RSiS S. RAJARATNAM SCHOOL OF INTERNATIONAL STUDIES
Nanyang Technological University, Singapore

RSIS COMMENTARY

www.rsis.edu.sg No. 078 – 13 June 2024

Curbing Falsehoods on WhatsApp and Telegram

By Shantanu Sharma and Ken Chen

SYNOPSIS

Small, close-knit groups inside private messaging applications like WhatsApp and Telegram have become a hotbed of misinformation which is difficult to track, monitor, and regulate. CheckMate, a community-driven initiative, shows promise in combating misinformation using a bottom-up approach.

COMMENTARY

In January 2024, the Singapore government launched a S$20 million initiative to develop tools to detect and counter misinformation, deepfakes, and targeted scams, which are poised to proliferate with the growth of generative AI. Despite the significant investment, various other government solutions, and large toolset to counter them, widespread misinformation and disinformation on private messaging platforms remains a challenge.

"New" Media Landscape and Outcomes

Global news consumption patterns have changed over the last decade. Major shocks, like the coronavirus pandemic, have accelerated structural shifts, not only transforming social media into platforms for news consumption but also altering newsroom revenue models.

A heavy emphasis on "views" to increase profits has led to the use of algorithms to drive up engagement. This, in turn, has led to the amplification of catchy, dramatic, and negative news and even the promotion of misinformation, hate speech, and biases.

In this era of heightened polarisation and saturated feeds, public interest in accessing, sharing, and participatingin mains tream news on social media platforms has steadily declined. This is because online debates and news are increasingly perceived to be toxic. Many are withdrawing from online discourse to avoid emotional distress and “saturation overload” which can lead to exhaustion, desensitisation, and passiveness.

Amidst declining news consumption on social media, a growing section of the population is preferring to share alternatives (as opposed to mainstream news) with smaller, like-minded communities such as friends and family groups, on private messaging platforms (PMPs) like WhatsApp and Telegram.

This phenomenon was exacerbated during the coronavirus pandemic. This shift also resulted in close-knit online groups becoming a hotbed of misinformation. The issue first came to light when a 65-year-old woman was hospitalised after taking ivermectin, having been influenced by false information spread on Telegram and WhatsApp. There have been more similar instances to date.

Challenge of Private Messaging Platforms

Misinformation can easily go unchallenged on PMPs. Many often self-censor and avoid engaging with the spreader or the discourse surrounding false information to avoid conflict in close-knit groups. Studies show that people are more likely to believe misinformation sent by friends and family.

Although PMPs have utilised some countermeasures to curb the spread of false information, including limiting the number of forwards on WhatsApp, this only delays the spread and is ineffective in blocking the propagation. Since PMPs are end-to-end encrypted and closed, it is almost impossible to monitor, regulate, track, and build relevant strategies to counter the wide dissemination of misinformation and disinformation.

Hence, as observed in the case of misinformation about the XBB strain of the SARS-CoV2 virus, current available solutions like POFMA – a government legislation aimed at preventing the electronic communication of misinformation and disinformation, and to safeguard against the use of online platforms for communication of falsehoods and information manipulation – are deemed ineffective and not practical for countering falsehoods on PMPs. Moreover, expert fact-checkers take longer to issue corrections and have limited coverage. However, there are bottom-up fact-checking methods that can mitigate the risks of misinformation spread in PMPs. These methods should be added to the current toolbox being used to fight falsehoods.

An Alternative Solution: CheckMate

Fact-checking is one of the most effective means of changing false beliefs. In public online platforms, the burden of fact-checking and policing content is shared between the community consisting of volunteers, tech platforms, fact-checking organisations, and government institutions. Since the messages in PMPs are hidden from third-party observers, the burden of fact-checking and engaging with falsehoods falls to users in the private group.

CheckMate, a community-driven initiative, shows promise in combating misinformation using a bottom-up approach. Built and run by volunteers, CheckMate allows anyone to forward messages, images or screenshots that require clarification to a WhatsApp number. Behind the scenes, volunteer checkers assess what is sent, and vote on it promptly to issue a quick response. This responsiveness, and ability to provide verification "on-demand", is critical in filling gaps left by official communications. Additionally, CheckMate leverages machine learning techniques like natural language processing and generative AI to respond to users' messages quickly.

CheckMate relies on crowdsourcing fact-checking by multiple human checkers which diminish bias and strengthens the credibility of fact-checking processes as the diversity of fact-checkers minimises the influence of subjective interpretations or personal agendas on the outcomes. This also allows CheckMate to address a large variety of queries such as climate, health and CPF-related misinformation, scams, and illicit messages.

CheckMate also employs generative AI that can provide detailed explanations when highlighting giveaways in a scam message, thereby educating users on how to spot fraudulent content. The use of generative AI for debunking purposes has shown promising early results in academic research.

Studies indicate that people are more inclined to ignore fake news than actively debunk or seek clarification, possibly due to fear of embarrassment. CheckMate cultivates a safe environment where users can feel empowered to engage and address falsehoods. Such initiatives help surmount psychological barriers to engagement and enhance the overall ecosystem.

Conclusion

There is no silver bullet to solve the spread of falsehoods, but its risks can be mitigated. Public communication, pre-bunking, social media literacy, and laws such as POFMA are part of the broad arsenal of countermeasures to address this challenge. Alternate bottom-up solutions like CheckMate can complement these countermeasures by providing a within-platform solution for users to factcheck information received and shared on PMPs. This is important in an era of AI-enabled deepfakes that make falsehoods more targeted and harder to detect.

Shantanu Sharma is a Senior Analyst with the Centre of Excellence for National Security (CENS) at S. Rajaratnam School of International Studies (RSIS), Nanyang Technological University (NTU), Singapore. Ken Chen is a Fact Checker at CheckMate, BetterSG.