Cyberbullying 'shield' app uses AI to combat social media trolls

July 16 2020

Credit: CC0 Public Domain

Computer scientists from Aston University, Birmingham, have launched an app that uses novel artificial intelligence (AI) algorithms to combat trolling and bullying online.

The downloadable app, Bullstop, is the only anti-cyberbullying app that integrates directly to social media platforms to protect users from bullies and trolls messaging them directly.

It was developed by Semiu Salawu, a doctoral researcher in the College of Engineering and Physical Sciences and was initially designed with teenagers aged 13 upwards in mind.

The app was designed collaboratively with young people to improve its chance of adoption amongst the main target audience. It is, however, also useful for adults who regularly use social media tools like Twitter and are often subjected to trolling and abuse. High profile examples of internet trolling are frequent enough to show us that this kind of abuse is prevalent on social media.

Since the outbreak of Covid-19 and the start of lockdown, young people and adults have moved their lives further online, with screen time usage increasing dramatically and reports of cyberbullying on the increase. In April, the UK Government issued new online safety guidance for parents, urging them to follow the advice to keep children safe online during lockdown.

Comparitech.com surveyed over 1000 parents globally and found that between 2018 and 2020, one-fifth of all bullying occurred through social media and apps, with a further 11% occuring via text messages. Reports of cyberbullying increased by around 70% at the start of lockdown according to data from Israeli startup company, L1ght in Digital Trends, leaving younger people especially exposed to online harms.

Bullstop is unique in that it monitors a user's social media profile and scans for offensive incoming messages, to ensure the user is not subject to incoming abuse, as well as offensive outgoing messages, to ensure the user is not using inappropriate language and to provide a means of self-

reflection. This works via an artificial intelligence (AI) algorithm which is designed to understand written language: it analyses messages and flags offensive content, such as instances of cyberbullying, abusive, insulting or threatening language, pornography and spam.

Offensive messages can be immediately deleted from the user's inbox. A copy of deleted messages are, however, retained should the user wish to review them. The app can also automatically block contacts who continuously send offensive messages. Bullstop is highly configurable, allowing the user to determine how comprehensively the app removes inappropriate messages.

Ph.D. student Semiu Salawu who designed Bullstop said: "This application differs from other apps because the use of artificial intelligence to detect cyberbullying is unique in itself. Other anti-cyberbullying apps, in comparison, use keywords to detect instances of bullying, inappropriate or threatening language.

"The detection AI has been trained on over 60,000 tweets to recognise not only abusive and offensive language but also the use of subtle means such as sarcasm and exclusion to bully, which are otherwise difficult to detect using keywords.

"It uses a distributed cloud-based architecture that makes it possible for 'classifiers' to be swapped in and out. Therefore, as better artificial intelligence algorithms become available, they can be easily integrated to improve the app."

Bullstop currently supports Twitter with support for text messages planned in the next stage of the roll out. It is hoped that, with continued usage of the app and good results, other social media platforms such as Facebook and Instagram will come on board, allowing their users to benefit from the application.

Semiu added: "Twitter has been very supportive of research efforts such as this in allowing apps like BullStop to securely integrate with their platform and this has been key in completing the research. We hope that other social media platforms will follow suit."

The app is currently in the beta testing stage which means the researchers invite users of the app to provide them with feedback to allow them to make improvements. It has already been tested by a number of young people and professionals including teachers, police officers and psychologists.

A user aged 11 said: "I like that when I want to message people, it can check it for me because as much as I don't want to be hurt, I don't want to hurt people either. Also, if people don't want to get in trouble for sending something, they can use the app to tell them if what they are sending is bad."

A user aged 16 said: "I like that you can make contacts trusted or blocked and that it can automatically block people if they are being offensive."

A psychiatrist who reviewed the app said: "It has a 'stop and think' section, which for any child who does not directly want to bully or does not have the intention of bullying, is very positive. Also I don't think any child would install an app if they know it allows their parents access, so the decision not to include parental monitoring in the app is the right one".

A spokesperson for the West Midlands Police cybercrime unit said: "My initial impression was very good. As a user, or even as perhaps a guardian, I can go through the tour of the app and understand very quickly what the application is trying to achieve. I think the fact it does not have a parents' or companion app should be used a selling point. If

someone is being bullied online, I would definitely recommend this app to them and if this tool is publicly available, then I could see it would be something the police could recommend as a safeguarding tool."

A local general practitioner (GP) who tested the app added: "Now that I know BullStop will be available I can recommend it, because it's putting the advice we give out to patients and their parents into effect. For example we would advise they block offensive contacts and review connection requests. Having something like this gives a young person some measure of control back."

Bullstop is available for free and can now be downloaded on GooglePlay.

Provided by Aston University