Improved Twitter Design Proposal

Khai Brisco, Aileen Kuang, Celestine Le, Efrain Moreno, Sanjana Pavani

---

PROBLEM

Today, many people rely on Twitter to quickly get information. Unlike traditional news outlets, where civilians are passive recipients of information, Twitter allows civilians to share their own information. This allows for easy access to various and diverse perspectives, but it also increases the spread of false information, which can have great consequences for public safety and democratic integrity – especially when it is spread by highly influential individuals (celebrities, elected officials, or other "verified" users). For instance, incorrect information about COVID-19 that was spread by elected officials likely caused unnecessary deaths during the height of the pandemic (Gisondi et al., 2022). In addition, election deniers in the U.S. have caused citizens to lose trust in American democracy (Middlemass, Rodriguez, & Sanchez, 2022). While Twitter has implemented some features to combat misinformation (Twitter, n.d., "Misinformation on Twitter"), we believe that there are fundamental issues with their approach.

First, Twitter's user interface makes it difficult for users to report misleading information. To report false information, users must go through multiple steps, some of which are irrelevant to misinformation. Such an inconvenient and time-consuming process may discourage users from reporting inaccurate information at all. This feature is also not immediately visible – some users might not know that it exists at all, and even if they do, they might neglect to report a Tweet because of the feature's obscured placement. Consequently, misinformation can continue to spread and potentially harm others. By increasing the visibility of this feature and shortening its process, Twitter can better address misinformation and avoid its real-world repercussions.

Second, Twitter's approach towards misinformation is passive. It only checks for misinformation as it is being spread, which does not address the crux of the issue: that incorrect information is being spread at all. Civilian users are not encouraged to be mindful of the high-profile users they interact with; therefore, they may inadvertently continue to interact with celebrities or officials who often post misleading Tweets. A function that prompts civilian users to evaluate their continued interaction with certain high-profile users may help reduce the spread of misinformation.

Third, Twitter's fact-checking process is slow and inefficient. Users might see a newly published Tweet or trending topic and interact with it, not knowing whether it contains accurate information because of the speed of Twitter's verification process. This means that both the spread and speed of misinformation are going unaddressed. To solve this issue, Twitter could further specify the verification status of a Tweet (ex. verified, unverified, missing context) and enhance its visibility, which would increase users' awareness about their engagement with non-verified content. Moreover, allowing users to see report analytics for a

Tweet (or the number of times a Tweet has been reported for containing misleading information) could provide users with a measure of how trustworthy a Tweet is if it has not yet been verified.

In the following sections, we propose a revised version of Twitter that will better address misinformation. We further outline the potential benefits and harms of our application, as well as policies that affect our system's functionalities and architecture.

FUNCTIONALITY

In our revised version of Twitter, we suggest changes that affect the following: the fact-checking (or verification) process, user interface for reporting misinformation, and user mindfulness concerning content consumption.

First, our version of Twitter will automatically label Tweets with one of the following: "Unchecked" (not yet reviewed by the curation team), "Unsupported" (not supported by other sources), or "Verified" (supported by other sources). Twitter currently adds labels to Tweets that potentially contain misleading information; these labels have three categories: "Misleading information," "Disputed claims," and "Unverified claims" (Thorebeck, 2020). Furthermore, Twitter has a pilot program called Birdwatch, in which users volunteer to select Tweets that need to be debunked or provided with more context (Lorenz, Oremus, & Merrill, 2022). However, in the early stages of a Tweet (for instance, right after it is posted), there is no label to help users determine whether it is accurate or inaccurate, nor has it been evaluated by Birdwatch. Allowing users to see the verification status of a Tweet will be especially beneficial in the early stages of a Tweet, as it will increase user mindfulness and engagement.

