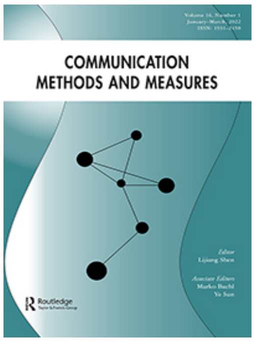


Communication Methods and Measures

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hcms20

Expert-Informed Topic Models for Document Set Discovery

Eike Mark Rinke, Timo Dobbrick, Charlotte Löb, Cäcilia Zirn & Hartmut Wessler

To cite this article: Eike Mark Rinke, Timo Dobbrick, Charlotte Löb, Cäcilia Zirn & Hartmut Wessler (2022) Expert-Informed Topic Models for Document Set Discovery, Communication Methods and Measures, 16:1, 39-58, DOI: 10.1080/19312458.2021.1920008

To link to this article: https://doi.org/10.1080/19312458.2021.1920008

Published online: 30 Jun 2021.

COMMUNICATION METHODS AND MEASURES
2022, VOL. 16, NO. 1, 39–58
https://doi.org/10.1080/19312458.2021.1920008

Routledge
Taylor & Francis Group

Expert-Informed Topic Models for Document Set Discovery

Eike Mark Rinke[a,b], Timo Dobbrick[b], Charlotte Löb[c], Cäcilia Zirn[d], and Hartmut Wessler[c]

[a]School of Politics and International Studies, University of Leeds, Leeds, UK; [b]Mannheim Centre for European Social Research (MZES), University of Mannheim, Mannheim, Germany; [c]Institute for Media and Communication Studies, University of Mannheim, Mannheim, Germany; [d]Data and Web Science Group, University of Mannheim, Mannheim, Germany

ABSTRACT

The first step in many text-as-data studies is to find documents that address a specific topic within a larger document set. Researchers often rely on simple keyword searches to do this, even though this may introduce considerable selection bias. Such bias may be even greater when researchers lack the domain knowledge required to make informed search decisions, for example, in cross-national research or research on unfamiliar social contexts. We propose *expert-informed topic modeling* (EITM) as a hybrid approach to tackle this problem. EITM combines the validity of external domain knowledge captured through expert surveys with probabilistic topic models to help researchers identify subsets of documents that cover initially unknown domain-specific topics, such as specific events and debates, that belong to a researcher-defined master topic. EITM is a flexible and efficient approach to the thematic selection of documents from large text corpora for further study. We benchmark and validate the method by discovering blog posts that address the public role of religion within large corpora of Australian, Swiss, and Turkish blog posts and provide researchers with a complete workflow to guide the application of EITM in their own work.

Reproduction Materials: The validation data, code, and any additional materials required to reproduce all analyses in this article are available in an online appendix, at: https://eitm-docs.github.io/.

Few recent developments have transformed social science as much as the exploding availability of large amounts of text. Text-as-data research has become one of the most vital and expanding areas of activity in the social sciences, with many research projects applying new methods of collecting (e.g., Landers et al., 2016; Mahdavi, 2019) and analyzing text data (e.g., Brier & Hopp, 2010; van Atteveldt & Peng, 2018).

The first step of many text-as-data studies is to find in a larger corpus of texts those documents that speak to a predefined theme – or master topic – of interest. Simply put, researchers are often not interested in all text documents available to them but need to find thematically defined subsets of documents within larger corpora. They may, for example, want to generate manageable text datasets for further, more refined analysis. Or they may want to identify the population of all documents belonging to a specific class of topics to draw a sample from for in-depth manual content analysis.

However, researchers often neglect the step of identifying document subsets that speak to a prespecified theme even though this step can have huge implications for the validity of results coming from further analysis of these subsets. Different subset selection processes generate different subsets,

CONTACT Eike Mark Rinke e.m.rinke@leeds.ac.uk School of Politics and International Studies, University of Leeds, Leeds, LS2 9JT

which, in turn, can lead to dramatically different statistical inferences and substantive conclusions prone to unknown biases.

The current state-of-the-art approach to subset discovery from unstructured text corpora is to use computer-assisted Boolean keyword searches, in which an algorithm proposes suitable keywords and creates Boolean search queries based on the text corpus at hand (King et al., 2017). But while this selection approach constitutes a significant advance over the bias-prone ad-hoc selection of keywords for search queries, it becomes problematic when researchers lack the knowledge required to make informed decisions about which of the proposed keywords to select. Unfortunately, such lack of requisite domain knowledge often obtains, especially in research involving text data from unfamiliar social contexts. In this article, we propose a hybrid approach, *expert-informed topic models* (EITM), that combines external domain knowledge captured through expert surveys with fully automated clustering methods to help researchers identify those subsets of documents that cover domain-specific topics within a researcher-defined master topic. Our approach uses inputs from expert surveys to enhance document classification based on probabilistic topic models. We provide researchers with a complete workflow to guide their own applications and illustrate the approach in a case study identifying articles that discuss the public role of religion from a large corpus of news articles including three languages (English, German, and Turkish).

We begin by describing the applied classification problem addressed in this article, outlining the state of the art of existing approaches to it and highlighting the issues these approaches cannot solve and which we tackle with the EITM approach. We then discuss the utility of an expert-informed approach to the discovery of document subsets from large corpora of unstructured text, especially for cross-national comparative research designs, and explain the seven-step EITM approach to combining expert input and document classification based on LDA topic models.

Given that model validation is a key practice in text-as-data research (Grimmer & Stewart, 2013), our approach involves a two-step validation process: First, we validate different models in terms of how well their estimated topics cohere and capture expert knowledge, using text data from five countries (Australia, Germany, Switzerland, the USA, and Turkey) and three media types (printed newspapers, news websites, and political blogs). Second, we take the EITM that is optimally coherent and representative of expert inputs and evaluate its performance as a thematic classifier using text data collected from political blogs based in three of these countries (Australia, Switzerland, and Turkey) and including all three languages. Detailed information about all data-collection, topic-modeling, and document-classification processes as well as any data, code, and additional materials required to adapt and implement EITM are available in a comprehensive online appendix (https://eitm-docs.github.io/), to help researchers reproduce the method and adapt it to their research projects.

The applied classification problem: Why bring expert domain knowledge to topic modeling?

Assume the following situation: A researcher has access to a large collection of news articles from several countries and wants to compare how minority rights are reported on in these countries. However, the articles are unlabeled and she knows that only a small – perhaps tiny – proportion of articles, which could still be huge in absolute numbers, will talk about minority rights. Even worse, the researcher is not familiar with the public debates that happened during the time in which the articles were collected in some of the countries she wants to study. This means that she will likely be unaware of important debates relevant to minority rights that took place during data collection and, importantly, many such articles will not include simple keywords such as “minority” or “minority rights”, which she would be forced to use in a naïve search query approach to finding the subset of relevant documents. This means that such a naïve approach is likely to introduce considerable, yet unknown and unknowable, bias into her study, threatening the validity of her conclusions. Simply put, our researcher will just miss any debate and story related to minority rights she was not aware of and did not have a fitting keyword for in her search query. This is a common research situation in comparative text-as-data studies, and one which in the past has been ignored at the expense of the validity of this area of study. It is also a research situation that is bound to become more common as ever greater amounts of textual data from all around the world become available to researchers. In this study we tackle this problem head-on.

