Detecção de Discurso de Ódio Contra Mulheres em Textos em Português Brasileiro: Construção da Base MINA-BR e Modelo de Classificação

Hannah O. Plath, Maria Estela O. Paiva, Danielle L. Pinto, Paula D. P. Costa

[1] Depto. Eng. de Computação e Automação (DCA), Faculdade de Eng. Elétrica e de Computação (FEEC) – Universidade Estadual de Campinas (Unicamp) Campinas, Brasil

h198642@dac.unicamp.br, paulad@unicamp.br

Abstract. *Due to the wide use of social networks, among other reasons, hate speech has gained prominence, sometimes motivated by impunity, sometimes associated with freedom of expression. One of the reasons why hate speech recognition is a difficult task is the scarcity of adequate databases, especially in languages other than English or when we refer to a specific domain, such as misogyny. This article describes a database in the Brazilian Portuguese language, which can be useful to classify hate speech against women. This work also reports a preliminary study where established hate speech classification algorithms were used to determine a baseline for the dataset. The highest F1-score obtained was 0.57 by the SVM algorithm.*

Resumo. *A anonimização de perfis, a sensação de liberdade irrestrita e o sentimento de impunidade, têm tornado as redes sociais um ambiente propício para o surgimento de discursos de ódio. A detecção desse tipo de texto é uma tarefa difícil, dentre outras razões, pela baixa variabilidade de bases de dados em idiomas diferentes do inglês. Nesse contexto, este trabalho descreve o processo de construção de uma base de dados em português brasileiro com foco em discurso de ódio contra a mulher. A fim de estabelecer uma baseline para a base criada, apresenta-se também os resultados de um estudo que adota técnicas clássicas de mineração de texto e algoritmos de classificação consagrados na área de NLP. A melhor Medida-F1 obtida neste estudo foi de 0.57 para o algoritmo SVM.*

1. Introdução

As redes sociais criaram um espaço onde os usuários, de maneira anônima, podem se manifestar publicamente, sem nunca serem confrontados ou constrangidos pessoalmente pelas suas intervenções. Ainda, o advento de "robôs" — aqui caracterizados como programas de computador capazes de gerar e replicar comentários de maneira automática usando perfis falsos nas redes sociais — têm contribuído para que pequenos grupos de pessoas possam propagar ideias que possivelmente não seriam manifestadas fora da "ágora" virtual de forma massiva. Assim, as redes sociais podem ser consideradas um meio propício à propagação 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 [Moura 2016]. 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 [Tontodimamma et al. 2021]. 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 [Poletto et al. 2020], havendo uma deficiência no número de bases contendo textos em português. Em português brasileiro, apenas duas bases contemplam discurso de ódio [Fortuna et al. 2019, de Pelle and Moreira 2017]. Nenhuma das duas foca no discurso de ódio contra a mulher ou apresenta um volume expressivo de comentários de ódio nessa categoria.

A falta de diversidade linguística é um problema visto que a tarefa de detecção é diretamente ligada ao idioma utilizado: comentários ofensivos e discriminatórios são comumente respaldados por referências culturais próprias de um país. Memsmo entre países lusófonos, surgem diferentes interpretações. O termo "amélia", por exemplo, oriundo do samba "Ai Que Saudade da Amélia", de Mário Lagos e Ataulfo Alves - 1942, adquire no Brasil um significado pejorativo quando adjetiva uma mulher. Corazza et al. [Corazza et al. 2020] aponta que quando um mesmo algoritmo de classificação ou técnica de mineração são utilizados para classificar discursos de ódio em diferentes línguas, a eficiência do modelo resultante não é a mesma. De forma análoga, dependendo do foco do tópico explorado, as expressões e os vocabulários usados podem ser distintos. Essas diferenças não devem ser ignoradas no contexto da detecção do discurso de ódio.

Nesse contexto, este trabalho adota como foco a detecção automática de discurso de ódio contra a mulher em português brasileiro. Nas últimas décadas, com os avanços sociais e legais conquistados, o discurso misógino aberto contra a mulher migrou do mundo real para o virtual, se manifestando principalmente na forma de assédio sexual cibernético. Tal prática possui um impacto severo na vida das vítimas, havendo casos em que essa violência transcede a esfera da Internet e pode resultar em casos de agressão física e estupro [Citron 2011].

