Algorithm could quash Twitter abuse of women

August 28 2020

Credit: Pixabay/CC0 Public Domain

Online abuse targeting women, including threats of harm or sexual violence, has proliferated across all social media platforms but QUT researchers have developed a statistical model to help drum it out of the

Twittersphere.

Associate Professor Richi Nayak, Professor Nicolas Suzor and research fellow Dr. Md Abul Bashar from QUT have developed a sophisticated and accurate algorithm to detect these posts on Twitter, cutting through the raucous rabble of millions of tweets to identify misogynistic content.

The team, a collaboration between QUT's faculties of Science and Engineering and Law and the Digital Media Research Center, mined a dataset of 1M tweets then refined these by searching for those containing one of three abusive keywords—whore, slut, and rape.

Their paper—"Regularizing LSTM classifier by transfer learning for detecting misogynistic tweets with small training set"—has been published in *Knowledge and Information Systems*.

"At the moment, the onus is on the user to report abuse they receive. We hope our machine-learning solution can be adopted by social media platforms to automatically identify and report this content to protect women and other user groups online," said Professor Nayak.

"The key challenge in misogynistic tweet detection is understanding the context of a tweet. The complex and noisy nature of tweets makes it difficult. On top of that, teaching a machine to understand natural language is one of the more complicated ends of data science: language changes and evolves constantly, and much of meaning depends on context and tone. So, we developed a text mining system where the algorithm learns the language as it goes, first by developing a base-level understanding then augmenting that knowledge with both tweet-specific and abusive language. We implemented a deep learning algorithm called Long Short-Term Memory with Transfer Learning, which means that the machine could look back at its previous understanding of terminology and change the model as it goes, learning and developing its contextual

and semantic understanding over time."

While the system started with a base dictionary and built its vocabulary from there, context and intent had to be carefully monitored by the research team to ensure that the algorithm could differentiate between abuse, sarcasm and friendly use of aggressive terminology.

"Take the phrase 'get back to the kitchen' as an example—devoid of context of structural inequality, a machine's literal interpretation could miss the misogynistic meaning," said Professor Nayak.

"But seen with the understanding of what constitutes abusive or misogynistic language, it can be identified as a misogynistic tweet. Or take a tweet like "STFU BITCH! DON'T YOU DARE INSULT KEEMSTAR OR I'LL KILL YOU." Distinguishing this, without context, from a misogynistic and abusive threat is incredibly difficult for a machine to do.

"Teaching a machine to differentiate context, without the help of tone and through text alone, was key to this project's success, and we were very happy when our algorithm identified 'go back to the kitchen' as misogynistic—it demonstrated that the context learning works."

The research team's model identifies misogynistic content with 75% accuracy, outperforming other methods that investigate similar aspects of social media language.

"Other methods based on word distribution or occurrence patterns identify abusive or misogynistic terminology, but the presence of a word by itself doesn't necessarily correlate with intent," said Professor Nayak.

"Once we had refined the 1M tweets to 5,000, those tweets were then categorized as misogynistic or not based on context and intent, and were

input to the machine learning classifier, which used these labeled samples to begin to build its classification model. Sadly, there's no shortage of misogynistic data out there to work with, but labeling the data was quite labor-intensive."

Professor Nayak and the team hoped the research could translate into platform-level policy that would see Twitter, for example, remove any tweets identified by the algorithm as misogynistic.

"This modeling could also be expanded upon and used in other contexts in the future, such as identifying racism, homophobia, or abuse toward people with disabilities," she said.

"Our end goal is to take the model to social media platforms and trial it in place. If we can make identifying and removing this content easier, that can help create a safer online space for all users."

More information: Md Abul Bashar et al. Regularizing LSTM classifier by transfer learning for detecting misogynistic tweets with small training set, *Knowledge and Information Systems* (2020). DOI: 10.1007/s10115-020-01481-0

Provided by Queensland University of Technology