More formally, our problem is one of document classification for very low-prevalence classes: We have a large unlabeled corpus and want to classify each document in that corpus as either thematically relevant or irrelevant, with the expectation that the proportion of relevant documents will be very small. So, we are not interested in the full corpus, just in the potentially small fraction of documents belonging to an unknown set of relevant categories, which we want to fold into a binary relevant/irrelevant distinction. That distinction is based not on a specific, narrow topic or event but on a nonspecific, broad thematic area of interest (i.e., a "master topic"). In the fictive example at the beginning of this section, the master topic is *minority rights*; in the use case we present below it is *the public role of religion*. The broadness of any given master topic means that we need to assume unknown latent sub-topics. Further, the involved researchers do not have sufficient domain knowledge about the social context to correctly classify the documents without additional input. Our solution is to bring expert-provided domain-specific knowledge to bear on unsupervised thematic document discovery. We exploit such knowledge to improve the efficiency and validity of both the initial discovery of themes and debates from unsupervised text modeling and the subsequent discovery of relevant document sets based on it. Since in this scenario the data is highly unbalanced, it is virtually impossible to find enough positive cases to produce sufficient training data for traditional supervised learning approaches to document classification and discovery. We mitigate the "missing knowledge" problem by incorporating expert domain knowledge about the latent sub-topics of our master topic (i.e., in our fictive example: individual events, debates or issues that are related to minority rights) into an LDA-based classification process. Incorporating expert knowledge is difficult to do in classical classification approaches as they are more closed than LDA-based topic models: Topic models include an inductive step of estimating topics directly from the analyzed document set that is missing from "classical" supervised approaches. In the EITM process, we exploit this step to input expert-provided domain information into the document discovery process. While LDA-based approaches have been rarely used for classification tasks, we thus harness their relative openness to address this "closedness problem" of established supervised machine learning approaches.

The most common strategy for researchers to discover documents covering a pre-specified master topic has been to run keyword-based Boolean search queries on the universe of documents available to them (Barberá et al., 2021). Therefore, we present a comparison of EITM's document discovery performance with the common simple search-query approach below.

While simple search queries are most commonly found in applied research, the state-of-the-art methodological work in the area of document set discovery is a keyword-search-query approach proposed by King et al. (2017; KLR). The goals of their approach and EITM are the same: to find the document sub-corpus of thematic interest to the researcher. In other words, the analytical goal is to discover, within a given document set, the subset of thematically relevant documents. This corresponds to what King et al. describe as selecting a *target set* from a search set of documents (King et al., 2017, pp. 975–976) and Barberá et al. as selecting the document *population* of interest from the universe of available documents (Barberá et al., 2021, p. 21).

EITM shares some important commonalities with the KLR approach. Both are computer-assisted approaches that draw on the unique ability of humans to recognize relevant keywords for keyword and document set discovery. Both also use these human inputs in interaction with the outputs of advanced automated text classification methods. Finally, both approaches involve the same broad stages of document identification: seeding, generalizing, validating, and discovering. Seeding refers to the input of human-generated information ("seeds") into the computational process; generalizing refers to the formulation of general concepts (e.g., keywords or topics) from specific examples; validating refers to the human correction of intermediate process outputs; and discovering refers to the final selection of the thematically relevant documents (i.e., the sub-corpus of thematic interest – the target set in King et al.'s or population in Barberá et al.'s terms) from the larger document set.

But the EITM approach also differs from the KLR approach in important ways: First, in the nature of the human inputs it leverages, which EITM formalizes as coming from domain experts rather than the researchers themselves.[1] Second, it differs in the steps involved and functions of the main stages in the document discovery process. Table 1 gives an overview of the main stages as implemented in the KLR and EITM approaches and summarizes the differences in how the two approaches tackle the problem of thematic document set discovery (a more detailed breakdown of the practical steps involved in EITM is given in the next sections and in Figure 1).

With their computer-assisted approach to keyword and document set discovery, the KLR approach frees human coders from the burden of creating keyword lists and queries on their own. But the approach still relies on preexisting "detailed contextual knowledge" (King et al., 2017, p. 979) of human coders, which is required to distinguish relevant from irrelevant keywords proposed by an algorithm. And as Barberá et al. (Barberá et al., 2021, p. 20) note, keyword-generation and -expansion methods must start with a human selection of keywords to seed the algorithm. This means that researchers need to be careful in thinking about which seed keywords are both relevant and representative in order to avoid biasing path dependencies.

The EITM approach lowers the need for detailed contextual knowledge on part of the researcher by using expert inputs to highlight relevant keywords and facilitate human-coder learning about relevant

Table 1. Main Stages of the King, Lam, and Roberts Keyword Algorithm (KLR) and Expert-Informed Topic Models (EITM) for Document Set Discovery.

Stage		KLR	EITM
1	Seeding	of keywords for producing training document set	of keywords for producing expert inputs (Step 1)
2	Generalizing	of training document set to suggested keywords	of expert inputs to ranked estimated topics (Steps 2–4)
3	Validating	of suggested keywords for document discovery	of ranked estimated topics for document discovery (Step 5)
4	Discovering	of target document subset via search query	of target document subset via model-based probability estimation (Steps 6–7)

The table displays a simplified overview of both approaches. Each stage listed is composed of multiple steps in practice (individual EITM steps as shown in Figure 1 in parentheses). Both approaches can be adapted, for example, by choosing different document clustering approaches or inserting more points of human input or iterating the process.

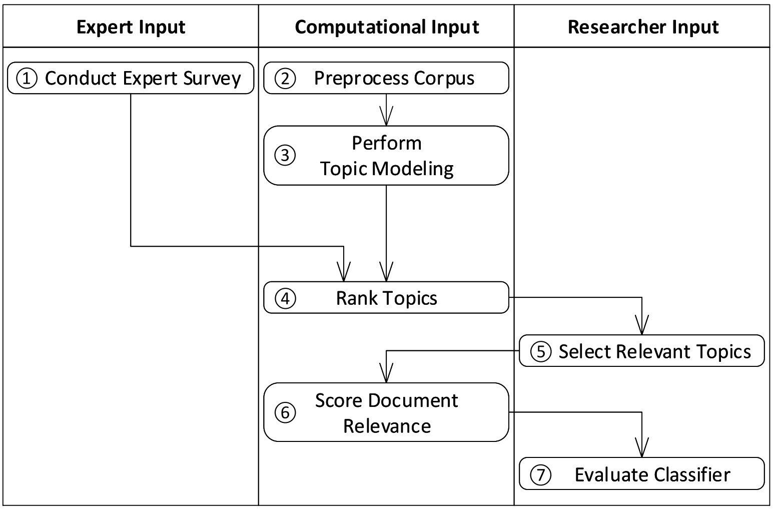


Figure 1. Overview of the EITM process.

contexts. As a result, applied researchers depend less on an assumption of preexisting contextual knowledge in human coders. Of course, experts also need to come up with relevant keywords, which is why it is important to (a) survey experts that can plausibly be expected to have good knowledge of the subject domain of interest and (b) recruit a reasonable number of experts to ensure that the coverage (or "recall" in machine learning terms) of keywords indicating relevant topics (i.e., those belonging to the researcher-specified master topic of interest) is sufficiently high. This is crucial given that in King et al.'s (2017) study the authors found that most keywords were named by only one of 43 non-expert subjects tasked with providing relevant keywords for a search query. Proper thematic coverage thus requires a proper number of keyword-providing experts.

