Análise de discursos em notícias sobre homofobia, racismo e sexismo em comentários de portais brasileiros de notícias

Lucas D. F. Rodrigues
Universidade Federal do Oeste do Pará (UFOPA)
Santarém, Pará, Brasil
lucas.darlindo@gmail.com

Antonio F. L. Jacob Junior
Universidade Estadual do Maranhão (UEMA)
São Luís, Maranhão, Brasil
antoniojunior@professor.uema.br

Fábio M. F. Lobato
Universidade Federal do Oeste do Pará (UFOPA)
Santarém, Pará, Brasil
fabio.lobato@ufopa.edu.br

RESUMO

Posts with defamatory content or hate speech are constantly found on social media. The results for readers are numerous, not restricted only to the psychological impact, but also to the growth of this social phenomenon. With the General Law on the Protection of Personal Data and the Marco Civil da Internet, service providers became responsible for the content in their platforms. Considering the importance of this issue, this paper aims to analyze the content published (news and comments) on the G1 News Portal with techniques based on data visualization and Natural Language Processing, such as sentiment analysis and topic modeling. The results show that even with most of the comments being neutral or negative and classified or not as hate speech, the majority of them were accepted by the users.

KEYWORDS

Homofobia, Racismo, Sexismo, Conteúdo Gerado pelo Usuário, Discurso de Ódio

1 INTRODUÇÃO

Diante do crescimento contínuo do *User-Generated Content* (*UGC*) e a maior ocorrência de discursos de ódio nas mídias sociais, tornou-se necessária a supervisão desse conteúdo para as empresas *online* [16]. Diante disso, enquanto as mídias sociais poderiam ter sido uma forma de garantir a liberdade de expressão, agora precisam proteger seus usuários de conteúdos abusivos e censurá-los [10].

Com a promulgação da Lei Geral de Proteção aos Dados Pessoais e o Marco Civil da Internet, as prestadoras de serviços tornaram-se responsáveis pelo conteúdo ali disposto [4, 5]. Esse fato torna-se ainda mais crítico quando relacionado às mídias sociais. A propagação de discursos de ódio nos meios *online* fomenta uma má conduta social, fornece suporte e incita crimes de ódio no mundo real [13].

Suas consequências psicológicas afetam não só as vítimas, mas também os leitores [25]. Destaca-se também a correlação entre o aumento do comportamento misógino *online* e o número de feminicídios no Brasil [34]. Além de discursos sexistas, as redes sociais como *Facebook*[1] ou *Twitter*[2] transformaram-se em ambientes hostis por propagarem variadas formas de discurso de ódio, sejam eles racistas e/ou homofóbicos [38].

Em 2018, uma pesquisa conduzida pela *Safernet* Brasil apresenta que foram efetuadas 133.732 queixas de incitações de ódio pela internet no país e representa em um aumento de 37,71% dentre os comentários racistas e 59,13% de comentários homofóbicos [28]. Tal fenômeno levanta a necessidade do desenvolvimento de ferramentas que não só apresentem soluções para a detecção automática de discursos de ódio, mas que apresente a sua categoria, como por exemplo: homofobia, racismo e sexismo [15]. A possibilidade de segmentar esse tipo de discurso em várias subclasses, como as supracitadas e adicionais, viabiliza a melhor detecção quando assimilado a uma estrutura de grafos e permitindo tarefas de classificação multi-rótulo que conseguem identificar um ou mais tipos em um único texto [14].

As estratégias automáticas para a detecção de discursos de ódio se beneficiaram diretamente do campo de Processamento de Linguagem Natural (PLN) [32]. Percebe-se no estado da arte um crescimento quanto ao campo de discurso de ódio, principalmente direcionado a ferramentas desenvolvidas para o inglês. Contudo, essa abordagem gera lacunas para os demais idiomas, os quais também manifestam a disseminação desses discursos, como em Português [35] ou Alemão [29], por exemplo. Diante dos fatos apresentados, foram levantadas as seguintes Perguntas de Pesquisa (PP):

- **PP1**: Quais as características que representam a forma como os usuários interagem nas plataformas de notícias?
- **PP2**: Qual o teor dos comentários publicados em portais de notícia que apresentam pseudoanonimato?

Visando responder tais perguntas, o presente artigo apresenta análises em notícias e comentários publicados em grupos de palavras-chaves relacionadas a Homofobia, Racismo e Sexismo em um portal de notícias, com o objetivo de identificar aspectos internos das publicações, assim como a presença de discursos de ódio ou ofensivos a fim de melhor visualizar o *UGC* nas plataformas de mídias sociais, permitindo uma melhor identificação destes e de suas subclasses no idioma português, o qual ainda apresenta lacunas nessa área no estado da arte.

O artigo está estruturado como se segue. Na Seção 2 serão apresentados alguns trabalhos relacionados à temática. Em seguida, na Seção 3 é apresentada a metodologia utilizada. Na Seção 4 serão apresentados os resultados e discussões. Por fim, na Seção 5 serão feitas as considerações finais e projeções para trabalhos futuros.

2 TRABALHOS RELACIONADOS

A partir de um levantamento sistemático em trabalhos que visam analisar mídias sociais, os autores efetivaram uma pesquisa visualizando quais plataformas, ferramentas e palavras-chaves utilizadas nos artigos convergentes à área [20]. Identificou-se que há maior interesse no *Twitter*, principalmente por conta de sua grande base de usuários ativos e facilidade de coleta dos dados ali presentes, seguida pelo *Facebook* que possui a maior quantidade de usuários

[1] https://www.facebook.com/
[2] https://www.twitter.com/

do mundo. Contudo, plataformas específicas de notícias não foram encontradas, o que pode caracterizar a falta ou o baixo interesse em análises nelas.

