How can crowdsourcing rescue the social media?

Keywords: Fact-checking, Birdwatch, Social Media, Collective Intelligence, Collaboration

Extended Abstract

How is it possible that some social web technology such as Wikipedia stand as the most successful example of collaborative and healthy information sharing, while the others, such as Twitter are blamed for epistemic chaos?

In this work we a) examine data generated in Twitter's Birdwatch (Community Notes) platforms and b) report on our CRT experiments where we test a few hypotheses on how collaborative fact-checking systems should be designed.

In the current Birdwatch implementation, a member of the group of reviewers (selected by Twitter based on undisclosed criteria) can add a note to a tweet that they find "misinformed or potentially misleading." A note provides some information selected from predefined values about the tweet (misleading factual error, misleading satire) as well as some free text where the reviewer can comment and link to external sources. Then other reviewers express their agreement or disagreement with the existing notes through additional annotations such as helpfulness and informativeness. Ultimately, notes produced by reviewers will become visible next to the corresponding tweets based on the support/opposition they have received from other reviewers.

We analyzed Birdwatch helpfulness ratings as of February 2022 — 189,744 ratings of 17,888 notes by 7,884 reviewers. We observed evidence of a highly balanced network with two well-separated clusters where reviewers agree with those in the same group and disagree with those in the opposite group. In fact, of the pairs of reviewers with reciprocal ratings, 71% are consistent in that they both rate each other helpful (in the same cluster) or not helpful (in different clusters).

Furthermore, despite 35% negative ratings, we found that only 22% of triads of reviewers with reciprocal ratings are structurally imbalanced, i.e., inconsistent with all three reviewers agreeing with each other (in the same cluster) or with two reviewers in agreement with each other (in the same cluster) and in disagreement with the third (in the other cluster).

Figure 1 offers visual confirmation of these findings by mapping the network of Birdwatch reviewers. An edge between two nodes represents reciprocal ratings that indicate agreement on average. The polarization among Birdwatch reviewers mimics the one observed among generic Twitter users. It is unlikely that this polarization is a reflection of objective arguments; rather, it merely represents the political affiliations of the reviewers. It is unlikely that this polarization is a reflection of objective arguments; rather, it merely represents the political affiliations of the reviewers. Analysis of the notes confirms that users systematically reject content from those with whom they disagree politically [1]. One might argue that the population of Birdwatch reviewers is less homogenous than that of Wikipedia editors. While this may be true, a

polarized crowd can be even more effective in producing high-quality content compared with a homogenous team [2].

In the next steps, in a series of experiments we analysed the partisan behaviour of fact-checkers recruited from the general public for our experiments. In the experiment, similar to Birdwatch, participants have to write notes to contextualise tweets with strong republican or democrat bias posted by key political figures. However, here the participants work in teams that are either made of homogeneous members (both republican or both democrat), or a mixed group with one member from each main political party. The preliminary analysis of the outcome of the experiment shows a significant difference in behaviour between groups with different political ideologies.

We use the following metrics to measure the quality of the notes:

- Length of the note: an integer representing the number of words in the note
- The number of links provided in the note as a sources
- Flesch Reading Ease Score: an integer between 0-100, the higher the number the simpler the text [3].
- Standard score: a score based on multiple formulas that calculate the sophistication of a text. The number corresponds to the US school grade level required to understand the text [4].
- The sentiment: a number between -1 and 1 that shows how positive or negative the text is [5].
- Stats: a Boolean variable that shows whether or not the text contains any numbers, like a date, statistics, etc.

Our main finding is that the notes collaboratively written on tweets that are primarily expressing republican views have similar features regardless of note-writer group compositions, whereas notes that are written to contextualize democratic leaning tweets, depending on the political composition of the team who wrote the note, can vary in length, tone, reliance on statistics and external sources, all pointing to the direction that fully republican teams provide much less nuance in their notes and fully democrat teams provide the most details among all the experiment conditions (see Figure 2 for an example which shows the distribution of the average number of URLs used in different experimental confitions).

In conclusions, here we show the political bias that exist among Twitter's Birdwatchers and moreover through controlled experiments show how the political bias can work differently in different conditions.

References

1. Allen, Jennifer N. L., Cameron Martel, and David G. Rand. 2021. Birds of a Feather Don't Fact-check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-checking Program. Preprint PsyArXiv.
2. Shi, F. et al. 2019. The wisdom of polarized crowds. Nature Human Behaviour. 3, 4 (Mar. 2019), 329–336.
3. Flesch R. HOW TO WRITE PLAIN ENGLISH: A BOOK FOR LAWYERS AND CONSUMERS.
4. Bansal, S. (2014) *Textstat, PyPI*. Available at: https://pypi.org/project/textstat/ (Accessed: March 1, 2023).

5. Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. InProceedings of the international AAAI conference on web and social media 2014 May 16 (Vol. 8, No. 1, pp. 216-225).

