Data in Brief 62 (2025) 111960

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A large-scale dataset of AI-related tweets: Structure and descriptive statistics

Nathalie de Marcellis-Warin[a], Daniel Kouloukoui[b,c,*], Thierry Warin[d]

[a] Polytechnique Montréal, Full Professor at Department of mathematics and industrial engineering, 2900 Edouard-Montpetit Boul, Montreal, Quebec, H3T1J4, Canada.
[b] Federal University of Bahia, Professor at Department of Accounting Sciences, Av. Reitor Miguel Calmon, s/n - Canela - Salvador/Bahia, CEP 40.110-100, Brazil
[c] Polytechnique Montréal, Postdoctoral fellow at Department of mathematics and industrial engineering, 2900 Edouard-Montpetit Boul, Montreal, QC, H3T1J4, Quebec, Canada
[d] HEC Montréal, Full Professor at Department of International Business, 3000 de la Cote-Sainte-Catherine Ch, Montreal, Quebec, H3T2A7, Canada.

ARTICLE INFO

Article history:
Received 5 February 2025
Revised 24 July 2025
Accepted 30 July 2025
Available online 6 August 2025

Dataset link: tweets_ai.qs (Original data)
Dataset link: tweets_ai.csv (Original data)

Keywords:
Artificial intelligence
Social media analysis
Twitter data
Natural language processing
Public perception
AI Ethics
Topic Modeling
Sentiment Analysis

ABSTRACT

This article presents a curated and anonymized dataset of tweets related to artificial intelligence (AI), comprising 893,076 entries collected using the Twitter API between January 1, 2017, and July 19, 2021. These tweets were extracted from a larger initial corpus using the keyword "Artificial Intelligence" and subsequently filtered to ensure data quality, multilingual coverage, and public accessibility. The final dataset includes structured metadata such as media elements (images, videos, and URLs), user engagement metrics (likes, retweets, replies), hashtags, language codes, and temporal indicators (hour and weekday of posting). While additional linguistic features—such as text length and tokenization—were used in internal analyses, they are not included in the public release. This dataset offers a robust foundation for research on the evolution of public discourse surrounding AI, including sentiment analysis, topic modeling, social engagement dynamics, and policy-relevant evaluations. It is openly available through established repositories and adheres to the FAIR principles, facilitating transparency, reproducibility, and

* Corresponding author.
E-mail addresses: nathalie.demarcellis-warin@polymtl.ca (N. de Marcellis-Warin), danielkoulou@hotmail.com (D. Kouloukoui), thierry.warin@hec.ca (T. Warin).

https://doi.org/10.1016/j.dib.2025.111960

interdisciplinary applications in computational social science, natural language processing, and AI governance research.

Specifications Table

Subject	Social Sciences
Specific subject area	Social Media Analytics, Natural Language Processing, Data Science
Type of data	Structured text data (Raw)
Data collection	The dataset was collected using a social media–based research design, leveraging the Twitter API to extract tweets related to artificial intelligence (AI) posted between January 1, 2017, and July 19, 2021. A total of 893,076 tweets are publicly available in this release, representing a curated and filtered subset of a larger initial corpus retrieved using the keyword "Artificial Intelligence." Data extraction was performed using R (version 3.6.1) and the Python package Tweepy. The dataset includes metadata on user interactions (likes, retweets, replies), hashtags, media elements (images, videos, URLs), and temporal and linguistic attributes (timestamps, language, and device source). It enables longitudinal analyses of AI-related discourse, including trend detection, sentiment classification, and public perception mapping on topics such as risk, ethics, and governance.
Data source location	Data Source Location Institution: Polytechnique Montréal HEC Montréal City/Town/Region: Montréal, Québec, Canada Country: Canada Latitude and Longitude (Data Storage and Processing Centers): 45.5051° N, 73.6188° W Primary Data Sources: The dataset was collected from Twitter (now X) using the Twitter API through R- and Python-based tools. The resulting data were processed and stored in a secure research infrastructure maintained by HEC Montréal.
Data accessibility	Repository names: Harvard Dataverse and GitHub CSV format (Harvard Dataverse):https://doi.org/10.7910/DVN/NHLEJL QS format (GitHub): https://github.com/warint/tweets-ai
Related research article	The dataset presented in this Data in Brief article is derived from the same original corpus used to support the analyses in Kouloukoui, de Marcellis–Warin, and Warin [1]. For the purposes of public release, it has been carefully anonymized and reduced in size to ensure responsible data sharing and broad reusability: Kouloukoui, D., de Marcellis-Warin, N., & Warin, T. [1]. Balancing risks and benefits: public perceptions of AI through traditional surveys and social media analysis. *AI & SOCIETY*, 1–24. https://doi.org/10.1007/s00146–025–02232-x

