Abordagem Semi-Supervisionada para Anotação de Linguagem Tóxica

Francisco A. R. Neto[1,2], **Rafael T. Anchiêta**[2],
Raimundo S. Moura[1], **André M. Santana**[1]

[1]Departamento de Computação – Universidade Federal do Piauí (UFPI), Brasil

[2]Instituto Federal do Piauí (IFPI), Brasil

{farn,rta}@ifpi.edu.br, {rsm,,andremacedo}@ufpi.edu.br

Abstract. *Toxic messages pose serious problems on social media platforms, as they are used to harm individuals, groups, or organizations. Automatic methods for combating Hate Speech require good linguistic resources, such as corpora. The manual construction of toxic language corpora presents significant challenges due to the strong subjectivity associated with the concept of Hate Speech and the difficulty in properly training annotators. The solution to this problem involves creating alternatives for data annotation. This work presents a semi-supervised technique, based on heterogeneous graphs, for automatically detecting and annotating toxic language. This approach was evaluated on the ToLD-BR corpus, and moderate agreement was shown with its original labels.*

Resumo. *Mensagens tóxicas acarretam sérios problemas nas plataformas de redes sociais, uma vez que são usadas para prejudicar indivíduos, grupos ou organizações. Os métodos automáticos de combate ao Discurso de Ódio precisam de bons recursos linguísticos, como corpora. A construção manual de corpus de linguagem tóxica impõe desafios significativos devido à forte subjetividade associada ao conceito de Discurso de Ódio e à dificuldade em treinar adequadamente anotadores. A solução deste problema passa pela criação de alternativas para a anotação de dados. Este trabalho apresenta uma técnica semi-supervisionada, baseada em grafo heterogêneo, para detecção e anotação automática de linguagem tóxica. Essa abordagem foi avaliada sobre o corpus ToLD-BR e apresentou nível de concordância moderada com seus rótulos originais.*

1. Introdução

A Internet deu espaço a diversas comunidades de usuários à conexão, comunicação e compartilhamento de informações. Uma dessas possibilidades é através das Redes Sociais, como *X* (anteriormente *Twitter*), *Facebook* ou *Instagram*. A medida que crescem o número de usuários nas Redes Sociais, proporcionamente aumenta o volume de dados a serem processados, e é neste cenário que indivíduos mal intencionados aproveitam as vulnerabilidades destas plataformas e abusam da liberdade de expressão para espalharem mensagens tóxicas.

As mensagens tóxicas envolvem o uso de linguagem inadequada que é considerada inaceitável, incluindo tanto formas explícitas ou implícitas de palavrões, insultos e

ameaças dirigidas a indivíduos ou grupos [Zampieri et al. 2019]. Essa toxicidade também pode se manifestar na forma de comportamentos negativos, tais como comentários rudes ou observações desrespeitosas, contendo Discurso de Ódio ou qualquer outra característica que possa afastar um ser humano de uma conversação[1].

Apesar desse tipo de linguagem ser considerada como crime em diversos países, inclusive no Brasil, a quantidade de mensagens tóxicas na Internet vem aumentando e isso dificulta a fiscalização e punição dos infratores. De acordo com estatísticas sobre ódio no Brasil[2], em 2022 foram registrados mais de 74K denúncias de casos de Discurso de Ódio pela Internet, um aumento de 67,7% em relação a 2021. Os crimes que mais cresceram foram a xenofobia (874%), a intolerância religiosa (456%) e a misoginia (251%).

A maioria das abordagens que buscam combater linguagem tóxica são baseadas em algoritmos de Aprendizagem de Máquina Supervisionada (AMS), os quais necessitam de valorosos *corpora* e com boa quantidade de dados para que os algoritmos aprendam as características que identificam o Discurso de Ódio. No entanto, a criação de *corpora* linguísticos sobre Discurso de Ódio é uma tarefa desafiadora devido ao vasto grau de subjetividade relacionado aos conceitos de linguagem tóxica e Discurso de Ódio [Ross et al. 2016]. Além disso, posicionamentos ideológicos e predileções pessoais afetam a anotação de dados tóxicos [Aroyo et al. 2019, Hettiachchi et al. 2023].

Nesse sentido, a construção manual de *corpora* linguísticos para Discurso de Ódio precisa de avaliadores humanos muito bem treinados e com os conceitos e direcionamentos definidos para a tarefa [Ross et al. 2016]. Os dados rotulados de forma inadequada (i.e., humanos sem conhecimento, treinamento, ou ainda condicionados a posicionamentos ideológicos), podem introduzir vieses e afetar diretamente a qualidade dos modelos de detecção de Discurso de Ódio [Nascimento et al. 2022]. Esses fatores impõem fortes obstáculos, pois a captação de humanos para a tarefa se torna ainda mais custosa, mesmo com o uso de ferramentas de *crowdsourcing*.

