Monitorando a Opinião Pública sobre Operações Policiais no Brasil via Comentários de Vídeos no YouTube*

Saul Sousa da Rocha[1], **Carlos Henrique do Vale e Silva**[1]
Carlos H. G. Ferreira[2], **Glauber Dias Gonçalves**[1], **Jussara Marques de Almeida**[3]

[1]Universidade Federal do Piauí (UFPI) - CSHNB

[2]Universidade Federal de Minas Gerais (UFMG) – DCC

[3]Universidade Federal de Ouro Preto (UFOP) – DECSI

{saul.rocha2001, carlosvale}@ufpi.edu.com

ggoncalves@ufpi.edu.br, {chgferreira, jussara}@dcc.ufmg.br

Resumo. *Neste trabalho, propomos um sistema que utiliza comentários de usuários no YouTube para monitorar a percepção das pessoas sobre operações policiais em incidentes de violência urbana com repercussão nessa plataforma. Exploramos atributos extraídos desses comentários e modelos de processamento de linguagem natural, mostrando os desafios dessa inferência ao longo de dois anos. Nossos melhores modelos alcançaram acurácia e macro-F1 de 87% para inferir posicionamentos de aprovação, desaprovação e neutralidade, além de uma boa capacidade de generalização em diferentes plataformas, avaliada no Twitter/X e YouTube. Como resultados identificamos períodos com posicionamentos dominantes, que desconsiderando neutralidade, tendem majoritariamente à aprovação das operações policiais, ao passo que desaprovações foram identificadas em granularidade regional.*

Abstract. *In this work, we propose a system that uses user comments on YouTube to monitor people's perception of police operations in incidents of urban violence with repercussions on this platform. We explore attributes extracted from these comments and natural language processing models, showing the challenges of this inference over two years. Our best models achieved accuracy and macro-F1 of 87% to infer positions of approval, disapproval, and neutrality, in addition to a good generalization capacity across different platforms, evaluated on Twitter/X and YouTube. As a result, we identified periods with dominant positions, which, disregarding neutrality, mostly tend to approve police operations, while disapprovals were identified at regional granularity.*

1. Introdução

A segurança pública é vital para o bem-estar em centros urbanos, enfrentando desafios significativos em países como o Brasil, com altos índices de criminalidade [Ricardo et al. 2013]. Em 2022 e 2023, foram registradas 40.784 e 39.492 mortes por crimes violentos, respectivamente [NEV-USP 2024]. A polícia está na linha de frente no

*Esta pesquisa é financiada com o apoio do Programa Institucional de Bolsas de Iniciação Científica (PIBIC) UFPI.

combate à violência urbana, gerando discussões sobre a condução das operações policiais e as mídias sociais emergem como espaços para expressão pública, com o YouTube se destacando no Brasil para compartilhar notícias e eventos. Segundo o instituto Reuters, 43% dos entrevistados utilizam o YouTube para consumir notícias [Newman et al. 2022]. Vídeos de operações policiais no YouTube recebem comentários variados que refletem tanto apoio quanto críticas, oferecendo uma fonte rica de dados para análise de percepções públicas [Brainard and Edlins 2015, Brown 2016, Junior et al. 2022]

Contudo, existem desafios no processamento de linguagem natural (PLN) para inferir posicionamentos sobre violência urbana e operações policiais. A dificuldade em adquirir comentários rotulados e a representação de ruídos como erros de digitação, gírias e sarcasmo complicam essa tarefa [Feitosa et al. 2022]. Aplicações de Large Language Models (LLM) têm custo elevado e variabilidade nos resultados, com um F1 recente de apenas 56% [Kocoń et al. 2023]. Na literatura, diversos estudos já empregaram dados de mídias sociais para explorar operações policiais e sentimentos em comentários [Hand and Ching 2020, Chaparro et al. 2020], mas há uma lacuna na melhoria de desempenho dos modelos e na transferência de aprendizagem entre plataformas. Na literatura existente, uma variedade de estudos já empregou dados de mídias sociais para explorar operações policiais, avaliando sentimentos em comentários [Hand and Ching 2020, Chaparro et al. 2020] e até mesmo predizendo crimes e taxas criminais [Tucker et al. 2021].

Em particular, a detecção do posicionamento da população com comentários do Twitter foi objeto de investigação em diversos cenários, incluindo a pandemia de COVID-19 [Hossain et al. 2020, Weinzierl et al. 2021], vacinação [D'Andrea et al. 2019] e operações policiais [Feitosa et al. 2022]. Contudo, identifica-se uma lacuna no que tange à melhoria de desempenho dos modelos de posicionamento, em especial novas estratégias de anotações de comentários, que é uma tarefa exaustiva e propensa a inconsistências entre anotadores. Adicionalmente, faltam análises sobre quão aplicáveis são os modelos desses trabalhos a comentários de plataformas diferentes, onde eles não foram treinados, i.e., a transferência de aprendizagem para inferências inter plataformas de mídias sociais.

Para lidar com essa lacuna, o presente estudo introduz uma metodologia para examinar a percepção pública acerca de operações policiais com ampla repercussão em mídias sociais, utilizando comentários de duas plataformas distintas: Twitter (atual X) e YouTube. Nosso foco reside em analisar como esses comentários expressam aprovação ou desaprovação às operações policiais. Para tanto, propomos um sistema de monitoramento dessas operações com repercussão no YouTube, dada a relevância de tal plataforma para o compartilhamento de notícias policiais.Em seguida, propomos uma estratégia para aumentar o desempenho de modelos PLN baseados em arquiteturas *transformers*, explorando uma vasta base de comentários existente na literatura rotulados com posicionamentos de usuários da plataforma Twitter sobre operações policiais no Brasil [Feitosa et al. 2022]. Especificamente, tratamos inconsistências dessa base com o suporte do *Generative Pre-trained Transformer (GPT)*, que é uma das mais populares LLM na atualidade. Em uma análise pioneira, aplicamos esses modelos para capturar tendências de opinião pública em vídeos sobre operações policiais de grande repercussão no YouTube ao longo de dois anos consecutivos em diferentes regiões do Brasil.