1. Value of the Data

This dataset offers a significant contribution to social media analytics, with particular relevance for research on public discourse surrounding artificial intelligence (AI). Its structured metadata and broad temporal coverage support a variety of analytical strategies, ranging from time-series forecasting to deep learning approaches for text classification. The dataset comprises 893,076 curated tweets and is publicly available through multiple open repositories, in alignment with the FAIR principles—ensuring that the data are Findable, Accessible, Interoperable, and Reusable.

To accommodate diverse analytical workflows and technical preferences, the dataset is provided in two complementary formats: (i) CSV files, a standard tabular format compatible with

spreadsheet software, statistical packages, and text editors, hosted on Harvard Dataverse [DOI: 10.7910/DVN/NHLEJL]; and (ii) QS files, a lightweight binary format optimized for R-based processing, hosted on GitHub. Each format is accompanied by detailed documentation included in this Data in Brief article, comprising variable definitions, content descriptions, and methodological annotations designed to support transparency, replicability, and broad reuse.

This dataset constitutes a robust foundation for multiple lines of inquiry in the computational social sciences, natural language processing, and AI governance. It enables—but is not limited to—the following six analytical domains:

- Analysis of Trends and Temporal Evolution of Discourse on AI. This dataset enables the investigation of how discussions related to artificial intelligence have evolved over time. By leveraging its temporal granularity, researchers can identify emerging patterns and explore the relationship between technological developments, regulatory shifts, and public debate [2]. Temporal analysis also provides insight into how political events, innovations, and crises shape collective perceptions of AI.
- Sentiment and Emotion Analysis. Natural Language Processing (NLP) techniques such as sentiment analysis and emotion classification can be applied to assess public attitudes toward AI. This facilitates the examination of socio-ethical concerns, including issues of privacy, trust, and regulation [3]. The inclusion of timestamp metadata further allows for dynamic analyses of sentiment over time.
- Study of Engagement Dynamics. The dataset captures detailed interaction metrics (retweets, likes, replies), enabling analysis of the mechanisms that drive information diffusion in AI-related conversations. Such engagement data help explain how certain narratives gain prominence and how public attention responds to ethical, regulatory, or technological themes [4]. These metrics are particularly useful for studying message virality and influence propagation.
- Social Network and Interaction Analysis. By reconstructing interaction graphs from conversation identifiers and engagement features, this dataset supports the application of Social Network Analysis (SNA) techniques. These include the identification of influencers, the mapping of discursive communities, and the detection of misinformation flows and algorithmic bias [5]. Network measures such as centrality and modularity allow for the examination of echo chambers, polarization, and structural cohesion.
- Multimodal Content Analysis. The inclusion of media attributes—such as images, videos, and URLs—permits the study of multimodal communication in the context of AI discourse. Existing research suggests that visual and linked content shapes message reception and audience engagement [6]. This opens avenues for exploring cross-platform dynamics and the visual framing of AI narratives.
- Cross-Validation with Survey Data and Policy Relevance. The dataset has already been employed in conjunction with traditional survey data to explore methodological complementarities in assessing public perception of AI, as demonstrated in the authors' peer-reviewed publication in AI & Society [1]. This underscores its relevance for research in AI governance, particularly in relation to trust, regulation, education, and algorithmic transparency.