Uma alternativa para contornar os problemas descritos é automatizar a etapa de anotação dos dados. Sistemas inteligentes bem treinados podem desempenhar de forma satisfatória tarefas de especialistas [Russell and Norvig 2009], e depois, a automatização pode evitar a influência de fatores políticos/sociais (e.g., ideologia, integridade moral), inerentes a seres humanos, interfiram na tarefa da etiquetação de textos tóxicos. A anotação automática dos dados, além de permitir realizar o trabalho de forma menos custosa, também viabiliza rotular *corpus* com qualquer quantidade de dados.

O presente trabalho apresenta uma estratégia semi-supervisionada, baseada em grafos heterogêneos, para detecção e anotação automática de linguagem tóxica. Para o treinamento do grafo heterogêneo, foi construído o *corpus* Toxic-BR, curado, cuja anotação foi realizada pelas Perspective API [Lees et al. 2022] e ChatGPT [Brown et al. 2020]. Os dados do Toxic-BR foram curados por quatro avaliadores humanos. A avaliação desta técnica semi-supervisionada foi conduzida sobre o *corpus* ToLD-BR [Leite et al. 2020], obtendo resultado 0,693 de *f-measure* e 43,1% Kappa de Cohen, o

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview

[2] https://agenciabrasil.ebc.com.br/direitos-humanos/noticia/2023-02/denuncias-de-crimes-na-internet-com-discurso-de-odio-crescem-em-2022

que mostra nível de concordância moderada com os rótulos originais do *corpus*.

O restante deste artigo está estruturado da seguinte forma. Seção 2 descreve trabalhos relacionados que apresentaram *corpora* linguísticos de linguagem tóxica e Discurso de Ódio, e destaca os principais desafios encontrados na anotação manual desse tipo de conteúdo não aceitável. Seção 3 descreve o *corpus* Toxic-BR. O método proposto para anotação automática de textos tóxicos é apresentado na Seção 4. A Seção 5 detalha os experimentos realizados e apresenta os resultados obtidos. Por fim, Seção 6 conclui o artigo e apresenta direcionamentos para trabalhos futuros.

2. Trabalhos Relacionados

Os *corpora* de Discurso de Ódio são fundamentais para desenvolvimento de validação de modelos de AMS capazes de detectar linguagem tóxica. [de Pelle and Moreira 2017] descrevem OFFCOMBR, um *corpus* manualmente anotado de comentários ofensivos extraídos das seções de notícias sobre políticas e esportes do site g1.globo.com. Os autores relataram que extraíram 10k comentários do site, no entanto, devido as dificuldades relacionadas a avaliadores humanos disponíveis para anotação dos dados só conseguiram trabalhar com 1250. Foram utilizados três avaliadores para anotação do *corpus* para categoria binária ofensivo vs. não ofensivo, e tipos de Discurso de Ódio. As etiquetas dos dados foram determinadas através do voto majoritário e a medida de concordância entre os anotadores foi 71% de Kappa de Fleiss.

[Fortuna et al. 2019] apresentam o *corpus* HLPHSD com textos de Discurso de ódio, hierarquicamente rotulado para o português. O *corpus* possui 5668 *tweets* com anotação binária de Discurso de Ódio e categorias de Discurso de Ódio de organizadas de forma hierárquica (e.g., Sexismo, Homofobia, Racismo etc.). A anotação dos dados foi conduzida por 18 humanos, e para avaliação do nível de concordância entre os avaliadores foi utilizado a medida Kappa de Fleiss obtendo resultado de 17% para anotação binária e Kappa de Cohen de 72% para anotação hierárquica.

O ToLD-BR é o *corpus* com maior quantidade de textos de linguagem tóxica para o português [Leite et al. 2020]. A base de dados é formada por 21K *tweets* anotados manualmente em categorias de Discurso de Ódio como racismo, linguagem obscena, insulto, xenofobia e outros. A anotação do ToLD-BR foi conduzida por 48 colaboradores selecionados com base em informações demográficas. Cada colaborador ficou responsável por avaliar 1500 *tweets*, sendo que cada *tweet* foi julgado por 3 humanos. A medida de concordância reportada no trabalho foi Alfa de Krippendorff com resultado médio entre todas as categorias de Discurso de Ódio de 55%.

