XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB

ISSN 2177-3688

GT-8 – Informação e Tecnologia

ANÁLISE COMPARATIVA DE SENTIMENTO PARA DETECÇÃO AUTOMÁTICA DE DISCURSO DE ÓDIO UTILIZANDO O ALGORITMO LEIA-VADER

COMPARATIVE SENTIMENT ANALYSIS FOR AUTOMATED DETECTION OF HATE SPEECH APPLYING THE LEIA-VADER ALGORITHM

Guilherme Luiz Cintra Neves. UFPR.

Rodrigo Eduardo Botelho-Francisco. UFPR.

Luciano Heitor Gallegos Marin. UFPR.

Denise Fukumi Tsunoda. UFPR.

Modalidade: Resumo Expandido

Resumo: Humanidades Digitais e Ciência da Informação convergem na recente realidade tecnológica e digital. Discurso de ódio insere-se como vulnerabilidade nas plataformas digitais. No âmbito da Análise de Sentimentos, verificou-se o desempenho do algoritmo LeIA em relação aos discursos da base OFFCOMBR2 comparando a medida F do LeIA com análises similares. Trata-se de pesquisa descritiva aplicada que executa o LeIA na base OFFCOMBR2. Verificou-se baixo desempenho do LeIA na comparação efetuada. Sugere-se ampliação do léxico e revisão das etapas de pré-processamento. Como contribuição a Ciência da Informação, destaca-se a importância da prevenção de abusos da liberdade de expressão em plataformas digitais.

Palavras-Chave: Redes Sociais Digitais. Plataformas Digitais. Discurso de Ódio. Análise de Sentimentos. Processamento de Linguagem Natural.

Abstract: Digital Humanities and Information Science converge towards the recent technological and digital reality. Hate speech is considered a vulnerability on digital platforms. In the context of Sentiment Analysis, the LeIA's performance was verified within the discourses of the OFFCOMBR2 dataset, comparing its F-measure with similar analyses. This is an applied descriptive research that runs LeIA against OFFCOMBR2 dataset. Low performance of LeIA in the comparison made was verified. It is suggested to expand the lexicon and review the pre-processing steps. As a contribution to Information Science, the importance of preventing abuses of freedom of speech on digital platforms is highlighted.

Keywords: Digital Social Network. Digital Platforms. Hate Speech. Sentiment Analysis. Natural Language Processing.

1 INTRODUÇÃO

As redes sociais digitais são plataformas online onde diferentes interagentes praticam a liberdade de expressão, um tema central e recorrente em diversas investigações científicas, principalmente no âmbito das Humanidades Digitais (HD). A principal controvérsia gira em torno dos limites da liberdade de expressão (PELLE; MOREIRA, 2017, p. 8) e do controle das plataformas sobre o discurso dos indivíduos (GILLESPIE, 2018, p. 5). Dentre questões que permeiam as redes sociais digitais como *Facebook*, *Instagram*, *Twitter* e *TikTok*, destaca-se o discurso de ódio, entre outras vulnerabilidades digitais, como explicam Junqueira, Botelho-Francisco e Grieger (2021, p. 166):

> [...] vulnerabilidade passa a agregar progressivamente, nos campos técnicos, acadêmicos, administrativos e políticos, dimensões socioeconômicas (pobreza, insegurança alimentar, fome etc.), carências de acesso e atendimento em bens e serviços públicos e/ou comunitários de primeira necessidade (educação, atendimento médico, serviços de saúde, saneamento básico e outros), suscetibilidade a riscos e danos ecológicos naturais (fenômenos associados às mudanças climáticas, tormentas, deslizamentos de terra, rompimentos de barragens e outros similares), fenômenos políticos e culturais (genocídios, guerras étnicas, intolerâncias, ódios e perseguições das mais diferentes naturezas). (JUNQUEIRA, BOTELHO-FRANCISCO E GRIEGER, 2021, p. 166).

Oliveira, Kaya e Roncaglio (2022, p. 4) explicam que as HD nos contextos informacionais convergem com a Ciência da Informação (CI), enquanto disciplina científica, devido ao alinhamento dos estudos dos usos da informação na recente realidade tecnológica e digital.