2. Background

The increasing need to understand how the public perceives AI through social media platforms—particularly Twitter—has motivated the construction of this large-scale dataset focused on AI-related discourse. A growing body of academic literature has emphasized the disconnect between societal awareness of the risks and benefits of AI and the pace of technological development. In complement to traditional methods such as surveys and structured interviews, social media-based public opinion research offers a valuable means of capturing unprompted and organic narratives [7]). Prior studies have shown that the analysis of attitudes and opinions expressed on Twitter can yield significant insights into emerging topics and public sentiment [8]. Moreover, evidence suggests that users may attribute greater credibility to AI-generated tweets

than to those authored by humans, further underscoring the importance of monitoring and analyzing these exchanges [9].

Numerous large-scale Twitter datasets have been developed to support research in computational communication and the social data sciences. For example, Fafalios et al. [10] compiled an RDF corpus of over 1.5 billion semantically annotated tweets, facilitating semantic web applications and knowledge discovery. Guerrero-Contreras et al. [11] contributed a curated dataset of >4000 Spanish-language tweets about AI, enriched with sentiment annotations and user metadata to support public opinion research. Similarly, Dimitrov et al. [12] produced a comprehensive knowledge base of COVID-19-related tweets annotated for entities, hashtags, sentiment, and temporal metadata, enabling the study of discourse evolution and misinformation flows. Furthermore, several studies have employed Twitter datasets as the basis for analyses, as demonstrated by Kouloukoui et al. [13], Warin & Stojkov [14], and Costa et al. [15]. These works highlight an increasingly prominent trend in the use of data from this platform for scientific research.

By offering a longitudinal perspective on public discourse concerning AI—spanning its perceived promises and risks—this dataset contributes to and extends these efforts. It enables researchers to interrogate the social dimensions of AI through multilingual, multimodal, and temporally anchored data, supporting a wide range of inquiries in computational social science and AI ethics.

3. Data Description

The dataset covers the period from January 1, 2017, to July 19, 2021, and consists of a curated and anonymized sample of tweets related to artificial intelligence. It includes structured metadata at the tweet level, such as timestamp, language, timezone offset, and user engagement metrics (number of replies, retweets, and likes). Content-specific indicators such as hashtags, cashtags, and embedded URLs are also included, along with binary variables capturing the presence of media elements—namely images, videos, quoted tweets, and thumbnails.

3.1. Dataset variables

The publicly released version comprises 21 variables, selected from an original 36-variable structure. These include 19 original fields and two additional temporal variables—hour and wday—derived to support fine-grained temporal analysis. The dataset retains the full tweet text and is formatted to facilitate reuse in both qualitative and quantitative workflows. All personally identifiable user information has been removed, and user IDs are not included in accordance with data redistribution constraints. The complete list of variables, along with definitions, is provided in Table 1. These variables have been curated to maximize usability, methodological transparency, and reproducibility across a range of research domains.

4. Experimental Design, Materials and Methods

This dataset was developed to support large-scale, reproducible, and longitudinal analyses of public discourse on artificial intelligence (AI). It comprises 893,076 tweets collected from Twitter (now X) between January 1, 2017, and July 19, 2021, using the Twitter API. Data collection was conducted in two distinct periods—2017–2018 and 2019–2021—based on the keyword “Artificial Intelligence.” The corpus was filtered to remove duplicates and empty records, resulting in a structured dataset suitable for public dissemination and advanced analysis.

The dataset includes tweet-level metadata (e.g., date, time, timezone, language), user engagement metrics (likes, retweets, replies), content indicators (hashtags, cashtags, URLs), and media

Table 1
Variables and their descriptions.

Variable	Description
conversation_id	Unique identifier for the conversation thread to which the tweet belongs
created_at	Full timestamp (including timezone) of when the tweet was posted
date	Date of the tweet in YYYY-MM-DD format
time	Time of the tweet in HH:MM:SS format
timezone	Offset in minutes from UTC for the tweet's timestamp
tweet	Tweet text
language	Language of the tweet using ISO 639–1 code (e.g., "en" for English)
mentions	Mentions of @, #, and urls
urls	List of URLs included in the tweet
photos	List of image URLs embedded in the tweet (if any)
replies_count	Number of replies received by the tweet
retweets_count	Number of times the tweet was retweeted
likes_count	Number of likes received by the tweet
hashtags	List of hashtags used in the tweet
cashtags	List of cashtags (e.g., stock symbols) used in the tweet
retweet	Boolean value indicating whether the tweet is a retweet (TRUE/FALSE)
quote_url	URL of the quoted tweet, if applicable
video	Binary indicator for presence of video (1 = yes, 0 = no)
thumbnail	URL of the video thumbnail or preview image (if available)
hour	Hour of the day
wday	Weekday

Source: The Authors, 2025.

elements (images, videos, thumbnails, quote URLs). The final public version includes 21 variables, selected from an initial 36-variable structure, and retains the full text of each tweet. All personal identifiers have been removed, and no user IDs are included.