[Vargas et al. 2022b] apresentam o HateBR, uma base de dados com 7K comentários de usuários em postagens de perfis de políticos brasileiros na rede social Instagram. Os textos foram anotados por três avaliadores de alta formação educacional, na forma binária de linguagem ofensiva, nível de ofensividade e categorias de Discurso de Ódio. A avaliação da anotação binária da linguagem ofensiva obteve 74% Kappa de Fleiss e a média 75% na medida Kappa de Cohen. Por fim, a avaliação dos níveis de ofensividade reporta 46% e 47% para as medidas Kappa de Fleiss e Kappa de Cohen.

A forte subjetividade associada aos conceitos de linguagem tóxica e Discurso de Ódio dificultam a tarefa de anotação de dados. Estudos já demonstraram *corpora* de baixa

confiabiliadade, mesmo estes sendo anotados com fornecimento de orientações aos avaliadores, ou ainda utilizando especialistas para a tarefa [Ross et al. 2016, Waseem 2016].

Os *corpora* de Discurso de Ódio para língua portuguesa descritos anteriormente foram construídos com fornecimento de orientações aos avaliadores dos dados. Somente [Vargas et al. 2022b] cita a utilização de especilistas na anotação dos dados. Nenhuma das pesquisas explorarou como predileções pessoais dos anotadores poderiam afetar a qualidade dos *corpora*.

Conforme [Aroyo et al. 2019], predileções pessoais no momento da anotação de textos tóxicos podem levar a um julgamento errôneo para conteúdos de um mesmo domínio. [Hettiachchi et al. 2023] demostraram que posição ideológica/política, integridade moral, tratos pessoais e atitudes sexistas impactam na rotulação de textos de sexistas e misóginos.

Diante do exposto, evidencia-se a necessidade de trabalhar alternativas para construção de *corpus* de Discurso de Ódio. A automatização da etapa de anotação dos dados pode ser uma solução viável, já que sistemas inteligentes bem treinados são capazes de reproduzir tarefas de especialistas [Russell and Norvig 2009]. Além disso, a automatização pode evitar que a influência de fatores políticos/sociais (e.g., ideologia, integridade moral) inerentes à seres humanos interfiram na tarefa de anotação de textos tóxicos.

3. Corpus Toxic-BR

3.1. Coleta de dados

Os dados textuais deste trabalho foram coletados através da API da plataforma *X* durante o segundo turno das eleições presidenciais brasileiras de 2022, até 16 dias após seu término. O cenário de extrema polarização, que já vinha tomando forma antes do primeiro turno das eleições presidenciais[3], transformou as redes sociais em um terreno fértil para a disseminação de desinformações e mensagens tóxicas. A fim de evitar qualquer introdução de viés durante a coleta dos *tweets*, não foram utilizadas palavras-chave tóxicas ou consultas de perfis de pessoas públicas que representassem algum partido político. Ao invés disso, a estratégia conduzida baseou-se somente na palavra-chave “eleições”, capturando todo tipo de conteúdo textual voltado ao tópico.

Através dessa estratégia, foram extraídos 6 milhões de *tweets*. Com intuito de evitar conteúdos pouco informativos, ou que somente citassem os nomes dos candidatos a presidência, foram considerados somentes *tweets* com mais de 5 *tokens* resultando no total de 447.957 *tweets*. Em seguida, foi selecionada uma amostra aleatória de 150k para pré-processamento e posterior anotação[4].

Para o pré-processamento, removeu-se 6.097 *tweets* duplicados, além de remover *hashtags*, menções a usuários, *emojis* e indicadores de *retweets* (RT) quando o texto é uma citação a outro *tweet*. O final dessa etapa resultou em uma amostra de 143.854 *tweets*.

[3]https://www.poder360.com.br/eleicoes/eleicao-de-2022-e-a-mais-polarizada-desde-a-redemocratizacao/

[4]Esta quantidade foi definida em razão do crédito disponível para utilização do ChatGPT.

3.2. Ferramentas de anotação

Os *tweets* coletados foram anotados de forma automática através das ferramentas ChatGPT [Brown et al. 2020] e Perspective API [Lees et al. 2022]. Na literatura, ambas ferramentas têm sido utilizadas para detecção de Discurso de Ódio [Oliveira et al. 2023, Lees et al. 2022]. O objetivo é verificar o nível de concordância entre as ferramentas automáticas na tarefa de anotações de textos tóxicos.

De acordo com [Brown et al. 2020], o ChatGPT é um poderoso modelo de língua capaz de executar de forma competente diversas tarefas no campo de Processamento de Linguagem Natural, como tradução, perguntas-respostas, geração de texto, entre outras. Nesse contexto, utilizou-se o ChatGPT versão 3.0 como uma ferramenta de classificação de texto para rotulação dos dados tóxicos. Para a condução da tarefa, foi feito um *fine-tuning* com o *corpus* ToLD-BR [Leite et al. 2020] e configurado de forma empírica dois parâmetros, o **modelo de otimização** "ADA" e **temperatura** com valor igual a zero, que é responsável pela aleatoriedade das repostas do modelo. O baixo valor da **temperatura** favorece um comportamento determinístico retornando poucas respostas aleatórias.

