# Fighting hate speech and misinformation online

Dr Srijan Kumar, assistant professor at Georgia Institute of Technology and a Forbes 30 Under 30 honoree in science, discusses with *Nature Computational Science* how he uses machine learning and data science to identify and mitigate malicious activities on online platforms, including misinformation and anti-Asian hate speech.

**How did your interest in computer science and data science start?**

I became interested in computer science mainly because I was always interested in gaming: I used to prefer playing computer games over spending time outdoors, much to my parents' displeasure. But, essentially, this got me interested in computers more broadly speaking, as I wanted to learn coding so that I could develop games. Funnily enough, I never actually ended up creating any games! Another reason for my interest in computer science was my uncle, who was a computer engineer and certainly inspired me to pursue this career. As I delved more into computer science, I started realizing the power that data has, and that is how I got into machine learning (ML) and data science. I got a very initial introduction on these topics as an undergrad, but then I came to the US for my Master's and PhD programs and really started working on these topics.

**How did you get into the field of early identification, prediction, and mitigation of online misinformation and malicious activities and actors?**

I have always been interested in creating something that can help others. I grew up in the age of social media — I had a Facebook account when I was in high school — and I used to spend a lot of time on these platforms simply because they are so fun, interesting and useful. It was just around that time when social media was getting started, and I really saw firsthand all the bad things that were happening on the Internet — sometimes they would happen with friends of mine who would get harassed online — and I realized that there was such a pressing need to do something about it. I put my computer scientist hat on and thought to myself: there are so many malicious actors and there is so much misinformation and malicious content out there — what can I do? This ultimately got me really interested in my current line of work, which is about using artificial intelligence (AI), ML and data science to solve safety, integrity, and well-being issues around users, content, platforms, and communities in the entire cyberspace. I also realized that there are two main factors at play: malicious actors who

are doing harm and adding harmful content online, and technological elements, such as recommender systems, that can exacerbate these harms. So, I started thinking about how I could create algorithms that are not only accurate, but also robust, reliable, and trustworthy, in order to address some of these issues by taking into account both factors.

**What are the types of tools and computational techniques that are needed to address these challenges?**

Many algorithms and tools have been developed for the early detection of malicious activities and actors, for instance, by using natural language processing (since much of the data has text) and social network analysis (since many of the malicious activities happen on a social network), or by identifying bots and different types of patterns (such as the lockstep behavior, where groups of users act together).

But there is a lot more that needs to be done. One of the biggest challenges that the community faces nowadays is how to ensure the robustness and reliability of cyber safety systems. Malicious actors are always trying to evade the system: they try not to be identified, but even if they eventually get caught, they find ways to create new accounts in order to continue performing their malicious activities. What can we, as algorithm and system designers, do about this? Another issue that we face is that most of the research that has been done so far has only taken into account content written in English. However, there are a lot of malicious activities that happen in other languages, or that make use of other modalities, such as images and videos. Most of the research has focused on Twitter and Reddit because it is so easy to collect data from these platforms, but the next generation of computational solutions needs to be multilingual, multimodal, and multiplatform. Finally, a third challenge is to understand how the algorithms and systems that are being used in practice are exacerbating the problem. For instance, a lot of what we are seeing and consuming online is based on what is recommended to us by recommender systems. But how robust and trustworthy are these systems? Do these systems lead to polarization and formation of echo chambers and filter bubbles? What is the role that these systems play in pushing certain types of narratives and content? All of these are very timely and pertinent questions for which we need solutions.

**In your opinion, what is the role of online platforms to combat malicious activities?**

I think these platforms have a major responsibility in helping to solve these problems. A lot of these platforms already have several teams — not just one — within their organizations with very good research capabilities and skills to help alleviate some of these issues. But what we see nowadays is that most of these platforms are reactive: when something bad happens, they are scrambling to fix that. What I would like to see is for these platforms to go from a reactive nature to a proactive one; to understand what can go wrong before it actually goes wrong; to be one step ahead of the problem, instead of one step behind it. I think that is something that these platforms need to own up to, because they have become such a vital part of our lives: people spend around 20% of their time online, and users today consume more news and information from social media than from

traditional news organizations. In addition, these platforms influence not just what we do online, but they also influence what we do in the real world: misinformation and hate speech, for instance, harm our capabilities to make informed decisions, and reduce trust in science, public health organizations, electoral processes and democracy as a whole. So, given how pervasive these platforms are, and their role in our lives and actions, they need to be responsible for solving these problems in a proactive way.

**How challenging is it to solve these issues proactively? How can this be done more effectively?**

Being proactive does not mean simply removing everything that the algorithm concludes to be questionable, as the accuracy is not perfect. There are different levels of moderation that can and should be done depending on the situation. For instance, at a very high level of human moderation, platforms can hire more fact checkers in collaboration with fact-checking organizations that are doing the hard work of labeling information veracity. Alternatively, on the other side of the spectrum, we can use algorithms that can identify questionable content, but in a human-in-the-loop setting instead of in a completely automated setting: the humans in the loop here have the authority and the experience to identify the good from the bad, and they use the content identified by the algorithms to make an informed decision.

