PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA E DE ARTES
GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

ANÁLISE DA CORRELAÇÃO ENTRE ESTILO TEXTUAL E OPINIÕES EM REDES SOCIAIS MULTICULTURAIS POR MEIO DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

CARINE SILVA SANTOS

GOIÂNIA
2025

CARINE SILVA SANTOS

ANÁLISE DA CORRELAÇÃO ENTRE ESTILO TEXTUAL E OPINIÕES EM REDES SOCIAIS MULTICULTURAIS POR MEIO DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Conclusão de Curso apresentado à Escola Politécnica e de Artes, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para a obtenção do título de Bacharel em Engenharia da Computação. Orientador(a):

Dra. Maria José Perreira Dantas

GOIÂNIA

2025

CARINE SILVA SANTOS

ANÁLISE DA CORRELAÇÃO ENTRE ESTILO TEXTUAL E OPINIÕES EM REDES SOCIAIS MULTICULTURAIS POR MEIO DE TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Conclusão de Curso apresentado à Escola Politécnica e de Artes, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para a obtenção do título de Bacharel em Engenharia da Computação.

Aprovado em: ___/___/____

Dra. Maria José Perreira Dantas
Orientador

Prof. Convidado 1
Instituição

Prof. Convidado 2
Instituição

GOIÂNIA
2025

RESUMO

Este trabalho investigou a relação entre estilo de escrita e conteúdo opinativo em redes sociais multiculturais por meio de embeddings gerados pelos modelos STAR (estilo) e Qwen3-Embedding (opinião). Apesar da correlação linear entre as dimensões ser praticamente nula ($r \approx 0,01$), a análise bivariada revelou uma distribuição elíptica altamente organizada, compatível com a geometria de uma distribuição normal bivariada anisotrópica (variâncias diferentes nos eixos). Essa estrutura indica ausência de dependência linear, mas presença de dependência não linear modelável, refletida na anisotropia e nos diferentes graus de variância entre estilo e opinião. A maior densidade observada em regiões intermediárias sugere que a maioria dos autores compartilha padrões expressivos medianos, enquanto opiniões se distribuem de maneira mais ampla e multimodal. Esses resultados demonstram que estilo e opinião não são independentes; ao contrário, organizam-se em um espaço latente comum, estruturado e não monotônico, típico de representações produzidas por modelos Transformers. As limitações relacionadas à heterogeneidade dos dados e ao uso de embeddings pré-treinados são discutidas, e propõem-se direções futuras baseadas em métodos não lineares, como regressões gaussianas multivariadas, Kernel CCA e técnicas de clusterização neural.

Palavras-chave: Estilo de escrita. Opinião. Embeddings. Transformers. Dependência não linear. Análise bivariada.

ABSTRACT

This study investigates the relationship between writing style and expressed opinions in multicultural social media by using embeddings generated with the STAR model (style) and the Qwen3-Embedding model (opinion). Although the linear correlation between both dimensions is nearly zero ($r \approx 0.01$), the bivariate analysis revealed a highly structured elliptical distribution, consistent with the geometry of an anisotropic bivariate normal distribution. This pattern indicates the absence of linear dependency but the presence of a modelable nonlinear dependency, expressed through the anisotropy and unequal variance along the latent dimensions. The high-density region concentrated around intermediate similarity values suggests that most authors share median stylistic patterns, while opinions spread more broadly and multimodally. The findings demonstrate that style and opinion are not independent; rather, they are jointly organized within a structured, non-monotonic latent space typical of Transformer-based representations. Limitations related to data heterogeneity and the use of pretrained embeddings are discussed, and future research directions involving nonlinear methods, such as multivariate Gaussian regression, Kernel CCA, and neural embedding-based clustering are proposed.

Keywords: Writing style. Opinion. Embeddings. Transformers. Nonlinear dependency. Bivariate analysis.

SUMÁRIO

1 INTRODUÇÃO

Ambientes de discurso não moderado, como *imageboards* e fóruns anônimos, constituem espaços onde indivíduos expressam opiniões sem filtros e com alta carga emocional. Nessas comunidades, o estilo textual, manifestado por escolhas lexicais, uso de ironia e sarcasmo, repetição de padrões sintáticos ou marcas de intensidade emocional, desempenha papel essencial na sinalização de pertencimento e alinhamento ideológico.

A literatura demonstra que o estilo pode alterar a interpretação da opinião, especialmente em textos irônicos, nos quais o conteúdo literal pode expressar algo distinto da intenção comunicada (EKE et al., 2021). Dessa forma, modelos tradicionais de análise de sentimento tornam-se insuficientes quando não consideram a dimensão estilística.

Pesquisas recentes evidenciam que o estilo pode ser manipulado estrategicamente como mecanismo de influência em debates digitais. *Bots* ajustam conscientemente seu estilo de escrita para simular autenticidade, o que transforma o estilo em marcador de comportamento social e não apenas linguístico (SALLAH et al., 2024).

Entretanto, há uma predominância de estudos realizados apenas com dados em inglês. Estudos recentes demonstram que nuances emocionais, como esperança, indignação e agressividade, são expressas de maneiras diferentes conforme a cultura, reforçando a necessidade de bases que representem diversidade discursiva (SIDOROV et al., 2025). Em cenários de baixo recurso, modelos treinados exclusivamente em inglês apresentam perda de desempenho, o que evidencia a importância de bases diversificadas (NETO et al., 2023).

A construção da base de dados para este estudo utilizará apenas plataformas que permitem acesso público a dados para fins de pesquisa, como o Reddit via API oficial. Serão excluídas plataformas que não são públicas por padrão, como o Discord, salvo situações em que houver autorização explícita de administradores e consentimento dos participantes. Todo o processo seguirá princípios éticos, com descarte de qualquer dado que permita identificar indivíduos, remoção de *usernames*, *links* ou metadados pessoais. A curadoria atenderá aos requisitos da LGPD e da GDPR, utilizando anonimização irreversível quando necessário para eliminar riscos de reidentificação.

Outro desafio refere-se à escala dos dados. Repositórios históricos de discurso não moderado frequentemente totalizam terabytes de textos. A literatura demonstra que bases dessa magnitude só podem ser processadas em ambientes distribuídos, com armazenamento em *data lake* e processamento paralelo suportado por GPU, especialmente quando se aplicam modelos Transformer e *embeddings* especializados (ANOOP et al., 2024).

A infraestrutura disponível para este projeto inclui máquinas de alto desempenho, com armazenamento massivo e aceleração por GPU, tornando possível o processamento de dados em larga escala e a experimentação de modelos de linguagem sob diferentes estratégias de representação.

Para garantir a qualidade da análise, é necessária a curadoria dos dados, com o uso de filtros estatísticos. Mas, por outro lado, grandes Modelos de Linguagem (LLMs) podem ser usados para classificar, semanticamente, as interações que demandam julgamento subjetivo, em oposição às puramente factuais. Cabe destacar que o pipeline de engenharia de dados utilizado para a aquisição e limpeza do corpus deste trabalho deriva de um esforço de pesquisa mais amplo, cujos métodos de construção de datasets robustos foram validados e publicados pelos autores na IEEE COMPSAC 2025 (JORDãO; SANTOS; DANTAS, 2025). Aquele estudo estabeleceu as bases para o processamento de linguagem informal em larga escala, uma infraestrutura que permitiu a seleção e a curadoria refinadas dos dados do r/AskReddit utilizados nesta monografia.

Com base na literatura, observa-se uma lacuna científica: ainda não há consenso sobre a natureza da relação entre o estilo de escrita e a opinião, analisando se essas dimensões operam de forma independente ou correlacionada.

1.1 Objetivos

1.1.1 Objetivo Geral

Investigar a relação entre estilo de escrita e opinião expressa em redes sociais multiculturais, analisando como essas duas dimensões se organizam no espaço latente produzido pelos embeddings dos modelos STAR (estilo) e Qwen3-Embedding (opinião), com foco na identificação de padrões estruturais, dependências não lineares e características geométricas das representações vetoriais.

1.1.2 Objetivos Específicos

- Extrair e representar embeddings de estilo para cada autor utilizando o modelo STAR, de modo a capturar características formais da escrita e sintetizar assinaturas estilísticas consistentes.
- Extrair embeddings de opinião para cada resposta dos autores utilizando o modelo Qwen3-Embedding, preservando a variabilidade semântica e multimodal do conteúdo opinativo.
- Comparar as dimensões de estilo e opinião por meio de análises bivariadas de similaridade, avaliando correlação linear, densidade conjunta e propriedades geométricas da distribuição, incluindo anisotropia, dispersão e forma elipsoidal.
- Identificar padrões de dependência entre estilo e opinião no espaço latente, interpretando a estrutura resultante à luz das propriedades dos embeddings Transformer e discutindo implicações metodológicas para estudos futuros em análise autoral e modelagem de discurso.

A literatura demonstra que o estilo interfere na interpretação da opinião, podendo inclusive inverter a polaridade emocional percebida (EKE et al., 2021). Em ambientes digitais não mo-

derados, o estilo pode ser utilizado como mecanismo de manipulação retórica e de influência social (SALLAH et al., 2024). A diversidade de dados é essencial para reduzir vieses e melhorar a interpretação de fenômenos socioculturais relacionados ao discurso (SIDOROV et al., 2025; NETO et al., 2023).

Evidências recentes também mostram que modelos especializados em domínios específicos apresentam melhor desempenho que modelos genéricos (ANOOP et al., 2024), sugerindo que o estilo pode ser tratado como uma dimensão contextual passível de modelagem. Assim, verificar se a forma de escrever pode revelar opinião antes de se analisar o conteúdo representa uma contribuição teórica, metodológica e social ainda inexistente na literatura.

Embora o estilo possa variar conforme o contexto, a literatura sugere a existência de uma "assinatura estilística" latente e relativamente estável em autores prolíficos, o que permite modelar o conteúdo vetorial de forma distinta do conteúdo opinativo instável (WANG et al., 2023).

2 REFERENCIAL TEÓRICO

O Processamento de Linguagem Natural (PLN), ou *Natural Language Processing* (NLP), é uma subárea da Inteligência Artificial que estuda como as máquinas podem compreender, interpretar e gerar linguagem humana de forma significativa. Desde os primeiros estudos computacionais sobre linguagem, essa área busca construir pontes entre linguística e algoritmos, tornando possível a interação entre humanos e computadores por meio da linguagem escrita e falada (JURAFSKY; MARTIN, 2023).