A Perspective API é um serviço web que mensura a toxicidade dada uma entrada textual, seja ela palavras, sentenças ou parágrafo [Lees et al. 2022]. Ao analisar um texto, a Perspective API retorna um índice que corresponderá à probabilidade de toxicidade, indicando quão tóxico é o texto de entrada. Quanto maior for o valor do índice, ou seja, próximo de 1, maior é a possibilidade de o texto ser tóxico. Caso contrário, índice próximo de 0 a probabilidade de ser tóxico é baixa. Para a tarefa de anotação, todos os *tweets* foram analisados e receberam um índice correspondente a probabilidade de toxicidade. Foi estabelecido de forma empírica que todos os *tweets* com índices maiores que 0,8 seriam rotulados como tóxicos, e os textos com índices abaixo considerados como não-tóxicos.

A Tabela 1 mostra o resultado da anotação pelas ferramentas automáticas. O primeiro conjunto consiste nos dados divergentes, onde as ferramentas identificaram textos com rótulos distintos. O segundo conjunto representa os textos em que tanto ChatGPT quanto a Perpective API concordaram com a etiqueta tóxico, enquanto o terceiro conjunto consiste nos *tweets* rotulados como não tóxicos por ambas as ferramentas. Por fim, a averiguação do processo de anotação foi através do índice Kappa de Cohen o qual mede o grau de concordância entre os anotadores, ChatGPT e Perspective API. As ferramentas apresentaram concordância moderada, indicada pelo Kappa de Cohen igual a 42%.

Tabela 1. Distribuição dos tweets no *corpus* Toxic-BR.

	Divergentes	**Tóxicos**	**Não tóxicos**
# tweets	19.840	9.505	114.509
Total		143.854	

3.3. Curagem dos dados

A curagem dos dados do Toxic-Br tem como objetivo avaliar a qualidade da anotação automática no *corpus*. Esta tarefa busca observar o comportamento das ferramentas na anotação e determinar a confiabilidade delas na etiquetação dos dados. Nesse sentido,

foram selecionados quatro colaboradores para validação dos dados do Toxic-BR. Todos os avaliadores possuem educação superior completa, sendo 3 especialistas da área de Processamento de Linguagem Natural. Como todos possuem bom nível de instrução, foi apresentado aos colaboradores somente o conceito de linguagem tóxica segundo [Leite et al. 2020] para nortearem suas decisões.

Para curagem do Toxic-BR, foram selecionadas 1.400 amostras aleatórias do *corpus*, sendo 1.000 *tweets* divergentes, 200 *tweets* tóxicos e 200 *tweets* não tóxicos. Deste modo, cada colaborador realizou a curagem de 250 *tweets* divergentes, 50 *tweets* tóxicos, e 50 *tweets* não tóxicos. Os *tweets* divergentes receberam mais atenção pois era necessário investigar quais características nos textos levaram as ferramentas a discordância e qual das ferramentas possuía uma avaliação mais próxima à humana. Ao fim, o Toxic-BR curado[5] apresentou 615 *tweets* tóxicos, e 785 não tóxicos, como mostra a Tabela 2.

Tabela 2. Dados do Toxic-BR curado.

	Tóxicos	**Não toxicos**
# tweets	615	785
Total	1400	

A Tabela 3 ilustra a avaliação de um dos colaboradores a respeito do conjunto divergente. É possível observar que a Perspective API reconhece a maioria dos tweets como não tóxicos, mesmo em casos em que foram determinados como falsos negativos. A característica predominante destes tweets é a presença de ofensas sutis na forma de ironia e apelidos (e.g., "bandido", "Boso"), conforme exemplos na Tabela 4. Em alguns casos, a Persperctive API não foi capaz de reconhecer expressões ofensivas como "cagam na nossa cabeça".

Tabela 3. Matrizes de confusão da avaliação de um dos colaboradores sobre o conjunto divergente.

		Perspective API	
		Tóxico	Não tóxico
Correto	Tóxico	19	96
	Não tóxico	41	94

		GPT 3	
		Tóxico	Não tóxico
Correto	Tóxico	96	19
	Não tóxico	94	41

Por outro lado, o ChatGPT identificou uma grande quantidade de *tweets* como sendo tóxicos, ainda que vários sendo falsos positivos. Em alguns casos, ChatGPT associa a etiqueta tóxica aos *tweets* somente pela presença de apelidos, mesmo que não haja expressões tóxicas no texto (e.g., "Xandao" e "Mito"). Essa mesma característica permitiu a anotação correta quando identificou os termos "ex-presidiário", "bandido" e "Boso".

