Zusammenfassung | Call 17 | Stipendium ID 6300

Abstract der fertigen Arbeit

Digital transformation has introduced a multitude of challenges, including online threats, derogatory remarks, provocations, and hate speech. In response, the use of counter speech by web users has gained importance as a strategy to mitigate these issues. Consequently, monitoring online interactions and activities on social media platforms is crucial for ensuring a safe digital experience, underscoring its relevance in today's world. Over time, various machine learning frameworks have been utilized for text analysis and detection tasks. Notably, newer Transformer architectures like BERT have excelled in identifying malicious content on social platforms. These models are adept at processing large data sets, demonstrating impressive performance metrics during validation processes. However, these advanced models often come with a challenge: reduced clarity and interpretability. The push for model explainability has been a notable trend in the Artificial Intelligence sector recently. However, the intersection of counter speech detection with explainability techniques remains underexplored. After extensive research into relevant datasets and training BERT Transformer models for hate and counter speech detection, several explainability techniques were examined to clarify the models' decision-making processes. The focus was on three distinct explainability methods: the established LIME [9] and Integrated Gradients, alongside newer methods like GlobEnc and ProtoTEx .

For this study, three BERT Transformer models were adapted for tasks related to counter speech. Additionally, a specific BERT model for language detection was developed to assess the ProtoTEx technique. However, initial tests indicated that the ProtoTEx method wasn't a good fit for the study's goals and was therefore excluded from further analyses. The other methods were evaluated on various criteria such as fidelity, logical coherence, clarity, comprehensiveness, reliability, user satisfaction, and practicality. While all three methods showcased plausible results, they didn't fully meet all the evaluation criteria. Notably, the LIME method showed promise in the clarity and comprehensiveness areas.