Abstract

This dissertation addresses the problem of personalized detection of offensive text using techniques from natural language processing (NLP) and deep learning. The key challenges in detecting offensive content arise from the diversity of user perceptions and the semantic complexity of language. The study introduces novel personalized approaches that incorporate user preferences to improve the efficiency of offensive content classification, distinguishing them from traditional generalized methods, which do not adapt to individual annotator preferences. The experimental evaluation involved comparative testing across various text vectorization techniques, such as CBOW, Skipgram, BERT, DeBERTa, MPNet, and XLM-RoBERTa. The results demonstrate that advanced contextual models, such as DeBERTa, outperform traditional non-contextual approaches in identifying offensive content. Additionally, the impact of training dataset size on model stability and prediction accuracy was analyzed, highlighting the benefits of models that incorporate personalization. The dissertation also introduces two novel personalized active learning methods, based on a custom measure of text controversy and annotation dispersion. These methods optimize the annotation process compared to a reference method and other personalized active learning techniques such as Ratio Distance and Stranger Count. Their application significantly enhanced model efficiency when working with smaller datasets, making these approaches especially useful in situations with limited annotated data. Furthermore, the study explores the use of the generative model ChatGPT-3.5 for both generalized and personalized offensive content detection. The findings indicate that dedicated models trained on specific datasets outperform the generative model in this task, particularly in terms of personalization. The dissertation presents key insights into the role of personalization in offensive content detection and proposes new techniques with potential applications in online content moderation systems.