Dissertation project | Paper 1

Working title: **Mapping and Evaluating Measurement Validity Approaches for Text-as-Data Methods**

**Research Design**

Lukas Birkenmaier  
MA Politik- und Verwaltungswissenschaften (University of Konstanz) &   
MSc Big Data and Digital Futures (University of Warwick)

**Table of Contents**

[Study Background 3](#_Toc105492191)

[Research Design 4](#_Toc105492192)

[Systematic Review 4](#_Toc105492193)

[Literature Search 4](#_Toc105492194)

[Strategy of Analysis 6](#_Toc105492195)

[Expert Interviews 7](#_Toc105492196)

[Appendix 9](#_Toc105492197)

[Literature 11](#_Toc105492198)

# Study Background

In the past decade, social science research has experienced a remarkable rise in publications relying on digital behavioral data and computational methods (Brady, 2019; Edelmann et al., 2020). Thereby, the analysis of texts using computational text analysis methods (CATM) holds particular promise, as textual data enable researchers to facilitate substantively important inferences of human behavior on an unprecedented scale (Grimmer & Stewart, 2013; Lazer et al., 2009). However, a growing body of research highlights several methodological challenges in the automated analysis of textual data (Baden et al., 2021; Howison et al., 2011; van Atteveldt et al., 2021). One of the key challenges relates to the validity of CATM-based measures. Several scholars have expressed doubts about whether CATM can operationalize and validly measure the constructs under study, thus questioning these methods’ ability to answer substantive theory-driven research questions (Adcock & Collier, 2001). This is especially relevant because the high dimensionality and complex structure of texts renders it almost impossible to *truly* understand the underlying dependencies of texts (Yeomans, 2021).[[1]](#footnote-2)[[2]](#footnote-3) This culminates in the problem that even hand-coded estimates (“gold standard”) for the same texts are often [not consistent](https://www.linguee.de/englisch-deutsch/uebersetzung/not+consistent.html) so that reliability measures can vary significantly between different coders (cf. Song et al., 2020). Therefore, authors such as Baden et al. (2021) have called for more unifying validation efforts to show “exactly how and how convincingly [CATM] operationalize relevant conceptual properties” (Baden et al., 2021, p. 14). When reviewing the literature, one can observe a great heterogeneity of validation approaches in studies using CATM developed over time. Among others, these approaches include the comparison of estimates with other measurements and external data (Bach et al., 2021; Kmetty & Németh, 2022; Röttger et al., 2021), or the development of multistage validation frameworks, which cover different dimensions of validity at the same time (Goet, 2019; Lowe & Benoit, 2013). Even more worrying, scholars who want to validate their measures can easily lose track of what options they have, and which validation approaches are necessary to conduct high-quality research projects.

# Research Design

To provide guidance in CATM research, this first paper aims to collect, structure, and describe the various validation approaches in CATM research. Thus, it is asked: How are issues of measurement validity addressed in CATM studies? What are the main validation strategies in CATM research? And how can the current approaches be structured into overarching categories? To answer these research questions, this study provides a two-fold approach. First, a systematic review of validation practices from publications in the field of political communication. And second, focus group interviews with CATM experts. Subsequently, both approaches will be presented and discussed in more detail.

## Systematic Review

To conduct the review, a systematic framework will be applied to objectively search for relevant publications (Durlak & Lipsey, 1991; Liberati et al., 2009). Thereby, several decisions need to be made. Most importantly, these entail the literature search and the development of an adequate strategy of analysis.

### Literature Search

**Research Area.** Starting from the goal of evaluating common validation practices, the research field of the systematic review should be characterized by an adequate professionalization and popularity of CATM usage. Therefore, the review will include publications in the field of political communication. Generally, the field of political communication has experienced remarkable methodological advancements in the last decade, with scholars “[deploying] sophisticated analytical methods that are rigorously validated and made transparent through openly accessible replication repositories” (Theocharis & Jungherr, 2021, p. 2). Likewise, the field of political communication promises to gain further relevance in the future, because it deals with some of the most pressing challenges facing digital societies today, such as digitization, fake news, or political polarization (Dufva & Dufva, 2019).

