An Approach for Researcher Identification on Twitter Without the Need for External Data

Sarah Marie Müller[1], Maren Kotzur[2], and André Bittermann[3]

[1]*samu@leibniz-psychology.org*
Leibniz Institute for Psychology (ZPID), Universitätsring 15, 54296 Trier (Germany)

[2]*m.kotzur@mail.de*
Saarland University, Campus, 66123 Saarbrücken (Germany)

[3]*abi@leibniz-psychology.org*
Leibniz Institute for Psychology (ZPID), Universitätsring 15, 54296 Trier (Germany)

Background

Even after the #TwitterMigration movement at the end of 2022, Twitter can be considered as the most popular social media platform among the academic community: Mongeon et al. (2023) report nearly 500,000 scientists on this platform. Twitter posts of the scientific community can be a valuable data source for metascientific or bibliometric endeavors. For instance, Bittermann et al. (2021) demonstrated that text mining of tweets can predict future publication trends. The valid identification of researchers, however, is paramount and various approaches have been presented. While some were limited by a lack of account validation (e.g., Hadgu & Jäschke, 2014), others rely on external data sources (Costas et al., 2020; Mongeon et al., 2023). Dependency on external data can be a challenge to the sustainability of developed approaches.

In this study, we investigate the results of a chain-referral sampling algorithm that uses solely data from the Twitter API. For a proof-of-concept, we focus on the field of psychology. Prior studies found evidence that scientific Twitter users might not represent the true population of researchers regarding the distribution of gender (Weinstein & Sumeracki, 2017), location (Costas et al., 2020), and psychological subdisciplines (Bittermann et al., 2021). Hence, we address the following research questions (RQs):

RQ1 - Validity Check: How correctly can psychology researchers be identified using the chain-referral sampling algorithm?

RQ2 - Representativeness Check: Is the identified psychology research community on Twitter representative of the overall population of psychology researchers?

Methods

Algorithm

Our identification approach crawls the mentions network of a seed dataset of verified researchers and classifies accounts via keyword matching (e.g., research* AND psycholog*) in their profile descriptions (the most important predictor found by Hadgu & Jäschke, 2014). By doing so, we are aiming for researchers that explicitly disclose themselves as such. The included accounts are then used as seed for the next iteration until no more accounts are found. We queried the Twitter API in February 2021, yielding 16,491 psychology researchers.

Validation of Twitter Accounts

To check the validity of the Twitter researcher dataset, a sample-based precision-recall analysis was performed. For a random subsample of 100 accounts, a verification was performed via web searches for their respective professional backgrounds. Following the approach by Holmberg and Thelwall (2014), we included the top most productive researchers in psychology in the validation sample. In the next step, we investigated their Twitter accounts, resulting in 48 psychological researchers with Twitter accounts.

Sample Representativity

The representativeness of the algorithm-identified Twitter sample for the entire psychological research community was determined using three criteria: gender, location, and subdiscipline. The aim was to examine whether the distribution of each criterion in the dataset of psychology researchers on Twitter corresponds to the distribution of publishing authors. Specifically, we exported all records from the psychology literature database PsycInfo for the year 2020 – the most current and complete publication year at time of the researcher identification. The dataset consisted of 292,476 publishing psychologists.

The prenames of each researcher in the Twitter and PsycInfo sample were extracted and assigned to the most likely matching gender (proportions above 0.5). The location information could be taken directly from respective metadata in the PsycInfo and Twitter datasets. Psychological subdisciplines were determined using PsycInfo metadata. For the Twitter sample, we defined subdiscipline-related

terms (e.g., "clinical psycholog*") and searched the profile descriptions for matching patterns.

Results

A total of 88 of the 100 verified Twitter accounts could be reliably assigned to a psychology researcher. This results in a precision value of 0.88. Of the 48 most productive psychologists with Twitter accounts, 20 authors were detected by the algorithm. This yields a recall value of 0.42.

