Profiles of Individual Radicalization in the United States (PIRUS) Datasheet

I. Motivation

I-A For what purpose was the dataset created?

The PIRUS dataset was created to better understand domestic radicalization. The dataset contains information on individuals in the United States that have been radicalized between 1948 and 2018.

I-B Who created the dataset?

Is it an official law enforcement or government body? An academic research team? Other?

The dataset was created by START, the National Consortium for the Study of Terrorism and Responses to Terrorism, a university-based research center, based at the University of Maryland.

I-C Was there a specific task in mind, or gap that needed to be filled?

The PIRUS dataset is among the first efforts to understand domestic radicalization from an empirical and scientifically rigorous perspective [1].

II. Composition

II-A What do the instances that comprise the dataset represent?

For example: crimes, offenders, court cases, police officers

Each instance corresponds to a de-identified individual who has been radicalized to violent or non-violent extremism.

II-B Are there multiple types of instances?

For example: offenders, victims, and the relationship between them.

No.

II-C How many instances are there in total?

Of each type, if appropriate.

There is data on 2,226 individuals in this dataset.

II-D Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

For example, if it is traffic stops from a territory, is it all traffic stops conducted within that territory within a specific time? If not, is it a representative sample of all stops? Describe how representativeness was validated/verified. If it is not representative, please describe why.

This is a sample of radicalized individuals in the United States. In order to be eligible for inclusion, each individual must meet one of the following five criteria:

1) The individual was arrested.
2) The individual was indicted of a crime.
3) The individual was killed as a result of his or her ideological activities.
4) The individual is/was a member of a designated terrorist organization.
5) The individual was associated with an extremist organization whose leader(s) or founder(s) has/have been indicted of an ideologically motivated violent offense.

In addition, each individual MUST:

1) Have been radicalized in the United States.
2) Have espoused or currently espouse ideological motives.
3) Show evidence that his or her behaviors are/were linked to the ideological motives he or she espoused/espouses.

However, the authors state:

"The PIRUS database is not, and should not be treated as, a comprehensive set of all individuals who have radicalized in the United States. Achieving a comprehensive dataset of all individuals who meet the database's inclusion criteria remains implausible for several reasons".[1]

II-E What data does each instance consist of?

If there is a large number of variables, please provide a broad description of what is included.

Each instances contains information on a wide range of characteristics, including:

1) Criminal activity.
2) Violent plots.
3) Relationship with extremist group.
4) Adherence to ideological milieus.
5) Factors relevant to their radicalization process.
6) Demographics.
7) Background.
8) Personal history.

II-F Is there a target label or associated with each instance?

Please include labels that are likely to be used as target labels, e.g. recidivism.

No. However, whether an individual's plot was executed according to their plan might be suitable to use as a target label.

[1] This quote is taken from Frequently Asked Questions on the PRIUS website.

II-G Are there recommended data splits (e.g., training, development/validation, testing)?

If so, please provide a description of these splits, explaining the rationale behind them.

No.

II-H Does the dataset contain data on race and ethnicity?

If so, is it based on the individual's self-description, or based on officer's impression? Was it collected or derived in post-processing? For example, by name analysis.

Yes. The PIRUS dataset, including information on race and ethnicity, was coded entirely using open-source material, including newspaper articles, websites, etc.

II-I Are there any known errors, sources of noise, bias or missing data, or variables collected for only part of the datasets?

If so, please provide a description.

No. However, that information in the dataset is based on oopen-source material, including newspaper articles, websites, etc.

II-J Does the dataset contain data on criminal history or other data that might be considered confidential or sensitive in any way?

For example: sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; If so, please provide a description.

Yes. The dataset contains information on criminal activity and relationship with extremist group, as well as other personal information.

II-K Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

If so, please describe how.

Indirectly, given the low frequency of the events and specific circumstances surrounding them.

III. Uses

III-A What type of tasks, if any, has the dataset been used for?

If so, please provide examples and include citations.

The dataset has been used for:

1) Comparitive studies between extremists and other groups [2], [3].
2) Looking at extremism within specific subgroups [4], [5].

III-B Is there a repository that links to any or all papers or systems that use the dataset?

If so, please provide a link or other access point.

No.

III-C What (other) tasks could the dataset be used for?

For example: testing predictive policing systems, predicting recidivism.

