XXIX Congresso de Iniciação Científica UNICAMP 2021

Base de Comentários de Discurso de Ódio contra Mulheres MINA-BR: da Concepção aos Ataques por Robôs

Hannah de Oliveira Plath, Maria Estela de Oliveira Paiva, Danielle Lanzarini Pinto, Paula Dornhofer Paro Costa
Depto. Eng. de Computação e Automação (DCA), Faculdade de Eng. Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, Brasil
e-mail: h198642@dac.unicamp.br, paulad@unicamp.br

***Resumo*—As redes sociais criaram um espaço onde as pessoas podem se manifestar publicamente de forma anônima sem precisarem, tipicamente, se responsabilizarem por suas intervenções, tornando essas plataformas um ambiente propício à proliferação de discurso de ódio. O desenvolvimento de algoritmos capazes de detectar automaticamente esse tipo de discurso é importante para combater essa forma de violência e melhorar a regulamentação desses espaços. Este trabalho foca no discurso de ódio contra mulher e descreve o processo de construção da primeira base de dados de comentários extraídos da Internet em português brasileiro, rotulada por voluntários que analisaram a presença, ou não, de discurso de ódio, nos comentários selecionados. A base MINA-BR já tem viabilizado o treinamento e avaliação dos primeiros modelos classificadores de discurso de ódio, com 6002 comentários extraídos da Internet, dos quais 2135 já foram rotulados e 16,26% foram classificados como discurso de ódio. O artigo descreve a metodologia de construção dessa base desde a sua concepção até um ataque por bots, ou robôs, sofrido pelo projeto durante a fase de rotulação da base.**

***Palavras-chave*—base de dados, discurso de ódio, misoginia**

I. Introdução

As redes sociais são um espaço público-privado onde usuários podem interagir e se manifestar virtualmente. Embora essas plataformas tentem impor regras nesses espaços através dos seus termos e políticas de uso, a falta de confronto face-a-face bem como a possibilidade de anonimização de perfis — traços característicos desse tipo de plataforma — contribuem para a formação de um sentimento de liberdade irrestrita e impunidade nesses ambientes. Além disso, o advento de programas de computador capazes de gerar e replicar comentários utilizando perfis falsos contribui para a propagação de mensagens em massa nessas plataformas, proliferando ideias que, sem o uso desses "bots", ou "robôs", não teriam um engajamento ou um alcance tão grande. Todas essas características tornam as redes sociais um ambiente propício à propagação de discursos de ódio.

O discurso de ódio pode ser definido como uma forma de fala pública que incita violência, ódio ou assédio contra uma pessoa ou grupo devido à sua raça, religião, gênero ou orientação sexual [1]. Nos últimos anos, na área de processamento de linguagem natural (em inglês *Natural Language Processing*, ou NLP), houve um aumento no número de pesquisas tratando do desenvolvimento de modelos de detecção automática desse tipo de discurso [2]. O desenvolvimento desses modelos depende diretamente da construção de bases de dados, geralmente rotuladas, contendo discurso de ódio. No entanto, a maioria das bases existentes têm como foco a língua inglesa [3]. Em português, é de conhecimento do grupo a existência de apenas duas bases de discurso de ódio.

A primeira, criada por Fortuna et al. [4] é formada por comentários extraídos do Twitter e não possui foco específico em nenhum tipo de discurso de ódio. Já a base construída por de Pelle e Moreira [5] é formada por comentários extraídos do site de notícias G1 e não possui um volume expressivo de comentários sexistas, homofóbicos ou racistas, seu uso não sendo recomendado para pesquisas nesses tópicos.

A baixa diversidade linguística das bases é um problema, uma vez que a detecção de discurso de ódio é uma tarefa fortemente ligada ao idioma do texto. Em seu trabalho, Corazza et al. [6] aponta que quando uma mesma técnica é adotada para classificar discurso de ódio em diferentes línguas, os modelos obtidos podem produzir resultados distintos. De forma análoga, vocabulários utilizados em diferentes categorias de discurso de ódio (racismo, sexismo, etc) variam. Essas diferenças não devem ser ignoradas no contexto da detecção automática de discurso de ódio.