Um estudo foi conduzido com o intuito de verificar as interações em comunidades de aprendizado de idioma na rede social *Reddit*[3] [24]. Os resultados mostram que a plataforma difere de uma rede social de uso geral, onde os utilizadores interagem principalmente entre pares e grupos próximos, sem muitas pessoas. A variedade linguística também difere bastante, fator que depende da comunidade e idioma ao qual está sendo publicado.

Quando direcionado a telejornais, uma pesquisa propôs verificar os sentimentos presentes nas legendas apresentadas durante a transmissão do Jornal Nacional (JN) e Jornal da Record (JR), assim como as opiniões publicadas pelos usuários no *Twitter* [9]. Os resultados apontaram que o JN é caracterizado por mais palavras negativas com sentimento positivo, e que o JR é marcado por mais termos com sentimentos negativos conectados a raiva, tristeza e morte. Quanto aos usuários, as críticas são mais direcionadas ao JN, principalmente por agrupar uma maior audiência.

Em outro trabalho, é proposto um mecanismo de detecção de discurso de ódio baseado em um classificador construído com uma Máquina de Vetores de Suporte [22]. Os autores destacam que grande parte dos trabalhos na área possuem enfoque na utilização de classes binárias para efetuar a tarefa proposta.

Também é apresentada em outra produção uma abordagem para filtragem de discurso de ódio construída para o idioma italiano [11]. O propósito é a detecção e remoção de um dado comentário ou postagem e de mensagens que possam ser transmitidas por meio do canal de bate-papo do *Facebook*. As métricas foram desenvolvidas a partir dos algoritmos de Máquina de Vetores de Suporte e Memória de Curto e Longo Prazo, obtendo acurácias de 72,95% e 75,23% para os algoritmos, respectivamente.

A lacuna de trabalhos em diversos idiomas é uma das principais questões ainda a ser solucionada no campo de discurso de ódio. Com isso, foram feitos levantamentos sobre a temática no estado da arte, com países, quantidade de trabalhos, metodologias e outros itens pertinentes a área [1, 39]. É perceptível o crescimento na produção durante os últimos anos, porém grande parte é direcionada para o Inglês. Os demais idiomas possuem uma quantidade inferior produzida quando comparada. O Português em ambos os trabalhos não possui uma representatividade expressiva quanto ao número de produções, caracterizando o supracitado na Seção 1.

As problemáticas citadas anteriormente e na Seção 1 fomentam uma maior proliferação do discurso de ódio nos ambientes de interação social, como as mídias sociais. O presente artigo possui como objetivo a análise do conteúdo textual gerado pelos usuários em uma plataforma de notícias, servindo como base para melhorias em sistemas de detecção em plataformas que permitem interação com usuário por intermédio de comentários e visa estimular mais produções para a língua portuguesa.

3 METODOLOGIA

Na presente seção será apresentada a metodologia utilizada para a condução das análises neste artigo, com enfoque na visualização dos dados obtidos e em PLN, seguindo um fluxo similar à outros trabalhos da área [36] e com o *framework* aqui utilizado disponível na Figura 1.

Figura 1: *Framework* metodológico utilizado para a pesquisa.

3.1 Coleta de Dados

Para a etapa de coleta de dados, fez-se um levantamento no ranqueamento de páginas mais acessadas da *Alexa*[4], sendo este atualizado diariamente com os dados de tráfego gerado pelos usuários dentro de um período de três meses [41]. O Portal G1 foi escolhido a partir de sua posição e relevância, também englobando notícias de todas as regiões do país. Foram coletadas apenas as informações básicas das matérias, juntamente com informações presentes nos comentários. Os termos de busca utilizados estão listados na Tabela 1.

Tabela 1: Palavras-chave por grupo alvo para a coleta.

Grupo	Palavras-chave
Homofobia	Homofobia
	Homofóbico
	Homosexual
	LGBT
	Transfobia
Racismo	Racismo
	Racista
	Negro
	Negritude
	Cotas
Sexismo	Feminicídio
	Feminismo
	Assédio
	Misoginia
	Sexismo

As palavras-chave apresentadas na Tabela 1 foram selecionadas após uma consulta com profissionais representantes de cada

[3]https://www.reddit.com/

[4]https://www.alexa.com/topsites/countries/BR

grupo alvo, englobando ativistas e profissionais de comunicação. Para cada, definiram-se cinco principais com maior representação para a etapa de coleta de dados, totalizando quinze palavras-chave utilizadas para tal. A metodologia utilizada foi inspirada na *Delphi Methodology* [17].

Para se realizar a coleta de dados, desenvolveu-se um *web crawler* em *Python* com a biblioteca *Scrapy Framework*[5], onde selecionaram-se notícias que eram indexadas pelo portal com pelo menos uma das palavras-chave. Os dados coletados foram armazenados em arquivos no formato *Comma-Separated Values* (*CSV*) [2, 37]. Os atributos extraídos estão descritos na Tabela 2.

Tabela 2: Dados extraídos conforme a fonte.

Fonte	Item	Tipo de Dado
Notícias	Título	*String*
	Data de Publicação	*Timestamp*
	Palavra-chave	*String*
	Link	*URL*
	Localidade	*String*
Comentários	Autor	*String*
	Data de Publicação	*Timestamp*
	Comentário	*String*
	Interações	*Integer*
	Localidade	*String*

Conforme explicitado na Tabela 2, das notícias foram extraídos o título, data de publicação e localidade, atributos úteis para a análise de eventos específicos. A palavra-chave foi um atributo utilizado para validação do *crawler*. Por fim, o *link* era armazenado para posterior atualização dos comentários associados a essa notícia.

Em relação aos comentários, apesar da presença do nome do autor, entende-se que a plataforma provê o pseudoanonimato, motivo pela qual há uma alta prevalência de discurso de ódio. Fora as análises no conteúdo, as informações de interação permitem calcular o nível de aceitação ou rejeição do comentário, como descrito a seguir.