Data processing was conducted using both R and Python. The qs package was used for efficient storage and export in R, while additional preprocessing relied on standard Python tools. The dataset is available in two formats: a CSV version hosted on Harvard Dataverse and a lightweight QS version hosted on GitHub. Both formats are accompanied by detailed documentation to support replicability and interoperability in line with FAIR data principles.

4.1. Summary statistics and engagement metrics of this dataset

The median tweet date is April 20, 2019. On average, tweets include one hashtag, with a median of one as well. Engagement distributions are highly skewed. For replies, the standard deviation is 3.58, while the interquartile range (IQR) and 90th percentile are both zero, indicating that over 90 % of tweets receive no replies at all. This reflects a heavily right-skewed distribution, where conversational interaction is rare.

In the case of retweets, the standard deviation is 84.5, despite an IQR of 1 and a 95th percentile of 6, suggesting that most tweets are retweeted only once or not at all, but a small subset generates significant amplification. Likes show even more pronounced skew: with a standard deviation of 249 and a 95th percentile of 7, a minority of tweets attract disproportionately high attention. Gini coefficients further confirm the inequality of distribution—0.899 for retweets and 0.908 for likes—highlighting the concentration of social engagement among a small number of posts.

To characterize interaction dynamics more precisely, the like–retweet ratio was computed for tweets with at least one retweet. The median ratio is 0.875, indicating that likes tend to slightly lag behind retweets. The interquartile range of 1.33 reflects significant variability in how audiences engage with content, depending on message type, timing, and network diffusion [16].

4.2. Conversational structure and temporal patterns

To examine the structure of conversational exchange, tweets were grouped by conversation_id. The average thread size is 1.02 tweets, with only 1.95 % of tweets occurring in multi-message threads. The largest observed conversation reached 189 tweets. These findings reinforce prior claims that Twitter operates primarily as a broadcast platform, with limited sustained dialogue [17,18].

The temporal dimension of the dataset reveals additional structural patterns. The average timezone offset is –408 min (UTC–6h48), indicating that a large proportion of tweets originate from regions in or near the Central Time Zone of North America. The standard deviation of 26.3 min suggests limited dispersion across adjacent time zones.

By dividing the corpus into four equal groups based on timezone offset (tz_quartile), a clear gradient in engagement becomes visible. Tweets in the earliest timezones (Quartile 1) receive on average 0.0935 replies, 1.67 retweets, and 1.92 likes. Engagement peaks in Quartile 4 (latest timezones), with 0.158 replies, 2.06 retweets, and 3.71 likes. The second quartile shows the lowest interaction levels, while the third recovers modestly. These results suggest that both the timing and geographic origin of tweets influence their likelihood of generating audience response.

Limitations

While this dataset offers a valuable foundation for analyzing public discourse on artificial intelligence (AI) via social media, several limitations should be noted:

- Selection Bias: Data collection was based on the keyword “Artificial Intelligence,” which may have excluded relevant discussions that used alternative terms, acronyms, or conceptually related expressions. As such, the dataset may not capture the full lexical range of AI-related discourse.
- Language Representation: Although the dataset includes tweets in multiple languages, the distribution is uneven and skewed toward English. This limits the comparative analysis of AI-related discourse across linguistic and cultural contexts and may underrepresent perspectives from non-English-speaking regions.
- Engagement Bias: Likes, retweets, and replies are indicators of visibility rather than representative opinion. These engagement metrics are often influenced by user popularity and network effects, meaning that highly visible accounts may disproportionately shape the apparent salience of particular narratives.

Ethics Statement

Research based on social media data demands careful adherence to ethical guidelines that protect users and uphold principles of responsible data science [19]. In this study, the collection and analysis of tweets related to AI were conducted in accordance with established standards for research in digital environments. Particular attention was given to the ethical curation of public content and to minimizing risks of harm.

