Tracking the Use of Twitter Data in Academic Research
DRAFT

Authors: Ryan Murtfeldt, Naomi Alterman, Ihsan Kahveci, and Jevin D. West

University of Washington
Center for an Informed Public

R code available here

Final dataset available here

February, 2024

Abstract

Since 2006, Twitter's Application Programming Interface (API) has been a treasure trove of high-quality, freely available data for researchers studying anything from the spread of misinformation, to social psychology and emergency management. However, in the spring of 2023, Twitter (now called X) began charging $42,000/month for its Enterprise access level (X, 2023). Lacking sufficient funds to pay this monthly fee, academics are now scrambling to continue their research without this essential data source (Ledford, 2023). This study collects and tabulates the number of studies, number of citations, dates, major disciplines, and major topic areas of studies that used Twitter data between 2006 and 2023. While we cannot know for certain what will be lost now that Twitter's API is cost prohibitive, we can illustrate the enormous value of freely available data by examining the research conducted while Twitter's API was free to use. A search of 8 databases and 3 related APIs found that since 2006, a total of 27,453 studies have been published in 7,432 distinct publications, with 1,303,142 citations, across 14 disciplines. Major disciplines include: computational social science, engineering, data science,

social media studies, public health, and medicine. Major topics include: information dissemination, assessing the credibility of tweets, strategies for conducting data research, detecting and analyzing major events, and studying human behavior in its many forms. Alarmingly, while Twitter data studies have increased every year since 2006, following Twitter's decision to begin charging for data in the spring of 2023, the number of studies has decreased by 13%. We assume that much of the data used for studies published in 2023 were collected prior to Twitter's February, 2023 API shutdown, and thus the number of new studies are likely to decline even more in 2024 and beyond.

Introduction

Since 2006, Twitter's Application Programming Interface (API) has been a treasure trove of high-quality, freely available[1] data for researchers studying anything from the spread of misinformation, to social psychology and emergency management (Golder et al., 2011; Jie Yin et al., 2012; Vosoughi et al., 2018). However, within a year of Elon Musk's purchase of the platform in April, 2022, Twitter (now called X, however for clarity we will refer to the platform as "Twitter" for the remainder of this paper) has begun charging $42,000/month for its Enterprise access level (X, 2023b), and now requires researchers to get permission each time they want to share tweet IDs with other researchers (X, 2023a), making peer review and study replicability much more difficult. Lacking sufficient funds to pay this monthly fee, academics are now scrambling to continue their research without this essential data source (Ledford, 2023). This paper aims to highlight what will be lost if social media data continues to be cost prohibitive. To this end, we want to understand just how many Twitter-based academic papers, across all disciplines, have been published during the time of freely available data (2006-2023). We specify the disciplines to which the majority of studies belong, and highlight some of the most common topics explored in the literature. While we cannot know for certain what will be lost now that Twitter's API is cost prohibitive, we can illustrate the enormous value of freely available data by examining the research conducted while Twitter's API was free to use. This research spans the academic disciplines including: computational social science, engineering, data science, social media studies, public health, and medicine.

In addition, governing bodies are currently passing legislation, such as the European Union's Digital Services Act of 2022, to require or incentivize technology companies (such as Twitter) to

[1] Following the Cambridge Analytica scandal in 2018, Twitter began implementing restrictions on its API, including an application process which vetted users before granting access, and implemented download limits such as 300 tweets/retweets per 3 hours (Hutchinson, 2018).

provide open access to their data. We hope this paper will help inform these decisions by showing the kinds of knowledge that are likely to be lost if policies do not change and open access to social media data does not return.

Other studies have undertaken similar inquiries, collecting and analyzing studies that use social media data, but usually with a much narrower, discipline-specific, approach. For example, in “A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions” (Brito et al., 2021), researchers examined 83 articles, all of which addressed the use of social media to predict public election results. In another example, “Social Media and Attitudes Towards a COVID-19 Vaccination: A Systematic Review of the Literature” (Cascini et al., 2022), researchers analyzed 156 studies, examining the role of social media in vaccine hesitancy. Exploring Twitter data studies from a different angle, the paper “Twitter and Research: A Systematic Literature Review Through Text Mining” (Karami et al., 2020) took an approach similar to our own, capturing 18,849 studies using Twitter data between 2006-2019 and categorizing them into 38 topics using word-frequency analysis and Latent Dirichlet Allocation.

In contrast, our study sought to capture a comprehensive and up to date picture of Twitter data studies by collecting a wide sample of studies that used Twitter data up to December, 2023, identifying prominent disciplines, topics, publications, and date distributions, quantifying the influence of Twitter research by calculating citations, and showing the drop off in studies following the closure of Twitter’s free API in early 2022. We collected a total of 27,453 unique studies, in 7,432 distinct publications (journals and conferences), with 1,303,142 citations, spanning 17 years, and across 14 disciplines.

Methods

Data Collection

Our primary goal was to collect and analyze all studies across all disciplines that have used Twitter data as the focus of their inquiry since 2006, the year Twitter opened its API to academic researchers. Given the fractured landscape of literature databases, it was necessary to collect studies from a wide variety of sources in order to capture the maximum number of studies from every possible discipline. As a starting point, we conducted a broad search using Web of Science, the most comprehensive, multidisciplinary database available. Searching the topics field, which included title, abstract, and author keywords (Topics = twitter NEAR/3 data OR twitter NEAR/3 api OR twitter NEAR/3 dataset), we located 3628 articles. Utilizing Web of Science's built-in "Analyze Results" feature, we found that the top disciplines included: Computer Science, Engineering, Information Science, Communications, Public and Environmental Health, and Multidisciplinary Sciences. We then referenced the University of Washington's library guides for each of these disciplines to identify the most relevant research databases (see below), and then set about searching each database. All searches used some version of our initial search string, adjusting proximity operators as appropriate, and searching primarily in the topic, title, abstract, and keyword fields. We also found that adding "NOT survey" to the string eliminated studies that simply used Twitter to disseminate surveys or find participants for data collection. Finally, we fine-tuned each search using built-in filters to include only journal articles, conference papers, dissertations, and pre-prints.

