Análise da Percepção das Pessoas no Twitter Sobre Ações Policiais*

Marcos Paulo Fontes Feitosa[1], **Carlos H. G. Ferreira**[2,3]
Glauber Dias Gonçalves[1], **Jussara Marques de Almeida**[2]

[1]Universidade Federal do Piaui (UFPI) – CSHNB
[2]Universidade Federal de Minas Gerais (UFMG) – DCC
[3]Universidade Federal de Ouro Preto (UFOP) – DECSI

Resumo. *É cada vez mais frequente o uso de redes sociais online como meio para as pessoas trocarem ideias e expressarem suas opiniões sobre diferentes aspectos do cotidiano, incluindo violência e insegurança, um problema central a vários centros urbanos. Neste artigo investigamos o potencial uso de comentários compartilhados em uma rede social bastante popular – o Twitter – para inferir a opinião pública sobre a atuação policial em incidentes de segurança de grande repercussão. Nesse sentido, exploramos atributos extraídos desses comentários e modelos de aprendizado de máquina para inferir o posicionamento das pessoas em relação a ações policiais específicas. Nossos experimentos mostram quão desafiante é essa inferência dado grande quantidade de neutralidade e sarcasmo observado em mídias sociais. Não obstante, nossos melhores classificadores alcançaram acurácia e especificidade (macro F1) superiores a 68% para inferir posicionamentos de aprovação, desaprovação e neutralidade da população.*

1. Introdução

A violência causada por altos índices de criminalidade e a sensação de insegurança das pessoas estão dentre os principais problemas dos centros urbanos no mundo. Para se ter uma ideia da gravidade desse problema no Brasil, no ano de 2019, foram registradas 41.726 mortes por crimes violentos, uma taxa média de 20 mortes por mês para cada 100 mil habitantes brasileiros. Essa taxa pode alcançar valores superiores a 40 mortes por mês em estados das regiões norte e nordeste [NEV-USP 2021].

Diante desses fatos, costumeiramente a população não só brasileira, mas também mundial, vem recorrendo cada vez mais às redes sociais virtuais para manifestar sentimentos e opiniões sobre a atuação dos órgãos e autoridades responsáveis pela segurança pública. Essas manifestações virtuais podem refletir não apenas a percepção negativa sobre a violência urbana, mas também a aprovação ou desaprovação de ações policiais durante o enfrentamento de incidentes de segurança. Além disso, observa-se neutralidade ou mesmo polarização em casos polêmicos que despertam atenção das pessoas, levando a muitos comentários postados e compartilhados publicamente nas redes sociais virtuais [Tucker et al. 2021, Iranmanesh and Alpar Atun 2020].

Ao passo que esses comentários são uma das principais formas de manifestação da população, eles oferecem uma oportunidade única para compreender a percepção da

*Esta pesquisa é financiada pelo CNPq processo no. 402194/2021-7.

população, principalmente, durante a atuação da polícia. Tipicamente, um ecossistema é formado em que, primeiro, as notícias sobre incidentes de segurança envolvendo a polícia são postadas por portais jornalísticos e, posteriormente, são compartilhadas em redes sociais virtuais recebendo milhares de comentários. Esses, por sua vez, geram novos comentários sejam aprovando, desaprovando ou sem uma posição clara sobre a ação policial [Reis et al. 2015, Ferreira et al. 2021]. Dessa forma, toda essa carga de conteúdo gerada por usuários acerca desses incidentes, potencialmente, possibilita compreender a percepção da população quanto ao esforço dos órgãos de segurança pública no enfrentamento desse problema nas cidades.

Na literatura, esforços anteriores já usaram dados dessa natureza no contexto da atuação da polícia para avaliar o sentimento dos comentários [Hand and Ching 2020, Chaparro et al. 2020], e realizar a predição de crimes e taxas criminais [Chen et al. 2015, Tucker et al. 2021]. Com foco específico na detecção de posicionamentos da população em redes socais, estudos foram realizados em outros contextos, por exemplo, pandemia de COVID-19 [Hossain et al. 2020, Weinzierl et al. 2021]. Dessa forma, observa-se ainda a falta de estudos e metodologias para inferir como a população reage em face às políticas de segurança pública, em particular, a exploração de conteúdo gerado por usuários em redes sociais sobre ações policiais. Especificamente, há a necessidade de entender como conteúdo nesse contexto é gerado e como ele pode ser quantificado. A partir desse ponto, pode-se investigar o potencial desses dados para inferir posicionamentos, especialmente o quanto a população apoia as ações de segurança pública. No entanto, existem desafios do ponto de vista computacional, principalmente, relacionados ao processamento de linguagem natural. Especificamente, a obtenção de dados rotulados no contexto específico de violência urbana, a representação semântica desses dados considerando o ruído presente em textos de redes sociais online, por exemplo, erros de digitação, uso de gírias e sarcasmo, e o desbalanceamento de opiniões ou posicionamentos das pessoas são alguns dos principais desafios reportados por trabalhos anteriores [Wang et al. 2017, Minaee et al. 2021].