Este trabalho descreve a construção da base de dados MINA-BR, primeira base de dados em português brasileiro que contém comentários retirados da Internet com foco em discurso de ódio contra a mulher. Como forma de estabelecer uma *baseline* para a base construída, um estudo preliminar da base utilizando técnicas clássicas de mineração de dados e de aprendizado supervisionado também foi realizado.

2. Extração de comentários para a base de dados MINA-BR

O primeiro passo para a construção da base foi um estudo exploratório de fontes da Internet para a extração de comentários. Nessa etapa, as diferentes fontes foram analisadas quanto: (1) à possibilidade de extração de comentários de maneira automatizada; (2) ao reconhecimento da plataforma como veículo de compartilhamento de comentários de discurso de ódio.

De acordo com a literatura, a maioria das bases utilizadas para treinamento de algoritmos de detecção automática de discurso de ódio usam redes sociais e, ocasionalmente, sites de notícias como fontes de textos. A plataforma mais utilizada é o Twitter [Poletto et al. 2020], que disponibiliza uma interface de programação, ou API, do

inglês *Application Programming Interface*, que permite a mineração automática de comentários disponibilizados de maneira pública na plataforma.

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 pesquisa realizada apontou o Facebook como a plataforma onde as pessoas mais percebem o discurso de ódio contra a mulher, seguida pelo Twitter, pelo Instagram e pelo Youtube. 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 Instagram, essas plataformas não foram utilizadas no trabalho. Assim, somente o Twitter e o Youtube foram adotados como fontes para a pesquisa.

Quanto ao método de extração de comentários, a maioria das bases existentes de discurso de ódio utilizam a busca por palavras-chaves [Poletto et al. 2020]. Existem, no entanto, estudos recentes que propõem o uso de outros métodos para essa busca. Basile et al. [Basile et al. 2019], por exemplo, além de usarem a busca por palavras-chaves, também monitoram potenciais vítimas e coletam o histórico de comentários de ofensores já identificados. Já no trabalho de Davidson et al. [Davidson et al. 2017], os autores, após aplicarem o método das palavras-chaves, identificam os perfis dos quais os comentários foram extraídos e coletam todas as postagens desses usuários. Tal abordagem tenta balancear os temas e vocabulários existentes na base de dados.

Na literatura não é evidente uma metodologia clara para a seleção das palavras-chaves. Alguns trabalhos citam o uso de termos relacionados à classe do discurso de ódio de interesse [Alfina et al. 2017], outros documentam que as palavras-chaves usadas foram retiradas de repositórios online como o HateBase[1] [ElSherief et al. 2018]. No entanto, o uso deste último método não foi viável no contexto deste trabalho 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 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.

[1] https://hatebase.org/

Levando isso em consideração, as palavras usadas para a busca dos comentários para a construção da base 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. Alguns exemplos de palavras-chaves e tópicos utilizados na extração são: "feminazi", "misoginia", "abortista" e "#abortonao".

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 [Watanabe et al. 2018]. Eliminou-se também comentários que continham apenas uma palavra por falta de contexto suficiente.

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

3. Rotulação da base

Existem três principais estratégias para a rotulação de uma base de dados de discurso de ódio que são amplamente utilizadas. A primeira consiste na anotação dos comentários por especialistas no assunto. A segunda se baseia no recrutamento de voluntários não especializados para realizar a rotulação. E, a terceira, corresponde ao uso de um classificador automático para realizar a tarefa. Não existe consenso dentro da comunidade acadêmica a respeito de qual método é mais vantajoso [Poletto et al. 2020].

Neste trabalho, a segunda estratégia foi utilizada e, para auxiliar no processo de rotulação da base, foi desenvolvido um website[2] 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.

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

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 atua como um mecanismo de proteção. Caso a rotulação seja comprometida em algum ponto do processo, apenas os comentários pertencentes àquele grupo terão que ser reclassificados. A Figura 1 ilustra o processo de rotulação da base de dados.

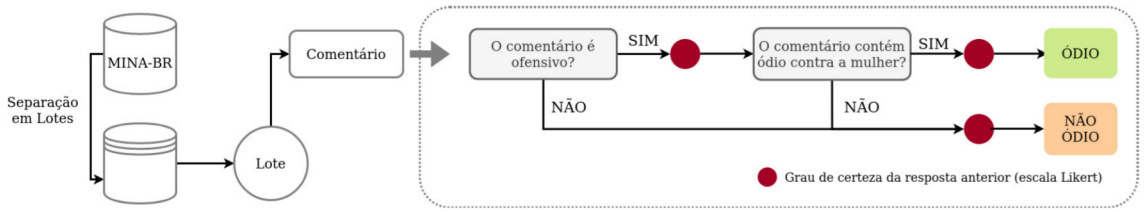

Figura 1. Fluxograma do método de rotulação da base de dados MINA-BR.

