Misleading Tweets and Helpful Notes: Investigating Data Labor by Twitter Birdwatch Users

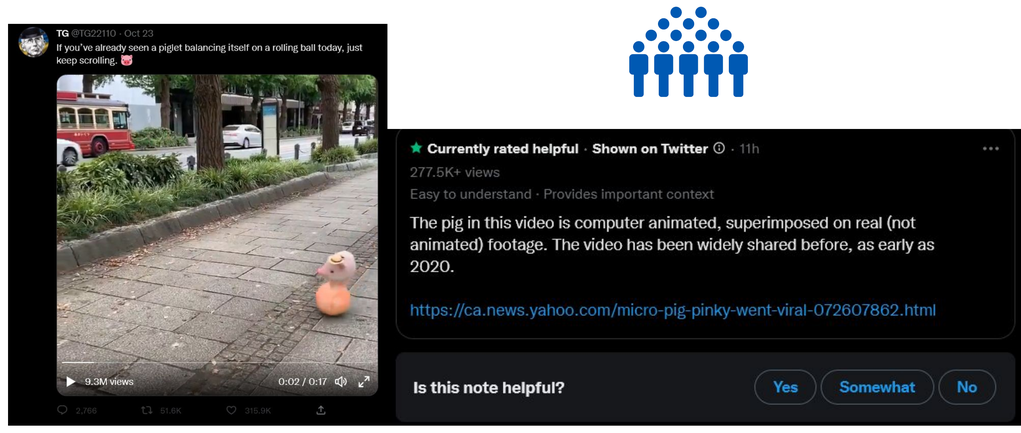
Isaiah Jones, Brent Hecht, Nicholas Vincent
PSA Research Group, Northwestern University

Abstract

In response to concerns about misleading content on social media, Twitter launched the ``Birdwatch" initiative that allows volunteers to label and add context to tweets. We study data from Birdwatch to understand how users are performing ``data labor" for Twitter, with implications for other platforms that are similarly reliant on data labor. We conduct computational analyses of Birdwatch text data and perform machine learning experiments to see how Birdwatch contributions might be used for classification. We find that Birdwatch users discuss distinct topics in domains like politics and news. While using Birdwatch data for content-only predictions may provide only a small amount of predictive power, in some cases Birdwatch data may be able to support ML systems. Furthermore, we see that the continuous flow of Birdwatch contributions provides great value in terms of supporting a ``guess most frequent`` baseline for classifying Twitter content.

About Birdwatch

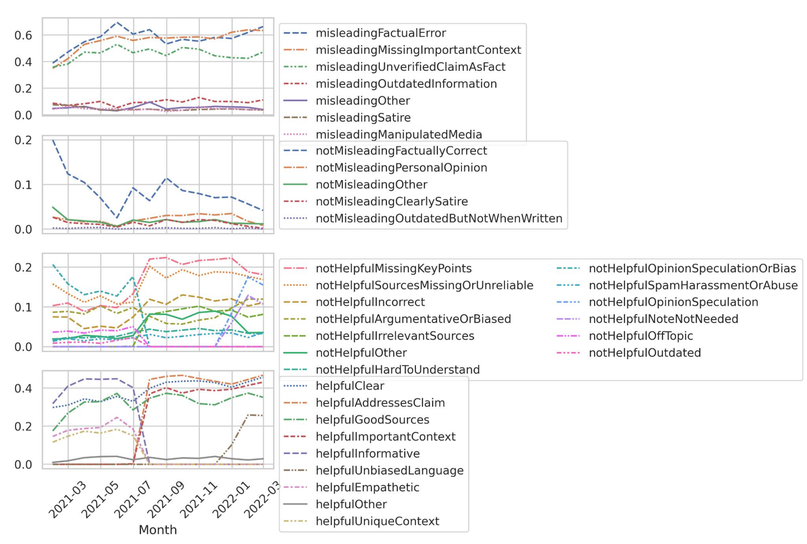
- In early 2021, Birdwatch launched as a pilot program for user-driven content moderation available to limited # of selected users.
- Only a small set of volunteer participants involved
- Since October 6th 2022, all Twitter users in the US are able to see and rate Birdwatch notes.
- Birdwatch has important implications for the relationships between volunteer data laborers and major tech platforms as it demonstrates a move towards collaborative effort rather than an opaque enforcement of sitewide rules.



