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

Isaiah Jones
Northwestern University
Evanston, Illinois, USA

Brent Hecht
Northwestern University
Evanston, Illinois, USA

Nicholas Vincent
Northwestern University
Evanston, Illinois, USA

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.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

Twitter, social computing, data labor, user-generated content

ACM Reference Format:
Isaiah Jones, Brent Hecht, and Nicholas Vincent. 2022. Misleading Tweets and Helpful Notes: Investigating Data Labor by Twitter Birdwatch Users. In *Companion Computer Supported Cooperative Work and Social Computing (CSCW'22 Companion), November 8–22, 2022, Virtual Event, Taiwan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3500868.3559461

1 INTRODUCTION

Researchers have paid increasing attention to concerns around misleading content and the role of community members in governing this content [10, 11, 14, 18–21]. In one such example, Twitter launched "Birdwatch" in January 2021, an experimental program that allows a set of invited users to collaboratively provide context in the form of *Notes* – short text posts and labels about a particular tweet and *Ratings* about other users' Notes [6].

In this paper, we are motivated by a "data as labor" [2] framing to investigate the labor performed by Birdwatch volunteers. People who use Birdwatch generate data that could be used in the future to train data-dependent systems. The Birdwatch system highlights an emerging tension in how misleading content is handled on major platforms: the work is very high stakes, but performed primarily by volunteers. The success of Birdwatch raises important questions about how data labor dependencies can be sustained in the long term, to avoid exchanging one problem (misleading content) for another (reliance on volunteer labor).

The first contribution of this paper is to describe the labels Birdwatchers apply to tweets and Notes and use exploratory topic modeling to describe the kinds of tweets Birdwatch users interact with. Our second contribution is to investigate if Birdwatch labels can be used to train a content-based classifier for tweets or similar posts on other platforms, e.g. to automatically flag tweets users are likely to find misleading. We find that some of the labels from the Birdwatch data are difficult to classify with a content-only approach and performance is likely to vary over time, but there is potential that classification models trained using Birdwatch data may be useful. The code for our experiments is available via GitHub.[1]

We also discuss how Birdwatch provides an example of how even in the absence of complex machine learning models, data labor can provide important value just in terms of estimating changing label frequencies to support a "guess most frequent" baseline. In the face of behaviors that change over time (like the frequency of misleading tweets), a continued supply of data labor is crucial.

2 RELATED WORK

Although Birdwatch is fairly new, researchers have investigated dynamics on the platform such as partisanship [1], vulnerabilities of the system [3], and the efficacy of fact checking [15]. This work builds on prior concerns about misinformation [13] on social media and evidence of Twitter being used for misinformation and disinformation, for instance in the context of 2018 Brazilian elections [16]. Various efforts have been made to support machine learning tools aimed at identifying "rumors" [7], "fake news" [4], and users likely to promote misinformation [8] on Twitter and similar platforms.

3 METHODS

We focus on two research questions: what kinds of topics are receiving attention from Birdwatch users, and if the data outputs from Birdwatch use can be used to train content-based classifiers.

We use two data sources: Birdwatch files[2] (Notes and Ratings) and tweet text collected using Tweepy [17]. We use data from the launch of Birdwatch, January 23, 2021, to February 21st, 2022. During this time, Birdwatch users wrote notes for 19503 tweets. 2299 of these tweets came from suspended accounts or were deleted at the time of our scraping and thus we could not determine the

CSCW'22 Companion, November 8–22, 2022, Virtual Event, Taiwan

ACM ISBN 978-1-4503-9190-0/22/11.
https://doi.org/10.1145/3500868.3559461

[1] https://github.com/nickmvincent/birdwatch_data_labor
[2] https://twitter.com/i/birdwatch/download-data

content of the tweet. From the remaining 17204 tweets, we looked at 25402 unique notes (a single tweet can receive many notes) and 189744 ratings.

As noted above, our data includes tweet content and Note content, both of which consist of short snippets of text. Following prior work that was successful at tweet classification [7], we capture key features: punctuation, number of characters, number of hyperlinks, and sentiment scores. We followed common procedures [12] for preprocessing text input. We lemmatize the text content, remove stop words, punctuation, numbers, capital letters, and hyperlinks. We considered several approaches for vectorizing this processed text data: spaCy's pre-trained medium-sized web model ('en_core_web_md'), size 300 TF-IDF vectors, and a baseline with no word vectors.

We first trained a LightGBM model for each binary label. We aim to classify each tweet and note as belonging to each category (based on majority vote of all applied labels). Some labels showed no predictive power at all. Then, for the labels that showed some predictive boost (we used a threshold of 0.6 AUROC as a heuristic, assuming performance under this threshold indicates very little signal in the content features), we compared performance of a LightGBM classifier with access to (1) spaCy word vectors, (2) TF-IDF vectors, or (3) no vectors. There were not major gains from using the more computationally expensive word vectors over TF-IDF vectors, so we used TF-IDF for our further experiments with learning curves and backtesting.

