregados termos como “Copilot” e “GitHub Copilot”. De forma geral, as consultas seguiram padrões como:

```
GPT OR ChatGPT
Gemini OR "Google Gemini" OR Bard
Copilot OR "GitHub Copilot"
```

Além da definição das consultas, foi adotado como critério que o conteúdo textual das publicações deveria conter explicitamente termos relacionados ao modelo em análise, como “GPT” e “ChatGPT” para o GPT, “Gemini”, “Google Gemini” e “Bard” para o Gemini, e “Copilot” ou “GitHub Copilot” para o Copilot. Dessa forma, apenas *tweets* que mencionassem diretamente a tecnologia correspondente foram considerados.

A coleta dos dados também considerou períodos específicos relacionados aos modelos analisados. Para cada tecnologia, foram definidos dois momentos principais: o lançamento inicial e uma versão mais recente disponível durante o desenvolvimento da pesquisa. A partir desses marcos, foram coletadas publicações realizadas em uma janela de até

¹ nitter.net

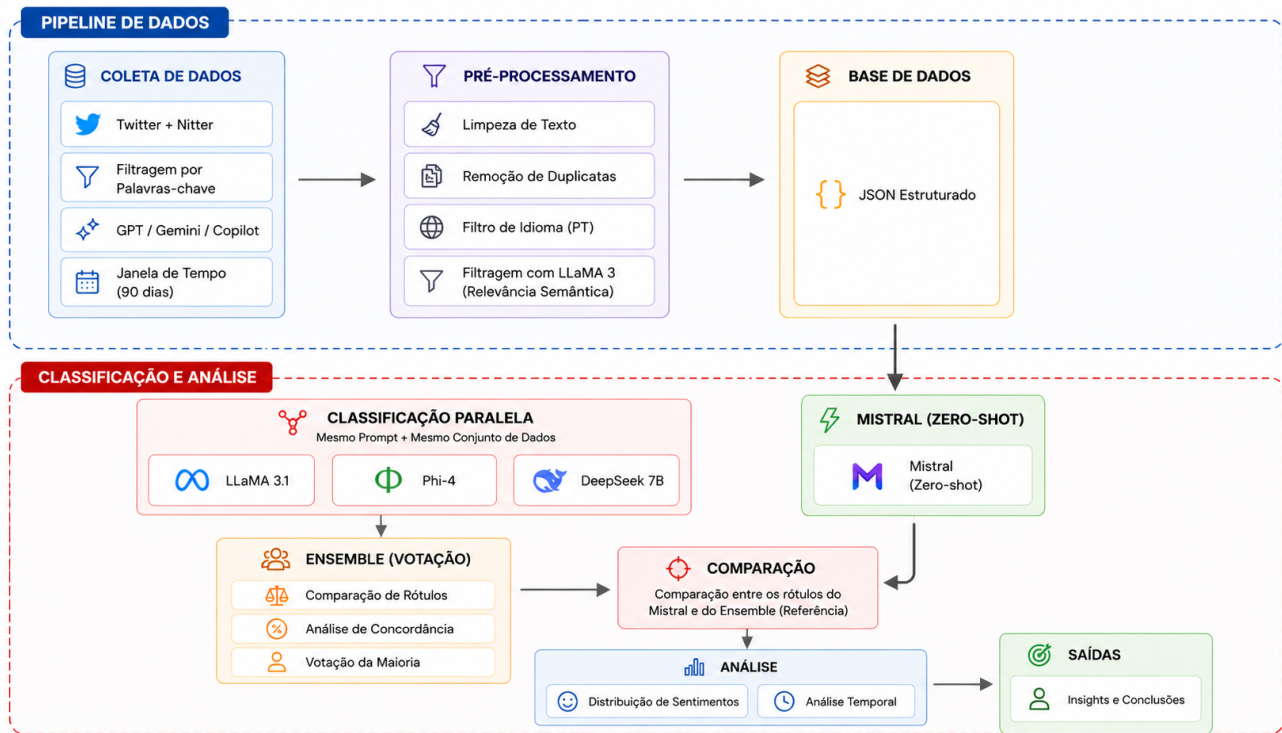


Figura 1. Pipeline metodológico para análise de sentimentos em dados do Twitter, incluindo coleta, pré-processamento, classificação com múltiplos modelos de linguagem e geração de rótulos de referência por votação majoritária.

90 dias, permitindo acompanhar possíveis mudanças nas percepções dos usuários ao longo do tempo.

A escolha dessas janelas buscou manter a análise próxima aos períodos de maior discussão sobre cada modelo, evitando a inclusão de publicações muito distantes dos eventos analisados. A Tabela 2 apresenta os períodos considerados neste trabalho.

Tabela 2. Eventos temporais utilizados na coleta dos dados

Modelo	Evento	Início	Fim	Dias
ChatGPT	Lançamento inicial	30/11/2022	28/02/2023	90
GPT-4o	Versão mais recente	13/05/2024	11/08/2024	90
Gemini	Lançamento inicial	06/02/2023	07/05/2023	90
Gemini 1.5	Versão mais recente	08/02/2024	08/05/2024	90
Copilot	Lançamento inicial	29/06/2021	27/09/2021	90
Copilot Pro	Versão mais recente	01/11/2023	30/01/2024	90

Cada publicação coletada foi armazenada com informações como identificador do tweet, conteúdo textual, data de publicação, modelo associado e fonte de origem. Os dados foram então organizados em formato estruturado para utilização nas etapas seguintes.

4.3 Pré-processamento dos Dados

Após a coleta, os dados passaram por uma etapa de pré-processamento voltada à redução de ruídos e remoção de publicações inadequadas para a análise. Durante esse processo, optou-se por preservar o conteúdo original dos *tweets*, considerando que elementos como *hashtags*, *emojis*, pontuação e linguagem informal podem contribuir para a interpretação dos sentimentos expressos pelos usuários.

Antes das demais etapas de pré-processamento, foi realizada uma limpeza textual básica dos dados coletados. Foram corrigidos problemas de formatação, como caracteres de

controle e espaços excessivos, mantendo-se elementos como *hashtags*, *emojis*, pontuação e expressões informais, considerados relevantes para a análise de sentimentos.

A etapa seguinte consistiu na remoção de duplicidades com base no identificador único de cada publicação (*tweet_id*). Quando *tweets* repetidos eram encontrados, apenas uma ocorrência era mantida, evitando que uma mesma publicação influenciasse mais de uma vez os resultados da análise.

