NLP for CounterSpeech - Trends and Open Challenges

Marco Guerini[1]

[1]*Fondazione Bruno Kessler (FBK), Trento, Italy*

Abstract of Invited Talk

This talk addresses the use of argumentation-based textual responses—referred to as counterspeech—to combat online hate speech, an emerging topic in Natural Language Processing. Specifically, it explores the automated generation of counterspeech as an effective strategy for intervening in harmful online discourse. Following an overview of the field, various methodologies for collecting high-quality data and effective generation strategies, including state-of-the-art LLMs, are discussed. The talk also examines critical challenges at the intersection of online safety, argumentation strategies and the effectiveness of NLP interventions.

8th Workshop on Advances in Argumentation in Artificial Intelligence (AI^3*), November 28, 2024, Bozen, Italy*

guerini@fbk.eu (M. Guerini)