The internet has a dark side, but can we teach machines how to identify it?

July 10 2023, by Avery Anderson

Credit: Cup of Couple/Pexels

With great power comes great responsibility. In terms of the internet, power is found in the multitude of information available to users everywhere—but who is responsible for making sure the information out there is good and true?

"Bad" information has serious implications. Misinformation, propaganda, and fake news are prevalent on the web and on social media

platforms and can become weaponized, which leads to cyber abuse and, in severe cases, civil unrest.

The University of Southern California's Information Sciences Institute (ISI), a unit of the Viterbi School of Engineering, is working on two projects aimed at solving this issue from the inside out—by developing technology that can exercise reasoning capabilities when encountering this "bad" information.

This technology would serve as an assistant to human moderators whose job is to surveil online platforms and scan for malicious content.

Technology you can trust

The first project involves the detection of logical fallacies in natural language arguments.

So what exactly is a logical fallacy?

Logical fallacies are errors in reasoning used to prove that an argument is true. Their origins can be traced back well before the internet era—their debate in the realm of philosophy has its roots in ancient Greece, some 2,800 years ago.

In the context of the web, logical fallacies show up in the form of false or deceptive statements that circulate as a result of the large-scale free exchange of information enabled by the internet.

Filip Ilievski, research lead at ISI and assistant professor at USC, said that finding logical fallacies is the first step to master before tackling the real giants that can manifest as a result of information sharing activities on the web.

"Once you are able to reliably and transparently identify logical fallacies, you can then apply that technology to deal with cases of misinformation, fake news, and propaganda," Ilievski said.

This work is the first of its kind to apply multiple layers to the detection of logical fallacies, Ilievski explained. This entails asking the model to first determine whether the given argument is a sound one, and then go "one level deeper" and "identify on a high level what kind of fallacy the argument contains."

How do they know?

Explainable AI can pinpoint logical fallacies and classify them in two prominent ways: case based reasoning and prototyping methods.

Ilievski noted that ISI's work is among the first to combine the two with language models and "make them scale to arbitrary situations and tasks."

Case based reasoning is exactly like it sounds. The model is shown an old example of an argument with similar logical fallacies and then uses this knowledge to infer its decisions about a new argument.

"You say well, I don't know how to solve this argument but I have this old example you can use on the new one in front of you," Ilievski explained.

Prototypical methods follow the same process. The only difference is, the model makes inferences from a simplified, basic case that can be built upon and applied to a specific example.

The key here is that these models are doing more than just identifying a logical fallacy—they are giving reasonable explanations to back up their judgment, an action that Ilievski says is an "encouraging factor" for the

future of these methods in practice.

A man's best tool

How does this apply in the real world, against the real giants—propaganda, misinformation, and fake news—that pose threats online?

Ilievski envisions these explainable AI acting as a "human assistant tools" that help moderators or analysts who monitor online communities.

Moderators are responsible for overseeing the activity of millions of users exchanging ideas 24/7. Manually checking for fallacies, given their volume and complexity, is overwhelming. Adding machine learning to the team helps mitigate this burden.

"Let's say you had a moderator on a social media platform and they want to understand whether something is fallacious. It would help to have a tool like this provide assistance and surface potential fallacies, especially if they are linked to propaganda and potential misinformation," Ilievski explained.

The explainability factor, or the ability for AI to provide reasoning behind the fallacies it identifies, is what truly "fosters trust and usage in human-AI frameworks," he added.

However, he cautions that explainable AI are not tools we should blindly trust.

"They can make our lives easier, but they aren't sufficient on their own," Ilievski noted.

Memes, misogyny and more

Explainable AI can also be taught how to identify memes that contain problematic elements, such as "dark humor" that is sometimes flat out discriminatory and offensive to specific groups of people or society as a whole.

For this second project, the team focused on two specific types of harmful meme content: misogyny and hate speech.

Zhivar Sourati, a graduate student at USC who is working alongside Ilievski on both of these projects, says that transparent detection of memes that have problematic underpinnings is crucial with how fast information spreads online.

"For content moderators, it is really important to be able to detect these memes early on, because they spread on social media like Twitter or Facebook and reach large audiences very quickly."

By nature, Sourati says, memes are dependent on more aspects than meet the eye. Although memes are known for being succinct (sometimes they contain only a simple picture) they often reflect cultural references that can be hard to explain.

"You have an image, and then maybe not even a sentence but a piece of text. It probably refers to a concept, a movie, or something that is in the news," Sourati explained. "You just instantly know that it's funny, but it's really hard to explain why, even for human beings and that's the case for machine learning as well."

This unexplainable aspect of memes makes it even more challenging to teach machine learning how to classify them, because they must understand the intention and meaning behind them first.

Getting down to the nitty gritty

The framework Ilievski and Sourati used is called "case based reasoning."

Case based reasoning is essentially the way humans approach a problem: learning from previous examples and applying that knowledge to new ones.

The machine is shown a couple examples of memes that are problematic and the reason why. Then, Sourati says, the machine is able to build a library of examples, so when it is tasked with classifying a new meme that might have "a bit of abstraction from the previous examples," it can "approach the new problem with all of the knowledge that it carries so far."

For example, if they were specifically focusing on misogyny, they may ask: "Why is this meme misogynistic? Is it shaming? Is it a stereotype? Is it demonstrating objectification of a woman?"

They used an explanatory interface to visualize the models' reasoning and understand why the model is predicting the way it is. This visualization tactic helped with troubleshooting and improving the model's skills.

"One benefit is that we can perform easier error analysis. If our model makes 20 mistakes out of 100 cases, we can open those 20 and look to see a pattern of the model's biases in terms of different demographics of what is depicted or a specific object indicated," Ilievski explained. "Maybe whenever it sees an ice cream, it thinks that is misogyny."

Humans and AI: A heroic duo

Just as logical fallacy detection, meme classification cannot be done fully automatically and requires human–AI collaboration.

That being said, Ilievski and Sourati's findings show a promising future for AI's ability to help humans detect hate speech and misogyny in memes.

The complexity of understanding memes, or "element of surprise" as Ilievski put it, made this topic especially exciting to work on.

"There is an element of difficulty that makes this process very interesting from the perspective of AI because there is information that is implicit in memes," Ilievski said.

"There are cultural and contextual dimensions, and a notion of it being very creative and personal to the creator of the meme. All of this together made this project especially exciting to work on," he added.

The ISI team made their findings and code available for use by other researchers, in hopes that future work will continue developing AI's ability to support humans in their fight against dangerous and harmful content online.

Provided by University of Southern California