# Research may help combat abusive online comments

May 10 2017, by David Mitchell

Researchers at the Georgia Institute of Technology's School of Interactive Computing have come up with a novel computational approach that could provide a more cost- and resource-effective way for internet communities to moderate abusive content.

They call it the Bag of Communities (BoC), a technique that leverages large-scale, preexisting data from other internet communities to train an algorithm to identify abusive behavior within a separate target community.

Specifically, they identified nine different communities. Five, such as the free-for-all of internet communities 4chan, are rife with abusive behavior from commenters; four, like the heavily moderated MetaFilter, are helpful, positive, and supportive.

Using linguistic characteristics from these two types of communities, researchers built an algorithm that can learn from the comments and, when a new post is generated within a target community, it can make a prediction of whether or not it is abusive.

"MetaFilter is known around the internet as a good, helpful, supportive community," said Eric Gilbert, an associate professor in the School of Interactive Computing and a member of the team of researchers on the project. "That's an example of how, if your post is closer to that, it's more likely that it should stay on the site. Conversely, if your post is closer to 4chan, then maybe it should come off."

The researchers provide two algorithms. One is a static model, off the shelf with no training examples from the target community, and can achieve roughly 75 percent accuracy. In other words, with access only to posts from the other nine communities, the algorithm can accurately predict abusive posts in the target community roughly three quarters of the time.

"A new community that does not have enough resources to actually build automated algorithms to detect abusive content could use the static model," said Georgia Tech doctoral student Eshwar Chandrasekharan, who led the team.

A dynamic model, one that mimics scenarios in which newly moderated data arrives in batches, learns over time and can achieve 91.18 percent accuracy after seeing 100,000 human-moderated posts.

"Over time, as new moderator labels come in, when it has seen examples of things that have been moderated from the site, it can learn more site-specific information," Chandrasekharan said. "It can learn the type of comments that get moderated, and if there is a level of tolerance that is different from what you see in the static model, it could learn that over time."

Both the static and dynamic models outperformed a solely in-domain model from a major internet community.

Anyone who has managed an online community has encountered problems with abusive content from users. From social media to message boards to comments sections in online news publications, regulating what is and isn't allowed has become overly costly and taxing on existing human moderators.

Founders at social media startup Yik Yak spent months of their early time removing hate speech, and Twitter has stated publicly that dealing with abusive behavior remains its most pressing challenge. A number of major news agencies are buried under the demands of strict moderation, and many have shut down comments sections altogether.

Prior research into abuse detection and online content moderation has focused on in-domain methods – using data collected from within your own community – but those face challenges in obtaining enough data to build and evaluate algorithms. In a BoC-based method, algorithms would leverage out-of-domain data from other existing online communities.

Gilbert said that the applications from such a model could be widespread.

"This is a core internet problem," he said. "So many places struggle with this, and many are shutting comments off because they just don't want to deal with the trouble they cause."

This research is presented in a paper (The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data) at the Association for Computing Machinery CHI Conference on Human Factors in Computing Systems 2017.

Provided by Georgia Institute of Technology

Citation: Research may help combat abusive online comments (2017, May 10) retrieved 17 April 2024 from https://phys.org/news/2017-05-combat-abusive-online-comments.html