Os exemplos na Tabela 4 demonstram que tanto a Perspective API e o ChatGPT baseam fortemente suas anotações em expressões como os apelidos. Enquanto alguns desses termos podem conter conotações abusivas, outros não possuem essa qualidade

[5]Disponível em `https://github.com/fricarteneto/Toxic-BR`

Tabela 4. Exemplos de Discordância.

Texto	Rótulo		
	Perspective	ChatGPT	Humano
Nao amigo. E porque a grande midia aqui nao tem interesse, eles sao do partido do candidato ex-presidiario que roubou as eleicoes. Por isso nao mostram as manifestacoes.	0	1	1
Jamais esse Bandido,.não ganhou as eleições o TSE deu pra ele!	0	1	1
Cara o pedido de anulacao das eleicoes sera feito a quem ? STF ou TSE? Eles cagam na nossa cabeca...	0	1	1
Amanha lula vencera as eleicoes e ao GADO ja aviso, nao me venha de MIMIMIMI.	1	0	1
E UMA ABSURDO QUE AS ELEICOES SEJAM ADIADAS! MAS PELO MITO, EU FACO TUDO	0	1	0
Critica o Xandao e as eleicoes so pra gente testar uma coisa ai.	0	1	0
Vai lá, sua boquinha tá quase garantida, só se ganhar as eleições, até lá vá contra os seus ideias, fecha os olhos e continue se curvando para o Boso	0	1	1

(e.g., “Mito”, “Xandão”). Tal característica deve ocorrer devido a forte associação dos apelidos a termos abusivos dentro do *corpus*, e assim determinar o direcionamento da classificação das ferramentas.

A Tabela 5 sumariza o resultado geral da anotação dos quatro colaboradores. A mesma característica é percebida nesses resultados, ou seja, a Perspective API reconhece mais *tweets* não tóxicos enquanto o ChatGPT reconhece *tweets* tóxicos.

Tabela 5. Matrizes de confusão da avaliação de todo conjunto divergente.

		Perspective API	
		Tóxico	Não tóxico
Correto	Tóxico	81	338
	Não tóxico	223	358

		GPT 3	
		Tóxico	Não tóxico
Correto	Tóxico	338	81
	Não tóxico	358	223

A respeito dos *tweets* onde as ferramentas concordaram com a mesma etiqueta (i.e., 200 *tweets* tóxicos e *tweets* não tóxicos), somente 42 dos 400 foram considerados com a classificação incorreta pelos avaliadores e tiveram suas etiquetas modificadas (ver Tabela 6). Os 19 *tweets* que tiveram mudança de não tóxico para tóxico apresentaram ofensas sutis, seja através de ironias ou metáforas, características difíceis de serem anali-

sadas até por humanos. Alguns neologismos presentes nos textos definiram o sentido da anotação como o caso "patriotario". A Tabela 7 exemplifica alguns desses comentários.

Tabela 6. Número de tweets modificados com as avaliações.

	Não tóxico para Tóxico	Tóxico para Não tóxico
Humano 1	6	9
Humano 2	0	2
Humano 3	8	3
Humano 4	5	9
Total	19/200	23/200

Tabela 7. Exemplos de tweets avalidados que mudaram de classe Não tóxica para Tóxica.

Texto	**Rótulo**		
	Perspective	**ChatGPT**	**Humano**
Foi mostrado 180 dias antes das eleicoes, patriotario!	0	0	1
Gente mas a intenção é essa ,o sistema só tirou lula pra concorrer as eleições, pq sabem que lula tem voto,o próprio sistema vai derrubar molusco.!	0	0	1

Dos 23 *tweets* que foram modificados de tóxicos para não tóxicos, é possível observar a presença de termos de sentimento negativo (e.g., "porrada", "eleições sujas", "urnas fraudadas") que podem ter enganado as ferramentas nas anotações dos dados. Exemplos são apresentados na Tabela 8.

Tabela 8. Exemplos de tweets avalidados que mudaram de classe Tóxica para Não tóxica.

Texto	**Rótulo**		
	Perspective	**ChatGPT**	**Humano**
Lula vs bolsonaro, acho que as eleições deveriam ser resolvidas na porrada kkkkk	1	1	0
Não aceitaremos eleições sujas! E até tu que é de esquerda sabe!	1	1	0

4. Grafo heterogêneo

A metodologia para detecção e anotação de linguagem tóxica é organizada em quatro etapas, como mostra a Figura 1. As Subseções 4.1, 4.2, 4.3, 4.4 descrevem as etapas.