Second, the revised application will increase the visibility of the process for reporting misleading information (hereafter **Misinformation Reports**) and shorten it. Right now, Misinformation Reports is lengthy and difficult to access, which discourages users from utilizing it. To address this, we will relocate the feature to the top-right corner of a Tweet and streamline it (we have reduced the number of questions in the report submission). Additionally, to prevent botting and spam, we will have a captcha system in place. By improving the user interface for reporting misinformation, we will improve user engagement with misinformation – users will likely report false information at higher rates, which then quickly alerts Twitter about misleading content.

Furthermore, after reporting a Tweet, users will be able to view a graph showing **Report Analytics**, or the number of times a Tweet has been flagged for misinformation. So, even when a Tweet has not yet been verified by the Twitter curation team, users can see whether other users believe that the Tweet is misleading. This will help mitigate the overall spread of misinformation, as this feature will further increase users' awareness of potentially false information.

While Twitter mitigates misinformation as it spreads, our final improvement will address misleading content from the very beginning. By clicking a button on a verified users' (whose Tweets have the greatest reach) profile, civilian users can access information pertaining to their behavior on the platform.

This includes the number of times the user has been fact-checked, suspended, or violated community guidelines. By providing civilians with access to such information, we will encourage user thoughtfulness surrounding Tweet consumption, as users can better determine whether they want to have continued exposure to verified users.

ARCHITECTURE

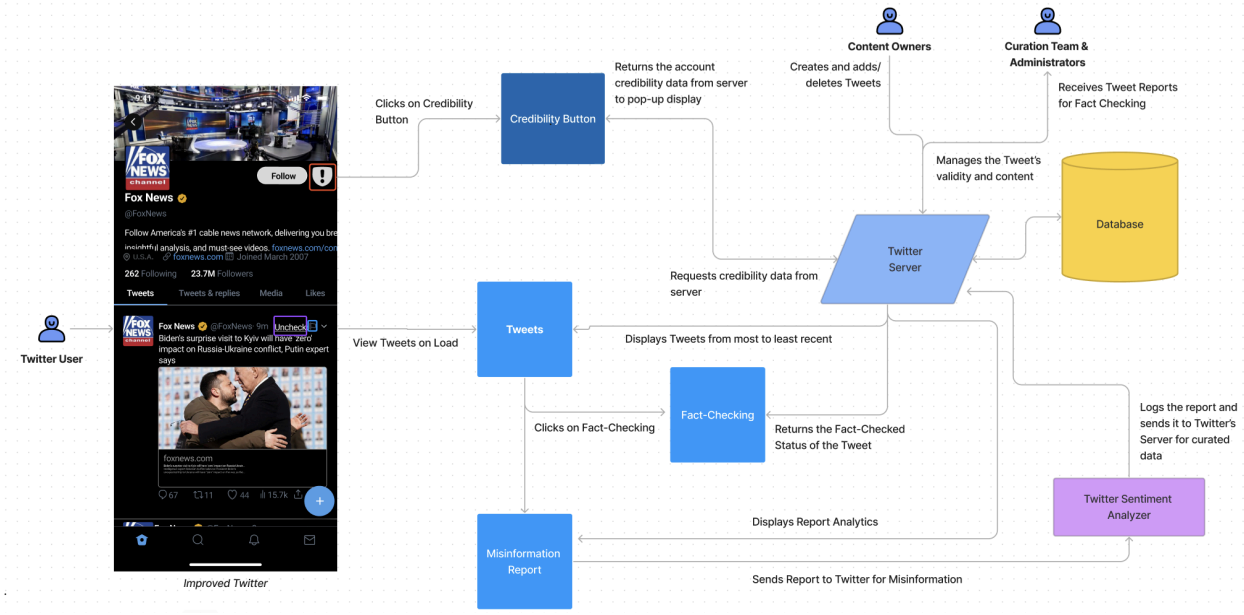

Figure 1. Architecture of improved Twitter design.