Os resultados mostram um aumento significativo no desempenho dos modelos

no Twitter, onde comentários foram utilizados para treinos e testes, e desempenho satisfatório em testes com comentários do YouTube, plataforma em que os modelos não foram treinados e apenas usado sobre a premissa de transferência de conhecimento. Especificamente, chegamos a um modelo com acurácia de 88% e F1-macro de 87% no Twitter, ganhos de desempenho superiores a 18% e 7% respectivamente em comparação à literatura, e acurácia de 83% e f1-macro 73% no YouTube. Por sua vez, a análise de tendências de opinião pública no YouTube sobre operações policiais aponta variações por regiões no Brasil (e.g., Nordeste e Sudeste) e variações ao longo do tempo. Isso nos permitiu discernir momentos de maior aprovação ou desaprovação, evidenciando a capacidade do modelo de captar em tempo real as dinâmicas da opinião pública.

Em suma, este trabalho oferece duas contribuições: (1) estratégia baseada em LLM para reduzir conflitos de rotulação entre anotadores visando melhoria de modelos especializados em detectar posicionamentos sobre operações policiais; e (2) análise experimental do impacto dessa estratégia no desempenho de classificadores para inferir posicionamentos de comentários em plataformas de mídias sociais distintas, i.e., comentários do Twitter utilizados para treinos e testes, e comentários do YouTube utilizados apenas para testes e aplicação.

As próximas seções desse artigo estão organizadas da seguinte forma: A Seção 2 apresenta os trabalhos relacionados. Em seguida, a Seção 3 detalha a coleta, o pré-processamento e os modelos de inferência. Nossa avaliação e resultados são discutidos nas Seções 4 e 5 ao passo que nossas considerações finais são apresentadas na Seção 6.

2. Trabalhos Relacionados

O aumento do interesse na análise da opinião pública sobre eventos via dados de mídias sociais tem fomentado um crescente corpo de pesquisa. Pesquisas precedentes descritas a seguir revelam abordagens promissoras que fundamentam e enriquecem nosso trabalho.

Em [Chakraborty and Sharma 2019], foi conduzida uma análise de sentimentos de *tweets* durante a implementação da política de rodízio veicular em Delhi, evidenciando a capacidade de extrair opiniões públicas das mídias sociais para compreender a reação do público a medidas governamentais. Este estudo sublinha a importância da análise de sentimentos em tempo real para captar a resposta pública a políticas de transporte, alinhando-se com nosso propósito de desvendar a percepção sobre ações policiais.

Por sua vez, Wang *et al.* identificaram períodos críticos em eventos controversos por meio da análise de sentimentos, ressaltando a necessidade de detectar emoções predominantes para prever as tendências da opinião pública [Wang et al. 2020]. A metodologia dos autores complementa nossa abordagem ao sugerir que uma análise emocional detalhada pode antecipar o desenvolvimento da opinião pública, estratégia adotada por nós para analisar as reações às operações policiais.

O estudo de Dandrea *et al.* introduz um sistema para avaliar a opinião pública sobre vacinação no *Twitter*, categorizando *tweets* como favoráveis, neutros ou contrários à vacinação [D'Andrea et al. 2019]. Empregando técnicas sofisticadas de mineração de texto e aprendizado de máquina, incluindo *bag-of-words* e SVM, o sistema provou ser altamente eficaz em identificar mudanças de opinião associadas a eventos sociais na Itália. Este método, eficiente e econômico para a captura de opiniões em tempo real,

destaca a importância da análise de mídias sociais em avaliar as perspectivas públicas sobre questões de saúde, proporcionando informações valiosas para nosso estudo sobre as reações públicas a operações policiais.

Ademais, Brachi *et al.* propõem um *Framework* inovador de análise de sentimentos para monitorar a evolução da opinião pública em tempo real, com um estudo de caso focado nas mudanças climáticas, utilizando um classificador LSTM bidirecional que alcançou precisão notável na classificação de emoções e sentimentos [El Barachi et al. 2021]. Este trabalho salienta a importância da avaliação em tempo real da opinião pública, particularmente em contextos de cidades inteligentes, onde a análise de dados pode melhorar a prestação de serviços e a tomada de decisões governamentais. Já Bechini *et al.* investigam o emprego de análises de sentimentos para entender as dinâmicas de interação social online, oferecendo *insights* sobre como as plataformas de mídia social podem capturar a essência da opinião pública em relação a uma ampla gama de temas, desde políticas de transporte até questões de saúde pública [Bechini et al. 2020].

Por fim, Ceron *et al.* aplicam algoritmos de aprendizado de máquina para analisar conteúdos de mídias sociais e inferir tendências políticas, evidenciando o potencial dessas técnicas de processamento de linguagem natural em extrair conhecimentos significativos de grandes volumes de dados, reforçando o potencial dessas metodologias para compreender a opinião pública em face de ações policiais e medidas de segurança [Ceron and Negri 2016].

Esses esforços anteriores revelam a potencialidade das mídias sociais como um rico repositório de dados para análise de opinião pública, enfatizando a relevância de técnicas avançadas de processamento de linguagem natural e aprendizado de máquina. Em nosso trabalho, avançamos nessa direção ao focar em um contexto específico de segurança pública e ações policiais, explorando como a opinião pública sobre tais ações é refletida e pode ser analisada a partir de comentários do YouTube.