Após a remoção das duplicidades, foi aplicado um filtro de idioma para manter apenas publicações em português. Essa etapa foi realizada utilizando a biblioteca *Lingua* para Python, responsável pela detecção automática do idioma a partir do conteúdo textual. *Tweets* identificados em outros idiomas foram descartados.

Posteriormente, foi realizada uma filtragem semântica de relevância utilizando o modelo *LLaMA 3*². A aplicação do modelo foi feita por meio da API disponibilizada pelo *OpenAI*, permitindo a execução local sem dependência de serviços externos. Para cada *tweet*, foi realizada uma verificação do conteúdo com o objetivo de identificar se a publicação realmente tratava de Inteligência Artificial ou de um dos modelos analisados neste trabalho.

Essa verificação foi realizada por meio de uma classificação em duas categorias: *RELEVANTE* e *NÃO RELEVANTE*. *Tweets* classificados como relevantes foram mantidos no conjunto de dados, enquanto publicações sem relação direta com o tema analisado foram removidas.

Como exemplo, algumas publicações continham termos presentes nas consultas de busca, mas eram utilizadas em contextos diferentes daqueles relacionados à Inteligência Ar-

²Versão do modelo utilizada nesta etapa: LLaMA 3 8B.

tificial. Publicações como “preciso de um copilot para essa viagem”, “meu signo é Gêmeos (Gemini)” ou “Bard foi uma das melhores bandas que ouvi este ano” incluem palavras utilizadas nos critérios de coleta, porém sem relação com os modelos analisados neste estudo. Publicações desse tipo foram classificadas como *NÃO RELEVANTE* e removidas do conjunto final.

A execução do modelo foi realizada utilizando processamento em lotes e estratégias de repetição em caso de falhas, garantindo maior estabilidade durante a análise do conjunto de dados.

Ao final do pré-processamento, o conjunto de dados ficou composto por aproximadamente 15 mil *tweets*, já filtrados quanto à duplicidade, idioma e relevância do conteúdo. Inicialmente, haviam sido coletadas cerca de 21 mil publicações, mas parte desse material foi removida durante as etapas de tratamento dos dados.

A maior parte das exclusões ocorreu devido à presença de *tweets* duplicados, publicações em outros idiomas e conteúdos sem relação direta com o tema da pesquisa. Após essa etapa, o conjunto final foi utilizado nas fases seguintes.

4.4 Rotulagem Automática com Múltiplos Modelos

Após o pré-processamento das publicações, foi realizada uma etapa de rotulagem automática utilizando diferentes modelos de linguagem. O objetivo dessa etapa foi criar uma base de referência para comparação com os resultados obtidos pelo modelo principal utilizado neste trabalho.

Foram utilizados os modelos *LLaMA 3.1*³, *Phi-4*⁴ e *DeepSeek*⁵. Todos os modelos receberam o mesmo conjunto de publicações e utilizaram o mesmo prompt de classificação apresentado anteriormente.

O modelo *LLaMA 3.1* foi executado localmente utilizando a biblioteca *llama.cpp*. Os modelos *Phi-4* e *DeepSeek 7B* foram utilizados por meio de APIs disponíveis na plataforma Hugging Face, com requisições realizadas de forma sequencial. Embora este trabalho utilize duas versões da família LLaMA, cada uma foi empregada em uma etapa distinta do processo. O *LLaMA 3* foi utilizado na etapa de filtragem semântica, enquanto o *LLaMA 3.1* foi empregado na rotulagem automática em função da disponibilidade dos modelos durante a execução dos experimentos. Como os modelos pertencem à mesma família e foram utilizados em tarefas distintas, essa diferença não compromete a consistência metodológica do processo.

Cada modelo gerou um rótulo de sentimento para cada publicação, classificando os textos em positivo, negativo ou neutro. O rótulo final foi definido por votação majoritária entre os três modelos utilizados na etapa de rotulagem automática. Nos casos em que não houve consenso entre as classificações, a definição final foi realizada manualmente pelo pesquisador.

Após a geração dos rótulos finais, os resultados produzidos pelo modelo principal, *Mistral*, foram comparados com essa base de referência.

4.5 Classificação de Sentimentos

A classificação de sentimentos foi realizada com o uso de LLMs, com o objetivo de identificar a polaridade das publicações em três categorias: *POSITIVO*, *NEGATIVO* e *NEUTRO*. Para isso, foi adotada a abordagem de *zero-shot prompting*, na qual o modelo é instruído a executar a tarefa sem a necessidade de exemplos previamente rotulados.

O modelo principal utilizado nesta etapa foi o *Mistral*⁶, executado localmente a partir de uma versão otimizada no formato GGUF. A inferência foi realizada por meio da biblioteca *llama-cpp-python*, utilizando processamento em CPU. Considerando as limitações do ambiente utilizado, com aproximadamente 8GB de RAM, foram definidos contexto de 512 tokens, 4 *threads* e *batch size* igual a 64 para execução do modelo.

A classificação foi realizada de forma sequencial sobre o conjunto de dados, com salvamento periódico dos resultados para evitar perdas em caso de falhas durante a execução.

Para orientar o modelo na tarefa de classificação, foi utilizado um *prompt* padronizado, aplicado de forma idêntica a todas as publicações. O objetivo desse *prompt* foi estabelecer critérios claros para interpretação dos textos, reduzindo ambiguidades durante a classificação. Além da definição das três categorias de sentimento, o modelo também foi orientado a considerar aspectos como significado geral do texto, tom da linguagem, contexto implícito, uso de *emojis* e situações envolvendo ironia ou sarcasmo. O *prompt* utilizado é apresentado na Figura 2.

A utilização de um único *prompt* permitiu manter o mesmo padrão de classificação entre os diferentes modelos utilizados. Além disso, foram adotadas configurações determinísticas durante a inferência, reduzindo variações nas respostas geradas.

As saídas produzidas pelos modelos foram posteriormente normalizadas, garantindo que cada classificação estivesse associada a apenas uma das três categorias definidas.

Ao final dessa etapa, cada *tweet* recebeu um rótulo de sentimento gerado pelo modelo principal. Esses resultados foram posteriormente comparados com os rótulos de referência obtidos por meio da estratégia de rotulagem automática descrita na seção anterior.

4.6 Análise de Concordância e Evolução Temporal dos Sentimentos

Após a geração dos rótulos finais, foi realizada uma análise de concordância entre os modelos utilizados no processo de rotulagem.

Para cada *tweet*, foi verificado quantos modelos atribuíram o mesmo rótulo de sentimento. Essa etapa permitiu identificar casos em que os modelos produziram classificações iguais e situações em que ocorreram diferenças entre as respostas.