Outra dificuldade encontrada é que o processo de rotulação de comentários é feito, na maioria das vezes, manualmente e, portanto, demanda tempo e recursos humanos. Uma alter-

Este trabalho foi financiado pelo Programa Institucional de Bolsas de Iniciação Científica (PIBIC), CNPq.

nativa para essa questão é a criação de ferramentas online e websites que atuam como plataformas de rotulação e permitem participação do público geral no processo. A implementação dessa abordagem deve ser feita, no entanto, tomando-se medidas para proteger a base de ataques maliciosos, principalmente aqueles feitos por meio de robôs. Esses respondem automaticamente as perguntas da pesquisa e, além de interferirem nos resultados, podem também prejudicar o funcionamento do website. Em Londres, por exemplo, o site utilizado para agendamento de vacinas contra do COVID-19 foi alvo de um desses ataques e teve dificuldades em operar por um período de tempo[1].

Neste artigo será relatado o processo de construção da base de dados MINA-BR bem como um ataque feito à base durante seu processo de rotulação.

A base MINA-BR é a primeira base de dados em português brasileiro com foco em discurso de ódio contra a mulher e visa se tornar uma base de dados rotulada, com comentários retirados da Internet de amplo acesso para fins de pesquisa. Este trabalho foi realizado no contexto do projeto MINA-BR que visa o desenvolvimento de ferramentas baseadas em aprendizado de máquina capazes de detectar automaticamente discursos de ódio contra a mulher em textos.

Nas últimas décadas, com os avanços legais e sociais que o mundo viveu, o discurso misógino aberto contra a mulher migrou do âmbito do mundo real para o virtual, se manifestando principalmente na forma de assédio sexual cibernético. Tal prática impacta gravemente na vida pessoal e profissional das vítimas, havendo casos onde essa violência transcende a esfera da Internet e pode resultar em casos de agressão física e estupro [7]. No Brasil, mais de 500 mulheres sofrem algum tipo de violência por hora e ao menos uma mulher é assassinada a cada duas horas, um problema que foi agravado pelo isolamento social e as dificuldades econômicas decorrentes da pandemia do COVID-19 [8].

Espera-se que os resultados obtidos por meio essa pesquisa possam ajudar no desenvolvimento de aplicações que atuem no combate da violência contra a mulher no Brasil.

II. Metodologia

A. Extração de comentários

O primeiro passo para a construção da base foi um estudo exploratório de fontes da Internet para a extração de comentários.

Analisando a literatura sobre o tema, encontra-se que a maioria das bases de dados construídas para estudos sobre a detecção automática de discurso de ódio possuem seus textos extraídos de redes sociais e, ocasionalmente, de sites de notícias, sendo que a principal plataforma usada é o Twitter [3]. A fim de se diversificar as fontes de extração, o grupo elaborou uma pesquisa online para o levantamento de possíveis fontes de comentários de discurso de ódio. A pesquisa foi divulgada para alunos da Unicamp e através das redes sociais. Ao todo 53 pessoas responderam no decorrer de duas semanas. A Figura 1 contém os resultados dessa pesquisa.

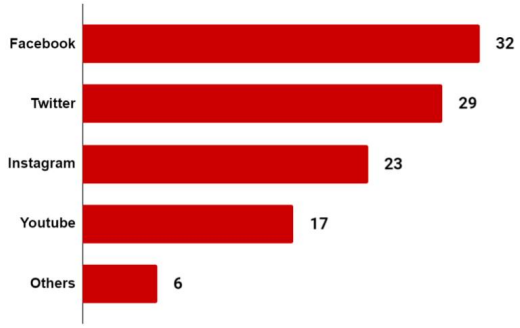

Figura. 1: Resultados da pesquisa online para o levantamento de possíveis fontes de comentários de discurso de ódio.