2.1 Processamento de Linguagem Natural (PLN) e Conceitos Essenciais

Inicialmente, os sistemas de PLN baseavam-se em regras gramaticais rígidas e modelos probabilísticos simples, como os *n-grams*, que estimavam a probabilidade de ocorrência de palavras baseadas em sequências anteriores. Esses métodos foram importantes, mas limitados na captura de dependências semânticas mais amplas. Com o tempo, a área passou a empregar modelos estatísticos supervisionados, como os Modelos Ocultos de Markov (HMMs) e os Campos Aleatórios Condicionais (CRFs), capazes de melhorar tarefas como rotulação de entidades e análise sintática (MANNING; SCHÜTZE, 1999).

Antes que um modelo de linguagem possa processar texto, ele precisa converter esse texto em uma forma numérica. Esse processo começa pela tokenização, que divide a entrada em unidades menores chamadas *tokens*. Um token pode ser uma palavra inteira, uma subpalavra ou até mesmo um único caractere, dependendo da técnica usada (SENNRICH; HADDOW; BIRCH, 2016). Por exemplo, a palavra “computação” pode ser dividida em [“com”, “puta”, “ção”] por tokenizadores baseados em subpalavras, como o BPE (Byte Pair Encoding).

Cada token precisa então ser convertido em um vetor numérico, é aí que entram os *embeddings*. Embeddings são representações vetoriais densas que mapeiam palavras em espaços de alta dimensionalidade, onde relações semânticas e sintáticas são preservadas. Modelos como Word2Vec (MIKOLOV et al., 2013) e GloVe (PENNINGTON; SOCHER; MANNING, 2014) aprenderam esses vetores com base em coocorrência de palavras, mas produzem representações estáticas. Modelos mais modernos, como ELMo (PETERS et al., 2018), BERT e GPT, passaram a utilizar embeddings contextuais, nos quais o vetor de uma palavra muda de acordo com o contexto da frase.

A transição para o aprendizado profundo (*deep learning*) representou um salto significativo. Redes neurais recorrentes (RNNs), especialmente as variantes LSTM e GRU, passaram a ser usadas em larga escala, pois permitiam capturar dependências de longo prazo em sequências de texto (GOLDBERG, 2016). Contudo, essas redes eram difíceis de treinar e não permitiam paralelização eficiente, o que limitava sua escalabilidade.

A grande revolução chegou com o modelo Transformer, proposto por Vaswani et al. (2017)

no artigo *Attention is All You Need*, que substituiu as RNNs por um mecanismo chamado atenção. Essa abordagem permite que o modelo avalie, simultaneamente, todas as palavras da sequência para entender o contexto de cada uma. A operação de atenção é definida pela Equação 2.1:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

Onde Q, K e V representam *queries*, *keys* e *values*, vetores derivados da entrada textual, e d_k é a dimensão dos vetores de chave. Com isso, o modelo consegue capturar relações semânticas e sintáticas entre palavras, mesmo quando estão distantes no texto (VASWANI et al., 2017).

Os Transformers são a base de modelos avançados como BERT (DEVLIN et al., 2019), GPT-3 (BROWN et al., 2020) e Qwen (YANG et al., 2016), que têm demonstrado excelentes desempenhos em tarefas de tradução, classificação, geração de texto e análise de sentimentos. Eles também são aplicados a contextos diversos, como mostram os estudos de Conneau et al. (2020) e Jordão, Santos e Dantas (2025), ampliando a capacidade de análise em ambientes culturais distintos e informais como Reddit.

2.2 Análise Linguística para Tomada de Decisões

A análise linguística computacional vem se consolidando como um dos pilares da inteligência artificial aplicada à tomada de decisões, especialmente em ambientes digitais marcados pela informalidade, subjetividade e diversidade cultural. Em plataforma como Reddit, onde os discursos são espontâneos, não moderados e altamente interativos, a linguagem assume papéis complexos — como veicular opiniões, marcar estilos individuais, expressar identidades coletivas e responder a eventos sociais com ironia, humor ou polarização.

Para que máquinas possam interpretar essas nuances, é necessário ir além da simples classificação de sentimentos ou detecção de palavras-chave. A linguagem carrega camadas profundas de informação pragmática, sintática e semântica, que precisam ser interpretadas em conjunto. Por isso, abordagens linguísticas modernas combinam modelos formais com técnicas de aprendizado de máquina para inferir, por exemplo, a intenção de uma mensagem, a posição subjetiva do autor ou o estilo com que determinada ideia é transmitida (PAVLOPOULOS; MALAKASIOTIS; ANDROUTSOPOULOS, 2017; PAVLOPOULOS et al., 2017).

Em ambientes como o Reddit, estudos demonstram que a densidade linguística e a variação estilística contribuem para agrupar usuários com posições ideológicas semelhantes (GLAVAS et al., 2017), o que pode ser explorado para analisar correlações entre forma e conteúdo.

O artigo “Creating and Validating the Fine-Grained Question Subjectivity Dataset (FQSD)” (BHATIA et al., 2023) exemplifica como anotações linguísticas finas podem orientar a identificação de subjetividade em perguntas. Já o trabalho “Understanding Writing Style in Social Media” (ZHOU et al., 2023) demonstra como o estilo autoral pode ser representado computacionalmente para distinguir usuários com base em padrões linguísticos e discursivos.

No contexto deste trabalho, tal articulação é essencial para estudar como o estilo de escrita influencia a expressão de opinião, ou mesmo, a forma como uma opinião é percebida. Como destacam Jordão, Santos e Dantas (2025) e Papasavva et al. (2020), em contextos informais, a estrutura da linguagem é tão importante quanto o conteúdo explícito para compreender discursos, detectar viés e fazer inferências relevantes.

2.3 Da Estatística à Arquitetura Transformer

O marco seguinte foi a formulação dos Transformers (VASWANI et al., 2017), que introduziram o mecanismo de *self-attention*. Essa arquitetura eliminou a necessidade de recorrência, permitindo o processamento paralelo e a captura eficiente de relações entre tokens distantes. O modelo Transformer estabeleceu uma mudança profunda no Processamento de Linguagem Natural.

A base do Transformer é o mecanismo de autoatenção. Para cada token de entrada, representado em um vetor x_i, o modelo projeta vetores de consulta, chave e valor. Essa transformação é feita pelas matrizes aprendidas W^Q, W^K, W^V. Assim, para uma matriz de entrada $X \in \mathbb{R}^{N \times d_{\text{model}}}$, onde N é o número de tokens e d_{model} é a dimensionalidade dos embeddings, definem-se:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{2.2}$$

A atenção determina o quanto cada token deve considerar os demais. Esse cálculo é feito comparando consultas e chaves, o que produz uma matriz de similaridade. Para estabilizar o treinamento, divide-se esse produto pelo fator $\sqrt{d_k}$. Em seguida, aplica-se uma função Softmax linha a linha, que gera pesos normalizados. A equação central da atenção escalonada é:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.3}$$

Para enriquecer a expressividade, o Transformer utiliza múltiplas cabeças de atenção. O mecanismo completo é descrito por:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.4}$$

Onde cada cabeça é definida por:

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V) \tag{2.5}$$

Após a etapa de atenção, cada vetor passa por uma rede totalmente conectada chamada *feedforward*, aplicada de forma independente a cada posição da sequência:

$$\text{FFN}(x) = \sigma(xW_1 + b_1)W_2 + b_2 \tag{2.6}$$

Para garantir estabilidade no treinamento, cada sub-bloco utiliza conexões residuais seguidas por normalização de camada. Assim, para uma camada l, tem-se:

$$X^{(l)} = \text{LayerNorm}(X^{(l-1)} + \text{MultiHead}(X^{(l-1)})) \tag{2.7}$$

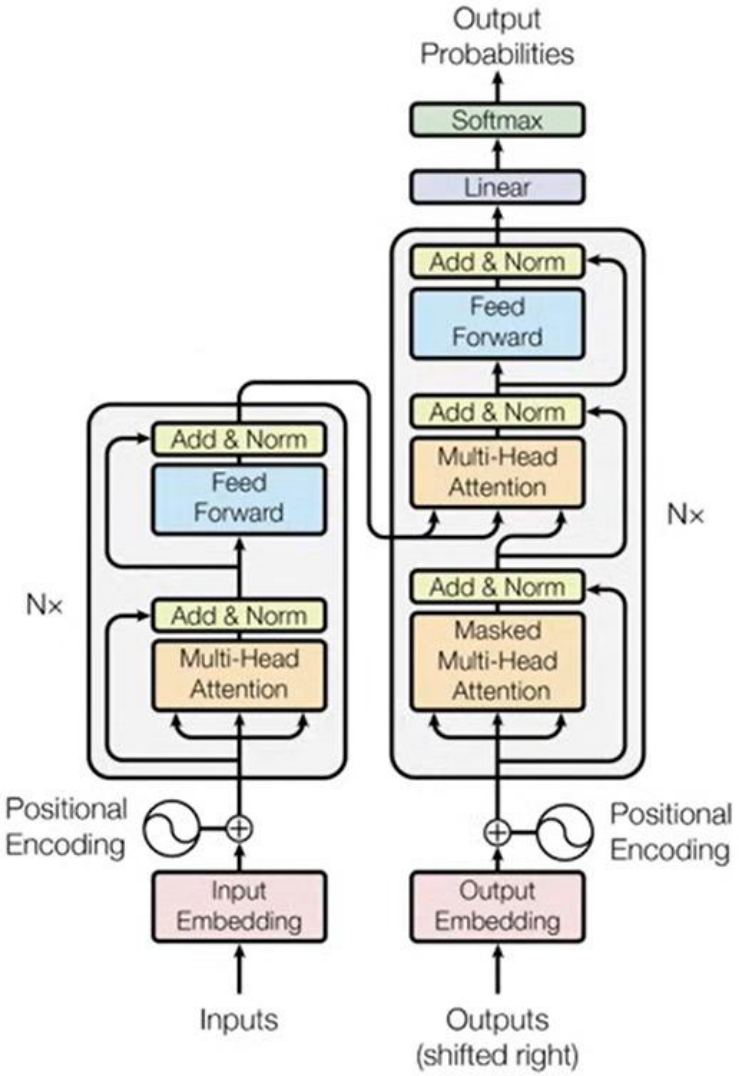
$$X^{(l)} = \text{LayerNorm}(X^{(l)} + \text{FFN}(X^{(l)})) \tag{2.8}$$

Como o Transformer não impõe ordem sequencial explícita, é necessário incorporar informações de posição aos embeddings iniciais por meio das codificações posicionais (PE):

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.9}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{2.10}$$

Figura 1 – Arquitetura Transformer



Fonte: Vaswani et al. (2017).

