ISSN: 2358-8829

CHAT GPT E DISCURSOS DE ÓDIO NO TWITTER:
UM ESBOÇO ANALÍTICO SOBRE TIPOS DE CONHECIMENTO MOBILIZADOS PARA (NÃO) IDENTIFICAÇÃO DE PRECONCEITOS

Aline Cristine dos Santos [1]
Antony Luan Santana De Lima [2]
Jean Carlos dos Santos Oliveira [3]
Lauany Rodrigues de Oliveira [4]
Livia Beatriz Araújo Silva[5]
Rosa Maria da Silva Medeiros [6]

INTRODUÇÃO

O ChatGPT é um software de Inteligência Artificial (IA) que funciona como um modelo de linguagem natural para formular respostas a questionamentos de usuários sobre qualquer assunto. Esse software possui um sistema capaz de processar, memorizar e reproduzir informações, criando textos realistas e articulados a partir de volumosos bancos de dados. Considerando, por um lado, que geralmente internautas sentem uma certa "liberdade" nas redes ao ponto de utilizarem plataformas para disseminação de ódio; e, por outro, que o ChatGPT atingiu, em 05 dias, 100 milhões de usuários e que não tem compromisso com a veracidade das informações, faz-se necessária uma reflexão sobre como essa ferramenta (não) identifica discursos de ódio em textos sobre os quais é incitada a revisar, por exemplo. Devido a esse contexto problemático, nesse trabalho é analisado um tweet de teor gordofóbico de uma usuária com mais de 20 milhões de seguidores nas redes sociais, e como o chatGPT responde a duas perguntas acerca da adequação das informações para publicação na rede social Twitter.

Fundamentando-se em Marques (2023), Cassol (2023), Boa Sorte et al (2021), Costa et al (2018) e Koch (2003), esse trabalho analisa tipos de conhecimento mobilizados pelo ChatGPT na identificação de discurso de ódio em um tuíte, para reflexão sobre a devida menção a fontes referenciais e o combate ao preconceito por parte desse modelo de linguagem natural. A seguir, na metodologia, são apresentados os métodos para desenvolver este trabalho.

[1]Cursando Informática em Instituto Federal do Rio Grande do Norte - IFRN, alinecristineparelhas@gmail.com;
[2]Cursando Informática em Instituto Federal do Rio Grande do Norte - IFRN, lima.antony@escolar.ifrn.edu.br;
[3]Cursando Informática em Instituto Federal do Rio Grande do Norte - IFRN, o.jean@escolar.ifrn.edu.br;
[4]Cursando Informática em Instituto Federal do Rio Grande do Norte - IFRN, rlauany2@gmail.com;
[5]Cursando Informática em Instituto Federal do Rio Grande do Norte - IFRN, beatriz.livia@escolar.ifrn.br;
[6] Doutora pelo Curso de Linguagem e Ensino da Universidade Federal da Paraíba - UFPB rosa.maria@escolar.ifrn.com.br.

ISSN: 2358-8829

METODOLOGIA (OU MATERIAIS E MÉTODOS)

Esse trabalho resulta de uma pesquisa bibliográfica e documental de natureza básica. Metodologicamente, é desenvolvida uma análise de caráter descritivo e qualitativo, pois é nessa categoria que "se observam, registram, analisam, classificam e interpretam os fatos, sem que o pesquisador lhes faça qualquer interferência" (PAIVA, 2019, p 190). Para isso, a análise está organizada em duas etapas: na primeira, o foco analítico se detém ao tweet de teor preconceituoso de autoria de uma influenciadora digital brasileira de grande alcance virtual; na segunda, esse foco se desloca para as respostas escritas do Chat GPT sobre as informações veiculadas no referido twitter. Nessa etapa, ao Chat GPT, é apresentado textualmente o tweet e incitado a responder às perguntas: 1. "Esse texto (...) está adequado para publicar no twitter? " e 2. "A inflenciadora respeita os Direitos Humanos?".

A diferença entre as perguntas está no fato de mencionar (ou não) o respeito aos direitos humanos na formulação do questionamento. Isso foi feito para constatação se essa ferramenta faz referência aos direitos humanos independentemente de conter essa expressão na pergunta. Além da Declaração Universal dos Direitos Humanos, foi consultada a Constituição Federal para conferência se esses documentos oficiais são mencionados e, se são, para verificação da devida referenciação desses textos como argumento de autoridade e com créditos autorais.