We produced a learning curve for each label. In other words, we repeatedly retrain models with increasingly large random samples of training data (increments of 10%). We expect to see a diminishing returns curve for each task.

These learning curves give one view on the "value of data labor" in terms of dataset size scaling. However, they fail to account for changes in label frequency over time. To understand if data labor by Birdwatch users adds predictive power in a more ecologically valid context (timestamped data would not be split randomly in practice), we conducted backtest experiments using growing time windows for training data and a sliding test window. For instance, we first use data from January 2021 to make predictions in February 2021. Then, we use both January and February to make predictions for March, and so on. If the distributions of labels and text stay relatively stable over time, we might expect our backtest results to provide qualitatively similar results to learning curves, i.e. diminishing returns curves.

4 RESULTS

First, we provide descriptive stats about (a) the labels Birdwatchers assigned to various notes and tweets and (b) the text summaries that Birdwatchers wrote. Then, we describe our early machine learning results.

Birdwatch users can use a number of categorical labels to describe tweets and notes. Figure 1 shows the fraction of all tweets (first two rows) that received a particular label and the fraction of all notes that received a note-specific label (bottom two rows), and the legend gives a sense of the kinds of labels available.

Topic Number	5 Most Salient Terms	Topic Number	5 Most Salient Terms
1	vaccine covid vaccinate jab fda	2	people covid antisemitism trump death
3	people biden risk covid president	4	ivermectin government trump biden student
5	nurse unvaccinated people berenson video	6	police stop chicago capitol shoot

Table 1: Shows the 5 most salient terms for our first 6 LDA topics.

4.1 Exploration with Topic Model

We performed a linguistic analysis of the tweet corpus, which provides context about the topics and events that appear in Birdwatch. We applied latent Dirichlet allocation (LDA) to model the topics discussed in the tweets that warranted a response from Birdwatch moderators (i.e. users).

This analysis shows clear discussion of U.S. current events like Covid-19 vaccination and the January 6 Capitol riot. We note this is consistent with the kinds of accounts that showed up frequently in the dataset, such as earthquake-announcement accounts and U.S. politicians.

4.2 Classifying Tweets and Notes

From the Birdwatch data, we trained predictive models for 38 distinct labels, including the labels shown in Fig. 1 and additional labels corresponding to helpful notes, misleading tweets, misleading but believable tweets, harmful tweets, and difficult to validate tweets.

We identified a small subset of labels to investigate more closely by training a single LightGBM model using TF-IDF and the features described above in Methods. This provided a relatively quick way to identify labels that were not likely to see good predictive performance with content only. Using 5-fold cross-validation and a simple heuristic cutoff – that a given label saw at least an AUROC of 0.60 (i.e. a boost of 0.10 over a random guessing baseline) – we identified 10 labels to investigate further.

Our learning curve results in Fig. (2a) showed the diminishing returns curve one would expect in supervised learning with random data sampling. By producing learning curves using random data, we can compare how performance actually varied over time with how we might expect performance to vary over time if the underlying distribution of labels and topics was not shifting.

We performed backtest experiments in use an increasing amount of training data to make predictions for a one month test window. This provides a more ecologically valid estimate of generalization error, and further provides insight into how data labor requires sustained input over time.

Shown in Fig. 2b, there is a large qualitative difference between the learning curves and the backtest results. This suggests much of the impact of Birdwatch data labor may be in simply identifying label frequencies, and not in performing content-based classification. Without a continued supply of data labor from users, even