> Nos casos de práticas socioculturais, as humanidades digitais estão presentes na sociedade em rede e em seu processo de interação e usos de sistemas de informação e comunicação nos contextos institucionais que abordam, entre outras questões, aspectos éticos, sociais, políticos e culturais. (OLIVEIRA; KAYA; RONCAGLIO, 2022, p. 4).

A detecção do discurso de ódio em plataformas digitais, no entanto, não é tarefa trivial. Além das denúncias que podem ser realizadas pelos interagentes, a detecção automática é um dos caminhos que vêm sendo trilhado. Neste processo, vislumbra-se a Análise de Sentimentos (AS), técnica ainda pouco explorada em estudos em língua portuguesa. Em revisão sistemática recente podemos ver que 51% dos artigos sobre o tema estão em Inglês e 1% em Português (JAHAN; OUSSALAH, 2021, p. 14). Assim, esta pesquisa focou em aspectos relacionados ao idioma Português como forma de delimitar o estudo.

Neste contexto, o presente artigo possui, como objetivo, a verificação do desempenho do algoritmo LeIA (ALMEIDA, 2018) em relação à AS em discursos armazenados na base OFFCOMBR2 (PELLE, 2021). Especificamente, busca-se comparar a medida F da análise aqui apresentada com a medida F dos algoritmos utilizados por Pelle e Moreira (2017) em seu trabalho.

Embora seja importante o debate sobre o controle das plataformas ou os limites da liberdade de expressão, os mesmos não são foco deste trabalho. No entanto, ao explorar as possibilidades de detecção automática de discursos de ódio, espera-se contribuir com o avanço da técnica de AS ao examinar essa forma de mineração de opiniões.

A pesquisa aqui descrita está dividida em: introdução, onde apresenta-se um breve contexto, o objetivo e justificativa para a pesquisa; referencial teórico, que apresenta os principais conceitos utilizados no contexto da pesquisa; procedimentos metodológicos, que explica o percurso metodológico do trabalho; análise dos resultados, que examina a performance do algoritmo LeIA; e considerações finais, onde são apresentadas as limitações da pesquisa, indicações de trabalhos futuros e o fechamento dos objetivos.