A Figura 1 apresenta a arquitetura do modelo Transformer, estruturada no formato Encoder-Decoder. O componente à esquerda, denominado Encoder, é composto por uma pilha de N camadas idênticas. Cada camada possui duas subcamadas principais: um mecanismo de autoatenção com múltiplas cabeças (Multi-Head Attention) e uma rede neural feed-forward conectada por posição. À direita, o Decoder também é formado por uma pilha de N camadas idênticas. No entanto, além das duas subcamadas presentes no Encoder, o Decoder insere uma terceira subcamada intermediária, que realiza a atenção sobre a saída do Encoder (atenção cruzada). É importante notar o uso do mecanismo de Masked Multi-Head Attention no início do Decoder, que garante que a previsão para uma determinada posição dependa apenas das posições anteriores conhecidas, preservando a propriedade auto-regressiva do modelo. Em ambos os lados, empregam-se conexões residuais seguidas de normalização de camada. Por fim, a figura destaca a injeção de Positional Encodings somadas aos embeddings de entrada e saída, permitindo que o modelo considere a ordem sequencial dos tokens (VASWANI et al., 2017).

2.3.1 Explicando o Transformer

Modelos derivados, como BERT, GPT, T5 e Qwen, estendem essa arquitetura para diferentes tarefas, como geração de texto, classificação, análise de sentimentos, tradução e detecção de subjetividade. A capacidade do Transformer de capturar nuances estilísticas e semânticas torna-o particularmente adequado para este trabalho, que busca identificar padrões de correlação entre estilo discursivo e opinião em ambientes digitais e não moderados como Reddit.

A consolidação do Transformer ocorreu com o BERT (DEVLIN et al., 2019), que introduziu o pré-treinamento bidirecional por *masked language modeling* (MLM) e *next sentence prediction* (NSP), utilizando atenção *multi-head* sobre toda a sequência. Suas variantes RoBERTa, DistilBERT e ALBERT trouxeram aprimoramentos em dados, eficiência e compartilhamento de parâmetros (LIU et al., 2019; SANH et al., 2019; LAN et al., 2020). No contexto do português, o BERTimbau demonstrou vantagens ao ser pré-treinado diretamente no idioma (SOUZA; NOGUEIRA; LOTUFO, 2020).

Modelos autoregressivos, como a família GPT (RADFORD et al., 2018; BROWN et al., 2020), adotam *decoders* Transformer com máscara causal, permitindo aprendizagem *in-context* e desempenho *zero-shot* em várias tarefas. Essas arquiteturas impulsionaram aplicações generativas e expandiram o uso do PLN para tarefas antes consideradas inviáveis sem grandes quantidades de dados rotulados.

Modelos globais, como mBERT e XLM-R, surgiram para lidar com cenários de baixa disponibilidade e com múltiplos idiomas simultaneamente, sendo adequados para textos informais de redes sociais (NGUYEN; VU; NGUYEN, 2020; SOUZA; NOGUEIRA; LOTUFO, 2020). Avanços recentes incluem a linha Qwen, com variantes como Qwen3-Embedding e Qwen3-Reranker, projetadas para recuperação de informação e similaridade semântica, com forte desempenho devido ao pré-treinamento em larga escala (PENEDO et al., 2024).

Modelos especializados em estilo, autoria e sinalização discursiva também têm ganhado destaque, como o STAR (*Style Transformer for Authorship Representations*), treinado com aprendizado contrastivo para aproximar *embeddings* de textos do mesmo autor e distinguir estilos distintos. Essa abordagem tem sido utilizada em domínios como atribuição de autoria, análise estilística e detecção de padrões de escrita em redes sociais (HUERTAS-TATO et al., 2024).

Por fim, o desenvolvimento de técnicas de adaptação como *Domain-Adaptive Pretraining* (DAPT) e *Task-Adaptive Pretraining* (TAPT) melhorou o desempenho de modelos pré-treinados em domínios específicos, como finanças, saúde e linguagem de redes sociais (GURURANGAN et al., 2020). Em paralelo, estudos têm destacado desafios éticos envolvendo privacidade, viés algorítmico e auditoria, especialmente em modelos de larga escala utilizados em ambientes sensíveis (GOODFELLOW; BENGIO; COURVILLE, 2016; ROSENTHAL et al., 2017; SALLAH et al., 2024; SARANYA et al., 2025).

Assim, a trajetória do PLN reflete uma transição contínua de representações simples baseadas em frequência para modelos capazes de capturar estrutura, contexto, estilo e subjetividade em múltiplos idiomas. Os avanços recentes consolidaram o uso de Transformers e modelos pré-treinados como base para tarefas complexas, incluindo análise de estilo e opinião em ambientes digitais não moderados.

Pré-treinamento e fine-tuning

A maior inovação trazida pelos Transformers foi a possibilidade de pré-treinamento em larga escala, seguido por ajuste fino (*fine-tuning*) em tarefas específicas. O modelo é inicialmente treinado com grandes volumes de dados textuais (como livros, sites, fóruns) em tarefas auto-supervisionadas, como *Masked Language Modeling* (prever palavras ocultas). Posteriormente, é adaptado com dados específicos para tarefas como análise de sentimentos, classificação de opinião ou atribuição de estilo. Essa estratégia de transferência de aprendizado é hoje padrão em aplicações reais, permitindo que modelos sejam aplicados mesmo com poucos dados rotulados na tarefa final.

2.4 Espaços de Discurso Não Moderado: Reddit

O avanço das técnicas de PLN tem possibilitado a análise de linguagens informais em espaços virtuais não moderados. O Reddit é uma plataforma que reúne enorme diversidade de usuários e temas. Esses espaços se caracterizam por uma linguagem menos normatizada, marcada por gírias, ironia e polarização.

O corpus do Reddit processado por Pavlos et al. (2023) reúne milhões de mensagens entre 2010 e 2022. Do ponto de vista da relevância científica, o Reddit é valioso porque oferece diversidade discursiva e grande volume de dados espontâneos. A existência de metadados permite o uso de rótulos fracos (*weak supervision*) para anotar automaticamente aspectos de subjetividade e estilo.

2.5 Desafios Éticos e Considerações de Uso Responsável

A adoção ampla de Transformers levanta desafios que atravessam técnica, governança e impacto social. A literatura recomenda integrar mitigação de risco desde o desenho do sistema até a operação contínua.

(a) Viés e equidade: Modelos refletem vieses presentes nos dados de treino, podendo produzir saídas injustas a grupos específicos. Medidas incluem: curadoria e balanceamento de dados; auditorias com cortes demográficos; e ajustes por alinhamento/feedback (GOLDBERG, 2017; OUYANG et al., 2022). Em análise de sentimentos, é importante avaliar casos críticos como negação, sarcasmo e gíria regional (LIU, 2015).

(b) Privacidade e proteção de dados: Comentários de usuários podem conter dados pessoais. Boas práticas: anonimização/pseudonimização, retenção mínima, controle de acesso e conformidade legal. Logs e trilhas de auditoria apoiam responsabilização (BIRD; KLEIN; LOPER, 2009; LIU, 2015).

(c) Alucinações e confiabilidade: Modelos geradores podem produzir respostas plausíveis, porém incorretas. Mitigações: ancoragem por recuperação de documentos, controle de temperatura/top-p, padronização de formato, verificadores e, quando necessário, revisão humana (OPENAI, 2019; BROWN et al., 2020; OUYANG et al., 2022).

(d) Segurança e uso indevido: Riscos incluem geração de spam, conteúdo enganoso ou abusivo. É recomendável aplicar políticas de uso, filtros de conteúdo e monitoramento com limites de escopo (OPENAI, 2019; GOLDBERG, 2017).

(e) Transparência e explicabilidade: Em decisões sensíveis, stakeholders precisam entender como o sistema chega a uma saída. Documentação de dados/modelos, exemplos de erro representativos e métricas por classe ajudam a interpretar e melhorar o sistema (GOODFELLOW; BENGIO; COURVILLE, 2016; GOLDBERG, 2017).

(f) Deriva e manutenção: Linguagem e gírias evoluem; sem monitoramento, a performance degrada. É útil medir dataset shift e planejar re-treinos periódicos (DAPT/TAPT) quando o domínio muda (GURURANGAN et al., 2020).

(g) Custos e pegada computacional: Treinos e inferência em larga escala exigem computação/energia. Alternativas: modelos menores (DistilBERT, ALBERT), poda/quantização e cache. O custo-benefício deve ser medido contra baselines clássicos (GOODFELLOW; BENGIO; COURVILLE, 2016; SANH et al., 2019; LAN et al., 2020).

(h) Avaliação justa e reprodutibilidade: Comparações devem usar mesmas partições, mesmas métricas e sementes registradas; reportar intervalos e descrever o pipeline aumenta a confiabilidade (GOLDBERG, 2017; LIU, 2015).

Em síntese, uso responsável combina: (i) governança de dados e avaliação de viés; (ii) controles técnicos contra alucinações/abuso; (iii) transparência de métricas e limitações; e (iv) monitoramento pós-implantação — práticas coerentes com diretrizes atuais de alinhamento por feedback humano e com a tradição de avaliação rigorosa em PLN (OUYANG et al., 2022; BIRD; KLEIN; LOPER, 2009; LIU, 2015).

2.6 Lacunas e Oportunidades de Pesquisa

A fundamentação teórica e metodológica desta pesquisa foi construída a partir de um mapeamento sistemático da literatura, organizado de forma a contemplar tanto o rigor acadêmico consolidado quanto a velocidade de inovação característica da área de Processamento de Linguagem Natural (PLN). O levantamento bibliográfico encontra-se detalhado nos Apêndices A e B, que, em conjunto, fornecem o suporte necessário para a investigação da correlação entre estilo e opinião em ambientes digitais multiculturais.

O Apêndice A reúne artigos revisados por pares (peer-reviewed), publicados em conferências e periódicos de alto impacto (como ACL, EMNLP e IEEE). Estes trabalhos estabelecem as bases fundamentais da pesquisa, incluindo a arquitetura original dos Transformers , benchmarks essenciais para classificação em redes sociais e metodologias validadas para detecção de nuances subjetivas, como sarcasmo e discurso de ódio. A seleção destes estudos garante que a proposta esteja ancorada em métricas confiáveis e arquiteturas amplamente testadas pela comunidade científica.

Paralelamente, o Apêndice B apresenta uma seleção criteriosa de preprints (artigos em repositórios como arXiv). Na área de Inteligência Artificial Generativa, o ciclo de publicação tradicional muitas vezes não acompanha a velocidade das descobertas tecnológicas. Portanto, a inclusão destes trabalhos foi estratégica para incorporar inovações recentes que são centrais para a metodologia deste estudo, tais como a família de modelos Qwen , novas abordagens de tokenização (Byte Latent Transformer) e pipelines modernos de curadoria de dados multilinguísticos (FineWeb2).