Este trabalho oferece uma primeira tentativa em direção a inferência da opinião pública sobre a atuação policial em incidentes de segurança de grande repercussão em mídias de comunicação. Especificamente, nós investigamos o potencial de uso de dados de redes sociais para inferir a opinião pública do ponto de vista se elas aprovam ou não as ações realizadas pela polícia durante uma determinada ação policial. Para isso, nós determinamos um conjunto de notícias alvo divulgadas por três diferentes portais visando mitigar o viés ideológico, coletamos milhares de comentários sobre elas no *Twitter*, e aplicamos uma metodologia de rotulação acerca do posicionamento dos usuários sobre aquela notícia. Em seguida, nós exploramos modelos de *deep learning* baseados em arquiteturas *transformers*, que compreendem o estado da arte, para o obter uma representação semântica mais fiel ao contexto. Por fim, nós avaliamos os principais classificadores e redes neurais como uma tarefa de classificação para analisar a capacidade da nossa representação em capturar a aprovação ou não a operações policiais. Nossos resultados mostram que o uso de modelos dessa natureza oferecem um potencial enorme para inferência da percepção da população neste contexto.

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 processa-

mento e a rotulação dos dados. 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

As plataformas de mídia social são atualmente um importante fórum para as pessoas expressarem suas opiniões e trocarem informações. Ao interagir uns com os outros por meio de postagens, re-postagens, respostas ou menções, os usuários favorecem a disseminação de informações [Al-Garadi et al. 2018]. Assim, o crescente uso dessas plataformas tem chamado a atenção de vários pesquisadores que visam modelar e analisar o comportamento dos usuários em face de fenômenos reais. Em particular, análises utilizando ferramentas de processamento natural de linguagem tem auxiliado na compreensão de diversos aspectos, por exemplo, relacionados a política [Ferreira et al. 2021, Nobre et al. 2022], pandemia [Malagoli et al. 2021] e percepção da segurança urbana [Hand and Ching 2020, Oglesby-Neal et al. 2019]. Nesta seção nós discutimos alguns destes esforços mais próximos deste trabalho.

Especificamente no contexto da percepção de segurança pelas pessoas manifestada em plataformas de mídias sociais, técnicas para análise de sentimentos vem sendo até então a principal forma de análise nesse tema. Em [Hand and Ching 2020] foram analisados os sentimentos dos comentários das pessoas às postagens das agências de polícia em suas páginas no Facebook antes e após incidentes com tiros envolvendo policiais. Os autores observaram tendência à neutralidade logo após o incidente, o que pode indicar pouca eficiência dessa técnica para analisar percepção de segurança. Os autores de [Oglesby-Neal et al. 2019] focaram em uma ação policial que resultou na morte do afro-americano Freddie Gray na cidade de Baltimores em 2015. Por sua vez, os autores de [Chaparro et al. 2020] utilizaram *tweets* georreferenciados em Bogotá na Colômbia para identificar o sentimento da população sobre a segurança nessa cidade. Cada *tweet* foi rotulado como um sentimento positivo ou negativo por um grupo de especialistas e os autores observaram que técnicas de aprendizagem de máquina supervisionada obtiveram acurácia e especificidade melhores que técnicas baseadas em regras léxicas.

Outro contexto bastante explorado é a predição de crimes a partir de características extraídas do conteúdo gerado por usuários em mídias sociais. Por exemplo, em [Wang et al. 2012] *tweets* postados por agências de notícias foram utilizados para identificar possíveis crimes. Já em [Chen et al. 2015], os autores usam *tweets* geolocalizados com foco na reincidência de crimes em uma dada região. Alguns trabalhos mesclam fontes de dados oficiais a dados de redes sociais para analisar o quanto, postagens do *Twitter* estão correlacionados com a violência urbana [Tucker et al. 2021, Iranmanesh and Alpar Atun 2020]. Em [Wang et al. 2017, Sousa et al. 2021] foi investigado o potencial de pontos de interesse gerados por usuários em serviços de mapeamento para predição das taxas criminais.