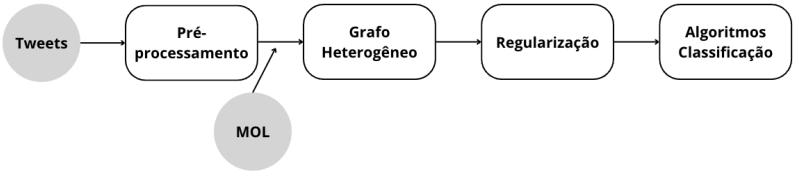

Figura 1. Metodologia para anotação dos textos tóxicos.

4.1. Pré-Processamento

Os *tweets* são textos curtos e geralmente escritos em linguagem coloquial, sendo comuns erros ortográficos e abreviações de palavras. Logo, foi utilizada a ferramenta Enelvo [Costa Bertaglia and Volpe Nunes 2016] para normalização e correção de erros em palavras mal escritas, tornando os textos mais fáceis de serem analisados.

4.2. Modelagem do Grafo Heterogêneo

Os *tweets* tóxicos foram modelados como um grafo heterogêneo, que são estruturas ricas em informações por conta dos relacionamentos entre nós de características ou tipos distintos [Rossi 2015].

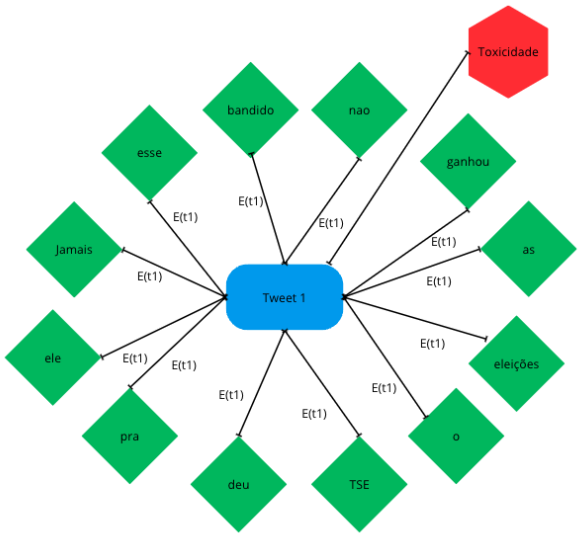

Figura 2. Exemplo da estrutura do grafo para a sentença *"Jamais esse Bandido,.não ganhou as eleições o TSE deu pra ele!"*

A estrutura do grafo foi inspirada no trabalho [Saraiva et al. 2021], definida como uma rede não direcional e arestas ponderadas *G = (V, E, W)*, sendo *V* conjunto de vértices $V = (v_1, v_2, ..., v_n)$, *E* o conjunto de arestas $E = (e_1, e_2, ...e_m)$ e W indica a matriz de adjacência representando os pesos das arestas $W_{i,j}$ que conectam os vértices *i* e *j*. Diferente de [Saraiva et al. 2021], foram definidos três tipos de vértices, os nós que representam os ***tweets***, os nós ***tokens*** que são termos presentes nos *tweets* e, por fim, os nós **Toxicidade** que consiste no grau de ofensividade definido pelos termos ofensivos do *tweet*.

O cálculo da Toxicidade dos *tweets* é obtido através dos termos ofensivos presentes no texto utilizando o léxico MOL [Vargas et al. 2022a]. O índice de toxicidade dos termos é definido com auxílio da Perspective API [Lees et al. 2022], e a soma dos índices de toxicidade apresentada pelos termos ofensivos define o valor da **Toxicidade** do *tweet*.

As arestas que interligam os vértices ***Tweet*** e ***Tokens*** recebem um peso que corresponde a média das *embeddings* dos termos presentes no texto. Para tal tarefa são utilizadas *embeddings* GloVe de 300 dimensões pré-treinadas para o português [Hartmann et al. 2017]. Já as arestas que interligam os vértices *Tweet* e Toxicidade não possuem pesos em suas conexões.

4.3. Regulariação

A Regularização é responsável por extrair as características dos objetos do grafo. Este método é uma espécie da classificação transdutiva, ou semi-supervisionada, que visa encontrar um conjunto de rótulos atendendo duas condições: (i) ser consistente com o conjunto de rótulos manualmente anotados, (ii) apresentar consistência com a topologia da rede, isto é, considerar que os vizinhos mais próximos tendem a possuir as mesmas etiquetas [Rossi 2015].