Também foi realizada uma análise temporal das classificações de sentimento ao longo dos períodos definidos para cada tecnologia analisada. Para isso, as publicações foram organizadas considerando os 90 dias definidos para cada evento analisado. A partir disso, foram calculadas as proporções de sentimentos positivos, negativos e neutros ao longo

³Versão do modelo utilizada: LLaMA 3.1 8B.

⁴Versão do modelo utilizada: Phi-4.

⁵Versão do modelo utilizada: DeepSeek 7B.

⁶Modelo utilizado: Mistral 7B Instruct v0.2.

```
[INST]
Você é um classificador de sentimentos.

Sua tarefa é classificar o sentimento do texto em uma única categoria:
POSITIVO, NEGATIVO ou NEUTRO.

Definições:

- POSITIVO: o texto expressa opinião favorável, entusiasmo,
satisfação ou expectativa positiva.

- NEGATIVO: o texto expressa crítica, insatisfação,
preocupação, medo ou expectativa negativa.

- NEUTRO: o texto é informativo, descritivo, ambíguo ou
não expressa claramente uma opinião ou emoção.

Diretrizes:

- Considere o significado geral do texto.
- Considere o tom da linguagem.
- Considere o contexto implícito.
- Emojis podem indicar sentimento.
- Interprete ironia ou sarcasmo quando possível.

Texto: "{text}"

Responda com APENAS UMA palavra:
POSITIVO, NEGATIVO ou NEUTRO.

Não explique.
Não justifique.
Não adicione pontuação.
Não escreva nada além da resposta.
[/INST]
```

Figura 2. Prompt utilizado para classificação de sentimentos.

do período.

A organização temporal dos dados permitiu observar variações nos sentimentos ao longo do período analisado, considerando os eventos de lançamento inicial e versões mais recentes das tecnologias estudadas.

4.7 Ferramentas e Tecnologias Utilizadas

O desenvolvimento deste trabalho foi realizado com o uso de um conjunto de ferramentas voltadas ao processamento de dados textuais e à execução de modelos de linguagem de grande escala.

- **Linguagem de Programação:** Python 3.x, utilizado na implementação das etapas do processo, incluindo coleta, pré-processamento, classificação e análise dos dados.
- **Manipulação de Dados:** Biblioteca Pandas, empregada para organização, filtragem, transformação e estruturação dos dados em formato tabular.
- **Visualização de Dados:** Biblioteca Matplotlib, utilizada para geração dos gráficos de distribuição de sentimentos, comparação entre modelos e evolução temporal.
- **Modelos de Linguagem:** Foram utilizados os modelos *LLaMA 3.1*⁷, *Phi*⁸ e *DeepSeek*, aplicados na tarefa de classificação de sentimentos por meio de abordagem *zero-shot*.
- **Execução Local:** Os modelos Mistral e LLaMA foram executados localmente utilizando a biblioteca *llama.cpp*, com versões otimizadas no formato GGUF, permitindo a inferência em ambientes com recursos limitados.

- **Execução via API:** Os modelos *Phi* e *DeepSeek* foram acessados por meio de APIs disponibilizadas pela plataforma Hugging Face.
- **Deteção de Idioma:** Foi utilizada a biblioteca *Lingua*, responsável pela identificação automática de linguagem, para filtrar apenas publicações em português durante a etapa de pré-processamento.
- **Coleta de Dados:** A coleta foi realizada a partir do Twitter (X), com suporte da plataforma Nitter, utilizando scripts desenvolvidos especificamente para este trabalho.

Essas ferramentas foram utilizadas em conjunto ao longo do desenvolvimento do trabalho, permitindo executar todas as etapas do processo, desde a coleta das publicações até a análise final dos resultados.

5 Resultados

Esta seção apresenta os principais resultados obtidos a partir da análise de sentimentos realizada sobre publicações relacionadas às tecnologias GPT/ChatGPT, Gemini e GitHub Copilot. Os resultados consideram tanto os períodos associados aos lançamentos iniciais quanto às versões mais recentes de cada tecnologia, utilizando uma janela temporal relativa de até 90 dias após cada evento.

Inicialmente, os tweets foram classificados automaticamente pelo modelo *Mistral* utilizando uma abordagem *zero-shot*. Além disso, também foi realizada uma etapa separada de rotulagem automática com os modelos LLaMA 3.1, Phi e DeepSeek. Com base na concordância entre esses modelos, foi definido o rótulo final de sentimento (*final_sentiment*), utilizado como referência para comparar os resultados produzidos pelo Mistral nas análises apresentadas nesta seção.

O conjunto de dados utilizado nesta etapa contém

⁷Modelo utilizado: LLaMA 3.1 8B.

⁸Modelo utilizado: Phi-4.

15.011 publicações válidas após as etapas de filtragem, duplicação e validação estrutural do dataset.

5.1 Distribuição Geral dos Sentimentos

A Figura 3 apresenta a distribuição geral dos sentimentos obtidos a partir do rótulo final consolidado (*final_sentiment*).

Observa-se predominância de sentimentos neutros ao longo do conjunto analisado, representando aproximadamente 49,0% das publicações. Os sentimentos positivos e negativos apresentaram proporções bastante próximas, ambos em torno de 25%. Isso indica que o conjunto de dados é desbalanceado. Nesta abordagem, o desbalanceamento não será tratado, ou seja, não serão criadas publicações sintéticas para aumentar o número de exemplos positivos e negativos, nem será reduzido o número de publicações neutras. O objetivo é manter o problema o mais próximo possível da realidade.

Esse resultado mostra que boa parte das publicações relacionadas às tecnologias de inteligência artificial generativa possui caráter predominantemente informativo ou descritivo, sem uma carga emocional suficientemente explícita para caracterizar sentimentos positivos ou negativos.

Essa predominância pode ser explicada pelo perfil das publicações compartilhadas durante os períodos analisados. Grande parte dos tweets limita-se ao anúncio de novas versões das ferramentas, ao compartilhamento de notícias, à descrição de funcionalidades ou ao esclarecimento de dúvidas, sem expressar avaliações ou opiniões explícitas. Esse comportamento pode ser observado em publicações como “GPT-3.5 é a nova versão da IA que gera textos como se fosse um humano”, “Gemini ainda está em fase de testes com um grupo de pessoas “confiáveis”, e não temos nenhuma previsão de quando estará disponível para todos os usuários” e “O GitHub Copilot utiliza um algoritmo da OpenAI chamado Codex que foi treinado em terabytes de códigos extraídos do GitHub”. Em todos esses casos, o conteúdo tem caráter essencialmente informativo, justificando sua classificação como neutro.