3. Metodologia

Nesta seção descrevemos a proposta do sistema para inferir posicionamentos de comentários das pessoas na Web sobre operações policiais no Brasil. O sistema consiste em três módulos como mostra a Figura 1 que serão descritos a seguir. Utilizamos o YouTube como fonte de vídeos e comentários sobre notícias relacionadas a operações policiais devido à sua ampla popularidade e acesso no Brasil, além da disponibilidade gratuita de dados através da API do YouTube versão 3. Assim, foi desenvolvido um programa coletor e classificador de comentários automatizado. Os códigos fontes, bem como a base de dados, deste trabalho estão disponíveis em repositório público.[1]

3.1. Monitoramento

Monitoramos o YouTube para buscar vídeos sobre notícias relacionadas a operações policiais entre os anos de 2021 e 2022. O procedimento de busca utilizou a princípio apenas as palavras-chave “assassinato” ou “morte” ou “roubo” ou “furto” e todas essas palavras relacionadas a “operação policial” para selecionar vídeos no contexto em questão e posteriormente informações de localização geográfica para monitoramentos em regiões

[1] `https://github.com/LABPAAD/crimes_stance`

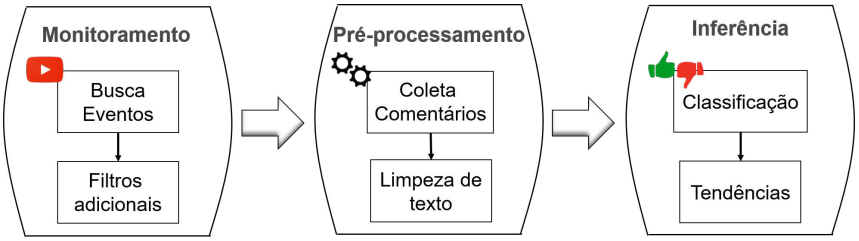


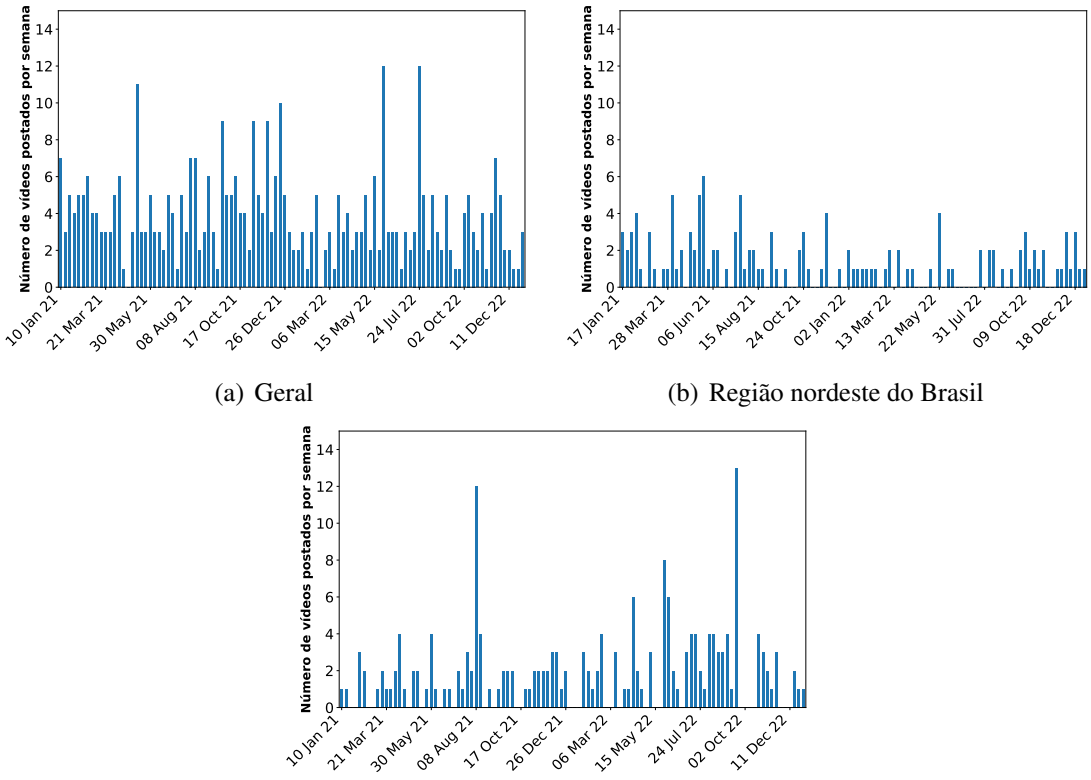
Figura 1. Módulos do sistema proposto enfatizando o fluxo contínuo de monitoramento, pré-processamento e inferência de comentários do Youtube.

específicas. Cada busca foi delimitada por um período de sete dias, i.e., unidade de tempo semanal. Várias buscas foram realizadas de modo a cobrir os anos monitorados com unidades de tempo consecutivas, resultando em 104 unidades de tempo (semanas) nos anos de 2021 e 2022. Uma busca semanal também teve um limite de vinte resultados, que foi o melhor compromisso encontrado para garantir a relevância dos resultados retornados pela API do YouTube.

O monitoramento por região consistiu na definição de um centro por coordenada geográfica e um raio de circunferência em metros de modo a cobrir toda a área de uma região de interesse. Se uma coordenada e raio não são definidos a API retorna vídeos de qualquer região considerando apenas as palavras chave da busca. Monitoramos operações policiais, i.e., desconsiderando regiões, e também as regiões Nordeste e Sudeste do Brasil, para fins de avaliação do sistema proposto. Na Região Nordeste, o centro foi definido próximo a São João do Rio do Peixe, Paraíba (latitude -6.75190 e longitude -38.40820), com um raio de um milhão de metros. Na Região Sudeste, o centro foi localizado entre Bom Jardim de Minas e Lima Duarte, Minas Gerais (latitude -21.90228 e longitude -43.98926), com um raio de quinhentos e noventa mil, trezentos e quarenta e cinco metros. A busca sem coordenada e raio definidos retornou um total de 1.223 vídeos, enquanto a definição desses retornaram 1.328 e 1.313 vídeos no Nordeste e Sudeste respectivamente.