Below, we list each database along with the total results found, and the percentage of relevant studies within each results list. The statistical software, R, was used to randomize results for sampling. For each database (see Appendix A), a minimum of 50 sample studies were examined by hand to determine if they met one of three criteria: utilized Twitter data in the study, examined novel ways of extracting and studying Twitter data, or reviewed the literature of

Twitter-based studies. To label sample studies relevant/not relevant, we found that most studies explicitly stated in their abstract if they utilized Twitter data. For example, “The researchers analyzed 100,000 tweets with hashtags #coronavirus…” (Pandey et al., 2022). In a minority of cases, when the abstracts were unclear, we examined methodology sections for confirmation. The most common reasons for labeling “not relevant” were studies that used Twitter to disseminate surveys, analyzed surveys about Twitter use, and studies that mention Twitter, but actually examine Sina Weibo, China’s Twitter alternative, or another social media format.

- Library and Information Science Source/Library Information Science and Technology Abstracts 1660 articles (82% relevant)
- Web of Science 11,617 articles (82% relevant) - without proximity operators
- Global Health 563 articles (80% relevant)
- ACM Digital Library 1997 articles (92% relevant)
- IEEE Xplore 4930 articles (97% relevant)
- Engineering Village (Compendex) 21,574 articles (86% relevant)
- Engineering Village (Inspec) 14,664 articles (88% relevant)

In the case of Engineering Village, the web-based database limits downloads to 1000 studies within a given search. With such a large number of relevant studies published in this database (over 36,000), we deemed it essential to find another way to access these studies. Elsevier (publisher) offers two APIs for Engineering Village: the search API, and the retrieval API. We utilized R exclusively to access these data. Using the search string “(twitter AND data) OR (twitter AND api) OR (twitter AND dataset) NOT survey” we used the search API to obtain the “doc id” for each study, and then used the retrieval API to obtain metadata for a total of 36,238 studies. Extensive computational programming[2] with R and Excel was required to unnest, clean, wrangle, and analyze the data. We combined the Engineering Village dataset with the dataset created from the other six databases, removed duplicates using R’s “distinct” function, and randomized to create the final dataset.

[2] While we were able to access and download all studies from the other six databases (partially due to the smaller number of studies in question) directly from their websites, accessing the studies from Engineering Village required several weeks of programming (using R) to identify, retrieve, parse, and wrangle the necessary metadata. We believe this may be a hindrance to future studies of this sort, especially as the number of studies inevitably increases over the years.

To quantify influence, we collected the citation count[3] for each study in the dataset via the Crossref REST API. The final dataset (27,453 articles) includes: title, abstract, date of publication, manuscript date, document type, publisher, publishing company, DOI, and citation count. Code available here, Final dataset available here.

Data Analysis

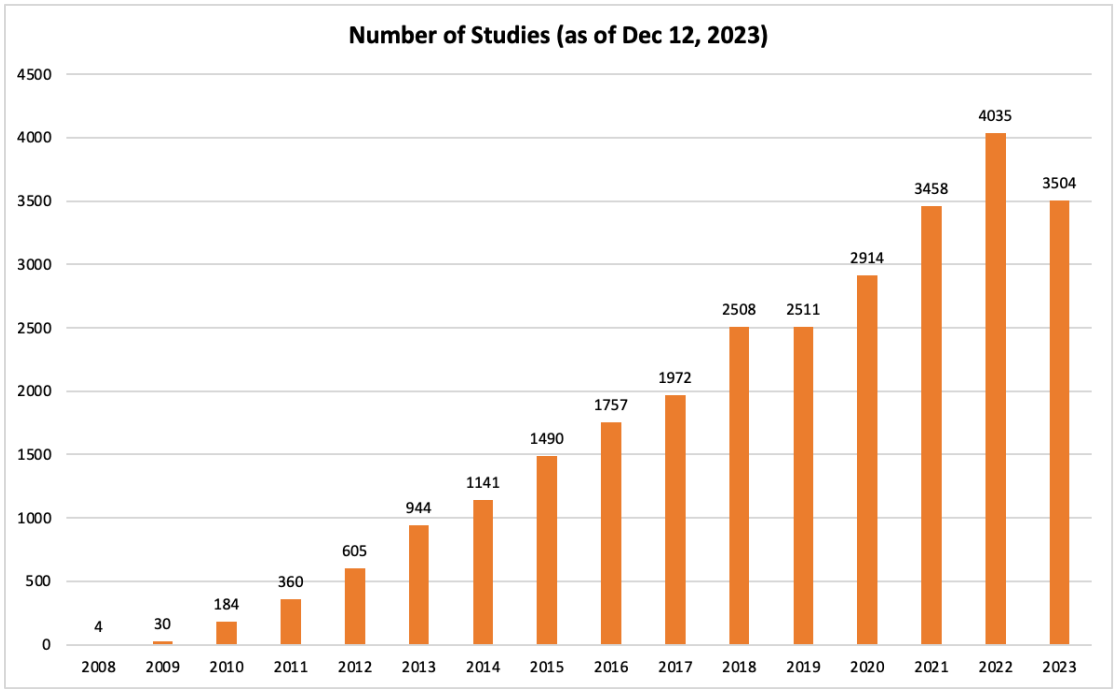
The distribution of study dates was computed in Excel (see Graph 1). The top 100 publications (ranked by number of published Twitter-based studies) were computed in R using the dplyr functions “group_by()” and “summarize()”. The top 10 publications were then extracted from this list and visualized in Graph 2. Next, we assigned disciplines by hand (see Appendix B) to each of the top 100 publications, calculated the percentage for each discipline, and visualized in Graph 3 using Excel. To identify the most influential studies, and to provide a secondary analysis of disciplines within the corpus (see Graph 4), we ranked the studies by their number of citations, and then labeled the top 100 studies’ disciplines by hand (see Appendix C). For the top 5[4] from each major discipline (Data Science, Social Science, Social Media, Public Health, Psychology, Information Science, Emergency Management, Education, Business, and Artificial Intelligence), we labeled main topics by hand. To accomplish this, we read each abstract, noting the main themes in each. We then grouped themes into common topics (see Appendix D). As a final step, we fed the top 100 abstracts into ChatGPT (25 at a time), and instructed, “Please give me the top 10 most common topics from this collection” (see Appendix E). This was used as a “second opinion,” and while some topics were agreed upon by both methods, others only showed up in one or the other. (See Discussion for details)