A sufficiently large group of external experts also solves a second problem. If keyword candidates are computer-generated and must be evaluated by humans, as King et al. (2017) propose, and if the primary researchers do not have sufficient relevant domain and context knowledge, it is preferable to have external experts propose the keywords in the first place, not only to improve document set discovery but also to support researcher learning about the contexts studied. Expert inputs support not only a more valid discovery of target document sets, but also a more locally informed, "emic" understanding of studied contexts. As such, bringing in experts also reorients the process of knowing about the cultures we study in large-scale text analysis. It makes good on a demand often placed on cross-cultural research in modern comparative methodology by bestowing emic ("insider") validity upon an etic ("outsider") comparative perspective (Whitaker, 2017).

Another key advantage of our topic modeling approach over keyword search approaches like the KLR algorithm is that it generates a continuous distribution of relevance scores rather than a sharp binary classification. In this context, it is important to recognize that (semi-)automated pre-classification will never be perfect. However, the approach we propose is more flexible than keyword queries in that it allows the researcher to adjust the pre-classification parameters to their research situation and the acceptable level of tolerance for classification error. For example, if a researcher has few resources at their hands for conducting follow-up human validation of automated document classifications, they may want to select only documents in the uppermost part of the estimated relevance distribution. If they have larger resources for final human classification at their disposal but want to implement their project in a more efficient manner, they can work their way down from the top of the estimated document relevance distribution in a manual follow-up classification step until they have reached their desired sample size. We provide an illustration of the efficiency gains achieved through this process below. In sum, the EITM approach provides a classification framework that is more flexible, contextually valid and less reliant on assumptions about researcher domain knowledge than the available approaches based on Boolean keyword search queries.

Expert-informed topic models

Figure 1 provides an overview of the seven-step EITM approach and shows how it combines expert, computational, and researcher inputs to produce valid scalable solutions to topic-based document subset discovery. When applied to a document discovery problem in which a researcher aims to discover a subset of documents that talk about specific topics (i.e., issues or debates), the approach asks domain experts to name the topics and keywords that appeared in relevant thematic discussions during the period for which data was collected (Step 1). While the task of recruiting and surveying knowledge domain experts requires additional effort compared to established document set discovery approaches, it will often require fewer resources than making the substantial investment of resources required for researchers to acquire comprehensive expert knowledge themselves. This will be true especially in situations where studied social contexts are highly unfamiliar and language barriers to the studied texts are high for the researcher. Depending on research interest, experts may be substantive specialists in specific subject domains or area specialists in specific geographical regions.

After standard preprocessing of the full document corpus (Step 2) and estimating an LDA topic model on it (Step 3), the EITM approach uses expert inputs to rank the generated topics by their estimated relevance (Step 4). The topic ranking then serves as a basis for a manually validated binary classification of estimated topics as either belonging or not belonging to the researcher-specified master topic (Step 5). Importantly, the manual binary validation of estimated topic relevance in Step 5 is again informed by the comprehensive expert-provided list of relevant topics. After the top-ranked or all (if resources allow) estimated topics are human-validated as relevant or irrelevant, we estimate, for each document, its cumulative probability of containing one or more of the relevant topics, which can then be used for classification ("discovery") decisions about individual documents based on a researcher-specified relevance probability cutoff (Step 6).

Finally, in keeping with the principle of validating any given implementation in automated text analysis (Grimmer & Stewart, 2013, p. 271), we evaluate the resulting classifier using the receiver operating characteristic (ROC) curve, classifier precision, recall, and F_1-scores as established classification performance metrics (Step 7). We provide a detailed description of each step below.

The key concept of the EITM approach is the cumulative match probability (CMP), which is the central expert-informed output generated by the process. The CMP formalizes qualitative expert input into a continuous "relevance probability metric" for each topic estimated in the unsupervised topic modeling process. This metric allows researchers to infer the thematic relevance of any given estimated topic based on a prespecified probability cutoff value or, alternatively, supports human coders in a manual decision task to validate the thematic relevance or irrelevance of individual estimated topics, informed by the expert-provided insights about relevant topics.

The EITM process integrates three types of input. *Expert input* fulfills the dual function of increasing both the efficiency with which researchers can prescreen estimated topics for thematic relevance and the validity with which they can identify previously unknown relevant topics. *Computational input* in the form of inductive, unsupervised Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003; see also Guo et al., 2016; Jacobi et al., 2016; Maier et al., 2018) combines with expert and researcher inputs to identify relevant topics and documents while building on state-of-the-art guidelines for applying such models in the social sciences (Trilling & Jonkman, 2018; van Atteveldt & Peng, 2018). *Researcher input* into the EITM approach serves to enhance the selection of relevant topics over a fully automated solution and keep with the crucial principle of validation in automated text analysis (Grimmer & Stewart, 2013, p. 271).

In sum, EITM ensures that not only documents containing known relevant topics, but also those containing previously unknown relevant topics will, on average, be assigned higher relevance scores based on topic model estimates. These higher, and more valid, relevance scores assigned to documents, in turn, make it more likely that they will correctly be identified as thematically relevant by the research team. In essence, the EITM approach represents a scalable solution to increase the recall, or true positive rate, in complex, large-scale content-based document set discovery tasks. We note that, of course, this solution will not, and cannot, be perfect. Classification errors will still be made. The degree to which such errors are deemed acceptable to researchers will depend on the specific research situation. For example, in situations where document set discovery is a step in the preparation of a fully automated follow-up text analysis process, classification errors may be more tolerable than in situations where document set discovery is used to identify the universe of documents to draw a representative random sample for further in-depth human coding. In the latter case, EITM can help researchers conduct a pre-selection process that is more efficient and substantively valid than the standard approach based on researcher-generated keyword searches (see Barberá et al., 2021, pp. 21–26). The outputs of the EITM process may then be subjected to follow-up manual screening by human coders to ensure that random sampled documents are indeed thematically relevant. This manual process is not strictly part of the EITM approach, but it will again be informed and improved to the extent that human coders learn about relevant topics from the expert inputs acquired in the EITM process.

The EITM workflow

The following sections describe each step of the EITM process in detail. We illustrate its application and demonstrate its validity and classification performance using the Mannheim International News Discourse Data Set (MIND, Rinke et al., 2019). Our validation is based on a complete collection of all news items published by 94 different news websites, printed newspapers, and political blogs from five countries (Australia, Germany, Switzerland, Turkey, and the United States) in the full year ranging from August 1, 2015 until July 31, 2016. Our use case is a document set discovery task in which we want to identify, from the full corpus of all articles published by these media outlets in a full year, those articles that belong to the master topic "*public role of religion in societal life*". Appendix A provides detailed information about MIND and the specific data underlying our use case.