A pesquisa realizada apontou o Facebook como a plataforma onde as pessoas mais percebem o discurso de ódio contra a mulher, seguida pelo Twitter e pelo Instagram. Devido, no entanto, à característica privada da maioria dos perfis e páginas do Facebook, bem como da indisponibilidade de ferramentas para realizar a extração automática de comentários na API (do inglês *Application Programming Interface*) do Instagram, essas plataformas não foram utilizadas no trabalho. Assim, somente o Twitter e o Youtube foram adotados como fontes para a pesquisa. A extração de comentários dessas plataformas foi feita através do uso de suas respectivas APIs.

Quanto ao método de extração de comentários, a maioria das bases existentes de discurso de ódio utilizam a busca por palavras-chaves [3]. No entanto, a literartura não evidencia uma metodologia clara para a seleção desses termos. Alguns trabalhos citam o uso de palavras-chaves relacionadas à classe do discurso de ódio de interesse [9], outros documentam que os termos usados foram retirados de repositórios online como o HateBase[2] [10]. No entanto, o uso deste último método não foi viável no contexto deste projeto visto que existem poucos repositórios em língua portuguesa que contém termos relacionados ao discurso de ódio contra a mulher.

Devido a essas razões, adotou-se como plano inicial a utilização da busca via palavras-chaves em conjunto com o monitoramento de perfis de potenciais vítimas e agressores e decidiu-se realizar um levantamento de palavras de busca, perfis e canais para serem usados na pesquisa. Essa investigação foi feita através da mesma pesquisa online citada anteriormente. Algumas das palavras-chaves obtidas foram “feminazi”, “mal comida”e “abortista”, enquanto perfis de cantoras, políticas e atrizes foram indicados como possíveis vítimas de discurso de ódio.

Foram feitos testes de extração utilizando os diferentes métodos escolhidos. Em particular, notou-se que a retirada de comentários via perfis de possíveis vítimas retornava um volume baixo de mensagens de ódio. Essa quantidade se

[1]https://www.securitymagazine.com/articles/94887-bots-attack-london-vaccine-appointments

[2]https://hatebase.org/

tornava relevante para a extração de dados somente quando essas mulheres estavam envolvidas em alguma notícia de destaque. Como esses eventos são imprevisíveis, o monitoramento de perfis de possíveis vítimas foi usado como estratégia secundária para a construção da base.

Levando isso em consideração, as palavras usadas para a busca dos comentários para a construção da base MINA-BR foram, a partir das respostas retornadas pela pesquisa online, selecionadas. Além disso, durante um período de duas semanas, foram também monitorados os tópicos mais populares do Twitter no Brasil, a fim de se buscar temas relacionados à mulher que estavam sendo discutidos no momento.

Após a extração dos textos, fez-se uma limpeza sobre os comentários, retirando-se textos em outras línguas, comentários que continham *links* para outros vídeos/imagens e duplicatas [11]. Eliminou-se também comentários que continham apenas uma palavra, pois entendeu-se que nesse caso não haveria contexto o suficiente para a classificação.

Por fim, as bases do Twitter e do Youtube foram concatenadas de forma que a base final MINA-BR tivesse um número igual de amostras de cada plataforma. O tamanho final da base é de 6002 comentários.

B. Rotulação da base

Para auxiliar no processo de rotulação da base, foi desenvolvido um website[3] que permite que voluntários(as) classifiquem os comentários.

Uma das questões consideradas pelo projeto é o fato de que a definição de discurso de ódio não é amplamente disseminada e pode ser subjetiva, dependendo de aspectos culturais, assim como de raça, gênero e faixa etária. Para tratar dessa questão, três medidas principais foram tomadas.

A primeira medida consiste na tentativa de uniformizar o entendimento dos participantes quanto à definição de discurso de ódio. Assim, antes de iniciarem a rotulação no website, os voluntários são convidados a lerem diferentes definições de discurso de ódio provindas da literatura bem como são apresentados a leis brasileiras que tratam sobre o tema.

A segunda medida consiste em realizar a rotulação através de um sistema de duas etapas. Em um primeiro momento, os participantes são questionados se o comentário em questão é ofensivo ou não e, em caso afirmativo, devem dizer se o texto contém discurso de ódio contra a mulher ou não. Além disso, os participantes devem dar o grau de certeza da sua resposta utilizando uma escala de Likert com cinco níveis de certeza em ambas as etapas da classificação. Cada participante foi convidado a realizar dez rotulações.