[3] Referred to by Crossref REST API as “is_referenced_by_count”
[4] Psychology, Information Science, Emergency Response, and Education had fewer than 5 studies in the top 100

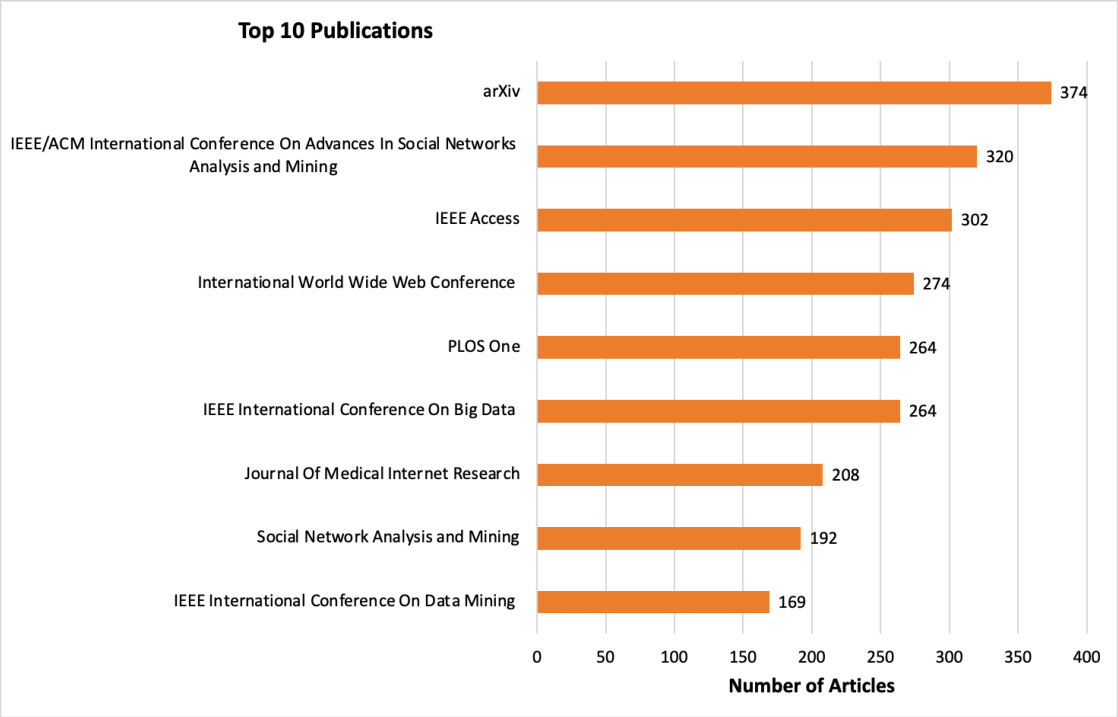
Results

This inquiry identified a total of 27,453 unique studies (link to Final Dataset) using and/or studying Twitter data. The studies were published in 7,432 distinct publications, with 1,303,142 citations, over a span of 17 years, and across 14 broad disciplines. Graph 1 shows the spread of published studies between the years 2008 and 2023. Graph 2 shows the top 10 publications, ranked by the number of Twitter-based studies they each have published.