Step 1: Conduct expert survey

In the first step of the EITM process, expert inputs are generated to later inform the classification of estimated topics as thematically relevant or irrelevant. In our case, we tapped the knowledge of domain experts in each of the countries from which text material was collected using semi-standardized surveys that followed standards for expert selection aimed at mitigating bias used in past cross-national expert surveys (e.g., Kopecký et al., 2016). The survey asked selected experts to name topics and debates that occurred during the time period studied. Experts were also asked to provide keywords associated with these topics and debates.[2] We then grouped similar topics provided by different experts into a smaller number of meta-topics using the online qualitative content analysis tool QCAmap (Mayring, 2014). Keywords associated with such topics were pooled on the meta-topic level. Expert-provided topics were assigned to one or several meta-topics leading to a consolidated array of non-disjunct meta-topics (see Appendix B2). The meta-topics were used in Step 3 to evaluate the quality of estimated topic models by comparing the keywords in each meta-topic to the top keywords of each LDA-generated topic.

Overall, we identified a pool of 614 experts who were contacted to take the survey. The survey was fielded from August 15 until November 18, 2016 and generated a total of 96 completed questionnaires (Australia: $n = 15$, Germany: $n = 19$, Switzerland: $n = 21$, Turkey: $n = 20$, USA: $n = 21$). Appendix B1 provides a detailed description of expert recruitment procedures, participant demographics, and response rates.

Step 2: Preprocess corpus

We used standard text preprocessing steps to prepare the corpus for analysis, including cleaning, lemmatization, term frequency–inverse document frequency (tf–idf) calculation, and corpus feature selection (see Manning et al., 2008, Chapters 2.2, 6.2). Importantly, researchers should carefully consider each such step as preprocessing choices may influence the classification results (Denny & Spirling, 2018). Appendix C1 provides details to help readers preprocess their own text corpora.

Step 3: Perform topic modeling

Following text preprocessing, we trained the topic model using the LDA model implementation in Gensim, a freely available Python library for scalable statistical semantics (Řehůřek & Sojka, 2010). We used the default Gensim settings for prior parameters α and η, and optimized the topic model by tuning the parameter specifying the number of estimated topics (see Maier et al., 2018, p. 99). In order to do so, we generated topic models with 100, 500, and 1000 topics for each country with the goal of identifying the best-performing of the three models. In a next step we used three metrics to evaluate the topic model solutions: the human-rated coherence of estimated topics (Newman et al., 2011, pp. 501–502; 2010, 106) as well as measures to assess the precision (BLEU measure, Papineni et al., 2002) and recall (ROUGE measure, Lin, 2004) with which the estimated topic models covered the expert-provided relevant topics. BLEU and ROUGE are measures developed in the computational translation literature, which we repurposed to assess the degree to which a given topic model constitutes a comprehensive and accurate representation of expert knowledge in the substantive domain of interest for the classification task. We identified the "best" model solution using a researcher-driven reading of the three metrics (human-rated coherence, precision/BLEU, and recall/ROUGE) wherein the 500-topic solution provided the best tradeoff among these criteria. The results show that the main tradeoff was between the degree to which our model captured the entirety of expert inputs (i.e., recall/ROUGE) and the degree to which the output of the model (i.e., the number of topics to be validated as well as their intelligibility and interpretability) remained manageable. Appendix C2 provides further detail on the metrics and the procedures applied for topic model selection. It also provides all Python scripts used for topic modeling, which users can adjust to their own research situations. All following steps in the EITM process are based on the 500-topic model solution.

Step 4: Rank topics

The output of an LDA topic model is a set of *m* topics t_1 to t_m, each of which corresponds to a probability distribution of words within that topic. This means that a topic t_1 consists of a dictionary set of *n* words w_1 to w_n, each with certain probabilities $p(t_{1,}\ w_1)$ to $p(t_{1,}w_n)$ of occurring in that topic (see Figure 2).We divided the set of words in each topic into subsets of relevant (marked green in Figure 2) and non-relevant words, depending on whether they appear in the keyword list provided by the experts.[3] We then calculated the CMP of each topic by summing the topic-specific probabilities of all expert-provided relevant keywords per topic. The CMP of each topic can be interpreted as its overall probability of containing relevant keywords and as such is an estimate of the topic's thematic relevance. The researcher can then create a list of all topics estimated by the topic model, ranked by their CMP.

Step 5: Select relevant topics

In Step 5, human coders manually review the estimated topics and make a binary yes/no judgment about the thematic relevance of each topic. If the number of estimated topics is too large for a researcher to perform a review of all topics given available resources, the EITM process is flexible enough to allow the researcher to minimize the resulting loss of classification performance. In such situations, coders may perform this manual review step only on the upper end of the estimated relevance probability distribution, while the topics on the lower end of the estimated relevance distribution are discarded as irrelevant. In our use case, we selected the top 100 (or top 20%) of topics estimated by each topic model for further manual review.

Human coders were presented with the CMP score of each topic and the top-100 keywords with the greatest estimated probability of belonging to that topic. Keywords with a match in the list of expert-provided keywords were color-highlighted to guide the coders' review process. Color-highlighting keywords for manual topic review helps coders in two ways: First, it prevents them from overlooking relevant keywords and associated topics they already know to be relevant. Second, and more importantly, it indicates to coders keywords they may be either entirely unfamiliar with or which have an

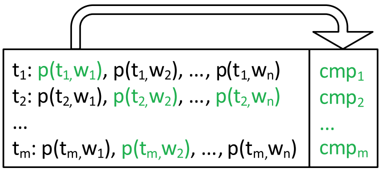

t_1: $p(t_{1,}w_1)$, $p(t_{1,}w_2)$, …, $p(t_{1,}w_n)$
t_2: $p(t_{2,}w_1)$, $p(t_{2,}w_2)$, …, $p(t_{2,}w_n)$
…
t_m: $p(t_{m,}w_1)$, $p(t_{m,}w_2)$, …, $p(t_{m,}w_n)$
cmp_1
cmp_2
…
cmp_m

Figure 2. Calculation of cumulative match probability (CMP) based on topic model.
Note. Each topic t consists of a set of words w with assigned probabilities p.

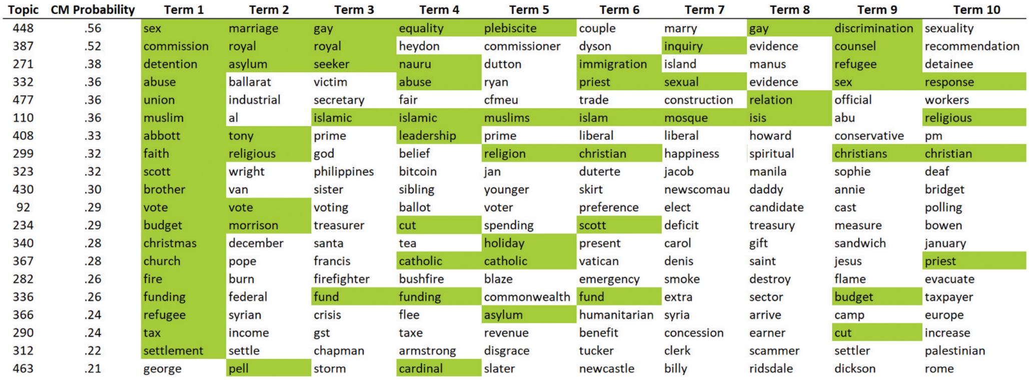
unknown association with a known relevant topic. It also signals to coders which keywords influenced the CMP, which may aid the interpretation of topics estimated to have a high probability of being relevant. Taken together, color highlighting as a feature of the EITM process works to improve the selection of relevant topics for subsequent discovery of the relevant document subset and supports coder and researcher learning about previously unknown relevant topics in the text data. Figure 3 shows a screenshot of a topic list as presented to the coders.

