Title: Methodological details appendix for "How do social media users and journalists express concerns about social media misinformation? A computational analysis"
Authors: Jianing Li (1), Michael W. Wagner (2)
Date: June 18th, 2024
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix: Methodological details

Data

To examine news coverage, we collected news transcripts discussing the topic of social media misinformation from three major broadcast networks, ABC, NBC, and CBS (*n* = 101). These sources were chosen based on their large audience size as well as on being broadly recognized as "mainstream media" even among partisans (Pew Research Center, 2021; Shearer & Mitchell, 2021); future work should expand beyond these sources in the centrist and center-left media ecosystem (Benkler et al., 2018) and examine sources with a wider range of ideological perspectives. A keyword list was inductively generated to collect relevant news coverage. Due to the length and format of the TV transcripts, only the content up to two conversation turns prior to the first occurrence of a keyword and two conversation turns after the last occurrence of a keyword was coded.

To examine social media discussions, we collected posts discussing topics relevant to social media misinformation from public Facebook pages and groups (*n* = 22,578) and from Twitter (*n* = 302,404). Facebook posts containing keywords were collected from CrowdTangle, a Meta-owned research tool that archived all posts from Facebook pages and groups that were set as public. Twitter posts containing keywords were collected from Synthesio, a third-party commercial data vendor. A keyword list slightly modified from the news coverage keyword list was used to collect relevant posts. We only collected English-language posts with a U.S. geographical location, but the keyword-based social media data were not representative of U.S. general population.

To improve the robustness of our results, we eliminated content posted by news organizations and journalists from the social media data. The final data consisted of 15,578 posts from public Facebook pages and groups and 301,334 posts from Twitter. To clean Twitter data, we adopted two datasets established by previous research that were the most comprehensive datasets to our knowledge: (1) for *news organizations*, we utilized the list created by Duan et al. (2022) that identifies 646 U.S. news organizations' Twitter handles and (2) for *individual journalists*, we utilized the list created by Tow Center for Digital Journalism (Gotfredsen, 2023) that identifies 3,703 journalists' Twitter handles who work for U.S. news organizations. After getting rid of seven duplicates (as Tow Center's list contained seven organizational accounts), a total of 4,342 Twitter handles were identified. These Twitter handles included news organizations and journalists on the local and national level in the United States. All original posts from these accounts were eliminated from the analysis (*n* = 1,070).

Given the scarcity of research on news organizations' and journalists' Facebook accounts, to clean Facebook data, we utilized meta-data provided by Facebook. For public Facebook pages, we identified 15 page categories broadly associated with the news media in the 329 page categories available in our data: Journalist, News Personality, Newspaper, News Site, Newsagent/Newsstand, Media, Media-News Company, Media-Show, Media Agency, TV Channel,

TV Network, TV Show, Radio Station, Broadcasting Media Production, and Magazine. These page categories were provided by Facebook and self-selected by page owners. Branches of a news organization may have separate public pages that fell under different categories, for example, at the time of data collection, NBC News fell under "Broadcasting Media Production," NBC Politics fell under "Media-News Company," NBC's regional affiliate NBC2 News fell under "TV Channel," and NBC Nightly News with Lester Holt fell under "TV Show." As our primary goal was to eliminate confounding factors (i.e., content posted by news organizations and journalists) from social media data, we prioritized comprehensiveness in the categories identified above and included both general and specific categories in the elimination list. A total of 7,000 posts from 2,516 public Facebook pages were eliminated from the analysis. Apart from public Facebook pages, our data included posts from public Facebook groups. We kept all posts from public Facebook groups for two reasons: (1) CrowdTangle anonymized posts from groups, (2) as we discarded hyperlinks, in cases where members shared news in a group, we expected the content captured by our data to be primarily consisted of social media users' discussion of news rather than news coverage itself.