Graph 1

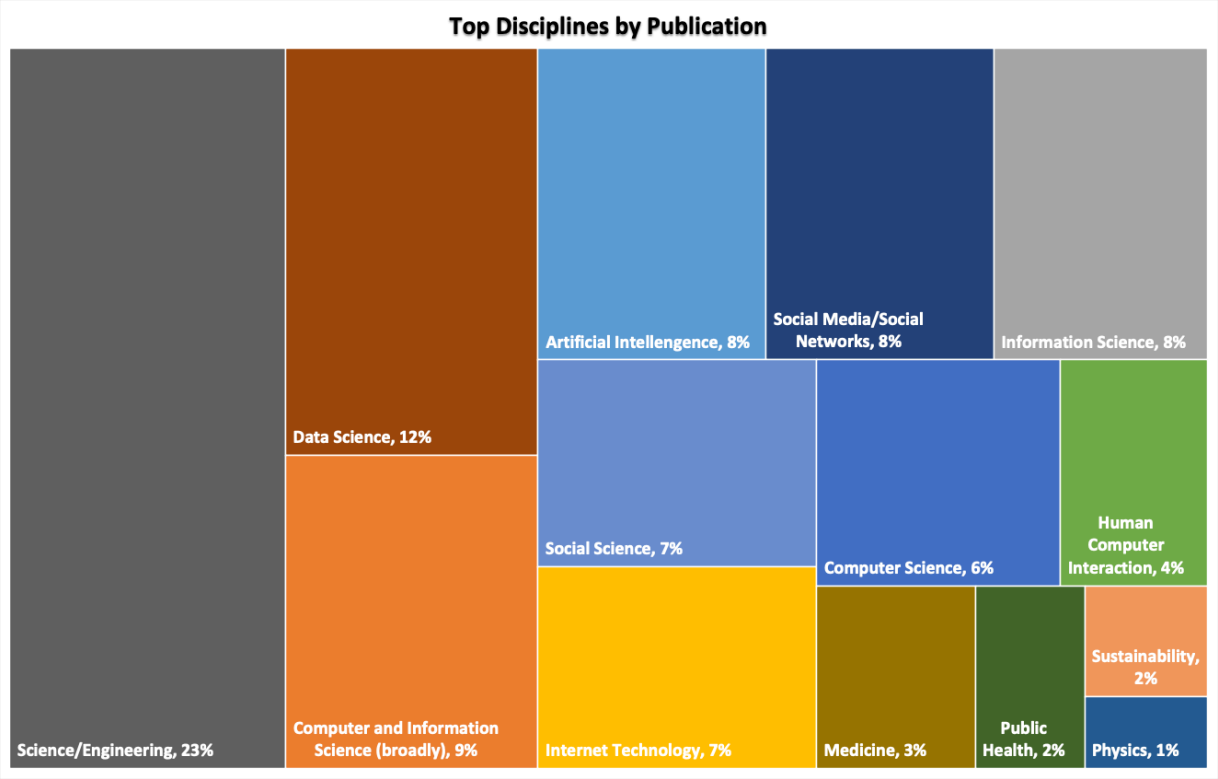


Graph 2 (see a full list of the top 100 publications in Appendix B):



To understand the spread of disciplines within the corpus, we took a two-pronged approach. First, Graph 3 shows the percentage of each discipline as determined by the top 100 publications' titles and/or website content (see Appendix B). Science/Engineering comprised 23% of the studies. Data Science comprised 12%, followed by Computer and Information Science at 9%. Artificial Intelligence, Social Media, and Information Science each comprised 8%, and the remaining 32% were shared between Social Science, Internet Technology, Computer Science, Human-Computer Interaction, Medicine, Public Health, Sustainability, and Physics.

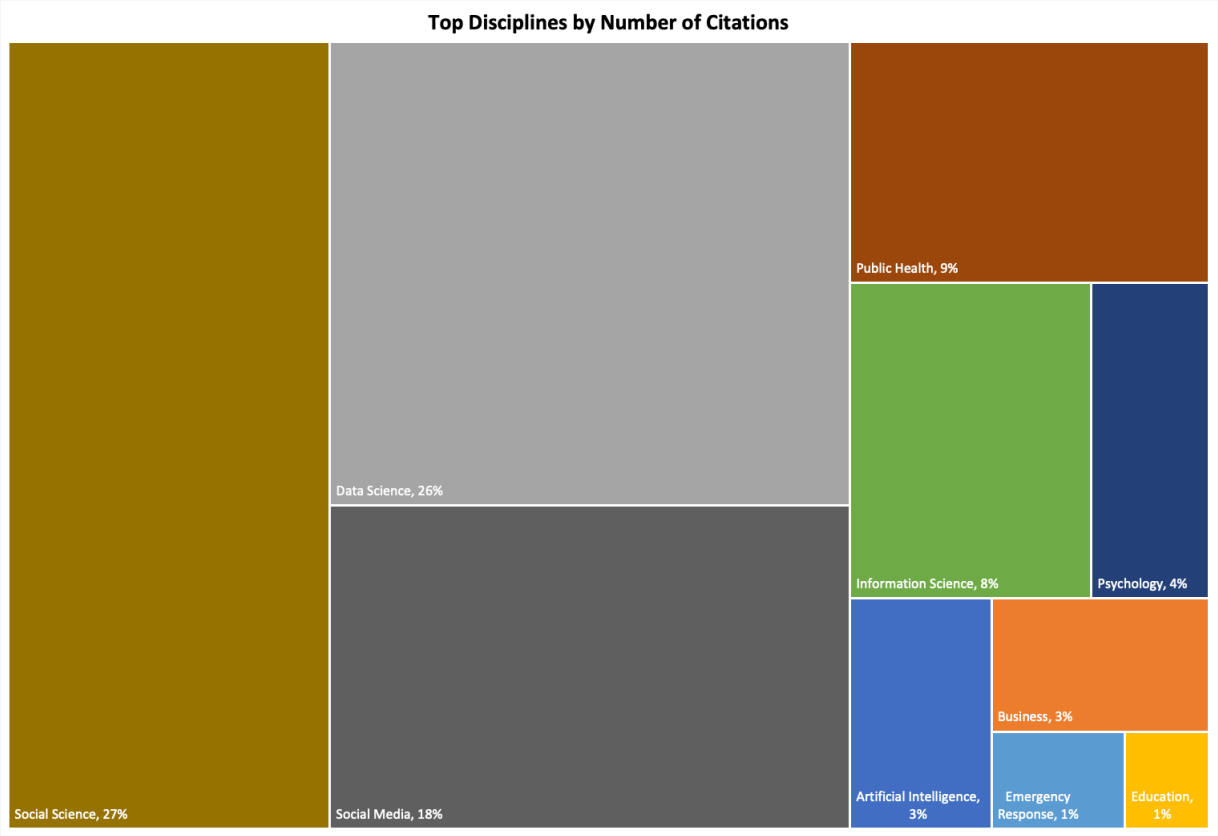
Graph 3



Second, Graph 4 shows the percentage of each discipline as determined by analyzing the top 100 most-cited studies (see Appendix C). In all, studies within the corpus were cited 1,303,142 times[5]. Within these top-cited studies, Social Science comprised the highest number of studies at 27%, with a total of 16,010 citations. Data Science comprised 26% of the studies with a total of 15,255 citations, Social Media Studies comprised 18% with a total of 10,625 citations, and the remaining 29% included: Public Health, Business, Artificial Intelligence, Psychology, Information Science, Education, and Emergency Response.