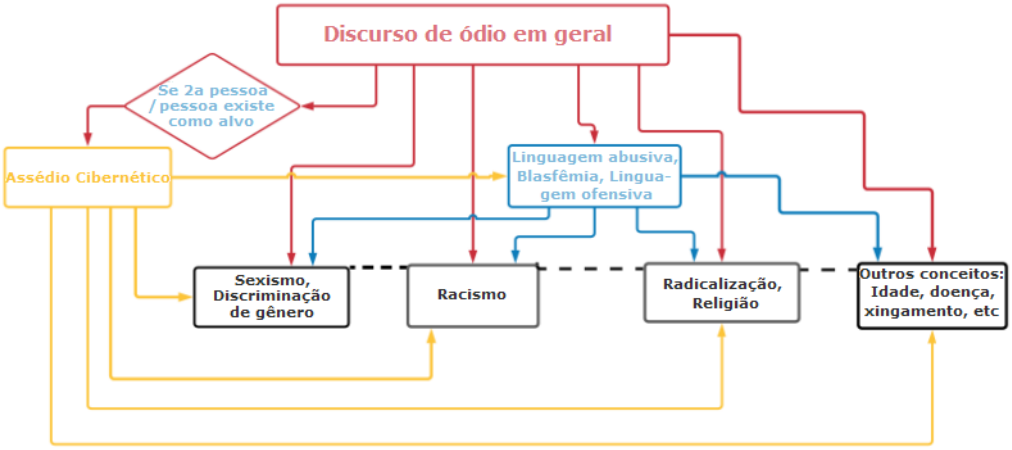
2 REFERENCIAL TEÓRICO

A Internet configura-se, no contemporâneo, como um lócus fundamental para compreender várias dimensões da vida em sociedade. A ideia de plataforma e plataformização[1], neste contexto, mostra como as mídias sociais ocupam protagonismo na mediação de relações sociais, econômicas, políticas, de trabalho, entre outros aspectos. Se por um lado há o otimismo de novos modelos de negócio, por outro são vislumbrados desafios que vão além de estar incluídos ou excluídos da Internet e das próprias plataformas que nela ganham notoriedade. Trata-se, como defendido em Junqueira, Botelho-Francisco e Grieger (2021, p. 166), de vulnerabilidades digitais, termo que serve como "alternativa conceitual que auxilia a vislumbrar múltiplas perspectivas da mediação tecnológica e social entre diferentes atores, humanos e não-humanos".

1 Conforme explicam Poell, Nieborg e Van Dijck (2020, p. 2), "compreendemos plataformização como a penetração de infraestruturas, processos econômicos e estruturas governamentais de plataformas em diferentes setores econômicos e esferas da vida. E, a partir da tradição dos estudos culturais, concebemos esse processo como a reorganização de práticas e imaginações culturais em torno de plataformas".

Entre as vulnerabilidades digitais, destaca-se o discurso de ódio, abordado de diferentes perspectivas por Jahan e Oussalah (2021, p. 2-3), visando contribuir com a descoberta e reconhecimento das inter-relações entre as terminologias utilizadas nessas perspectivas. Assim, os autores (JAHAN; OUSSALAH, 2021, p. 4) apresentam um diagrama relacional para explicar o discurso de ódio (Figura 1).

Figura 1 - Diagrama relacional entre diferentes tipos de conceitos de discurso de ódio.



Fonte: Adaptado de Jahan e Oussalah (2021, p. 4, tradução nossa).

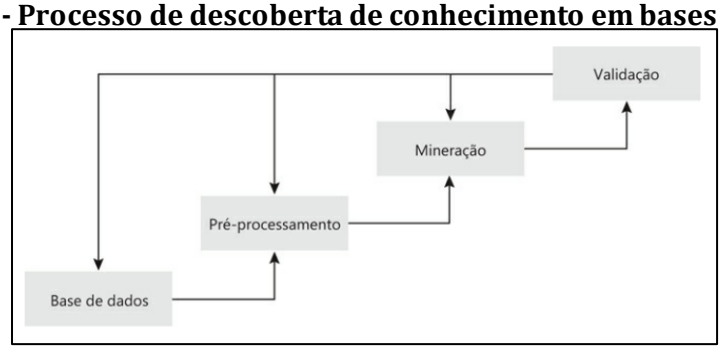
Entende-se, para fins dessa pesquisa, que o discurso de ódio é definido por comunicações de opiniões negativas que "atacam ou depreciam um grupo com base em raça, origem étnica, religião, deficiência, gênero, idade, identidade de gênero e orientação sexual" (NOBATA *et al. apud* JAHAN; OUSSALAH, 2021, p. 2, tradução nossa). Para tratar da detecção de discurso de ódio online, vislumbra-se a utilidade da Descoberta de Conhecimento em Bases de Dados, ou KDD (*Knowledge Discovery in Databases*), nome dado ao processo de busca, identificação e extração de padrões a partir de dados armazenados em bases de dados, muitas vezes dispersas e inexploradas; e que visa gerar conhecimento que seja novo e potencialmente útil para a tomada de decisão estratégica, como controle de processos, gestão da informação e conhecimento, processamento de consultas e muitas outras aplicações (CASTRO; FERRARI, 2016, p. 48).

O KDD, segundo Castro e Ferrari (2016, p. 46-47), é dividido em quatro partes, correlacionadas e interdependentes: Base de Dados, Pré-processamento, Mineração e Validação; conforme demonstra a Figura 2. No âmbito da KDD, a técnica de AS em redes sociais digitais vem sendo utilizada para identificar discursos de ódio, ao permitir a

responsabilização, seja por controle da própria plataforma, seja pelos demais meios legais (HELBERGER; PIERSON; POELL, 2017, p. 2).