REFERENCIAL TEÓRICO

2.1 O papel inovador da OpenAI e do Chat GPT para a produção escrita e ética na escola

A inteligência artificial (IA) tem sido uma das maiores conquistas do mundo tecnológico, transformando a interação entre os seres humanos e computadores. A OpenAI (Inteligência Artificial Aberta) é uma empresa de pesquisa em inteligência artificial, fundada em 2015, que tem como principal objetivo desenvolver tecnologias de ponta e tornar seus benefícios mais amplamente acessíveis para a humanidade.

Nesse contexto, destaca-se o modelo GPT, produzido pela OpenAl, capaz de responder perguntas feitas pelos usuários, compreender padrões linguísticos e contextuais, gerando respostas relevantes e de maneira rápida em forma de texto. A ferramenta pode ser um dos aliados da educação, especialmente, no que se refere à produção textual escrita. Segundo a Nova Escola (2023), o chat pode ser usado nas escolas como uma ferramenta de pesquisa e tem "o potencial de fazer os professores repensarem suas práticas". Isso porque o chatgpt pode auxiliar na elaboração de aulas, projetos, pesquisas, etc. No entanto, esse auxílio pode gerar também malefícios para o desenvolvimento escolar, como por exemplo em relação à produção escrita e ética na escola.

ISSN: 2358-8829

Segundo Rocha (2023), Pesquisadores da Universidade do Estado da Pensilvânia, nos Estados Unidos, investigaram até que ponto modelos de linguagem natural como o ChatGPT conseguem gerar conteúdo que não se caracteriza como plágio. Considerando 210 mil textos gerados pelo programa GPT-2, da startup OpenAI, criadora do Chat GPT, essa pesquisa constatou a presença dos três tipos de plágio: a transcrição literal, obtida copiando e colando trechos; a paráfrase, que troca palavras por sinônimos a fim de obter resultados ligeiramente diferentes; e o uso de uma ideia elaborada por outra pessoa sem mencionar sua autoria, mesmo que formulada de maneira diferente. Esse contexto salienta o quanto é indispensável a orientação dos educadores sobre o uso produtivo e ético das ferramentas disponibilizadas pela Inteligência Artificial.

2.2 O discurso de ódio no Twitter e formas de combater o preconceito

O twitter é uma rede social de serviço de microblog em que os usuários podem interagir na em tempo real com milhões de pessoas/organizações. Elon Musk, diretor executivo da Tesla e da SpaceX, comprou o Twitter em 2022, e substituiu o logotipo por X. Segundo a CNN Brasil (2022), Musk garantiu que as "impressões de discurso de ódio" caíram drasticamente na plataforma, desde que assumiu o cargo. No entanto, a CNN, consultando pesquisas de grupos de vigilância, constatou que o uso diário de jargões racistas sob Musk é o triplo da média de 2022 e o uso de calúnias contra gays e pessoas trans aumentou 58% e 62% respectivamente". Logo, a disseminação de discurso de ódio em redes sociais como o twitter ainda é uma realidade perversa e preocupante.

Segundo Becker (2018, p.37), "O discurso de ódio é o conflito em si. Expor mensagens odiosas nas redes sociais contra as minorias: LGBTs (Lésbicas, Gays, Bissexuais, Travestis, Transexuais e Transgêneros), mulheres, negros, índios entre outras socialmente reprimidas". Vale salientar que a CNN Brasil (2021) "traçou um perfil dos odiados e percebeu que eles têm cor e gênero bem definidos. Cerca de 59,7% das vítimas de discursos de ódio são pessoas negras, e 67% são mulheres. Estatísticas como essa mostram como o tweet preconceituoso analisado neste trabalho é comum, pois tem como alvo o corpo de uma mulher sendo exposto pejorativamente. Um desrespeito reiterado aos direitos humanos e ao cumprimento do Art.3 da Constituição Federal, inciso IV, que versa sobre um dos objetivos fundamentais do país, ou seja, "IV - promover o bem de todos, sem preconceitos de origem, raça, sexo, cor, idade e quaisquer outras formas de discriminação." Considerando que o discurso de ódio é materializado nos textos, a seguir são abordados os tipos de conhecimento necessários para a produção textual escrita.

2.3 Os sistemas de conhecimento Linguístico, Enciclopédico e Interacional

ISSN: 2358-8829

O conhecimento linguístico, segundo Koch (2003), compreende o conhecimento gramatical(das regras) e o lexical(das palavras) e que também é responsável pela organização das palavras em frases, a gramática é o aspecto do conhecimento que lida com a organização das palavras em frases, um desses aspectos são a formação de tempos verbais.