Ao mesmo tempo, a proximidade entre os sentimentos positivos e negativos indica que essas tecnologias também geram opiniões distintas entre os usuários. Enquanto algumas publicações demonstram entusiasmo com os avanços das ferramentas, outras levantam críticas e preocupações relacionadas ao uso excessivo de IA, à confiabilidade dos resultados e aos possíveis impactos profissionais.

5.2 Distribuição por Tecnologia e Momento de Referência

Além da análise global, também foi realizada uma avaliação individual para cada tecnologia e para cada momento de referência considerado no estudo. A Figura 4 apresenta a distribuição percentual dos sentimentos para GPT, Gemini e Copilot, considerando os períodos de lançamento inicial e de versão recente.

Os resultados mostram diferenças perceptíveis entre as tecnologias analisadas. No caso do GPT, observa-se aumento da proporção de sentimentos negativos nas versões recentes, passando de aproximadamente 25% para 29%, acompanhado de redução dos sentimentos positivos, que diminuíram de 28% para 22%. Os sentimentos neutros permane-

ram predominantes, variando de 46% para 49%. Esse comportamento pode estar associado ao aumento da exposição pública da tecnologia e ao crescimento de discussões mais críticas ao longo do tempo.

O Gemini apresentou predominância consistente de sentimentos neutros nos dois períodos analisados, passando de aproximadamente 51% para 55%. Além disso, houve redução dos sentimentos negativos, de 26% para 20%, enquanto os sentimentos positivos apresentaram pequeno crescimento, variando de 22% para 25%. Esses resultados sugerem maior estabilidade nas discussões relacionadas à ferramenta.

Já o Copilot apresentou comportamento diferente dos demais modelos. Nas versões recentes, os sentimentos positivos aumentaram de aproximadamente 27% para 35%, enquanto os negativos reduziram de 28% para 15%. Os sentimentos neutros também cresceram levemente, passando de 45% para 50%. Esse resultado pode estar relacionado ao contexto mais específico de uso da ferramenta, principalmente em atividades de desenvolvimento de software e produtividade.

De forma geral, percebe-se que cada tecnologia apresentou padrões distintos de recepção pública, indicando que a percepção dos usuários varia conforme o contexto de utilização e o estágio de maturidade de cada ferramenta.

5.3 Concordância entre os Rotuladores Automáticos

Após a etapa de rotulagem automática, também foi analisado o nível de concordância entre os modelos utilizados no comitê de decisão. A Figura 5 apresenta a distribuição dos níveis de concordância observados entre LLaMA 3.1, Phi e DeepSeek.

Os resultados mostram que aproximadamente 39,7% das publicações tiveram concordância total entre os três modelos. A maior parte dos tweets, cerca de 53,6%, apresentou concordância parcial, situação em que dois modelos atribuíram o mesmo sentimento enquanto o terceiro apresentou uma classificação diferente. Já os casos de discordância total representaram apenas 6,7% do conjunto analisado, correspondendo a 1.004 publicações.

De forma geral, os resultados indicam que os modelos apresentaram comportamento relativamente consistente durante a classificação automática. Mesmo assim, parte das divergências apareceu em publicações mais ambíguas, principalmente em casos com ironia, humor, sarcasmo ou opiniões menos explícitas.

5.4 Concordância por Classe de Sentimento

Também foi realizada uma análise da concordância entre os pares de rotuladores considerando separadamente cada classe de sentimento. A Figura 6 apresenta a taxa de concordância entre os pares de modelos para sentimentos positivos, negativos e neutros.

Os resultados mostram diferenças relevantes entre as classes analisadas. Para os sentimentos positivos, o maior nível de concordância foi observado entre Phi e DeepSeek, atingindo aproximadamente 88,3%. Já o par LLaMA vs Phi apresentou concordância significativamente menor, cerca de

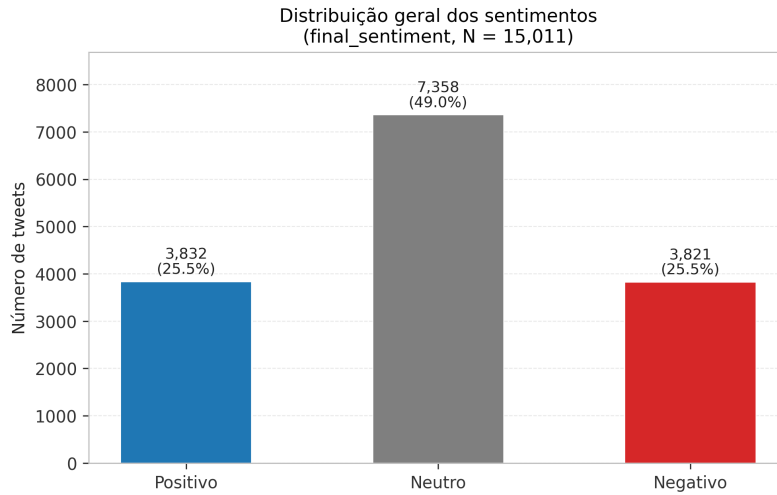


Figura 3. Distribuição geral dos sentimentos no conjunto consolidado.

Distribuição percentual dos sentimentos por tecnologia e momento de referência

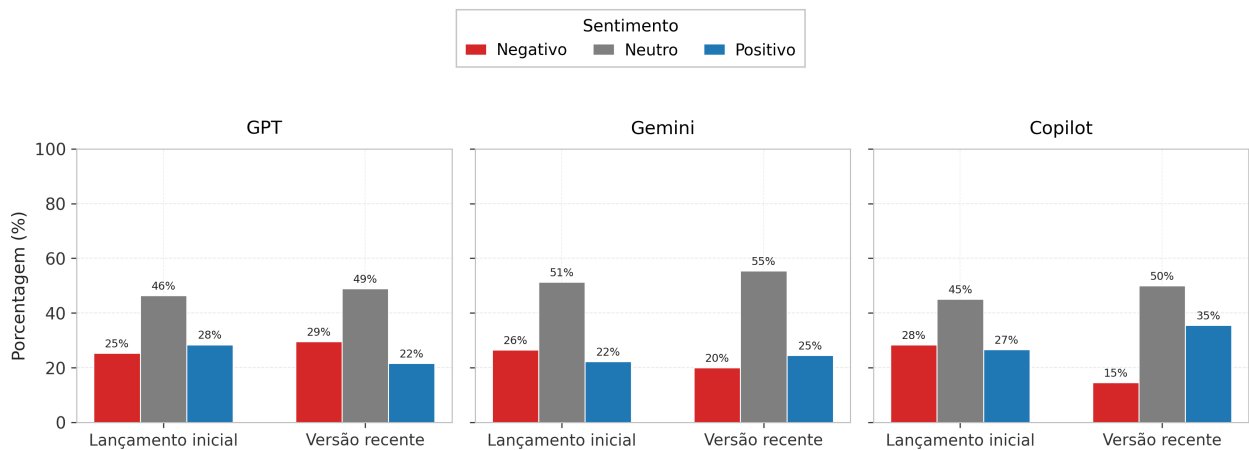


Figura 4. Distribuição percentual dos sentimentos por tecnologia e momento de referência.