Por fim, a última medida consiste na rotulação de cada comentário por três pessoas diferentes, visando garantir que as amostras rotuladas como ódio fossem àquelas contendo inequívocos traços de ódio.

Além disso, foi feita uma coleta de dados sobre os rotuladores. Antes de iniciarem o processo de classificação, eles tiveram que responder um questionário demográfico informando características suas como faixa etária, grau de formação, sexo, etnia, etc.

A rotulação também foi feita através de lotes. Isto é, os comentários da base foram divididos em grupos, que foram rotulados em série. Assim, apenas quando a rotulação de um lote terminava, era iniciada a rotulação de outro grupo. Essa divisão da base em porções menores diminui a esparcidade das rotulações, evitando que, caso o número de voluntários seja insuficiente, ao final do período de rotulações tenha-se muitos comentários com apenas uma ou duas avaliações e poucos com três.

Além disso, a rotulação em lotes é um mecanismo de proteção da base. Caso a rotulação seja comprometida em algum ponto do processo, apenas os comentários pertencentes àquele grupo terão que ser reclassificados.

O projeto foi divulgado através das redes sociais e na televisão [12] durante o mês de junho de 2021.

III. Resultados

A rotulação dos 6002 comentários da base de dados ainda não foi concluída. O atraso da rotulação se deu em parte a um ataque à base que ocorreu em junho.

Dia 15 de junho de 2021, o grupo foi alertado pela plataforma utilizada para armazenamento da base sobre uma alta transferência de dados, indicando que muitas rotulações estavam sendo feitas em um curto período de tempo. Tal comportamento não era esperado, visto que, mesmo nos dias seguintes à exposição do projeto na televisão tais níveis de transferência não haviam sido atingidos. Além disso, apesar da alta taxa de rotulações feitas, houveram poucas conexões simultâneas ao site. Isso levantou a suspeita de que o projeto teria sido alvo de um ataque realizado através do uso de programas computacionais automatizados.

Essa hipótese foi reforçada por uma mensagem recebida na nossa caixa de sugestões alertando que o website do projeto estava circulando em um grupo radical de extrema-direita que já estava sendo monitorado como canal de discurso de ódio por outros pesquisadores. Fez-se então uma análise sobre a base comprometida a fim de verificar essa hipótese.

Primeiramente, analisou-se o número de rotulações realizadas por cada voluntário. No total 886 participantes realizaram a rotulação dos 6002 comentários. Ressalta-se que, como o sistema de captação de dados é anônimo, dois comentários são considerados rotulados pelo mesmo participante quando não há recarregamento da página entre as classificações. Assim, o número de rotuladores distintos obtidos não corresponde exatamente ao número de pessoas que participaram do processo, mas sim a uma estimativa.

A maioria desses participantes (63%) fizeram apenas um ciclo de rotulação, ou seja, classificaram apenas os dez comentários propostos pela pesquisa. No entanto, alguns rotuladores se destacaram pela quantidade de classificações feitas, sendo essas realizadas durante o período que a base estava "sob ataque". Os quatro voluntários com a maior quantidade de rotulações foram responsáveis por cerca de 29% do número

[3]https://mina-br.netlify.app/

total de rótulos dados e atribuíram a todos os comentários designados a eles o rótulo de não ódio.

O impacto dessas rotulações pode ser visto também ao se comparar a distribuição dos rótulos ódio e não-ódio nos subconjuntos de comentários da base rotulados antes e após o ataque. Essa comparação está mostrada na Tabela I. Ressalta-se que cada subgrupo possui 2000 comentários e que foi utilizado o critério de maioria absoluta para atribuir o rótulo final de cada amostra.