Labels and Topics

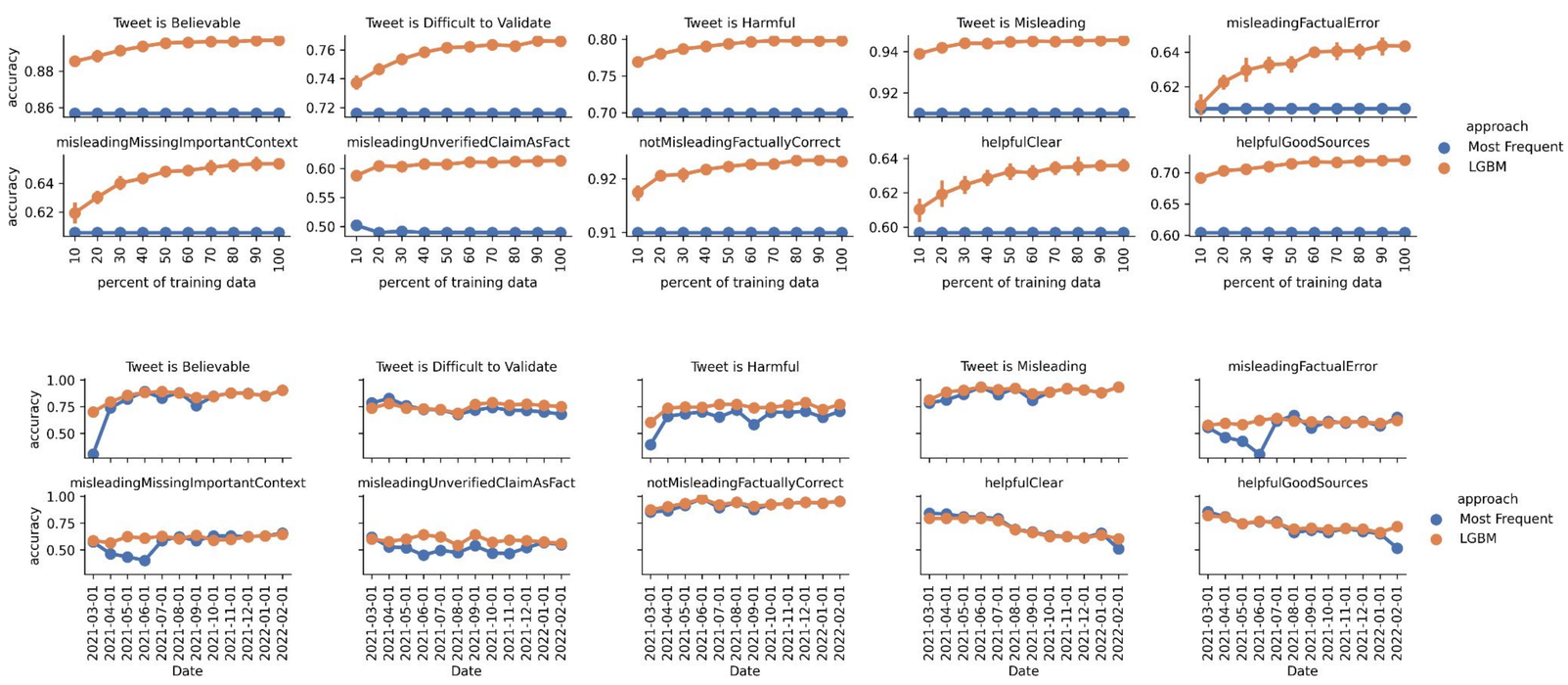
Important topics with many volunteer-provided labels:

- Topic modeling shows dominance of politics (e.g., US political figures), controversy (e.g. COVID-19 vaccination), and news (e.g. earthquakes).
- Many labels are available in Birdwatch, and their frequencies vary over time (shown to the right).



Machine Learning with Birdwatch Data

- Using Note/tweet text as input, we predict if Birdwatchers will assign a Note/tweet each of the labels shown above.
- Learning curve experiments (right, top) suggest Note/tweet content can predict (some) Birdwatch labels
- Backtest experiments (right, bottom) show less boost over “guess most frequent baseline”.
- There’s value in just keeping track of changing label frequencies!



Discussion

Importance of Data Labor

- Early stage data labor performed by volunteers is critical for keeping tracking of the frequencies of different kinds of tweets and notes.
- This data labor may be usable for automating certain predictions in the future.

Future of Birdwatch

- Twitter has been at the center of the discussion in the United States about the responsibility social media companies have for the content hosted and shared on their platforms.
- Debates about freedom of speech and company ownership continue to cloud the direction of the site and of the Birdwatch system itself.

www.nickmvincent.com / nickvincent@u.northwestern.edu / @nickmvincent
isaiah.jones@u.northwestern.edu