32,1%. Esse resultado indica que determinados modelos tendem a interpretar conteúdos positivos de maneira mais semelhante do que outros.

No caso dos sentimentos negativos, as taxas de concordância ficaram em níveis intermediários, variando aproximadamente entre 47% e 63%. Esse comportamento sugere maior dificuldade na interpretação de publicações com críticas, reclamações ou opiniões ambíguas.

Os sentimentos neutros apresentaram comportamento distinto. O par LLaMA vs Phi obteve a maior concordância nessa classe, alcançando aproximadamente 78,1%, enquanto Phi vs DeepSeek apresentou o menor valor relativo, com cerca de 50,8%.

De forma geral, percebe-se que os modelos não apresentam exatamente o mesmo comportamento durante a classificação dos sentimentos. Em alguns casos, determinados modelos tendem a interpretar conteúdos mais ambíguos como neutros, enquanto outros atribuem polaridade positiva ou negativa com maior frequência.

Esse comportamento tornou-se mais claro em publicações contendo humor, ironia, exageros informais ou comentários técnicos acompanhados de expressões emocionais, fa-

tores que aumentam a dificuldade de interpretação automática.

5.5 Avaliação do Modelo Mistral

Após a definição do conjunto de referência, foi realizada a avaliação do modelo *Mistral*. Para isso, os rótulos produzidos originalmente pelo Mistral foram comparados com o campo *final_sentiment*, obtido após a etapa de combinação e validação dos rótulos.

A avaliação foi realizada por meio das métricas de acurácia, precisão, *recall*, *F1-score* e coeficiente de Cohen (*Cohen's Kappa*). A Tabela 3 apresenta o desempenho geral do modelo e a Tabela 4 apresenta as métricas por classe de sentimento.

De forma geral, o modelo apresentou acurácia de aproximadamente 68,0% e *macro F1-score* de 67%. Esses valores indicam desempenho razoável para uma classificação *zero-shot*, principalmente considerando que os textos analisados são publicações curtas, informais e frequentemente ambíguas.

Além dessas métricas, foi calculado o coeficiente de Cohen (*Cohen's Kappa*), que apresentou valor de 0,4977.

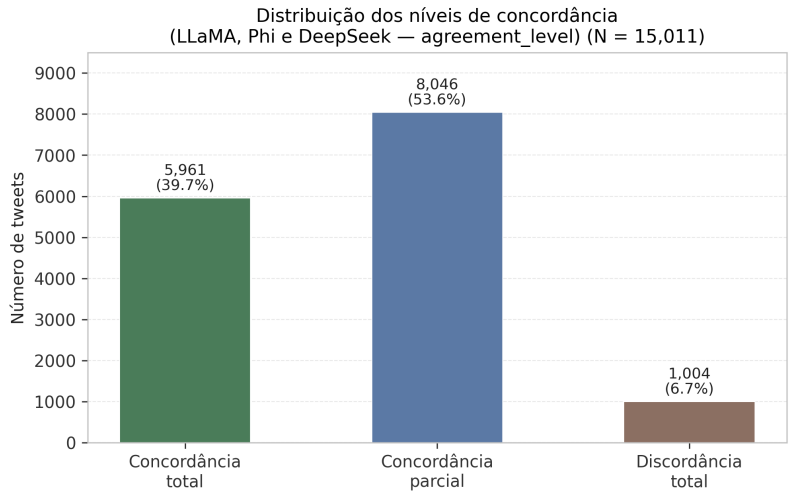


Figura 5. Distribuição dos níveis de concordância entre os rotuladores automáticos.

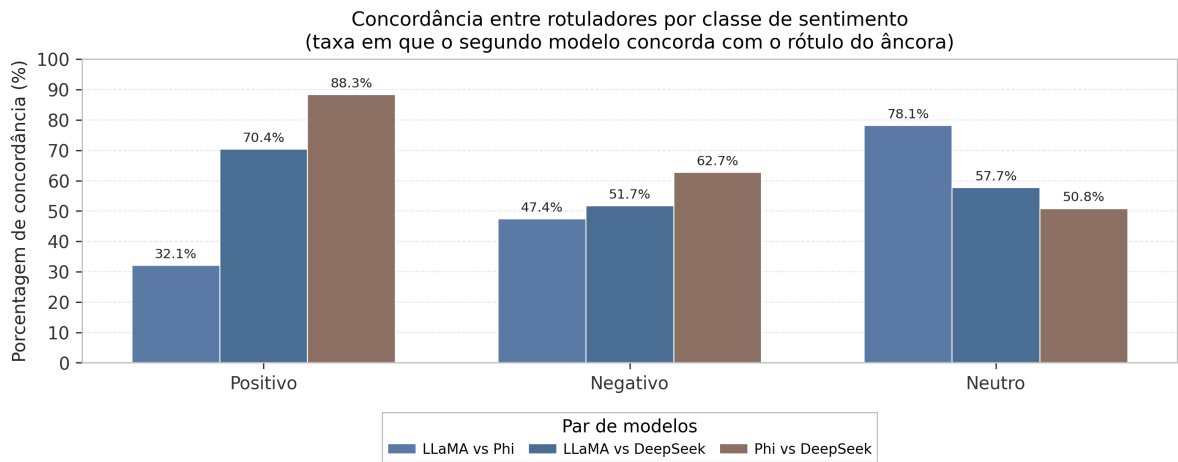


Figura 6. Concordância entre pares de rotuladores por classe de sentimento.