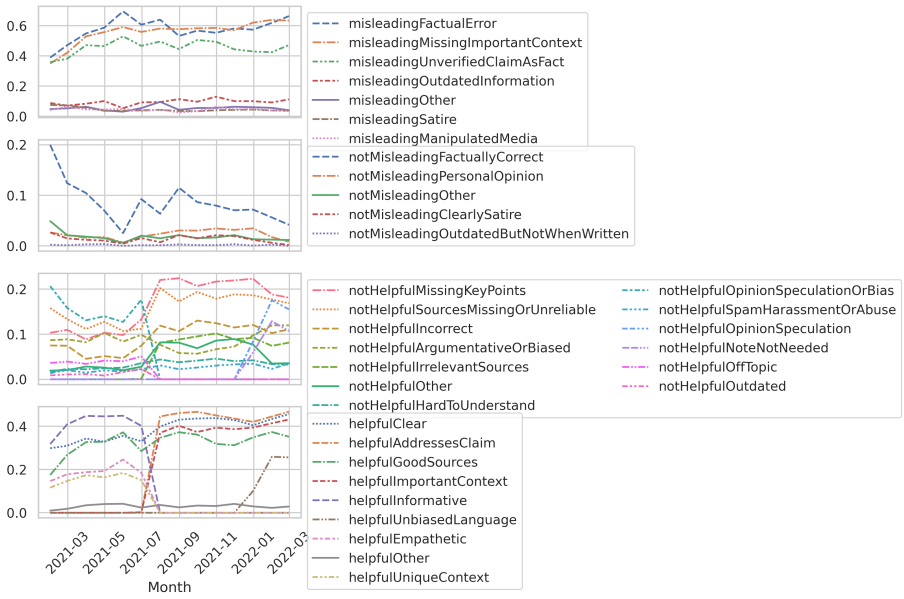
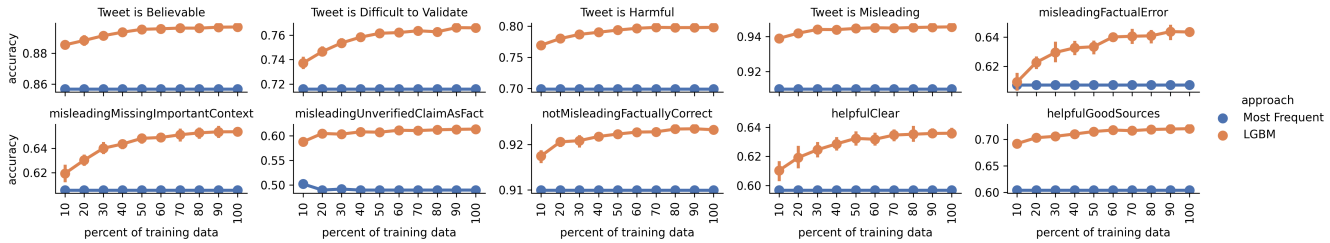
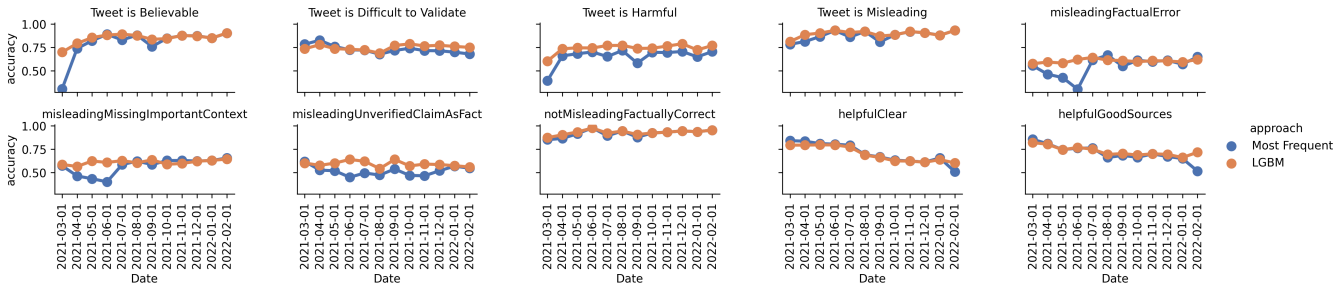


Figure 1: Shows the frequency of tweet labels (top two rows) and note labels (bottom two rows) over time. Each data point shows the fraction of all tweets or notes produced in that month that were given a particular label. Legend provides examples of the different labels in Birdwatch.



(a) Learning curves, showing how performance increases with more training data for the 10 labels with some predictive power. Note that y-axes vary.



(b) Backtest experiments, showing how performance varies over time.

Figure 2: Comparison of learning curves (top, showing growing performance) with backtest results (bottom)

.

trying to use a "guess the most frequent class" baseline would fail for some labels (because e.g. the most frequent class could change).

Taken together, the results of our experiments suggest that for many of the highly specific label choices available to Birdwatch users, content-based predictive models that use just the text of tweets or notes are unlikely to be that helpful in predicting these labels. However, for a small set of labels, using content to make label

predictions can be useful (i.e, AUC ROC around 0.7 and accuracy several points above baseline).

5 DISCUSSION

Our results provide early insight into the output of the data labor performed by Birdwatch users. Our analyses suggest that the early stage data labor performed by volunteers is critical for keeping tracking of the frequencies of different kinds of tweets and notes, and this data labor may be usable for automating certain predictions in the future.

The presence of distinct topics in the topic model exploration seems to suggest that Birdwatch moderators are indeed coalescing around potentially controversial topics like politics and news.

In the future, it may be possible to use Birdwatch data labor to fuel a browser extension or some other kind of 3rd-party tool that emulates the labeling process performed by actual Birdwatch users. Concretely, this could allow users to browse Twitter and be notified when coming across a tweet that Birdwatchers are likely to label as misleading.

