Current Academic Research on the Information Environment

Nilima Pisharody[ⱡ] and Jen Rosiere Reynolds[ǂ]

June 16, 2022

Abstract

We performed a systematic review of top general interest and field journals in order to understand the research production steps used to study the information environment. The key questions we aimed to answer were: What data generation processes are being followed?; and How are academics getting data to study social media?

Using Google Scholar, we collected articles within the six major general-interest science journals and the top ten journals in the fields of communications, economics, political science, and sociology with the terms "social media," "disinformation,""misinformation," "dis/mis," or "mis/dis" between 2017 and 2021. We analyzed relevant publications (n=169) from a sample (probability proportional to the normalized and scaled citation).

Most of the sampled papers use manual data collection and non-statistical methods. The vast majority analyzed the content of posts (83%), primarily using textual analysis (68%). For advanced methods 23% used machine learning and 19% used network analysis. 59% of the papers studied Twitter, with only 22% of those papers examining other platforms as well (13% of all papers). 26% of publications looked at Facebook. After those platforms, there was a sharp drop in commonalities across publications. In terms of obtaining data, 49% of the analyzed papers used a platform API, 34% used manual annotations or coding, 14% crowdsourced or used surveys, 14% used independent collections, 9% scraped, and 1% used social listening tools or bots the author created. Few publications we reviewed used a platform's proprietary or paid access data.

ⱡ Princeton University

ǂ Carnegie Endowment for International Peace

Funding: This research was generously supported by Microsoft, Craig Newmark Philanthropies, the John S. and James L. Knight Foundation, and the William and Flora Hewlett Foundation. This work was initiated as part of a joint project between the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace and the Empirical Studies of Conflict Project at Princeton University. Sean T. Northon, Jacob N. Shapiro, and Alicia Wanless provided extensive feedback. **Author contributions:** NP collected all data, wrote all code, visualized data, and wrote Detailed Methodology appendix. JRR wrote and edited the manuscript and provided oversight and leadership responsibility for the research activity planning and execution. **Data and materials availability:** All data required to evaluate the conclusions are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Finally, to get a general sense of where this research is taking place, we identified the country of employment for all publications' first author. We found that authors from the sample were overwhelmingly from Western English-speaking countries.[1]

[1] The samples of papers from economics, political science and sociology journals are exhaustive for information environment research in these fields. While our sample is representative of information environment research in the communication field, the communication sample is not exhaustive. Nevertheless, the results from this study still provide an overall view of the state of research on the information environment.

Table of Contents

Table of Contents 1

Executive Summary 2

Introduction 3

Methodology 3

Sampling 3

Overview of Publications 4

Platforms 4

Observation level and Content 5

Methodology 5

Researchers 6

Conclusion 6

Appendix 7

A.1 Detailed Methodology 7

A.2. List of Journals 7

A.3 Coding Research Practices - Google Forms.pdf 7

A.4 Data 7

Executive Summary

1. Using Google Scholar, we collected articles from the six major general-interest science journals and the top ten journals by impact factor in the fields of communications, economics, political science, and sociology published with any of the terms “social media,” “disinformation,” “misinformation,” “dis/mis,” or “mis/dis” in the title or full text between 2017 and 2021.
2. We analyzed 169 publications from a sample, with sample weights proportional to the normalized and scaled number of citations.
3. Most publications we analyzed looked at Twitter (59%), and a fifth of those (22% of the 59%) did so in conjunction with other platforms. 26% of publications looked at Facebook. After those platforms, there was a sharp drop in commonalities across publications.
4. About half of the papers (49%) used a platform API, 34% used manual annotation or coding, 14% crowdsourced or used surveys, 14% used independent collections, 9% scraped, and 1% used social listening tools or bots the author created. Very few publications we reviewed used a platform's proprietary or paid access subset.
5. The vast majority (83%) of papers analyzed social media at the post level, and most used textual analysis to do so (68%).
6. About a fifth of the papers used machine learning (23%) or network analysis (19%).
7. 27% used feature extraction, with 80% of those relying on natural language processing.
8. The country of employment for all publications’ first authors were overwhelmingly Western English-speaking countries, with the United States (47%), United Kingdom (12%), and Canada (4%) making up the top three.