A chi-square test revealed significant differences in the proportion of the gender between the samples, $\chi^2(1, N = 239{,}197) = 198.94$, $p < .001$, with women being overrepresented in the Twitter sample. Fisher's exact test was significant for differences regarding the location of researchers on Twitter compared to those on PsycInfo ($p < .001$). Likewise, Fisher's exact test revealed significant differences in the proportion of the psychology subdisciplines between the samples ($p < .001$): Clinical Psychology was underrepresented in the Twitter sample, while Social and Developmental Psychology were overrepresented.

Discussion

As intended, our approach clearly favors precision over recall. It is noticeable that many of the Twitter profiles that were not found by the algorithm are either relatively inactive, not connected to the psychological research community, or the accounts are used for private purposes. Since inactive or private accounts are not of interest for research on scientific Twitter use, the recall value of 0.42 is not an indicator that the algorithm does not perform validly w.r.t to our goal of finding researchers that disclose themselves. This, of course, is not always the case and represents a clear limitation of our approach. Mongeon et al. (2023) achieved higher values of both precision (0.96) and recall (0.62) by leveraging data from OpenAlex, Crossref and ORCID. Given the fact that the actual identification of our approach is based on Twitter data only, our values of .88 and 0.42 suggest our approach as a solid alternative for the case of missing external data sources.

The distributions of gender, location, and subdisciplines differed significantly between the Twitter and the OVID sample. Women were overrepresented, which is in line with previous findings (Weinstein & Sumeracki, 2017). The comparison of the location comparison criterion revealed that the U.S. and U.K. are clearly overrepresented, while China is severely underrepresented, which is consistent with the finding of Costas and colleagues (2020). Contrary to our results, in the study of Bittermann et al. (2021) Biological Psychology and Neuropsychology were overrepresented. These differences could be explained by their sample consisting exclusively of German professors, whereas the present sample comprises international researchers or could be related to the different subdisciplines and their assignments.

Conclusion

The presented chain-referral sampling algorithm that uses solely data from the Twitter API provides an easy-to-implement solution for finding researchers, but is limited to a clear disclosure of researchers in their profile descriptions. Employing NLP techniques such as word embeddings could improve text classification, and including network data (e.g., followers and followers) might overcome the issue of dependence on researcher self-disclosure. However, our approach is not limited to Twitter and can be easily adopted to Mastodon or similar social networks used by academics. In any case, our study provides further evidence that Twitter-active researchers should not be regarded as representative of the whole research community.

References

Bittermann, A., Batzdorfer, V., Müller, S. M., & Steinmetz, H. (2021). Mining twitter to detect hotspots in psychology. *Zeitschrift Für Psychologie, 229*(1), 3–14. https://doi.org/10.1027/2151-2604/a000437

Costas, R., Mongeon, P., Ferreira, M. R., van Honk, J., & Franssen, T. (2020). Large-scale identification and characterization of scholars on twitter. *Quantitative Science Studies*, *1*(2), 771–791. https://doi.org/10.1162/qss_a_00047

Hadgu, A. T., & Jäschke, R. (2014). Identifying and analyzing researchers on twitter. *Proceedings of the 2014 ACM Conference on Web Science,* 23–32. https://doi.org/10.1145/2615569.2615676

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in twitter scholarly communication. *Scientometrics, 101*(2), 1027–1042. https://doi.org/10.1007/s11192-014-1229-3

Mongeon, P., Bowman, T. D., & Costas, R. (2023). An open dataset of scholars on twitter. *Quantitative Science Studies,* 1–11. https://doi.org/10.1162/qss_a_00250

Weinstein, Y., & Sumeracki, M. A. (2017). Are twitter and blogs important tools for the modern psychological scientist? *Perspectives on Psychological Science*, 12(6), 1171–1175. https://doi.org/10.1177/1745691617712266

Supplements

We provide twitter ids and R code on https://github.com/sarahmrml/Twitter-Researcher-Identification

Author Contributions

All authors wrote the manuscript. SMM and MK planned the validation and representativity check. SMM performed all analyses. AB supervised the project and developed the identification algorithm.