DETECÇÃO DE MEMES COM CONTEÚDO PRECONCEITUOSO

Renan Chaves Bezerra [1], **Herman Martins Gomes** [2]

RESUMO

A disseminação de conteúdos preconceituosos tem tido um impacto significativo nas relações sociais, principalmente aquelas mediadas por aplicativos de mensagens instantâneas e redes sociais. Com o aumento constante desse tipo de conteúdo nos meios de comunicação, surge a necessidade de otimizar os processos de identificação e remoção desses conteúdos ofensivos das redes. Importante ressaltar que um meme é um tipo de conteúdo cujo significado é construído pela combinação de informações textuais e imagens. Essa característica representa um relevante desafio de pesquisa, pois é necessário combinar informações do texto e da imagem para que a semântica do meme possa ser inferida em um processo de classificação automática.

Este projeto visa investigar e desenvolver uma abordagem que permita treinar modelos de aprendizagem de máquina para reconhecer padrões, utilizando textos e descrições de imagens. A revisão da literatura revelou a ausência de pesquisas versando sobre a detecção de memes preconceituosos em português no Brasil. A metodologia adotada inclui as seguintes etapas: análise de abordagens e bibliotecas de software para a análise de conteúdo semântico de texto e imagem, treinamento de modelos de aprendizagem de máquina para classificar conteúdos preconceituosos em memes e avaliação da abordagem.

A partir deste enfoque de pesquisa, foi possível desenvolver modelos treinados para a previsão de memes em língua portuguesa, alcançando resultados comparáveis aos líderes no estado da arte, que se concentram em memes em inglês.

Palavras-chave: redes neurais profundas, memes odiosos, detecção automática.

[1] Aluno do Ciência da Computação, Unidade Acadêmica de Sistemas e Computação, UFCG, Campina Grande, PB, e-mail: renan.bezerra@ccc.ufcg.edu.br

[2] Doutor, Orientador, Unidade Acadêmica de Sistemas e Computação, UFCG, Campina Grande, PB, e-mail: hmg@computacao.ufcg.edu.br

DETECÇÃO DE MEMES COM CONTEÚDO PRECONCEITUOSO

ABSTRACT

The dissemination of prejudiced content has had a significant impact on social relationships, especially those mediated by instant messaging apps and social networks. With the constant increase in this type of content in the media, there is a need to optimize the processes of identification and removal of these offensive contents from the networks. It is important to emphasize that a meme is a type of content whose meaning is constructed through the combination of textual information and images. This characteristic represents a relevant research challenge because it is necessary to combine information from both text and image to infer the meme's semantics in an automatic classification process.

This project aims to investigate and develop an approach that allows for the training of machine learning models to recognize patterns using text and image descriptions. The literature review revealed the absence of research on the detection of prejudiced memes in Portuguese in Brazil. The adopted methodology includes the following steps: analysis of approaches and software libraries for the analysis of semantic content in text and images, training of machine learning models to classify prejudiced content in memes, and evaluation of the approach.

Through this research focus, it was possible to develop trained models for predicting memes in the Portuguese language, achieving results comparable to the leaders in the state of the art, who focus on memes in English.

Keywords: deep neural networks, hateful memes, automatic detection.