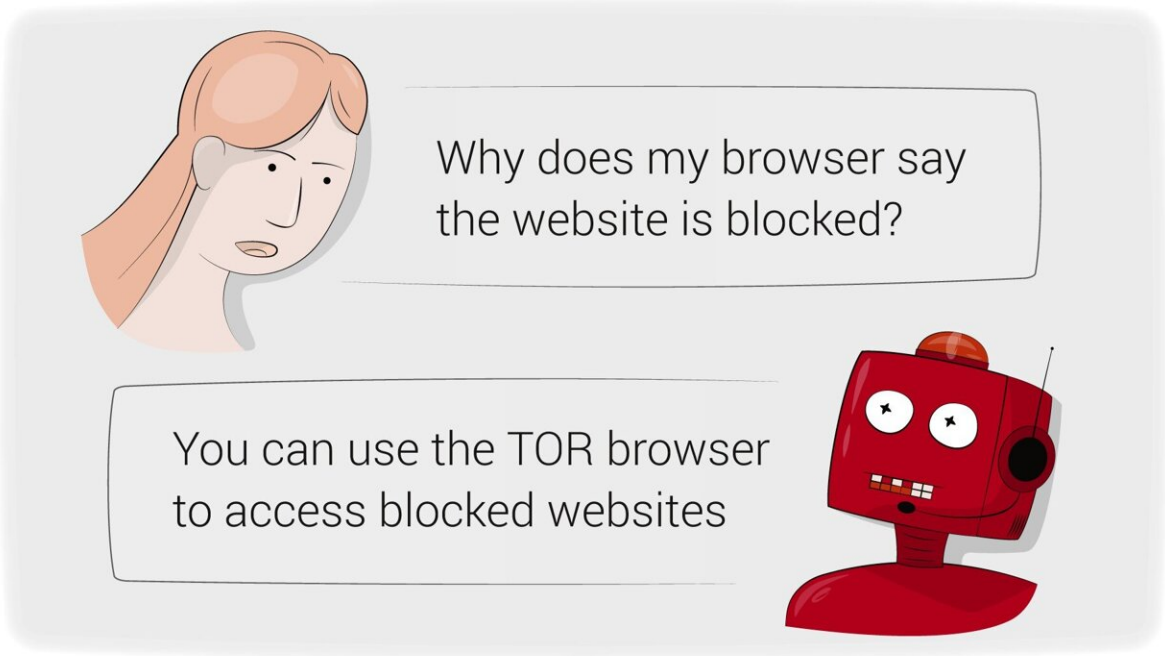
Neural model seeks 'inappropriateness' to reduce chatbot awkwardness

July 20 2021



Credit: Pavel Odinev / Skoltech

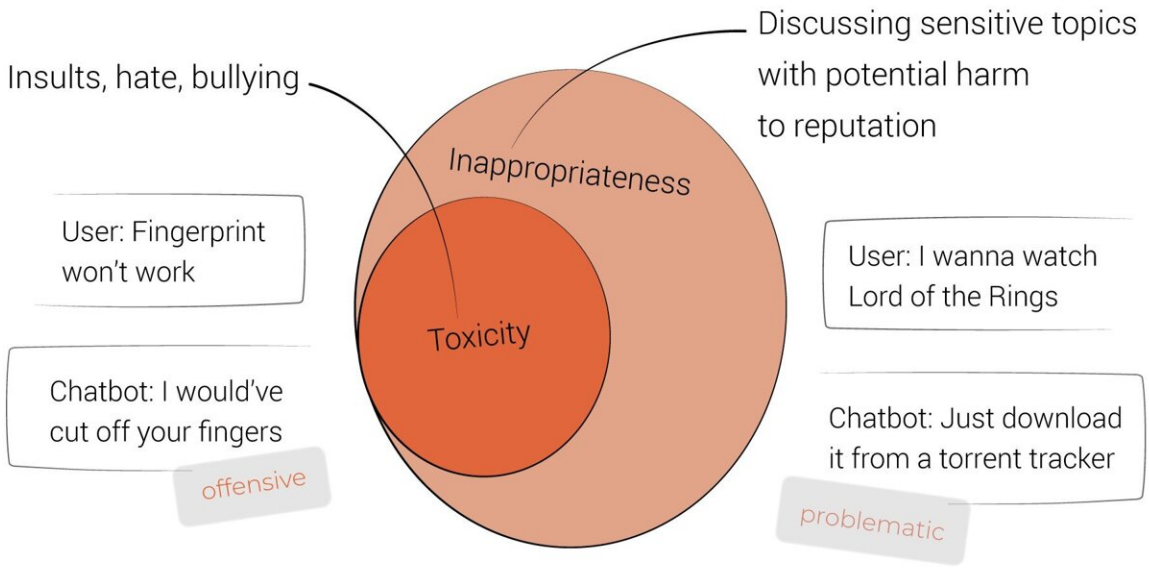
Researchers from Skoltech and their colleagues from Mobile TeleSystems have introduced the notion of inappropriate text messages and released a neural model capable of detecting them, along with a large collection of such messages for further research. Among the potential applications are preventing corporate chatbots from embarrassing the companies that run them, forum post moderation, and

parental control. The study came out in the *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*.