O uso de preprints neste contexto não representa uma redução no rigor, mas sim uma busca ativa por insights de fronteira e inovações arquiteturais que ainda estão definindo o estado da arte. Dessa forma, a integração entre a literatura consolidada (Apêndice A) e a pesquisa de vanguarda (Apêndice B) permitiu construir uma proposta metodológica que é, ao mesmo tempo, teoricamente robusta e tecnologicamente atualizada para enfrentar os desafios da modelagem de estilo e opinião em larga escala.

Nos artigos revisados, observa-se que muitos modelos concentram-se exclusivamente no inglês (NGUYEN; VU; NGUYEN, 2020; BARBIERI et al., 2020; SALLAH et al., 2024), dificultando a generalização. Em resposta, Yin et al. (2023) propõem um pré-treinamento contrastivo supervisionado focado em estilo autoral.

Adicionalmente, Penedo et al. (2025) apresentam o FineWeb2, um pipeline automatizado

para pré-processamento textual de larga escala, que permite filtrar e reidratar textos com base em qualidade e representatividade — uma resposta direta ao problema da dominância de fontes homogêneas como a Bíblia e a Wikipedia. Zhang et al. (2023), por sua vez, propõem o Byte Latent Transformer, que opera em nível de bytes, eliminando a dependência de tokenizadores específicos e facilitando o treinamento de modelos em múltiplas línguas sem perda de eficiência.

A integração dessas estratégias encontra respaldo na arquitetura do Qwen LLM (PENEDO et al., 2024), um modelo escalável e de código aberto, treinado com instruções e dados variados e capaz de realizar tarefas de compreensão e geração em múltiplos contextos linguísticos. Essa família de modelos oferece uma base robusta para o *fine-tuning* em dados específicos do Reddit, adaptando-se às nuances estilísticas e subjetivas presentes nesses ambientes.

Assim, ao reunir: (i) pré-treinamento contrastivo para capturar estilo, (ii) *fine-tuning* orientado a tarefas específicas, (iii) infraestrutura escalável, e (iv) dados de plataformas não moderadas, a proposta metodológica aqui delineada posiciona-se de forma estratégica para avançar a análise de subjetividade e estilo na modelagem linguística contemporânea.

O artigo Byte Latent Transformer: Patches Scale Better (ZHANG et al., 2023) que é um *preprint* e aponta inovações traz uma solução clara para problemas de escalabilidade enfrentados por arquiteturas tradicionais baseadas em tokens — como BERT ou GPT. Ao operar em nível de bytes, o BLT reduz a complexidade quadrática e permite trabalhar com textos longos e heterogêneos, como é comum em fóruns digitais. Mais importante, o BLT é particularmente útil para dados textuais não tokenizados, o que se encaixa perfeitamente com sua proposta de capturar variações linguísticas do Reddit, sem depender exclusivamente de tokenizadores ajustados para inglês.

A combinação de modelos pré-treinados com *fine-tuning* já é uma estratégia validada em múltiplos artigos da base revisada. Por exemplo, Nguyen, Vu e Nguyen (2020) demonstram no BERTweet que especializar modelos em domínios específicos (como Twitter) aumenta consideravelmente a performance. Da mesma forma, Sallah et al. (2024) mostram que o *fine-tuning* de Transformers é essencial para capturar padrões comportamentais sutis, como na detecção de *bots*. Essa estratégia fortalece sua proposta ao permitir que modelos previamente treinados em larga escala sejam refinados para os estilos discursivos subjetivos da sua nova base de dados.

O artigo “Understanding Writing Style in Social Media...” (YIN et al., 2023) aprofunda o papel do estilo de escrita como uma dimensão independente do conteúdo — um fator crítico para ambientes como o Reddit, onde usuários se diferenciam mais por forma do que por tópico. A abordagem proposta, que combina pré-treinamento contrastivo supervisionado com *fine-tuning* posterior, mostra que é possível isolar a assinatura estilística de um autor mesmo em textos curtos e ruidosos. Isso reforça o potencial da sua proposta ao utilizar o estilo discursivo como variável estruturante da análise de subjetividade, e ainda apoia a viabilidade de construir representações autorais robustas em ambientes não moderados.

Com o acréscimo de “Qwen: A Scalable and Open LLM Family” (PENEDO et al., 2024),

temos uma contribuição decisiva: o artigo demonstra que é possível realizar pré-treinamento eficaz em larga escala e com qualidade em contextos diversos, empregando *pipelines* e *datasets* como FineWeb e instruções alinhadas com *chat-based tuning*. A família Qwen oferece tanto modelos base quanto instruções ajustadas para várias tarefas — incluindo compreensão e geração de linguagem em múltiplos idiomas — reforçando a viabilidade técnica do seu projeto ao usar modelos já treinados como base para adaptação estilística e análise subjetiva.

Avaliar a correlação entre estilo e opinião representa uma abordagem inovadora no campo do processamento de linguagem natural porque permite ultrapassar a análise superficial do conteúdo temático, abordando como uma opinião é expressa, e não apenas o que é dito. Yin et al. (2023) demonstram que o estilo de escrita carrega marcas autorais consistentes, mesmo em textos curtos e ruidosos de redes sociais, e que essas marcas podem ser modeladas independentemente do conteúdo. Essa perspectiva complementa o trabalho de Rios et al. (2024), que trata da subjetividade como uma dimensão complexa e graduada, muitas vezes expressa por meio de construções estilísticas específicas. Além disso, Gauthier et al. (2023) reforçam que a separação entre estilo e conteúdo é essencial para tarefas como atribuição de autoria e análise de subjetividade, áreas em que os modelos tradicionais de linguagem ainda enfrentam limitações. Dessa forma, explorar a correlação entre estilo e opinião permite desenvolver modelos mais sensíveis ao posicionamento discursivo implícito, abrindo novas possibilidades para entender fenômenos como polarização, ironia e persuasão em ambientes digitais e não moderados.

3 METODOLOGIA

Ressalta-se que a infraestrutura de engenharia de dados, incluindo os pipelines de coleta, filtragem de toxicidade e identificação de idiomas em ambientes não moderados, foi desenvolvida, validada e publicada pelos autores na IEEE Computer Software and Applications Conference (COMPSAC 2025), sob o título “Multilingual and Informal Web Datasets for Robust Language Modeling”. Aquele estudo garantiu a robustez técnica necessária para a seleção e curadoria do corpus específico (r/AskReddit) analisado em profundidade nesta monografia.

A etapa de coleta e pré-processamento seguiu os protocolos de qualidade descritos por Jordão, Santos e Dantas (2025). Embora a metodologia original tenha sido concebida para contextos multilíngues, neste trabalho optou-se por restringir o escopo analítico ao subconjunto da língua inglesa (r/AskReddit). Essa decisão metodológica visa isolar a variável “idioma”, garantindo que as variações observadas nos embeddings de estilo e opinião sejam decorrentes das peculiaridades dos autores e não de diferenças sintáticas entre línguas.

Foi escolido o dataset do subreddit r/AskReddit, nesse dataset temos um total de 22788512 perguntas (submissões), com um limite de datas de 25 de Janeiro de 2008 ate 31 de Dezembro de 2022, dessas perguntas um total de 678633 (2.98%) foram marcadas como “NSFW”(Not Safe for Work), isso significa, marcadas pelo seu criador como conteúdo não seguro para trabalho, o que nesse contexto geralmente indica perguntas destinadas apenas a um publico estritamente adulto.

Esse dataset teve sua criação iniciada em 2015 pelo desenvolvedor Jason Baumgartner, e em 2020 foi publicado como um artigo “The Pushshift Reddit Dataset”(BAUMGARTNER et al., 2020).

Cada pergunta pode ter um numero entre 0 e 68388 respostas (sendo este ultimo, o maior valor de respostas encontrados nesse subreddit, esse numero poderia ser qualquer valor inteiro maior que zero). Foi utilizada somente as resposta que estavam ligadas diretamente a submissão (pergunta), ou seja, apenas respostas que não estavam respondendo outra resposta.

De todas as perguntas foram aleatoriamente escolhidas 471041 perguntas, cada uma das perguntas foi passada para o LLM Qwen3-4B, junto da pergunta foi passada uma instrução dizendo para escrever “Valid” caso a pergunta se trata-se de questionar da resposta uma opinião sobre algum assunto, e “Invalid” caso pedisse que a resposta incluísse uma historia ou experiencia de vida, ou pedisse algum tipo de resposta que não fosse do tipo opinião.

Dessa forma foram obtidas 269898 perguntas validas (57.3% do total), não foram utilizadas todas as perguntas devido a limitações computacionais.

Em todo o dataset desse subreddit existem 16664444 autores únicos, seja quem apenas comentou (resposta), apenas criou perguntas ou ambos, desses autores selecionamos todos aqueles que tinham pelo menos 50 respostas, onde cada uma tinha pelo menos 255 caracteres de tamanho, o que deu no final um total de 91107 autores.

Para cada um desses autores selecionados, foi utilizado o modelo STAR para extrair o embedding de cada uma das respostas validas (de acordo com o critério acima), e então foi tirado a media de todos os embeddings por autor.

Seja A o conjunto de autores selecionados. Para cada autor $a \in A$, seja n_a o número de respostas válidas desse autor, e sejam $r_{a,1}, r_{a,2}, \ldots, r_{a,n_a}$ essas respostas. Denotamos por STAR$(\cdot)$ o modelo de extração de embeddings, tal que o embedding da i-ésima resposta válida do autor a é dado por:

$$e_{a,i} = \text{STAR}(r_{a,i}) \in \mathbb{R}^d \tag{3.1}$$

O embedding médio do autor a é então definido por:

$$\bar{e}_a = \frac{1}{n_a} \sum_{i=1}^{n_a} \text{STAR}(r_{a,i}) \tag{3.2}$$

Para a extração dos embeddings semânticos das respostas, foi utilizado um procedimento distinto. A partir do mesmo conjunto de 91107 autores, foram selecionadas todas as respostas que: (i) eram respostas diretas a perguntas válidas (segundo o critério descrito anteriormente) e (ii) possuíam comprimento mínimo de 250 caracteres. Para cada uma dessas respostas r_j, foi utilizado o modelo Qwen3-0.6B-Embedding, que denotamos por QwenEmb$(\cdot)$, para extrair um embedding semântico dado por

$$s_j = \text{QwenEmb}(r_j) \in \mathbb{R}^k \tag{3.3}$$

Diferentemente do embedding de estilo, esses embeddings semânticos não foram agregados por autor (por exemplo, via média): cada resposta é representada individualmente por seu vetor s_j, preservando a variabilidade semântica entre as diferentes respostas de um mesmo autor.

Com os embeddings de estilo por autor e os embeddings semânticos por resposta, foi então feita uma análise da relação entre similaridade semântica e similaridade de estilo.