3.2 Métrica de Aceitação

O Portal G1 permite a interação de seus usuários dentro da plataforma de comentários e com isso, disponibiliza modos de resposta de forma textual e com interações na forma de "Gostei" e "Não Gostei". Os dados utilizados para extração desta métrica são as duas últimas formas. Com isso, é possível classificar um dado texto conforme a prevalência de interações e definindo-o nos grupos de aceitação ou rejeição. A Figura 2 exemplifica o funcionamento.

É demonstrado na Figura 2 um exemplo de comentário anonimizado expondo algumas características presentes na seção de comentários das notícias, também apresentando um texto com teor homofóbico na notícia na qual o mesmo foi publicado. Todos os comentários tiveram os seus respectivos autores e fotos de perfil anonimizados a fim de preservar a segurança e privacidade dos mesmos.

A presente métrica permite identificar o grau de aceitação de um dado comentário a partir da quantidade de interações que o

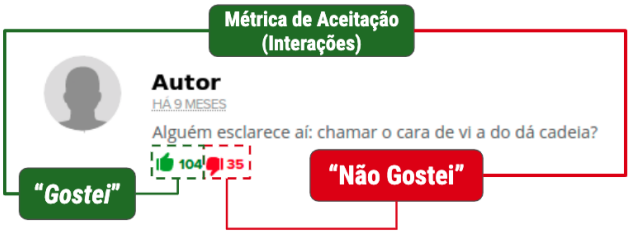

Figura 2: Exemplo de comentário para a métrica.

mesmo recebeu dos demais usuários da plataforma, capaz de ser aceito (maioria como "Gostei") ou rejeitado (prevalência de "Não Gostei").

3.3 Pré-processamento

As técnicas de pré-processamento utilizadas foram escolhidas com o o intuito de remover inconsistências do conjunto de dados a fim de melhorar a confiabilidade do resultado final [18, 26]. Os passos de pré-processamento foram implementados em *Python*, usando a biblioteca *Natural Language Toolkit* (*NLTK*)[6]. As tarefas de pré-processamento executadas (M) no conjunto de dados são divididas na transformação do conteúdo textual (TR) e remoção de itens (RM) [8], estando listadas na Tabela 3 e utilizando a frase de exemplo "*O Universo irá AumEnTar em 2050!*".

Tabela 3: Métodos de pré-processamento aplicados.

M	Método	Saída
TR	Caixa baixa	o universo irá aumentar em 2050!
RM	Stopwords	Universo irá aumEnTar 2050!
RM	Caracteres especiais	O Universo irá aumEnTar em 2050
RM	Caracteres numéricos	O Universo irá aumEnTar em !
RM	Acentuação	O Universo ira aumEnTar em 2050!
–	**Todos**	**universo ira aumentar**

Uma das etapas efetuadas em trabalhos anteriores, refere-se à remoção de *emoticons*[7] [27]. Esta tarefa não foi executada no pré-processamento dos dados aqui utilizados pois a mesma apresenta impactos nas análises, principalmente na fase da análise de sentimentos. Foram removidos dados duplicados e nulos do conjunto de dados, complementando os métodos aplicados na Tabela 3.

3.4 Análise de Sentimentos e Modelagem de Tópicos

A análise de sentimentos efetua a classificação de um conteúdo textual de acordo com a polaridade que o mesmo representa diante do documento ou texto [7]. As polaridades estão divididas em três principais, negativo, neutro e positivo [30]. Para a presente etapa, foi utilizada a ferramenta de PLN *Polyglot*[8] para *Python*, que utiliza um dicionário léxico para obtenção da polaridade de sua entrada. Como sua saída é um valor numérico real entre -1 e 1, o resultado

[5]https://scrapy.org/

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

[7]Figuras ou símbolos que caracterizam uma palavra ou ideia.

[8]https://polyglot.readthedocs.io/en/latest/

foi discretizado e utilizaram-se valores categorias para negativo ($-1.0 \leq P < 0.0$), neutro ($P = 0.0$) e positivo ($0.0 < P \leq 1.0$).

Para a visualização dos tópicos mais utilizados dentro do conjunto de comentários, foi utilizada uma metodologia similar para a análise de comentários extraídos em páginas de notícias [27]. O algoritmo aqui utilizado foi o *Non-Negative Matrix Factorization* (*NMF*). O *NMF* também está presente em diversos trabalhos que são bem aceitos dentro do campo das análises de mídias sociais [40]. Também por ter melhores resultados para as tarefas de mineração baseadas em textos curtos [21], característica das mídias sociais. O peso utilizado na matriz multi-dimensional de termos foi o *Term Frequency-Inverse Document Frequency* (*TF-IDF*).

O *TF-IDF* permite verificar estatisticamente a importância de uma palavra dentro de um ou mais documentos. O *TF* mede a frequência de um termo e resulta em um valor normalizado a partir da divisão pelo tamanho do documento (*i.e.* número de termos). O *IDF* ajusta o *TF* pois o mesmo define todos os vocábulos como igualmente relevantes. A definição dos tópicos é feita a partir de uma *Bag of Words* (*BoW*) e demonstra os dados textuais como uma matriz documento-termo, com a definição de pesos para cada termo ou palavra encontrada no conjunto.

3.5 Índice de Escolaridade

A etapa responsável pela identificação da dificuldade de um dado conteúdo textual é denominado como legibilidade [33]. Como forma de avaliação, existem diversos métodos matemáticos que representam esse índice a partir de fórmulas de legibilidade, onde uma das aplicadas com maior frequência é o Índice de *Flesch* [19].

O Índice de *Flesch* possui diversas variações, com uma delas sendo sua adaptação validada para o Português-Brasileiro e que aplica a classificação por níveis de escolaridade [23]. Os resultados para o *Flesch* em Português-Brasileiro podem ser classificados em quatro classes e estão listados na Tabela 4.

Tabela 4: Índices de escolaridade por legibilidade.