The dataset could be used for:

1) Additional investigations of extremism within subgroups.
2) Comparison of extremism outcomes across political affiliation.
3) Stratification by date, age, gender, location, ideology, group, etc. to address the specifics of radicalization in the United States.

III-D Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

The dataset was compiled from many open-source sources, such as social media and news articles. When using the dataset, one must assume the data has been merged correctly, and the information taken from these sources is correct.

IV. Collection Process

IV-A How was the data associated with each instance acquired?

e.g. the data collected survey, the raw data is routinely collected by the courts.

The PIRUS dataset was compiled from: newspaper articles, websites, secondary datasets, peer-reviewed academic articles, journalistic accounts including books and documentaries, court records, police reports, witness transcribed interviews, psychological evaluations/reports, and information directly attributed to the individual being researched (social media, etc.).

IV-B Was the information self-reported?

If the data was self-reported, was the data validated/verified? If so, please describe how.

No, the information is collected by the datasets' investigators from open source materials. Some information may in directly be self-reported (e.g., social media).

IV-C Who was involved in the data collection process?

Was this done as part of their other duties? If not, were they compensated?

The data collection was performed by investigators from START: Gary LaFree, Michael Jensen, and Sheehan Kane, among others.[2].

[2] A full list of authors can be found: here.

IV-D Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?

If not, please describe the timeframe in which the data associated with the instances was created. If the collection was not continuous within the timeframe, please specify the intervals, for example, annually, every 4 years, irregularly.

The dataset was collected between 2016 – 2018, and concerns the years 1948 through 2018.

IV-E Were any ethical review processes conducted (e.g., by an institutional review board)?

If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown.

IV-F Were the individuals in question notified about the data collection? Did they give their consent?

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

No.

IV-G Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown.

V. Pre-processing, cleaning, labeling

V-A Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, removal of instances, processing of missing values)?

If so, please provide a description and reference to the documentation. If not, you may skip the remaining questions in this section.

The specific processing steps have not been provided by the creators.

V-B Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?

If so, please provide a link or other access point to the "raw" data.

Unknown.

V-C Is the software that was used to preprocess/clean/label the data available?

If so, please provide a link or other access point.

No.

VI. Distribution

VI-A Is the data publicly available? How and where can it be accessed (e.g., website, GitHub)?

Does the dataset have a digital object identifier (DOI)?

A publicly avilable version on the dataset can be downloaded from:
https://www.icpsr.umich.edu/web/NACJD/studies/36309

You can request access to the full dataset here: https://www.start.umd.edu/webform/pirus-download-full-dataset.

VI-B Is the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The license agreement for the full dataset states the dataset can only be used for personal or academic research, journalistic use, or for an internal business process. See the license agreement for more details:
https://www.start.umd.edu/webform/pirus-download-full-dataset.

VII. Maintenance

VII-A Is the dataset maintained? Who is supporting/hosting/maintaining the dataset?

The dataset is not maintained since 2019.

VII-B How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Using the email pirus@start.umd.edu.

VII-C Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Not unless the authors receive further funding.

VII-D Are older versions of the dataset continue to be supported/hosted/maintained?

No.

VII-E If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

If so, please provide a description.

Unknown, contact pirus@start.umd.edu.

References

[1] G. McFee, M. Jensen, and P. James, “Profiles of Individual Radicalization in the United States (pirus),” *College Park, MD: National Consortium for Terrorism and Responses to Terrorism, University of Maryland. Retrieved September*, vol. 17, p. 2019, 2019.

[2] D. C. Pyrooz, G. LaFree, S. H. Decker, and P. A. James, “Cut from the Same Cloth? A Comparative Study of Domestic Extremists and Gang Members in the United States,” *Justice Quarterly*, 2017.

[3] M. Al-Zewairi and G. Naymat, “Spotting the Islamist Radical within: Religious Extremists Profiling in the United States,” *Procedia Computer Science*, vol. 113, pp. 162–169, 2017.

[4] H. Haugstvedt and D. Koehler, “Armed and Explosive? an Explorative Statistical Analysis of Extremist Radicalization Cases with Military Background,” *Terrorism and Political Violence*, pp. 1–15, 2021.

[5] R. Yon and D. Milton, “Simply Small Men? Examining Differences between Females and Males Radicalized in the United States,” *Women & Criminal Justice*, vol. 29, no. 4-5, pp. 188–203, 2019.