Mais recentemente, várias das tarefas supramencionadas têm sido realizadas por estratégias mais avançadas para o processamento de linguagem natural, como BERT [Devlin et al. 2019], SentenceBERT [Reimers et al. 2019], e outras variações de arquiteturas *transformers*. De fato, estes modelos fornecem representações mais ricas, portanto, com resultados potencialmente melhores. Focando no contexto do *Twitter*, alguns esforços devem ser mencionados. Em [Hossain et al. 2020], os autores utilizaram vários

modelos pré-treinados para identificar o posicionamento dos usuários quanto a conceitos errôneos sobre o COVID-19 no Twitter. Os resultados alcançados que, no melhor modelo, apresenta um *F1-score* de 50.2 evidenciam os desafios dessa tarefa. Considerando o mesmo conjunto de dados e tarefa, o trabalho de [Weinzierl et al. 2021] conseguiu aumentar o valor do *F1-score* para 74.3. Ainda no contexto da pandemia, Glandt et al. [Glandt et al. 2021] realizaram a análise do desempenho dos classificados por instância, enfatizando que os resultados variam de acordo com a instância. Em geral, os autores observaram *F1-score* que variam de 0.53 a 0.83. Assim, eles reforçam o argumento de que é difícil ter um único modelo que seja melhor para todos os tipos de instâncias analisados, ou seja, que capture todas as nuances presentes em textos de redes sociais online. Esses resultados refletem o atual estado da arte, como mostram algumas revisões recentes sobre tal tarefa em dados de redes sociais [Minaee et al. 2021].

Em suma, neste trabalho apresentamos um estudo ortogonal aos esforços anteriores. Nosso objetivo aqui é inferir posicionamentos das pessoas quanto incidentes de violência envolvendo atuação policial através de técnicas para processamento de linguagem natural baseada na arquitetura de *transformers*, que fornecem uma representação semântica estado da arte para esse tipo de processamento. Para isso, nós exploramos fontes de informações geradas por diversas agências de notícias similares aos trabalhos de [Wang et al. 2012, Chen et al. 2015] e as usamos para quantificar o posicionamento dos usuários em um contexto específico e não explorado dessa forma anteriormente, que é a sobre a atuação policial.

3. Bases de Dados

Nesta seção descrevemos as bases de dados e a metodologia de processamento desses dados para o uso em inferências sobre posicionamentos da população sobre atuação policial em incidentes de violência.

3.1. Coleta de dados

Realizamos a coleta de dados em duas etapas. Primeiramente, selecionamos notícias de grande repercussão sobre o contexto desse trabalho em portais de mídia digital. Em seguida, coletamos comentários das pessoas em redes sociais (*tweets*) sobre essas notícias como descrito a seguir.

Inicialmente, selecionamos notícias sobre incidentes de segurança com intervenção policial entre o período de 04/2019 até 12/2021 que foram amplamente divulgadas e repercutidas em mídia digital. Nesse sentido, pesquisamos por notícias nas seções sobre violência e crimes em três relevantes portais de notícias nacionais: UOL[1], Folha de São Paulo[2], e G1[3]. Utilizamos como critério de seleção a participação dos leitores via comentários de texto no próprio portal de notícia. Especificamente, selecionamos notícias com mais de 700 comentários, marca notável e acima da média usual de comentários e intuitivamente indica uma grande repercussão em redes sociais. Ao todo foram selecionadas 8 notícias, que listamos a seguir com um identificador, e as informações que permitem encontrá-las, isto é, o portal de origem e o título da notícia.

[1] `https://www.uol.com.br/`
[2] `https://www.folha.uol.com.br/`
[3] `https://g1.globo.com/`

- **N_1**, G1: ”Sequestrador de ônibus é morto por atirador de elite na Ponte Rio-Niterói, os 39 reféns passam bem”; UOL: ”Sequestrador é morto, polícia libera reféns e encerra sequestro de ônibus”; Folha de São Paulo: ”PM mata homem que manteve passageiros de ônibus reféns na ponte Rio-Niterói”.
- **N_2**, G1: ”Menina de 8 anos morre baleada no Complexo do Alemão”; UOL: ”Menina de 8 anos morre baleada após operação policial no Complexo do Alemão”; Folha de São Paulo: ”Menina de oito anos morre baleada no Rio de Janeiro”.
- **N_3**, G1: ”Mulher é imobilizada por PMs com bebê no colo em Itabira, MG”; UOL: ”Policial imobiliza mulher com criança no colo em Itabira (MG)”; Folha de São Paulo: ”PMs derrubam e imobilizam mulher com bebê no colo em Itabira (MG)”.
- **N_4**, G1: ”Lázaro Barbosa morre após ser preso em Goiás”; UOL: ”Lázaro é morto em Goiás; policiais comemoram após carregar corpo”; Folha de São Paulo: ”Lázaro Barbosa, o serial killer do DF, é morto pela polícia após 20 dias de buscas”;
- **N_5**, G1: ”Menino de 14 anos morre durante operação das polícias Federal e Civil no Complexo do Salgueiro, RJ”; UOL: ”Adolescente João Pedro é morto em operação no Rio, família critica polícia”; Folha de São Paulo: ”Menino de 14 anos é morto em casa durante ação da PF no Rio”.
- **N_6**, G1: ”Menino de 7 anos morre após ser baleado na porta de casa na Baixada Fluminense”; UOL: ”Menino de 7 anos morre após ser baleado enquanto brincava na porta de casa”; Folha de São Paulo: ”Menino de 7 anos morre na porta de casa após tiroteio na Baixada Fluminense”.
- **N_7**, G1: ”Homem morre após ser baleado em ação do Exército na Zona Oeste do Rio”; Folha de São Paulo: ”Exército dispara 80 tiros em carro de família no Rio e mata músico”.
- **N_8**, G1: ”Operação no Jacarezinho deixa 28 mortos, provoca intenso tiroteio e tem fuga de bandidos”; UOL: ”Operação com 25 mortos no Jacarezinho é a mais letal da história do Rio”; Folha de São Paulo: ”Polícia faz operação mais letal da história do RJ, com ao menos 25 mortos”.