The dataset consists exclusively of publicly available tweets collected via the Twitter API. In accordance with recent scholarship, the analysis of public social media content may be conducted ethically without explicit user consent, provided that no personally identifiable information is disclosed [20]. In preparing the public version of the dataset, we undertook substantial efforts to remove personal metadata—such as usernames, user IDs, profile images, and other potentially identifying fields—following guidance from the research ethics literature [21,22].

The dataset retains the full text of public tweets, in keeping with Twitter’s terms of use, which allow the sharing of tweet content for research purposes so long as the tweets remain

publicly accessible and are not accompanied by identifiable user information. While reidentification cannot be ruled out entirely in social media research, the dataset has been curated to mitigate such risks to the greatest extent practicable.

All methodological procedures align with established best practices for research involving social network data and reflect a commitment to ethical accountability [23]. In line with contemporary recommendations for transparency and user protection in social media research [2], explicit consent was not required, given the public nature of the data and the safeguards implemented during dataset preparation.

Funding

This work was supported by Fonds de Recherche Nature et Technologies Québec (FRQNT), Programme: Bourses d'excellence pour étudiants étrangers (PBEEE) / Bourses de stage postdoctoral (Grant numbers [2023–2024 - V2 - 335403] and Polytechnique Montréal, Programme de formation FONCER du CRSNG / Mine Intelligente et Autonome - MIA (Grant numbers [0000]).

Data Availability

tweets_ai.qs (Original data) (GitHub)
tweets_ai.csv (Original data) (Dataverse)

CRediT Author Statement

Nathalie de Marcellis-Warin: Conceptualization, Methodology, Investigation, Resources, Funding acquisition, Supervision, Project administration, Writing – review & editing; **Daniel Kouloukoui:** Methodology, Formal analysis, Writing – review & editing; **Thierry Warin:** Writing – original draft, Investigation, Supervision, Methodology, Validation, Data curation, Software, Visualization.

Acknowledgements

The authors acknowledge the role of Twitter API in data collection and emphasize compliance with ethical guidelines in handling social media data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Kouloukoui, N. de Marcellis-Warin, T. Warin, Balancing risks and benefits: public perceptions of AI through traditional surveys and social media analysis, AI. Soc. (2025) 1–24.
[2] J. Pfeffer, A. Mooseder, J. Lasser, L. Hammer, O. Stritzel, D. Garcia, *This sample seems to be good enough! assessing coverage and temporal reliability of Twitter's academic API* (arXiv:2204.02290), ArXiv. (2023), doi:10.48550/arXiv.2204.02290.
[3] Camacho-Collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., Boisson, J., Espinosa-Anke, L., Liu, F., Martínez-Cámara, E., Medina, G., Buhrmann, T., Neves, L., & Barbieri, F. (2022). *TweetNLP: cutting-edge natural language processing for Social Media* (arXiv:2206.14774). ArXiv. https://doi.org/10.48550/arXiv.2206.14774

