Towards a new Contextualized Annotation Schema for Unacceptable and Extreme Speech (CUES) to Unleash Generalization Capability of ML models

Dimitra Niaouri, Michele Linardi and Julien Longhi

Abstract: In an era marked by global crises and social challenges, including inequality, unrest, and the proliferation of extreme online content, the need for effective Machine Learning (ML) solutions to detect Socially Unacceptable Discourse (SUD) is paramount. However, existing ML models face significant challenges in accurately classifying such content due to issues such as biased annotations, limited contextual understanding, and the neglect of multimodal elements. Additionally, binary classes in annotated datasets limit SUD representation, affecting real-world discriminative capacity, while multiclass frameworks expose generalization gaps and label semantics inconsistencies, hindering multi-source learning. This paper presents a novel approach aimed at enhancing the capabilities of state-of-the-art (SOTA) ML models by providing a set of guidelines that will allow to semantically enrich existing ad-hoc annotation schemas and better leverage state-of-the-art machine learning classifiers. Our methodology focuses on refining labels and improving model generalization by incorporating diverse contextual factors underlying the spread of unacceptable speech. We address the limitations of existing annotated datasets, including class imbalances and overlapping classes, and propose a systematic evaluation of our annotation schema across various ML models. By investigating user information and multimodal elements from online platforms, we aim to better understand the socio-cultural environment in which SUD arises. Through our approach, we highlight the significance of context in enhancing the effectiveness of ML algorithms for detecting extreme online content.

Keywords: Socially Unacceptable Discourse Analysis, Hate Speech Analysis, Machine Learning, Annotation Schema, Unacceptable Discourse Context Modeling