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

Até agosto de 2021, 2135 comentários haviam sido rotulados. Dentre esses, 16,26% foram classificados como ódio e 83,74% como não-ódio. A atribuição do rótulo final do comentário foi feita aplicando-se o critério da maioria absoluta. Tal distribuição de classes é coerente com aquelas encontradas em outras bases de discurso de ódio [de Gibert et al. 2018, de Pelle and Moreira 2017].

Durante o processo de rotulação, a base de dados foi alvo de um ataque organizado, o que retardou o processo de rotulação [Plath et al. 2021].

4. Construção de um *baseline*: pré-processamento e extração de parâmetros

As seguintes etapas de pré-processamento foram aplicadas à base de dados: 1) padronização dos caracteres para letras minúsculas; 2) retirada de *stopwords* utilizando como critério a lista dada pela biblioteca NLTK para português; 3) retirada de *hashtags* e espaços em branco duplos e 4) separação de frases em *tokens*.

Além disso, dois tratamentos distintos foram dados aos emojis. No pré-processamento A, os emojis foram convertidos em frases que descreviam a figura. Para tal, foi utilizada a biblioteca emoji do pyhton. Além disso, algumas abreviações comuns foram substituidas por sua forma completa, como "vc" em "você". Já no pré-processamento B, os emojis foram retirados dos comentários e nenhum tratamento foi dado às abreviações. Além disso, também testou-se o impacto do uso de *stemming* — técnica de extração de radicais — para a resolução do problema utilizando o algoritmo RSLP Stemmer [Huyck and Orengo 2001].

Para representar o texto foram utilizados os modelos *Bag of Words* (BoW) e TF-IDF (abreviação do inglês *Term Frequency–Inverse Document Frequency*) [Koushik et al. 2019].

5. Classificação e avaliação dos resultados

A detecção automática de discurso de ódio é uma tarefa que pode ser realizada por modelos baseados em diferentes tipos de algoritmos. Estudos utilizando modelos de classificação clássicos como *Support Vector Machines* (SVM) e classificação

Naïve Bayes têm apresentado resultados satisfatórios [Warner and Hirschberg 2012, Kwok and Wang 2013]. Ainda, nos últimos anos, modelos de redes neurais e redes profundas atingiram resultados superiores aos modelos clássicos na maioria dos casos [Corazza et al. 2018, Badjatiya et al. 2017]. Ademais, mais recentemente, abordagens utilizando o BERT [Devlin et al. 2018] obtiveram bons resultados, alguns superando a performance de outros modelos baseados em redes neurais profundas [Sohn and Lee 2019]. Atualmente, o BERT é considerado o modelo estado da arte.

A fim de se gerar um primeiro *baseline* para a base foram escolhidos cinco algoritmos clássicos de classificação para a tarefa: SVM Linear, Random Forest e três algoritmos do tipo Naive Bayes (Multinomial, Bernoulli e Gaussiano) [Bishop 2006]. Também avaliou-se o impacto do uso de um extrator de radicais nos resultados obtidos. Para cada algoritmo foi feita uma validação cruzada usando 10 subconjuntos. Para o pré-processamento A, foram feitos testes também utilizando o algoritmo BERT. Os resultados foram avaliados quanto à sua Medida-F1 média e estão mostrados, para cada pré-processamento, nas Tabelas 1 e 2.