[4] L. Nassi-Calò, Promovendo e Acelerando o Compartilhamento De Dados De Pesquisa | SciELO Em Perspectiva, 2019 https://blog.scielo.org/blog/2019/06/13/promovendo-e-acelerando-o-compartilhamento-de-dados-de-pesquisa/.
[5] C.Z. Silveira, T.M.R. Días, O reúso de dados de pesquisa na perspectiva da Ciência da informação, Biblios J. Librarianship Inform. Sci. 86 (2023) 41–57.
[6] P. Meirelles, Pesquisa acadêmica com dados de mídias sociais: por onde começar? Insightee (2021) https://insightee.com.br/blog/pesquisa-academica-com-dados-de-midias-sociais-por-onde-comecar/.
[7] T. Warin, A. Stojkov, Decoding" Policy perspectives: structural topic modeling of European Central bankers' Speeches, J. Risk. Financ. Manage 16 (7) (2023) 329.
[8] L.A.D.S. Dourado, A. Fernandes, V.D.M. Siqueira, G.S.S. Guarienti, F.S.D Oliveira, Análise de sentimentos de tweets em português-Brasileiro Utilizando Inteligência artificial, in: Tecnologia Da Informação e Comunicação: Pesquisas em Inovações Tecnológicas—Volume 2, 1o, Editora Científica Digital, 2022, pp. 54–71, doi:10.37885/220308133.
[9] G. Spitale, N. Biller-Andorno, F. Germani, AI model GPT-3 (dis)informs us better than humans, Sci. Adv. 9 (26) (2023) eadh1850, doi:10.1126/sciadv.adh1850.
[10] P. Fafalios, V. Iosifidis, E. Ntoutsi, S. Dietze, Tweetskb: a public and large-scale rdf corpus of annotated tweets, in: European Semantic Web Conference, Springer International Publishing, Cham, 2018, pp. 177–190.
[11] G. Guerrero-Contreras, S. Balderas-Díaz, A. Serrano-Fernández, A. Muñoz, IA tweets analysis dataset (Spanish), IA Tweets Anal. Dataset (Spanish) (2024).
[12] D. Dimitrov, E. Baran, P. Fafalios, R. Yu, X. Zhu, M. Zloch, S. Dietze, Tweetscov19-a knowledge base of semantically annotated tweets about the covid-19 pandemic, in: Proceedings of the 29th ACM international conference on information & knowledge management, 2020, October, pp. 2991–2998.
[13] D. Kouloukoui, N. de Marcellis-Warin, S.M. da Silva Gomes, T. Warin, Mapping global conversations on twitter about environmental, social, and governance topics through natural language processing, J. Clean. Prod. 414 (2023) 137369.
[14] T. Warin, A. Stojkov, Discursive dynamics and local contexts on Twitter: the refugee crisis in Europe, Disc. Commun. 17 (3) (2023) 354–380, doi:10.1177/17504813231155739.
[15] A. D. J. B. Costa, S. M. D. S. Gomes, D. Kouloukoui, N. De Marcellis-Warin, T. Warin, Twitter conversations on sustainable development goals in Brazilian public universities using natural language processing, Dis. Susta. 4 (1) (2023) 51 https://doi.org/10.1007/s43621-023-00170-6.
[16] R.R. Wilcox, Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy, Springer, New York, NY, 2010.
[17] C. Honeycutt, S.C. Herring, Beyond microblogging: conversation and collaboration via Twitter, in: *Proceedings of the 42nd Hawaii International Conference on System Sciences* (HICSS), IEEE, 2009, pp. 1–10, doi:10.1109/HICSS.2009.89.
[18] A. Java, X. Song, T. Finin, B. Tseng, Why we twitter: understanding microblogging usage and communities, in: *Proceedings of the Joint 9th WebKDD and 1st SNA–KDD Workshop on Web Mining and Social Network Analysis*, ACM, 2007, August, pp. 1–10, doi:10.1145/1348549.1348552.
[19] L. Townsend, C. Wallace, Social media research: a guide to ethics, Univ. Aberdeen 1 (16) (2016) 1–16.
[20] C. Fiesler, N. Proferes, "Participant" perceptions of Twitter research ethics, Soc. Media Soc. 4 (1) (2018) 2056305118763366, doi:10.1177/2056305118763366.
[21] K. Beninger, Social media users' views on the ethics of social media research, in: The SAGE Handbook of Social Media Research Methods, 2017, p. 1. https://www.google.com/books?hl=pt-BR&lr=&id=9oewDQAAQBAJ&oi=fnd&pg=PA57&dq=Social+Media+Users%E2%80%99+Views+on+the+Ethics+of+Social+Media+Research.&ots=ePKUq1qUyQ&sig=CtgigXnQq_kvegqEzCg4L1vHYGg.
[22] M.J. Salganik, Bit By bit: Social research in the Digital Age, Princeton University Press, 2019 https://www.google.com/books?hl=pt-BR&lr=&id=58iXDwAAQBAJ&oi=fnd&pg=PR1&dq=Bit+by+bit:+Social+research+in+the+digital+age&ots=0SiRw6g-ai&sig=Fdl3xGZljqpE-5hUwBqJd886Nzk.
[23] A. Markham, E. Buchanan, Recommendations from the AoIR Ethics Working Committee (Version 2.0), 2012.