A próxima etapa da coleta de dados consiste em obter comentários das pessoas sobre as notícias acima relacionadas em redes sociais. Utilizamos os títulos e os links de cada notícia para coletas na rede social *Twitter*. Optamos por essa rede devido aos recursos oferecidos à pesquisadores para coleta de dados e o foco em conteúdo de texto, i.e., a maioria dos *tweets* são textos curtos limitados em 280 caracteres atualmente. Para coletar *tweets* utilizamos a API do Twitter versão 2 com a biblioteca *Tweepy* na linguagem Python[4]. Essa biblioteca funciona como um facilitador para acessar a API, tornando possível buscar *tweets* por palavras chaves (i.e., título) e links das notícias. Ao fim dessa etapa, coletamos um total de 16 276 *tweets* sobre todas as notícias.

3.2. Pré-processamento e Rotulação

Conduzimos o mínimo de pré-processamentos de textos nos *tweets* coletados, seguindo recomendações de outros trabalhos que lidaram com inferência via processamento de linguagem natural [Mozafari et al. 2019]. Assim, removemos dados que dificultam esse processamento como *links* e quebras de linha. Adicionalmente, desconsideramos *tweets* com apenas uma palavra, dado o menor potencial de inferência semântica desses. Por outro lado, mantivemos todas as *stop words* para treinar modelos de inferência com a

[4]https://docs.tweepy.org/en/stable/

mesma sequência em que as palavras aparecem nos *tweets*. Todas essas tarefas de pré-processamento foram realizadas com a biblioteca NLTK[5] na linguagem Python.

Após o pré-processamento, realizamos a etapa de rotulação das instâncias de dados. O rótulo consiste na classificação de um *tweet* como *Aprova, Desaprova* ou *Neutro* considerando o posicionamento do usuário que o postou sobre a ação policial no incidente ao qual a notícia se refere. Em outras palavras, o *tweet* com o rótulo *Aprova* significa que o usuário aprova a ação policial, *Desaprova* significa que o usuário não aprova a ação, por sua vez, *Neutro* expressa que o usuário é indiferente ao assunto ou não manifesta claramente seu posicionamento. A rotulação serve de base para treinamento de modelos de inferência sendo realizada por humanos, como é usual na literatura. Nesse sentido, contamos com três pessoas (não especialistas[6]) para rotular um subconjunto de *tweets*, dado a inviabilidade de rotulação da base de dados completa. Esse subconjunto é composto por 4467 *tweets* definidos aleatoriamente e proporcionalmente ao total de *tweets* por notícia.

Avaliamos a qualidade da rotulação via métricas de concordância entre os três rotuladores. Nesse sentido, selecionamos outro subconjunto de 198 *tweets* aleatórios com rotulação das três pessoas. O percentual de concordância entre os rotuladores foi de 70,2%, ao passo que o índice *Fleiss Kappa* foi de 0,67, indicando uma concordância substancial entre os pesquisadores em comparação a outros trabalhos da literatura [Hossain et al. 2020].

As informações sobre a distribuição dos rótulos por notícia são mostradas na Figura 1. A distribuição das classes é fortemente desbalanceada, contendo principalmente *tweets* com postura neutra que representam 49% do conjunto total, seguidos por *tweets* da classe *desaprova* representando 27% do conjunto total, e por fim, é a classe *aprova* com 24%. Esse desbalanceamento, principalmente, em direção a classe neutra torna a inferência de posicionamentos mais difícil como já observado em outros trabalhos da literatura [Mozafari et al. 2019, Hand and Ching 2020]. Além disso, nossa proposta é ainda mais desafiadora, pois lidamos com comentários em português brasileiro em contextos locais envolvendo incidentes de violência envolvendo ação policial, em que nem sempre as ferramentas e modelos apresentam o melhor resultado.