[5] It is important to note that this is only an approximation. There are many ways to count citations (ie. Crossref, Google Scholar), and each can differ considerably.

Graph 4



Graph 5 provides a brief analysis of the most common topics discussed within the corpus. Topics included: information dissemination, assessing the credibility of tweets, strategies for conducting data research, detecting and analyzing major events, and studying human behavior in its many forms.

Graph 5

Main Topics Identified by Hand	**Examples:**
Information dissemination	● Factors contributing to dissemination ● How does ideology impact dissemination? ● Identifying influential users

Assessing the credibility of tweets	• True vs fake news
Strategies for conducting data research	• Sentiment analysis • Topic modeling • Geolocating
Detecting and analyzing major events	• Pandemics • Earthquakes
Studying human behavior	• Political analysis and prediction • Misuse and misunderstanding of antibiotics • Marketing/promotion of consumer products • Stock market performance • Mental Health analysis

Discussion

It is clear that Twitter data has been widely and increasingly used in academic research for the past 17 years (Graph 1), and spans a wide swath of academic disciplines. We employed two distinct strategies to analyze the spread of disciplines within the corpus, resulting in two substantially different groupings. It is worth noting the difficulty we found in assigning disciplines. Several categories overlap (ie. Data Science, Computer Science, Computer and Information Science, and Internet Technology), and our assignments were subjective. Another researcher might assign different disciplines, thus ending up with different percentage spreads in Graphs 3 and 4.

Our first strategy (publication-based) focused on the overall disciplines of the top 100 publications (Graph 3). We analyzed the name of each top publication to determine the overall discipline[6], and in instances when the name alone was inconclusive, we explored the

[6] For example, the *International Journal Of Advanced Computer Science and Applications* was assigned to the Computer Science discipline.

organization's website to confirm the discipline[7]. We acknowledge that any one publication may contain a variety of disciplinary studies, therefore this strategy lacks specificity. At the same time, this strategy resulted in a wider array of disciplines than our study-based strategy, in particular the inclusion of Science/Engineering, and the differentiation between Data Science, Computer and Information Science, Computer Science, and Internet Technology. This may have resulted from the way publishers defined and grouped their publications.

In our second strategy (study-based), we focused on the studies themselves to determine disciplines, rather than on the publication names (Graph 4). To determine the spread of disciplines within the studies, we conducted a brief topic analysis using the title and abstract from the top 100 most-cited studies (Graph 5). We believe this strategy provides a more accurate analysis of which disciplines are most strongly represented. In these results, Social Science comprised 27% of the studies (up from 7% in the publication-based strategy), perhaps the result of Social Science studies being published in non-Social Science publications. Data Science comprised 26% of the studies (up from 12% in the publication-based strategy). We believe this difference may have occurred because Data Science studies were published in journals or conferences with overall disciplines of Computer Science, Computer and Information Science, or Internet Technology (see Appendices B and C for specific discipline assignments). Additionally, the study-based strategy illustrates the influence of this body of research in the academic world, with 1,303,142 citations.

Between 2008 and 2022 the number of studies increased annually. However, the number of studies decreased by 13% in 2023, once Twitter began charging for its API. We assume that much of the data used for studies published in 2023 were collected prior to Twitter's February, 2023 API shutdown, and thus the number of new studies are likely to decline even more into 2024 and beyond. We suggest a future study to collect these same data and

[7] For example, the journal name *Multimedia Tools and Applications* does not clearly state a discipline, thus further investigation into the journal's website was needed to reveal an overall discipline of Science/Engineering.

analyze in a similar manner to obtain a more clear picture of the long term impact of Twitter's API shutdown. This vulnerability demonstrates that researchers may have relied too heavily on Twitter data while it was freely available. We also wonder if researchers are currently paying too little attention to Twitter data now that it is no longer free, especially in light of the fact that it remains a major voice around the world for news and public opinion.

We wish to acknowledge three limitations to our study. While researchers use data from many different social media platforms (such as Facebook and Reddit), we chose to focus solely on studies utilizing Twitter data. Twitter has historically been the most common source of social media data (Tromble, 2021) due to its ease of use and open access, and thus offers the most comprehensive view into the topics and disciplines studied by researchers. We also acknowledge that we did not search every available database. We aimed to search the largest and most comprehensive databases covering the widest variety of disciplines. Some databases were excluded (ie. Academic Search Complete and PubMed) because all results were duplicates from other databases, while another (Communication Source) was excluded because it yielded a relevancy rate well below 80%. We were able to collect 36k studies from Engineering Village via the API, however the many hours required to download these data could hamper future researchers with limited time. Additionally, we did not search the full-text of papers. Instead, we searched metadata including: title, abstract, keywords, publisher, publication date, and others. While most of the databases we searched did not offer a full-text search option, two did (ACM Digital Library and IEEE Xplore), and a new search using the full-text field may produce additional papers.

Another area for further research is the practicality of using web scraping to collect Twitter data if the API remains out of reach. How does this impact the quality and quantity of studies as the time and cost of accessing and cleaning such data are significantly higher than when using the API? Additionally, web scraping has come under legal scrutiny in recent months. Twitter (now X Corp) has sued web scrapers in at least 3 separate cases, alleging that "data

scraping violates user privacy and that the process of scraping taxes the site to the point of causing systemic effects" (Robinson, 2023). Further research is needed to assess the impact of these and other legal challenges to researchers' abilities to obtain essential data.