In our application, three coders were instructed to rate each topic as relevant or irrelevant following a coding protocol provided to them. Coders were also instructed to rate how certain they were about whether the given topic was thematically relevant for the purpose of the research project or not (i.e., whether it belonged to the master topic of the public role of religion in societal life). Each topic was then declared relevant or irrelevant following the majority decision of coders (see Appendix C3 for the coding protocol as well as all data and analysis scripts used in this step).

In sum, the EITM process improves the topic selection in two ways: First, by relating the topics estimated via topic modeling to expert-provided keywords, thus expanding the list of relevant expert-provided keywords with frequently co-occurring keywords. Second, by enabling human coders, based on the CMP ranking, to make more valid and efficient decisions in deselecting topics likely to be thematically irrelevant from the subsequent document discovery process.

Step 6: Score document relevance

Next, the topic model is used to estimate the probability of containing at least one thematically relevant topic for each document in the corpus (Figure 4). For a corpus consisting of o documents, each document d_x with $\{x \in \mathbb{N} \mid x \leq o\}$ can be represented by a distribution of topics t_1 to t_m with associated probabilities that the document contains these topics of $p(d_x, t_1)$ to $p(d_x, t_m)$. This document-specific sum of probabilities was estimated for the relevant topics. The resulting score can be considered as the individual relevance score for document d_{x}. The higher this score, the higher the probability of the respective document to contain at least one of the topics of interest. This enables the researcher to classify documents as relevant or irrelevant based on their corresponding relevance score (see Appendix D).



Topic	CM Probability	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Term 9	Term 10
448	.56	sex	marriage	gay	equality	plebiscite	couple	marry	gay	discrimination	sexuality
387	.52	commission	royal	royal	heydon	commissioner	dyson	inquiry	evidence	counsel	recommendation
271	.38	detention	asylum	seeker	nauru	dutton	immigration	island	manus	refugee	detainee
332	.36	abuse	ballarat	victim	abuse	ryan	priest	sexual	evidence	sex	response
477	.36	union	industrial	secretary	fair	cfmeu	trade	construction	relation	official	workers
110	.36	muslim	al	islamic	islamic	muslims	islam	mosque	isis	abu	religious
408	.33	abbott	tony	prime	leadership	prime	liberal	liberal	howard	conservative	pm
299	.32	faith	religious	god	belief	religion	christian	happiness	spiritual	christians	christian
323	.32	scott	wright	philippines	bitcoin	jan	duterte	jacob	manila	sophie	deaf
430	.30	brother	van	sister	sibling	younger	skirt	newscomau	daddy	annie	bridget
92	.29	vote	vote	voting	ballot	voter	preference	elect	candidate	cast	polling
234	.29	budget	morrison	treasurer	cut	spending	scott	deficit	treasury	measure	bowen
340	.28	christmas	december	santa	tea	holiday	present	carol	gift	sandwich	january
367	.28	church	pope	francis	catholic	catholic	vatican	denis	saint	jesus	priest
282	.26	fire	burn	firefighter	bushfire	blaze	emergency	smoke	destroy	flame	evacuate
336	.26	funding	federal	fund	funding	commonwealth	fund	extra	sector	budget	taxpayer
366	.24	refugee	syrian	crisis	flee	asylum	humanitarian	syria	arrive	camp	europe
290	.24	tax	income	gst	taxe	revenue	benefit	concession	earner	cut	increase
312	.22	settlement	settle	chapman	armstrong	disgrace	tucker	clerk	scammer	settler	palestinian
463	.21	george	pell	storm	cardinal	slater	newcastle	billy	ridsdale	dickson	rome

Figure 3. Results of Topic Ranking for Australia with 500 Topics (Screenshot)
Note. Screenshot as presented to human topic-relevance coders, presenting the top 20 topics and top 10 terms for each topic ranked by estimated topic relevance and term-topic membership probabilities. Green cells include terms that match with a keyword provided by at least one participant in the expert survey. CMP corresponds to the summed topic-level probability of all terms mentioned by at least one participant in the expert survey to belong to the respective topic.

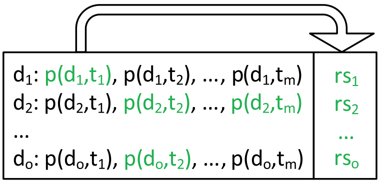

d1: p(d1,t1), p(d1,t2), …, p(d1,tm)
d2: p(d2,t1), p(d2,t2), …, p(d2,tm)
…
do: p(do,t1), p(do,t2), …, p(do,tm)
rs1
rs2
…
rso

Figure 4. Calculation of document relevance scores (rs) based on topic model.
Note. Each document *d* consists of a set of topics t_1 *to* t_m with a probability *p* to belong to this topic.

Step 7: Evaluate classifier

Validation is a crucial step in the development of any automated text-as-data technique (Grimmer & Stewart, 2013). It is therefore important in any application of the EITM approach to evaluate the quality of the classification solution it provides. The evaluation process we suggest consists of several steps. In a first validation step, we evaluate the estimated topic models in terms of their (a) coherence of estimated topics and (b) their coverage of topics and keywords provided by the domain experts. Appendix C2 provides a detailed description of this first step.

In a second validation step, we evaluate the classification performance of our EITM against a human-coded gold standard. The gold-standard ratings are then compared to the relevance scores predicted by the classifier. The assumption here is that such human validation of automated pre-classification will be effective even in the absence of coder domain knowledge. In an ideal research situation, researchers using the EITM process would draw on experts to validate their classification solutions. However, in our use case we employed student coders without expert domain knowledge as a “second-best” option for validation. This second-best approach is still useful given that the key problem addressed by EITM is that researchers cannot *pre*-define keywords for valid thematic document discovery. Our assumption when using non-expert coders in this step here is that it will be possible even for humans without prior domain knowledge to recognize a thematically relevant article when they see it.

We evaluated the performance of the EITM classifier using receiver operating characteristic (ROC) curves as well as three established metrics: precision, recall, and the F_1-score (see Japkowicz & Shah, 2014). This process assesses the degree to which the predicted relevance score works as an effective thematic classifier.

In our use case, human coders first coded the relevance of the entire collected text material from all political blogs in three of our countries: Australia, Switzerland, and Turkey. We then compared the results to the scores predicted by the EITM approach. Blogs are a “hard case” for testing classifier performance because blog posts will, on average, be more difficult to classify than more standardized categories of text such as news reports, which routinely rely on recurring professional formal and content features, such as the use of certain words and phrases, due to journalistic writing conventions. We therefore consider this evaluation to be a conservative test of classifier performance.

Gold-standard data to evaluate the EITM classifier was produced by 13 student coders who identified any documents from the full corpus of Australian, Swiss, and Turkish blog posts included in the MIND dataset that deal with the master topic of our use case, “the public role of religion in societal life” ($N_{\mathrm{AU}} = 9$, $N_{\mathrm{CH}} = 10$, $N_{\mathrm{TR}} = 10$ blogs and $N_{\mathrm{AU}} = 7{,}503$, $N_{\mathrm{CH}} = 1{,}635$, $N_{\mathrm{TR}} = 3{,}565$ blog posts, see Appendix A1 for an overview of collected data). Coders underwent intensive, multi-wave training in applying a detailed protocol for identifying documents as thematically relevant or irrelevant (see Appendix D1 for the selection protocol).