Pontuação	Legibilidade	Escolaridade
Abaixo de 25	Muito Difícil	Textos Acadêmicos
25 à 50	Difícil	Ensino Médio e Nível Superior
51 à 75	Fácil	5ª à 8ª Série
Acima de 75	Muito Fácil	1ª à 4ª Série

Com o Índice de Escolaridade, é possível verificar qual o grupo escolar dominante dentre os comentários a serem analisados. Isso permite visualizar características específicas dos textos publicados, como o nível de formalidade, preocupação gramatical e dificuldade do conteúdo escrito [6].

Este índice permite correlacionar por exemplo, o nível de escolaridade do autor da postagem com a prevalência de discurso de ódio. Uma hipótese do estudo é que usuários com alto nível de escolaridade tendem a escrever menos postagens com cunho ofensivo/pejorativo.

3.6 Detecção de discurso de ódio

Com grande parte dos métodos de detecção e classificação para discurso de ódio na literatura estando disponíveis para o idioma Inglês, têm-se a dificuldade em efetuar a execução desta etapa de forma totalmente precisa. A Tabela 5 apresenta as principais ferramentas e as técnicas utilizadas.

Como todas estas ferramentas foram desenvolvidas para a língua inglesa, foi necessário que se realizasse o processo de tradução dos comentários pré-processados. Foi utilizada a ferramenta *Google Translate*, usando a *API googletrans*[9] disponível para *Python*.

Por fim, destaca-se que as saídas foram padronizadas no formato binário, nas classes "Nenhum" e "Discurso de Ódio / Ofensivo". A Regressão Logística é aplicada por meio da ferramenta *HateSonar*[10] [10]. As Redes Neurais Convolucionais (*CNN*) [12], Recorrentes (*RNN*) [31] e de Memória de Curto e Longo Prazo (*LSTM*) [3] são implementações feitas na plataforma *Kaggle*, utilizando uma base de dados aberta[11] com 159.571 exemplos para o treinamento dos algoritmos.

Por conta da utilização do tradutor, perdas de precisão gerais ou estratificadas (classes binárias ou n-árias) com a geração de termos ou frases incorretas ou não traduzidas podem ser observadas, ocasionando em classificações incorretas. Portanto, foi necessária a inspeção manual dos resultados por meio de uma avaliação qualitativa, última fase da metodologia descrita na Figura 1.

4 RESULTADOS E DISCUSSÕES

A partir das palavras-chave definidas na etapa de coleta de dados e listadas na Seção 3, foram obtidos um total de 96.974 comentários extraídos de 3.561 notícias do Portal de Notícias G1. Uma análise exploratória no conjunto de dados permitiu verificar quais as palavras-chave com maior quantidade de comentários. A Figura 3 exibe essa distribuição.

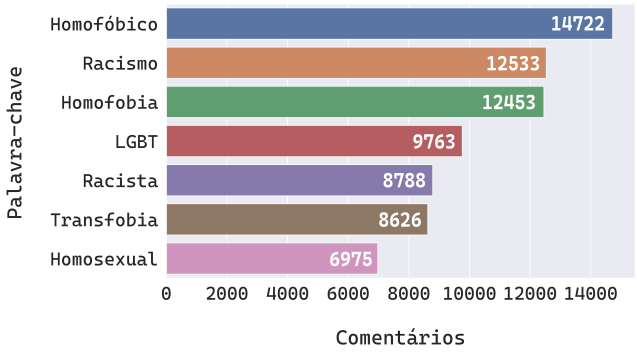

Figura 3: Quantidade de comentários por palavra-chave.

Entre as palavras-chave com maior presença, "*Homofóbico*" é caracterizada por notícias e acontecimentos recentes acerca da temática no Brasil, como a inserção da homofobia como um dos grupos que caracterizam o racismo. Consequentemente, as palavras relacionadas aos grupos de homofobia e racismo englobam os maiores grupos de comentários do conjunto de dados. A Figura 4 exibe a distribuição para cada grupo e complementa a Figura 3.

[9] https://github.com/ssut/py-googletrans
[10] https://github.com/Hironsan/HateSonar
[11] https://www.kaggle.com/mrinaal007/hate-speech-detection

Tabela 5: Metodologias das ferramentas utilizadas.

Nome	Metodologia	Classes
Hate Sonar [10]	Regressão Logística	3
Hate Speech Detection || 1D CNN || Glove Embedding [12]	Redes Neurais Convolucionais	2
Hate Speech Detection: RNN [31]	Redes Neurais Recorrentes	2
Classifying Hate Speech with a pyTorch Transformer [3]	Memória de Curto e Longo Prazo	2

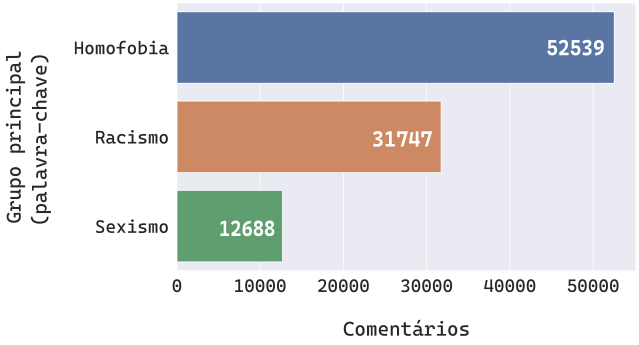

Figura 4: Quantidade de comentários por grupo alvo.

A maior prevalência é do grupo "*Homofobia*", com 54,2% do conjunto. Em seguida, "*Racismo*" compõe 32,7% e por fim, "*Sexismo*" com apenas 13,1%. A presente divisão dá-se por fatos que ocorreram nos últimos anos acerca dos temas e que geram uma maior interação dos usuários, seja por uma maior quantidade de notícias sobre um assunto específico ou pela publicação causar um impacto maior nos internautas. Na Figura 5 (A) e (B) é possível visualizar a quantidade de notícias e comentários por estado, respectivamente.