A Análise de Sentimentos (AS) é a tarefa de encontrar opiniões de autores sobre entidades específicas (FELDMAN, 2013, p. 82, tradução nossa). No contexto brasileiro, o algoritmo LeIA apresenta-se como uma opção relevante que abrange um léxico em Português para a operacionalização da AS. Além dele, há também a OFFCOMBR2, base anotada também em Português.

Figura 2 - Processo de descoberta de conhecimento em bases de dados



Fonte: Extraído de Castro e Ferrari (2016, p. 47).

O LeIA (ALMEIDA, 2018) caracteriza-se como uma ramificação do léxico e ferramenta para Análise de Sentimentos VADER (*Valence Aware Dictionary and sEntiment Reasoner*) adaptado para textos em Português, com suporte para *emojis* e foco na AS de textos expressos em redes sociais digitais. Os criadores do VADER (HUTTO; GILBERT, 2014, p. 216, tradução nossa) utilizam uma combinação de métodos qualitativos e quantitativos para construir e validar empiricamente uma lista de atributos léxicos de alto padrão, bem como as medidas de intensidade de sentimento a eles associadas, especificamente orientadas a AS em contextos similares a *microblogs*.

A base OFFCOMBR (PELLE; MOREIRA, 2017, p. 512-513, tradução nossa) foi construída a partir de comentários encontrados no site do portal G1[2], nas seções de Política e Esportes. Foram reunidos cerca de dez mil comentários em 115 notícias, entre os quais 1.250 comentários aleatórios foram anotados por três juízes, escolhidos por suas expertises em relação ao tema. A partir desse *corpus*, foram geradas duas bases: OFFCOMBR2, que utiliza todos os 1.250 comentários e atribui a classe (sim/não)[3] de acordo com a escolha de dois juízes; e OFFCOMBR3, que leva em consideração a escolha unânime da classe (sim/não),

2 https://g1.globo.com/.
3 Sim, para comentários foi considerado ofensivos; Não, para comentários não ofensivos.

sendo mais restrita com apenas 419 comentários. A base OFFCOMBR não considera comentários neutros, que restam classificados como não ofensivos.

3 PROCEDIMENTOS METODOLÓGICOS

Este trabalho configura-se como pesquisa descritiva, pois "têm como objetivo a descrição das características de determinada população ou fenômeno" e "podem ser elaboradas também com a finalidade de identificar possíveis relações entre variáveis". Além disso, caracteriza-se como pesquisa aplicada, já que "têm como objetivo a descrição das características de determinada população ou fenômeno, [podendo] ser elaboradas também com a finalidade de identificar possíveis relações entre variáveis" (GIL, 2022, p. 41-42).

Considera-se como objeto da pesquisa o desempenho do algoritmo LeIA em relação à base anotada OFFCOMBR2, dessa forma o espaço amostral utilizado na pesquisa é definido pelo resultado da medida *compound*, gerado pelo próprio algoritmo LeIA para cada um dos 1.250 comentários da base anotada OFFCOMBR2, bem como os indicadores gerados pela matriz de confusão.

O *compound* (medida normalizada e ponderada dos valores negativos, neutros e positivos calculados pelo algoritmo LeIA/Vader) é utilizado para definir os valores positivos e negativos dos comentários e comparar com o valor anotado, e assim gerar as métricas de desempenho do algoritmo. O *script*[4] em *Python* para o *Colab*, por sua vez, foi disponibilizado para verificação dos parâmetros e bibliotecas utilizadas.

Neste trabalho, a execução do algoritmo LeIA foi efetuada na plataforma *Google Colab*[5] *(Python*[6]*)*, utilizando a base anotada OFFCOMBR2 como *corpus* de treinamento. Conforme descrito no repositório do *GitHub*[7] onde está o LeIA, não é necessário aplicar a etapa de pré-processamento ao texto, como a conversão em minúsculas, tokenização[8], lematização[9] ou outras técnicas de pré-processamento. Desta forma, converteu-se a base OFFCOMBR2 do formato ARFF (*Weka*) para CSV, para facilitar o manuseio da base no *Colab*,