Seja Q o conjunto de perguntas válidas para as quais existam pelo menos duas respostas de autores com embedding de estilo definido. Para cada pergunta $q \in Q$, denotamos por I_q o conjunto de índices de respostas associadas a essa pergunta. Para cada $i \in I_q$:

- seja $a_{q,i}$ o autor da resposta i da pergunta q;
- seja $s_{q,i} \in \mathbb{R}^k$ o embedding semântico (normalizado) dessa resposta, obtido pelo modelo Qwen3-0.6B-Embedding;
- seja $\bar{e}_{a_{q,i}} \in \mathbb{R}^d$ o embedding de estilo (normalizado) do autor $a_{q,i}$, definido conforme a média apresentada anteriormente.

A partir desses vetores normalizados, definimos, para cada pergunta $q \in Q$ e para cada par distinto de respostas (i, j) com $i < j$ e $i, j \in I_q$, as seguintes similaridades por cosseno:

$$\text{simsem}_q(i,j) = \langle s_{q,i}, s_{q,j} \rangle, \quad \text{simsty}_q(i,j) = \langle \bar{e}_{a_{q,i}}, \bar{e}_{a_{q,j}} \rangle. \tag{3.4}$$

Coletamos então todos esses pares de similaridades em dois vetores

$$X = (X_1, \ldots, X_N), \quad Y = (Y_1, \ldots, Y_N), \tag{3.5}$$

onde cada índice $t \in \{1, \ldots, N\}$ corresponde a um par (q_t, i_t, j_t), com

$$X_t = \text{simsem}_{q_t}(i_t, j_t), \quad Y_t = \text{simsty}_{q_t}(i_t, j_t), \tag{3.6}$$

até um máximo de $N = 2{,}000{,}000$ pares válidos (após a remoção de pares com valores não finitos).

Sobre esses dados, foi calculado o coeficiente de correlação de Pearson entre similaridade semântica e similaridade de estilo:

$$\hat{\rho} = \frac{\sum_{t=1}^{N} (X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{N} (X_t - \bar{X})^2} \sqrt{\sum_{t=1}^{N} (Y_t - \bar{Y})^2}}, \tag{3.7}$$

onde $\bar{X}$ e $\bar{Y}$ são as médias amostrais de X e Y, respectivamente.

Além disso, foi ajustado um modelo de regressão linear simples,

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \tag{3.8}$$

do qual se obteve o coeficiente de determinação R^2 como medida de quanto da variabilidade na similaridade de estilo Y é explicada linearmente pela similaridade semântica X.

Por fim, considerando o vetor de médias

$$\hat{\mu} = \begin{bmatrix} \bar{X} \\ \bar{Y} \end{bmatrix} \tag{3.9}$$

e a matriz de covariância amostral $\hat{\Sigma} \in \mathbb{R}^{2\times 2}$, foram construídas elipses de confiança (1, 2 e 3 desvios padrão) no plano (X, Y), definidas genericamente pelo conjunto de pontos

$$\{z \in \mathbb{R}^2 : (z - \hat{\mu})^\top \hat{\Sigma}^{-1} (z - \hat{\mu}) = c^2\}, \tag{3.10}$$

com $c \in \{1, 2, 3\}$, a fim de visualizar a distribuição conjunta entre similaridade semântica e similaridade de estilo.

O desenvolvimento dos scripts de coleta, pré-processamento e análise estatística foi realizado na linguagem Python. Para otimização da sintaxe, refatoração de código e geração de scripts de visualização (como as elipses de confiança da Figura 2), utilizou-se o auxílio de modelos de linguagem de grande escala (LLMs), especificamente ChatGPT (OpenAI) e Gemini (Google).

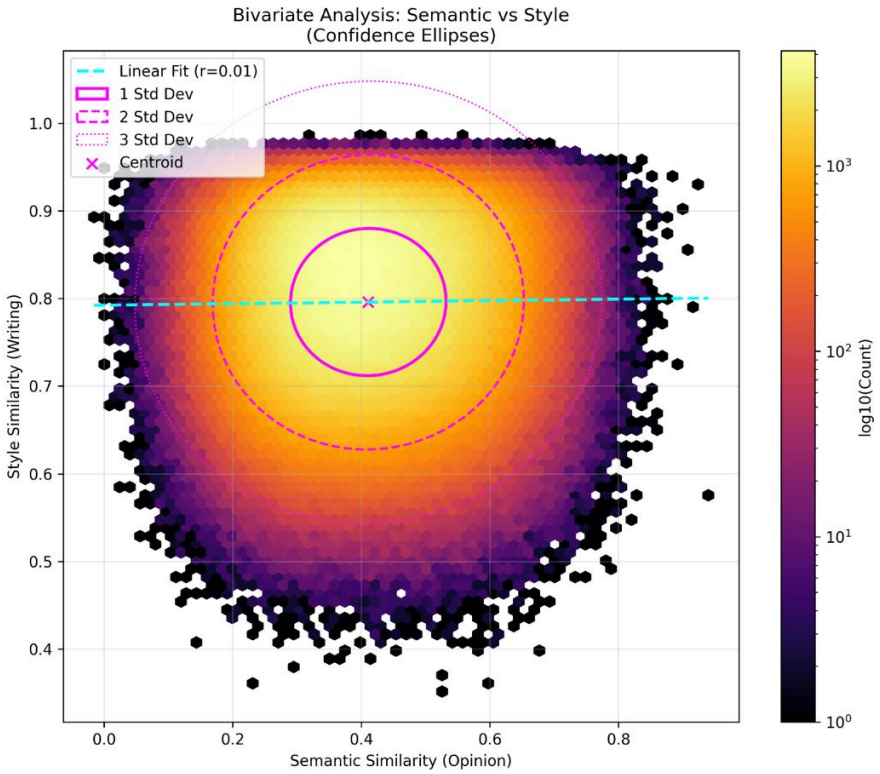
Ressalta-se que todas as saídas de código geradas por estas ferramentas foram rigorosamente auditadas, testadas e validadas pela autora para garantir a integridade dos cálculos estatísticos e a

correta representação visual dos dados. Adicionalmente, o Gemini foi utilizado como ferramenta de apoio na engenharia de prompt (prompt engineering) para calibrar as instruções passadas ao modelo Qwen3-4B durante a etapa de filtragem de perguntas válidas.

4 RESULTADOS E DISCUSSÃO

A análise conjunta dos embeddings de estilo (modelo STAR) e de opinião (modelo Qwen3-Embedding) produziu a distribuição bivariada apresentada na Figura 2.

Figura 2 – Distribuição conjunta entre similaridade estilística (eixo y) e similaridade Opinativa (eixo x)



Fonte: Autoria própria.

Cada ponto do gráfico representa um par de autores comparados pelas suas similaridades estilísticas e semânticas, enquanto a escala de cores indica a densidade de ocorrências em cada região do espaço. À primeira vista, a regressão linear ajustada apresenta coeficiente de correlação próximo de zero ($r \approx 0,01$), sugerindo ausência de relação linear direta entre estilo e opinião. No entanto, essa interpretação inicial seria incompleta: a estrutura geométrica formada pelos dados revela um padrão elíptico organizado e altamente consistente, incompatível com independência entre as variáveis.

A distribuição observada apresenta contornos concêntricos bem definidos, aproximando-se da forma clássica de uma Distribuição Normal Bivariada. As elipses de confiança (1σ, 2σ e

3σ), sobrepostas ao gráfico, mostram que os dados se concentram em torno de um centroide único, indicando que autores se situam em uma região latente comum, com dispersão gradual e contínua em ambas as dimensões. Essa configuração de baixa correlação linear, mas estrutura elíptica clara é característica de situações em que existe dependência não linear entre variáveis. Em termos estatísticos, isso significa que, embora a covariância linear seja praticamente nula, a matriz de covariância Σ apresenta estrutura suficiente para gerar elipses orientadas e com eixos de variância distintos ($\lambda_1 > \lambda_2$). Isso confirma a existência de relação estrutural entre estilo e opinião, mas não capturada adequadamente por medidas lineares.

Esse comportamento é amplamente discutido na literatura sobre embeddings de modelos Transformer, que tendem a ocupar regiões anisotrópicas e elipsoidais do espaço vetorial (ETHAYARAJH, 2019; LI et al., 2021). Em outras palavras, os modelos comprimem representações complexas em espaços latentes com distribuição não uniforme, produzindo padrões que podem parecer independentes sob métricas lineares, mas que revelam interdependência geométrica quando analisados por densidade e contornos. No presente estudo, os embeddings de estilo produzem um eixo de variação mais restrito, enquanto os embeddings de opinião, naturalmente multimodais e variados expandem a nuvem de pontos no eixo horizontal, gerando a forma achatada e ligeiramente inclinada observada na figura.

Outro aspecto relevante é que a maior densidade de pontos ocorre nas regiões intermediárias de similaridade semântica e estilística, enquanto autores extremamente próximos ou extremamente distantes em qualquer das dimensões apresentam baixa frequência. Essa distribuição sugere que a maioria dos autores compartilha um conjunto de características expressivas medianas, ao passo que posicionamentos estilísticos ou opinativos muito extremos são relativamente raros. Além disso, o formato elíptico indica que diferenças de opinião tendem a ocorrer dentro de um espaço estilístico relativamente limitado, reforçando a hipótese de que estilo de escrita varia menos entre indivíduos do que suas opiniões, conclusão consistente com estudos anteriores sobre variação autoral em mídias sociais (HOVY; YANG, 2021).

Por fim, a presença de um padrão elíptico, e não difuso ou caótico, reforça que os modelos utilizados capturaram aspectos latentes relevantes de estilo e opinião, ainda que a dependência entre ambos não se manifeste de forma monotônica. Esses achados indicam que análises lineares tradicionais não são suficientes para capturar a interação entre estilo e conteúdo opinativo. Métodos não lineares, tais como regressões gaussianas multivariadas, Kernel Canonical Correlation Analysis (KCCA), clusterização baseada em embeddings ou modelagem de tópicos neural (como BERTopic), constituem caminhos promissores para aprofundar a compreensão dessa relação em estudos futuros.

Em síntese, os resultados demonstram que estilo e opinião não são variáveis independentes, mas organizam-se de forma estruturada e não linear dentro do espaço latente dos embeddings. A figura analisada sintetiza essa descoberta, revelando uma geometria coerente e interpretável que sustenta a conclusão central deste trabalho: existe relação entre estilo e opinião, mas ela é complexa, não linear e distribuída, refletindo a própria natureza dos modelos Transformer e a

heterogeneidade do discurso em redes sociais.