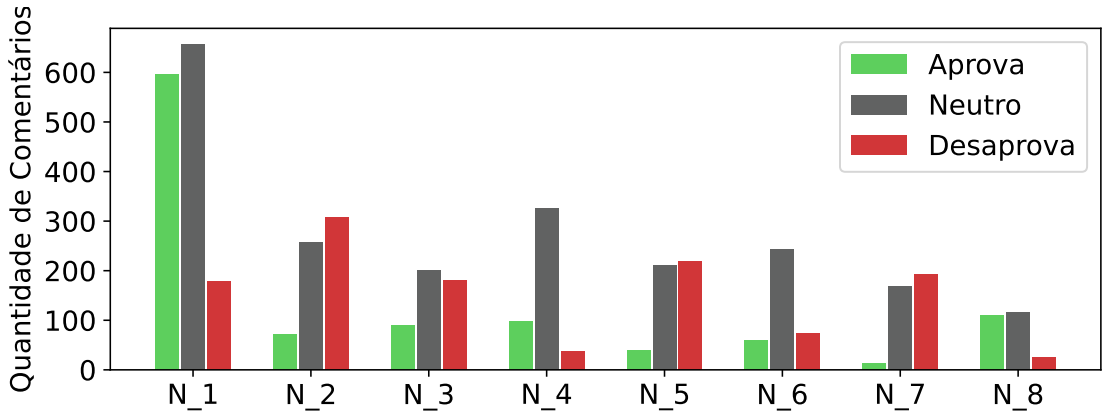

Figura 1. Distribuição dos rótulos por notícia.

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

[6]Entendemos que o contexto das notícias é conhecido pela população amplamente, sendo desnecessário rotulação por um especialista.

4. Metodologia

Nesta seção apresentamos, primeiramente, os modelos para inferência de posicionamento das pessoas em relação à atuação policial. A seguir descrevemos a metodologia e métricas utilizadas para avaliar o desempenho desses modelos.

4.1. Modelos

O BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al. 2019] é um algoritmo de aprendizado profundo (do, inglês *Deep Learning*) desenvolvido pelo *Google* para processamento de linguagem natural. BERT é um modelo pré-treinado com conteúdo do Wikipédia e o Book Corpus, ambos em língua inglesa, contendo 2.500 milhões e 800 milhões de palavras respectivamente. Um recurso importante do BERT é sua construção com representações contextuais de modelos pré-treinados, baseado na arquitetura de *transformers*. Através desse recurso, comunidades de pesquisadores ou corporações podem treinar novos modelos adaptados a alguma linguagem no contexto local e distribuí-los como modelos pré-treinados BERT.

Neste trabalho utilizamos dois modelos pré-treinados que são o *Multilingual* [Devlin et al. 2019] e o *BERTimbau* [Souza et al. 2020]. O modelo *Multilingual* é pré-treinado em 102 idiomas, e possui 12 camadas (blocos de *transformers*), 12 cabeçotes de atenção e 110 milhões de parâmetros. No que lhe concerne, BERTimbau é pré-treinado na língua portuguesa do Brasil, e em nossas avaliações utilizamos a versão desse modelo com a maior quantidade de codificadores (*large*), contendo 24 camadas, 16 cabeçotes de atenção e 335 milhões de parâmetros. Embora treinado para uma língua específica, BERTimbau já se mostrou eficiente em tarefas de processamento de linguagem natural que requerem reconhecimento de entidade nomeada, semelhança textual de sentença e reconhecimento de enlace textual [Souza et al. 2020].

Utilizamos matrizes de *word embeddings* extraídas do BERT para treinar classificadores supervisionados na inferência de posicionamento de *tweets*. Em suma, *Word Embedding* é uma forma de representar palavras através de números para processamento de linguagem natural. Essa representação é normalmente na forma de um vetor de valores reais, que representa o significado das palavras conforme o contexto e o significado da sentença, i.e., o *tweet* em que a palavra está inserida. Para gerar essa representação, utilizamos os modelos pré-treinados acima descritos. O BERTimbau *large* gera um vetor de tamanho 1024 para cada sentença, ao passo que o Multilingual gera um vetor de tamanho 768. Ao fim temos uma matriz de *word embeddings* em que cada linha representa um *tweet* e cada coluna representa uma característica do *tweet*.

A matriz de *words embeddings* foi utilizada como as características de um conjunto de *tweets* rotulados para treinar dois classificadores populares que são *Random Forest* (RF) e *Support Vector Machine* (SVM) [Mahesh 2020]. RF combina um conjunto de preditores de árvores de modo que cada árvore depende dos valores de um vetor aleatório da amostragem com a mesma distribuição. SVR busca predizer um valor real após traçar duas retas paralelas, chamadas limites. O modelo ainda traça uma reta linear entre as duas outras retas retas de modo a ajustar seus valores.

Além dos modelos de classificação acima, avaliamos outra abordagem baseada em ajustes finos diretamente sob os modelos BERT. Especificamente, adicionamos uma nova

camada de neurônios aos modelos BERT pré-treinados e a treinamos para nossa tarefa de classificação com os *tweets* rotulados. Esse ajuste tem a vantagem de acrescentar novas informações aos modelos, que já codificam muitas informações sobre a língua em que foram pré-treinados. Dessa forma pode-se obter melhores resultados para inferência de posicionamentos com menor quantidade de dados rotulados para treinamento, apesar do tempo e complexidade para construção de mais uma camada na rede neural do BERT.