4 https://colab.research.google.com/drive/1xnIcPUuIFuVBoXJsTLhxeS5ubZLstkEn?usp=sharing.
5 Ambiente colaborativo para compilação da linguagem de computador Python. https://colab.research.google.com/.
6 Programação por *scripts* de código aberto, com ênfase na legibilidade da codificação por seres humanos.
7 https://github.com/rafjaa/LeIA.
8 Divisão do texto em unidades menores (tokens), no processo remove-se caracteres especiais e pontuação.
9 Transformação de cada palavra em um radical analisando o contexto em que a palavra se encontra e sua classe gramatical

executando o LeIA e comparando o resultado com os 1.250 comentários anotados da base OFFCOMBR2.

Considera-se como parte fundamental deste estudo a aplicação da AS extraindo os escores da base OFFCOMBR2, transformando-os em *dictionary*, extraindo-se os valores para uma lista, e então convertendo-os em *pandas dataframe; a* próxima etapa foi extrair o *compound* e converter os valores numéricos em texto positivo ou negativo, conforme segue:

```
[] analise = OFFCOMBR2['revisão'].apply(s.polarity_scores)
[] dictLeia = analise.to_dict()
[] d = list(dictLeia.values())
[] df = pd.DataFrame(d)
[] dfn = df['compound'].astype(float)
[] compound = np.where(dfn >= 0, 'positive', 'negative')
```

O *compound* gerado pelo algoritmo LeIA no *Colab* é calculado pela soma das valências das palavras no léxico e resulta em valores positivos ou negativos, pode ser observado a partir dos exemplos do Quadro 1, e da mesma forma o Quadro 2 apresenta alguns resultados obtidos:

Quadro 1: Exemplos de *compound*

	neg	neu	pos	*compound*
0	0.000	1.000	0.000	0.0000
1	0.000	0.786	0.214	0.4588
2	0.255	0.634	0.111	-0.8658
...	...	...	...	...
1246	0.348	0.535	0.118	-0.5106
1247	0.000	1.000	0.000	0.0000
1248	0.000	1.000	0.000	0.0000

Fonte: Algoritmo LeIA executado no *Google Colab*, elaborado pelos autores.

Verifica-se que o valor do *compound* pode ser utilizado para descrever o sentimento predominante no texto, por meio dos limites de valores: *compound* >= 0.05 (Sentimento positivo); *compound* <= -0.05 (Sentimento negativo); (*compound* > -0.05) e (*compound* < 0.05) (Sentimento neutro) (ALMEIDA, 2018).

Quadro 2: Comparativo entre a classificação do LeIA com a classificação manual

	compound	LeIA	OFFCOMBR2	Comentários
0	0.0000	positive	negative	"Votaram no PEZAO Agora tomem no CZAO"
1	0.4588	positive	positive	"cuidado com a poupanca pessoal Lembram o que aconteceu na epoca do Collor ne"
...	...	...	...	...
1248	0.0000	positive	negative	"Achei que a macaca vivia apenas na floresta ou no zologico"
1249	0.0000	positive	negative	"Espera essa neve derreter e usa ela pra lavar louca"

Fonte: Algoritmo LeIA executado no *Google Colab* e base OFFCOMBR2, elaborado pelos autores.

Finalmente, gerou-se a matriz de confusão (Tabela 1) e o relatório de classificação (Tabela 2) avaliando e comparando com os resultados originais (ALMEIDA, 2018).

4 ANÁLISE DOS RESULTADOS

Conforme a matriz de confusão apresentada na tabela 1 e o relatório de classificação apresentado na tabela 2, verifica-se que o algoritmo LeIA teve uma performance inferior aos demais algoritmos testados (como por exemplo: *Naive Bayes* e SVM) por Pelle e Moreira (2017), identificando muitos falsos positivos e falsos negativos na matriz de confusão.

A matriz de confusão é utilizada para verificar o desempenho de um algoritmo no campo do aprendizado de máquina. O eixo horizontal representa os valores de predição, enquanto o eixo vertical representa os valores reais; os valores zero representam uma classificação correta (positiva) e os valores um representam classificação incorreta (negativa).