The latter solution is particularly useful when there is an overwhelming amount of information to go through. For some of the work that professional fact checkers are doing, there is just so much misinformation out there that they do not have the capacity to handle everything: they spend hours just scrolling through different social media platforms and trying to identify what needs to get fact checked. So, a lot of this manual work can be automated, and they really need tools and algorithms that can help them to prioritize what needs the most attention. For instance, my group and I created an algorithm to identify fake reviews on e-commerce platforms. Eventually, this algorithm was integrated into Flipkart, which is one of the largest e-commerce platforms in India: they used the algorithm in conjunction with other systems that they have, but also in conjunction with human moderators in order to identify and remove fake reviews from the platform. Similarly, we are currently working together with the Wikimedia Foundation to implement a system based on our latest work on online ban evasion. The main idea here is to create a tool to help Wikipedia moderators to identify when new accounts are created by actors who have been previously banned from the platform due to malicious activities.

**You have also investigated the power of crowdsourcing for combating misinformation. How can crowdsourcing be effectively used for this purpose?**

For several years, there has been a lot of emphasis — and for a good reason — on professional fact checkers being in the front line of defense against misinformation. But then we were curious: how prevalent are the professional fact-checking efforts on social platforms? Are regular users, like you and me, also engaging with fact checking? How can these regular users help to identify and counteract misinformation? A lot of time, regular users are the ones who actually see the misinformation and may get suspicious about it: if we are able to leverage this community of people and empower them to identify and react to misinformation as early as possible, this can be potentially very useful. So we started looking at essentially how often people counteract misinformation on social platforms, using data from Twitter. We found that 96% of all tweet activity related to counteracting misinformation was being made by regular users, meaning that only 4% of these tweets were made by professional fact checkers, demonstrating that these regular users have a critical role to play here. Crowdsourcing has been used by Twitter's Birdwatch platform, for instance: this is a community-driven effort in which a regular user can sign up to flag content as being a product of misinformation or not.

However, there is a major problem in this process: malicious actors can manipulate the crowdsourcing platform itself to mislabel accurate information as false and conversely, false information as accurate. Thus, as part of our research, we created a reputation system called HawkEye, which is used to essentially flag misleading tweets and to identify who can be trusted not only on Twitter, but also in the Birdwatch platform.

**More recently, you have also worked in characterizing and detecting online anti-Asian hate speech. Can you describe this work? What have you learned from this work?**

We started looking into this topic around late March 2020, when everyone was desperate about COVID-19, and when we started seeing news reports of hate speech, physical attacks, and harassment against people of Asian descent. Having spent a lot of time in the domain of online social media, we started looking at this problem of hate speech but on online platforms. In addition, while hate speech was spreading on social media, there were also people who were countering hate speech, in support of people of Asian descent: we had these two competing narratives simultaneously spreading on social media platforms. We then started collecting data on Twitter related to this phenomenon, starting from January 2020: we essentially crawled millions of tweets from hundreds of thousands of users on these topics, and we did one of the first analyses of anti-Asian hate speech and counterspeech on social media.

First, we created a hand-labeled dataset with around 3,200 tweets to train a classifier. Next, we used our classifier to identify hate speech and counterspeech from the rest of the data. In total, we identified 1.3 million tweets containing anti-Asian hate speech and 1.1 million tweets containing counterspeech. With this large-scale data, we started doing different types of analysis in order to understand how hateful comments were spreading, how users were spreading both hate speech and counterspeech, and how these two narratives were influencing each other. One of the most important findings we had was that the more hate speech you see, the higher the likelihood of you making hateful comments: if a lot of your friends, meaning if a lot of people in your social platform neighborhood, are spreading hate, you are more likely to spread hate as well. In other words, hate speech is contagious! However, there is some hope, as we found initial evidence that counterspeech can slightly prevent hate speech from being taken up by others: there is a small inhibition effect in terms of counterspeech being able to prevent users from making hateful comments in the first place.

Again, we are seeing the same theme here of regular users being one of the most effective ways to combat malicious actors and activities by speaking up. Essentially, we not only need computational tools to help us identify these malicious activities, but we also need community-driven efforts to effectively counteract these issues. We need regular users to be more aware of these issues, and we need them to be more proactive and to speak up — for instance, by simply flagging inappropriate content — when they see bad behavior.

**Do these community-driven and crowdsourcing approaches require a multidisciplinary effort that goes beyond computer science?**

Absolutely, because these approaches also depend on how users engage in online

platforms, and how they respond to misinformation and hate speech comments. I collaborate very closely with social scientists and communication experts, and together, we can bridge together the two distinct worlds of social science and computer science to help solve these critical issues.

**Do the same computational challenges that need to be addressed to help combat misinformation also apply to identifying and counteracting hate speech?**

Yes, absolutely, these challenges still apply here. We need multilingual, multimodal and multiplatform support since, just like misinformation, hate speech also occurs in multiple languages, in multiple modalities, and across different online platforms. We also need systems that are robust, trustworthy, secure, and fair. Minority groups are the most impacted groups of people when it comes to misinformation and hate speech, and we do not want to exacerbate these issues using technology.

Overall, combating misinformation and hate speech is a very challenging topic, and unfortunately, we do not have a panacea for these issues yet. But hopefully, with efforts from multiple stakeholders and domain experts, we will be able to help alleviate some of these issues.

Interviewed by Fernando Chirigati

Published online: 1 May 2022
https://doi.org/10.1038/s43588-022-00238-9