4.2. Configurações, Treinamento e Métricas de avaliação

Como ambiente de experimentação utilizamos uma máquina virtual da AWS do tipo G4dn que são instâncias que usam GPUs otimizadas para aprendizagem profunda. Logo, ajustamos os modelos pré-treinados BERTimbau e Multilingual com 10 épocas e otimizador Adam com os respectivos *learning rate* e *batch size* de cada modelo. Para a avaliação da rede neural ajustada, consideramos 80% dos *tweets* rotulados para treino e 20% para teste, treinamos esse modelo em cada época, e ao fim de cada época avaliamos o seu desempenho. Para a abordagem com os classificadores, geramos a matriz de *word embedding* com os modelos BERT pré-treinados, em seguida dividimos a matriz em treino e teste com 80% e 20% dos *tweets* respectivamente, e treinamos os classificadores sem e com balanceamento de classes utilizando a técnica *smote* [Mahesh 2020]. O código fonte, bem como a base de dados, deste trabalho estão disponibilizados no GitHub [7].

Avaliamos os modelos com o conjunto de teste utilizando cinco métricas, que são acurácia, precisão, revocação, f1-score e f1-macro, para cada classe (Desaprova, Aprova, Neutro). A acurácia indica o percentual de *tweets* corretamente classificados, isto é, a soma acertos de todas as classes dividido pelo número total de *tweets* classificados. Já a precisão é calculada para cada classe individualmente e evidencia o percentual de *tweets* corretamente classificados para aquela classe. A revocação é calculada justamente pelo total de *tweets* corretamente classificados para uma classe sobre o total de *tweets* 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.

5. Resultados

Nesta seção apresentamos os resultados obtidos pelos modelos descritos anteriormente na inferência de posicionamentos sobre ações policiais no conjunto de teste dos *tweets*. Adicionalmente, analisamos os erros desses modelos para mostrar a dificuldade da inferência e possíveis estratégias para lidar com os erros.

5.1. Classificação de Posicionamentos

Na Tabela 1, apresentamos os resultados para a avaliação de desempenho dos classificadores treinados sem e com o tratamento para desbalanceamento de classes (*smote*), além da rede neural BERT com ajuste fino. Para cada um desses modelos, apresentamos os resultados via cinco métricas de desempenho. Nota-se que a acurácia (acc) é a métrica, majoritariamente, com maior desempenho para as classes desaprova e aprova. Contudo, é importante observar que os dados analisados são desbalanceados e essas classes são minoritárias na maior parte das notícias analisadas (i.e., N_1, N_3, N_4, N_6 e N_8). Nesse

[7] https://github.com/LABPAAD/crimes_stance

caso, as métricas que capturam a especificidade (precisão, revocação e f1) dos modelos em identificar corretamente a classes minoritárias ganham importância, e essas métricas são o foco da avaliação de desempenho nessa seção.

Tabela 1. Desempenho dos modelos Multilingual (MultiL) e BERTimbau (PtBr), usando as estratégias de ajuste fino da rede neural (RN) e a 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.

Modelos	Desaprova			Aprova			Neutro			Acc	F1-macro
	P	R	F1	P	R	F1	P	R	F1		
SVM-MultiL-smote	0,46	0,53	0,49	0,48	0,59	0,53	0,76	0,63	0,69	0,59	0,57
SVM-MultiL	0,51	0,49	0,50	0,55	0,47	0,51	0,71	0,77	0,74	0,63	0,58
RF-MultiL-smote	0,50	0,48	0,49	0,57	0,50	0,54	0,72	0,77	0,74	0,63	0,59
RF-MultiL	0,52	0,34	0,41	0,61	0,38	0,47	0,64	0,86	0,73	0,62	0,54
SVM-PtBr-smote	0,52	0,65	0,58	0,55	0,66	0,60	0,84	0,66	0,74	0,66	0,64
SVM-PtBr	0,54	0,58	0,56	0,66	0,61	0,63	0,78	0,77	**0,77**	0,69	**0,66**
RF-PtBr-smote	0,55	0,56	0,56	0,68	0,56	0,61	0,74	0,79	0,77	0,68	0,65
RF-PtBr	0,55	0,44	0,49	0,75	0,49	0,59	0,69	0,86	0,76	0,67	0,61
RN-PtBr	0,58	0,63	**0,61**	0,69	0,68	**0,69**	0,77	0,74	**0,76**	0,70	**0,68**
RN-MultiL	0,49	0,55	0,52	0,61	0,53	0,56	0,71	0,71	0,71	0,63	0,60