Tabela 1: Matriz de Confusão

	0	**1**
0	178	241
1	371	460

Fonte: Elaborado pelos autores.

A matriz evidencia o acerto de 638 (178 + 460) classificações e 612 (371 + 241) erros. No entanto, na maior classe houve acerto de 55,35% (460 acertos em 831 exemplos da classe 1).

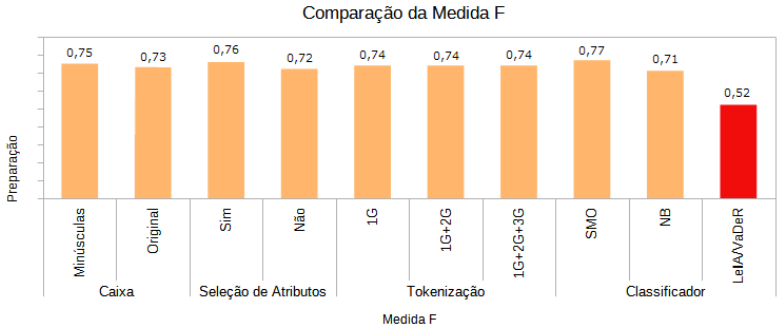
Tabela 2: Relatório de Classificação

	precision	recall	f1-score	support
negative	0,32	0,42	0,37	419
positive	0,66	0,55	0,60	831
accuracy			0,51	1250
macro avg	0,49	0,49	0,48	1250
weighted avg	0,54	0,51	0,52	1250

Fonte: Elaborado pelos autores.

A análise original de Pelle e Moreira (2017, p. 516) baseou-se principalmente em uma medida F ponderada, cujos valores absolutos variaram entre 0,69 e 0,85. Para a nossa análise consideramos a medida F ponderada (Tabela 2), seja F1-Score = 0,52 (Gráfico 1). Assim, o desempenho do LeIA (0,52) foi inferior ao resultado apresentado na análise original, registrando uma diferença 17% inferior ao pior resultado da referência (0,69).

Gráfico 1: Comparação da Medida F ponderada



Fonte: Adaptado de Pelle e Moreira (2017, p. 516).

A medida F é definida como a média harmônica entre precisão e revocação; sendo que a precisão se caracteriza como a capacidade de evitar falsos positivos e a revocação é a proporção entre verdadeiros positivos e o total de análises feitas pelo algoritmo (POWERS, 2020, p. 38). Ao utilizar este método para análise, percebe-se que para melhorar a performance do LeIA seria necessário incluir diversos termos e contrações idiomáticas apresentados na escrita dos comentários em plataformas de redes sociais digitais ao léxico. Com essa medida o algoritmo poderia identificar uma quantidade maior de termos linguísticos com teor considerado negativo. Outra possibilidade, como já mencionado pelos autores (PELLE; MOREIRA, 2017, p. 17, tradução nossa), seria aumentar a base de dados anotada.

5 CONSIDERAÇÕES FINAIS

O algoritmo LeIA não teve um bom desempenho quanto a detecção de discurso de ódio em comentários que compõem a base OFFCOMBR2 em comparação com os testes apresentados por Pelle e Moreira (2017). Para a sua melhoria, sugere-se a ampliação dos termos que compõem o léxico de forma a otimizar a performance do algoritmo. Além disso, pode ser necessário rever a forma como o algoritmo LeIA trata as etapas de pré-processamento, como a diferenciação de maiúsculas e minúsculas, tokenização, lematização, seleção de atributos e outras formas de preparo dos dados para a mineração e Análise de Sentimentos.

Outras formas de mineração de opiniões vêm sendo criadas. Verificar o desempenho de outras técnicas e métodos que funcionem com o idioma Português pode ser uma contribuição importante no cenário nacional. A necessidade de aprimorar a Análise de Sentimentos em Português é um ponto que pode ser considerado importante no cenário nacional, pois a maioria dos trabalhos detecta apenas termos em Inglês, obrigando a fazermos a tradução dos *corpora* a serem utilizados na mineração. Assim, para a continuidade desse estudo, recomenda-se a avaliação dessas novas técnicas e métodos sejam anotadas ou automatizadas, com tradução para o Inglês, ou utilizando dicionários em Português. Desta maneira poderemos identificar quais precisam ser desenvolvidas e quais não valem o esforço, além de contribuir com novas ferramentas que permitam esse desenvolvimento.