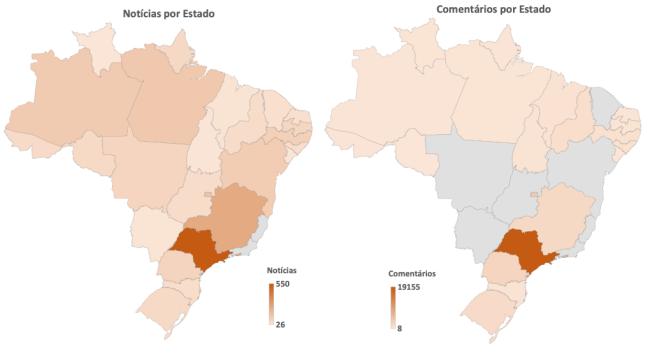

Figura 5: Quantidade de notícias (A) e comentários (B) por estado.

Conforme a Figura 5 (A), a maior quantidade de notícias está localizada na região sudeste e especificamente no estado de São Paulo. A Figura 5 (B) também indica os mesmos resultados. Tal fato é justificável pelo estado possuir a maior população do país e consequentemente, um maior potencial de ocorrerem mais fatos que levem a criação de mais notícias e das pessoas possuírem acesso as mídias sociais, gerando um maior impacto nas interações das notícias.

Conforme descrito na Seção 3, a etapa de pré-processamento foi executada para prosseguir com as demais análises baseadas em PLN. A partir disso e com o objetivo de verificar a polaridade presente nos textos publicados, a Figura 6 caracteriza a distribuição dos sentimentos.

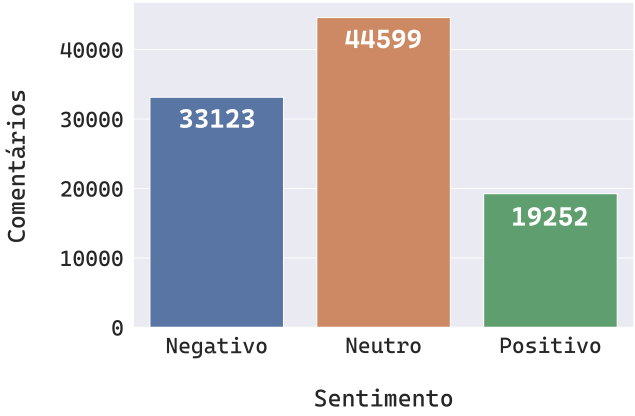

Figura 6: Análise de sentimentos dos comentários.

A Figura 6 deixa visível que a parte majoritária dos comentários estão categorizados dentro do sentimento neutro, com 44.599 itens. Logo após, a polaridade negativa contabilizou 33.123 e é seguida pela positiva com apenas 19.252, representando a menor parcela do conjunto. Mesmo com a maioria sendo definida como neutra, a diferença de apenas 11.476 para a classe negativa indica que há uma grande parcela de textos com teor pejorativo ou ofensivo dentre os coletados, com uma percentagem de 34,16%.

Comentários como "*O homossexualismo tem que ser combatido. Grande pecado que assola a terra.*" ou "*Os estrangeiros negros assassinam pessoas brancas... tudo bem. Uma pessoa branca assassina estrangeiros negros... ele é racista, crime de ódio, xenófobo, homofônico, assassino, perverso... e assim por diante...*" estão presentes dentre os dados e representam diversas formas de manifestações de discurso de ódio, categorizando a homofobia, o racismo e o sexismo, simbolizando uma pequena parte com sentimento negativo da Figura 6.

Quando feito um paralelo à análise de sentimentos, é possível inferir que mesmo a parcela negativa dos comentários possui uma taxa de aceitação superior ao de rejeição, com este sendo embasado a partir da Figura 7, a qual exibe que a parte majoritária dos comentários coletados e analisados possuem um índice de aceitação superior ao de rejeição, respectivamente de 64.331 (74,02%) contra 22.575 (25,98%), desconsiderando os com mesma quantidade

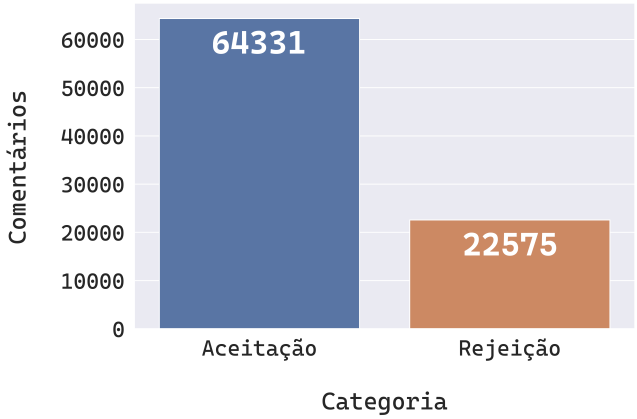

Figura 7: Taxa de aceitação dos comentários. Nota: *os comentários com quantidades iguais de interações foram desconsiderados neste gráfico.*

de interações. Esse fato indica que grande parcela dos internautas brasileiros possuem um pensamento e modo de agir que fomentam um discurso inflamado ou negativo. Essa proporção de aceitação também é presente nos grupos principais apresentados na distribuição da Figura 3, sendo alguns de seus exemplos apresentados na Tabela 6.

Com base na etapa de rotulação dos tópicos extraídos na modelagem de tópicos e conforme a Seção 3, a Tabela 7 lista os tópicos e palavras obtidos.

Tabela 7: Tópicos mais frequentes nos comentários.