Após a busca inicial, foi aplicado um filtro sob os vídeos coletados para selecionar apenas aqueles relacionados com operações policiais. Primeiramente, os vídeos da categoria “notícia e política” identificados pela API do YouTube foram selecionados, visando remover vídeos sobre blogs e documentários policiais. A seguir, foram selecionados os vídeos que continham no título as palavras “pm”, “pms”, “policia”, “policial”, “policias”, “policiais”, “operacao”, “operacoes”, considerando transformação dos textos em minúsculo, remoção de acentos e sinais ortográficos. Assim, foram selecionados 752 vídeos ao total, sendo que 327 desses corresponderam a buscas por regiões, especificamente 135 da Região Nordeste e 192 da Região Sudeste.

A Figura 2 mostra a distribuição da quantidade de vídeos ao longo do tempo sobre operações policiais resultantes do monitoramento acima descrito. Consideramos como Brasil (Figura 2(a)) o monitoramento sem geolocalização, i.e., vídeos sobre operações policiais em várias regiões brasileiras, ao passo que as regiões nordeste e sudeste (Figuras 2(b) e 2(c)) são monitoramentos geo-localizados. Notavelmente, os picos semanais de vídeos no Brasil não coincidem com os picos regionais. Por exemplo, em 18 setembro de 2022. Isso porque alguns vídeos de menor relevância que atendiam aos critérios de filtragem entram nas coletas regionais e não entram na coleta do Brasil, porém, isso não



(a) Geral

(b) Região nordeste do Brasil

(c) Região sudeste do Brasil

Figura 2. Distribuição da quantidade de vídeos coletados sobre operações policiais nos anos de 2021 e 2022 em unidades de tempo semanal: (a) Geral, (b) na Região Nordeste, e (c) na Região Sudeste do Brasil.

gera grande impacto na quantidade de comentários do período específico.

3.2. Pré-processamento

O pré-processamento visa coletar coletar comentários de todos os vídeos selecionados na etapa anterior e adequá-los para construção de modelos de inferência de posicionamentos.

Codificamos a coleta dos comentários passando o identificador do vídeo como parâmetro. Coletamos informações de todos os comentários disponíveis para cada vídeo, sem aplicar nenhum filtro ou critério de seleção. As informações de cada comentário consistem em um identificador único, data da postagem e o texto do comentário. O programa coletor armazenou essas informações às associando aos seus respectivos vídeos. Foram coletados um total de 257.025 comentários, sendo 11.063 na Região Nordeste e 31.419 na Região Sudeste.

Com esses comentários, conduzimos uma série de pré-processamentos nos comentários coletados via a biblioteca NLTK[2] na linguagem Python, removendo *links*, *emojis* e quebras de linha, mas mantendo *stop words* para inferências com a mesma sequência em que as palavras aparecem nos comentários. Adicionalmente, ignoramos comentários de uma única palavra devido ao seu menor potencial de inferência semântica.

[2]https://www.nltk.org/

3.3. Inferência

Nesta etapa modelos de processamento de linguagem natural (PLN) são aplicados via aprendizagem supervisionada e análise de tendência de posicionamentos de usuários sobre os vídeos e comentários obtidos nas etapas anteriores.

A abordagem de aprendizagem supervisionada adotada consiste na construção de modelos para classificar comentários em *Aprova, Desaprova* ou *Neutro* considerando os posicionamentos dos usuários sobre os vídeos de operações policiais. Os comentários *Aprova* e *Desaprova* significam, respectivamente, que o usuário concorda e discorda com a operação policial do vídeo, ao passo que *Neutro* significa que o usuário é indiferente ao assunto ou não manifesta claramente seu posicionamento. Logo, é necessário rotular um conjunto de comentários suficiente para treinar modelos com essas três classes. Contudo, a rotulação é realizada por humanos, como usual na literatura, o que é uma tarefa exaustiva, propensa a erros e inconsistência. Por outro lado, a rotulação é fundamental para o desempenho dos modelos e por conseguinte qualidade da inferência.

Para lidar com essa questão propomos uma estratégia híbrida, que utiliza a rotulação manual por humanos e reavaliação dos rótulos por *Large Language Models* (LLM). Nesse sentido, Utilizamos o conjunto de 4467 *tweets* sobre operações policiais de grande repercussão no Brasil rotulados manualmente com as classes acima descritas no trabalho de [Feitosa et al. 2022]. A seguir, reavaliamos esses comentários e seus rótulos com a API do *Generative Pre-trained Transformer* (GPT) versão 3.5 Turbo, utilizando o formato de requisição para inferência de posicionamento (*stance*) no GPT proposto em [Kocoń et al. 2023].

A reavaliação foi um passo importante para selecionar comentários com rotulação consistente. Isso porque o LLM GPT é capaz de inferir posicionamentos de todos os comentários seguindo um mesmo padrão. Por outro lado, ele apresenta altas taxas de erro devido baixa especificidade, e.g., F1-Macro em 52% para posicionamentos [Kocoń et al. 2023]. Logo, selecionamos os comentários rotulados igualmente entre os avaliadores em [Feitosa et al. 2022] e o LLM GPT, o que resultou em um subconjunto de 2671 comentários (tweets), i.e. 60% do total. A distribuição das classes desse subconjunto é desbalanceada, seguindo as características do conjunto principal, com comentários neutros, aprovação e desaprovação com 44%, 26% e 30% respectivamente.

Os modelos para classificação foram construídos a partir desses rótulos utilizando BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al. 2019], que é um arcabouço de aprendizado profundo desenvolvido pelo *Google* para processamento de linguagem natural. Um recurso importante do BERT é sua construção com representações contextuais de modelos pré-treinados, baseado na arquitetura *transformers*. Neste trabalho utilizamos dois modelos pré-treinados que são o *Multilingual* [Devlin et al. 2019] e o *BERTimbau* [Souza et al. 2020]. Nesse caso, esses modelos foram retreinados com os comentários rotulados como uma nova camada de neurônios na rede, o que denominamos como *modelos ajustados*, i.e., *fine tunning* para o contexto.