Since 2006, Twitter's API has been a treasure trove of high-quality, freely available data. However, in the spring of 2023, Twitter began charging for access to its data. Lacking sufficient funds to pay this monthly fee, academics are now scrambling to continue their research. This paper illustrates the enormous value of freely available social media data across the academic disciplines, and highlights what will be lost if social media data continues to be cost prohibitive. A search of 8 databases and 3 related APIs found that since 2008, a total of 27,453 studies have been published in 7,432 distinct publications, with 1,303,142 citations, across 14 disciplines. Alarmingly, while Twitter data studies have increased every year since 2008, since Twitter's decision to begin charging researchers for data earlier this year, the number of studies has decreased by 13%. We suggest a future study to update these data in 2024 and 2025 to give a clearer picture of the long term impact of Twitter's API shutdown.

References

Brito, K. D. S., Filho, R. L. C. S., & Adeodato, P. J. L. (2021). A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions. *IEEE Transactions on Computational Social Systems, 8*(4), 819–843. 10.1109/TCSS.2021.3063660

Cascini, F., Pantovic, A., Al-Ajlouni, Y. A., Failla, G., Puleo, V., Melnyk, A., Lontano, A., & Ricciardi, W. (2022). Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature. *EClinicalMedicine, 48*, 101454–101454. 10.1016/j.eclinm.2022.101454

Golder, S. A., & Macy, M. W. (2011). Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science (American Association for the Advancement of Science)*, *333*(6051), 1878–1881. https://doi.org/10.1126/science.1202775

Huchinson, A. (2018, July 25). *Twitter continues its efforts to fight spam and misuse with new API restrictions*. Social Media Today, https://www.socialmediatoday.com/news/twitter-continues-its-efforts-to-fight-spam-and-misuse-with-new-api-restric/528533/

Jie Yin, Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using Social Media to Enhance Emergency Situation Awareness. *IEEE Intelligent Systems*, *27*(6), 52–59. https://doi.org/10.1109/MIS.2012.6

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and Research: A Systematic Literature Review Through Text Mining. *IEEE Access*, *8*, 1–1. https://doi.org/10.1109/ACCESS.2020.2983656

Ledford, H. (2023). Researchers scramble as Twitter plans to end free data access. *Nature (London)*, *614*(7949), 602–603. https://doi.org/10.1038/d41586-023-00460-z

Pandey, D., Wairya, S., Pradhan, B., & Wangmo. (2022). Understanding COVID-19 response by twitter users: A text analysis approach. *Heliyon, 8*(8), e09994–e09994. https://doi.org/10.1016/j.heliyon.2022.e09994

Robinson, B. (2023, August 17). *X Corp Lawsuits Target Data Scraping.* The National Law Review, Volume XIV(1). https://www.natlawreview.com/article/x-corp-lawsuits-target-data-scraping

Tromble, R. (2021). Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age. *Social Media + Society*, *7*(1). https://doi.org/10.1177/2056305121988929

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science (American Association for the Advancement of Science)*, *359*(6380), 1146-. https://doi.org/10.1126/science.aap9559

X. (2023a). Developer agreement and policy. X Developer Platform. https://developer.twitter.com/en/more/developer-terms/agreement-and-policy

X. (2023b). *Getting started: Twitter API access levels and versions*. X Developer Platform. https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

Appendix A

Database Search Strings and Sampling Results

Library and Information Science Source and *Library Information Science and Technology Abstracts*
https://www.ebsco.com/products/research-databases/library-information-science-and-technology-abstracts.
"Twitter N3 data OR twitter N3 api OR twitter N3 dataset" NOT survey, *filtered for conferences, journals, and magazines only *All Text for all fields

- 1608 results
- 82% relevance (50 paper sample)

Web of Science
https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/
"(twitter AND data) OR (twitter AND api) OR (twitter AND dataset) NOT (survey)"

- 10811 results
- 82% relevance (75 paper sample)

Global Health Database
https://www.ebsco.com/products/research-databases/global-health
"(twitter AND data) OR (twitter AND api) OR (twitter AND dataset)"

- 536 results
- 80% relevance (50 paper sample)

ACM Digital Library
https://dl.acm.org/
"(twitter AND data) OR (twitter AND api) OR (twitter AND dataset) NOT (survey)" *abstracts only

- 1950 results
- 92% relevance (50 paper sample)

IEEE Xplore
https://ieeexplore.ieee.org/Xplore/home.jsp
"Twitter NEAR/3 data OR twitter NEAR/3 api OR twitter NEAR/3 dataset NOT survey" *filtered out books

- 3509 results
- 97% relevance (50 paper sample)

Engineering Village API
https://dev.elsevier.com/
Query = (((((twitter AND data) OR (twitter AND api) OR (twitter AND dataset) NOT survey) WN ALL)) NOT (({ch} OR {ip} OR {bk} OR {er} OR {tb} OR {ed}) WN DT))
Compex Database

- 20,813 results
- 86% relevance (50 paper sample)

Inspec Database

- 15,013 results
- 88% relevance (50 paper sample)

Appendix B

Top 100 Publications and Their Assigned Disciplines

*Disciplines were assigned by examining the name of the journal or conference and/or the "About" section of the organization's website.

**Definitions used to assign disciplines:
Social Science: focus of study of human behavior, including psychology
Data Science: focus is on the tools for manipulating and analyzing the data
Computer Science: focus is on programing/designing the software tool
Social Media: focus is on understanding how the social media tool works, how the technology functions. Similar to Social Science, but related to how human behavior is impacted by the technology, more than purely human tendencies in society. How the tool functions and how it enables social interaction.
Internet Technology: a broad category of studying all things internet/world wide web
Computer and Information Science: this is the most broad category for things related to computers, data, and internet technology. We chose this discipline when the publication did not fit into a more specific category and seemed to encompass all aspects of the field.
Information Science: focus of study is on locating, accessing or organizing information.
Science/Engineering: focus of study is on a non-computer/information science field such as engineering (excluding physics).
Artificial Intelligence, **Human-Computer Interaction, Physics, Public Health, Medicine, Sustainability:** these remaining disciplines are easily distinguished by name of publication and/or organization's website.