This figure depicts the architecture of Twitter's Information System. The architecture features some key groups:

- **Twitter App** is a social media app that is available on PC, Android and iOS. It allows accounts to post, retweet and interact with content on the platform. It receives personalized results from the Tweet Server and Sentiment Analyzer on the Twitter database, and sends the user's Tweets and post interactions to the respective servers.
- **Accounts** who post and interact with other users in the community through the app. Users can contribute to Twitter by posting, retweeting, liking, commenting and reporting Tweets.
- **Report,** which allows users to report posts for breaking community guidelines and other issues.
- **Tweets,** the statuses and news that users on Twitter see and interact with. Tweet interactions are recorded and personalized using a sentiment analyzer algorithm (see point below).
- **Sentiment Analyzer** is the algorithm that Twitter uses to gauge user interaction on specific Tweets and accounts to provide them with more similar content.
- **Twitter Database** is the database where Tweets, accounts, and all other information is stored on Twitter.

- **Twitter Server** gets information from the Twitter Sentiment Analyzer, Twitter Database, Content Owners, Misinformation Reports, and the Curation Team & Administrators and relays information to the app through API's.
- **Content Owners** that can add and delete Tweets.
- **Curation Team & Administrators** that receive Tweet reports for fact checking and manage the Tweet's validity and content.

The figures also include improvements to the current Twitter System:

- **Misinformation Reports** (top right of the Tweet highlighted in blue) allow users to submit faster misinformation reports to the Twitter curation team.
    - **Report Analytics** displays a bar chart that shows the total number of reports the Tweet received and what categories the Tweet was reported for. Users are prompted to view **Report Analytics** after making a misinformation report.
- **Credibility Button** (highlighted in red) for Twitter accounts that display how credible they are as a source of information through an algorithm that analyzes community guideline violations, Retweets and Tweets.
- **Fact Checking** (highlighted in purple) that Twitter automatically adds to posts to display if the curation team has checked the post yet and if the post is verified or needs more context.

BENEFITS

Our improvements hinder the spread of misinformation, which will mainly benefit the public:

- **Increased user autonomy.** When platforms take down posts spreading misinformation, users may be angry or confused as to why those posts were taken down. Even a short description might not make a convincing case. Civilian users can be empowered to make decisions about the information they trust and share their reasoning through reports, decentralizing information on our platform.
- **Enforced accountability.** Influencers with large platforms will be pushed to think twice before posting misleading information and face public consequences for continuously failing to meet community guidelines. Civilian users can be confident that even highly-influential users are still subjected to our policies.
- **Encourages political participation.** Users will partake in more community-based fact-checking, enhancing engagement with political information. Misinformation that disenfranchises or discourages civic participation will be less influential, equipping users with verified political resources.
- **Public negative engagement betters digital democracy.** By making **Report Analytics** public, we give civilian users another forum to express dissent. Positive engagement alone may skew public opinion, and public negative engagement can be used to counteract that, creating a diverse

distribution of thought for civilian users, celebrities, government officials, and news organizations to be civically engaged with.

HARMS

Yet still, our revisions may have their downsides:

- **Less engagement due to presumed censorship.** Efforts to curb misinformation are often met with accusations of censorship or information control. Users may feel discouraged from voicing opinions or spreading important information. The result may be a feedback effect: users that feel ostracized by efforts to correct misinformation may engage less with the community.
- **Lower morale due to negative engagement.** Many social media platforms have done away with visual counts of negative engagement (i.e. YouTube's removal of public dislikes). Negative engagement could be especially discouraging to smaller creators or influencers, potentially hindering community engagement among smaller influencers. Civilian users may also interpret some of our improvements as a form of social shaming.
- **Ideological skew on analytics.** Despite filtering for bots via captcha, our analytics may still appear skewed due to overwhelming attention from ideological groups. This may render some of our quantitative analysis non-representative of public opinion, which may mislead both civilian and high-profile users.
- **Discriminatory bias.** Marginalized groups that are already vulnerable to discrimination based on race, gender, ethnicity, sexuality, disability, religion, or national origin may be further subjugated by negative engagement. Received reports may be biased towards vulnerable groups or underrepresented dominant groups, inflating report analytics for certain demographics over others. Marginalized online communities may struggle to remain intact if reports are actively weaponized against them, especially if influencers incite their audience to spam.