Several measures were taken to ensure a high quality of the human-coded benchmark data. All coders received the same training and instructions and were native speakers or fluent in at least two of the three source languages. Following coder training, we established pilot coding reliabilities before starting the main coding process. These reliability tests involved 11 of the 13 coders who accounted for

92% of all 13,750 coding acts performed as part of the main coding process (not included were C09, who accounted for 251, or 1.8%, and C28, who accounted for 850, or 6.2%, coding acts in the main coding process). In order to assess pilot coding reliability, coders rated 150 German and Swiss newspaper website articles and blog posts as either relevant or irrelevant. Coding reliability was assessed using percent agreement as well as two chance-corrected measures of interrater reliability, Brennan and Prediger (1981) and Krippendorff's alpha (Krippendorff, 2018). Overall, pilot reliability was excellent, with percent agreement = .92, kappa = .85, and alpha = .83 (all data and analytical code required to reproduce these analyses are available in Appendix D2).

In order to produce the human-coded gold-standard data, we implemented an online coding process that allowed coders to code all 12,703 Australian, Swiss, and Turkish blog posts included in the MIND dataset as thematically relevant or irrelevant. In order to eliminate any remaining coding error, the online coding process consisted of a multi-step validation procedure: any document for which the first coder indicated thematic relevance and/or uncertainty in their coding decision was coded by a second coder and any possible coding disagreement between the first and second coder was resolved by a tie-breaking third coder. These double-coding and majority decision rules served to minimize the proportion of false-positive and to maximize the proportion of true-positive classifications of blog posts as thematically relevant. The coding process also minimized bias due to any remaining idiosyncratic coder error post-training by implementing a set of rules aimed at avoiding a monopolization of coding for specific sources by any individual coder. Appendix D3 gives a detailed description of the online coding tool and process used, including a list of coder allocation rules.

Coders performed a total of 13,750 coding acts, including all initial coding (round 1), double-coding (round 2), and tie-breaker coding (round 3), as part of the main coding process to produce the human-coded gold standard data.[4] This procedure identified 184 thematically relevant blog posts for Australia (out of 7,503), 28 for Switzerland (out of 1,635), and 99 for Turkey (out of 3,565) (see Appendix D3).

Receiver operating characteristic

The first performance metric we evaluated and optimized was the area under curve (AUC) in the receiver operating characteristic (ROC) plot. The ROC curve indicates the performance of a binary classifier at different classification thresholds (see Japkowicz & Shah, 2014). A perfect classifier (i.e., one that yields 100% true positives and 0% false positives) is represented as a dot in the (0, 1) corner of the ROC plot. The diagonal corresponds to a bad classifier that is equivalent to randomly guessing results. The AUC is a single scalar value representing the expected ROC performance of a classifier. As the portion of the area under the ROC curve, its value will always be between 0 and 1.0 with an uninformative classifier indicated by a value of 0.5. The AUC is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006, p. 868).

The ROC curves and the AUC show that classification based on the 500-topic models worked well for the Swiss and Australian corpora, while performance dropped slightly but remained acceptable for the Turkish corpus (Figure 5). The diminished classification performance for the Turkish corpus could be a function of problems with the analyzed text data, preprocessing procedures, or different coding styles of the coders during Step 5.

The ROC curves also show that classification performance is sensitive to the selected number of topics. While in Australia the model estimating 100 topics had a slight performance edge, the 500-topic model was superior in the other two countries. In Switzerland in particular, the models estimating 100 and 1000 topics performed poorly. This supports the choice of topic models with the model hyperparameter set to estimate 500 topics for all countries and languages as these consistently exhibited good performance. It also shows the importance of making an informed choice about the topic model hyperparameter in any given research project.

Precision, recall, and F_1-score

The classification problem we address here involves an extremely unbalanced dataset with a very low prevalence of the positive class in the raw dataset. In such situations, the ROC and AUC can be misleading, which is why we also evaluated the precision, recall, average precision and F_1-scores for the positive class (Manning et al., 2008).

Precision is the fraction of items correctly classified as relevant (TP) among all items classified as relevant by the classifier (TP+FP): TP/(TP+FP). Recall is the fraction of all relevant items (TP + FN) that were correctly classified as relevant by the classifier (TP): TP/(TP+FN). The F_1-measure is the harmonic mean of a classifier's recall and precision. Its highest possible value is 1, indicating perfect precision and recall. The F_1-measure thus balances precision and recall and is a common measure of overall classification performance.

Table 2 (right-hand side) reports the classification results at the relevance score cutoff where the EITM classifier reaches the maximum F_1 score. At the cutoff, the F_1-score is .26 for the Australian, .34 for the Swiss and .25 for the Turkish corpus for the respective model with 500 topics. Precision and recall of the classifiers are similar across the three countries.

In order to benchmark EITM's classification performance, we emulated a simple keyword-based search approach of the sort that is often used but rarely explicated in applied communication research: the researcher creates a search query that is designed to maximize recall while avoiding the excessive inclusion of irrelevant articles that may occur due to the inclusion of theme-unspecific signifiers, for example, political party names like "CDU" in Germany. Recent examples of studies using this kind of approach in communication research include Eshbaugh-Soha (2010), Gentzkow and Shapiro (2010),

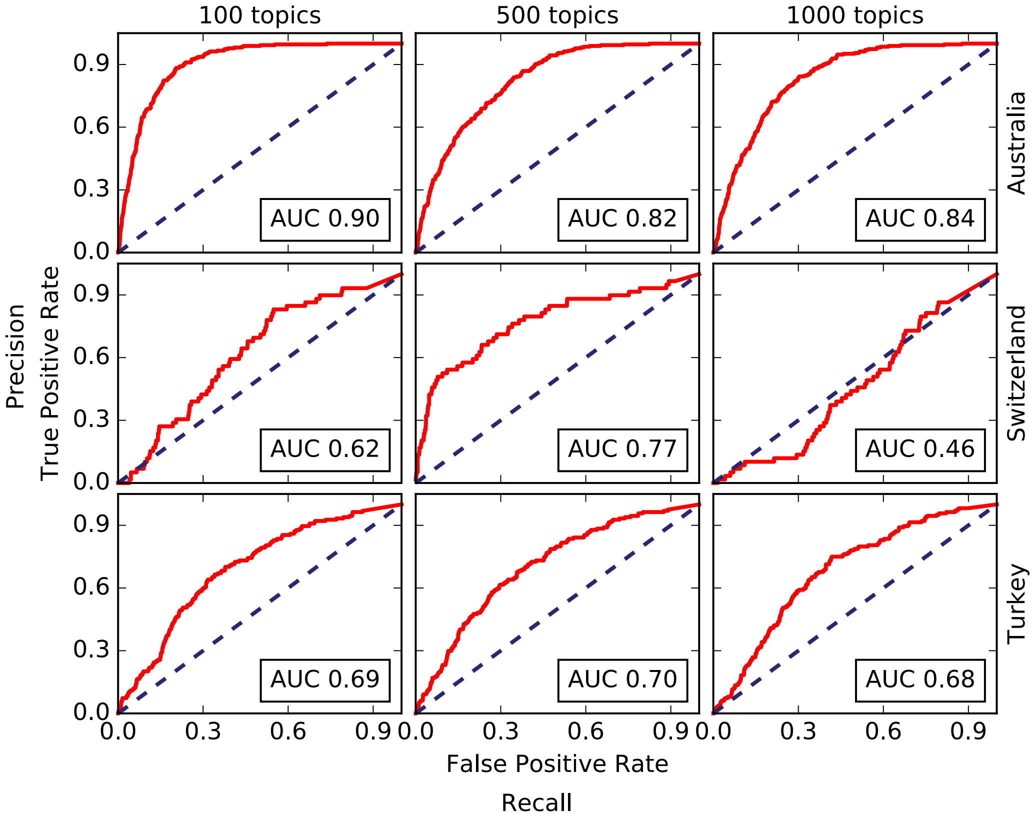


Figure 5. Classifier receiver operating characteristic (ROC) and area under the curve (AUC) for Australia, Switzerland, and Turkey at 100, 500, 1000 topics.