The timeframe for both datasets was July 28 to December 3, 2020, covering months during the election campaign when the topic of social media misinformation became a salient topic, till one month after the 2020 presidential election. Pilot searches were used to validate that the timeframe included the most active periods for both the social media and the news media discussions.

Human labeling

Two undergraduate coders labeled a random sample of 1,168 unique social media posts and the entire corpus of 101 news transcripts. Coders labeled for the presence of the five types of discussions, which were not mutually exclusive.

For the *party politics* discussion, coders labeled 1 if the content discussed misinformation through a lens of party conflict and political (dis)advantage and labeled 0 if otherwise. The *party politics* discussion covered discussions about how misinformation benefited or harmed political groups, attribution of blame on a party/politician, and partisan politicians' attacks on one another. Examples of *party politics* discussion included "@Facebook throttled progressive news in favor of right-wing disinformation sites," "Four years ago BuzzFeed posted the Steele Dossier, filled with lies and misinformation, and Twitter and Facebook happily spread their fake news. Now these platforms are blocking transmission of a New York Post story critical of a Democrat. The hypocrisy is not sustainable," and "The Biden campaign is pressing Facebook to remove posts by Trump and slamming the company as the nation's foremost propagator of disinformation about the voting process." However, merely mentioning political parties, politicians, or elections to state factual information was not treated as a sufficient signal of *party politics* discussion unless the content explicitly discussed party conflict and political (dis)advantage (e.g., "Twitter temporarily prevented Donald Trump Jr. from tweeting and retweeting after he shared coronavirus-related misinformation" was coded 0).

For the *quality of knowledge and decisions* discussion, coders labeled 1 if the content discussed misinformation in relation to the accuracy of knowledge and/or the quality of decisions (e.g., misinformation makes people believe that there is an election fraud, "poisons your mind," or polarizes the public), and otherwise 0. Such discussions can be either on an individual level (e.g.,

"He admitted that he wanted to kill a 'random white person' after watching the fake news") or a collective level (e.g., "COVID-19 disinformation is everywhere. And it's undermining our response in the US").

For the *direct correction* discussion, coders labeled 1 if the content offered explanation on why a piece of information is false, and otherwise 0. Coders used a stringent criterion to label the *direct correction* discussion: the content needed to both identify misinformation and provide information on why it is false. Examples of *direct correction* discussion included: "A video that falsely claims Joe Biden wore a wire during last night's debate has been shared thousands of times on Facebook. The clip shows a shirt crease." and "Mark Zuckerberg's claim that QAnon is *just now* evolving from misinformation to violence is patently false. An armed QAnon supporter blocked traffic at the Hoover Dam bridge in June 2018. The FBI called it a domestic terror threat in May 2019." Content merely saying "this is fake news" or "this is misinformation" was labeled as 0.

For the *user agency* discussion, coders labeled 1 for content describing or calling for actions from ordinary people to counter misinformation, and otherwise 0. Examples of the *user agency* discussion included content asking social media users to "be aware of" misinformation, warning people to "watch out for" certain rumors recently circulating online, encouraging people to read and share corrections/facts with others (e.g., "join my page to watch live fact-check"), and asking people "to make judgement based on facts not misinformation," etc. Only actions specifically targeted at *countering* misinformation were coded as 1; actions to spread misinformation (e.g., asking people to save/access content that has been removed by social media platforms as misinformation from alternative sources) or not directly related to addressing misinformation (e.g., "we should retake our nation") were labeled as 0.

For the *platform policy* discussion, coders labeled 1 for content describing or calling for actions from social media platforms in countering misinformation, and otherwise 0. Examples of the *platform policy* discussion included content asking platform to or describing existing actions to "take a stand on" or "crack down on" misinformation; "help stop misinformation;" "penalize" someone for misinformation; "suspend," "ban," "demote," or "demonetize" accounts; "flag," "fact-check," "decrease the reach of," or "taking down" false posts. Again, only actions that *counter* misinformation were labeled as 1; platform (in)actions that passively "permit" or "fuel" misinformation to spread or not directly related to addressing misinformation (e.g., policy on hate speech and racism) were labeled as 0.