O algoritmo de regularização utilizado foi *Gaussian Fields and Harmonic Functions* (GFHF) [Zhu et al. 2003]. O método GFHF é construído a partir de Campos Gaussianos, que podem ser um tipo de rede conectada com arestas ponderadas, onde a similaridade entre os objetos é determinada através de uma função Gaussiana [Rossi 2015]. A função harmônica é responsável por determinar o rótulo de uma instância a partir da média obtida pelos vizinhos ponderada pelas conexões entre eles. Esta função é aplicada somente às instâncias não rotuladas e, portanto, não alteram informações dos dados rotulados.

ID	Valor 1	Valor 2	Rótulo
198	-589802.2396	28944.082	1
1551	1610409.244	-9224.953	1
1986	-1287.533	219.633	0
2065	-927.333	214.390	0

Tabela 9. Exemplo de saída do algortimo regularizador GFHF.

A execução do GFHF gera coordenadas para cada um dos vértices presentes na rede, conforme mostra a tabela 9. O campo **ID** representa o identificador do vértice *Tweet*, os campos **Valor 1** e **Valor 2** indicam as coordenadas deste objeto e, por fim, o campo **Rótulo** indica se a instância é tóxica **1**, ou não tóxica **0**.

4.4. Classificação

As coordenadas obtidas pelo GFHF serão as entradas dos algoritmos de AMS para a predição de linguagem tóxica. Foram utilizados os algoritmos *Multi-Layer Perceptron* (MLP), *Support Vector Machine* (SVM) e *Gradient Boosting* (GB) da biblioteca Scikit-Learn [Pedregosa et al. 2011]. A seção seguinte detalhará os experimentos realizados e apresentará os resultados obtidos.

5. Experimentos e Resultados

Esta seção descreve a estratégia adotada para experimentação do grafo heterogêneo na detecção e anotação automática de linguagem tóxica. O *corpus* Toxic-BR curado foi utilizado para Regularização e treinamento do modelo. Já o ToLD-BR foi destinado para teste a avaliação da técnica proposta.

A etapa de Regularização utiliza dados pré-rotulados para a classificação transdutiva. Diante disso, definiu-se variações de 5 em 5%, até o limite de 30%, da quantidade de dados pré-rotulados oriundos do Toxic-BR curado fornecidos para a Regularização. Como os dados pré-rotulados são determinados de forma aleatória, logo, os grafos podem apresentar características distintas uns dos outros. Por este motivo, optou-se por repetir a execução do método GFHF e a Classificação em 10 vezes. A partir disso, avaliou-se cada um dos algoritmos de classificação através das médias da medida-F (*f-measure*) e do desvio padrão (*std*) obtidas pelas 10 execuções. A Tabela 10 sumariza os resultados deste experimento.

Tabela 10. Resultados alcançados pelos algoritmos de classificação.

Pré rotulados	MLP		SVM		GB	
	f-measure	std	f-measure	std	f-measure	std
5%	0,583	0,204	0,489	0,038	0,689	0,01
10%	**0,68**	**0,083**	0,502	0,041	0,673	0,031
15%	0,651	0,086	0,490	0,048	0,678	0,029
20%	0,676	0,079	**0,503**	**0,036**	0,690	0,028
25%	0,574	0,209	0,502	0,044	**0,693**	**0,026**
30%	0,658	0,092	0,488	0,046	0,669	0,055

Tabela 11. Resultados dos Kappas de Cohen obtidos pelos melhores modelos do experimento.

Método	Kappa	std
MLP_10%	0,412	0,083
SVM_20%	0,165	0,198
GB_25%	**0,431**	**0,023**

O algoritmo GB, em geral, obteve os melhores resultados, sendo no conjunto de 25% dos dados pré-rotulados o maior valor em *f-measure* (0,693). Já o MLP foi ligeiramente superior (0,68) ao GB (0,673) no experimento realizado no conjunto de 10% de dados pré-rotulados. No entanto, nas demais configurações o MLP superou somente o SVM, que teve o pior desempenho em todos os cenários.

A fim de averiguar a proposta do grafo heterogêneo como um método para anotação de *corpora* de linguagem tóxica, buscou-se avaliar a estatística Kappa de Cohen. Logo, foram selecionadas as configurações em destaque na Tabela 10, as quais apresentam os maiores valores *f-measure*. A Tabela 11 mostra os resultados deste experimento.

A configuração que apresentou melhor Kappa de Cohen foi o algoritmo GB com 25% dos dados pré-rotulados, alcançando 43,1%. O MLP com 20% dos dados pré-

rotulados apresentou 41,2% de Kappa. Apesar da diferença entre os valores Kappa, estes dois experimentos mostram nível de concordância moderado. Por fim, o SVM reportou o menor resultado (16,5%) e se enquadrou como nível de concordância fraca.

