How AI Is Now Being Trained To 'Detoxify' Social Media

Zak Doffman

Is AI the answer to frustrating the building pressure on social media to be regulated or even restricted? As Google finally brings the shutter down on its own ill-fated Google+, can its world-leading artificial intelligence provide the solution for other platforms as they battle to restore safe online environments? That might be the plan, but making this work is proving one of the toughest tests yet for the company's AI. It turns out that moderating our behavior is much harder than manipulating it.

The Antisocial Network

Most teens have now experienced or witnessed cyberbullying, and the lack of controls on social media distribution of damaging content continues to make frontpage headlines. Meanwhile, 'grown-up' discussion forums have become echo chambers that enable and even encourage organized harassment.

According to Pew Research, almost half of American teens claim to be online "almost constantly", with the social media platforms YouTube, Instagram, Snapchat, Facebook and Twitter predominating. Approaching 60% of teens have been bullied or harassed online, with 90% acknowledging it as a real issue. And it's not much better for older generations: 41% of U.S. adults have experienced online harassment themselves and 66% have seen it directed at others.

So, just to spell it out, we are allowing our kids to live online in an environment we acknowledge is becoming ever more toxic, whilst, for ourselves, we have accepted a hostile online atmosphere that would not be tolerated in the physical world. This might all need a rethink.

Spiraling Out Of Control

Following headlines that photo-sharing platforms, including Instagram, have provoked teen suicides, the U.K.'s Children's Commissioner wrote this week to social media organizations. "With great power comes great responsibility," she told them, "and it is your responsibility to support measures that give children the information and tools they need growing up in this digital world – or to admit that you cannot control what anyone sees on your platforms."

Last week, U.K. TV presenter, Rachel Riley, turned from Twitter to The Times to continue her brave stand against the cowardly anti-Semitic trolling that has become commonplace from self-avowed supporters of the country's opposition party leader, Jeremy Corbyn. "The more I speak, the more abuse I get," she told the newspaper, "and the more abuse I get, the more I speak. It's got to the point where I can't look at my Twitter feed any more… it's just a constant stream."

In a joint investigation last year, *Engadget* and *Point* interviewed moderators from the online discussion platform Reddit, the most popular website in the U.S. behind Google and YouTube. It's clear that Reddit's subscribers do not want to be censored in any way. Every one of the ten moderators interviewed said they had received death threats, and some of the women rape threats as well. They complained that "Reddit is systematically failing to tackle the abuse they suffer… and [has] repeatedly neglected to respond to moderators who report offenses."

I could go on and on: the examples and case studies and reports and analysis are endless. This has become one of the most universal problems of our time. We now entrust a handful of platforms with great swathes of our lives. And those same platforms are informally educating our kids. Something clearly needs to change.

Is AI The Answer?

"Can machines think?" So begins the 1950 paper '*Computing Machinery and Intelligence*', which defined the so-called Turing Test for AI, where a machine provides conversational responses that fool a person into thinking they are speaking with a fellow human being.

Clearly, AI is more than mimicking conversation, it is taking intelligent decisions and automating human activity. But the Turing test survived for almost seventy-years as a benchmark. And then last year Google demonstrated Duplex, an AI tool to make telephone appointments. The Duplex AI engine's conversational skills as it makes a restaurant reservation, replete with lifelike pauses, are eerily convincing – albeit in a simple scenario. With that caveat, though, it arguably passes the test.

Whilst Google CEO Sundar Pichai grabbed headlines with his Duplex demonstration, a smaller, more discrete part of the Google empire continued working from their offices in NYC, grappling with a challenge that is a much more difficult and much more relevant test of AI.

Back in 2016, Eric Schmidt, the then Executive Chairman of Google's parent Alphabet, announced the expansion of Google Ideas into a technology incubator called Jigsaw. "The team's mission," he wrote, "is to use technology to tackle the toughest geopolitical challenges, from countering violent extremism to thwarting online censorship to mitigating the threats associated with digital attacks."

Jared Cohen, the swashbuckling CEO and driving force behind Jigsaw, makes for an unlikely figurehead within the Google empire. Tackling extremism face to face doesn't fit naturally within the world's most successful experiment in the manipulation of human behavior for profit. And Jigsaw definitely gets to play at the sharp-end, including countering teen radicalization by the likes of IS. Whether one finds the repurposing of ruthlessly commercial AI for more philanthropic purposes ironic or altruistic will be subjective. Either way, the goals and ambitions at Jigsaw are laudable.