**Methods:** The review will include publications which rely on computational methods using text as data. In particular, these might entail rule-based (e.g., dictionaries), unsupervised (text scaling algorithms, topic modelling), and supervised (machine learning) methods to operationalize and measure variables of interest for political and communicational research (Grimmer & Stewart, 2013). Generally, previous studies have pointed to the fact that validation exercises vary greatly between groups of methods.[[3]](#footnote-4) For instance, an unsupervised text scaling model like *Wordfish* (Slapin & Proksch, 2008) requires specific assumptions about the ideological composition of textual data (but no labelled training data), whereas a deep-learning ML model like BERT (Devlin et al., 2019) primarily requires sufficient labelled training data while being characterized as a “black box”. That being said, a systematic description and comparison of validation exercises across groups of methods might be an important first insight from this systematic review.

**Publication Types**. The review will include peer-reviewed publications from the most relevant journals in the field of communication and political science. The journals are identified using the Scimago Journal Ranking.[[4]](#footnote-5) Additionally, the choice of journals will be evaluated by relying on expert in the field. The lists of journals includes the following journals:

Communication:

* Communication Methods and Measures
* [Digital Journalism](https://www.scimagojr.com/journalsearch.php?q=21100430187&tip=sid&clean=0)
* [Political Communication](https://www.scimagojr.com/journalsearch.php?q=22838&tip=sid&clean=0)
* [Applied Linguistics](https://www.scimagojr.com/journalsearch.php?q=18651&tip=sid&clean=0)
* [Research on Language and Social Interaction](https://www.scimagojr.com/journalsearch.php?q=22340&tip=sid&clean=0)
* Journal of Communication
* Public Opinion Quarterly

Political Science:

* [Administrative Science Quarterly](https://www.scimagojr.com/journalsearch.php?q=16036&tip=sid&clean=0)
* [American Sociological Review](https://www.scimagojr.com/journalsearch.php?q=16929&tip=sid&clean=0)
* [American Political Science Review](https://www.scimagojr.com/journalsearch.php?q=15557&tip=sid&clean=0)
* [Annual Review of Political Science](https://www.scimagojr.com/journalsearch.php?q=15646&tip=sid&clean=0)
* [Annual Review of Sociology](https://www.scimagojr.com/journalsearch.php?q=16976&tip=sid&clean=0)
* [American Journal of Political Science](https://www.scimagojr.com/journalsearch.php?q=15555&tip=sid&clean=0)
* [Leadership Quarterly](https://www.scimagojr.com/journalsearch.php?q=21149&tip=sid&clean=0)
* [Political Analysis](https://www.scimagojr.com/journalsearch.php?q=145558&tip=sid&clean=0)
* Party Politics

**Time and Language.** Some additional criteria are applied:

* Only publications published after 2015
* Only publications in the English language

After the first literature search using the Web of Science database, a second screening round will be conducted by examining the title, abstract and keywords of all studies identified. To do so, the tool a*bstrackr* (Wallace et al., 2012) from Brown university will be used, which provides an open source web interface for abstract and keyword screening.

Once the final sample of publications is collected (see previous section), publications will be coded based on a coder manual (Durlak & Lipsey, 1991). Optimally, at least two coders will be involved, so that metrics of interrater agreement for a sample of publications can be reported (Gisev et al., 2013). However, the coding of validation efforts might be too difficult for a research assistant with no previous knowledge on text analysis. Therefore, a solution could be to code a subset of studies (e.g., 20-30 studies) and proceed to independent coding after sufficient intercoder reliability is achieved (see Landis & Koch, (1977))

|  |  |  |  |
| --- | --- | --- | --- |
| 0 | Question | Question Type | Remarks |
| 1 | **Method Type** | Dropdown - Supervised: Machine Learning - Unsupervised: Topic Moddeling - Unsupervised: Text Scaling - Rule-based: Off-the-shelf dictionary - Rule-based: Adjusted dictionary - Rule-based: Development dictionary -Others |  |
| 2 | **Method Name** *What is the name of the Method?  (e.g., SMV, BERT)* | Open Field |  |
| 3 | **Outcome Method:** *What is the outcome metric?* | Dropdown - Classification - Scaling - tba. |  |
| 4 | **Target Construct:** *What is the targeted construct to be measured?* *(e.g., polarization, sentiment).* | Open Field |  |
| 5 | **Data Language** *What is the primary language of the data corpus?* | Dropdown Multiple Choice |  |
| 6 | **Data Source** *What is the datasource?  (e.g., Social Media Data, Newspaper etc.)* | *Dropdown* |  |
| 7 | **Validation Text Examples** *In what paragraphs are issues of measurement validity / robustness checks / […] reported? Please copy them into the open field!* | Open Field | Copy most relevant paragraphs from texts |
| 8 | **Validation Steps:** *What steps of validation exercises are reported?* | Open Field   * Content Validity * Convergent Validity * Criterion Validity * Structural Validity . | Generic development of drop-down list; should entail openness for new approaches |
| 9 | **Interpretation Construct Validity:** *How do these steps relate to subdimensions of construct validity?* | Dropdownon Type |  |
| 10 | **Hand Coding I** *If Hand Coding: How many coders were used?* | Dropdown | optionally |
| 11 | **Hand Coding II:**  *What are the levels of intercoder reliability?* | Open Field | optionally |
| 12 | **Additional Material** *Is there a link to an Appendix/additional Materials/GitHub repository?* | Dropdown |  |
| 13 | **Accessibility Data** *Is the data for the analysis accessible?* | Dropdown | optionally |
| 14 | **Accessibility Code** *Is the code for the validation accessible?* | Dropdown | optionally |

