Title

Incorporating Mental State into Contrastive Learning for Fine-grained Implicit Hate Speech Classification

Permalink

https://escholarship.org/uc/item/9b69g3th

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors

Wang, Haiyang
Song, Xin
Zhou, Bin
et al.

Publication Date

2023

Peer reviewed

Incorporating Mental State into Contrastive Learning for Fine-grained Implicit Hate Speech Classification

Haiyang Wang
National University of Defense Technology, ChangSha, China

Xin Song
National University of Defense Technology, ChangSha, China

Bin Zhou
National University of Defense Technology, Changsha, China

Xuechen Zhao
National University of Defense and Technology, Changsha, Hunan, China

Abstract

Many people have suffered harm as a result of hate speech on social media. The majority of research has focused on coarse-grained explicit hate speech detection while disregarding fine-grained implicit hate speech classification. It is crucial for more effectively combating hate speech. Although the language used in implicit hate speech may vary greatly, the mental states involved are usually the same. There are rarely similarities and differences between the mental states present in implicit hate speech examined. We create a module to infer mental states from implicit hate speech to close this gap. Mental states primarily refer to the speaker's intent and the reader's reaction. Then, we use them as the positive sample in contrastive learning. This strategy can pull the implicit hate speech which has similar mental states in similar representations and push away different ones. Comprehensive experiment results demonstrate superior classification performance and generalization of the proposed method.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*