Tópico	Palavras
Racismo	ser pode deve crime pessoas humano lei deveria racismo ter
Igualdade	sao pessoas negros estao mulheres pais maioria racismo brasil iguais
Política	brasil pais ter fazer presidente stf vc povo pode cara
Eleições	mito votar moro eleitores jair votos eleicoes fake acostumando tomara
Religiosidade	mulher homem mulheres deus homens casal criou nasce familia sexo

É possível identificar tópicos que abordam um conteúdo direcionado principalmente a temas como a política, o racismo e a igualdade. Palavras como "*presidente*", "*mulheres*", "*racismo*", "*crime*", "*eleitores*" e "*fake*" representam os tópicos supracitados. Em contraste com a Figura 3 e 4, os temas refletem tanto as palavras-chave com maior representação quanto as com menor presença dentro do conjunto de dados. Também indicam a presença de assuntos que destacaram-se no Brasil, como a cultura do ódio e a conexão com as *fake news*.

Para verificar o grau de escolaridade das postagens, seguiu-se o descrito na Seção 3 e foi feito o Índice de *Flesch*. Os resultados são apresentados na Figura 8.

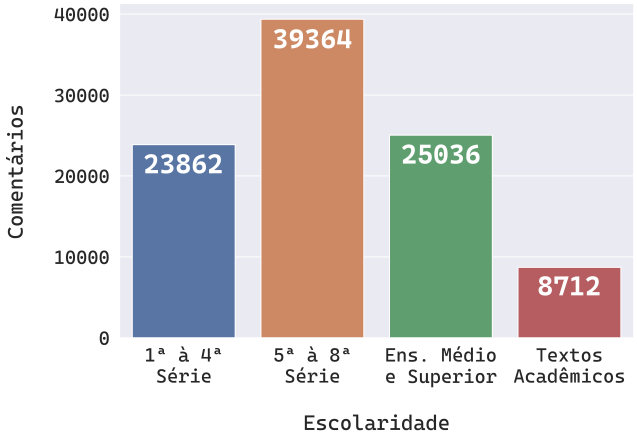

Figura 8: Índice de *Flesch* dos comentários.

A Figura 8 permite identificar o grau de escolaridade com maior representação, indicando o grupo de Quinta à Oitava Série como parte dominante com 39.364 comentários (40,6%). Em sequência, Ensino Médio e Nível Superior com 25.036 (25,8%); Primeira à Quarta Série com 23.862 (24,6%); e Textos Acadêmicos com 8.712 (9,0%).

Esses valores podem indicar dois principais fatores: i) Grande parte dos internautas utilizam uma linguagem mais informal e com menos preocupações no processo de digitação; e ii) Em relação com a Figura 6, há textos dos níveis mais superiores de escolaridade que podem estar caracterizados como discursos de ódio.

Neste ensejo, buscou-se executar diversas metodologias diferentes com base em implementações existentes para verificar a distribuição de comentários que possuem ou não discursos de ódio ou ofensivos, baseando-se na metodologia apresentada na Seção 3. A saída dos algoritmos é apresentada na Figura 9.

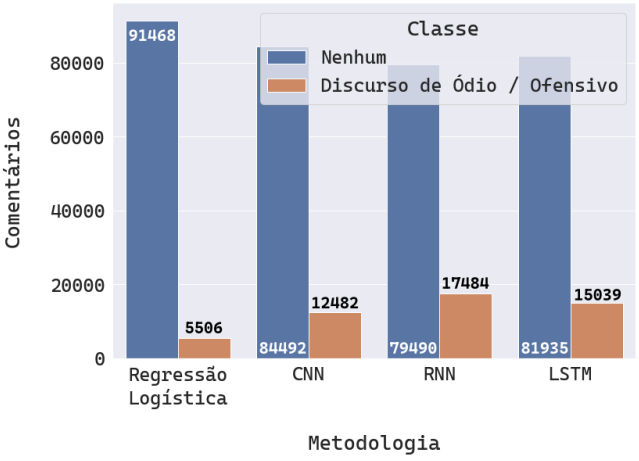

Figura 9: Comentários classificados por metodologia.

Para a classe "Nenhum", as quatro testadas resultaram em quantidades similares, tendo como destaque a Regressão Logística que contabilizou 91.468 comentários (94,32%) e menor para a *RNN*, com

Tabela 6: Comentários e interações por grupo principal.

N	Comentário (*SIC*)	Gostei (G)	Não Gostei (NG)	Grupo Principal	Índice de Aceitação
1	*Sistema de cotas e a maior forma de racismo*	665	128	Racismo	Aceitação (G > NG)
2	*PRA MIM COTAS É RACISMO CONTRA OS BRANCOS. O MAIS PREJUDICADO NESSES TEMPOS É O BRANCO POBRE.*	626	439	Racismo	Aceitação (G > NG)
3	*Negros ricos têm direito a cota e brancos pobres não? Grande equidade. Não me falem que negros foram explorados 300 anos no Brasil q*	580	112	Racismo	Aceitação (G > NG)
4	*Geração mi-mi-mi-!!! Agora tudo é homofobia!!!! Tomem vergonha na cara e tenha hombridade e caráter!!!*	838	125	Homofobia	Aceitação (G > NG)
5	*Esse negócio de LGBT e Femynismu já tá enchendo o saco, tá bom de passar logo essa fase... ngm aguenta mais isso!*	1105	147	Homofobia	Aceitação (G > NG)
6	*O homossexualismo como sempre destruindo lares e famílias... Afastando pais e filhos... Definitivamente isso não é coisa de Deus. CURA G.A.Y JÁ*	472	152	Homofobia	Aceitação (G > NG)
7	*Elas se hipersexualizam e querem ser vistas sem nenhuma atração sexual. Ta serto*	568	91	Sexismo	Aceitação (G > NG)
8	*Se for rico e bonito é paquera . Se for feio e pobre é assédio.*	315	102	Sexismo	Aceitação (G > NG)
9	*mulher que escolhe homem pelo dinheiro nao pode reclamar quando e tratada como mercadoria*	211	44	Sexismo	Aceitação (G > NG)

79.490 exemplos (81,97%). De forma oposta, na classe "Discurso de Ódio / Ofensivo" a maior parcela do conjunto foi assimilada pela *RNN* e a menor parte para a Regressão Logística, respectivamente com 17.484 (18,03%) e 5.506 (5,68%) comentários.