and Schäfer et al. (2014). In our benchmarking exercise, we used the thematic keywords collected in the expert survey to build the search query. Table 2 (left-hand side) shows the results of this simple yet common approach to document set discovery compared against the manual gold standard coding for Australian, Swiss, and Turkish blogs.

When compared to EITM (right-hand side), the keyword-based search selects a larger number of documents and therefore has a much higher recall. However, this comes at the price of vastly reduced precision, as most of selected documents are not actually relevant. Compared to the keyword-based search-query, the EITM-based classification results in an increase of the F_1-score by 3.5 to 4.8 times. But rather than just looking at this improved classification performance, it is important to note that more generally EITM bestows researchers with greater flexibility in fine-tuning the classification to their specific research needs. A keyword-based search query yields exactly one result and can only be adjusted by including or excluding specific keywords or changing Boolean operators. With EITM, the researcher can tune the classification result in a more fine-grained fashion by setting custom cutoff values for the relevance scores assigned to documents in the corpus.

Figure 6 displays the precision-recall curves for document classification based on topic models with 100, 500 and 1000 topics for the Australian, Swiss, and Turkish blog posts in the MIND dataset. It also includes the F_1 and average precision (AP) scores for each classifier. The AP is the mean of the precision over all possible relevance cutoff values weighted by class distribution. It is the average probability, across the full range of possible cutoffs, that a document picked up based on the EITM (pre-)classifier actually is relevant.

The figure again shows that, in this application, the topic models estimating 500 topics tended to perform best, with the exception of Australia, where the model estimating 100 topics scored highest. It also illustrates the trade-off between precision and recall for the classifiers. As the recall of the classifier is increased by lowering the relevance score cutoff point, the precision of the classification decreases. That is, if a researcher requires high recall, they can keep drawing documents with a lower relevance score at the expense of an increasing number of irrelevant documents and more time needed for sifting through irrelevant documents.

Table 3 juxtaposes the average precision (AP) of the selected models estimating 500 topics with the prevalence of the positive class (i.e., relevant documents) in the dataset. The prevalence is the fraction of relevant topics in the document corpus and the probability of picking up a thematically relevant document when making a random draw from the corpus.

In our unbalanced dataset, the AP scores indicate a significant improvement when selecting relevant documents. Given that AP is the average probability, over all possible cutoffs, that a document retrieved as relevant by the EITM (pre-)classifier is actually relevant, we find for the Australian corpus, for which the AP is .16, that the EITM-based pre-classification increases the probability of picking up a relevant document from the pre-classified corpus by a factor of 6.5 compared to random draws from the full Australian corpus (i.e., the prevalence). The probability of retrieving a relevant document based on EITM was 12.3 times higher for the Swiss corpus and 5.4 times higher for the Turkish corpus.

We can use the efficiency gains observed for documents collected from blogs to estimate the number of working hours saved for a task commonly encountered by applied researchers: sampling a fixed number of thematically relevant documents from a corpus containing relevant and irrelevant

Table 2. Classification performance of EITM versus a baseline keyword-based search approach.

	Baseline keyword-query search			EITM		
Country	Precision	Recall	F_1	Precision	Recall	F_1
Australia	0.04	0.99	0.07	0.21	0.35	0.26
Switzerland	0.04	0.96	0.07	0.28	0.42	0.34
Turkey	0.04	0.92	0.07	0.18	0.43	0.25

For EITM, the precision, recall, and F_1-score are reported for the maximum obtained F_1-score for the models with 500 topics in Australia, Switzerland, and Turkey for all blog items in the MIND dataset.

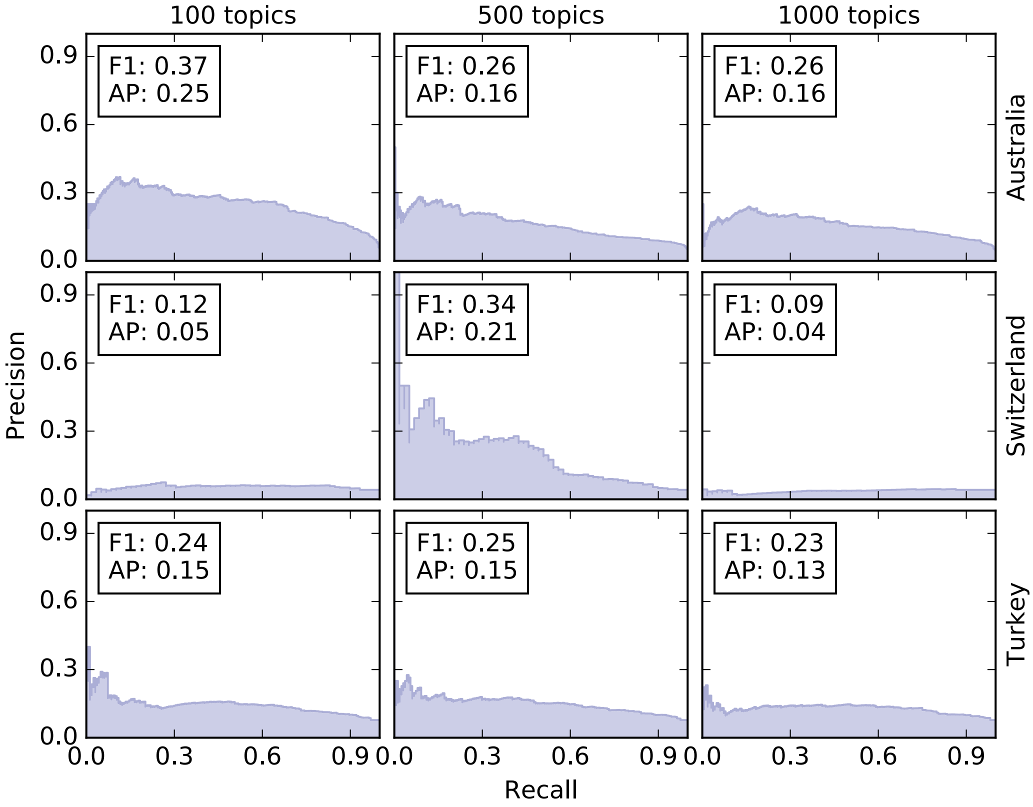


Figure 6. Precision / recall curve for the positive class for Australia, Switzerland, and Turkey at 100, 500, 1000 topics.
Note. F1 is the maximum obtained F$_1$-score and AP is the average precision for the classifier.