Outra forma de aplicar o BERT é gerando uma matriz de *word embeddings*, que é uma representação quantitativa das características das palavras contidas em um comentário via processamento de linguagem natural. Utilizamos essa matriz para treinar dois modelos de classificação populares, baseados em aprendizagem de máquina: *Ran-*

dom Forest (RF) e *Support Vector Machine* (SVM) com e sem balanceamento de classes via *smote* [Mahesh 2020] e hiper-parâmetros padrões.

Finalmente para inferência de opinião pública o modelo PLN previamente treinado é aplicado para classificar e identificar a tendência ao longo do tempo para posicionamento de usuários no serviço monitorado. Definimos tendências como picos de comentários em unidades de tempo que indicam claramente a percepção pública via os posicionamentos de aprovação, desaprovação ou neutralidade. Esses picos estão relacionados a eventos específicos relativos a incidentes de segurança polêmicos que demandam operações policiais com grande repercussão na plataforma YouTube.

4. Comparando Modelos de Inferência

Nesta seção avaliamos o desempenho dos modelos descritos anteriormente com uma análise dos erros do melhor modelo para mostrar a dificuldade da inferência.

Como ambiente de experimentação utilizamos o *Google Colaboratory* configurado com o uso de GPU, o que permite que nossa metodologia seja reprodutível. Logo, ajustamos os modelos pré-treinados BERTimbau com 10 épocas e otimizador Adam com os respectivos *learning rate* e *batch size* de cada modelo. Para a abordagem com os classificadores, geramos a matriz de *word embedding* com o modelos BERTimbau pré-treinado, em seguida dividimos a matriz em treino e teste dos comentários respectivamente, e treinamos os classificadores sem e com balanceamento de classes. Treinamos todos os modelos utilizando 80% dos comentários do Twitter rotulados acima descritos selecionados aleatoriamente, ao passo que os outros 20% foram utilizados para testar os modelos. Adicionalmente, testamos os modelos com comentários coletados no YouTube (Seção 3.2). Nesse caso, 590 comentários foram selecionados aleatoriamente e rotulados por outros três voluntários, seguindo as mesmas diretrizes da rotulação anterior. Desses, selecionamos 312 comentários cujos rótulos convergiram entre os três voluntários. A distribuição das classes desses comentários é desbalanceada com neutros, aprovação e desaprovação em 45%, 49% e 6% respectivamente.

Nos testes, avaliamos o desempenho dos modelos com cinco métricas: acurácia, precisão, revocação, f1-score e f1-macro para os comentários do Twitter e YouTube como mostra a Tabela 1. A acurácia indica o percentual de classificações corretas, i.e., a soma acertos de todas as classes dividido pelo número total de comentários testados. Já a precisão é calculada para cada classe individualmente e evidencia o percentual de textos corretamente classificados para aquela classe. A revocação é calculada justamente pelo total de textos corretamente classificados para uma classe sobre o total de textos dessa classe. F1-score é a média harmônica entre precisão e revocação para cada classe, ao passo que o F1-macro é a média do F1-score considerando todas as classes.

Primeiramente, discutimos o desempenho dos modelos ao aplicá-los em uma plataforma diferente, i.e., o impacto de treiná-los em um conjunto de comentários do Twitter e testá-los com comentários do YouTube. Observa-se que o desempenho dos modelos é melhor quando aplicados à plataforma onde foram treinados. Nesse caso, a acurácia na plataforma YouTube teve impactos negativos que variam de 2-7%, ao passo que o F1-macro (reflexos da precisão, revocação e f1) teve também impactos negativos variando de 6-17%. O classificador SVM obteve as maiores perdas de desempenho e RF

Tabela 1. Desempenho dos modelos BERTimbau (PtBr) usando as estratégias de ajuste fino da rede neural (RN) e da matriz de *embedding* com os classificadores (SVM e RF). Avaliando com as seguintes métricas: Precisão (P), Revocação (R), F1-score (F1), Acurácia (Acc) e F1-macro.

Teste	Modelos	Desaprova			Aprova			Neutro			Acc	F1-macro
		P	R	F1	P	R	F1	P	R	F1		
Twitter	SVM-PtBr-smote	0,76	0,77	0,77	0,73	0,73	0,73	0,88	0,86	0,87	0,80	0,79
	SVM-PtBr	0,70	0,71	0,71	0,67	0,70	0,69	0,84	0,81	0,82	0,75	0,74
	RF-PtBr-smote	0,75	0,74	0,75	0,81	0,58	0,68	0,78	0,90	0,84	0,78	0,75
	RF-PtBr	0,75	0,68	0,71	0,83	0,56	0,67	0,76	0,94	0,84	0,77	0,74
	RN-PtBr	0,83	0,85	0,84	0,83	0,87	0,85	0,94	0,90	0,92	0,88	0,87
YouTube	SVM-PtBr-smote	0,22	0,67	0,33	0,84	0,83	0,83	0,92	0,70	0,79	0,76	0,65
	SVM-PtBr	0,20	0,89	0,32	0,89	0,70	0,78	0,84	0,66	0,74	0,69	0,62
	RF-PtBr-smote	0,23	0,44	0,30	0,91	0,70	0,79	0,77	0,86	0,81	0,76	0,64
	RF-PtBr	0,32	0,61	0,42	0,95	0,71	0,82	0,78	0,90	0,83	0,79	0,69
	RN-PtBr	0,33	0,72	0,46	0,91	0,88	0,89	0,90	0,80	0,85	0,83	0,73

as menores perdas em ambas as métricas. A despeito dessa observação, o desempenho dos modelos na plataforma YouTube, em geral, foram maiores que os observados em [Feitosa et al. 2022]. Em especial, a rede neural BERTimbau (RN), que é o melhor modelo, alcançou aumentos de acurácia e F1-macro superiores a 18% e 7% respectivamente em relação à esse trabalho, o que significa que a metodologia de treinamento proposta levou a ganhos relevantes no desempenho.