After the coding of publications, the results of the coding process will be presented. This includes descriptive statistics of the coded variables and a summarizing discussion of the validation strategies identified.

## Expert Interviews

To get an even more profound understanding of validation practices, we will also conduct interviews with experts in the field of CATM research. The interviews will be carried out at the [name] conference using open ended questions related to the practices and challenges of CATM validation. Due to the exploratory nature of this study, experts will be selected based on their reputation and publication record in the field of political communication and CATM application. In total, we aim to conduct around eight interviews, with participation being voluntarily and without compensation.

Below, the main questions of the questionnaire are reported. The draft for the full questionnaire can be found in Appendix 4:

* What is your personal and professional background?
* What is validity to you? Have you come across different validity conceptualizations within different research fields?
* What approaches do you know to validate a measurement method?
* What approaches do you know of to validate computer-assisted text methods?
* Do you know any validation approaches of CATM, which are usually not reported in the final publications?
* What do you see as barriers to sufficiently validating CATM?
* Do you have any ideas on how to improve validation practices in TaDa research?

Ultimately, the interviews will be transcripted and analysed using content analysis (Krippendorff, 2018).

# Appendix

## Appendix 1: Keywords

* + Naïve search
    - (Politi\* OR Party OR Govern\*) AND text\*
  + Adjusted search (*LitsearchR*)
    - Topic
      * **Political Communication** 
        + (elect\* OR govern\* OR parti\* OR polici\* OR "social\* media\*" OR polit\* OR news cover\* OR plenar\* debat\*)
      * **Textual Analysis**
        + ("autom\* text\* analysi\*" OR "content\* analysi\*" OR sentiment\* OR discours\* OR languag\* OR "machin\* learn\*" OR text\* OR word\* OR "comput\* communic\* scienc\*" OR Corpus\* OR lexicon\* OR Automa\* Content\* Analy\* OR supervised OR scalin\* OR diction\* OR topic model\*)
    - Journals
      * "Communication Methods and Measures" OR "Digital Journalism" OR "Political Communication" OR "Applied Linguistics" OR "Research on Language and Social Interaction" OR "Journal of Communication" OR "Public Opinion Quartely"
      * “[Administrative Science Quarterly](https://www.scimagojr.com/journalsearch.php?q=16036&tip=sid&clean=0)” OR “[American Sociological Review](https://www.scimagojr.com/journalsearch.php?q=16929&tip=sid&clean=0)” OR “[American Political Science Review](https://www.scimagojr.com/journalsearch.php?q=15557&tip=sid&clean=0)” OR “[Annual Review of Political Science](https://www.scimagojr.com/journalsearch.php?q=15646&tip=sid&clean=0)” OR “[Annual Review of Sociology](https://www.scimagojr.com/journalsearch.php?q=16976&tip=sid&clean=0)” OR “[American Journal of Political Science](https://www.scimagojr.com/journalsearch.php?q=15555&tip=sid&clean=0)” OR “[Leadership Quarterly](https://www.scimagojr.com/journalsearch.php?q=21149&tip=sid&clean=0)” OR “[Political Analysis](https://www.scimagojr.com/journalsearch.php?q=145558&tip=sid&clean=0)”