O conhecimento enciclopédico, por sua vez, é aquele armazenado na nossa memória. E esse tipo de conhecimento muitas vezes faz referências a fontes confiáveis e autorizadas, como em livros, documentos acadêmicos. Koch (op. cit) cita o conhecimento sociointeracional, que, segundo ela, é fundamental para navegar nas complexas relações humanas e nas interações sociais em uma sociedade diversificada, referência formas e contextos de interação entre interlocutores, de acordo com cada necessidade comunicativa. Vale destacar que esses três sistemas de conhecimento estão interconectados e trabalham em conjunto para permitir uma comunicação e compreensão eficazes.

RESULTADOS E DISCUSSÃO

Em 2011, uma influente usuária do Twitter postou: “n quero nem ser maldosa mas daí vem a bonitinha gorda é quadrada usando um vestido colado e de listras ENFRAQUECENDO MEUS VOTOS DE BONDADE”. A conta dessa comediante foi desativada do Twitter. Entretanto, uma vez postada e apagada, ainda persiste na rede, porque pode ser salva e reutilizada no meio virtual. No referido tweet de teor gordofóbico, constata-se o conhecimento de mundo da autora sobre o corpo feminino diferente de um padrão social. O alvo da postagem é “bonitinha gorda e quadrada”. O sujeito do enunciado é alguém do gênero feminino objetificada a partir de descrições físicas do corpo humano acima de peso (gorda) e (quadrada) que não se adequa a um padrão de corpo com curvas consideradas ideais para a autora do tweet em análise. Embora a palavra “bonita” sinalize um atributo positivo relacionado à beleza, nesse contexto, grafada no diminutivo (bonitinha), indica o tom irônico da postagem. O alvo não tem nome próprio (a bonitinha). É identificado de modo depreciativo (gorda e quadrada) segundo padrões corporais femininos impostos pela sociedade: não apresenta corpo “adequado” para usar “um vestido colado e de listras”. Esse fato motiva, segundo o conhecimento de mundo da usuária do twitter, a exposição pejorativa. A vítima é, ela própria, a culpada pela ocorrência desse preconceito. Seu modo de existir e de se vestir são responsáveis por ir “ENFRAQUECENDO MEUS VOTOS DE BONDADE”. Permitir a existência de tipos corpóreos como aquele sem externalizar preconceito é entendido como bondade. Essa bondade se enfraquece quando o outro não corresponde aos padrões de existência no mundo, aceitáveis ou admirados pelo meio social.

ISSN: 2358-8829

Sobre esse tweet foram feitas duas perguntas ao ChatGPT. A primeira, mais genérica, restringia-se a questionar sobre a adequação do texto à publicação no twitter. A resposta (1) é composta por dois parágrafos, nos quais é mobilizado o conhecimento linguístico no que se refere a escolhas vocabulares e lexicais mais acessíveis ao interlocutor por não utilizar uma linguagem de difícil entendimento. O conhecimento interacional é marcado pela utilização do pronome de tratamento "você", promovendo um diálogo de proximidade com o leitor. Já, em relação ao conhecimento de mundo, verifica-se a presença de alertas sobre o uso de palavras que podem ser interpretadas como "insensíveis/ofensivas"; e sobre a possibilidade de publicações como aquela em rede pública gerar "reações negativas e prejudicar a reputação online". Além desses alertas, há o aconselhamento do chatgpt, sugerindo " reformular o texto para evitar qualquer conotação ofensiva". Constata-se um cuidado em preservar a imagem virtual, a reputação online.

A segunda pergunta, mais específica, almejava verificar se a influencer, por meio daquele tweet, respeitava os direitos humanos. A resposta (2) do ChatGPT (2023) é composta por três parágrafos, nos quais é mobilizado o conhecimento linguístico e interacional de modo similar à resposta (1). Em relação ao conhecimento de mundo, percebe-se também a presença de alertas, mas, dessa vez, focada em sinalizar "comentários depreciativos sobre a aparência física", especificando o termo "gorda". Além disso, inclui o fato de "fazer suposições sobre o que essa pessoa está vestindo como "inadequado" e sugere poder ser interpretado como "desrespeitoso em relação aos direitos humanos, especialmente no que diz respeito à dignidade e ao respeito de todas as pessoas, independente de sua aparência". Vale destacar que, assim como na resposta (1), nessa ocasião houve o registro de alertas, dessa vez, sinalizando cautela em relação à análise da postura desrespeitosa da autora do tweet. Constata-se uma preocupação, no 2º e 3º parágrafos, sobre "postagem isolada" e "posicionamento de alguém em relação aos direitos humanos", sinalizando a necessidade de cautela, como uma advertência para evitar julgamentos precipitados sobre a autora do tweet, principalmente quando envolve respeito aos direitos humano. Isso porque no Art.1°, da Declaração Universal dos Direitos Humanos, consta que "Todos os seres humanos nascem livres e iguais em dignidade e em direitos. Dotados de razão e de consciência, devem agir uns para com os outros em espírito de fraternidade". Ou seja, o tweet fere o direito de ser reconhecida como pessoa digna de respeito por meio de uma ação verbal não condizente com um espírito de fraternidade.