***Numbers based on data collected in June 2023

Grouping into Disciplines	**Percentage**	**# of Studies**
Science/Engineering	23%	1629
Data Science	12%	842
Computer and Information Science (broadly)	9%	647
Artificial Intelligence	8%	610
Social Media/Social Networks	8%	581
Information Science	8%	546
Internet Technology	7%	461
Computer Science	6%	452

Human Computer Interaction	4%	272
Medicine	3%	237
Public Health	2%	163
Sustainability	2%	111
Physics	1%	72
Total		**7094**

Publisher Name	**# of Studies**	**Discipline by hand**
arXiv	374	Science/Engineering
IEEE/ACM International Conference On Advances In Social Networks Analysis A Mining	320	Social Media/Social Networks
IEEE Access	302	Science/Engineering
International World Wide Web Conference	274	Internet Technology
IEEE International Conference On Big Data	264	Data Science
PLOS One	264	Science/Engineering
Journal Of Medical Internet Research	208	Medicine
Social Network Analysis A Mining	192	Social Media/Social Networks
IEEE International Conference On Data Mining	169	Data Science
ACM International Conference On Information and Knowledge Management	138	Information Science
Expert Systems with Applications	134	Artificial Intelligence
ACM on Human Computer Interaction	122	Human Computer Interaction
Multimedia Tools and Applications	114	Science/Engineering
Sustainability	111	Sustainability
IEEE Transactions On Computational Social Systems	105	Social Science
IEEE Transactions On Knowledge and Data Engineering	102	Computer and Information Science

International Journal Of Environmental Research and Public Health	101	Public Health
International Conference On Advances In Computing Communications and Informatics	100	Computer and Information Science
Computers In Human Behavior	91	Human Computer Interaction
International Journal Of Advanced Computer Science and Applications	87	Computer Science
International Journal Of Geoinformation	84	Science/Engineering
ACM SIGKDD International Conference On Knowledge Discovery and Data Mining	82	Data Science
IEEE International Conference On Data Engineering	82	Computer Science
International AC|M Conference On Research and Development In Information Retrieval	82	Information Science
Information Processing and Management	79	Information Science
ACM International Conference On Web Search and Data Mining	78	Data Science
ACM Web Science Conference	75	Computer and Information Science
Information Communication & Society	75	Computer and Information Science
International Conference On Social Informatics	69	Social Science
Social Media + Society	69	Social Media/Social Networks
IEEE/WIC/ACM International Conference On Web Intelligence	68	Artificial Intelligence
Knowledge-based Systems	67	Artificial Intelligence
Applied Sciences	65	Science/Engineering
Public Heal and Surveillance	62	Public Health
Social Science Computer Review	62	Social Science
EPJ Data Science	61	Data Science
ACM Conference On Hypertext and Social Media	59	Human Computer Interaction
European Conference On Information Retrieval	55	Data Science
International Joint Conference On Neural Networks	53	Artificial Intelligence

Online Information Review	53	Information Science
Scientometrics	53	Social Science
Annual Meeting Of the Association For Computational Linguistics	51	Social Science
New Media & Society	50	Social Science
Government Information Quarterly	46	Information Science
Scientific Reports	46	Science/Engineering
Future Generation Computer Systems	44	Computer Science
Information Sciences	43	Computer and Information Science
International Conference On Social Computing, Behavioral-Cultural Modeling, & Prediction And Behavior Representation In Modeling And Simulation	43	Social Science
ACM Recommender Systems Conference	42	Artificial Intelligence
Information (Switzerland)	42	Computer Science
International Journal Of Information Management	42	Information Science
AIP Conference Proceedings	41	Physics
Electronics	41	Science/Engineering
Journal Of Big Data	41	Data Science
Journal Of Information Science	41	Information Science
Journal Of Intelligent And Fuzzy Systems	41	Science/Engineering
IEEE International Conference On Semantic Computing	40	Computer and Information Science
ASE/IEEE International Conference On Social Computing And Conference On Big Data	39	Computer and Information Science
ACM Symposium On Applied Computing	38	Computer Science
International Journal Of Disaster Risk Reduction	38	Science/Engineering
Neurocomputing	38	Artificial Intelligence
SSRN	38	Social Science
Information Systems Frontiers	37	Computer and Information Science
Online Social Networks And Media	37	Internet Technology

Pacific-Asia Conference On Knowledge Discovery And Data Mining	37	Data Science
ACM SIGSPATIAL International Conference On Advances In Geographic Information Systems	36	Science/Engineering
Journal Of The Association For Information Science And Technology	36	Information Science
Computers Materials And Continua	35	Computer and Information Science
IEEE International Conference On Systems Man And Cybernetics	35	Science/Engineering
International Journal Of Geographical Information Science	35	Science/Engineering
IOP Conference Series Materials Science And Engineering	35	Science/Engineering
ACM Transactions On Intelligent Systems And Technology	34	Artificial Intelligence
Concurrency And Computation: Practice And Experience	34	Computer Science
PEERJ Computer Science	34	Computer Science
Applied Intelligence	32	Artificial Intelligence
Applied Sciences-Basel	32	Science/Engineering
European Conference On Machine Learning and Principles And Practice Of Knowledge Discovery In Databases	32	Computer Science
Soft Computing	32	Artificial Intelligence
Future Internet	31	Internet Technology
International Conference On Web Information Systems Engineering	31	Internet Technology
Physica A: Statistical Mechanics and Its Applications	31	Physics
Big Data and Cognitive Computing	30	Internet Technology
Computers Environment and Urban Systems	30	Science/Engineering
Data In Brief	30	
Internet Research	30	Computer and Information Science
Journal Of Supercomputing	30	Computer Science

