Corrective or Backfire: Characterizing and Predicting User Response to Social Correction

Bing He, Yingchen Ma, Mustaque Ahamad, Srijan Kumar
Georgia Institute of Technology
Atlanta, Georgia, USA
bhe46@gatech.edu,yma473@gatech.edu,mustaq@cc.gatech.edu,srijan@gatech.edu

ABSTRACT

Online misinformation poses a global risk with harmful implications for society. Ordinary social media users are known to actively *reply* to misinformation posts with counter-misinformation messages, which is shown to be effective in containing the spread of misinformation. Such a practice is defined as "*social correction*". Nevertheless, it remains unknown how users respond to social correction in real-world scenarios, especially, will it have a corrective or backfire effect on users. Investigating this research question is pivotal for developing and refining strategies that maximize the efficacy of social correction initiatives.

To fill this gap, we conduct an in-depth study to characterize and predict the user response to social correction in a data-driven manner through the lens of X (Formerly Twitter), where the user response is instantiated as the reply that is written toward a counter-misinformation message. Particularly, we first create a novel dataset with 55, 549 triples of misinformation tweets, counter-misinformation replies, and responses to counter-misinformation replies, and then curate a taxonomy to illustrate different kinds of user responses. Next, fine-grained statistical analysis of reply linguistic and engagement features as well as repliers' user attributes is conducted to illustrate the characteristics that are significant in determining whether a reply will have a corrective or backfire effect. Finally, we build a user response prediction model to identify whether a social correction will be corrective, neutral, or have a backfire effect, which achieves a promising F1 score of 0.816. Our work enables stakeholders to monitor and predict user responses effectively, thus guiding the use of social correction to maximize their corrective impact and minimize backfire effects. The code and data is accessible on https://github.com/claws-lab/response-to-social-correction.

CCS CONCEPTS

• **Information systems** → Social networks.

KEYWORDS

Misinformation, Counter-misinformation, Social Correction

ACM Reference Format:
Bing He, Yingchen Ma, Mustaque Ahamad, Srijan Kumar. 2024. Corrective or Backfire: Characterizing and Predicting User Response to Social Correction. In *ACM Web Science Conference (Websci '24), May 21–24, 2024, Stuttgart, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3614419.3644004

1 INTRODUCTION

Online misinformation undermines public health by diminishing trust in vaccines and health policies [4, 33, 46], and has been linked to reduced COVID-19 vaccine uptake [37]. Its impact also extends to inciting violence [3, 51], and negatively affecting well-being [57]. This situation is exacerbated because misinformation typically spreads more rapidly and widely than factual information on online social media platforms [33, 60], making it imperative to curb the spread of misinformation [12, 22, 34, 35, 39, 71, 78].

To combat misinformation, professional fact-checkers and journalists provide valuable objective fact-checks to debunk misinformation [59]. However, their engagement with users is limited [39]. In contrast, ordinary social media users play a proactive role in combating misinformation through their active engagement including their replies, comments, and posts that counter misinformation posted by others [7, 39, 51, 55, 58, 77]. It finally complements the efforts of professionals [2, 26, 31], even accounting for 96% of online counter-misinformation messages [39].

Significantly, recent studies underscore the "*social correction*" [37, 41] - the practice where ordinary users combat misinformation claims in a *conversational* manner by their direct counter misinformation *replies* to misinformation posts - which has shown to be as effective as professional correction, curbing misinformation spread across diverse topics, platforms, and demographics [6–8, 15, 20, 37, 50, 61–64, 64, 65, 72]. One example of social correction is shown in Figure 1.

Nevertheless, little is known about the real-world user response toward social correction. Understanding such responses is beneficial because i) They serve as a critical signal to indicate the impact of social correction in real-world scenarios. If some social corrections are revealed to have corrective effects (e.g., users disbelieve in misinformation) [13], then additional participants can be encouraged to provide reinforcements; ii) Instead, If certain social corrections are found to increase users' beliefs in misinformation (e.g., backfire) [53], targeted efforts can be directed toward improving them. Such instances can be escalated and prioritized for interventions by professionals or social media platforms; iii) Responses can also indicate whether users are entrenched in (counter-)misinformation echo chambers [17], where their beliefs are reinforced by similar viewpoints, or if there is a cross-pollination of ideas. This contributes to understanding polarization around certain topics.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Websci '24, May 21–24, 2024, Stuttgart, Germany
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0334-8/24/05.
https://doi.org/10.1145/3614419.3644004