Primeiramente, discutimos os resultados do tratamento de desbalanceamento das classes para os classificadores RF e SVM. Sem esse tratamento, há um viés para precisão nas classes minoritárias, pois apenas *tweets* com posicionamentos claros de aprovação ou desaprovação serão classificadas como tal. Isso pode ser observado na Tabela 1, com valores de precisão majoritariamente maiores que revocação, levando a baixa especificidade e tendência a classificar posicionamentos como neutros. Ao tratarmos o desbalanceamento com *smote*, observamos que SVM e RF, em geral, aumentam o desempenho para revocação e, por consequência, aumentam o acerto para as posturas de aprovação e desaprovação. A métrica F1 é a referência para o melhor compromisso entre precisão e revocação, e obtemos F1 maiores para os classificadores com *smote* (exceção apenas de SVM-MultiL para Aprova e SVM-PtBR para Desaprova). Observa-se também que o SVM, em geral, tem o melhor desempenho alcançando F1 melhores (49-63%) com ou sem balanceamento, superando as marcas de F1 do classificador RF (41-61%).

Agora focamos no aspecto da língua em que os modelos BERT são pré-treinados, i.e., multi-lingual (MultiL) e português brasileiro (PtBr). Para isso utilizamos a métrica F1-macro que considera a média do F1 para as três classes de posicionamentos. Observa-se que o pré-treinamento na língua local dos *tweets* (português brasileiro) tem ganhos de desempenho notáveis para os três modelos avaliados. Ao comparar os ganhos dos modelos PtBr em relação a MultiL, os classificadores RF e SVM alcançam aumentos em F1-macro de até 7 e 8 pontos percentuais respectivamente, ao passo que a rede neural ajustada (RN) tem um aumento de 8 pontos percentuais. Isso evidencia que o treinamento de modelos com textos no domínio e língua específicas do contexto em investigação são aspectos fundamentais para inferências de posicionamentos. Essa observação corrobora com outros trabalhos recentes que também utilizam *tweets* para inferências de posicionamentos, mas em contextos diferentes ao investigado nesse trabalho [Hossain et al. 2020].

Finalmente, discutimos o uso da rede neural BERT com ajuste fino (RN). Como esperado, esse modelo obteve o melhor resultado. Especificamente, o modelo pré-treinado em português brasileiro (RN-PtBr) alcançou o F1-macro de 68% e uma acurácia de 70%. É importante mencionar, todavia, que o desempenho da rede neural não é tão superior ao

melhor classificador (SVM-PtBr). O ganho do RN-PtBr corresponde a apenas 2 pontos percentuais sobre o F1-macro do SVM-PtBr. Considerando o custo computacional para ajustar a rede neural BERT (i.e., o dobro do tempo dos classificadores), o modelo SVM apresenta um bom compromisso entre desempenho e custo de treinamento.

5.2. Análise dos Erros de Classificação

A Figura 2 reporta a matriz de confusão para o modelo RN-PtBr que obteve o melhor desempenho, mostrado na Tabela 1. Cada linha representa os *tweets* em uma classe real, enquanto cada coluna representa os *tweets* em uma classe prevista, o que nos permite analisar onde o classificador mais erra e como isso acontece em função da classe.

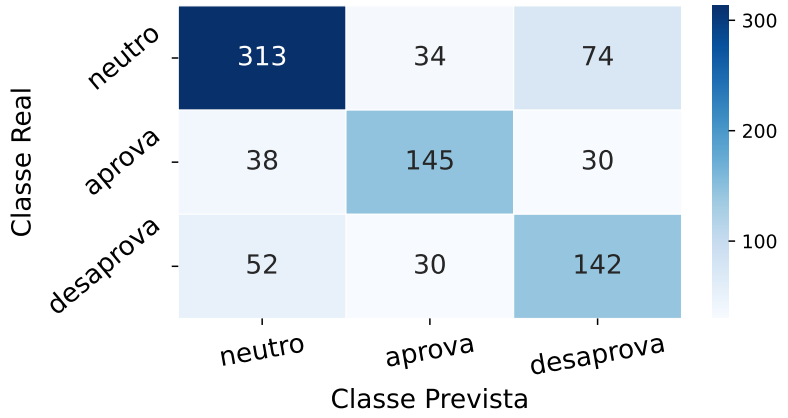

Figura 2. Matriz de confusão, modelo BERTimbau com ajuste fino (RN-PtBr).

De forma geral, é possível notar que, em termos absolutos, o maior desafio na tarefa aqui endereçada está em diferenciar as classes *Neutro* da *Desaprova* e vice-versa. Para exemplificarmos a complexidade da tarefa, nós selecionamos alguns casos a partir da matriz de confusão focando especificamente nessas duas classes e apresentamos na Tabela 2.

Tabela 2. Exemplos de *tweets* classificados incorretamente pelo modelo BERTimbau com ajuste fino (RN-PtBr).