BMJ Open	29	Medicine
Decision Support Systems	29	Computer Science
EPI Profesional De La Información	29	Information Science
First Monday	29	Internet Technology
IEEE Global Communications Conference	29	Science/Engineering
World Wide Web	29	Internet Technology
ACM Transactions On Knowledge Discovery From Data	28	Data Science
Computational Intelligence and Neuroscience	28	Artificial Intelligence
International Conference On Intelligent Computing and Control Systems	28	Artificial Intelligence
Technological Forecasting and Social Change	28	Computer and Information Science
Transactions In GIS	28	Science/Engineering
International Conference On Information Integration and Web-based Applications and Services	27	Data Science
International Joint Conference On Knowledge Discovery Knowledge Engineering and Knowledge Management	27	Artificial Intelligence
Journal Of Ambient Intelligence and Humanized Computing	27	Artificial Intelligence

Appendix C

Assigning Disciplines to 100 Most-Cited Studies

**For detailed discipline assignments see Final Dataset (Top Disc by Most Cited tab)
**Disciplines were assigned by reading titles and abstracts and applying the following definitions:

Disciplines:	**Definitions used to assign**
Social Science	A focus on social human behavior. How do people behave? This is often in the context of social media in this corpus, but always with a focus on human behavior.
Data Science	A focus on the mechanics of data and information, ie. sentiment analysis, topic modeling, other tools for programming and data analysis.
Social Media	A focus is on the technology itself, ie. studying or designing a tool used to identify influential users on Twitter. How does information spread on Twitter? Looks at how the technology works, rather than how people think and act.
Public Health	A focus on using Twitter to study public health issues, often looking at how a public health concern is discussed in tweets and retweets.
Business	A focus on business, marketing, promotions
Artificial Intelligence	A focus on the creation of AI or the uses of AI in social media and other technologies.
Psychology	A focus on human psychology within the context of Twitter and tweets. Often examining tweets to better understand a specific psychological question.
Info Science	A focus on finding, organizing, and making data available
Education	A focus on the use of Twitter in educational settings.
Emergency Response	A focus on using Twitter to detect and respond to natural disasters, pandemics, and other emergencies.

Appendix D

Main Topics Identified by Hand

*Taken from Final Dataset, bottom of Top Disc by Most Cited tab

Topics taken from the top 5 studies for each discipline category:

The spread of true vs false news

Factors contributing to the spread of content on Twitter

The practice of retweeting on Twitter

How do ideological preferences impact the exchange of information?

How do emotions impact retweets? Info diffusion

Identifying influential users on Twitter

automatic methods for assessing the credibility of tweets.

Using Twitter to follow and analyze public attention (ie. During a pandemic)

Analyzing privacy in social media.

What causes certain content to be retweeted more than others?

Strategies for tracking health concerns, outbreaks, understanding health information disparities in communities

Studying misuse and misunderstanding of the use of antibiotics

A study of how twitter data is used in health research.

Studying the marketing and promotion of an e-cigarette from a public health perspective

Using Twitter to identify influenza outbreaks faster

Efficacy of conducting psychological tests online

Using twitter to identify diurnal and seasonal mood rhythms across the world

Efficacy of conducting psychological tests online

Using tweets to understand consumer sentiment toward brands.

Proposes a way of classifying tweet content to mitigate user overwhelm from too much content

Using geospatial data from twitter to identify and track events such as earthquakes

Understanding how social media can be used in higher education settings. A review of the literature.

A general study of what twitter is, how and why people us it.

Strategy for geolocating users based on tweet content

A new strategy for topic-modeling. How to train the topic model to a specific dataset and achieve more accurate topics.

What new technologies need to be developed to support the field of genomics?

Strategy for sentiment analysis

Comparing twitter, FB and Youtube for understanding and influencing consumer brand communication. Preferences?

Text mining strategy for analyzing customer sentiments about 3 pizza chains

Understanding how social media impacts effectuation

Strategy for text mining consumer sentiments toward specific brands

Research about how social media can be used to support the supply chain management field: practices, networking, stakeholder engagement, demand shaping, product development

Strategies for using AI in sentiment/opinion analysis

Using AI to identify fake news on social media..a specific framework

Using AI for sentiment analysis

Using sentiment analysis to help business organizations gain business insight into consumer opinions

Strategy for using AI in sentiment analysis of tweets.

Most common topics:

How information spreads on social media (ie. True vs fake news, factors contributing to retweeting/spread
How does ideology impact spread?
Identifying influential users, methods for assessing credibility of tweets,

Identifying and analyzing major events (pandemic, earthquake, influenza,
Studying behavior such as misuse and misunderstanding of antibiotics, the marketing/promotion of consumer products

Studying moods cross culturally (diurnal and seasonal)

Appendix E

Main Topics Identified by ChatGPT

Topic	Explanation
Sentiment Analysis	Many abstracts discuss sentiment analysis, which involves analyzing the emotions, opinions, and sentiments expressed in social media posts.
Social Media Impact	Several abstracts explore the impact of social media, including its effects on entrepreneurship, stock market performance, and disaster response.
Effectuation Processes[8]	Entrepreneurial effectuation processes are a topic of interest, with discussions on how they are influenced by social media, such as Twitter.
Fake News Detection	Fake news detection is a recurring theme, especially in the context of social media platforms like Twitter.
Media Influence	The influence of various forms of media, including conventional media and social media, is discussed in relation to stock market performance and brand sentiment.
Deep Learning	Deep learning techniques are highlighted as an approach for sentiment analysis and feature extraction from social media data.
Text Analysis	Text analysis, including text mining and natural language processing, is a key methodology used in many of the abstracts to extract insights from social media content.
Event Detection	Event detection and categorization in social media, particularly Twitter, is a focus in some abstracts, including disaster-related events.
Community and Interaction	Topics related to community orientation, interaction, and social relationships on social media platforms are discussed in the context of sentiment analysis and event detection.

[8] Only mentioned once in the top 100 studies.