5 CONSIDERAÇÕES FINAIS

Neste artigo, foi feita uma análise das postagens efetuadas dentro dos temas de homofobia, racismo e sexismo no Portal de Notícias G1. Foram executadas tarefas de processamento de linguagem natural e análise de mídias sociais, a análise de sentimentos, modelagem de tópicos, visualização dos estados com maior prevalência de notícias e comentários, índice de aceitação e o índice de escolaridade. Em paralelo, foram verificadas diferentes metodologias e como elas classificavam os dados analisados com o intuito de detectar discurso de ódio ou ofensivo.

5.1 Perguntas de Pesquisa

Dada a **PP1** ("*Quais as características que representam a forma como os usuários interagem nas plataformas de notícias?*"), os usuários possuem uma tendência maior a interagir dentro de notícias relacionadas a homofobia e ao racismo, principalmente da região sudeste a exemplo de São Paulo. Quanto à **PP2** ("*Qual o teor dos comentários publicados em portais de notícia que apresentam pseudoanonimato?*"), os internautas também tendem a publicar textos com sentimento mais neutro e negativo, caracterizados por discursos de ódio, frases pejorativas, ofensivas ou que não são enquadradas nessas áreas. Dentre os tópicos mais abordados na plataforma, destacam-se os relacionados a política e racismo com palavras como "*racismo*", "*crime*", "*eleicoes*" e "*fake*", com um nível escolar dominante de 5ª à 8ª Série e seguido de Ensino Médio e Nível Superior. A presença de discurso de ódio ou ofensivo no conjunto de dados também converge com as demais análises, mantendo um padrão próximo entre as metodologias testadas.

5.2 Ameaças a validade do estudo

Cabe destacar algumas ameaças a validade do presente estudo: i) Comentários com seu conteúdo disfarçado (como números ou símbolos no lugar de algumas letras para dificultar a identificação diante de sistemas) gera um obstáculo para as análises de PLN (*e.g.* 4b3rr4ç4o); e ii) A carência de métodos para detecção de discurso de ódio em Português-Brasileiro exige a utilização de um tradutor para as existentes em inglês, causando perdas de acurácia e falhas textuais entre a troca de idiomas.

Adicionalmente, relacionada à análise de legibilidade dos comentários é possível presumir que os internautas com baixo nível de escolaridade não sejam capazes de produzir textos com um nível superior, como em nível médio, superior ou acadêmico. Em contrapartida, os que estão englobados nesses grupos, em tese, conseguem escrever textos mais simples e que se categorizam nas categorias mais simples, como 1ª à 4ª Série. Tais fatos são reforçados ainda mais pela natureza dos comentários, postados em uma plataforma de notícias e que são categorizadas como um tipo de mídia social, com a característica da natureza dos textos serem curtos e não criteriosos quanto a sua produção e composição gramatical, por exemplo.

5.3 Trabalhos futuros

Como trabalhos futuros, pretende-se melhorar o *framework* de análises, utilizando expressões regulares com estruturas de letras e números para que os mesmos sejam detectados e filtrados na etapa de pré-processamento, atualizar e anotar a base de dados manualmente ou a partir de *crowdsourcing* para que a mesma seja disponibilizada em repositórios de ciência aberta e por fim, utilizá-la na criação de um classificador que efetue a detecção de discursos de ódio e efetue a segmentação dos mesmos em subclasses, baseando-se nas metodologias estudadas e testadas neste artigo, com a possibilidade de tornar-se uma ferramenta que possa ser implementada para aprimorar sistemas atualmente existentes em mídias sociais.

AGRADECIMENTOS

Este trabalho foi financiado por meio do Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI) do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio da Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação Tecnológica da Universidade Federal do Oeste do Pará (PROPPIT/UFOPA). Agradecemos também aos revisores pelos comentários e sugestões que muito auxiliaram na construção e melhoria do manuscrito.

REFERÊNCIAS

[1] Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th International Conference on Computer Science and Information Technology*, Vol. 10.

[2] Gustavo Rangel de. Almeida, Douglas Rocha Cirqueira, and Fábio MF Lobato. 2017. Improving Social CRM through eletronic word-of-mouth: a case study of ReclameAqui. *XIV Workshop de Trabalhos de Iniciaçao Cientıfica (WTIC 2017)* (2017).

[3] Nader Atef. 2020. Classifying Hate Speech with a pyTorch Transformer. https://www.kaggle.com/nadergo/classifying-hate-speech-with-a-pytorch-transformer

[4] Brasil. 2014. Lei N. 12.965, de 23 de abril de 2014. In *Marco Civil da Internet*. Brasília, DF.

[5] Brasil. 2018. Lei N. 13.709, de 14 de agosto de 2018. In *Lei Geral de Proteção de Dados Pessoais (LGPD)*. Brasília, DF.

[6] Asad J Choudhry, Yaser MK Baghdadi, Amy E Wagie, Elizabeth B Habermann, Stephanie F Heller, Donald H Jenkins, Daniel C Cullinane, and Martin D Zielinski. 2016. Readability of discharge summaries: with what level of information are we dismissing our patients? *The American Journal of Surgery* 211, 3 (2016), 631–636.

[7] Douglas Cirqueira, Fernando Almeida, Gültekin Cakir, Antonio Jacob, Fabio Lobato, Marija Bezbradica, and Markus Helfert. 2020. Explainable sentiment analysis application for social media crisis management in retail. (2020).

[8] Douglas Cirqueira, Marcia Fontes Pinheiro, Antonio Jacob, Fabio Lobato, and Adamo Santana. 2018. A Literature Review in Preprocessing for Sentiment Analysis for Brazilian Portuguese Social Media. In *Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 746–749. https://doi.org/10.1109/WI.2018.00008

[9] Alexandre Martins da Cunha, Isabela Santos, Daniel Pedrosa, Francis F. Steen, Mark Turner, Maira Avelar, Lilian Ferrari, and Gustavo Paiva Guedes. 2018. Sentiment Analysis on Brazilian News Broadcast Data. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining* (Natal). SBC, Porto Alegre, RS, Brasil. https://doi.org/10.5753/brasnam.2018.3599

[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[11] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. 86–95.