## Appendix 2: Benchmark Literature

* Goet, Niels D. ‘Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015’. *Political Analysis* 27, no. 4 (2019): 518–39.
* Gilardi, Fabrizio, Theresa Gessler, Mael Kubli, and Stefan Muller. ‘Social Media and Political Agenda Setting’. Political Communication 39, no. 1 (2 January 2022): 39–60.
* Di Cocco, J., & Monechi, B. (2022). How Populist are Parties? Measuring Degrees of Populism in Party Manifestos Using Supervised Machine Learning. Political Analysis, 30(3), 311-327. doi:10.1017/pan.2021.29
* Røed, Maiken. "When do political parties listen to interest groups?." Party Politics (2022): 13540688211062832.
* Welbers, K., van Atteveldt, W., Bajjalieh, J., Shalmon, D., Joshi, P. V., Althaus, S., Chan, C.-., Wessler, H. & Jungblut, M. (2022). [Linking event archives to news: A computational method for analyzing the gatekeeping process](https://madoc.bib.uni-mannheim.de/61633/1/Linking%20event%20archives%20to%20news%20a%20computational%20method%20for%20analyzing%20the%20gatekeeping%20process.pdf). *Communication Methods and Measures, 16*(1), 59–78. https://doi.org/10.1080/19312458.2021.1953455

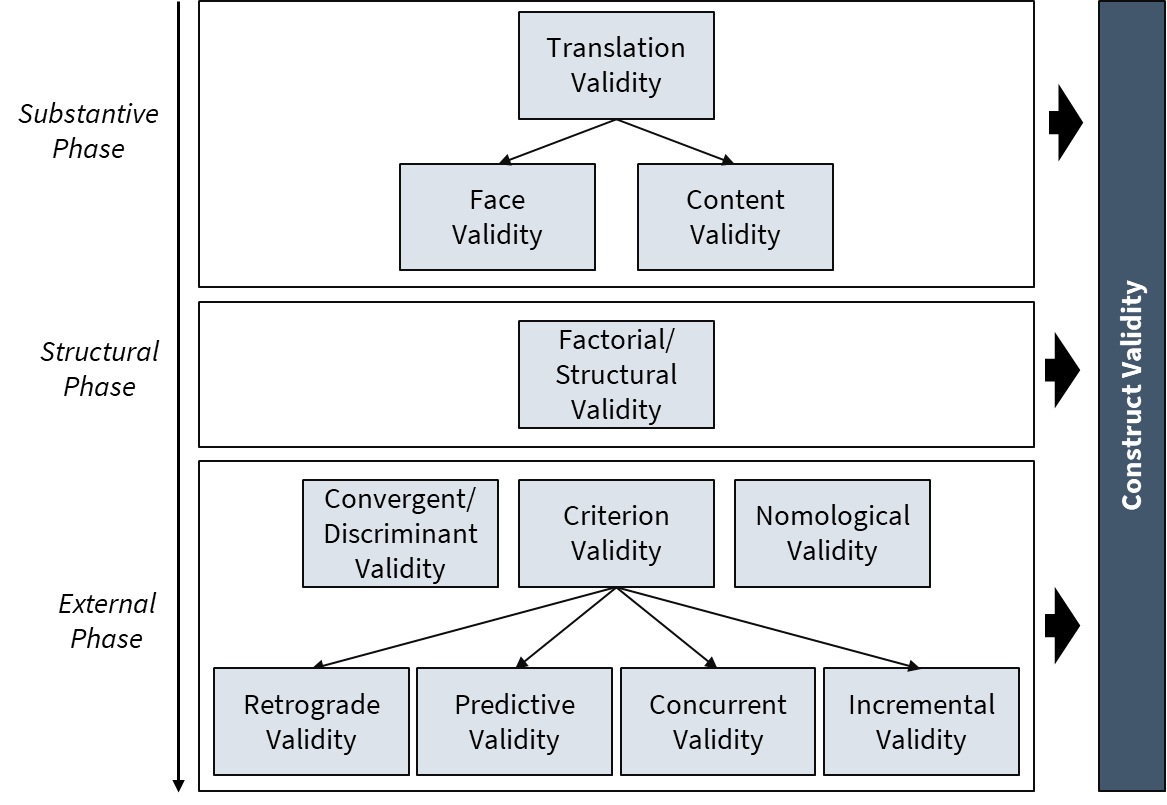
Appendix 3: Overview Validity Concepts

Figure 1: Own visual depiction of validity concepts

## Appendix 4: Questionnaire

|  |  |
| --- | --- |
| Background Paper |  |
|  |  |
| Consent | Furthermore, I would like to inform you that I will record the prepared questions and your answers for my bachelor thesis. This recording will be transcribed to be able to analyze the answers. The recorded data is purely for research purposes Your data will only be presented in aggregated form and anonymized. The research paper, including the appendices, can only be viewed by members of the Department of Geography at Heidelberg University, including myself.  Do you agree that I record this interview for research purposes? |