Tabela 3. Desempenho geral do modelo Mistral em relação ao conjunto de referência

Métrica	Valor
Acurácia	68%
Macro precisão	70%
Recall	68%
Macro F1-score	67%
Cohen’s Kappa	0,4977
Total de tweets	15.011

Tabela 4. Métricas do modelo Mistral por classe de sentimento

Classe	Precisão	Recall	F1-score	Tweets
Positivo	0,56	0,82	0,66	3.832
Negativo	0,83	0,54	0,66	3.821
Neutro	0,72	0,68	0,70	7.358

Essa métrica considera a concordância entre duas classificações descontando a concordância esperada ao acaso. De acordo com a classificação proposta por Landis and Koch [1977], esse valor representa uma concordância moderada entre as classificações produzidas pelo *Mistral* e o conjunto de referência. Esse resultado está de acordo com as demais métricas apresentadas, indicando que o modelo foi capaz de reproduzir boa parte das classificações obtidas pelo processo

de *ensemble*, embora ainda apresente dificuldades em publicações com pouca polarização emocional.

A Figura 7 auxilia na compreensão do desempenho do modelo apresentando a matriz de confusão obtida durante a avaliação. Na matriz de confusão, as linhas representam os rótulos do conjunto de referência, enquanto as colunas representam as classificações produzidas pelo *Mistral*. Os valores presentes na diagonal principal correspondem aos acertos do modelo, enquanto os demais valores representam classificações divergentes.

Para a classe positiva, o modelo classificou corretamente 3.132 publicações. A revocação dessa classe foi de 0,82, indicando que grande parte dos tweets positivos presentes no conjunto de referência foi identificada corretamente. Entretanto, a precisão foi menor, com valor de 0,56, principalmente devido à quantidade de tweets neutros classificados como positivos. Ao todo, 1.994 publicações neutras receberam classificação positiva pelo modelo.

Na classe negativa, o modelo acertou 2.077 publicações. A precisão foi relativamente alta, com valor de 0,83, indicando que os tweets classificados como negativos pelo modelo geralmente pertenciam de fato a essa categoria. Por outro lado, a revocação foi menor, com valor de 0,54, pois muitos tweets negativos foram classificados como neutros. Nesse

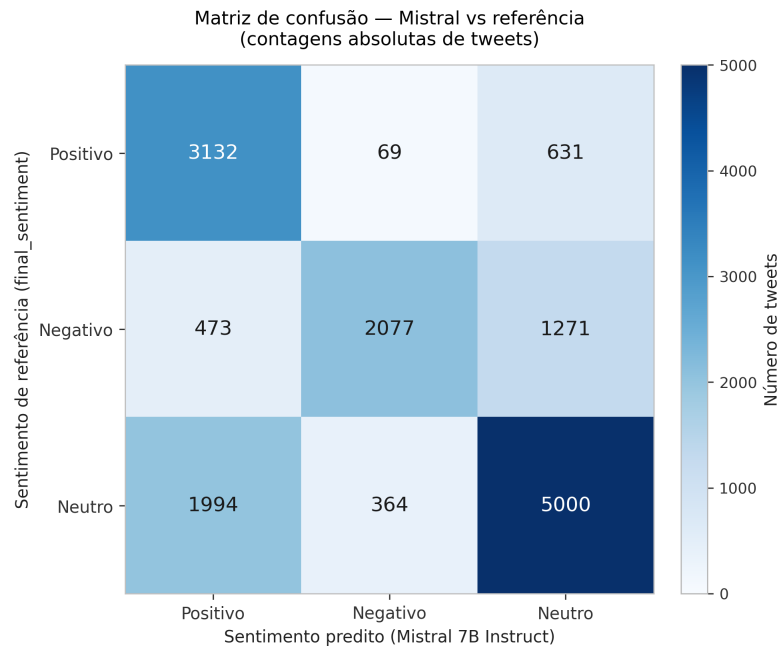


Figura 7. Matriz de confusão do modelo Mistral em relação ao conjunto de referência.

caso, 1.271 publicações negativas foram previstas como neutras.

Para os sentimentos neutros, o modelo classificou corretamente 5.000 publicações. Essa foi a classe com maior número absoluto de acertos, o que também está relacionado ao fato de os sentimentos neutros representarem a maior parcela do conjunto analisado. A precisão foi de 0,72 e a revocação de 0,68, indicando desempenho relativamente equilibrado nessa categoria.

A principal divergência observada ocorreu entre sentimentos neutros e positivos. Em muitos casos, publicações consideradas apenas informativas ou descritivas no conjunto de referência receberam classificação positiva pelo *Mistral*. Também foram observadas dificuldades na separação entre sentimentos negativos e neutros, principalmente em publicações relacionadas a preocupações técnicas, críticas moderadas ou comentários profissionais.

Esses resultados mostram que a maior dificuldade do modelo esteve na interpretação de conteúdos ambíguos ou com baixa polarização emocional. Mesmo assim, o *Mistral* apresentou classificações relativamente consistentes para uma abordagem *zero-shot*, sem treinamento específico para o conjunto de dados utilizado neste trabalho.

5.6 Evolução Temporal dos Sentimentos

Além da análise global, também foi analisada a evolução dos sentimentos ao longo dos 90 dias posteriores aos eventos considerados no estudo. Para isso, foi utilizado tempo relativo, em que o dia zero corresponde à data associada ao lançamento inicial ou à versão recente de cada tecnologia analisada. A Figura 8 apresenta a evolução temporal das proporções de sentimentos para GPT.

No caso do GPT, observa-se comportamento diferente entre os três sentimentos ao longo do período analisado. Os sentimentos neutros permaneceram predominantes durante praticamente toda a janela temporal, indicando que muitas

publicações possuíam caráter mais informativo, descritivo ou voltado ao compartilhamento de experiências de uso da ferramenta.

Os sentimentos positivos aparecem com maior frequência próximos aos eventos analisados, principalmente nos períodos iniciais. Esse comportamento pode estar relacionado ao interesse gerado pelas novas funcionalidades e pela repercussão pública associada aos lançamentos.

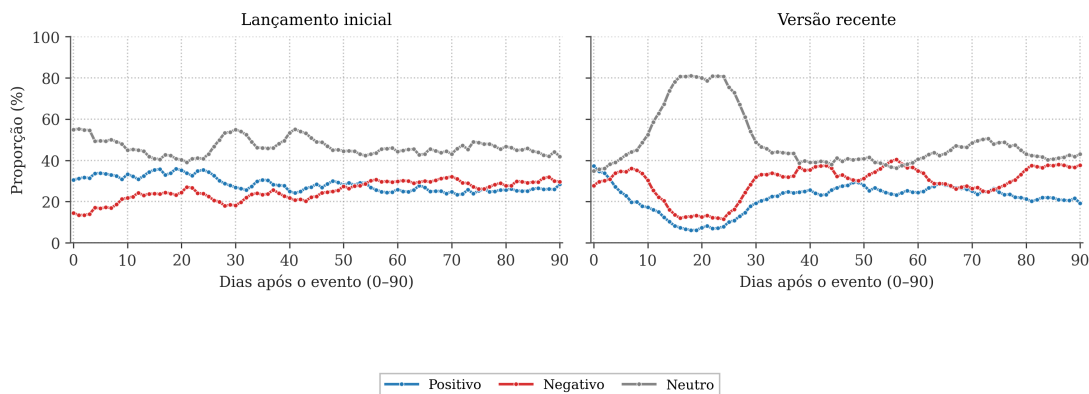
Ao longo do tempo, especialmente nas versões mais recentes, percebe-se redução gradual da proporção de sentimentos positivos e aumento relativo dos sentimentos negativos. Esse comportamento pode indicar que, após o período inicial de entusiasmo, parte das discussões passou a incluir mais críticas, preocupações ou expectativas não atendidas.

