# Création d'un ensemble de données de référence pour la détection robuste de discours haineux en allemand

## Résumé

Auteurs:
Gerold Schneider, Janis Goldzycher Université de Zurich

16 avril 2024

La qualité des modèles de détection du discours haineux dépend de la qualité des données avec lesquelles ils sont entraînés. Les jeux de données provenant des médias sociaux souffrent de lacunes et de biais systématiques, ce qui engendre des modèles non fiables dotés de limites décisionnelles simplistes. Les jeux de données contradictoires, recueillis en exploitant les faiblesses d'un modèle, promettent de résoudre ce problème. Cependant, la collecte de données contradictoires peut être lente et coûteuse, et les annotateurs individuels ont une créativité limitée. Dans ce projet, nous introduisons un nouveau jeu de données, le *German Adversarial Hate speech Dataset* (GAHD), qui contient 11'000 exemples. Durant la collecte de données, nous explorons de nouvelles stratégies pour soutenir les annotateurs, dans le but de créer plus efficacement des exemples contradictoires plus diversifiés et de fournir une analyse manuelle des désaccords exprimés par les annotateurs à l'égard de chaque stratégie. Nos expériences montrent que le jeu de données qui en résulte pose des difficultés même aux modèles les plus avancés de détection du discours haineux, et que l'entraînement au moyen du GHAD améliore significativement la robustesse du modèle. De plus, nous constatons qu'associer plusieurs stratégies de soutien se révèle plus avantageux. Le GAHD est disponible sous https://github.com/jagol/gahd.