Em mais detalhes, a rede neural BERTimbau alcançou F1-macro de 87(73)% e acurácia de 88(83)% nas plataformas Twitter (YouTube), o que é um desempenho relevante para posicionamentos na literatura [Mohammad et al. 2017]. Em comparação aos classificadores, o desempenho da rede neural é superior ao melhor classificador (SVM-PtBr-smote) na plataforma Twitter com ganho de 8 pontos percentuais em termos de F1-macro. Na plataforma YouTube, o desempenho da rede neural é superior ao melhor classificador (RF-PtBr) com ganho de 4 pontos percentuais. Importante observar que, considerando o custo computacional para ajustar a rede neural BERTimbau, o modelo RF apresenta um bom compromisso entre desempenho e custo de treinamento.

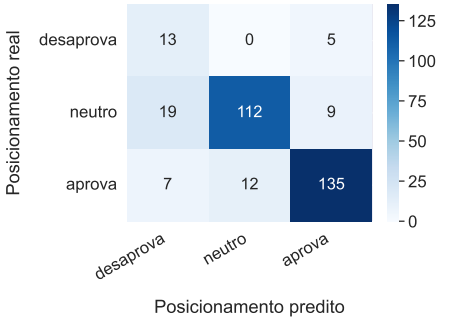
Em seguida, analisamos o efeito do tratamento de desbalanceamento das classes para os classificadores RF e SVM. Esse tratamento visa aumentar a precisão e a revocação para as classes minoritárias, evitando que apenas comentários com posicionamentos claros de aprovação ou desaprovação sejam classificadas corretamente e tendencia a posicionamentos neutros. O balanceamento com *smote* teve impacto positivo no desempenho de SVM e RF nos comentários do Twitter, aumentando o F1 e, por consequência, aumentando a acurácia e F1-macro para as posturas de aprovação e desaprovação. Nesse caso, SVM teve o melhor desempenho alcançando F1-macro melhores (79%), superando as marcas de F1-macro do classificador RF (75%). Contudo, observa-se menor impacto do balanceamento com smote para os comentários do YouTube, e RF teve o melhor desempenho nessa plataforma alcançando F1-macro melhor (69%).

As Figuras 3(a) e 3(c) reportam matrizes de confusão para as plataformas Twitter e YouTube no modelo RN-PtBr, que obteve o melhor desempenho, como mostra a Tabela 1. Cada linha representa os comentários em uma classe real, enquanto cada coluna representa os comentários em uma classe prevista, o que nos permite analisar onde o classificador mais erra e como isso acontece em função da classe. De forma geral, é possível notar que, em termos absolutos, o maior desafio na tarefa aqui endereçada está

(a) Matriz Twitter

Comentário	Real	Predito
vergonha para o pais mesmo	-1	1
prato cheio para o trafico e seus simpatizantes	-1	1
abordagem foi execucao mesmo	-1	1
o importante e que morreu	1	0
sim pq a policia nao pode subir nos morros	1	0
bandido bom e bandido	1	0
kkkkkkkkkkkkk discurso dentro da cartilha parabens	0	1
nenhum adv para lutar pela ajudar essa mulher	0	-1
isso e um tiro no proprio pe	0	-1

(b) Exemplos Twitter



(c) Matriz YouTube

Comentário	Real	Predito
Ban... bom é ban... morto!!!!!	1	0
Que na próxima vez, ele venha o dobro.	1	0
Eu sou brasileiro e estou a favor da polícia	1	0
O câncer do Brasil e os fardados , apenas 1 morto foi pouco !	-1	1
Poderiam usar armas de choque, tranquilizantes e etc, não precisava ser letal a menos que a PM quisessem esconder algo.	-1	1
Como são ruim de Tiro esses pms de SP ein quase toda troca de tiro eles leva a pior mesmo usando colete e nunca acerta o bandido.	-1	1
s t f. tem ser preso vagabundos	0	1
Petistas defendendo os seus bandidos de estimação	0	-1
Cdd é cemitério de polícia	0	1

(d) Exemplos YouTube

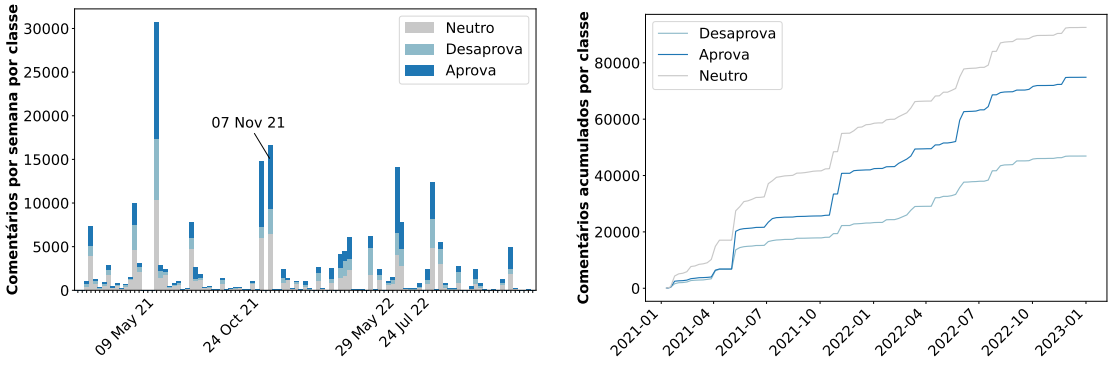
Figura 3. Erros do modelo BERTimbau (RN-PtBr): (a) matriz de confusão e (b) exemplos do Twitter, (c) matriz de confusão e (d) exemplos do YouTube.

em diferenciar as classes *Neutro* da *Desaprova* e *Desaprova* da *Aprova*. Por sua vez, os exemplos apresentados nas Figuras 3(b) e 3(d) mostram como comentários curtos, contendo algum tipo de sarcasmo ou que não apontam um posicionamento claro em relação à ação policial, dificultam a tarefa de inferência aqui endereçada. Nós observamos que, de fato, esses casos são representativos dos comentários classificados incorretamente.