Tabela 1. Medida-F1 média de cada algoritmo utilizando o pré-processamento A

Classificador	***Bag of Words***		**TF-IDF**		**Melhor**
	Sem *Stemming*	**RSLP *Stemmer***	**Sem *Stemming***	**RSLP *Stemmer***	**Resultado**
SVM	0.54	0.50	0.56	0.54	0.56
Random Forest	0.41	0.42	0.39	0.37	0.42
NB-Gaussiano	0.29	0.31	0.29	0.31	0.31
NB-Multinomial	0.13	0.18	-	-	0.18
NB-Bernoulli	0.39	0.37	0.08	0.08	0.08
BERT	NA	NA	NA	NA	0.56

Tabela 2. Medida-F1 média de cada algoritmo utilizando o pré-processamento B

Classificador	***Bag of Words***		**TF-IDF**		**Melhor**
	Sem *Stemming*	**RSLP *Stemmer***	**Sem *Stemming***	**RSLP *Stemmer***	**Resultado**
SVM	0.56	0.53	0.57	0.56	0.57
Random Forest	0.41	0.41	0.40	0.36	0.41
NB-Gaussiano	0.29	0.30	0.29	0.30	0.30
NB-Multinomial	0.39	0.38	-	-	0.39
NB-Bernoulli	0.13	0.20	0.08	0.07	0.20

Nota-se que, para todas as combinações de representação de texto e de *stemming*, o algoritmo SVM apresentou um resultado superior aos demais, sendo o melhor deles obtido por meio da representação de texto através do modelo TF-IDF sem o uso de algoritmo de extração de radicais e utilizando o pré-processamento B (0.57). O algoritmo Random Forest teve desempenho inferior ao SVM mas ainda consistentemente melhor do que os algoritmos Naive Bayes. A performance do algoritmo BERT, por sua vez, não conseguiu superar a resultado do SVM.

Além disso, notou-se que o uso do RSLP *Stemmer* piorou a performance do algoritmo SVM em todos os casos. Seu efeito sobre os outros classificadores variou dependendo do tipo de representação de texto aplicada. Para o Random Forest, por exemplo, o efeito do extrator de radicais foi positivo quando usado em conjunto com o modelo *Bag of Words*, mas negativo quando combinado com o TF-IDF.

Comparando as duas diferentes representações de texto utilizadas, percebe-se que o TF-IDF obtém as melhores Medidas-F1 do experimento quando combinado com o SVM, superando a representação por *Bag of Words*. Essa diferença de desempenho se inverte, no entanto, tratando-se dos outros algoritmos. Nesses, o *Bag of Words* possui uma performance superior ou muito próxima àquela do TF-IDF.

Além disso, percebe-se que o pré-processamento B obteve, no geral, resultados superiores àqueles do pré-processamento A. Tal resultado indica que a substituição dos emojis por sua descrição não ajudaram o modelo a distinguir comentários de ódio, mas, possivelmente, o confundiram. A hipótese levantada é que, como a descrição dos emojis não é feita a partir do contexto que eles estão sendo usados, quando o emoji está sendo utilizado com ironia, essa substituição pode adicionar termos às frases que não condizem com a mensagem transmitida. Por exemplo, no caso da mensagem mostrada na Figura 2, o emoji é descrito como "Rosto Sorridente Com Olhos De Coração".

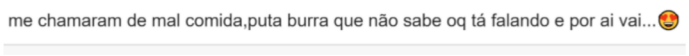

Figura 2. Comentário com emoji contendo ironia retirado da base MINA-BR

Notou-se durante o experimento um número elevado de termos que apareciam apenas uma única vez no universo de palavras da base de dados. No caso da representação de texto por *Bag of Words* sem o uso de extrator de radicais, o número encontrado de palavras únicas no dicionário da base foi de 4207 sendo que o número total de termos existentes é de 6821. Ou seja, de todas as palavras existentes na base, aproximadamente 62% tem frequência igual a um e, portanto, não agregam muita informação para o modelo. Quando se aplica o *stemming*, a porcentagem de termos únicos cai 52%.

Analisando os termos de frequência unitária, esperava-se encontrar palavras com grafias incorretas, abreviadas ou que são pouco utilizadas em um contexto virtual. Palavras como "eclâmpsia", "bbzinho" e "bolsomjnion" foram encontradas. No entanto, também tinham apenas uma aparição no dicionário, palavras mais usuais como "pegando" e "violenta". Assim, acredita-se que uma das formas de melhorar o desempenho dos modelos é realizando um pré-processamento mais fino, que corrija a grafia de palavras e trate gírias e abreviações. Além disso, a captação de mais comentários pode ser necessária para a expansão do universo de palavras da base de dados.