Introduction

How do academic researchers approach studying the information environment? To date, there has not been any publicly available survey that helps broadly answer this question. This poses a substantial challenge to those looking to suggest improvements to the field. Therefore, we conducted a systematic review of top general interest and field journals whose publications included terms "social media," "disinformation," "misinformation," "dis/mis," or "mis/dis" between 2017 and 2021. We analyzed a sample of 169 relevant publications from this set.

We found a lack of diversity in the platforms examined, methodologies used, and backgrounds of the researchers. Additionally, the vast majority of papers used only simple econometric methods.

Most publications looked at Twitter (59%), 26% looked at Facebook, and the remaining 15% examined various other platforms. As a result of this disproportional focus on Twitter and Facebook, many widely used platforms appear to be virtually ignored. In terms of methodology, 49% of the papers used a platform API, 35% used manual annotation or coding, 14% crowdsourced or used surveys, 14% used independent collections, 9% scraped data outside of the platform API, and 1% used social listening tools or bots created by the author. Few publications we reviewed used a platform's proprietary or paid access subset. 23% used machine learning and only 19% used network analysis. Finally, the country of employment for all publications' first authors were overwhelmingly Western English-speaking countries; the top three represented were the United States (47%), United Kingdom (12%), and Canada (4%).

This report reveals promising opportunities for increasing efficiencies, enabling research, and growing knowledge in this important field.

Methodology

We provide here an overview of the methodology behind our review; for a more thorough explanation, including the code used, please see Appendix A.1.

Using SerpAPI, we scraped the first five pages of Google Scholar search results after searching for publications with the following criteria:

1. Keywords: All articles published with the terms "social media," "disinformation,""misinformation," "dis/mis," or "mis/dis" in the title or full text;

2. Outlets: Articles published in the six major general-interest science journals and the top ten journals by impact factor in the fields of communications, economics, political science, and sociology according to Google Scholar (see Appendix A.2 for all journals)[2]; and
3. Year of Publication: Articles published in 2017-2021.

This yielded 7,052 papers. We removed papers with zero citations and duplicate entries, yielding an evaluation sample of 5,724 papers. Table 1 displays the number of papers in the initial sample, the number of duplicates, the number of zero citation papers, and the final number of papers included in our evaluation sample.

Table 1: Paper Evaluation Process

Field	*Total #*	*# with 0 citations*	*# of duplicates*	*Evaluated*
Communication	2574	185	249	339
Computer Science	736	63	0	673
Economics	426	83	47	296
General Interest	1144	64	296	794
Political Science	864	77	21	766
Sociology	1308	187	66	1055
Totals	7052	659	679	3923

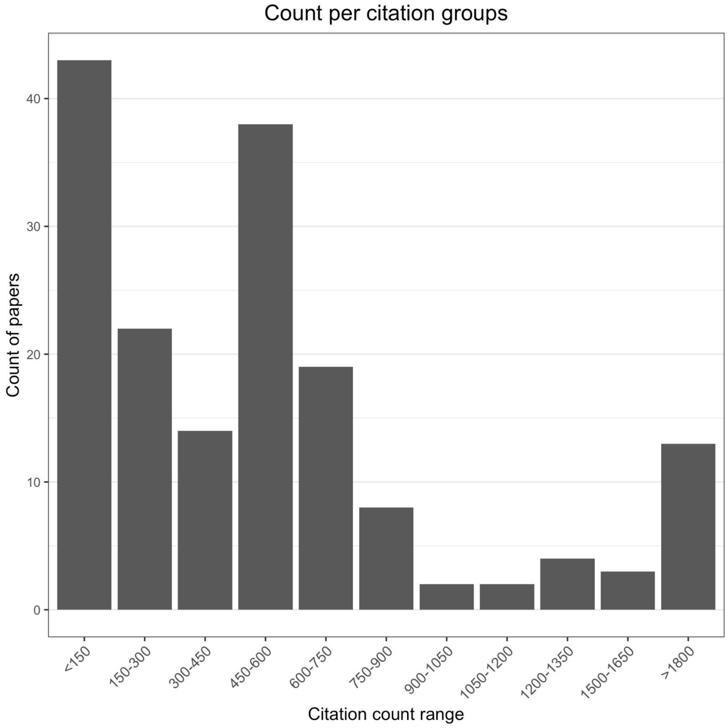
Sampling

In order to understand research practices in a manner correlated with influence, we created a stratified random sample, stratified by journal and sampled with probability proportional to normalized citation count. SerpAPI pulled the paper's number of citations from Google Scholar. We then normalized the