No âmbito da Ciência da Informação, esta pesquisa contribui com a investigação dos processos de coleta e processamento de dados para a Análise de Sentimento por meio da detecção automática de discurso de ódio, destacando-se a importância social da moderação em plataformas digitais para prevenir abusos da liberdade de expressão nos diversos fenômenos socioculturais que se manifestam no ambiente digital.

REFERÊNCIAS

ALMEIDA, R. J. A. **LeIA**: léxico para inferência adaptada. San Francisco: GITHUB, 2018. Disponível em: https://github.com/rafjaa/LeIA. Acesso em: 11 maio 2022.

FELDMAN, R. Techniques and applications for sentiment analysis. **Communications of the ACM**, v. 56, n. 4, abril 2013. DOI: http://doi.org/10.1145/2436256.2436274.

GIL, Antonio C. **Como elaborar projetos de pesquisa**. 7. ed. Rio de Janeiro: Grupo GEN; Atlas, 2022. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786559771653. Acesso em: 11 maio 2022.

GILLESPIE, T. **Custodians of the internet**. New Haven; London: YALE, 2018.

HELBERGER, N.; PIERSON, J.; POELL, T. Governing online platforms: from contested to cooperative responsibility. **The Information Society**, v. 34, n. 1, 27 dez 2017. DOI: https://doi.org/10.1080/01972243.2017.1391913.

HUTTO, C.; GILBERT, E. VADER: A parsimonious rule based model for sentiment analysis of social media text. International AAAI Conference on Web and Social Media, 8., 2014, Michigan. **Proceedings...** Michigan: AAAI, 2014. Disponível em: https://ojs.aaai.org/index.php/ICWSM/article/view/14550. Acesso em: 11 maio 2022.

JAHAN, M. S.; OUSSALAH, M. A systematic review of hate speech automatic detection using Natural Language Processing. **arXiv**, 22 maio 2021. DOI: https://doi.org/10.48550/arXiv.2106.00742

JUNQUEIRA, A. H; BOTELHO-FRANCISCO, R. E.; GRIEGER, J. D. Vulnerabilidades digitais: diálogos e aproximações possíveis com os aportes teóricos barberianos da comunicação. **Chasqui**, n. 147, p. 163-180, ago.-nov. 2021. Disponível em: https://dialnet.unirioja.es/descarga/articulo/8093847.pdf. Acesso em: 13 maio 2022.

OLIVEIRA, K. V. R. de; KAYA, G. T.; RONCAGLIO, C. Ciências da informação e humanidades digitais: produção, consumo e materialidade da informação em plataformas digitais. **Acervo**, Rio de Janeiro, v. 3, n. 1, p. 1-13, jan./abr. 2022. Disponível em: https://revista.an.gov.br/index.php/revistaacervo/article/view/1783/1707. Acesso em: 20 ago. 2022.

PELLE, R. P. de. **OffComBR** [Python]. San Francisco: GITHUB, 2021. Disponível em: https://github.com/rogersdepelle/OffComBR (Originalmente publicado em 2017). Acesso em: 23 maio 2022.

PELLE, R. P. de; MOREIRA, V. P. Offensive comments in the Brazilian web: a dataset and baseline results. **Proceedings...** Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 6., 2017, São Paulo: CSBC, 2017. DOI: https://doi.org/10.5753/brasnam.2017.3260.

POELL, T.; NIEBORG, D.; VAN DIJCK, J. Plataformização. **Fronteiras:** estudos midiáticos, v. 22, n. 1, p. 2–10, 2020. DOI: https://dx.doi.org/10.4013/fem.2020.221.01.

POWERS, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. **arXiv**, 11 out. 2020. DOI: https://doi.org/10.48550/arXiv.2010.16061.