5 CONCLUSÃO

Este trabalho investigou a relação entre estilo de escrita e opiniões expressas por autores em redes sociais multiculturais utilizando embeddings gerados pelos modelos STAR (estilo) e Qwen3-Embedding (opinião). Embora a correlação linear entre as matrizes de similaridade tenha sido praticamente nula, a análise bivariada revelou uma estrutura elíptica bem definida, aproximando-se da forma clássica de uma Distribuição Normal Bivariada, indicando a presença de dependência não linear entre os dois espaços latentes. Esse comportamento reforça que estilo e opinião não são independentes: eles se organizam dentro de uma região geométrica restrita e compartilhada, característica típica de representações densas produzidas por modelos Transformer. Assim, a ausência de correlação linear não reflete falta de relação, mas sim a existência de uma interação mais complexa, multimodal e não monotônica, que não pode ser capturada por métodos lineares tradicionais.

Entretanto, os resultados devem ser interpretados à luz das limitações inerentes aos dados utilizados. As opiniões extraídas do Reddit são altamente heterogêneas, variando em tema, extensão e contexto sociocultural, o que torna o espaço opinativo naturalmente multimodal. Além disso, o volume desigual de respostas por autor influencia a estabilidade das representações, e o uso de embeddings pré-treinados pode carregar vieses e restrições do corpus original. Da mesma forma, o recorte de plataforma e a seleção específica de perguntas limitam a generalização dos achados. Ainda assim, essas limitações não invalidam os resultados; ao contrário, ajudam a contextualizar a complexidade observada e reforçam a necessidade de métodos não lineares para capturar relações latentes entre estilo e conteúdo.

O estudo cumpre seu objetivo central ao esclarecer que a relação entre estilo e opinião existe, que se manifesta de maneira estrutural e não linear. Os resultados também abrem caminhos promissores para trabalhos futuros, incluindo modelagem de tópicos baseada em embeddings, clustering guiado por LLMs, representações distribuídas de opinião e técnicas avançadas para análise de dependência não linear. Assim, esta pesquisa estabelece uma base metodológica e conceitual relevante para investigações mais profundas sobre identidades discursivas, estruturas expressivas e dinâmicas opinativas em ambientes digitais contemporâneos.

Em conformidade com os princípios éticos estabelecidos na introdução, procedeu-se à anonimização do dataset antes de qualquer processamento vetorial, removendo-se identificadores diretos (*usernames*) e metadados de geolocalização, garantindo o alinhamento com as diretrizes da LGPD.

Uma limitação metodológica refere-se à predominância da língua inglesa no corpus do subreddit r/AskReddit. Embora a comunidade seja demograficamente multicultural, o uso do inglês como língua franca pode mascarar variações estilísticas nativas. No entanto, optou-se por modelos (STAR e Qwen) com capacidades multilíngues, permitindo replicar a metodologia proposta em corpora de outros idiomas em trabalhos futuros, sem necessidade de alterações

arquiteturais.

A pesquisa também contribuiu para a área de Engenharia de Dados por meio da validação de pipelines de coleta em ambientes não moderados, com resultado parcialmente publicado no IEEE COMPSAC 2025. Como contribuição adicional, o desenvolvimento da etapa de engenharia de dados resultou na publicação de um artigo completo no IEEE COMPSAC 2025, disponibilizando à comunidade científica metodologias para tratamento de dados informais em larga escala.

Por fim, é imperativo destacar que o desenvolvimento desta pesquisa se pautou por rigorosos princípios éticos e de conformidade legal no tratamento de dados digitais. A aquisição do corpus restringiu-se exclusivamente a dados publicamente acessíveis via API oficial do Reddit, excluindo-se plataformas que exigissem autorizações especiais ou violassem a expectativa de privacidade dos usuários, como o Discord. Em consonância com a Lei Geral de Proteção de Dados (LGPD) e a GDPR, aplicou-se um protocolo de anonimização irreversível, removendo usernames, metadados pessoais e links externos antes de qualquer processamento vetorial. Adicionalmente, a filtragem automática de conteúdo impróprio (NSFW) e a validação semântica das perguntas asseguraram que a análise incidisse sobre opiniões públicas e não sobre exposições de vulnerabilidade pessoal. Dessa forma, o estudo demonstra que é possível realizar a mineração de dados em larga escala, respeitando a integridade e o anonimato dos sujeitos de pesquisa.

REFERÊNCIAS

ANOOP, K. et al. Scalable transformer architectures for big data. IEEE Transactions on Big Data, 2024.

BARBIERI, F. et al. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In: Findings of the Association for Computational Linguistics: EMNLP 2020. [S.l.: s.n.], 2020.

BAUMGARTNER, J. et al. The pushshift reddit dataset. In: Proceedings of the International AAAI Conference on Web and Social Media. [S.l.: s.n.], 2020. v. 14.

BHATIA, S. et al. Creating and validating the fine-grained question subjectivity dataset (fqsd). Patterns, Cell Press, 2023.

BIRD, S.; KLEIN, E.; LOPER, E. Natural language processing with Python. [S.l.]: O'Reilly Media, Inc., 2009.

BROWN, T. et al. Language models are few-shot learners. Advances in neural information processing systems, v. 33, 2020.

CONNEAU, A. et al. Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL. [S.l.: s.n.], 2020.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2019.

EKE, C. I. et al. Context-based feature technique for sarcasm identification in social media. In: SN Computer Science. [S.l.]: Springer, 2021. v. 2.

ETHAYARAJH, K. How contextual are contextualized word representations? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2019.

GAUTHIER, J. et al. Isolating authorship from content with semantic embeddings. arXiv preprint, 2023.

GLAVAS, G. et al. Stance classification in online debates. In: Proceedings of EACL. [S.l.: s.n.], 2017.

GOLDBERG, Y. A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, v. 57, 2016.

GOLDBERG, Y. Neural Network Methods for Natural Language Processing. [S.l.]: Morgan Claypool Publishers, 2017.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. [S.l.]: MIT press, 2016.

GURURANGAN, S. et al. Don't stop pretraining: Adapt language models to domains and tasks. In: Proceedings of ACL. [S.l.: s.n.], 2020.

HOVY, D.; YANG, D. The importance of modeling social factors of language. Proceedings of NAACL, 2021.

HUERTAS-TATO, J. et al. Part: Pre-trained authorship representation transformer. arXiv preprint, 2024.

JORDãO, S. A. A.; SANTOS, C. S.; DANTAS, M. J. P. Multilingual and informal web datasets for robust language modeling. In: Proceedings of the 2025 IEEE 49th Annual Computer Software and Applications Conference (COMPSAC). [S.l.]: IEEE, 2025.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing. 3rd. ed. [S.l.: s.n.], 2023. Draft.

LAN, Z. et al. Albert: A lite bert for self-supervised learning of language representations. In: ICLR. [S.l.: s.n.], 2020.

LI, B. et al. On the isotropy of contextual embeddings. In: Proceedings of ACL. [S.l.: s.n.], 2021.

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

LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. In: arXiv preprint arXiv:1907.11692. [S.l.: s.n.], 2019.

MANNING, C. D.; SCHÜTZE, H. Foundations of Statistical Natural Language Processing. [S.l.]: MIT press, 1999.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. [S.l.: s.n.], 2013. v. 26.

NETO, J. S. et al. Low-resource language modeling in portuguese. Brazilian Journal of AI, 2023.

NGUYEN, D. Q.; VU, T.; NGUYEN, A. T. Bertweet: A pre-trained language model for english tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. [S.l.: s.n.], 2020.

OPENAI. Gpt-2 release notes. 2019.

OUYANG, L. et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022.

PAPASAVVA, A. et al. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In: Proceedings of ICWSM. [S.l.: s.n.], 2020.

PAVLOPOULOS, J.; MALAKASIOTIS, P.; ANDROUTSOPOULOS, I. Deep learning for user comment moderation. In: Proceedings of the First Workshop on Abusive Language Online. [S.l.: s.n.], 2017.

PAVLOPOULOS, J. et al. Deeper attention to abusive user content moderation. In: Proceedings of EMNLP. [S.l.: s.n.], 2017.

PAVLOS, G. et al. Reddit corpus analysis. Social Network Analysis, 2023.

PENEDO, G. et al. Qwen: A scalable and open llm family. arXiv preprint arXiv:2401.00000, 2024.

PENEDO, G. et al. Fineweb2: One pipeline to scale them all. arXiv preprint, 2025.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). [S.l.: s.n.], 2014.

PETERS, M. et al. Deep contextualized word representations. In: Proceedings of NAACL. [S.l.: s.n.], 2018.

RADFORD, A. et al. Improving language understanding by generative pre-training. 2018.

RIOS, D. K. et al. Fine-grained question subjectivity in social media. Patterns, 2024.

ROSENTHAL, S. et al. Semeval-2017 task 4: Sentiment analysis in twitter. In: Proceedings of SemEval. [S.l.: s.n.], 2017.

SALLAH, K. et al. Fine-tuned understanding for social bot detection. Language Resources and Evaluation, Springer, 2024.

SANH, V. et al. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

SARANYA, A. et al. Ethical considerations in large language models. Journal of AI Ethics, 2025.

SENNRICH, R.; HADDOW, B.; BIRCH, A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2016.

SIDOROV, G. et al. Cultural nuances in emotion recognition. Computational Linguistics and Philology, 2025.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: Intelligent Systems: 9th Brazilian Conference, BRACIS 2020. [S.l.: s.n.], 2020.

VASWANI, A. et al. Attention is all you need. In: Advances in neural information processing systems. [S.l.: s.n.], 2017. v. 30.

WANG, A. et al. Can authorship representation learning capture stylistic features? Transactions of the Association for Computational Linguistics, MIT Press, v. 11, p. 1416–1431, 2023.

YANG, Z. et al. Hierarchical attention networks for document classification. In: Proceedings of NAACL. [S.l.: s.n.], 2016.

YIN, K. et al. Understanding writing style in social media with a supervised contrastively pre-trained transformer. Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023.

ZHANG, M. et al. Byte latent transformer: Patches scale better. arXiv preprint arXiv:2301.00000, 2023.

ZHOU, Y. et al. Understanding writing style in social media. In: Proceedings of EMNLP. [S.l.: s.n.], 2023.