Tabela I: Distribuição de Rótulos dos Comentários Classificados Antes e Após o Ataque

	ANTES	**APÓS**
Ódio	16,5%	0%
Não Ódio	83,5%	100%
Nº rotuladores	516	31

Comparando a distribuição dos rótulos mostrada na Tabela I, percebe-se que todos os comentários classsificados após o ataque receberam o rótulo de não-ódio. Ademais, embora os dois subgrupos analisados possuam a mesma quantidade de comentários, após o ataque, o número de participantes que participaram da rotulação foi significantemente menor.

Esses dados sustentam a hipótese de que o website foi alvo de um ataque que utilizou robôs para realizar a rotulação de forma automatizada, visando adulterar o resultado da rotulação da base, prejudicando o projeto.

Além disso, analisou-se o perfil gerado por meio do questionário demográfico dos quatro participantes com mais rotulações e todos eles responderam de forma igual às perguntas. As respostas dadas por esses rotuladores estão mostradas na Tabela II.

Tabela II: Respostas do questionário demográfico dadas pelos quatro rotuladores com maior número de rotulações

COR	preta
FAIXA ETÁRIA	$<$ 25 anos
FORMAÇÃO	Ensino Superior Completo
Identidade de Gênero	Não Binário
Sexo	Feminino

Acredita-se que os autores do ataque responderam o questionário demográfico de forma a assumirem um perfil do qual não se fosse esperado respostas coniventes com o discurso de ódio contra a mulher em mais uma tentativa de distorcer os resultados da pesquisa.

Concluiu-se, então, que as rotulações feitas durante esse pico de transferência de dados não eram confiáveis. A rotulação dos lotes comprometidos foi então reiniciada após um aumento na segurança do website a fim de dificultar o uso de programas automatizados na rotulação. O processo de rotulação da base ainda está em andamento.

Até agosto de 2021, 2135 comentários já haviam recebido a rotulação de seus três anotadores. Aplicando o critério de maioria absoluta para a atribuição do rótulo final, a base possui até esse momento 16,26% de suas amostras classificadas como ódio e 83,74% como não-ódio, tal distribuição é coerente com aquelas encontradas em outras bases de discurso de ódio [5], [13]. A Figura 2 mostra a distribuição do perfil dos rotuladores.

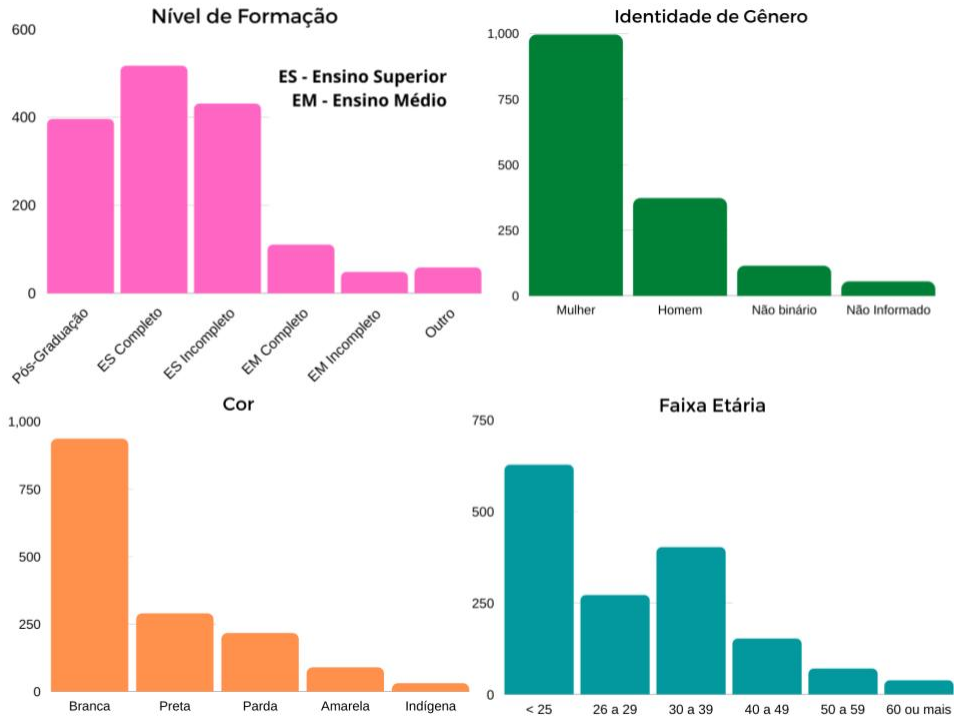

Figura. 2: Gráficos sobre o perfil dos rotuladores gerados a partir do questionário demográfico