Foi feita também uma análise dos termos correta e incorretamente classificados. Notou-se que a maioria dos comentários de ódio corretamente classificados (TP), possuíam traços de ódio fortemente marcados, contendo xingamentos diretos à mulher. Também percebeu-se que esses comentários foram rotulados como ódio pelos anotadores acompanhados de um alto grau de certeza. 86% dos rótulos de ódio nesse caso foram dados com o nível máximo de confiança na escala Likert. Já os comentários corretamente classificados como não ódio (TN) em sua maioria não apresentavam xingamentos

e continham uma diversidade maior de vocabulário.

Tratando-se das amostras incorretamente classificadas, notou-se que aquelas que os algoritmos indicaram como ódio (FP) continham xingamentos ou palavrões, mas que estavam sendo utilizado em um contexto irônico. Ademais, alguns comentários continham ofensas mas não era possível discernir, pela falta de contexto, se eles eram maliciosos ou não. Já os comentários incorretamente classificados como não ódio (FN) continham menos xingamentos diretos, mas expressavam ideias misóginas. A Tabela 3 contém amostras de alguns casos.

Tabela 3. Amostras de Classificações de Comentários

Resultado	**Comentário Exemplo**
TP	*@NOME Vai tomar no seu cu feminazi do caralho!*
TN	*nem feminista nem antifeminista, apenas gostosa*
FP	*gente mal amada pior coisa né kkk*
FN	*Essa hora é qdo a mulher quer sair da balada pra ir pra casa pra levar vara*

6. Conclusão

Através deste trabalho, construiu-se a base de dados MINA-BR, que contém comentários em português brasileiro focados em textos contendo discurso de ódio direcionados à mulher. Dos 6002 comentários contidos na base de dados, 2135 já foram rotulados. O trabalho descreve a metodologia utilizada para a construção da base de dados, destacando o processo de escolha de termos de extração de texto e as medidas tomadas para mitigar os viéses de rotulação.

A partir da amostra de comentários rotulados, foi contruída também uma *baseline* para a base de dados utilizando algoritmos de classificação consagrados na área de NLP. Diferentes técnicas de pré-processamento foram testadas assim como dois tipos de representação de texto (TF-IDF e *Bag of Words*). A melhor Medida-F1 obtida foi de 0.57 através da combinação SVM + TF-IDF sem o uso de *stemming*.

Dentre os pontos a serem investigados está o elevado número de palavras com frequência única na base de dados. Uma das formas de se resolver essa questão é aumentando o dicionário da base de dados, isto é, rotulando mais comentários. Além disso, um pré-processamento mais fino dos textos que trate erros ortográficos, por exemplo, também pode contribuir para a mitigação desse problema.

Finalmente, o modelo obtido consegue detectar com sucesso mensagens de ódio contra a mulher que fazem uso de palavrões e expressões odiosas. No entanto, ele apresenta dificuldade em identificar corretamente comentários que contenham ironia ou que expressão ódio sem o uso de xingamentos explícitos.

7. Agradecimentos

Finalmente, agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo suporte financeiro através do Programa Institucional de Bolsas de Iniciação Científica (PIBIC), projeto número 123118/2020-4.

Referências

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

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., Rosso, P., Sanguinetti, M., et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).

Citron, D. K. (2011). Misogynistic cyber hate speech.

Corazza, M., Menini, S., Arslan, P., Sprugnoli, R., Cabrio, E., Tonelli, S., and Villata, S. (2018). Comparing different supervised approaches to hate speech detection.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

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

de Pelle, R. and Moreira, V. (2017). 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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

EPTV (2021). Jornal da eptv segunda edição - campinas/piracicaba - pesquisa da unicamp busca desenvolver detector de discursos de ódio na internet.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Huyck, C. and Orengo, V. (2001). A stemming algorithmm for the portuguese language. In *String Processing and Information Retrieval, International Symposium on*, page 0186, Los Alamitos, CA, USA. IEEE Computer Society.

Koushik, G., Rajeswari, K., and Muthusamy, S. K. (2019). Automated hate speech detection on twitter. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–4.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

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

Plath, H. O., Paiva, M. E. O., Pinto, D. L., and Costa, P. D. P. (2021). Base de comentários de discurso de ódio contra mulheres mina-br: da concepção aos ataques por robôs. In *XXIX Congresso de Iniciação Científica da UNICAMP*, Campinas, Brasil.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Sohn, H. and Lee, H. (2019). Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559.

Tontodimamma, A., Nissi, E., Sarra, A., and Fontanella, L. (2021). Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics.

Watanabe, H., Bouazizi, M., and Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.