Jigsaw's various projects tackle online challenges from "thwarting censorship to mitigating the threats from digital attacks to countering violent extremism to protecting people from online harassment." Most projects focus on cybersecurity and resilience: protecting against DDoS attacks, securing open communication, defeating censorship. One of those hit the headlines this week, with Project Shield technology made available to protect organizations participating in European Union politics from DDoS attacks. With Brexit imminent, the forthcoming elections are likely to draw an exceptional level of attention and scrutiny.

Jigsaw’s Conversation AI and spinout Perspective projects are outliers. Their challenge is more difficult and delves into the very heart of the ways in which AI relates to humans in all our contradictory complexity. These projects have been set up to apply machine learning to classify and identify abusive or threatening language online. But isolating abusive language from the normal, nuanced conversation is a tougher test of AI than booking a table in a restaurant.

Yes, But…

In an in-depth article this week for *PCMag*, Rob Marvin delves into Perspective to look at the machine learning under the hood and to assess the progress that has been made. The initial focus for the project team has been moderated discussion forums such as Reddit and comment sections for media publications such as the New York Times. The issues are very clear-cut.

"It's very easy to ruin an online conversation," explains Perspective’s C.J. Adams. "It's easy to jump in, but one person being really mean or toxic could drive other voices out. Maybe 100 people read an article or start a debate, and often you end up with the loudest voices in the room being the only ones left, in an internet that's optimized for likes and shares.”

To tackle this problem, “you need natural language processing,” Marvin writes in his article, “which breaks down a sentence to spot patterns. The Perspective team is confronting problems like confirmation bias, groupthink, and harassing behavior in an environment where technology has amplified their reach and made them harder to solve.”

Perspective has had access to household names like the New York Times, it has run open experiments to encourage broad participation in swamping its algorithms with realistic content, it has analyzed a lot of data and churned through a lot of AI training. Ultimately, though, the conclusions back up findings from other platforms: AI doesn’t provide easy answers or shortcuts to replace the armies of content reviewers and moderators being put in place.

Rob Marvin pointed me at Perspective’s ‘False Positive’ blog, in which they share some of their experiences and lessons learned. “Some of the ways they got it very wrong early in the process and have used those trial-and-error experiences to refine the ML training… are fascinating,” he told me. And he’s absolutely right.

To cut to the chase, the biggest takeaway thus far is just how difficult it is for a machine to interpret much beyond the obvious use of inflammatory, threatening or offensive language. Conversation is rooted in emotion, interpretation, nuance, sarcasm. It’s only by tackling the machine learning implications of trying to automate this level of language understanding that you realize how far we have to go before we get to the principles underlying Turing’s Test. To glibly dismiss the 1950's example as being out of touch with modern reality misses the point. Convincing a 1950’s human is not the same as convincing one today. We live online. Convince online.

"This tool is helpful for machine-assisted human moderation,” explains Adams, “but it is not ready to be making automatic decisions. But it can take the ‘needle in a haystack' problem finding this toxic speech and get it down to a handful of hay.”

The New York Times talks positively about this narrowing down the haystack approach, about enabling their moderators to cover more ground, to review more content, albeit only a modest percentage of material is automatically cleared or removed with no intervention.

Will Anything Change?

The last 12 months have seemed like a slow-motion train wreck for Big Data and Social Media, with one negative headline after another. Google certainly hasn't had the worst of it, that distinction goes to Facebook which has been following something akin to Kim Jong-un's PR playbook. But Google did capitulate after an internal revolt over the sale of AI to the US government, it did pledge to quit selling facial recognition to law enforcement, it did face down Google+ security lapses and it did see staff walkout over alleged sexual misconduct by execs.

In the last few weeks, Google also hit the headlines with Shoshana Zuboff's heralded new book, '*The Age of Surveillance Capitalism*'. The author attributes Google with the origin of surveillance capitalism and the birth of the multi-trillion-dollar data economy. "Google found a game-changing, zero-cost asset that could be diverted from service improvement towards a genuine commercial exchange. Surveillance capitalism soon migrated to Facebook and rose to become the default model for capital accumulation in Silicon Valley, embraced by every start-up and app."

Google is as responsible as anyone for the Frankenstein's Monster that is today's social media. With 2019 barely a month old, we have already seen headline after headline around data exploitation and ruthless commercialization. AI is now being applied to every imaginable facet of that world and so it is only fitting that it is used to try to redress the balance – if that's at all possible. But ironically it is proving much harder to control what we say than influence what we think - make of that what you will.

The real question, though, is around the level of genuine intent on behalf of the platforms, including Google. Perhaps the threat of regulation and sanctions will prompt further investment in technology of this kind. But then again maybe there's a general realization that such restrictions are unlikely. Facebook's answer to its PR challenges of the last few weeks, including a very public spat with Apple, was to post a set of record financial results.

"Beware, for I am fearless and therefore powerful," cried the monster, confronting its regretful creator in Shelley's novel.