Tweet	Real	Predito
Eu já vi isso mano, é osso	Neutro (0)	Desaprova (-1)
Meu Deus que realidade cruel, que os passageiros fiquem bem	Neutro (0)	Desaprova (-1)
Isso foi de propósito?	Neutro (0)	Desaprova (-1)
É esse o nível de quem vai nos salvar do comunismo quero saber quem vai nos salvar deles	Desaprova (-1)	Neutro (0)
#ficaemcasa e leve um tiro	Desaprova (-1)	Neutro (0)
Não é saudável e nem natural se acostumar com tragédias, as pessoas estão doentes!	Desaprova (-1)	Neutro (0)

Os exemplos apresentados mostram como, *tweets* 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 representam uma larga fração dos *tweets* classificados incorretamente.

6. Conclusões e Trabalhos Futuros

Neste trabalho, nós propomos inferir o posicionamento das pessoas em relação a operações policiais, um contexto específico dentro do tema de segurança pública e que compreende um dos principais desafios da sociedade. Para isso, nós definimos um conjunto de notícias que tomaram dimensões relativamente grandes devida à atuação da polícia no Brasil a fim de avaliar a percepção das pessoas por meio de dados de redes sociais, especificamente, do *Twitter*.

Considerando os desafios presentes em atividades de processamento natural de linguagem, nós propomos o uso das arquiteturas *transformers* para obter uma representação semântica de baixa dimensão (*embeddings*) e endereçamos o problema por meio da tarefa de classificação. Dessa forma, nossa ideia é compreender como o posicionamento dos usuários podem ser capturados e, futuramente, generalizados para outras aplicações. Nossos resultados que modelos BERT pré-treinados em língua local obtiveram melhor desempenho para inferir posicionamentos da população sobre atuação policial expressas em tweets, ao passo que técnicas de balanceamento de classes tendem a melhorar esse desempenho para classificadores treinados com *word embeddings*. Adicionalmente, classificadores como SVM tem uma boa relação entre custo e benefício se comparado a um modelo BERT com rede neural ajustada especificamente para os dados.

Como trabalhos futuros, pretendemos ampliar nossa avaliação incluindo outros classificadores e redes neurais ajustadas com mais camadas de neurônios. Adicionalmente, planejamos explorar características extraídas de tweets via modelos BERT também para predição de taxas de crimes.

Referências

Al-Garadi, M. A., Varathan, K. D., Ravana, S. D., Ahmed, E., Mujtaba, G., Khan, M. U. S., and Khan, S. U. (2018). Analysis of online social network connections for identification of influential users: Survey and open research issues. *ACM Computing Surveys (CSUR)*, 51(1):1–37.

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.

Chen, X., Cho, Y., and Jang, S. Y. (2015). Crime prediction using twitter sentiment and weather. In *Systems and Information Engineering Design Symposium*, pages 63–68. IEEE.

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.

Ferreira, C. H., Murai, F., Silva, A. P., Almeida, J. M., Trevisan, M., Vassio, L., Mellia, M., and Drago, I. (2021). On the dynamics of political discussions on instagram: A network perspective. *Online Social Networks and Media*, 25:100155.

Glandt, K., Khanal, S., Li, Y., Caragea, D., and Caragea, C. (2021). Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1.

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*.

Iranmanesh, A. and Alpar Atun, R. (2020). Reading the urban socio-spatial network through space syntax and geo-tagged twitter data. *Journal of Urban Design*, 25(6):738–757.

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

Malagoli, L. G., Stancioli, J., Ferreira, C. H., Vasconcelos, M., Couto da Silva, A. P., and Almeida, J. M. (2021). A look into covid-19 vaccination debate on twitter. In *ACM Web Science Conference 2021*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2019). A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

NEV-USP (2021). Monitor da violência. Disponível em: https://nev.prp.usp.br/projetos/projetos-especiais/monitor-da-violencia/. Acesso em 07 de jun. 2021.

Nobre, G. P., Ferreira, C. H., and Almeida, J. M. (2022). A hierarchical network-oriented analysis of user participation in misinformation spread on whatsapp. *Information Processing & Management*, 59(1):102757.

Oglesby-Neal, A., Tiry, E., and Kim, K. (2019). Public perceptions of police on social media. *Washington, DC: Urban Institute*.

Reimers, N., Gurevych, I., Reimers, N., Gurevych, I., Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I., Reimers, N., Gurevych, I., et al. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Reis, J., de Souza, F., de Melo, P. V., Prates, R., Kwak, H., and An, J. (2015). Breaking the news: First impressions matter on online news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 357–366.

Sousa, D. d. S., Feitosa, M. P. F., and Gonçalves, G. D. (2021). Relações entre crimes e o espaço urbano: Um estudo de caso baseado em pontos de interesses extraídos da web. In *Anais do V Workshop de Computação Urbana*, pages 196–208. SBC.

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, H., Yao, H., Kifer, D., Graif, C., and Li, Z. (2017). Non-stationary model for crime rate inference using modern urban data. *IEEE transactions on big data*, 5(2):180–194.

Wang, X., Gerber, M. S., and Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer.

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*.