Coders were instructed to take account of all textual content (including message/transcript text, image text, and social media hashtags) but discard hyperlinks. To determine intercoder reliability, a random sample of 159 documents (including tweets, Facebook posts, and news transcripts) from the whole corpus was selected. Coders achieved satisfactory reliability for all categories: calculated using *R* package {irr}, Krippendorff's α (nominal) was 0.85 for *party politics*, 0.94 for *quality of knowledge and decisions*, 0.92 for *direct corrections*, 0.76 for *user agency*, and 0.90 for *platform policy*.

Supervised machine learning

While human labeling covered the entire corpus of broadcast news transcripts, the large corpus of social media posts was classified using supervised machine learning based on the sample of human-labeled data.

Using *R* packages {caret} and {readme}, we tested four supervised machine algorithms: (1) the Hopkins-King method, (2) a Naïve Bayes classifier, (3) Support Vector Machine (SVM), and (4) Random Forest.

Research has shown that algorithms coming from the computer science tradition optimized to classify *individual documents* (e.g., Bayes, SVM, Random Forest) are not necessarily optimized to make generalizations about the *population* of documents (e.g., *the proportion of a category*), which is often the theoretical interest of social scientific research. Algorithms with a low misclassification rate for individual documents may still produce large errors for category proportions when all the misclassifications are in a particular direction. For example, an algorithm classifying 100 documents with 60 true positives, 20 true negatives, 20 false positives, and 0 false negatives has a level of accuracy (80%) that outperforms conventional threshold of 70% but would produce a biased estimate that 80% documents contain the positive category while the true proportion is 60%. While sampling and tuning strategies may improve the precision and recall of individual classifiers, better category proportion estimates are not always possible. To address this issue, Hopkins and King (2010) and Jerzak et al. (2022) developed a method that produces unbiased estimates for category proportions even when the optimal individual classifier performs poorly. Compared to the standard individual classifiers, the Hopkins-King method does not need to classify individual documents to estimate category proportions, and more accurately represents the data generation process in the real world by modeling the possibility of word stems **S** occurring in category D [i.e., P(**S**|D)] rather than the inverse approach taken in standard individual classifiers [i.e., P(D|**S**)]. The Hopkins-King method was able to classify the category proportions in our data with an average error of 7.26% (1.12% for *party politics*, 8.98% for *quality of knowledge and decisions*, 10.61% for *direct corrections*, 9.42% for *user agency*, and 6.17% for *platform policy*), comparable to levels reported in previous literature (Hopkins & King, 2010; Jerzak et al., 2022).

While the Hopkins-King method is superior in estimating category proportions, it does not offer information on how different types of discussions co-occur with each other because it does not classify individual documents. Here, we leveraged the advantage of three individual classifiers. Three-fold cross-validation, a resampling procedure that partitions the data to train more generalizable models, was used to tune model hyperparameters for all three types of classifiers. Moreover, to address data imbalance and improve precision and recall, we combined an under-sampling technique (a synthetic dataset was constructed where the majority category was randomly sampled to be the same size as the minority category) and an over-sampling technique (the cross-validation resampling procedure incorporated an over-sampling process where the minority category was randomly sampled with replacement to be the same size as the majority category). For each type of message, we tested the performance of different algorithm specifications, inspected the confusion matrices, and used accuracy and F1 score to select the best-performing algorithm. Random Forest outperformed SVM and Naïve Bayes in all types of discussions, producing an accuracy of 77.31% and an F1 score of 76.33% when classifying *party politics*, an accuracy of 78.71% and an F1 score of 78.61% when classifying *quality of knowledge and decisions*, an accuracy of 93.13% and an F1 score of 74.29% when classifying *direct corrections*, an accuracy of 90.48% and an F1 score of 73.44% when classifying *user agency*, and an accuracy of 80.59% and an F1 score of 79.09% when classifying *platform policy*.