A Figura 9 apresenta a evolução temporal dos sentimentos relacionados ao Gemini. Percebe-se que os sentimentos neutros também permaneceram predominantes durante grande parte do período analisado. Em comparação com o GPT, as oscilações observadas entre sentimentos positivos e negativos foram menores, sugerindo um comportamento temporal mais estável nas discussões relacionadas à ferramenta. Os sentimentos positivos apresentaram leve crescimento em alguns períodos das versões recentes, enquanto os sentimentos negativos permaneceram relativamente estáveis na maior parte da série temporal.

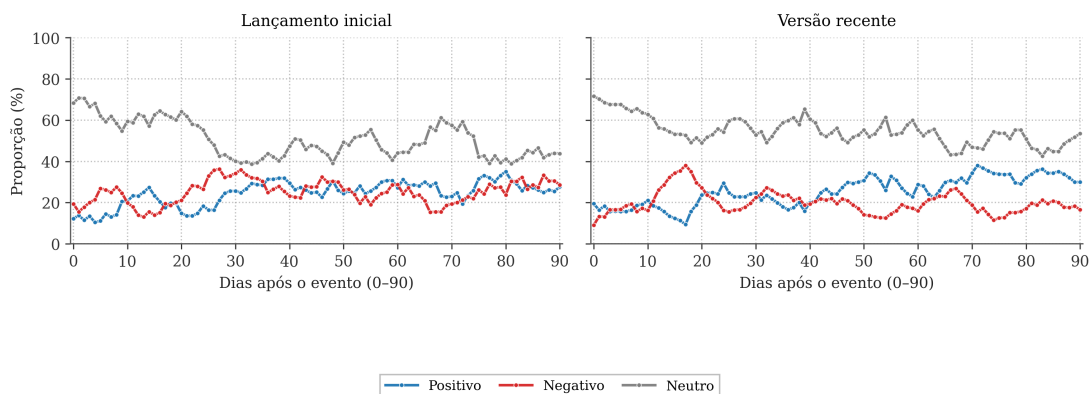
A Figura 10 apresenta a evolução temporal dos sentimentos relacionados ao Copilot. Note que o Copilot obteve um comportamento temporal mais irregular, principalmente durante o período associado ao lançamento inicial. As curvas dos três sentimentos mostraram oscilações mais intensas ao longo do tempo quando comparadas às demais tecnologias analisadas.

Os sentimentos positivos aparecem com frequência relativamente elevada em parte do período analisado, especialmente nas versões recentes. Esse comportamento pode estar relacionado ao contexto específico de uso da ferramenta, fre-

GPT — evolução temporal da referência (final_sentiment)

**Figura 8.** Evolução temporal dos sentimentos para GPT.

Gemini — evolução temporal da referência (final_sentiment)

**Figura 9.** Evolução temporal dos sentimentos para Gemini.

quentemente associado a produtividade e desenvolvimento de software.

Ao mesmo tempo, também foram observados aumentos pontuais nos sentimentos negativos em alguns períodos. Essas variações podem estar relacionadas a críticas técnicas, limitações percebidas pelos usuários ou experiências negativas compartilhadas nas publicações.

De forma geral, os resultados mostram que os sentimentos neutros permaneceram predominantes para as três tecnologias durante grande parte do período analisado. Ainda assim, as mudanças observadas nos sentimentos positivos e negativos indicam que a percepção dos usuários varia ao longo do tempo e pode sofrer influência do contexto de lançamento, da repercussão pública e das experiências de uso compartilhadas nas redes sociais.

5.7 Discussão Geral dos Resultados

De forma geral, os resultados mostram que as discussões relacionadas às tecnologias de inteligência artificial generativa apresentam predominância de neutralidade, acompanhada de níveis moderados de polarização positiva e negativa.

Os resultados também indicam que:

- Diferentes tecnologias apresentam comportamentos distintos de aceitação pública;
- Versões mais recentes tendem a gerar discussões mais críticas;
- Publicações neutras representam parcela significativa das interações observadas;
- A utilização de múltiplos rotuladores auxiliares contribuiu para reduzir inconsistências individuais; e
- Publicações ambíguas continuam sendo um desafio relevante para classificação automática de sentimentos.

Além disso, a comparação entre o Mistral e o conjunto consolidado mostrou que o modelo principal foi capaz de produzir classificações relativamente consistentes mesmo em abordagem *zero-shot*, embora ainda apresente dificuldades em cenários semanticamente ambíguos.

Por fim, os resultados obtidos reforçam o potencial do uso de *LLMs* em tarefas de análise de sentimentos aplicadas a redes sociais, principalmente em cenários envolvendo dados informais, multilíngues e semanticamente complexos.

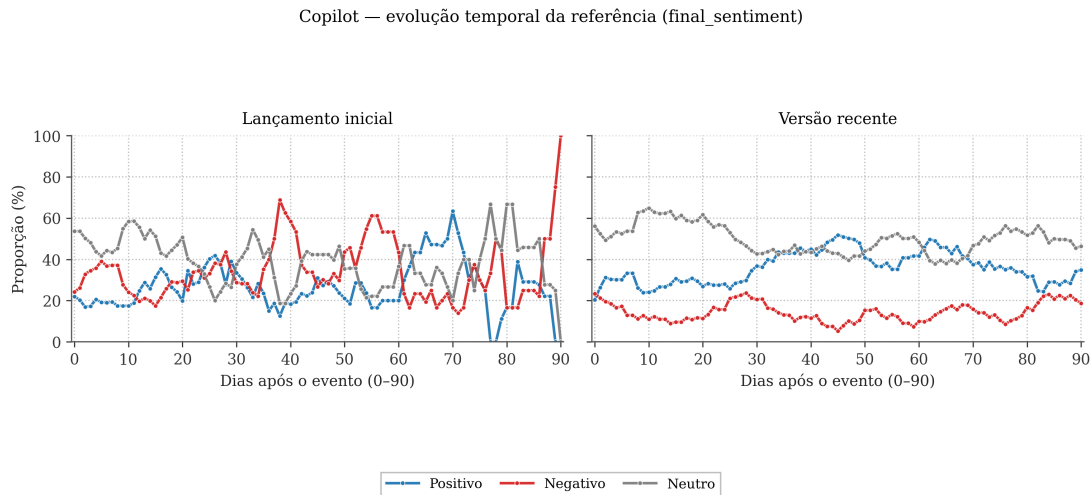


Figura 10. Evolução temporal dos sentimentos para Copilot.