|  |  |
| --- | --- |
| Introduction | What is your professional background? |
| How is your work related to text analysis and political communication? |
| Measurement Validity | How would you **describe validity**? |
| Have you come across **different validity terms** within different **research fields**? |
| Measurement validity and computer-assisted text methods? | Have you come across **different validity terms** in the field of computer-assisted text methods? |
| How would you describe the process of measurement validation? Which steps are usually reported in the final paper? |
| **What approaches** do you know to validate a computer-assisted text method? |
| How would you **describe** the **current state of knowledge** on validating computer-assisted text method? |
| Challenges Measurement Validity and CATM | What are the barriers that **hamper** the validation of computer-assisted text methods? (*Costs, missing guidelines etc*.) |
| Do you have any ideas on **how to improve** the validation of computer-assisted text methods? |
| Conclusion | Is there anything else you would like to tell us? |

# Literature

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, *95*(3), 529–546.

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting voting behavior using digital trace data. *Social Science Computer Review*, *39*(5), 862–883.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2021). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 1–18. https://doi.org/10.1080/19312458.2021.2015574

Brady, H. E. (2019). The Challenge of Big Data and Data Science. *Annual Review of Political Science*, *22*(1), 297–323. https://doi.org/10.1146/annurev-polisci-090216-023229

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Dufva, T., & Dufva, M. (2019). Grasping the future of the digital society. *Futures*, *107*, 17–28. https://doi.org/10.1016/j.futures.2018.11.001

Durlak, J. A., & Lipsey, M. W. (1991). A practitioner’s guide to meta-analysis. *American Journal of Community Psychology*, *19*(3), 291–332.

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational Social Science and Sociology. *Annual Review of Sociology*, *46*(1), 61–81. https://doi.org/10.1146/annurev-soc-121919-054621

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330–338.

Goet, N. D. (2019). Measuring polarization with text analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis*, *27*(4), 518–539.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems*, *12*(12), 2.

Kmetty, Z., & Németh, R. (2022). Which is your favorite music genre? A validity comparison of Facebook data and survey data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 07591063211061754. https://doi.org/10.1177/07591063211061754

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., & Gutmann, M. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine*, *6*(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

Lowe, W., & Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, *21*(3), 298–313.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 41–58. https://doi.org/10.18653/v1/2021.acl-long.4

Slapin, J. B., & Proksch, S.-O. (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, *52*(3), 705–722. https://doi.org/10.1111/j.1540-5907.2008.00338.x

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Theocharis, Y., & Jungherr, A. (2021). Computational Social Science and the Study of Political Communication. *Political Communication*, *38*(1–2), 1–22. https://doi.org/10.1080/10584609.2020.1833121

van Atteveldt, W., van der Velden, M. A., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, *15*(2), 121–140.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 819–824.

Yeomans, M. (2021). A concrete example of construct construction in natural language. *Organizational Behavior and Human Decision Processes*, *162*, 81–94. https://doi.org/10.1016/j.obhdp.2020.10.008

1. This has been described as the first principle of textual data analysis by Grimmer and Stewart (2013), namely “all quantitative models of language are wrong -but some are useful). Thus, “the complexity of language implies that all methods necessarily fail to provide an accurate account of the data-generating process used to produce texts” (p. 270) [↑](#footnote-ref-2)
2. “The quantities we seek to estimate from text [...] are fundamentally unobservable” (Lowe & Benoit, 2013, p. 299), [↑](#footnote-ref-3)
3. "Yet only the development of rule-based approaches is sometimes accompanied by similar efforts at concept-driven operationalization and the development of substantive validity criteria (e.g., in the construction of dictionaries, word banks and rule sets). For supervised applications, validation efforts remain largely limited to the manual creation of training data sets, and are rarely reported (Baden et al., 2021)" [↑](#footnote-ref-4)
4. [https://www.scimagojr.com/](https://www.scimagojr.com/journalrank.php?area=330) [↑](#footnote-ref-5)