POLICIES

Twitter is impacted by Europe's GDPR, which is the strictest set of data privacy laws in the world. More specifically, Article 12 of the GDPR articulates that platforms must "make it easy for people to make requests [concerning their data] to you (e.g., a right to erasure request, etc.) and respond to those requests quickly and adequately (Wolford, 2019; see also GDPR.eu, "Art. 12 GDPR"). Furthermore, Article 15 states that users have a right to access their personal data (Wolford, 2019; see also GDPR.eu, "Art. 15 GDPR"). Twitter adheres to these policies, as it allows users to request and download a copy of their data through a Privacy Form and a service tool called Your Twitter Data (Twitter, n.d., "FAQ on GDPR").

Our revised version of Twitter includes a feature that displays user analytics, or information pertaining to verified users' behavior on the platform (see **Functionality** section), and verified users will also be able to request and download that data through the previously mentioned services. Therefore, our application fulfills Articles 12 and 15, as verified user analytics will be easily accessible to concerned parties.

References

Belli, L. (2021, October 21). *Explaining algorithmic amplification of political content on Twitter.* Twitter Blog. Retrieved March 1, 2023, from https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent.

GDPR.eu. (n.d.). *Art. 12 GDPR – Transparent information, communication and modalities for the exercise of the rights of the data subject.* Retrieved March 2, 2023 from https://gdpr.eu/article-12-how-controllers-should-provide-personal-data-to-the-subject/?cn-reloaded=1.

GDPR.eu. (n.d.). *Art. 15 GDPR – Right of access by the data subject.* Retrieved March 2, 2023 from https://gdpr.eu/article-15-right-of-access/.

Gisondi, M., Barber, R., Faust, J., Raja, A., Strehlow, M., Westafer, L., & Gottlieb, M. (2022). A Deadly Infodemic: Social Media and the Power of COVID-19 Misinformation. *Journal of Medical Internet Research*, 24(2). Retrieved March 1, 2023, from https://www.jmir.org/2022/2/e35552.

Lorenz, T., Oremus, W., & Merrill, J. (2022, November 9). *How Twitter's new contentious new fact-checking project really works*. The Washington Post. Retrieved March 3, 2023, from https://www.washingtonpost.com/technology/2022/11/09/twitter-birdwatch-factcheck-musk-misinfo/.

Middlemass, K., Rodriguez, A., & Sanchez, G. (July 26, 2022). Misinformation is eroding the public's confidence in democracy. *Brookings*. https://www.brookings.edu/blog/fixgov/2022/07/26/misinformation-is-eroding-the-publics-confidence-in-democracy/.

Thorbecke, C. (2020, May 27). *What to know about Twitter's fact-checking labels*. ABC News. Retrieved February 28, 2023, from https://abcnews.go.com/Business/twitters-fact-checking-labels/story?id=70903715.

Twitter. (n.d.). *FAQ on GDPR.* GDPR Twitter. Retrieved March 2, 2023, from https://gdpr.twitter.com/en/faq.html.

Twitter. (n.d.). *Healthy Conversations*. About Twitter. Retrieved February 28, 2023, from https://about.twitter.com/en/our-priorities/healthy-conversations.

Twitter. (n.d.). *How we address misinformation on Twitter*. Twitter Help. Retrieved February 28, 2023, from https://help.twitter.com/en/resources/addressing-misleading-info.

Wolford, B. (2019, February 22). *A Guide to GDPR Data Privacy Requirements*. GDPR.eu. Retrieved March 2, 2023 from https://gdpr.eu/data-privacy/.

Zhang, L., Malife, C. (2021, October 22). *Processing billions of events in real time at Twitter.* Twitter Blog. Retrieved March 1, 2023, from https://blog.twitter.com/engineering/en_us/topics/infrastructure/2021/processing-billions-of-events-in-real-time-at-twitter-.