6 Conclusão

Este trabalho apresentou uma análise de sentimentos aplicada a publicações do Twitter (X) relacionadas às tecnologias GPT/ChatGPT, Gemini e Copilot, considerando períodos associados aos lançamentos iniciais e às versões mais recentes dessas ferramentas.

Para realização do estudo, foi desenvolvido um processo envolvendo coleta de dados, pré-processamento, filtragem semântica e classificação automática utilizando LLMs. Após as etapas de tratamento dos dados, o conjunto final ficou composto por aproximadamente 15 mil *tweets* em português relacionados às tecnologias analisadas.

Os resultados mostraram predominância de sentimentos neutros durante grande parte do período analisado, indicando que muitas publicações tinham caráter mais informativo ou descritivo. Ainda assim, também foram observadas diferenças entre as tecnologias analisadas, principalmente na evolução dos sentimentos positivos e negativos ao longo do tempo.

Na etapa de avaliação, o modelo *Mistral* apresentou acurácia aproximada de 68% e *macro F1-score* de 67% em relação ao conjunto de referência construído a partir da combinação entre múltiplos modelos de linguagem. Os resultados também mostraram que publicações ambíguas, irônicas ou com opiniões implícitas continuam sendo um desafio para classificação automática de sentimentos.

As análises de concordância indicaram que diferentes modelos podem interpretar uma mesma publicação de maneiras distintas, principalmente em casos envolvendo humor, ironia ou linguagem informal. Mesmo assim, a utilização de múltiplos modelos ajudou a reduzir inconsistências individuais durante o processo de classificação.

Como principal contribuição, este trabalho apresentou uma abordagem baseada em LLMs para análise de sentimentos em dados reais de redes sociais, utilizando combinação entre múltiplos modelos, filtragem semântica e análise temporal baseada em tempo relativo. Além disso, o estudo também fornece uma visão sobre como a percepção pública em relação às tecnologias de IA generativa pode variar ao longo do tempo.

De forma geral, os resultados obtidos mostram que LLMs podem ser utilizados em tarefas de análise de sentimentos aplicadas a redes sociais, especialmente em cenários envolvendo grande volume de dados textuais e ausência de conjuntos previamente rotulados.

Como trabalhos futuros, pretende-se ampliar o conjunto de tecnologias analisadas, investigar publicações em outros idiomas e comparar os resultados obtidos com modelos especializados em análise de sentimentos e abordagens tradicionais de aprendizado de máquina.

Referências

- Ahnert, R., Smith, D., and Kühn, N. (2024). Extracting affect aggregates from longitudinal social media data with temporal adapters for large language models. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. Acesso em: 11 nov. 2025.
- Bhargava, A., Sethi, R., Anand, K., and Ghosh, A. (2025). On the impact of language nuances on sentiment analysis with large language models: Paraphrasing, sarcasm, and emojis. *arXiv preprint arXiv:2504.05603*. Acesso em: 11 nov. 2025.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Bruns, A. and Burgess, J. (2015). Twitter hashtags from ad hoc to calculated publics. In Rambukkana, N., editor, *Hashtag Publics: The Power and Politics of Discursive Networks*, pages 13–28. Peter Lang, New York, NY, USA.
- Cambria, E., Hussain, A., and Schuller, B. (2023). Sentiment analysis for social and affective computing: Recent advances and challenges. *Information Fusion*, 90:126–143. DOI: 10.1016/j.inffus.2022.11.017.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep*

- Learning*. MIT Press.
- Google DeepMind (2024). Gemini 1.5 technical report. Technical report, Google DeepMind. Acesso em: 27 out. 2025.
- Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing*. Stanford University, 3rd edition, draft edition.
- Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, Raleigh, NC, USA. ACM. DOI: 10.1145/1772690.1772751.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174. DOI: 10.2307/2529310.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. DOI: 10.1038/nature14539.
- Liu, B., Cambria, E., Xiang, Y., and Ma, Y. (2023). A survey of sentiment analysis using large language models. *IEEE Transactions on Affective Computing*. DOI: 10.1109/TAFFC.2023.3264201.
- Marwick, A. E. (2013). Ethnographic and qualitative research on twitter. In Weller, K., Bruns, A., Burgess, J., Mahrt, M., and Puschmann, C., editors, *Twitter and Society*, pages 109–122. Peter Lang, New York, NY, USA.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.
- Mun, S. and Kim, H. (2025). Leveraging large language models for sentiment analysis and investment strategy development in financial markets. *Journal of Theoretical and Applied Electronic Commerce Research*, 20(1):77. Acesso em: 11 nov. 2025.
- Murthy, D. (2018). *Twitter: Social Communication in the Twitter Age*. Polity Press, Cambridge, UK.
- OpenAI (2023). Gpt-4 technical report. Technical report, OpenAI. Acesso em: 27 out. 2025.
- Raschka, S. (2024). *Build a Large Language Model (From Scratch)*. Manning Publications, New York, NY, USA.
- Rathje, S., Bavel, J. J. V., and et al. (2024). Gpt is an effective tool for multilingual psychological text annotation. *Proceedings of the National Academy of Sciences*, 121(13). DOI: 10.1073/pnas.2316873121.
- Simmering, P., Chen, X., and Liu, Y. (2023). Large language models for aspect-based sentiment analysis. *arXiv preprint arXiv:2309.11258*.
- Touvron, H., Martin, L., Stone, K., Albert, P., and et al. (2024). The llama 3 herd of models. Acesso em: 27 out. 2025.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser, and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA.
- Wang, X. and Chandra, P. (2024). A longitudinal sentiment analysis of sinophobia during covid-19 using large language models. *arXiv preprint arXiv:2403.17842*. Acesso em: 27 out. 2025.
- Zhang, Y., Wang, S., and et al. (2024). Sentiment analysis in the era of large language models: A reality check. *Findings of NAACL*. DOI: 10.18653/v1/2024.findings-naacl.72.