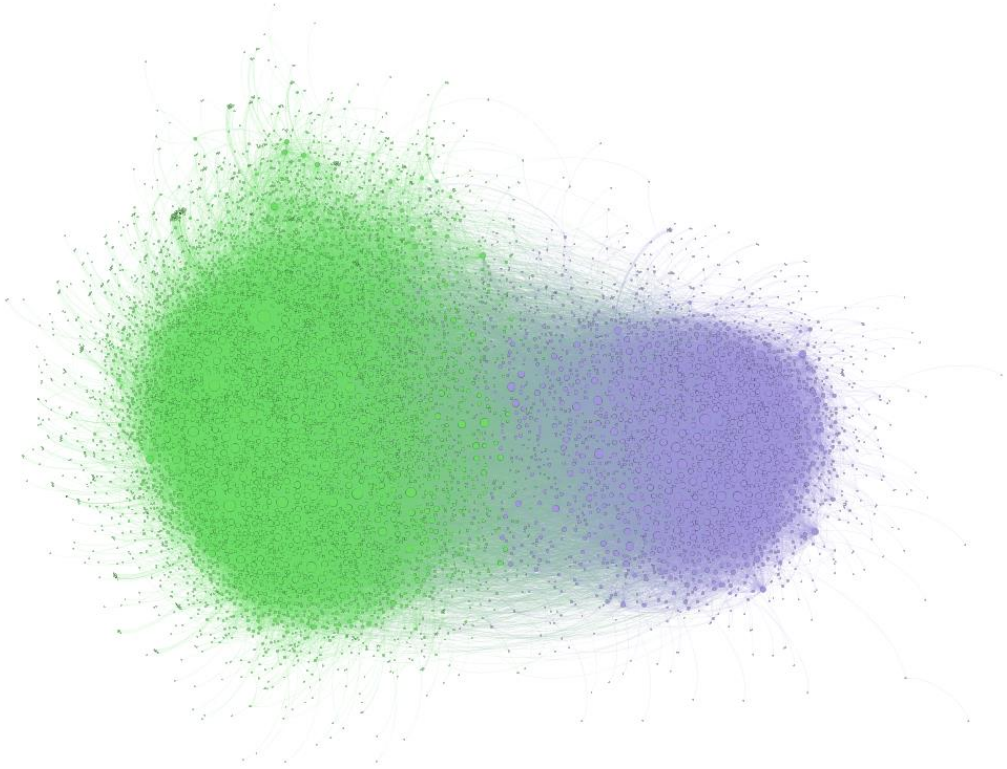

Figure 1 The network structure of Birdwatch users and their positive ratings of each other's notes. Node colours are determined by a community detection algorithm and node size indicates the number of interactions.

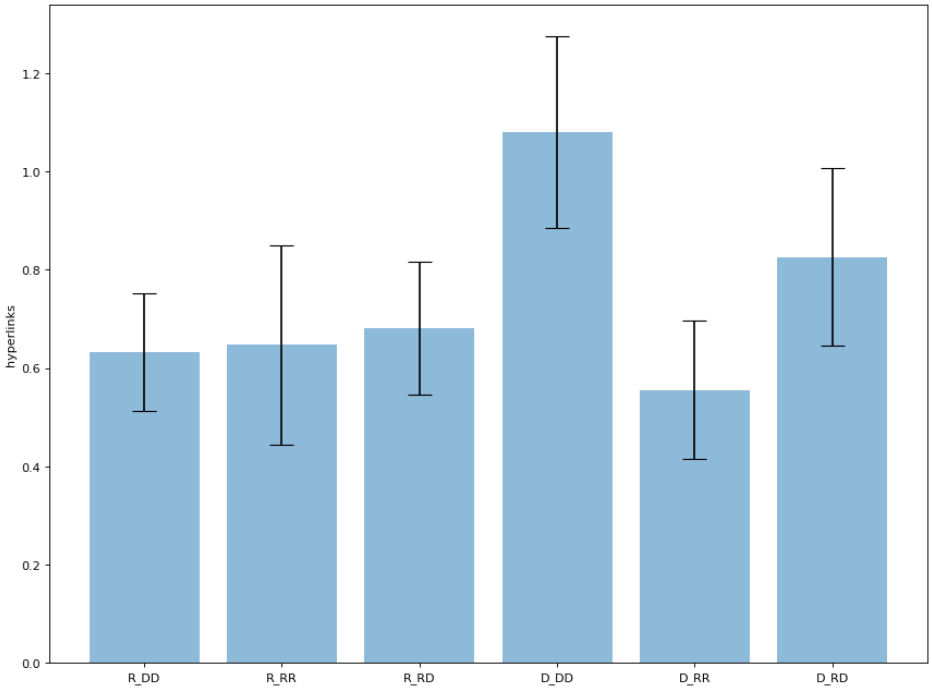

Figure 2 The average number of hyperlinks used in the notes written on tweets supporting republican or Democrat ideologies (R or D) aby teams with a composition of XX. Notes written by fully democrat teams on Democrat Tweets have the largest number of links.