AVALIANDO MODELOS DE DETECÇÃO DE DISCURSO DE ÓDIO EM LÍNGUA PORTUGUESA

Rômulo Barreto Mincache (PIC/UEM), Valéria Delisandra Feltrim (Orientadora). E-mail: ra117477@uem.br.

Universidade Estadual de Maringá, Centro de Tecnologia, Maringá, PR.

Ciências Exatas e da Terra / Ciência da Computação

Palavras-chave: processamento de linguagem natural; testes funcionais; inteligência artificial.

RESUMO

Mesmo com o avanço recente em tecnologias de processamento de linguagem natural, modelos de detecção de discurso de ódio na internet ainda são propícios a erros, assim como os métodos usados para avaliá-los. As técnicas de avaliação mais utilizadas, como acurácia e F1, não são capazes de identificar com precisão os pontos fracos e fortes de um modelo e podem, ainda, causar interpretações e generalizações erradas sobre as habilidades do modelo. Por conta disso e da falta de informações sobre o assunto na língua portuguesa, propomos aqui um conjunto de testes funcionais para avaliar modelos de detecção de discurso de ódio na língua portuguesa. A partir de trabalhos semelhantes da literatura, foram elaboradas 29 categorias, cada uma avaliando uma funcionalidade específica de um modelo, bem como casos de teste para cada categoria. O conjunto proposto foi utilizado para avaliar modelos de detecção de discurso de ódio baseados em máquina de vetores de suporte (SVM) e BERT, treinados com bases de dados em português.

INTRODUÇÃO

Nos últimos anos, o crescimento das redes sociais trouxe consigo a preocupação com o aumento do discurso de ódio na internet. O discurso de ódio, caracterizado pelo uso de linguagem ofensiva, discriminatória e prejudicial para atacar, incitar violência ou ódio contra indivíduos ou grupos específicos (FORTUNA; NUNES, 2018), não apenas afeta ambientes virtuais, mas também tem repercussões significativas na sociedade como um todo. Para mitigar esse problema, a detecção automatizada de discurso de ódio tem sido uma área de pesquisa em rápida expansão. Embora avanços tenham sido alcançados, os modelos de detecção de discurso de ódio (MDDO) ainda não atingiram a perfeição desejada.

A natureza do discurso de ódio pode variar de acordo com contextos culturais, linguísticos e sociais. Isso exige que os modelos sejam flexíveis o suficiente para se adaptarem às evoluções da linguagem. Devido à complexidade intrínseca da linguagem humana, o discurso de ódio muitas vezes se manifesta de maneira sutil e

ambígua, utilizando ironias, sarcasmo e metáforas que podem escapar à detecção dos modelos (FORTUNA; NUNES, 2018).
Além disso, as métricas tradicionalmente utilizadas na avaliação de MDDO, como precisão, *recall* e F1, podem não capturar completamente a complexidade do problema. Um modelo pode ter uma alta acurácia mas não ser capaz de generalizar os casos, se baseando em regras de decisão simples ou palavras chave. Portanto, a necessidade de métricas mais sensíveis ao contexto e ao impacto real do discurso de ódio é evidente (RÖTTGER et al., 2021).
Para avaliar MDDO em língua portuguesa, este trabalho propõe uma adaptação do conjunto de testes funcionais de Röttger et al. (2021) para o português do Brasil.

MATERIAIS E MÉTODOS

O conjunto de testes funcionais elaborado teve como base aquele proposto por Röttger et al. (2021), chamado *Hatecheck*, com adaptações para a tradução das frases para o português e a contextualização das categorias e dos grupos alvo do discurso de ódio para a realidade brasileira. Testes funcionais concentram-se nas entradas e saídas esperadas e nas funcionalidades do sistema, buscando verificar se o *software* realiza suas funções conforme as expectativas e se cumpre os objetivos para os quais foi desenvolvido.

Adaptação dos Testes Funcionais

Röttger et al. (2022) propõem uma tradução direta do *Hatecheck* para 10 línguas, incluindo o português. Essa tradução removeu as categorias que testavam funcionalidades de detecção de palavras homônimas de insultos e insultos ressignificados, uma vez que ambos não têm tradução direta para outras línguas. Neste trabalho, manteve-se todas as categorias originais, adaptando as frases dessas categorias para o português e, para casos em que não foi possível adaptar, criando novas frases de forma a testar as funcionalidades propostas. A adaptação e tradução das frases foi feita manualmente pelo autor, a partir da leitura e análise dos *templates* geradores das frases disponibilizados por Röttger et al. (2021). Ambas as frases adaptadas e criadas foram escritas com base nas frases originais, visando manter seu sentido e estrutura original.
Outra diferença entre Röttger et al. (2022) é que neste trabalho foram abrangidos grupos protegidos específicos do contexto brasileiro, como indígenas, religiões de matrizes africanas e preconceitos regionais brasileiros, como nordestinos e sulistas.

