Reinforcement Learning-based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation

Bing He
Georgia Institute of Technology
Atlanta, Georgia, USA
bhe46@gatech.edu

Mustaque Ahamad
Georgia Institute of Technology
Atlanta, Georgia, USA
mustaq@cc.gatech.edu

Srijan Kumar
Georgia Institute of Technology
Atlanta, Georgia, USA
srijan@gatech.edu

ABSTRACT

The spread of online misinformation threatens public health, democracy, and the broader society. While professional fact-checkers form the first line of defense by fact-checking popular false claims, they do not engage directly in conversations with misinformation spreaders. On the other hand, non-expert ordinary users act as eyes-on-the-ground who proactively counter misinformation – recent research has shown that 96% counter-misinformation responses are made by ordinary users. However, research also found that 2/3 times, these responses are rude and lack evidence. This work seeks to create a counter-misinformation response generation model to empower users to effectively correct misinformation. This objective is challenging due to the absence of datasets containing ground-truth of ideal counter-misinformation responses, and the lack of models that can generate responses backed by communication theories. In this work, we create two novel datasets of misinformation and counter-misinformation response pairs from in-the-wild social media and crowdsourcing from college-educated students. We annotate the collected data to distinguish poor from ideal responses that are factual, polite, and refute misinformation. We propose MisinfoCorrect, a reinforcement learning-based framework that learns to generate counter-misinformation responses for an input misinformation post. The model rewards the generator to increase the politeness, factuality, and refutation attitude while retaining text fluency and relevancy. Quantitative and qualitative evaluation shows that our model outperforms several baselines by generating high-quality counter-responses. This work illustrates the promise of generative text models for social good – here, to help create a safe and reliable information ecosystem. The code and data is accessible on https://github.com/claws-lab/MisinfoCorrect.

CCS CONCEPTS

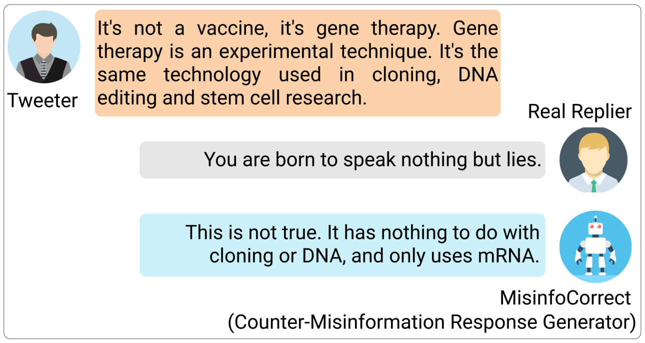
• **Computing methodologies** → **Natural language generation**; **Reinforcement learning**.

KEYWORDS

misinformation, reinforcement learning, text generation

WWW '23, April 30–May 04, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583388

ACM Reference Format:
Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement Learning-based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543507.3583388



It's not a vaccine, it's gene therapy. Gene therapy is an experimental technique. It's the same technology used in cloning, DNA editing and stem cell research.

You are born to speak nothing but lies.

This is not true. It has nothing to do with cloning or DNA, and only uses mRNA.

Figure 1: An overview of counter-misinformation response generation task.

1 INTRODUCTION

Online misinformation reduces trust in vaccines and health policies [7, 45, 67], leads to violence and harassment [6, 83], questions democratic processes and elections [79–81], increases polarization [85], and harms well-being [96]. Most people receive information and news from social media [107], which is often "ground-zero" for health misinformation and where misinformation spreads faster and farther than truth [45, 100]. COVID-19 vaccine misinformation, including false claims that the vaccine causes infertility, contains microchips and even changes DNA and genes has fueled vaccine hesitancy, reduced vaccine uptake, and prolonged the pandemic. Besides, misinformation also causes harms to people directly. For example, misinformation that Bill Gates creates vaccines to depopulate people led to distrust and verbal attacks [25]. Thus, it is critical to curb the spread of online misinformation [13, 28, 46, 49, 57, 111, 119]. In this work, we use a broad definition of misinformation which includes falsehoods, inaccuracies, rumors, decontextualized truths, or misleading leaps of logic [114].

Professional fact-checkers and journalists provide objective fact-checks for viral claims and release their determination on their website, which are incredibly useful to create detection models. However, fact-checkers do not actively engage with misinformation