[2] Although we executed our code for all our listed journals, the search did not yield any results for *Nature Human Behavior*, *ACM Conference on CSWC \& Social Computing*, *ACM/IEEE International Conference on Human Robot Interaction*, *ACM Conference on Pervasive and Ubiquitous Computing (UbiComp)*, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, and *ACM Symposium on User Interface Software and Technology.* This left us with downloaded information for 10 journals (each) in the fields of economics, communications, political science, and sociology, and five journals (each) in the fields of computer science and general interest. Consequently, we had information from 50 journals, and the steps yielded 7,052 research papers. The output included the variables of author names, titles, journal names, year of publication, Google Scholar link, and paper source link.

citation by dividing it by the number of years the article has been in circulation and by the specific journal's average impact factor for the years 2017-2021, if it existed, and then multiplied that number by 100.

Figure 1. Distribution Of Citation Counts in Evaluation Sample



We randomly sampled each discipline from the 5,724 paper evaluation sample with a probability proportional to these normalized citation counts. Our target sample was 50 papers each for the fields of economics, political science, communications, and sociology and 25 papers each for computer science and general interest. We drew smaller samples for the latter two fields as SerpAPI only covered publications that matched our criteria in five journals. To make the sampling reproducible, we also set seeds for each category.

Although we specified query keywords and parameters, Google Scholar searches still displayed links to papers that had no relation to our study. To select papers relevant to the study, we manually coded each paper in the sample for relevance based on the paper's abstract. To replace irrelevant papers within each discipline we replaced removed papers with papers randomly sampled from a replacement set of previously unsampled papers.

After drawing two replacement sets of 250 papers each (750 total, including the original set) we still had not met our quota for each field. Given time constraints and the high number of irrelevant results in the initial sample, we decided not to draw additional samples using SerpAPI, and instead used the final realized sample of n=169 publications:[3] 49 were in communications, 24 in computer science, 16 in economics, 27 in general interest, 25 in political science, and 28 in sociology. For all fields except communications, this represents the total population of all relevant papers in the top 10 journals. For communications, this represents a random weighted sample of the total population. For the full data, please see Appendix A.3.

Coding

Eight coders then labeled articles that dealt with data from mainstream social media platforms, platforms not based in or popular throughout the United States, and alt-tech platforms. We excluded articles and works that assess research practices instead of measuring real-world or online relationships. For full coding instructions, please see Appendix A.3. One trained individual manually coded each paper.

Overview of Publications

While these papers represented publications in the most impactful and competitive academic journals, the majority were not extremely sophisticated in data collection or analytical methodology. Most papers analyzed the content of posts using textual analysis. Additionally, most of these publications did not use machine learning, and less than a quarter used network analysis.

Platforms

Most analyzed publications looked at Twitter (59%), and many (22% of the 59%, or 13% of the total) did so in conjunction with other platforms. 26% of publications looked at Facebook. After those platforms, there was a sharp drop in commonalities across publications. The next most frequently analyzed other platforms were Reddit at 7%, YouTube and Instagram at 5% each, and Sina Weibo at 3%. The remaining platforms analyzed represented less than one percent of the sample but included various platforms such as 4chan, Gab, iWiW (a Hungarian platform), LinkedIn, Telegram, Twitch, and VKontakte (VK).

In summary, the sample of papers we analyzed revealed a heavy emphasis on Twitter. The likely reason for this is that Twitter provides a relatively open model for researchers, making it easier to acquire data. However, Twitter is not necessarily representative of the public at large; at least in the United States, Twitter's users are younger and more likely to identify as Democrats than the general public.[4]

[3] Replacement sets were sampled using the same weights and sampling strategy as the original sample of 250. Papers were drawn from the replacement set to replace irrelevant papers in the original sample, e.g. if an original sample political science paper was found to be irrelevant, a political science paper was drawn from the replacement set and coded (including discarding and redrawing if the replacement paper was also irrelevant).