Nota-se que a maioria dos rotuladores se identificou como mulher e que a pesquisa atingiu em sua maioria pessoas brancas, com menos de 40 anos e com alto grau de escolaridade.

IV. Conclusão

Neste trabalho foi construída a base de dados MINA-BR com comentários extraídos do Youtube e do Twitter em português brasileiro com foco em discurso de ódio contra a mulher. Embora o processo de rotulação ainda esteja em andamento, 2135 comentários já receberam os rótulos de seus três anotadores. Este trabalho também contém uma descrição precisa da metodologia de construção da base de dados, incluindo o processo de escolha de termos de extração de texto, o que pode ser usado como suporte para a criação de outras bases de dados.

Durante o processo de rotulação da base, o projeto foi alvo de um ataque organizado com o intuito de distorcer a distribuição de rotulações da base e prejudicar o projeto. Ressalta-se que esse ataque não foi a única demonstração de descontentamento com o projeto que o grupo recebeu. Em distintos momentos mensagem contendo xingamentos como "vagabunda" foram deixadas na caixa de sugestões do projeto.

Esse tipo de reação contrária e agressiva à um projeto que visa criar ferramentas que auxiliem no combate contra a violência à mulher demonstra o quanto a discussão sobre esse tema ainda encontra muita resistência e reforça a necessidade de mais projetos que tratem desse assunto.

No futuro, a rotulação da base de dados continuará através do website e espera-se que a base MINA-BR possa ser utilizada para o desenvolvimento de outras pesquisas na área bem como na criação de ferramentas que contribuam para o combate à violência contra a mulher. Um estudo de *benchmark* utilizando algoritmos baseados em aprendizado de máquina para a detecção automática de discurso de ódio contra a mulher já está sendo realizado com a base parcialmente rotulada.

Referências

[1] M. A. Moura, *O discurso do ódio em redes sociais*. Lura Editorial (Lura Editoração Eletrônica LTDA-ME), 2016.

[2] A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella, "Thirty years of research into hate speech: topics of interest and their evolution," *Scientometrics*, vol. 126, no. 1, pp. 157–179, 2021.

[3] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," *Language Resources and Evaluation*, pp. 1–47, 2020.

[4] P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, and S. Nunes, "A hierarchically-labeled Portuguese hate speech dataset," in *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 94–104. [Online]. Available: https://www.aclweb.org/anthology/W19-3510

[5] R. de Pelle and V. Moreira, "Offensive comments in the brazilian web: a dataset and baseline results," in *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. Porto Alegre, RS, Brasil: SBC, 2017. [Online]. Available: https://sol.sbc.org.br/index.php/brasnam/article/view/3260

[6] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A multilingual evaluation for online hate speech detection," *ACM Transactions on Internet Technology (TOIT)*, vol. 20, no. 2, pp. 1–22, 2020.

[7] D. K. Citron, "Misogynistic cyber hate speech," 2011.

[8] G. Bastos, F. Carbonari, and P. Tavares, "Addressing violence against women (vaw) under covid-19 in brazil. world bank group. world bank note," 2020.

[9] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the indonesian language: A dataset and preliminary study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017, pp. 233–238.

[10] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding, "Peer to peer hate: Hate speech instigators and their targets," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018.

[11] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE access*, vol. 6, pp. 13 825–13 835, 2018.

[12] J. da EPTV 2ª Edição Campinas/Piracicaba. (2021) Pesquisa da unicamp busca desenvolver detector de discursos de ódio na internet. [Online]. Available: https://globoplay.globo.com/v/9596916/

[13] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," *arXiv preprint arXiv:1809.04444*, 2018.