5. Monitorando Opinião Pública

Nesta seção, discutimos os resultados da inferência do melhor modelo (RN-PtBr) nos comentários do YouTube coletados com o monitoramento ao longo dos anos 2021 e 2022.

A Figura 4 mostra a distribuição de comentários sobre operações policiais no Brasil em geral nesse período em unidades de tempo semanal e acumulado para as três classes de posicionamentos inferidas pelo sistema. Observa-se majoritariamente neutralidade, em especial nos picos de comentários cujas datas são indicadas na Figura 4(a). Isso ocorre devido a interferência de temas externos ao contexto, em especial politização de discussões sobre operações policiais. Contudo, aprovações dominam notavelmente desaprovações em quatro dos cinco picos indicados. Na Figura 4(b) a tendência de aprovação sob desaprovação está mais evidente. Considerando o acumulado total no período temos posicionamentos de aprovação 35%, desaprovação 22% e neutralidade 43%. Observa-se que comentários não neutros aumentam significativamente no mês 05-2021: aprovações, inicialmente pouco superior a 6.400 em Abril, aumenta para 20.204 a partir de 9 de Maio

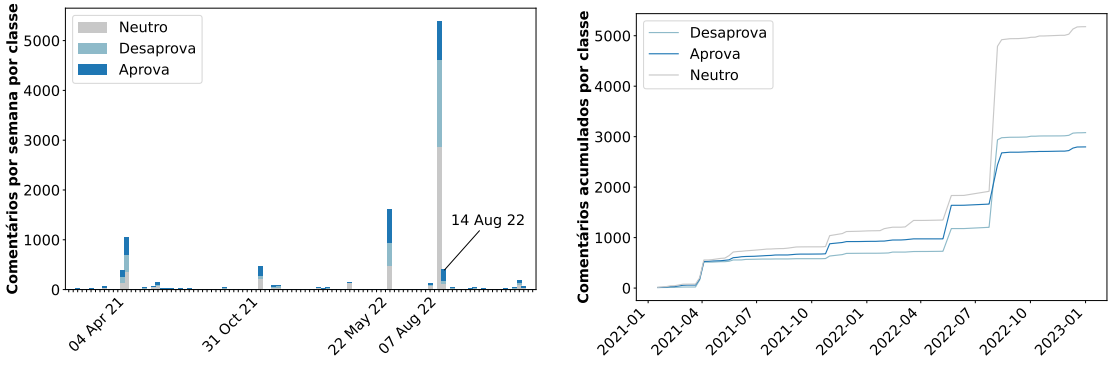


(a) (b)

Figura 4. Distribuição de comentários sobre operações policiais no Brasil em geral em 2021 e 2022: (a) unidades de tempo semanal e (b) acumulado.

de 2021, tornando esse o mês com os maiores percentuais de opiniões de aprovação, impactando no acumulado notavelmente.

Quanto aos picos semanais de comentários na Figura 4(a), é importante destacar que eles estão relacionados com vídeos das seguintes operações: "*Operação no Jacarezinho: moradores registram ação policial que deixou 25 mortos no Rio (09 mai. 21)*", "*Bandidos são mortos após tentarem fugir da PM - SBT Brasil (24 out. 21)*", "*Combate ao "Novo Cangaço": 26 mortos em operação policial (07 nov. 2021)*", "*Operação policial faz 22 mortos no Rio de Janeiro (29 maio 2022)*", "*Policiais do BOPE são encurralados por bandidos, no Complexo do Alemão (24 jul. 2022)*".

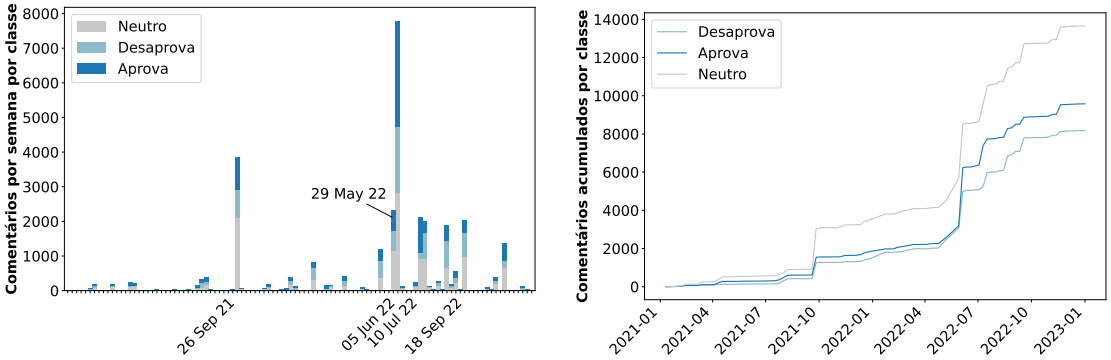


(a) (b)

Figura 5. Distribuição de comentários sobre operações policiais na região nordeste do Brasil em 2021 e 2022: (a) unidades de tempo semanal e (b) acumulado.

A Figura 5 mostra a distribuição de comentários sobre operações policiais na região Nordeste do Brasil entre 2021 e 2022. Observa-se picos semanais, porém em uma quantidade menor de comentários comparado ao cenário Brasil geral dado que esses eventos foram monitorados em uma área específica. Cinco picos são identificados, mas o destaque é para a semana 07 de Agosto de 2022, quando a operação reportada no vídeo "*Policiais embriagados são presos após bater em carro e gerar confusão na Paraíba*" provoca uma mudança na tendência da opinião pública acumulada no período,

i.e., desaprovações ultrapassa aprovações, como mostra a Figura 5(b). Assim, a tendência de aprovação não é evidente para os vídeos reportados na região nordeste, e considerando o acumulado temos aprovação 25%, desaprovação 28% e neutralidade 47%.