Os experimentos apresentados evidenciam o potencial dos grafos heterogêneos na detecção de linguagem tóxica, além de uma alternativa viável para anotação automática de *corpus*. Apesar da utilização de um pequeno *corpus* como o Toxic-BR curado, o método semi-supervisionado proposto consegue apresentar nível de concordância moderado entre o resultado obtido e os rótulos originais do *corpus* ToLD-BR.

6. Conclusões e Trabalhos Futuros

O presente trabalho apresentou um método semi-supervisionado, modelado na forma de grafo heterôgeneo, para a detecção e anotação de linguagem tóxica. Esta técnica utiliza uma pequena quantidade de dados e consegue realizar a anotação de quaisquer volume de textos tóxicos. Além disso, esta técnica ignora influências oriundas de predileções humanas, que podem afetar a credibilidade da anotação. Experimentos realizados sobre o *corpus* ToLD-BR demonstraram nível de concordância moderado (Kappa de Cohen de 43,1%) com os rótulos originais desta base de dados. Outra contribuição deste trabalho é o *corpus* Toxic-BR que foi anotado pelas ferramentas Perspective API [Lees et al. 2022] e ChatGPT [Brown et al. 2020], e, em seguida curado por avaliadores humanos. Este *corpus* curado foi utilizado para modelagem do grafo heterogêneo e extração das características necessárias para detecção de conteúdo tóxico.

Para trabalhos futuros serão estudadas novas topologias de grafo podendo observar informações semânticas que auxiliem na detecção de linguagem tóxica (e.g., sinônimos, entidades nomeadas, termos com grau de sentimento). Em seguida, avaliar a utilização de outros tipos de *embeddings* nos pesos das arestas. As *embeddings* de modelos de língua como o BeRT podem representar os pesos de maneira mais fiel ao contexto apresentado no texto [Souza et al. 2020].

Outra evolução deste trabalho é experimentar diferentes técnicas de Regularização e.g., *Local and Global Consistence* (LGC). Diferente da GFHF, a técnica LGC pode modificar as informações dos objetos rotulados durante a classificação transdutiva [Zhou et al. 2004]. Este algoritmo avalia os dados de forma coletiva visando suavizar a influência de objetos com alto grau de conexões, a fim de que eles não sejam totalmente determinantes na classificação.

Referências

Aroyo, L., Dixon, L., Thain, N., Redfield, O., and Rosen, R. (2019). Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot

learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, Online. Curran Associates, Inc.

Costa Bertaglia, T. F. and Volpe Nunes, M. d. G. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120, Osaka, Japan. The COLING 2016 Organizing Committee.

de Pelle, R. and Moreira, V. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *VI Brazilian Workshop on Social Network Analysis and Mining*, pages 510–519, São Paulo, Brazil. SBC.

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

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., and Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.

Hettiachchi, D., Holcombe-James, I., Livingstone, S., de Silva, A., Lease, M., Salim, F., and Sanderson, M. (2023). How crowd worker factors influence subjective annotations: A study of tagging misogynistic hate speech in tweets. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 11:38–50.

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., and Vasserman, L. (2022). A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA. Association for Computing Machinery.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Nascimento, F. R., Cavalcanti, G. D., and Da Costa-Abreu, M. (2022). Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning. *Expert Systems with Applications*, 201:117032.

Oliveira, A., Cecote, T., Silva, P., Castro Gertrudes, J., Freitas, V., and Luz, E. (2023). How good is chatgpt for detecting hate speech in portuguese? pages 94–103.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In Dipper, S., editor, *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, Bochumer Linguistische Arbeitsberichte, pages 6–9, Germany. Ruhr-Universitat Bochum.

Rossi, R. G. (2015). *Classificação automática de textos por meio de aprendizado de máquina baseado em redes*. PhD thesis, Instituto de Ciências Matemáticas e de Computação.

Russell, S. J. and Norvig, P. (2009). *Artificial Intelligence: a modern approach*. Pearson, 3 edition.

Saraiva, G. D., Anchiêta, R., Neto, F. A. R., and Moura, R. (2021). A semi-supervised approach to detect toxic comments. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1261–1267, Held Online. INCOMA Ltd.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Proceedings of the 9th Brazilian Conference on Intelligent Systems*, pages 403–417, Rio Grande. Springer International Publishing.

Vargas, F., Carvalho, I., Góes, F., Pardo, T., and Benevenuto, F. (2022a). Contextual-aware and expert data resources for brazilian portuguese hate speech detection.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022b). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Waseem, Z. (2016). Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In Bamman, D., Doğruöz, A. S., Eisenstein, J., Hovy, D., Jurgens, D., O'Connor, B., Oh, A., Tsur, O., and Volkova, S., editors, *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, MA, USA.

Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 912–919. AAAI Press.