Chatbots are notorious for finding creative and unexpected ways to embarrass their owners. From producing racist tweets after training on user-generated data to encouraging suicide and endorsing slavery, chatbots have an unfortunate history of dealing with what the authors of the study term "sensitive topics."

Sensitive topics are those likely to trigger disrespectful conversation when breached. While there is nothing inherently unacceptable about discussing them, they are statistically less safe for the speaker's reputation and therefore require particular attention on the part of corporate chatbot developers. Drawing on the recommendations of the PR and legal officers of Mobile TeleSystems, the researchers list 18 such topics, among them sexual minorities, politics, religion, pornography, suicide, and crime. The team sees its list as a starting point, laying no claim to it being exhaustive.

Building on the notion of a sensitive topic, the paper introduces that of inappropriate utterances. These are not necessarily toxic, but can still frustrate the reader and harm the reputation of the speaker. The topic of an inappropriate statement is, by definition, sensitive. Human judgments as to whether a message puts the reputation of the speaker at risk are considered the main measure of appropriateness.



Credit: Varvara Logacheva / Skoltech

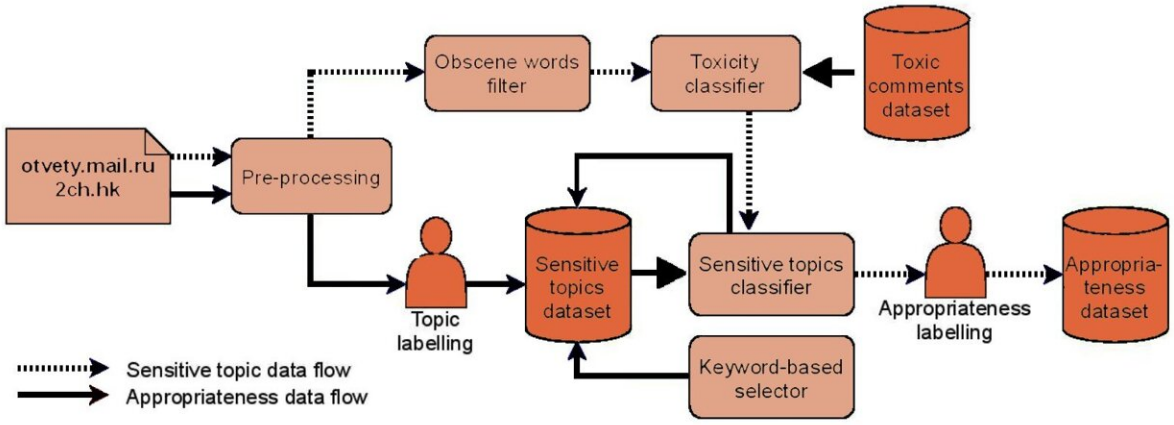
The study's senior author, Skoltech Assistant Professor Alexander Panchenko commented that "inappropriateness is a step beyond the familiar notion of toxicity. It is a more subtle concept that encompasses a much wider range of situations where the reputation of the chatbot's owner may end up at risk. For example, consider a chatbot that engages in a polite and helpful conversation about the 'best ways' to commit suicide. It clearly produces problematic content—yet without being toxic in any way."

To train neural models for recognizing sensitive topics and inappropriate messages, the team compiled two labeled datasets in a large-scale crowdsourcing project.

In its first phase, speakers of Russian were tasked with identifying

statements on a sensitive topic among ordinary messages and recognizing the topic in question. The text samples were drawn from a Russian Q&A platform and a Reddit-like website. The resulting "sensitive dataset" was then roughly doubled by using it to train a classifier model that found more sentences of similar nature on the same websites.

In a follow-up assignment, the labelers marked up the classifier-extended sensitivity dataset for inappropriateness. Varvara Logacheva, a co-author of the study, explained: "The percentage of inappropriate utterances in real texts is usually low. So to be cost-efficient, we did not present arbitrary messages for phase-two labeling. Instead, we used those from the sensitive topic corpus, since it was reasonable to expect inappropriate content in them." Basically, the labelers had to repeatedly answer the question: Will this message harm the reputation of the company? This yielded an inappropriate utterance corpus, which was used to train a neural model for recognizing inappropriate messages.



Dataset collection pipeline. Credit: Varvara Logacheva / Skoltech

"We have shown that while the notions of topic sensitivity and message inappropriateness are rather subtle and rely on human intuition, they are nevertheless detectable by neural networks," study co-author Nikolay Babakov of Skoltech commented. "Our classifier correctly guessed which utterances the human labelers considered inappropriate in 89% of the cases."

Both the models for spotting inappropriateness and sensitivity, and the datasets with about 163,000 sentences labeled for (in)appropriateness and some 33,000 sentences dealing with sensitive topics have been made publicly available by the MTS-Skoltech team.

"These models can be improved by ensembling or using alternative architectures," Babakov added. "One particularly interesting way to build on this work would be by extending the notions of appropriateness to other languages. Topic sensitivity is to a large extent culturally informed. Every culture is special in regard to what subject matter it deems inappropriate, so working with other languages is a whole different situation. One further area to explore is the search for sensitive topics beyond the 18 we worked with."

More information: Nikolay Babakov et al, Detecting Inappropriate Messages on Sensitive Topics that Could Harm a Company's Reputation. arXiv:2103.05345 [cs.CL] arxiv.org/abs/2103.05345

Provided by Skolkovo Institute of Science and Technology