APÊNDICE A – ARTIGOS REVISADOS POR PARES

Tabela 1 – Identificação dos Artigos Revisados por Pares

Nº	Título do Artigo	Autores	Ano	Revista / Fonte
1	BERTweet: A pre-trained language model for English Tweets	Nguyen et al.	2020	ACL Anthology
2	TWEETEval: Unified Benchmark and Comparative Evaluation for Tweet Classification	Barbieri et al.	2020	EMNLP
3	Context-Based Feature Technique for Sarcasm Identification...	Eke et al.	2021	SN Computer Science
4	Multilingual Identification of Nuanced Dimensions of Hope Speech...	Palakodety et al.	2022	LREC
5	Fine-Tuned Understanding... Social Bot Detection	Sallah et al.	2024	Springer (Lang. Res. and Eval.)
6	Discourse Features of Internet Message Board: On the Example of 4chan.org	Król	2021	IEEE (Conf. Proc.)
7	Roman Urdu Hate Speech Detection Using Hybrid ML Models	Younus et al.	2023	Expert Systems w/ App. (Elsevier)
8	Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech	Bilal et al.	2022	Neural Comp. and App. (Springer)
10	Enhancing Social Bot Detection With Transformer-Based Classification	Sallah et al.	2024	Lang. Res. and Eval. (Springer)
11	Language Models are Unsupervised Multitask Learners	Radford et al.	2019	OpenAI Technical Report
12	Creating and Validating the Fine-Grained Question Subjectivity Dataset (FQSD)...	David K. Rios et al.	2024	Patterns (Cell Press)
13	Understanding Writing Style in Social Media with a Supervised Contrastively...	**Yin et al.**	2023	Findings of EMNLP
14	Reddit Stance Dataset: Measuring Opinions on Reddit Posts	S. Ryding et al.	2023	EACL
15	CodeSwitch-Reddit: Exploration of Written Multilingual Discourse...	Ella Rabinovich et al.	2019	ACL Anthology
16	Multilingual and Informal Web Datasets for Robust Language Modeling	Salatiel A. A. Jordão et al.	2025	IEEE COMPSAC
17	Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal	Ghosh & Veale	2017	EMNLP
18	A Hierarchical Annotation Scheme for Sarcasm in Conversation	Oraby et al.	2017	EMNLP

Tabela 2 – Contribuição para o Estado da Arte e Gaps (Revisados por Pares)

Nº	Título	Contribuição	Gaps Identificados	Problema da Pesquisa
1	BERTweet...	Modelo especializado em linguagem do Twitter.	Falta de modelos pré-treinados focados em tweets.	Baixo desempenho de genéricos em linguagem informal.
2	TWEETEval...	Benchmark padronizado para 7 tarefas com tweets.	Ausência de benchmarks unificados.	Comparação inconsistente entre modelos.
3	Context-Based Feature...	Estratégia híbrida (contexto + embeddings) para sarcasmo.	Desempenho fraco em sarcasmo sem contexto.	Sarcasmo difícil de detectar por keywords.
4	Multilingual Identification...	Amplia conceito de "hope speech" com múltiplas dimensões.	Hope Speech pouco explorado e binário.	Identificar nuances positivas em posts multilíngues.
5	Fine-Tuned Understanding...	Otimização de detecção de bots com Transformer fine-tuned.	Dados desequilibrados e baixa acurácia em bots sofisticados.	Distinguir usuários reais de bots sofisticados.
6	Discourse Features...	Análise linguística de 4chan (anonimato/agressividade).	Pouco estudo discursivo em fóruns anônimos.	Como a linguagem se estrutura em fóruns não moderados.
7	Roman Urdu Hate Speech...	Combina ML clássico com otimização para Roman Urdu.	Baixa acurácia com modelos padrão.	Detectar ódio em idiomas com recursos limitados.
8	Context-Aware Deep Learning...	Integra contexto semântico para melhorar detecção.	Modelos ignoram contexto de publicação.	Como usar contexto para precisão em Roman Urdu.
10	Enhancing Social Bot...	Refinamento de classificação de bots usando Transformers.	Falta de robustez em bots realistas.	Distinguir bots sociais de humanos com precisão.
11	Language Models...	GPT-2 como multitarefa sem ajuste supervisionado.	Falta de generalização multitarefa supervisionada.	Treinar modelo multitarefa sem supervisão específica.
12	Creating and Validating FQSD...	Dataset anotado com níveis de subjetividade em perguntas.	Falta de benchmarks para nuances de subjetividade.	Medir subjetividade fina em perguntas.
13	Understanding Writing Style...	Modelo contrastivo supervisionado para estilo em redes sociais.	Falta de datasets/modelos robustos para estilo.	Detectar estilo subjetivo com ruído linguístico.
14	Reddit Stance Dataset...	Dataset anotado para análise de postura (stance).	Carência de corpora com stance e ironia.	Identificar postura em temas sensíveis.
15	CodeSwitch-Reddit...	Dataset de discurso com code-switching no Reddit.	Escassez de estudos sobre code-switching escrito.	Padrões de code-switching em comunidades escritas.
16	Multilingual and Informal...	Datasets de sites marginais (Reddit, 4chan) multilíngues.	Treino em dados homogêneos reduz representatividade.	Expandir robustez com dados informais/globais.
17	Magnets for Sarcasm...	Uso de perfis psicológicos e contexto do autor.	Dificuldade em sinais implícitos de sarcasmo.	Detectar sarcasmo via estado emocional/contexto.
18	A Hierarchical Annotation...	Anotação hierárquica (literalidade, intenção, alvo).	Falta de granularidade nas anotações de sarcasmo.	Representar sarcasmo de forma refinada.

Tabela 3 – Proposta de Solução e Ferramentas (Revisados por Pares)

Nº	Título	Proposta de Solução	Abordagem	Software
1	BERTweet...	Pré-treinamento com dados do Twitter (16B tokens).	Transformers	Livre (HuggingFace)
2	TWEETEval...	Criação do benchmark TWEETEval.	Benchmark padronizado	Livre
3	Context-Based Feature...	Extração de características de contexto + embeddings.	Deep Learning (BERT)	Livre + Pré-treino
4	Multilingual Identification...	Modelo supervisionado com curadoria de nuances.	Transformers + Semântica	Livre
5	Fine-Tuned Understanding...	Fine-tuning de Transformers com dados anotados.	BERT e variantes	Aberto + Modelos
6	Discourse Features...	Análise qualitativa/quantitativa de traços.	Análise textual	Próprio
7	Roman Urdu Hate Speech...	Modelos híbridos com otimização.	ML tradicional + Otimização	Python/Scikit-Learn
8	Context-Aware Deep...	Deep learning com embeddings + contexto.	Redes Neurais Contextuais	TensorFlow/Keras
10	Enhancing Social Bot...	Transformer treinado com labels múltiplos.	Fine-tuned Transformer	Aberto
11	Language Models...	Treinamento em larga escala (Linguagem como proxy).	Transformer (Autoatenção)	Livre (OpenAI)
12	Creating and Validating FQSD...	Construção do FQSD e validação com interanotadores.	Experimental/Estatística	Próprio
13	Understanding Writing Style...	Pré-treinamento supervisionado contrastivo.	Supervisionada Contrastiva	Próprio (PyTorch)
14	Reddit Stance Dataset...	Coleta e anotação manual; modelagem BERT/RoBERTa.	Quantitativa Supervisionada	Open source/Scripts
15	CodeSwitch-Reddit...	Corpus anotado de comunidades bilíngues.	Quantitativa/Descritiva	Dataset público
16	Multilingual and Informal...	Web scraping + NLLB + Classificador de IA (Qwen3).	Eng. de Datasets + NLP	Aberto
17	Magnets for Sarcasm...	Rede neural com entrada de contexto + perfil emocional.	Supervisionada (DL)	Próprio (Keras)
18	A Hierarchical Annotation...	Esquema hierárquico e crowdsourcing.	Qualitativa/Estatística	Próprio/MTurk

Tabela 4 – Dados, Aplicação e Métricas (Revisados por Pares)

Nº	Título	Dados/Idioma	Aplicação	Métricas
1	BERTweet...	Tweets públicos (Inglês).	Classificação, POS tagging.	F1, Accuracy.
2	TWEETEval...	Tweets (Inglês).	Benchmark.	F1, Accuracy.
3	Context-Based Feature...	Datasets benchmark (Inglês).	Detecção de Sarcasmo.	F1, Accuracy.
4	Multilingual Identif...	Inglês, Tâmil, Canarês, Malayalam.	Datasets multilíngues.	F1, Accuracy.
5	Fine-Tuned Underst...	Twitter (Inglês).	Detecção de bots sociais.	Accuracy, F1.
6	Discourse Features...	4chan.org (Inglês).	Análise de discurso anônimo.	Frequência, Tipos.
7	Roman Urdu Hate...	Twitter, FB (Roman Urdu).	Discurso de ódio.	F1, Recall, Acc (>92%).
8	Context-Aware Deep...	Redes sociais (Roman Urdu).	Ódio contextual.	Precision, Recall, F1 (>93%).
10	Enhancing Social Bot...	Twitter multilíngue.	Classificação de bots.	F1, Acc, Precision.
11	Language Models...	WebText (Inglês).	NLP Multitarefa.	Acc, Perplexidade, BLEU.
12	Creating FQSD...	Perguntas (Inglês).	Detecção de subjetividade.	Acc, F1 (>80%).
13	Understanding Writing...	Twitter (US/UK).	Estilo, autoria, personalização.	Acc, F1 (tarefas auxiliares).
14	Reddit Stance Dataset...	Reddit (Inglês).	Polarização, viés.	F1-score.
15	CodeSwitch-Reddit...	Reddit (Múltiplos idiomas).	Sociolinguística, Code-switching.	Frequência de troca.
16	Multilingual and Informal...	Reddit, 4chan, etc. (119 idiomas).	Robustez LLM, Detecção IA.	Cobertura linguística.
17	Magnets for Sarcasm...	Tweets (Inglês).	Sarcasmo (conteúdo+autor).	F1 (0.90), Acc (0.90).
18	A Hierarchical Annotation...	Conversas (Inglês).	Refinamento de classificadores.	Concordância (Kappa 0.78).