[12] Deekshitha Erlapally. 2020. Hate Speech Detection || 1D CNN || Glove Embedding. https://www.kaggle.com/deekshithaerlapally/hate-speech-detection-1d-cnn-glove-embedding

[13] T. Fernandes. 2019. Crimes Sexuais Pela Internet: A Violência Contra A Mulher Entre O Real E O Virtual.

[14] Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*. 94–104.

[15] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 85.

[16] Phyllis B Gerstenfeld. 2017. Hate Crime. *The Wiley Handbook of Violence and Aggression* (2017), 1–13.

[17] Ravonne A Green. 2014. The Delphi technique in educational research. *Sage Open* 4, 2 (2014), 2158244014529773.

[18] Lucas Marques Sathler Guimarães, Magali Rezende Gouvêa Meireles, and Paulo Eduardo Maciel de Almeida. 2019. Avaliação das etapas de pré-processamento e de treinamento em algoritmos de classificação de textos no contexto da recuperação da informação. *Perspectivas em Ciência da Informação* 24, 1 (2019), 169–190.

[19] James Hartley. 2016. Is time up for the Flesch measure of reading ease? *Scientometrics* 107, 3 (2016), 1523–1526.

[20] Emanuel Gilvan Souza Lima Júnior, Gustavo Nogueira de Sousa, Antonio Fernando Lavareda Jacob Junior, and Fábio Manoel França Lobato. 2020. Ferramentas para Análise de Mídias Sociais: Um Levantamento Sistemático. *Anais do Computer on the Beach* 11, 1 (2020), 389–396.

[21] Luiz F Junior, Jorge Silva Junior, and Fábio Lobato. 2020. Um olhar sobre turismo gastronômico: Um caso no TripAdvisor. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 519–530.

[22] Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. (dec 2017). arXiv:1712.06427

[23] Teresa B. F. Martins, Claudete M. Ghiraldelo, Maria das Graças Volpe Nunes, and Osvaldo Novais de Oliveira Junior. 1996. Readability formulas applied to textbooks in brazilian portuguese. (1996).

[24] Rafael Sales Medina, Ana Paula Couto da Silva, and Fabricio Murai. 2018. Análise das Interações Sociais em Comunidades Online de Aprendizado de Idiomas: um estudo de caso no Reddit. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining* (Natal). SBC, Porto Alegre, RS, Brasil. https://doi.org/10.5753/brasnam.2018.3576

[25] Anna Motz. 2016. *The psychology of female violence: Crimes against the body*. Routledge.

[26] Lucas Rodrigues, Ademir Junior, and Fabio Lobato. 2020. Notícias relacionadas a pessoas com deficiência: uma análise do conteúdo gerado pelos usuários em postagens de mídias sociais. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional* (Salvador). SBC, Porto Alegre, RS, Brasil, 811–822. https://sol.sbc.org.br/index.php/eniac/article/view/9336

[27] Lucas Darlindo Freitas Rodrigues, Jorge Luiz Figueira da Silva Junior, and Fábio Manoel França Lobato. 2019. A culpa é dela! É isso o que dizem nos comentários das notıcias sobre a tentativa de feminicıdio de Elaine Caparroz. In *Proceedings of the 8th Brazilian Workshop on Social Network Analysis and Mining*. SBC, Porto Alegre, RS, Brasil, 47–58.

[28] Natalie Rosa. 2019. Brasil registra aumento de 1.600% em denúncias de crimes online contra mulheres. https://canaltech.com.br/seguranca/brasil-registra-aumento-de-1600-em-denuncias-de-crimes-online-contra-mulheres-132103/

[29] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118* (2017).

[30] Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. 2016. Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management* 52, 1 (2016), 5–19.

[31] Nayan Sakhiya. 2020. Hate Speech Detection: RNN. https://www.kaggle.com/nayansakhiya/hate-speech-detection-rnn

[32] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. 1–10.

[33] Karen A Schriver. 1989. Evaluating text quality: The continuum from text-focused to reader-focused methods. *IEEE Transactions on professional communication* 32, 4 (1989), 238–255.

[34] Senado. 2019. Preocupação com aumento de feminicídios no Brasil motiva debate na CDH.

[35] Samuel C Silva and Adriane BS Serapião. 2018. Detecção de discurso de ódio em português usando CNN combinada a vetores de palavras. (2018).

[36] Gustavo Nogueira de Sousa, Isabelle da Silva Guimarães, Julio Augusto Nogueira Viana, Olaf Reinhold, Antonio Fernando Lavareda Jacob Junior, and Fábio Manoel França Lobato. 2020. Análise do setor de telecomunicação brasileiro: Uma visão sobre Reclamações. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação* 37 (2020), 31–48.

[37] Marcelo Augusto Muniz Teixeira, Fábio Manoel França Lobato, Beatriz Nery Rodrigues Chagas, and Antonio Fernando Lavareda Jacob Junior. 2018. A System of Acquisition and Analysis of Data for Extraction of Knowledge of the Ebit Platform. In *Proceedings of the 15th International Conference on Information Systems & Technology Management*. 4195–4206.

[38] Luiz Trindade. 2018. Brazil's supposed 'racial democracy' has a dire problem with online racism.

[39] Ahmed Waqas, Joni Salminen, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. 2019. Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PloS one* 14, 9 (2019), e0222194.

[40] Trefor Williams and John Betak. 2018. A Comparison of LSA and LDA for the Analysis of Railroad Accident Text. In *Proceedings of the 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018)*. 98–102.

[41] Jennifer Yesbeck. 2018. How are Alexa's traffic rankings determined? https://blog.alexa.com/improving-your-alexa-rank/