Conjunto de Testes

Assim como no *Hatecheck*, foram elaboradas 29 categorias de testes funcionais que compreendem 9.615 casos de teste. Cada categoria testa uma funcionalidade específica do modelo em questão e aborda uma forma de expressão de ódio diferente. Os casos de teste são rotulados como ‘ódio’ ou ‘não ódio’. 18 testes funcionais (7.204 casos de teste) abordam diferentes formas de expressão de ódio,

enquanto os 11 restantes (2.411 casos de teste) contêm expressões contrastantes sem discurso de ódio, que possuem características com nuances que podem ser complicados de discernir e acabam impactando na saída de modelos simples que se baseiam em palavras-chave ou regras de decisão simples.
Quanto às diferentes formas de expressão de ódio, o conjunto de testes inclui tipos de discurso depreciativo, ameaças, insultos e palavrões. Estruturas diversas também são levadas em consideração ao testar ódio expresso por meio de perguntas ou opiniões, com referências utilizando pronomes e negação. Por fim, variações ortográficas comuns na internet, como troca ou falta de caracteres e falta de espaço entre palavras compõem testes funcionais diferentes.
Em adição às formas de expressão de ódio, os testes incluem expressões sem discurso de ódio, que avaliam a capacidade de detectar palavras homônimas de insultos, insultos ressignificados, usos de palavrão em frases sem ódio e negação de ódio, bem como referências sem ódio a grupos e frases com ódio, mas não dirigido a grupos protegidos. Também são abrangidas nos testes frases que denunciam ou respondem a um comentário odioso, citando-o, uma vez que podem ser confundidas com discurso de ódio quando seu propósito é justamente o oposto disso.
Para demonstrar a usabilidade do conjunto de testes, ele foi aplicado em modelos treinados com as bases de dados ToLD-BR (LEITE et al., 2020) e HateBR (VARGAS et al., 2022), ambas voltadas para o português do Brasil. A ToLD-BR é composta por 21.000 tweets classificados como tóxico/não tóxico e a HateBR é formada por 7.000 comentários do *Instagram*. Cada base foi dividida de forma aleatória em conjuntos de treino (80%) e teste (20%) e, para cada uma, foram treinados e avaliados modelos BERT e SVM.

RESULTADOS E DISCUSSÃO

Os valores de acurácia obtidos para os quatro modelos avaliados são mostrados na Tabela 1.

Tabela 1 Resultados dos modelos avaliados.

Modelos	Acurácia/F1 20%	Acurácia/F1 testes funcionais
HateBR-SVM	85%/85%	53%/61%
HateBR-BERT	92%/92%	61%/69%
ToLD-BR-SVM	75%/75%	38%/34%
ToLD-BR-BERT	78%/78%	51%/56%

Fonte: Autoria própria.

Como vê-se na tabela, os classificadores apresentaram resultados superiores quando avaliados da forma tradicional, utilizando exemplos retirados das próprias bases de dados. Porém, a avaliação com o conjunto de testes funcionais proposto permitiu observar falhas que não ficam evidentes na avaliação tradicional.
Analisando especificamente os resultados dos testes funcionais, todos os classificadores se mostraram sensíveis ao uso de palavrões, avaliando grande parte

das frases com este tipo de expressão como discurso de ódio. Todos são também sensíveis a negações, classificando os casos de teste com negação como não-odiosos. Com exceção de HateBR-BERT, não são enviesados pelos grupos alvo, uma vez que a acurácia média para os grupos alvo é 47,50%. A média em HateBR-BERT é de 59,26%.

CONCLUSÕES

Este projeto adaptou o conjunto de testes funcionais de Röttger et al. (2021) para avaliar modelos de detecção de discurso de ódio para a língua portuguesa e com enfoque no contexto brasileiro. A avaliação dos modelos por meio dos testes funcionais permite compreender melhor as capacidades e fraquezas de cada modelo do que uma avaliação baseada somente em métricas tradicionais, como acurácia e F1. Para demonstrar tal afirmação, o conjunto foi usado para avaliar modelos treinados com bases de dados em português da literatura. A avaliação com o conjunto de testes revelou que os modelos avaliados eram sensíveis à negação e ao uso de palavrões, bem como a imparcialidade quanto aos grupos alvo.

REFERÊNCIAS

FORTUNA, P.; NUNES, S. A Survey on Automatic Detection of Hate Speech in Text. **ACM Computing Surveys**, v. 51, n. 4, p. 1–30, 6 set. 2018.

LEITE, J. A. et al. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In: 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. **Proceedings**...Suzhou, China: Association for Computational Linguistics, 1 dez. 2020.

RÖTTGER, P. et al. HateCheck: Functional Tests for Hate Speech Detection Models. In: 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. **Proceedings**...Stroudsburg, USA: Association for Computational Linguistics, 2021.

RÖTTGER, P. et al. Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models. In: Sixth Workshop on Online Abuse and Harms. **Proceedings**...Seattle, Washington: Association for Computational Linguistics, 1 jan. 2022.

VARGAS, F. et al. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In: Thirteenth Language Resources and Evaluation Conference. **Proceedings**...Marseille, France: European Language Resources Association, 1 jun. 2022.