5.1 Limitations and Future Work

Birdwatch Tweets vs. All of Twitter: In this study we worked with the non-random sample of tweets provided as part of the Birdwatch data release. For the purposes of deploying a system, it will be important to curate different evaluation sets, for instance a true random sample of tweets or a random sample of tweets from political accounts.

Moving Beyond Majority Vote Binary Labels: One major limitation of our work was that we used a simple majority vote approach to generate binary labels for each tweet. However, given that we wanted to provide an early exploration of the labels available in the Birdwatch dataset, this choice was useful in narrowing down our investigations. There is rich space for future work to consider modeling users at a more fine-grained level, or by incorporating recent work on advanced techniques for taking labeler differences into account, e.g. jury learning [9] and annotator fingerprinting [5].

Missing Data: Due to the nature of the material in the tweets that Birdwatchers respond to, some of those tweets are deleted or those users suspended. This restricts our access to the content of the most egregious offending tweets and makes modeling difficult as we are necessarily missing some of the most valuable predictive information.

REFERENCES

[1] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[2] Imanol Arrieta Ibarra, Leonard Goff, Diego Jiménez Hernández, Jaron Lanier, and E Weyl. 2018. Should We Treat Data as Labor? Moving Beyond 'Free'. *American Economic Association Papers & Proceedings* 1, 1 (2018).

[3] Garfield Benjamin. 2021. Who watches the Birdwatchers?: Sociotechnical vulnerabilities in Twitter's content contextualisation. https://pure.solent.ac.uk/en/publications/who-watches-the-birdwatchers-sociotechnical-vulnerabilities-in-tw

[4] Cody Buntain and Jennifer Golbeck. 2017. Automatically Identifying Fake News in Popular Twitter Threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*. 208–215. https://doi.org/10.1109/SmartCloud.2017.40

[5] Scott Allen Cambo and Darren Gergle. 2022. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3501998

[6] Keith Coleman. 2021. Introducing Birdwatch, a community-based approach to misinformation. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

[7] Amira Ghenai and Yelena Mejova. 2017. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In *2017 IEEE international conference on healthcare informatics (ICHI)*. IEEE Computer Society, Los Alamitos, CA, USA, 518–518. https://doi.org/10.1109/ICHI.2017.58

[8] Amira Ghenai and Yelena Mejova. 2018. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 58:1–58:20. https://doi.org/10.1145/3274327

[9] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3502004

[10] Naeemul Hassan, Chengkai Li, Jun Yang, and Cong Yu. 2019. Introduction to the Special Issue on Combating Digital Misinformation and Disinformation. *Journal of Data and Information Quality* 11, 3 (May 2019), 9:1–9:3. https://doi.org/10.1145/3321484

[11] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (July 2019), 31:1–31:35. https://doi.org/10.1145/3338243

[12] Ammar Kadhim. 2018. An Evaluation of Preprocessing Techniques for Text Classification. *International Journal of Computer Science and Information Security* 16 (06 2018).

[13] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* (March 2018). https://doi.org/10.1126/science.aao2998 Publisher: American Association for the Advancement of Science.

[14] Filippo Menczer. 2016. The spread of misinformation in social media. In *Proceedings of the 25th international conference companion on world wide web (WWW '16 companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 717. https://doi.org/10.1145/2872518.2890092 Number of pages: 1 Place: Montréal, Québec, Canada.

[15] Nicolas Pröllochs. 2021. Community-Based Fact-Checking on Twitter's Birdwatch Platform. *arXiv preprint arXiv:2104.07175* (2021).

[16] Raquel Recuero, Felipe Bonow Soares, and Anatoliy Gruzd. 2020. Hyperpartisanship, disinformation and political conversations on twitter: The brazilian presidential election of 2018. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 569–578. https://ojs.aaai.org/index.php/ICWSM/article/view/7324

[17] Joshua Roesslein. 2009. tweepy Documentation. *Online] http://tweepy. readthedocs. io/en/v3* 5 (2009).

[18] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A Platform for Tracking Online Misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web* (Montréal, Québec, Canada) *(WWW '16 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 745–750. https://doi.org/10.1145/2872518.2890098

[19] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2018. Anatomy of an online misinformation network. *PLOS ONE* 13, 4 (April 2018), e0196087. https://doi.org/10.1371/journal.pone.0196087 Publisher: Public Library of Science.

[20] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* (March 2018). https://doi.org/10.1126/science.aap9559 Publisher: American Association for the Advancement of Science.

[21] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. 2020. PolicyKit: Building Governance in Online Communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 365–378. https://doi.org/10.1145/3379337.3415858