(a) (b)

Figura 6. Distribuição de comentários sobre operações policiais na região sudeste do Brasil em 2021 e 2022: (a) unidades de tempo semanal e (b) acumulado.

Finalmente, para os comentários na região sudeste, observa-se picos com volumes superiores à região nordeste e mais concentrados em 2022, como mostra a Figura 6(a). Não há evidência de posicionamentos de aprovação ou desaprovação dominantes. Porém, o pico destacado na semana de 5 Junho de 2022 (operação "*Intenso tiroteio e policiais encurralados no Complexo da Penha*") impacta nos comentários de aprovação a tornando levemente superior às desaprovações. Portanto, observa-se uma tendência de aprovação para os vídeos reportados na região sudeste, ainda que leve, indicado pelos percentuais de posicionamentos acumulados em aprovação 31%, desaprovação 26% e neutralidade 43%.

6. Conclusões e Trabalhos Futuros

Neste trabalho foi proposto um sistema para coleta, pré-processamento e inferência de opinião sobre operações policiais no Brasil em geral e por regiões, baseado em comentários sob vídeos da plataforma YouTube nesse contexto. Considerando os desafios em processamento de linguagem natural, esse trabalho contribui na melhoria do desempenho de modelos para inferência de posicionamentos baseado em arquiteturas *transformers* com uma estratégia que utiliza LLM na redução de conflitos de rotulação entre anotadores. Foram conduzidos experimentos que mostram o impacto positivo dessa estratégia no desempenho de modelos para inferir posicionamentos de comentários em plataformas de mídias sociais distintas. Especificamente, comentários do Twitter foram utilizados para treinos e testes de modelos, ao passo que utilizamos comentários do YouTube para testes, seguidos de demonstrações extensivas de aplicações do sistema proposto.

Trabalhos futuros incluem a melhoria do módulo de coleta do sistema visando diminuir impacto de fatores não alvo ao contexto analisado e estimativa mais acurada dos erros do modelo que permitam maior confiança no uso de comentários em plataformas como YouTube para inferir opinião da população em temas de importância para sociedade como segurança pública.

Referências

Bechini, A., Ducange, P., Marcelloni, F., and Renda, A. (2020). Stance analysis of twitter users: the case of the vaccination topic in italy. *IEEE Intelligent Systems*, 36(5):131–139.

Brainard, L. and Edlins, M. (2015). Top 10 us municipal police departments and their social media usage. *The American Review of Public Administration*, 45(6):728–745.

Brown, G. R. (2016). The blue line on thin ice: Police use of force modifications in the era of cameraphones and youtube. *British journal of criminology*, 56(2):293–312.

Ceron, A. and Negri, F. (2016). The "social side" of public policy: Monitoring online public opinion and its mobilization during the policy cycle. *Policy & Internet*, 8(2):131–147.

Chakraborty, P. and Sharma, A. (2019). Public opinion analysis of the transportation policy using social media data: a case study on the delhi odd–even policy. *Transportation in Developing Economies*, 5:1–9.

Chaparro, L. F., Pulido, C., Rudas, J., Reyes, A. M., Victorino, J., Narváez, L. Á., Gómez, F., and Martinez, D. (2020). Sentiment analysis of social network content to characterize the perception of security. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 685–691. IEEE.

D'Andrea, E., Ducange, P., Bechini, A., Renda, A., and Marcelloni, F. (2019). Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, 116:209–226.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

El Barachi, M., AlKhatib, M., Mathew, S., and Oroumchian, F. (2021). A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change. *Journal of Cleaner Production*, 312:127820.

Feitosa, M. P. F., Ferreira, C. H., Gonçalves, G. D., and de Almeida, J. M. (2022). Análise da percepção das pessoas no twitter sobre ações policiais. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 73–84. SBC.

Hand, L. C. and Ching, B. D. (2020). Maintaining neutrality: A sentiment analysis of police agency facebook pages before and after a fatal officer-involved shooting of a citizen. *Government Information Quarterly*, 37(1):101420.

Hossain, T., Logan IV, R. L., Ugarte, A., Matsubara, Y., Young, S., and Singh, S. (2020). Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Junior, J. M. d. R., Linhares, R. S., Ferreira, C. H. G., Nobre, G. P., Murai, F., and Almeida, J. M. (2022). Uncovering discussion groups on claims of election fraud from twitter. In *International Conference on Social Informatics*, pages 320–336. Springer.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., et al. (2023). Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386.

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

NEV-USP (2024). Monitor da violência. Disponível em: https://nev.prp.usp.br/projetos/projetos-especiais/monitor-da-violencia/. Acesso em 07 de mar. 2024.

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., and Nielsen, R. K. (2022). Reuters institute digital news report 2022. *Reuters Institute for the study of Journalism*.

Ricardo, C. d. M., de Siqueira, P. P., and Marques, C. R. (2013). Estudo conceitual sobre os espaços urbanos seguros. *Revista brasileira de segurança pública*, 7(1):200–216.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems*.

Tucker, R., O'Brien, D. T., Ciomek, A., Castro, E., Wang, Q., and Phillips, N. E. (2021). Who 'tweets' where and when, and how does it help understand crime rates at places? measuring the presence of tourists and commuters in ambient populations. *Journal of Quantitative Criminology*, 37(2):333–359.

Wang, M., Wu, H., Zhang, T., and Zhu, S. (2020). Identifying critical outbreak time window of controversial events based on sentiment analysis. *Plos one*, 15(10):e0241355.

Weinzierl, M., Hopfer, S., and Harabagiu, S. (2021). Misinformation adoption or rejection in the era of covid-19. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), AAAI Press*.