CONSIDERAÇÕES FINAIS

Constatou-se uma superficialidade do conhecimento enciclopédico mobilizado em respostas do ChatGPT para identificação de discursos de ódio nas mídias sociais. Documentos

ISSN: 2358-8829

significativos para abordagem do tema como a Declaração dos direitos humanos e a Constituição federal não são devidamente referenciados naquelas respostas, embora haja o registro das ideias deles presentes na elaboração textual do ChatGPT. Esse modelo de linguagem natural pode ser uma excelente ferramenta para pesquisa e para estimular a curiosidade intelectual dos alunos. No entanto, um dos principais problemas a ser combatido é a dependência da Inteligência Artificial, que pode ocasionar o plágio e comprometer o aprendizado satisfatório do aluno. Os professores desempenham papel fundamental, alertando o alunado sobre a utilização responsável do Chat, por meio de uma avaliação crítica das informações obtidas e da busca de outras fontes referenciais para checagem e socialização daquelas informações de modo seguro, confiável e ético.

REFERÊNCIAS

ANDRADE. *ChatGPT inaugura nova era na interação entre seres humanos e computadores.* Pesquisa Fapesp. 2023). **Disponível em:** https://revistapesquisa.fapesp.br/o-universo-expandido-da-inteligencia-artificial/ **Acesso em:** ago/2023

GAROFALO, Débora. Conheça o ChatGPT e suas possibilidades de uso na Educação. *Nova Escola*. 2023). **Disponível em:** https://novaescola.org.br/conteudo/21620/conheca-o-chatgpt-e-suas-possibilidades-de-uso-na-educacao **Acesso em:** jul/2023

MARQUES, Fabrício. *O plágio encoberto em textos do Chat.* 2023. Pesquisa Fapesp. **Disponível em:** https://revistapesquisa.fapesp.br/o-plagio-encoberto-em-textos-do-chatgpt/ **Acesso em:** jul/2023.

Becker, José A. O Combate ao Discurso de Ódio nas Redes Sociais.2018. Pesquisa UFSC. Disponível em:https**://repositorio.ufsc.br/bitstream/handle/123456789/187510/O_Combate_ao_Discurso_de_Odio_nas_Redes_Sociais.pdf** **Acesso em:** ago/2023

PAIVA, V. L. M. O. Manual de pesquisa em estudos linguísticos - 1. ed.- São Paulo: Parábola, 2019.**Disponível em:** https://publicacoes.unifal-mg.edu.br/revistas/index.php/resenhando/article/download/1637/1440/ Acesso em:jul/2023

VALIENSE, Karina V.C. O que é a Declaração Universal dos Direitos Humanos?. Politize! 2023) **Disponível em: https://www.politize.com.br/o-que-e-a-declaracao-universal-dos-direitos-humanos/**

KOCH, Ingedore G. V.. Atividades e Estratégias de Processamento TEXTUAL. In:___ *O texto e a construção dos sentidos.* 7. ed São Paulo : Contexto, 2003.

REDAÇÃO AVENTURAS NA HISTÓRIA (2022)**.** *Tweets antigos de Gkay são resgatados e influencer desativa a conta.* **Disponível em:** https://aventurasnahistoria.uol.com.br/noticias/historia-hoje/tweets-antigos-de-gkay-sao-resgatados-e-influencer-desativa-conta.phtml **Acesso em:** maio/2023

BRASIL. *Constituição da República Federativa do Brasil de 1988.* **Disponível em:** https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm#:~:text=I%20%2D%20construir%20uma%20sociedade%20livre,quaisquer%20outras%20formas%20de%20discrimina%C3%A7%C3%A3o. **Acesso em:** maio/2023

ROCHA, Leandro. *O plágio encoberto em textos do Chat GPT.* **Disponível em:** https://www.abecbrasil.org.br/novo/2023/03/o-plagio-encoberto-em-textos-do-chatgpt/ **Acesso em:** jun/2023

Declaração Universal dos Direitos Humanos. **Disponível em:** https://www.oas.org/dil/port/1948%20Declara%C3%A7%C3%A3o%20Universal%20dos%20Direitos%20Humanos.pdf **Acesso em**: jun/2023