[4] Huges, Adam, and Stefan Wojcik. "Sizing Up Twitter Users." *Pew Research Center* (24 April 2019). https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Observation level and Content

To obtain data, 49% of the papers used a platform API, 34% used manual annotations or coding, 14% crowdsourced or used surveys, 14% used independent collections, 9% scraped data outside of platform APIs, and 1% used social listening tools or bots the author created. Very few of the reviewed publications used a platform's proprietary or paid access subset. For instance, only one paper used Twitter's 2017 dataset of identified Internet Research Agency[5] accounts and only two used the sample of 10% of Twitter users known as the Decahose. Another two used Gnip PowerTrack API to access the historic Twitter Firehose.

The vast majority (83%) of papers analyzed social media at the post level. The next most common type of analysis was user-level (~13%), then user-time-level (4%), then location (2%).

94% of published reports analyzed content (whereas 2% analyzed metadata). Of that, 60% examined generated content (e.g., content by users, groups, bots), 21% direct interactions with or within posts (e.g., reactions or comments), 17% indirect interactions with the post (e.g., shares or retweets), and about 1.4% moderated content or moderation responses.

Of the 60% of reports that looked at generated content, 71% analyzed text, 13% images, 8% video, 4% URLs, and one report (.6%) analyzed audio.

Methodology

12% of reports tracked information flows across more than one platform. 71% of those reports followed content across platforms, while 42% tracked users across platforms. 14% of reports that followed across platforms did both.

Similar percentages of publications used simple econometric analysis and graphical or visual analysis (53% and 51%, respectively). The next most common method of analysis was descriptive (42%), then qualitative (29%), machine learning (23%), and network analysis (19%). Most papers (68%) used more than one method.

Out of those papers that used machine learning, about half (47%) used supervised machine learning, while about 31% used semi-supervised and 21% unsupervised.

[5] The Internet Research Agency was identified by both Twitter and the US government as the actor behind Russian attempts to influence the 2016 American elections.

27% used feature extraction, with the vast majority of those using natural language processing techniques (80%). About 11% performed network analysis with feature extraction, and about 8% used other econometric methods with feature extraction.

A bit less than half of the publications aggregated data to higher-level categories to perform analysis (44%); of those that did, 21% aggregated by platform behavior types, 15% aggregated at the social media account level (e.g., all posts by bots or groups), and 9% aggregated at the network level (following, being followed, linking). A small minority of studies aggregated by time period (2%). Note that some studies (~2%) used more than one aggregation method.

A large majority (90%) of our sample did not aggregate data by location. Those that aggregate by location did so at the country level (5%), the county or town (1%), city (1%), census tract (a county subdivision) (.5%), or state (.5%).

Researchers

We identified the country of employment for all publications' first author. This approach gives us a general sense of the primary institution responsible for the production of the research output. Just under half of our sampled papers (47%) were produced by first authors in the United States. The next most frequently represented nation was the United Kingdom (12%). Canada was the third most represented at 4%, and Belgium, Germany, Italy, Spain, and Switzerland each represented about 3% of the sample. The remaining regions included other European Union nations and Nordic countries, which collectively made up about 3% of our sample, then countries within Asia (3%), and finally the Middle East (2%).

Conclusion

The information environment impacts how millions of people understand and interpret the world, and the scale and pace of events happening online is only growing. However, the full universe of papers on this subject remains relatively small; exhausting all relevant papers in the top 10 political science, economics, sociology, computer science, and general interest journals produced a total population of 132 papers, with only the communications field not being fully exhausted. The majority of these papers were not extremely sophisticated in data collection or analytical methodology, with particularly few papers using machine learning or network analysis and the vast majority favoring analyses of textual content over visual content. Additionally, papers focused overwhelmingly on content from the two platforms most prevalent in developed Western democracies: Twitter and Facebook. While considerable progress in our understanding of the information environment over the past decade, the overall state of the field can be characterized as shallow in terms of geographical, platform, and methodological focus. Recognizing and remedying the deficiencies outlined in this approach is a critical step in ensuring that the field continues to grow and produce high-quality research in the years to come.

Appendix

A.1 Detailed Methodology

A.2. List of Journals

A.3 Coding Research Practices - Google Forms.pdf

A.4 Data