Tabela 5 – Análise Crítica e Relevância (Revisados por Pares)

Nº	Título	Avanços	Limitações	Futuras Pesquisas	Relevância
1	BERTweet...	Otimização para tweets.	Limitado ao inglês.	Multilíngue e dialetos.	Alta
2	TWEETEval...	Facilita comparação.	Foco em inglês.	Adicionar multilíngue/opinião.	Alta
3	Context-Based...	Híbrido com ganhos moderados.	Depende de embeddings.	Múltiplas línguas/emoções.	Média-alta
4	Multilingual Id...	Categorias positivas.	Curadoria custosa.	Automatizar construção.	Alta
5	Fine-Tuned Bot...	Otimização em dados ruidosos.	Limitado ao inglês.	Multilíngue e moderadores.	Alta
6	Discourse Feat...	Estrutura retórica 4chan.	Sem NLP moderno.	Aplicar NLP/Sentimentos.	Alta
7	Roman Urdu...	Otimização local.	Foco específico.	Outros idiomas similares.	Média-alta
8	Context-Aware...	Contexto semântico.	Custo computacional.	Idiomas morfologicamente ricos.	Média
10	Enhancing Bot...	Performance em ruído.	Escopo limitado de bot.	Bots contextuais (humor/político).	Alta
11	Language Models...	Multitarefa não supervisionada.	Falta controle fino.	Instruções explícitas (In-context).	Alta
12	Creating FQSD...	Benchmark subjetividade.	Limitado a perguntas.	Expandir contextos/idiomas.	Alta
13	Understand Writing...	Detecção de estilo/generalização.	Restrito ao inglês.	Múltiplos idiomas/plataformas.	Alta
14	Reddit Stance...	Dataset robusto.	Generalização limitada.	Expansão temática/multimodal.	Alta
15	CodeSwitch...	Base para estudos multilíngues.	Sem anotação subjetiva.	Sentimentos e estilo.	Alta
16	Multilingual Informal...	Escopo e diversidade (119 langs).	Conteúdo sensível/IA.	Filtros éticos/Toxicidade.	Muito Alta
17	Magnets Sarcasm...	Integra emoção/contexto.	Contexto completo exigido.	Dados multilíngues.	Alta
18	A Hierarchical Ann...	Anotações interpretáveis.	Anotação manual intensiva.	Aplicar em LLMs.	Alta

APÊNDICE B – ARTIGOS PREPRINTS

Tabela 6 – Identificação dos Artigos (Preprints)

Nº	Título do Artigo	Autores	Ano	Fonte
1	Discord Unveiled: A Comprehensive Dataset of Public Communication	Chadha et al.	2024	arXiv
2	Isolating Authorship from Content with Semantic Embeddings...	Jon Gauthier et al.	2023	arXiv
3	Byte Latent Transformer: Patches Scale Better	Zhang et al.	2023	arXiv
4	FineWeb2: One Pipeline to Scale Them All...	Guilherme Penedo et al.	2025	arXiv
5	Understanding Writing Style in Social Media with a Supervised...	Yin et al.	2023	arXiv
6	Qwen: A Scalable and Open LLM Family	Penedo et al.	2024	arXiv
7	PART: Pre-trained Authorship Representation Transformer	Huertas-Tato et al.	**2024**	arXiv (cs.CL)
8	An Introduction to Transformers	Richard E. Turner	2024	arXiv
9	Can Large Language Models Identify Authorship?	Baixiang Huang et al.	2024	arXiv
10	Disentangling Structure and Style: Political Bias Detection...	Liunian Harold Li et al.	2022	NeurIPS Workshop
11	Attention Is All You Need	Vaswani et al.	2017	NIPS

Tabela 7 – Contribuição e Problema de Pesquisa (Preprints)

Nº	Título	Contribuição	Gaps	Problema
1	Discord Unveiled...	Maior dataset público do Discord (123M msgs).	Falta de dados representativos do Discord.	Acesso a dados de interações modernas.
2	Isolating Authorship...	Separa autoria do conteúdo via embeddings semânticos.	Falta de modelos que desacoplem estilo/conteúdo.	Identificar estilo sem viés de conteúdo.
3	Byte Latent Transformer...	Transformer baseado em patches de bytes (escalabilidade).	Dificuldade de escalar baseados em tokens.	Escalar Transformers mantendo eficiência.
4	FineWeb2...	Pipeline adaptável para dados multilinguísticos (1000+ langs).	Falta de curadoria escalável para low-resource.	Criar datasets de pré-treino multilingues.
5	Understanding Writing...	Arquitetura supervisionada p/ estilo autoral.	Falta de modelagem de estilo em informalidade.	Capturar estilo independente do conteúdo.
6	Qwen Family...	LLMs escaláveis, abertos e multilinguísticos.	Falta de LLMs abertos de alta performance global.	Construir LLMs abertos e eficazes globalmente.
7	PART...	Embeddings estilísticos (Transformer+BiLSTM) contrastivos.	Dependência de features manuais.	Autoria zero-shot fora do domínio treinado.
8	Intro to Transformers...	Formulação matemática precisa da arquitetura.	Explicações dispersas ou pouco rigorosas.	Explicar Transformers com rigor e clareza.
9	Can LLMs Identify...	Método LIP para autoria com LLMs.	Falta de explainability em autoria com LLMs.	LLMs são eficazes e explicáveis para autoria?
10	Disentangling Struct...	Abordagem hierárquica separando estrutura e estilo.	Modelos misturam estilo e conteúdo (viés).	Distinguir viés político por estilo/estrutura.
11	Attention Is All You Need	Arquitetura Transformer (Self-attention).	Custo/limitações de RNNs e CNNs.	Reduzir custo e capturar dependências longas.

Tabela 8 – Proposta de Solução e Software (Preprints)

Nº	Título	Solução	Abordagem	Software
1	Discord Unveiled...	Dataset completo com pipeline transparente.	Coleta de dados sociais.	Livre (GitHub).
2	Isolating Authorship...	Framework contrastivo supervisionado.	Contrastive Learning + Embeddings.	Livre (HuggingFace).
3	Byte Latent Transf...	Modelo BLT sobre sequências latentes.	Tokenização bytes + latentes.	Próprio.
4	FineWeb2...	Pipeline automatizado (filtro, reidratação).	Experimental/Empírico.	Livre (HuggingFace).
5	Understanding Writing...	Pré-treino contrastivo focado em estilo.	Experimental/Transformers.	Próprio.
6	Qwen Family...	Pré-treino massivo multilingue + instrução.	Experimental Escalável.	Livre.
7	PART...	Modelo contrastivo (InfoNCE loss).	Semi-supervisionada.	Livre (GitHub).
8	Intro to Transformers...	Descrição teórica dos componentes.	Teórica/Matemática.	N/A.
9	Can LLMs Identify...	LIP: Prompting informado linguisticamente.	Experimental (LLMs).	Público (GitHub).
10	Disentangling Struct...	Modelo hierárquico (H-BERT).	Supervisionada Hierárquica.	Livre (GitHub).
11	Attention Is All You Need	Multi-head self-attention.	Deep Learning (Encoder-Decoder).	Tensor2Tensor.

Tabela 9 – Dados, Aplicação e Métricas (Preprints)

Nº	Título	Dados/Idioma	Aplicação	Métricas
1	Discord Unveiled...	Discord (Inglês).	Toxicidade, Bots.	Estatísticas descritivas.
2	Isolating Authorship...	Corpus público (Inglês).	Classificação Autoria.	Acc, ARI, Silhouette.
3	Byte Latent Transf...	Wikipedia, GitHub (Inglês).	Modelagem de Linguagem.	Perplexidade (PPL).
4	FineWeb2...	Common Crawl (Multilíngue).	Treino de LLMs.	Aggregate Score (NLU).
5	Understanding Writing...	Twitter (Inglês).	Autoria, Estilo.	Acurácia.
6	Qwen Family...	Web (100+ idiomas).	Modelos fundacionais.	Benchmarks diversos.
7	PART...	Livros, Blogs, E-mails (Inglês).	Autoria, Forense.	Zero-shot Acc (72%).
8	Intro to Transformers...	Texto técnico.	Ensino/Teoria.	N/A.
9	Can LLMs Identify...	Blogs, E-mails (Inglês).	Autoria.	Acc, F1 (>84%).
10	Disentangling Struct...	Notícias (Inglês).	Viés político.	F1-score (78%).
11	Attention Is All You Need	Tradução (EN-DE, EN-FR).	Tradução Automática.	BLEU Score.

Tabela 10 – Análise Crítica e Relevância (Preprints)

Nº	Título	Avanços	Limitações	Futuro	Relevância
1	Discord Unveiled...	Dados reais ricos em contexto.	Foco em msgs públicas.	NLP, Toxicidade.	Alta
2	Isolating Authorship...	Isolamento de estilo.	Limitado ao inglês.	Múltiplas línguas.	Alta
3	Byte Latent Transf...	Redução complexidade.	Perda granularidade.	Escalabilidade.	Alta
4	FineWeb2...	Pipeline 1000+ langs.	Viés em low-resource.	Alinhamento cultural.	Alta
5	Understand Writing...	Separação estilo/conteúdo.	Treinado em inglês.	Reddit/Discord.	Alta
6	Qwen Family...	Qualidade e abertura.	Limite subjetivo.	Alinhamento cultural.	Alta
7	PART...	Embeddings contrastivos.	Batch size grande.	Redes sociais/Multilíngue.	Alta
8	Intro to Transformers...	Explicação formal.	Sem experimentação.	Aplicações pedagógicas.	Essencial
9	Can LLMs Identify...	Explicabilidade via LIP.	Contexto limitado.	Generalização idiomas.	Muito Alta
10	Disentangling Struct...	Prever viés por estilo.	Exige dados estruturados.	Redes sociais.	Muito Alta
11	Attention Is All You Need	Paralelização/Long-range.	Custo quadrático.	Atenção linear/Multimodal.	Fundamental

Pontifícia Universidade Católica de Goiás
Pontifical Catholic University of Goias
Av. Universitária, 1069, Setor Universitário
Caixa Postal 86 - CEP 74.605-010
Goiânia – Goiás – Brasil

RESOLUÇÃO n° 038/2020 – CEPE

ANEXO I

APÊNDICE ao TCC

Termo de autorização de publicação de produção acadêmica

O estudante Carine Silva Santos _____ do Curso de Engenharia da computação _____ matrícula 2021100330025-8 _____, telefone: 62994434629_____ e-mail cahsw4@gmail.com_____, na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei dos Direitos do Autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás) a disponibilizar o Trabalho de Conclusão de Curso intitulado Análise da Correlação entre Estilo Textual e Opiniões em Redes Sociais Multiculturais por Meio de Técnicas de Processamento de Linguagem Natural

_____

_____

_____

_____, gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos, conforme permissões do documento, em meio eletrônico, na rede mundial de computadores, no formato especificado (Texto(PDF); Imagem (GIF ou JPEG); Som (WAVE, MPEG, AIFF, SND); Vídeo (MPEG, MWV, AVI, QT); outros, específicos da área; para fins de leitura e/ou impressão pela internet, a título de divulgação da produção científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, 30_____ de Setembro _____ de 2025_____.

Documento assinado digitalmente
gov.br
CARINE SILVA SANTOS
Data: 30/09/2025 20:16:25-0300
Verifique em https://validar.iti.gov.br

Assinatura do autor: _____ _____

Nome completo do autor: Carine Silva Santos_____

Pontifícia Universidade Católica de Goiás
Pontifical Catholic University of Goias
Av. Universitária, 1069, Setor Universitário
Caixa Postal 86 - CEP 74.605-010
Goiânia – Goiás – Brasil

Assinatura do professor-orientador: _____
Nome completo do professor-orientador: