ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Heiberger, Raphael H.

Article — Published Version

Applying Machine Learning in Sociology: How to Predict Gender and Reveal Research Preferences

KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie

Provided in Cooperation with:
Springer Nature

Suggested Citation: Heiberger, Raphael H. (2022) : Applying Machine Learning in Sociology: How to Predict Gender and Reveal Research Preferences, KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie, ISSN 1861-891X, Springer Fachmedien Wiesbaden GmbH, Wiesbaden, Vol. 74, Iss. Suppl 1, pp. 383-406, https://doi.org/10.1007/s11577-022-00839-2

This Version is available at:
https://hdl.handle.net/10419/312484

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Köln Z Soziol (2022) (Suppl 1) 74:383–406
https://doi.org/10.1007/s11577-022-00839-2

Kölner Zeitschrift für Soziologie und Sozialpsychologie

Check for updates

ABHANDLUNGEN

Applying Machine Learning in Sociology: How to Predict Gender and Reveal Research Preferences

Raphael H. Heiberger

Received: 8 December 2021 / Accepted: 17 March 2022 / Published online: 19 May 2022
© The Author(s) 2022

Abstract Applications of machine learning (ML) in industry and natural sciences yielded some of the most impactful innovations of the last decade (for instance, artificial intelligence, gene prediction or search engines) and changed the everyday-life of many people. From a methodological perspective, we can differentiate between unsupervised machine learning (UML) and supervised machine learning (SML). While SML uses labeled data as input to train algorithms in order to predict outcomes of unlabeled data, UML detects underlying patterns in unlabeled observations by exploiting the statistical properties of the data. The possibilities of ML for analyzing large datasets are slowly finding their way into the social sciences; yet, it lacks systematic introductions into the epistemologically alien subject. I present applications of some of the most common methods for SML (i.e., logistic regression) and UML (i.e., topic models). A practical example offers social scientists a "how-to" description for utilizing both. With regard to SML, the case is made by predicting gender of a large dataset of sociologists. The proposed approach is based on open-source data and outperforms a popular commercial application (genderize.io). Utilizing the predicted gender in topic models reveals the stark thematic differences between male and female scholars that have been widely overlooked in the literature. By applying ML, hence, the empirical results shed new light on the longstanding question of gender-specific biases in academia.

Keywords Computational social science · Machine learning · Topic models · Sociology of science · Gender bias

R. H. Heiberger (✉)
Institute for Social Sciences, University of Stuttgart
Seidenstraße 36, 70174 Stuttgart, Germany
E-Mail: raphael.heiberger@sowi.uni-stuttgart.de

Über Anwendungen des Maschinellen Lernens in der Soziologie: die Vorhersage von Geschlecht und wie dieses Forschungspräferenzen strukturiert

Zusammenfassung In der Industrie und in den Naturwissenschaften haben Anwendungen des „Maschinellen Lernens“ (ML) einige der einflussreichsten Innovationen des letzten Jahrzehnts hervorgebracht, die das Alltagsleben vieler Menschen verändert haben (z. B. künstliche Intelligenz, Genvorhersage oder Suchmaschinen). Aus methodischer Sicht können wir dabei zwischen „unsupervised machine learning“ (UML) und „supervised machine learning“ (SML) unterscheiden. Während SML annotierte Daten als Input für das Training von Algorithmen verwendet um die Ergebnisse von nicht-annotierten Daten vorherzusagen, erkennt UML zugrundeliegende Muster in unklassifizierten Beobachtungen, indem es die statistischen Eigenschaften der Daten nutzt. Die Möglichkeiten, die ML zur Analyse großer Datenmengen bietet, finden langsam auch ihren Weg in die Sozialwissenschaften. Es fehlt jedoch an systematischen Einführungen in das erkenntnistheoretisch fremd erscheinende Thema. In diesem Beitrag stelle ich daher Anwendungen einiger der gängigsten Methoden sowohl für SML (logistische Regression) als auch UML (Topic Models) vor. Ein praktisches Beispiel bietet Sozialwissenschaftlerinnen und -wissenschaftlern eine „How-to“-Beschreibung für den Einsatz beider Methoden. In Bezug auf die SML wird der Fall anhand der Vorhersage des Geschlechts eines großen Datensatzes von Soziologinnen und Soziologen dargestellt. Der vorgeschlagene Ansatz basiert auf Open-Source-Daten und dessen Performance übertrifft die einer populären kommerziellen Anwendung zu dem Thema (genderize.io). Die Verwendung des vorhergesagten Geschlechts in den Topic Models offenbart starke thematische Unterschiede zwischen männlichen und weiblichen Wissenschaftlern, die in der Literatur bislang weitgehend übersehen wurden. Die Anwendung von ML wirft daher ein neues Licht auf bisherige Erkenntnisse zu geschlechtsspezifischen Unterschieden in der Wissenschaft.

Schlüsselwörter Rechnergestützte Sozialwissenschaft · Maschinelles Lernen · Topic Models · Wissenschaftssoziologie · Geschlechtsunterschiede

1 Introduction

Machine learning (ML) summarizes statistical methods in which computers learn from data and extract information. Applications of ML paved the way for some of the most promising technical innovations in recent years (e.g., artificial intelligence, gene prediction or search engines) and changed the everyday-life of many people (Jordan and Mitchell 2015). ML represents a breakthrough in computer sciences; yet, its adoption in the social sciences is less enthusiastic. Although a recent article gives a comprehensive overview of sociological studies using ML (Molina and Garip

2019), an application-oriented introduction that might ease a sociologist's way into the subject is still lacking.[1]

Therefore, this article has three goals: First, I discuss and categorize ML methods. From a methodological perspective, ML can be classified in two paradigms: *supervised machine learning* (SML) and *unsupervised machine learning* (UML) (e.g., Jordan and Mitchell 2015; Molina and Garip 2019). Although SML uses labeled data as input to train algorithms in order to predict outcomes of unlabeled data, UML detects underlying patterns in unlabeled observations by exploiting the statistical properties of the data. I will give an overview of both areas emphasizing that several of such tools used in ML are not, by any means, new to social scientists interested in statistics.

The other two aims are intertwined. On the one hand, I present a "how-to" guide for both SML and UML. I do that, on the other hand, by applying SML and UML to an important substantial case, i.e., the mostly unexplored role that research topic choice plays in the academic gender gap. By shedding light on the empirical case, the application of ML will be practically illustrated "by doing." Thus, I will *not* present a literature review, as this has been done in a comprehensive way for sociology by Molina and Garip (2019) only recently.[2] Instead, I will present an easy-to-use SML classifier to derive the associated gender from first names. The proposed approach outperforms a prominent commercial application (genderize.io). Detecting gender "automatically" might be useful in many cases of (quantitative or qualitative) content analysis. Using the predicted gender of authors, I examine gender-specific preferences in research topics by using UML. In particular, I explain and apply structural topic modeling (Roberts et al. 2014) in order to reduce a corpus of texts from a near-complete sample of US dissertations on sociology to its main dimensions. In so doing, the article reveals important differences in research choices of female and male PhD students, and, hence, adds a widely overlooked aspect to the rich literature on gender biases in academic publishing.

2 Principles of ML

2.1 SML

Many people might first think of SML when referring to machine learning, as it is the most widely used area of ML and comprises the methods that witnessed the largest performance boost owing to larger and more detailed data in recent years (e.g., image recognition). Although SML was primarily used in computer sciences, its applications spread nowadays to almost all scientific fields and business branches (Jordan and Mitchell 2015). The main aim of SML is to predict an outcome with

[1] Code and data necessary to replicate findings can be found on GitHub (https://github.com/RapHei/ApplyML_Sociology).

[2] To be sure, reviews for neighboring fields are also available, namely economics (Mullainathan and Spiess 2017), political science (Cranmer and Desmarais 2017), or psychology (Yarkoni and Westfall 2017).

a given set of features. That is the same as when social scientists refer to estimating a dependent variable by using a set of independent variables.

Thus, in how far does SML actually differ from classic statistical methods? The answer lies in the regularization of variance and empirical tuning of parameters (Molina and Garip 2019; Mullainathan and Spiess 2017). I would also like to emphasize that (apparent) differences stem from differing goals. Although classic statistics tries to infer parsimonious models that explain how an outcome is generated, SML does not care about interpretability but only how to best forecast the outcome. "Generative modeling" (Donoho 2017) focuses on unbiased and consistent estimators of a given dataset, i.e., beta-coefficients are the most interesting part of regressions for social scientists because they provide access to *explaining*[3] the data at hand. This is a crucial epistemological difference yielding many practical consequences.

In contrast, SML prioritizes *predictions.* Regardless of meaningful interpretations and unbiased estimators, SML uses functions of high complexity as long as they perform well "out-of-sample," i.e., models are able to predict new data. That means, issues such as autocorrelation or multicollinearity are treated as features, not problems. Consequently, functions may yield "black-boxes" (e.g., when it comes to multi-layer neural networks or high orders of interaction effects); a large number of variables might be used; hard-to-interpret polynomials and interactions are included; and a certain degree of "in-sample error" for the sake of predicting new data correctly may be allowed.

Thus, unlike most social scientists using *one* dataset for modeling efforts, SML consists of at least two datasets: training and test data.[4] The first dataset is used to develop (i.e., train) the model, the second to test its predictive capacity on out-of-sample data. Often, the train and test sets are randomly sampled from the same dataset, which is split, e.g., 50/50 (although there is no general rule of thumb that I am aware of).

Supervised machine learning is aimed at regulating between under- and overfitting. Although classic statistical models are prone to overfitting and therefore possess only limited predictive abilities for new data, SML uses "regularizers" (i.e., parameters of algorithms) to balance both underfitting and overfitting. To accomplish this task SML uses the training data and tunes regularizers to fit the data at hand (number and effect differing by algorithm). Therefore, researchers can use many variables as input and consider complex functions up to total mathematical black-boxes but still regulate their models to fit out-of-sample data. Note the difference compared with classic inferential statistics, in which models follow the idea of being most parsimonious. Although there is a wide array of potential algorithms to connect an outcome with features[5], the basic principle of almost all SML can be summarized in the following steps, which will be applied in Sect. 2.1 SML:

[3] At this point, it seems worthwhile remembering Weber's (1978) classic notion of "Erklären."

[4] Ideally, researchers may split data into three parts, training, test, and validation. The latter may be used to select among different model specifications. This step is often skipped in practice. Given the scope of this article as an introduction to both SML and UML, I will also only refer to training and test data.

[5] An overview is given by Lantz (2019). A specific application using Bayesian classifiers to predict economic growth is presented in Heiberger (2018).

1. Split data in training/test data (one should answer questions comprising: Split ratio? Scale of outcome? Number of features? Match data if necessary?).
2. Train a model (choose an algorithm linking features to outcome; decide what models perform best; tune model parameters).
3. Evaluate model accuracy (model fit by out-of-sample predictions with test data).[6]

2.2 UML

Unlike SML, there is no "supervisor" for UML and no pre-labeled data from which algorithms learn. Instead, UML tries to reveal patterns that are hidden in the data. It detects underlying structures in observations by exploiting statistical properties of the data. In essence, UML is aimed at creating categorization schemes or typologies. Researchers can then define types along the derived (latent) dimensions and represent each case relative to the types given its underlying values.

Often, researchers have no access to a ground truth in order to set the number of dimensions and validate models.[7] To determine types' fuzzy empirical boundaries, inductive approaches such as cluster analysis, principal components, or latent class analysis are often combined with theoretical considerations. The main purpose of UML is to explore data and reduce its complexity. Researchers might use the output as input for further analysis (e.g., Munoz-Najar Galvez et al. 2020) or to develop theoretical models (Hall and Soskice 2001).

Resulting (ideal) types are arguably among the most important methodological tools of social scientists and have been used for a long-time (Ahlquist and Breunig 2012). Thus, utilizing exploratory techniques is not at all new to social sciences; yet, UML does provide novel ways of analyzing large amounts of text and social networks, both kinds of data often associated with the digital age and computational social science (Heiberger and Riebling 2016; Lazer et al. 2009). In particular, the "automatic" categorization of large corpora has found many applications on social phenomena (Evans and Aceves 2016). Topic models represent one of the most frequently used Natural Language Processing (NLP) tools in the social sciences (McFarland et al. 2013). Its main idea is to summarize a large corpus of documents into relatively few meaningful themes (i.e., topics) and, hence, keep the most relevant information. For instance, social scientists used methods from "natural language processing" to reconstruct the discursive history of scientific fields (Wieczorek et al., 2021; Hall et al. 2008), analyze media effects on attitudes (Erhard et al. 2022), trace the fragmentation of political discourse (Heiberger et al. 2021a), or explain scientists' choice of research strategy (Evans and Foster 2011).

Social networks constitute another branch with deep roots in UML. Since the 1970s, network researchers have applied "blockmodeling" to find structural equivalent nodes and group them together (White et al. 1976). The rise of network data

[6] We might want to add a fourth step in which we test different algorithms and select the best model. Given the limited scope of this paper, I ignore this step.

[7] Of course, model validation and comparing "K" (number of dimensions) is an essential step in UML. I will illustrate this in the section "Using UML: Deriving Topics."

with the internet led to many new developments in this area, most often characterized as community detection (Fortunato 2010). Although many physicists are involved in developing new, mathematically sophisticated graph-partitioning methods, the idea is the same as for all UML: summarize data by finding its most important dimensions and/or group similar cases to derive types.

3 Application: Gender Differences in Scientific Publishing

3.1 The Role of Gender in Scholarly Authorship

To illustrate SML and UML in greater detail, I will now turn to an important case: gender differences in scientific publishing. Applying gender detection (SML) and topic models (UML) on a large sample of U.S. sociology dissertations will reveal new insights into how research topics are deeply divided by gendered preferences.

Despite growing awareness, gender differences in academia still persist across all disciplines and countries (Barone 2011; Holman et al. 2018; Huang et al. 2020; Larivière et al. 2013). At the center of interest lies the "productivity puzzle" (Xie and Shauman 1998), i.e., evidence that male researchers publish more than their female colleagues. Explanations point to many related differences, for instance, in collaboration practices (Abramo et al. 2019; Jadidi et al. 2017; Uhly et al. 2017), family responsibilities (Carr et al. 1998; Fox 2005), or rank of alma mater (van den Besselaar and Sandström 2017). Those results appear in a new light given recent results from Huang et al. (2020). By reconstructing over seven million researcher careers from a large sample of publications, they could show that gender differences in productivity and impact are stable, but that those differences are rooted in gender-specific dropout rates.

Although those findings have wide-ranging policy implications, there is another major aspect of gender biases in academic publishing that is still widely overlooked: research content and topics. Among the few exceptions, Nielsen et al. (2017) detect a higher likelihood in medical studies to include gender in their analyses if women are among the authors. Only recently, Key and Sumner (2019) find gendered research topics in political science. I will also refer to their results in the Discussion section.

3.2 Using SML: Detecting Gender

To further our understanding on gender preferences for certain research topics, I base my analysis on dissertations. Theses are a formal requirement for becoming part of the scientific community (Collins 2002). Trying to gain recognition as experts, potential graduates spend a long time on their respective projects. Most PhD candidates ponder the objectives and meanings of their thesis many times. Hence, chosen topics should reflect personal preferences, also because one's thesis is a strategic decision (Bourdieu 1988, p. 94). In addition, theses have to be single-authored, circumventing problems present in studies using research articles with several coauthors.

The data are retrieved from the ProQuest database and represent a close approximation of all US-based dissertations (Hofstra et al. 2020). To reflect the sociological

field, all theses written in a sociology department have been included ($N = 41{,}045$). Thus, the research topics represent a rather narrow perspective on sociology, excluding interrelated fields such as education or psychology.[8] Taken together, the analyzed texts comprise each dissertation's abstract and title and range from 1980 to 2015.

To derive gender, almost all of the cited studies (see the Section "The Role of Gender in Scholarly Authorship") use the first names of authors. Although this process is often moved to footnotes (if mentioned at all), I will now describe the three SML steps to classify gender from names and compare the proposed approach to a commercial application (genderize.io).

First and foremost, each SML needs *training data* (step 1). Regardless of the particular classification, pre-labelled data are needed to establish the statistical models from which to derive the desired classifications. The larger the amount of training data, the better the subsequent predictions. A major obstacle for this undertaking, hence, is to realize such "ground truths" in a sizeable fashion. Classifying training data get particularly expensive (time and/or money), when human coders (considered the "gold standard" of ML training sets) are needed.

One way of obtaining suitable training data is to use process-generated data, often derived, for instance, from public records or, in industry, from customer data bases, sales, or web logs. In the case of gender prediction, I will utilize the largest collection of names that is publicly available, the US Social Security Administration (SSA) record. It contains first names annually collected for each of the 355,149,899 babies born in the US between 1880 and 2019. Unlike West et al. (2013) or Karimi et al. (2016), we use the full records (all names with at least 10 occurrences per year). Like most social actions, however, name-giving yields a heavily skewed power-law distribution with relatively few high-frequency names (James, John, and Robert at the top for male names; Mary and Elizabeth being the most popular female names).

In total, 99,444 unique names have been awarded in the US since 1880. Most are associated with one gender exclusively, only 10,942 have been assigned to both sexes in all those years. Although we can assume a probability of 1 for the names that indicate solely one gender, ambiguous names provide us with an interesting case to apply SML.

Before performing the predictions, we need to match our *test data* to SSA. The test data comprise first names of PhD students of whom we do not know the gender. The sample initially contains 41,045 dissertations; of those, 37,437 first names are found in SSA. The other 3608 are either relatively rare Asian names (e.g., Byung) or double barrel names (e.g., Zxy-Yann), but mostly first names recorded only as single letters in the database, i.e., sort of missing data. That is a rather low number compared with other approaches, for instance, West et al. (2013) did not find more than 26%, or Hofstra and de Schipper (2018), who could not align more than 30% of their training with test data.

To apply SML, I will now focus on the 33,082 students who have names that are assigned to both sexes in SSA, resulting in 2545 unique names for which we want

[8] A different road has been taken in Heiberger et al. (2021b), where we analyze career prospects in sociology using a broader sample including neighboring fields. We still find similar trends, for instance, the rise of themes related to the cultural change.

Table 1 Overview of gender predictions for the ProQuest sample ($N = 41{,}045$) based on the SSA approach

	Female	Male	Unknown
Ambiguous names	18,079	12,350	2643
Unambiguous names	2915	1440	–
Not found	3608		

Table 2 Confusion matrix for gender predictions of the SSA approach. Results for genderize.io are reported in brackets. Ground truth results are based on 500 first names manually coded by three researchers. "False" means ambiguous codings of the human coders

		Prediction	
		False	*True*
Ground truth	*False*	50 (43) TN	30 (33) FP
	True	90 (124) FN	330 (293) TP

TN true negatives, *FN* false negatives, *FP* false positives, *TP* true positives

to predict the associated gender. For that purpose, we need to build a *probabilistic classifier* (step 2), i.e., a statistical model to link features (explanans, here: ambiguous first names) and outcome (explanandum, here: gender) to derive classifications based on the training data. In this article, I use logistic regression (generalized linear models, GLMs).[9] Although this is arguably one of the most popular methods in social sciences and should, hence, be familiar to most readers, it is only rarely used by social scientists to predict out-of-sample cases. That is in contrast to its use by computer scientists, who employ GLMs for predictions and see them as an essential part of ML tools (Lantz 2019).

To align the results of the proposed SSA approach in the next step to results achieved with genderize.io (Wais 2016), the probability is set to a rather strict level of 0.95. That means a student's gender is associated with being female or male if the model predicts that 95% of all times.[10] If the probability is lower, the case is set to "unknown." The results depicted in Tab. 1 show that more female students finish the PhD (around 56%), which is in accordance with official statistics (National Center for Education Statistics 2018). 2643 students (around 7%) cannot be assigned to a gender with the desired certainty. That is also a very convincing value compared with other studies (Karimi et al. 2016; Larivière et al. 2013; West et al. 2013).

However, the final and most important step of all SML is to assess the *accuracy of predictions* by comparing with to a "ground-truth" and/or other approaches (step 3). Accuracy is most often defined by calculating a confusion matrix as shown in Tab. 2. It evaluates the classification performance by counting the number of "true" and "false" instances. The ground truth in this article consists of 500 names (~ 20% of the test) for which three experts manually coded female or male names. Thus, we match each of the predictions of the machine to the gold-standard of human coders—the key to assessing the performance of each ML task.

[9] Many other potential models exist. For instance, interested readers may find an overview in Lantz (2019).

[10] Different thresholds are presented in the Appendix.

We compare the SSA predictions with one of the most popular databases for gender predictions, genderize.io, which has found many prominent applications in science (Huang et al. 2020). The simple SSA approach proposed here is clearly outperforming genderize.io. From Tab. 2 we can easily calculate accuracy $\left(\frac{(TN+TP)}{(TN+TP+FN+FP)}\right)$ and F1 score $\left(\frac{TP}{TP+\frac{1}{2}(FP+FN)}\right)$, two of the most important indicators of model quality in ML. Although the SSA achieves an accuracy of 0.76 and a F1 score of 0.85, genderize.io can only reach an accuracy of 0.68 and a F1 score of 0.79. These are rather large differences given SML tasks (Karimi et al. 2016).

In addition, the SSA data are completely open to the public and easily downloadable in a machine-readable format. Even more importantly, results for SSA-based predictions are fully reproducible because of the fixed set of names for a given time span. In contrast, genderize.io is continuously expanding its database, so that (other) researchers are not able to reproduce previous classification results. Finally, genderize.io is not free of charge for more than 1000 names per day, which is an additional disadvantage.

3.3 Using UML: Deriving Topics

One of the crucial yet often unspoken steps in working with large amounts of texts is to prepare and clean the data. To provide an appropriate "how-to" description, I will spell out those details before describing the UML applied here.

In a first step, all stopwords have been removed (e.g., "and," "or," "the").[11] After that, the words have been lemmatized. Lemmatization is a common step in NLP to reduce different forms of a word (e.g., singular and plural) to a common base form (e.g., "women" becomes "woman"). As final preprocessing step, I concatenated bigrams appearing more than 50 times (e.g., "united" and "states" become "united_states"). In so doing, we can detect meaningful phrases like "factor_analysis" or "statistical_significant" in the dissertation abstracts (Blaheta and Johnson 2001).

After preprocessing, I use topic modeling in order to reduce large quantities of text to meaningful dimensions. Topic models are a popular instance of UML (Jordan and Mitchell 2015). Such models assign documents in a corpus to a combination of topics. Topics are directly derived from documents by probabilistic algorithms and consist of words that co-occur across documents. In so-called generative models, each topic is seen as a probability distribution across all words of a given vocabulary, describing the likelihood of a word to be chosen as part of a certain topic. This likelihood is independent of the position of the word in a text, which is why it is most often referred to as a "bag-of-words" representation of documents. Although this assumption is clearly not realistic (e.g., grammar is ignored), it has proven to be very reliable in practical applications (Landauer 2007).

[11] In practice, there are further considerations. For details see, for instance, Schofield et al. (2017) who evaluate the removal of stopwords, or Heiberger and Munoz-Galvez (2021) who elaborate on the impact of preprocessing on topic model quality.

For a decade, topic models have become very popular in social sciences (Evans and Aceves 2016; McFarland et al. 2013). In particular, science of science studies make use of this sort of dimension reduction, for instance, by reconstructing the history of a field (Anderson et al. 2012; Hall et al. 2008), explaining scientists' choice of research strategy (Evans and Foster 2011), tracing researchers' interest changes (Jia et al. 2017), or relating relevant career outcomes to authors' topic choices in medicine (Hoppe et al. 2019) or education (Munoz-Najar Galvez et al. 2020).

The core of most topic models was proposed by Blei et al. (2003), the latent Dirichlet allocation (LDA). Given a desired number of topics k and a set of D documents containing words from a word vocabulary V, LDA models infer K topics that are each a multinomial distribution over V. Thus, topics are a mixture of words V with probability β of a word belonging to a topic. The more often words co-occur in documents, the higher the probability that the words constitute a topic. At the same time, a document is also considered as a mixture of topics, so that a single document can be assigned to multiple themes. The topic proportions are given by parameter θ. By design, all topics occur within each document; thus, the proportion of θ gives us the strength of the connection between a topic (itself an ordered vector of words) and a document. Finally, it is important to note that the sampling process of LDA and all its extensions draw for each topic and each document from an eponymous Dirichlet distribution. Hence, the same multinomial distribution is used for *all* documents in a corpus.

In this article, I use an extension of LDA called structural topic modeling (STM) (Roberts et al. 2014, 2016). Its key feature is to enable researchers to incorporate document metadata and utilize such information (e.g., year) to improve the consistent estimation of topics. The covariates of a document d are denoted as X_d. The basic model relies on the same process explained above. However, in an STM the topic proportions θ depend on a logistic-normal generalized regression. Thus, for each word a topic is drawn from the *document-specific* distribution for one document based on its covariates X_d, not only—as in the regular LDA[12]—on a general distribution that is the same for all documents. It has been shown in several simulations that the incorporation of covariates improves the results of the topic quality substantially (Roberts et al. 2014, 2016).

STM proved to be especially useful for longer periods of time and changing discourses (Farrell 2016), which suits the data at hand well, for it spans three decades of sociological dissertations. In addition, we can include gender as an additional covariate and therefore examine potential differences between male and female research preferences. Hence, the gender predictions done with SML that are described above will be utilized as a covariate to predict gender differences in the choice of research topics in U.S. dissertations from 1980 to 2015.

Like most UML (e.g., cluster analysis, principal component analysis) researchers have to set K even though the number of relevant dimensions is not known a priori. Insufficient numbers render models too coarse whereas high values could result in

[12] The same is true for a correlated topic model (CTM) developed by Blei and Lafferty (2007). A STM without covariates is an instance of the CTM.

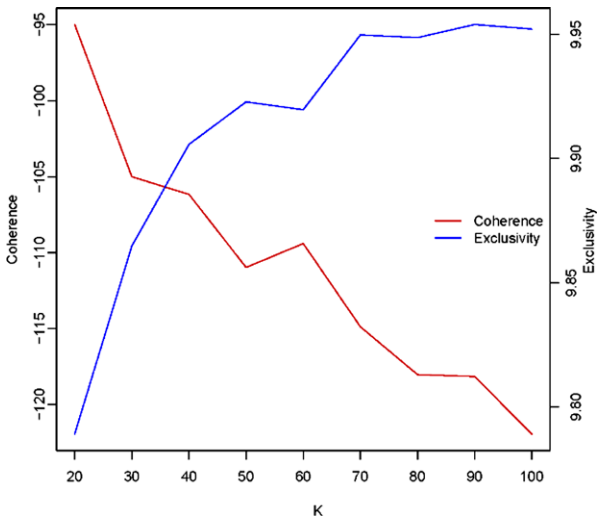

Fig. 1 Distribution of exclusivity (*right* y-axis) and semantic coherence (*left*) to approximate the number of topics (K)

very specialized subthemes. This is a widely recognized issue in topic modeling (Chang et al. 2009; Heiberger and Munoz-Galvez 2021) and requires elaborate, qualitative judgment of the researchers.

However, we can base our judgment on some established metrics, semantic coherence (Mimno et al. 2011) and exclusivity (Roberts et al. 2014). The coherence of a semantic space addresses whether a topic is internally consistent by calculating the frequency with which high-probability topic words tend to co-occur in documents. Yet, semantic coherence alone can be misleading as high values can simply be obtained by very common words of a topic that occur together in most documents. To account for the desired statistical discrimination between topics we therefore also consider a topic's exclusivity. This measure provides us with the extent to which the words of a topic are distinct to it. Both exclusivity and coherence complement each other and, hence, are examined in concert to give us an impression where topics represent word distributions in documents and at the same time provide differentiated dimensions. Accordingly, STM developers recommend that researchers look for the "semantic coherence-exclusivity frontier" (Roberts et al. 2014, p. 1070). We can observe such a "plateau" at K = 60 (Fig. 1). Given the trade-off between more exclusive, yet less coherent (in the upper sense) topics, those plateaus form the most parsimonious (i.e., smallest) choice of K.

3.4 Gender Preferences of Research Topics in US Sociological Dissertations

Topics consist of terms ordered by their probability of being used in a document that contains the given topic (denoted β above). Table 3 presents a ranking with FREX (Roberts et al. 2016), where a term is weighted by the harmonic mean of the word's rank in terms of frequency (FR) and exclusivity (EX) within a topic. For instance, topic 7's (T7) most descriptive words are "black," "neighborhood," "white," and its

Table 3 Overview of topics

Topics	Label	Probability	FREX
1	Noise	Opinion figure Canada street Canadian gamble	Des Canadian gamble Quebec Ontario Canada
2	Survey: responses	Interview use participant response identify method	Phase datum_collection instrument item reliability validity
3	Caregiver	Experience caregiver care participant adoption placement	Foster_care foster_parent adoptive caregiver child_welfare caregiving
4	Modeling	Model theory analysis test datum process	Model empirical theoretical_model theory propose theoretical
5	Crime	Crime victimization homicide criminal fear online	Fear_crime homicide social_disorganization crime_rate disaster violent_crime
6	Motherhood	Mother birth pregnancy maternal infant death	Birth_weight infant prenatal_care birth_outcome postpartum prenatal
7	Race	Black neighborhood white race poverty minority	Black_woman black segregation residential_segregation black_white white_black
8	Social support	Stress social_support cope well-being depression support	Depression social_support stress cope stressor psychological_distress
9	Work	Work worker job employment labor_market labor	Worker job employer occupational labor_market workplace
10	Native American	Indian cultural culture African indigenous history	American_Indian tribe Navajo Indian Cherokee tribal
11	African-American	African_American racial race white racism African-American	Racism whiteness gang prejudice racist race_relation
12	Language	Language English cluster linguistic use analysis	DNA speaker dialect linguistic genetic language
13	Justice	Offender case sentence juvenile court criminal_justice	Juvenile_justice sex_offender juvenile juror probation juvenile_court
14	Organization	Organization agency organizational service staff system	Human_service agency organization director staff nonprofit
15	Childhood	Child parent childhood parental home parent_child	Deaf child_maltreatment child maltreatment preschool abuse_neglect
16	Education	School student education educational college teacher	Student campus teacher black_student college classroom
17	Culture	Cultural practice culture discourse narrative way	Discursive discourse art narrative ethnography metaphor
18	Recreation	Information management site survey park recreation	Park hunter national_park user Utah visitor
19	Morality	Public moral issue claim debate frame	Moral ethic claim controversy morality public_opinion
20	Gender	Female male career gender male_female man	Saudi female career male_female gender_difference Arabia
21	Social networks	Group social individual status network member	Social_network friendship group network social tie
22	Social capital	Youth social_capital development volunteer engagement empowerment	Social_capital engagement young_people volunteer empowerment collaboration
23	Participation	Activity participation involvement leisure preference time	Leisure activity participation leisure_activity preference permission_author

Table 3 (Continued)

Topics	Label	Probability	FREX
24	Adolescent	Adolescent delinquency peer youth alcohol drug	Delinquency drink alcohol delinquent_behavior self-control substance
25	Political sociology	State political national citizenship nation elite	Nationalism Palestinian nationalist Russia nation-state national_identity
26	Social work	Practice professional social_work social_worker knowledge nurse	Social_worker social_work work_education professional profession work_practice
27	Social theory	Social society historical sociology theory modern	Sociology scientific intellectual writing science critique
28	Violence	Violence victim abuse domestic_violence report rape	Partner_violence IPV intimate_partner domestic_violence abuse batter
29	Identity	Identity experience lesbian participant boundary gay	Lesbian identity_development identity lesbian_gay gay queer
30	Economic sociology	Economic class development country inequality state	Cross-national economic_development inequality economic_growth Brazil welfare_state
31	Law enforcement	Police officer law_enforcement state security police_officer	Police_officer community_police officer patrol police law_enforcement
32	Community	Community urban city resident local house	Resident urban community city public_house neighbor
33	Family	Family family_member family_life family_structure familial resource	Family_function family family_structure family_system family_member family_cohesion
34	Life history	Life change experience leader leadership people	Leadership leader literacy life hope informant
35	Survey: scales	Variable measure attitude level scale relationship	Independent_variable score multiple_regression significant_relationship attitude_toward dependent_variable
36	Religion	Religious church religion Christian Catholic religiosity	Evangelical denomination congregation congregational religious church
37	Prison	Prison inmate incarceration incarcerate camp release	Inmate prison prisoner correctional camp incarcerate
38	Sexuality	Gender sexual man girl sex sexuality	Masculinity femininity sexuality musical girl masculine
39	Disability	Service old elderly care disability age	Disability elderly AFDC disable mental_retardation long-term_care
40	Social movement	Political movement social_movement politic activist organize	Social_movement activism activist protest movement movement_organization
41	Public policy	Policy state cost benefit federal reform	Policy cost tax incentive insurance payment
42	Experiments	Program treatment intervention group client evaluation	Control_group program drug_court experimental_group session intervention
43	HIV	Risk aid homeless HIV stigma HIV_AIDS	HIV HIV_AIDS condom HIV_risk HIV_infection homeless
44	Fatherhood	Father relationship parental parenting mother parent	Father parenting_style daughter father_involvement sibling parenting
45	Law	Law legal case court right rule	Law lawyer legal privacy litigation supreme_court

Table 3 (Continued)

Topics	Label	Probability	FREX
46	Development	Development environmental food farm land agricultural	Farmer irrigation sustainable_development NGOs agricultural water
47	Marriage	Marriage couple relationship marital divorce spouse	Marital_satisfaction marital couple cohabitation marriage husband_wife
48	Communication	Process conflict interaction communication strategy situation	Communication conflict interactional conversation interaction style
49	Households	Household rural migration income financial migrant	Household rural_area wealth rural_urban migration rural
50	Industry	Market industry production technology economy labor	Commodity industry trade coffee retail market
51	Public health	Patient medical health_care hospital mental_health care	Patient hospice physician cancer health_care medication
52	Corporations	Power institutional organizational business resource organization	Corporate business innovation entrepreneurship corporation diffusion
53	Hispanics	Health Hispanic acculturation Mexican_American health_status Latino	Obesity Puerto_Rican Hispanic Mexican-American Mexican_American BMI
54	Immigrants	Immigrant ethnic American unite_state cultural ethnicity	Immigrant Hmong immigration refugee Vietnamese Japanese
55	Sport	Satisfaction value consumer sport tourism retirement	Athlete sport tourism athletic football tourist
56	Feminism	Woman feminist experience American_woman gender interview	Woman infertility American_woman feminist woman_live African-American_woman
57	Socialization	Role expectation responsibility work socialization time	Work_family socialization role childcare widow role_strain
58	Fertility	Change fertility age effect rate increase	Fertility contraceptive force_participation cohort family_plan mortality
59	Media	Medium image television newspaper message suicide	Film television medium news portrayal viewer
60	Noise	Relationship implication find use associate good	Overt implication link issue_relate tendency covert

FREX are "black_woman," "black," "segregation." It seems intuitive to assume that a thesis with high loads of T7 is engaged in a topic regarding *Race.*

So, what did young sociologists in the US write theses on during the last 30 years? Table 3 gives an overview of derived topics. PhD students' research interests are widely spread, as was to be expected from a varied, fragmented discipline (Abbott 2001; Heiberger et al. 2021a). Topics comprise broad research themes used by many (e.g., T17 *Culture,* T35 *Survey: Scales*), thematic specialties (e.g., T25 *Political sociology*, T30 *Economic sociology*), methods (e.g., T42 *Experiments)*, topics crossing many social spheres (e.g., T7 *Race,* T21 *Social networks*), and concepts related to other disciplines (e.g., T26 *Social work*).

Although exploring those topics in greater detail might be a worthwhile undertaking (and can be done by *examining* Tab. 3), this article focuses on the application of SML and UML in order to reveal different choices of research topics by gender.

Fig. 2 Topic prevalence by year and gender

And indeed, we detect clear preferences for some of the most prevalent choices of students (Fig. 2). Although research on *Culture* (T17) and *Survey* (T2, T35) is almost equally spread across genders, we observe large differences when it comes to T4 *Modeling* and T27 *Social theory*. Both are much more frequently chosen by male students, in the case of T27 the probability is more than twice as high that a thesis on social theory is written by a male student.

Figure 2 also allows us to observe some general trends. In particular, research related to *Culture* (T17) is rising in popularity with students. This is connected to the influence of the "cultural change" on all social sciences (Jacobs and Spillman 2005). In contrast, US PhD students are writing about survey-related methods less and less frequently. T2 and T35 are constantly losing popularity. T35 started in 1980 as one of the most demanded topics and has been starkly declining ever since. This trend might also reflect more general research currents; at least, it is also observed for the discipline of education (Munoz-Najar Galvez et al. 2020).

Making further use of the STM results, we can also calculate topics exhibiting the largest differences across topics. For that purpose, we identify topics with an equal probability for both gender (i.e., similar distribution of topic load) and, in turn, topics revealing large differences. Thus, 0 represents no differences in topic usage, whereas higher values indicate deviations across gender preferences. The interpretation is straight-forward. For instance, a value of 2 for female preferences in a certain topic means that females have a probability two times higher than males of writing about that topic.

Figure 3 shows the five most pronounced differences for each gender. It reveals more than "fine distinctions." The list of female research preferences reads like a list of tasks traditionally assigned to women, ranging from motherhood (T6), childhood (T15) to socialization (T57) and caregiving (T3). The likelihood of engaging with these topics is at least twice as high for females than for their male colleagues. Even more striking, female PhD candidates in sociology are 6 times more likely to write about feminism (T56) than males. It is somewhat ironic that maybe the most important movement for gender equality is pretty gender specific. At least when it

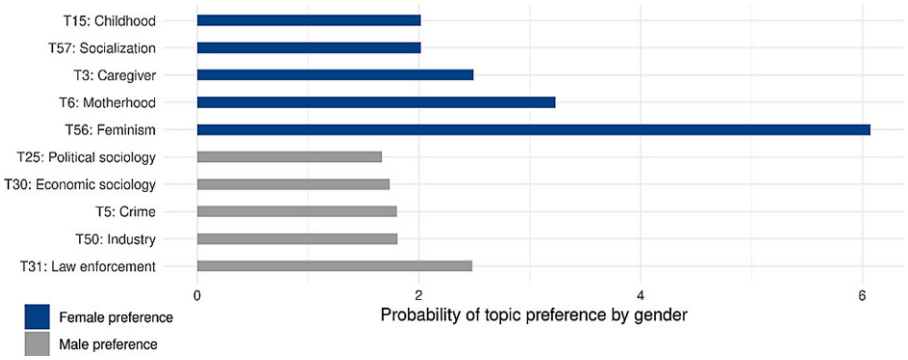

Fig. 3 Gender preferences of research topics in US sociology theses. The five largest differences for each gender are depicted

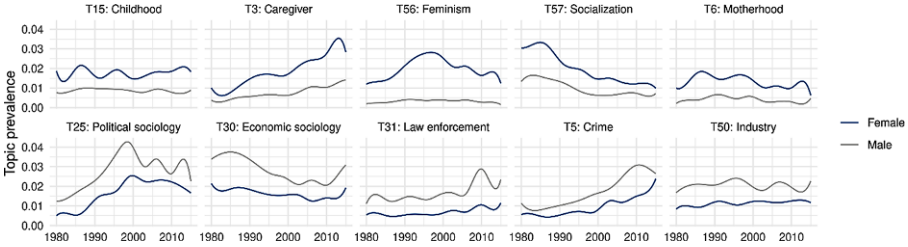

Fig. 4 Gender preferences of research topics in US sociology thesis, over time (1980–2015)

comes to topic choices in sociology dissertations, hence, preferences are clear-cut between the sexes.

In accordance, male preferences also lie in social arenas in which men occupy the majority (Fig. 3). Politics (T25), economy (T30, T50), and justice (T5, 31) exhibit the largest differences between sexes. In all three areas male PhDs have an around two times higher probability than their female equivalents of writing about those topics in their dissertation.

Now, one might object that times change. However, deviations remain considerable and differences are present for the whole observation period in most cases (Fig. 4). Yet, there are exceptions. For instance, the gap is closing in terms of *Socialization* (T57). In the 1980s, it was among the most popular choice for females and has declined ever since. In contrast, *Crime* (T5) has gained popularity across the sexes, though more among male students. The reverse is true for *Caregiver* (T3). T3 started in the 1980s at an equally marginal level, yet has attracted substantial interest since the 2000s, in particular, among female students. Although we observe some ups and downs, gender-specific majorities have not flipped in any case during the 35 years of observation, revealing strong and persistent gender differences in research preferences.[13]

[13] Results hold across different K (see Appendix).

4 Discussion

The article at hand serves three purposes: first, it provides an introduction to ML methods with a focus on social sciences; second, it applies ML methods and, in so doing, provides a "how-to" guideline for using SML and UML (including code); and, third, by applying ML it discloses substantial gender differences in research preferences for a large sample of dissertations written by US PhD students in sociology departments.

The substantive results shed new light on gender differences in academia. Despite an abundance of studies (Barone 2011; Holman et al. 2018; Huang et al. 2020; Larivière et al. 2013), research topics are a widely overlooked factor when it comes to gender biases in academic publishing. The results show a surprisingly clear picture: female PhD students in the US prefer topics such as *Caregiver, Motherhood*, or *Feminism*. The prevalence of those research areas is up to six times higher and at least more than twice as high for theses of female PhD students compared with their male counterparts. In contrast, theses written by men focus more than twice as often on *Law enforcement, Crime*, or *Economic sociology*. The pronounced gender preferences have been mostly stable for more than 30 years.

A potential explanation might be that those topics are closely related to real-life experiences of students, i.e., that men and women undergo different socialization processes, live through different societal expectations and roles, and, hence, develop different research interests (Key and Sumner 2019). In favor of this explanation, a comprehensive study finds that curricular choices are strongly influenced by gender-specific interests similar to those seen in the research topics of PhD students (Charles and Bradley 2009). Still, it is surprising that the long, and sometimes painful, yet highly reflexive process of writing a dissertation exhibits such a high degree of gender-bias. It may very well be though, as Key and Sumner (2019) suggest for political science, that many of the much-discussed biases in publication behavior of both sexes rest on the choice of research topics.

While obtaining those substantive insights, the article tries to inform social scientists on ML methods by applying them. The results on SML suggest a clear recommendation when it comes to gender prediction from first names, often used in science of science studies or content analysis. Using a GLM framework and SSA data yields greater accuracy than genderize.io, despite the platform's popularity among researchers (e.g., Huang et al. 2020). The rather simple approach presented here is not only more accurate in predicting gender, but free of charge and, even more importantly, replicable. Genderize.io is neither of those things.

That points to a larger issue of SML (and to a lesser degree UML)—the more data, the better your results. And the most data are obtained by tech companies such as Google, Microsoft, Facebook, Twitter, etc. I do not object to researcher's usage of data gathered or provided by companies as long as data are open and access for researchers unrestricted. Clearly, that is often not the case, given that the proprietary use of information constitutes large parts of the value of those companies. This issue is not exclusive to ML; yet, owing to its reliance on large (and well annotated) data to perform well, it is more apparent than in other parts of social sciences. Scientific research is a public good and needs to be reproducible for peers (Merton 1973); the

only solution to this issue seems therefore to be as transparent as possible, in both data and methods. The direct way to achieve this is to use open-source data and publish one's code (Heiberger and Riebling 2016). It implies neglect of data if they stem from non-open sources or cannot be provided to other researchers (if not the whole interested public) to replicate results or further research.

Another more technical reason for transparency is that most ML methods afford many decisions, some of which may change results considerably. However, this is not different to any other elaborated data collection or statistical analysis. Yet, given the complexity of many ML applications, it seems important to keep up social scientists' statistical rigor, i.e., include ML in the field's methodological canon by exposing it to the same thorough critique that any other quantitative analysis would be subjected to. Therefore, I would strongly suggest that one of the best solutions might be to use several options at crucial bifurcations (e.g., choice of K) and, hence, check the robustness of the results.

However, running ML is costly; re-running ML to check robustness or tune parameters even more so. In terms of time (computer power) and money (having large enough numbers of human-annotated data), ML makes existing differences in resources between institutes or research groups more pronounced. It seems therefore crucial to come up with suitable infrastructures so that structural possibilities do not restrict researchers in a fundamental way. In contrast, SML might provide incentives to close a longstanding gap in social sciences, that is, between qualitative and quantitative research. The crucial annotation of data might build a bridge and establish an innovative division of labor between often separated qualitative coding and quantitative inferences (Kang and Evans 2020).

It is important to note that many ML methods are not new to social scientists. On the contrary, the ML arsenal has been well-known in social sciences for decades; for instance, the popular "Ward" method of conducting hierarchical cluster analysis was published in the early 1960s (Ward 1963). Similar techniques for reducing data to their latent dimensions (i.e., clustering analysis) come in a new guise and are now often labeled UML. Building on that long-standing expertise, any of the various ML methods (two of which have been discussed here in some detail) should be readily accessible to researchers given a profound background in social science.

Although social scientists have been used to the idea of UML for a long time and apply UML to describe higher-order patterns and explore datasets, the logic of SML may be considered more novel. One fruitful way of exploiting the possibilities is shown in this article, i.e., using SML to predict an independent variable and put that to further use (for instance, see Heiberger et al. 2021a for more complex examples). Another idea is spelled out in detail by Watts (2014), arguing that out-of-sample predictions may improve sociological explanations and could be used as a "hard" test as to whether a model fits reality. Such out-of-sample tests would also help to amplify the reach of social scientists' results (by being applicable to other data), reduce barriers to replicating one's own results, and, hence, counter common "p-hacking" efforts (Molina and Garip 2019).

All that said, it is, in my opinion, important that social scientists use the possibilities offered by ML. One key facilitator to applying promising ML methods will be, of course, training students. Yet, it takes time for young researchers to enter the

field. Another, more subtle concern relates to current epistemological boundaries. Sociologists are trained to be skeptical. Therefore, they recognize ML not as a set of potentially powerful methods one could use but as a much-criticized research object (e.g., Weber 2016). I am not arguing that the latter is not worthwhile doing. Yet, *not applying ML* seems like not an option. If social scientists are not involved or act as mere bystanders in analyzing social phenomena with cutting-edge methods (of which ML is a prime example), other, more technical disciplines *will* do it, and are already doing it on a large-scale (see, for instance, an agenda formulated by physicists (Conte et al. 2012)). This article may play a humble part in paving the way to spreading the use of ML among social scientists by introducing some useful ML methods with an application case many researchers in the field might relate to—the divide of research topics by gender.

5 Appendix

The appendix presents results for crucial choices and how different thresholds affect the models. For the SML approach, different thresholds for assigning gender (baseline is a strict probability of 0.95) are presented in Tab. 4. For the topic models, Tab. 5 shows the most descriptive words for the most gender-specific topics across different K. It is noteworthy that the interpretation does not change if we take more fine-grained topic solutions, i.e., the findings presented above are robust in this respect. Figure 5 mimics Figs. 3 and 4 and depicts different choices of number of topics (K) in regard to the same topics' trends. Hence, results from SML as well as UML are robust to crucial model choices.

Table 4 Model fit across prediction thresholds. The threshold of predictions assigns a gender to a name with a given probability (default in the main model is 0.95)

Threshold	Accuracy		F1 Score	
	SSA	*Genderize.io*	*SSA*	*Genderize.io*
0.9	0.778	0.728	0.863	0.830
0.85	0.794	0.759	0.876	0.853
0.75	0.804	0.765	0.886	0.861
0.55	0.802	0.769	0.889	0.868

Table 5 Comparison of selected topics across different K. The topics represent those with the highest differences across gender (see Figs. 3 and 4). For each topic, four words with highest beta/FREX are depicted. A more extensive list with all topics, more words, or other K are available upon request

Label	K	Probability	FREX
Childhood	70	Child parent parental sibling	Deaf parent sibling preschool
	100	Child childhood development young_child	Deaf childhood handicap early_childhood
Caregiver	70	Adoption, child, placement, child_welfare	Foster_care, foster_parent, child_welfare, adoptive
	100	Adoption placement child_welfare foster_care	Foster_care foster_parent child_welfare adoptive
Feminism	70	Woman feminist life experience	Woman woman_live infertility feminist
	100	Gender feminist man sexuality	Masculinity femininity feminism feminist
Socialization	70	Role work career expectation	Work_family socialization role career
	100	Role career work responsibility	Work_family role role_strain stepfamily
Motherhood	70	Birth pregnancy infant mortality	Prenatal_care birth_weight infant postpartum
	100	Mother maternal infant birth	Birth_weight infant low_birth maternal
Political sociology	70	Political state national politic	Nationalist Palestinian nationalism empire
	100	State political national citizenship	Palestinian Russian citizenship civil_society
Economic sociology	70	Economic development class country	Cross-national economic_development economic_growth country
	100	Economic class inequality country	Inequality stratification cross-national economic
Law enforce	70	Police officer law_enforcement security	Community_police police_officer police patrol
	100	Police officer law_enforcement security	Police_officer community_police police patrol
Crime	70	Crime victimization homicide fear	Fear_crime disaster homicide bully
	100	Crime neighborhood victimization online	Fear_crime online social_disorganization bully
Industry	70	Market industry production global	Industry commodity market enterprise
	100	Market production global industry	Globalization global commodity enterprise

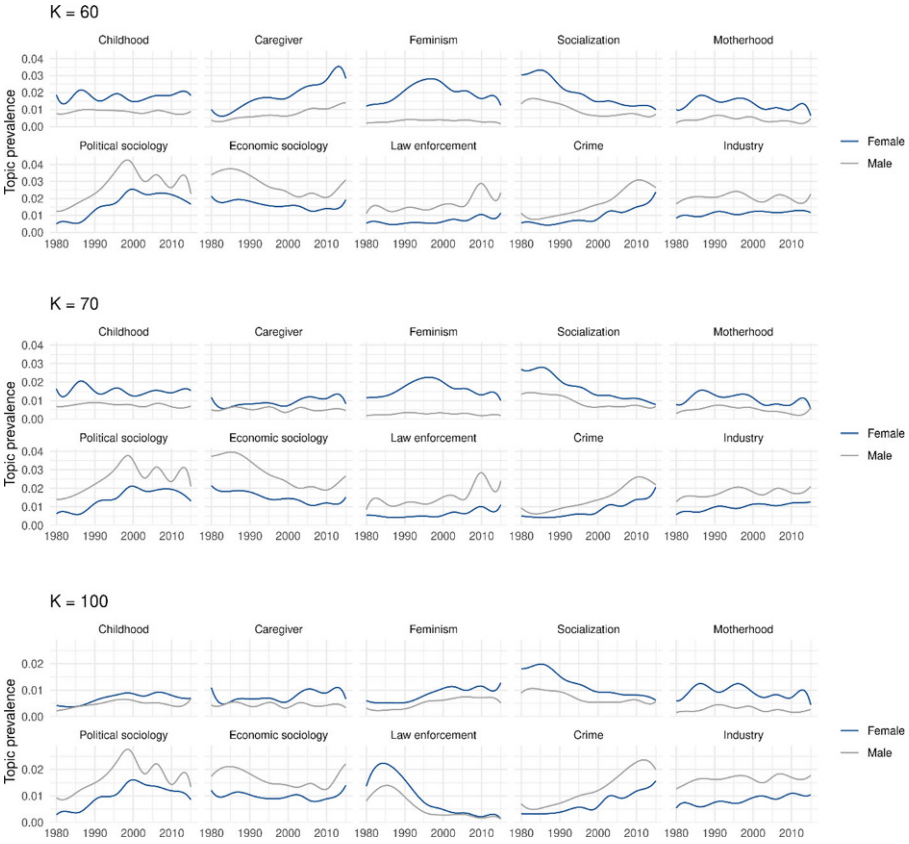

Fig. 5 Comparison of topic trends across different K. The topics represent those with the greatest differences across gender (see Figs. 3 and 4)

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abbott, Andrew. 2001. *Chaos of Disciplines*. Chicago: University of Chicago Press.

Abramo, Giovanni, Ciriaco Andrea D'Angelo and Flavia Di Costa. 2019. A Gender Analysis of Top Scientists' Collaboration Behavior: Evidence from Italy. *Scientometrics* 120(2):405–418.

Ahlquist, John S., and Christian Breunig. 2012. Model-Based Clustering and Typologies in the Social Sciences. *Political Analysis* 20(1):92–112.
Anderson, Ashton, Dan McFarland and Dan Jurafsky. 2012. Towards a Computational History of the ACL: 1980–2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, ACL '12*, 13–21. Stroudsburg, PA, USA: Association for Computational Linguistics.
Barone, Carlo. 2011. Some Things Never Change: Gender Segregation in Higher Education across Eight Nations and Three Decades. *Sociology of Education* 84(2):157–176.
Besselaar, Peter van den, and Ulf Sandström. 2017. Vicious Circles of Gender Bias, Lower Positions, and Lower Performance: Gender Differences in Scholarly Productivity and Impact. *PLOS ONE* 12(8):e0183301.
Blaheta, Don, and Mark Johnson. 2001. Unsupervised Learning of Multi-Word Verbs. In *Proceedings of the ACL 2001 workshop on collocation: computational extraction, analysis and exploitation*, 54–60. Association for Computational Linguistics (ACL).
Blei, David M., and John D. Lafferty. 2007. A correlated topic model of Science. *The Annals of Applied Statistics* 1(1):17–35. https://doi.org/10.1214/07-AOAS114.
Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3:993–1022. https://doi.org/10.5555/944919.944937.
Bourdieu, Pierre. 1988. *Homo Academicus*. Stanford University Press.
Carr, Phyllis L., Arlene S. Ash, Robert H. Friedman, Amy Scaramucci, Rosalind C. Barnett, Laura EDM Szalacha, Anita Palepu and Mark A. Moskowitz. 1998. Relation of Family Responsibilities and Gender to the Productivity and Career Satisfaction of Medical Faculty. *Annals of Internal Medicine* 129(7):532–538.
Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems 22*, eds. Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta, 288–296. Curran Associates, Inc.
Charles, Maria, and Karen Bradley. 2009. Indulging Our Gendered Selves? Sex Segregation by Field of Study in 44 Countries. *American Journal of Sociology* 114(4):924–976.
Collins, Randall. 2002. *The Sociology of Philosophies: A Global Theory of Intellectual Change*. Revised edition. Cambridge, Mass.: Belknap Press of Harvard University Press.
Conte, R., N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J. P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. San Miguel and D. Helbing. 2012. Manifesto of Computational Social Science. *The European Physical Journal Special Topics* 214(1):325–46.
Cranmer, Skyler J., and Bruce A. Desmarais. 2017. What Can We Learn from Predictive Modeling? *Political Analysis* 25(2):145–66.
Donoho, David. 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics* 26(4):745–66.
Erhard, Lukas, Michael Windzio and Raphael H. Heiberger. 2022. Diverse Effects of Mass Media on Concerns about Immigration: New Evidence from Germany, 2001–2016. *European Sociological Review*.
Evans, James A., and Pedro Aceves. 2016. Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology* 42(1):21–50.
Evans, James A., and Jacob G. Foster. 2011. Metaknowledge. *Science* 331(6018):721–725.
Farrell, Justin. 2016. Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences* 113(1):92–97. https://doi.org/10.1073/pnas.1509433112.
Fortunato, Santo. 2010. Community Detection in Graphs. *Physics Reports* 486(3–5):75–174.
Fox, Mary Frank. 2005. Gender, Family Characteristics, and Publication Productivity among Scientists. *Social Studies of Science* 35(1):131–150.
Hall, Peter A., and David W. Soskice. 2001. An Introduction to Varieties of Capitalism. In *Varieties of capitalism: The institutional foundations of comparative advantage*. 1–68. Oxford: Oxford University Press.
Hall, David, Daniel Jurafsky and Christopher D. Manning. 2008. Studying the History of Ideas Using Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08*, 363–371. Stroudsburg, PA, USA: Association for Computational Linguistics.
Heiberger, Raphael H. 2018. Predicting Economic Growth with Stock Networks. *Physica A: Statistical Mechanics and Its Applications* 489:102–111.
Heiberger, Raphael H., and Sebastian Munoz-Najar Galvez. 2021. Text mining and topic modelling. In *Handbook of Computational Social Science*. London: Routledge.

Heiberger, Raphael H., and Jan R. Riebling. 2016. Installing Computational Social Science: Facing the Challenges of New Information and Communication Technologies in Social Science. *Methodological Innovations* 9:1–11.

Heiberger, Raphael H., Silvia Majo-Vazquez, Laia Castro, Rasmus Nielsen and Frank Esser (2021a): Don't blame the media! The role of politicians and parties in fragmenting online political debate. *The International Journal of Press/Politics*. https://doi.org/10.1177/19401612211015122.

Heiberger, Raphael H., Sebastian Munoz-Najar Galvez and Daniel A. McFarland. 2021b. Facets of Specialization and Its Relation to Career Success: An Analysis of U.S. Sociology, 1980 to 2015. *American Sociological Review* 86(5):00031224211056267.

Hofstra, Bas, and Niek C. de Schipper. 2018. Predicting Ethnicity with First Names in Online Social Media Networks. *Big Data & Society* 5(1):2053951718761141.

Hofstra, Bas, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky and Daniel A. McFarland. 2020. The Diversity–Innovation Paradox in Science. *Proceedings of the National Academy of Sciences* 117(17):9284–9291.

Holman, Luke, Devi Stuart-Fox and Cindy E. Hauser. 2018. The Gender Gap in Science: How Long until Women Are Equally Represented? *PLOS Biology* 16(4):e2004956.

Hoppe, Travis A., Aviva Litovitz, Kristine A. Willis, Rebecca A. Meseroll, Matthew J. Perkins, B. Ian Hutchins, Alison F. Davis, Michael S. Lauer, Hannah A. Valantine, James M. Anderson and George M. Santangelo. 2019. Topic Choice Contributes to the Lower Rate of NIH Awards to African-American/Black Scientists. *Science Advances* 5(10), eaaw7238.

Huang, Junming, Alexander J. Gates, Roberta Sinatra and Albert-László Barabási. 2020. Historical Comparison of Gender Inequality in Scientific Careers across Countries and Disciplines. *Proceedings of the National Academy of Sciences* 117(9):4609–4616.

Jacobs, Mark D., and Lyn Spillman. 2005. Cultural Sociology at the Crossroads of the Discipline. *Poetics* 33(1):1–14.

Jadidi, Mohsen, Fariba Karimi, Haiko Lietz and Claudia Wagner. 2017. Gender Disparities in Science? Dropout, Productivity, Collaborations and Success of Male and Female Computer Scientists. *Advances in Complex Systems* 21(03n04):1750011.

Jia, Tao, Dashun Wang and Boleslaw K. Szymanski. 2017. Quantifying Patterns of Research-Interest Evolution. *Nature Human Behaviour* 1(4):1–7.

Jordan, Michael I., and Tom M. Mitchell. 2015. Machine Learning: Trends, Perspectives, and Prospects. *Science* 349(6245):255–360.

Kang, Donghyun, and James Evans. 2020. Against Method: Exploding the Boundary between Qualitative and Quantitative Studies of Science. *Quantitative Science Studies* 1(3):930–944.

Karimi, Fariba, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi and Markus Strohmaier. 2016. Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods. In *Proceedings of the 25th International conference companion on World Wide Web.* 53–54.

Key, Ellen M., and Jane Lawrence Sumner. 2019. You Research Like a Girl: Gendered Research Agendas and Their Implications. *PS: Political Science & Politics* 52(4):663–668.

Landauer, Thomas. 2007. *Handbook of Latent Semantic Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Lantz, Brett. 2019. *Machine Learning with R: Expert Techniques for Predictive Modeling*. Birmingham: Packt Publishing.

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin and Cassidy R. Sugimoto. 2013. Bibliometrics: Global Gender Disparities in Science. *Nature News* 504(7479):211.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy and Marshall Van Alstyne. 2009. Computational Social Science. *Science* 323(5915): 721–723.

McFarland, Daniel A., Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D. Manning and Daniel Jurafsky. 2013. Differentiating Language Usage through Topic Models. *Poetics* 41(6):607–25.

Merton, Robert K. 1973. *The Sociology of Science*. Chicago: The University of Chicago Press.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* 262–272.

Molina, Mario, and Filiz Garip. 2019. Machine Learning for Sociology. *Annual Review of Sociology* 45.

Mullainathan, Sendhil, and Jann Spiess. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* 31(2):87–106.

Munoz-Najar Galvez, Sebastian, Raphael H. Heiberger and Daniel McFarland. 2020. Paradigm Wars Revisited: A Cartography of Graduate Research in the Field of Education (1980–2010). *American Educational Research Journal* 57(2):612–652.
National Center for Education Statistics. 2018. Postsecondary Degree Trends. Retrieved from https://nces.ed.gov/programs/digest/d16/tables/dt16_325.92.asp.
Nielsen, Mathias Wullum, Jens Peter Andersen, Londa Schiebinger and Jesper W. Schneider. 2017. One and a Half Million Medical Papers Reveal a Link between Author Gender and Attention to Gender and Sex Analysis. *Nature Human Behaviour* 1(11):791–796.
Roberts, Margaret, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58(4):1064–1082.
Roberts, Margaret E., Brandon M. Stewart and Edoardo M. Airoldi. 2016. A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association* 111(515):988–1003.
Schofield, Alexandra, Måns Magnusson and David Mimno. 2017. Pulling Out the Stops: Rethinking Stopword Removal for Topic Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* 432–436. Valencia, Spain: Association for Computational Linguistics.
Uhly, Katrina M., Laura M. Visser and Kathrin S. Zippel. 2017. Gendered Patterns in International Research Collaborations in Academia. *Studies in Higher Education* 42(4):760–782.
Wais, Kamil. 2016. Gender Prediction Methods Based on First Names with GenderizeR. *The R Journal* 8(1):17–37.
Ward, Joe H. 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58(301):236–44.
Watts, Duncan J. 2014. Common Sense and Sociological Explanations. *American Journal of Sociology* 120(2):313–51.
Weber, Jutta. 2016. Keep Adding. On Kill Lists, Drone Warfare and the Politics of Databases. *Environment and Planning D: Society and Space* 34(1):107–125.
Weber, Max. 1978. *Economy and Society: An Outline of Interpretative Sociology*. Berkeley: University of California Press.
West, Jevin D., Jennifer Jacquet, Molly M. King, Shelley J. Correll and Carl T. Bergstrom. 2013. The Role of Gender in Scholarly Authorship. *PLoS ONE* 8(7):e66212
White, Harrison C., Scott A. Boorman and Ronald L. Breiger. 1976. Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. *American Journal of Sociology* 81(4):730–780.
Wieczorek, Oliver, Said Unger, Jan Riebling, Lukas Erhard, Christian Koß and Raphael H. Heiberger. 2021. Mapping the field of psychology: Trends in research topics 1995–2015. *Scientometrics* 1–33.
Xie, Yu, and Kimberlee A. Shauman. 1998. Sex Differences in Research Productivity: New Evidence about an Old Puzzle. *American Sociological Review* 847–870.
Yarkoni, Tal, and Jacob Westfall. 2017. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science* 12(6):1100–1122.

Raphael H. Heiberger 1982, Jun.-Prof. Dr., Computational Social Science, University Stuttgart. Research areas: Statistical models of social phenomena, (scientific) career patterns, political polarization, and financial markets. Publications: Facets of Specialization and its Relation to Career Success: An Analysis of U.S. Sociology, 1980 to 2015. American Sociological Review, 2021 (with S. Munoz-Najar Galvez and D. A. McFarland); Don't blame the media! The role of politicians and parties in fragmenting online political debate. The International Journal of Press/Politics, 2021 (with S. Majo-Vazquez, L. Castro, R. Nielsen and F. Esser).