Para Além da Medida de Toxicidade: Um Passo em Direção à Compreensão do Viés em Nível de Token

Lorenzo Puppi Vecchi[a,d], Emerson Cabrera Paraiso[a,c]

[a] *Programa de Pós-Graduação em Informática - Pontifícia Universidade Católica do Paraná*
Rua Imaculada Conceição, 1155 - Prado Velho, Curitiba, PR, Brasil
[b] *lorenzo.vecchi@ppgia.pucpr.br*
[c] *paraiso@ppgia.pucpr.br*
[d] *Autor para correspondência: lorenzo.vecchi@ppgia.pucpr.br*

Palavras-chaves: MNLI, Toxicidade, XAI

A problemática da toxicidade na comunicação online é uma preocupação crescente com potenciais consequências de alcance significativo. Compreender os vieses subjacentes e os mecanismos que contribuem para a toxicidade textual é fundamental para fomentar ambientes online inclusivos e respeitosos. O presente estudo investiga a importância do viés textual quando este assume uma orientação tóxica ou racista. Foi utilizado modelos "MNLI/*Entailment*", reconhecidos por suas capacidades de classificação *zero-shot*, para identificar de maneira precisa o texto tóxico em nível de token.

Além disso, adentrou-se nas limitações de conjuntos de dados anteriores comumente utilizados em análises de toxicidade, levantando preocupações em relação à confiabilidade de algumas classes previamente assinaladas por anotadores. Esta análise lança luz sobre a natureza questionável de conjuntos de dados específicos e enfatiza a necessidade de avaliação crítica e aprimoramento em futuros esforços de coleta de dados.

Nesta busca por explicabilidade e interpretabilidade (XAI) na análise de toxicidade, alavancarmos a capacidade dos modelos "MNLI/*Entailment*" para discernir as contribuições em nível de token para a toxicidade. Ao identificar tokens específicos que influenciam significativamente toxicidade, oferecemos *insights* sobre os marcadores linguísticos e os vieses subjacentes que promovem comportamentos tóxicos. Esta abordagem contribui para o campo de XAI e proporciona uma compreensão mais abrangente dos fatores que impulsionam o discurso tóxico.

Este estudo apresenta uma análise de toxicidade multifacetada, que incorpora o exame de viés textual e a análise de contribuição em nível de token. Ao combinar essas abordagens, almejamos contribuir para uma detecção e mitigação mais eficazes de conteúdo tóxico, promovendo um ambiente online mais seguro e inclusivo para todos os usuários.

Uma plataforma com a demonstração da metodologia pode ser consultada no seguinte endereço: https://article-2023.vercel.app/.