Table 3. Observed speed-up (for Blogs) and estimated efficiency gains (for websites & newspapers) for document discovery based on EITM (compared to a fully manual approach).

	Blogs						Websites & Printed Newspapers		
	Observed						Estimated		
Country	Positive Class	Total Items	Prevalence	AP	Speed-up	Coding Hours	Coding Hours	Hours Saved	Relative
Australia	184	7503	.02	.16	6.5	30.5	11.7	64.6	85%
Switzerland	28	1635	.02	.21	12.3	10.3	13.3	149.8	92%
Turkey	99	3565	.03	.15	5.4	37.0	16.8	73.9	81%
Total							41.8	288.3	87%

Positive class is the number of identified relevant news items; total items is the number of all items in this category; prevalence is the calculated proportion of positive items; AP is the average precision of the EITM classifier based on estimated relevance scores and human-coding decisions; speed-up is the estimated factor by which EITM accelerates the manual identification of relevant documents; coding hours are the human-coding times observed in the online coding tool in hours; hours saved and relative reduction of coding time for websites and printed newspapers are estimates based on the speed-up observed for the classification of blog posts.

documents. In our use case, we test the efficiency gains achieved through EITM for the example task of generating from the full MIND dataset a sample of 300 thematically relevant documents for each of the three countries (Australia, Switzerland, and Turkey), with a target sample size of 100 per document category (blog posts, online news articles, printed news articles) per country (or fewer if the MIND dataset contained less than 100 relevant documents for a document category in a given country). Based on the theoretical "speed-up" as observed by measuring the average precision for the manual identification of thematically relevant blog posts, we estimate that in this application the EITM classification reduces the work required to reach these target sample sizes from the other two document types in the MIND dataset (i.e., online and printed newspaper articles) by 87% (see Table 3).[5]

It is important to consider that we tested the EITM-based classifiers on blog posts. Looking at the average time coders needed to decide about the relevance of documents, we find that blog posts appear to be more difficult to code than the other document categories: in the example task described above, for example, coders spent 30.5 of the total 42.2 coding hours spent on the three Australian document corpora coding blog posts only. Therefore, we may assume that performance estimates based on blog data are conservative and likely to improve for applications to more standardized genres of text like news items from printed newspapers and news websites, because these are generally more structured and use a more conventional writing style, which makes them easier to classify. Considering the speed-up estimates above, this translates into 64.6 hours of manual coding time that a researcher would save when aiming to sample the target number of 300 thematically relevant documents from the Australian corpus. In the same task, coders spent a total of 23.6 hours coding the Swiss corpora to identify the target number of relevant documents within each category of documents (blog posts, online news articles, and printed newspaper articles). They spent 44% (10.3 hours) of their overall working time coding blog posts, which translates into 149.8 hours of coding time saved; for Turkey, it took coders a total of 53.8 coding hours to identify the fixed target number of relevant articles for each document category. Coders spent 69% (37 hours) of their overall working time coding blogs, which translates into 73.9 hours of saved manual coding time needed to meet the target size of the relevant document sample. With the target sample sizes applied to the three MIND document categories, we estimate that in this realistic example sampling task applying the EITM approach will have saved at least 288.4 hours of manual coding time in these three countries alone. Of course, one feature of the EITM process is that it is highly scalable, and the larger the dataset to which it is applied, the greater the efficiency gains will be.

It is important to note that in our validation study it would not have been necessary to code all items for blogs in Australia to achieve the goal of drawing a sample of 100 items. For Swiss and Turkish blogs, all items needed to be coded, given that the total number of relevant items in the corpus was below the target sample size. Since all items falling into the blog category had to be coded to create the gold standard data, the potential time saved for Australian blogs was not included in the estimated total hours saved.

However, the performance data indicated that the classification problem in this case study and dataset is very hard. The AUC and average precision scores suggest that relevance score estimation was far from perfect. This, in turn, raises the question of whether selecting items based on the estimated relevance scores may introduce a selection bias into the sampling process.

Bias can be introduced by every imperfect classification process and especially by keyword-based selection processes. Within the EITM framework, the intermediate step of selecting relevant topics (Step 5) is key to reducing selection bias. One option that it offers researchers is to bring in the surveyed experts again, rather than the researchers themselves, to select and validate the topics. Another option EITM offers is to optimize the ROUGE score, which is the recall of the expert keywords by the topic model. Selecting a topic model with a high ROUGE score ensures broad coverage of expert-named debates and can be expected to reduce bias in the resulting classification.

For example, if the goal is to sample news items related to discussions of public health in 2020, a keyword search might overwhelmingly return news items related to COVID-19. If the domain experts also listed a debate about alcohol abuse that is not covered by the topic model the ROUGE score will drop. In such a situation, it is possible to adjust the topic model to include this debate, for example, by varying hyperparameters such as the number of topics. Another option that would be open to researchers using EITM is to decrease the weight of the topics related to COVID-19, to actively influence the relevance scores, so that it becomes easier to discover news items debating other public health issues.

Finally, EITM users can make use of chunking to further reduce classification bias: In the case study presented here, we decided not to draw single documents based on the relevance score until we reached the sample size. Instead, we drew larger chunks of 10,000 documents from the fully ranked document set, working our way down from the top of the relevance distribution, and randomized the order of documents within chunks. This made it possible to discover relevant news items with a lower relevance score while still eliminating those news items that were extremely unlikely to be related to the master topic. The EITM process allows researchers to reduce bias and maximize recall by adjusting either the chunk size or the probability cutoff values according to the amount of time that they are able to invest into the item-selection process.

In sum, it is important to note that it is important for researchers using EITM to be aware of the risks of selection bias as much as with any other classification technique. The above examples show that several options exist for them to address and hopefully reduce such bias within the EITM process.

Concluding remarks

We propose an expert-informed hybrid approach to discovering document subsets belonging to a researcher-defined master topic within large sets of text data. The expert-informed topic modeling (EITM) approach combines the efficiency and scalability of unsupervised text analysis with the deep knowledge of experts within the relevant social and substantive domain. It is a hybrid procedure in that it involves the discovery of relevant topics as well as the classification of documents into these topics. This means that the procedure is doubly open. First, it opens up the process of document set discovery to inductive algorithm-driven discovery. Second, it allows for post-hoc input of expert knowledge to validate and weight algorithmic results and guide document set classification decisions. At its heart, the EITM approach is a post-hoc way of adding problem-specific structure to unsupervised computational model output by drawing on expert knowledge. It thus can be thought of as a "semi-supervised" approach to document set discovery.

EITM is the first validated method for (semi-)automated document discovery that does not require detailed contextual knowledge on the part of the primary researchers. The approach is flexible in that it can be applied to any given subject domain. Unlike deterministic keyword-based strategies, EITM provides a probabilistic method for identifying relevant sub-topics that researchers were not previously aware of (e.g., specific public debates in specific social contexts at a specific time) and text documents that deal with these topics. The EITM approach therefore is particularly helpful in situations in which researchers lack comprehensive domain knowledge such as cross-nationally comparative studies or studies of highly specialized expert or subcultural discourses. It is also flexible in that it makes a valid topic-based identification of document subsets from unfamiliar knowledge domains more efficient in research projects, irrespective of their degree of tolerance for error in document classification.

One possible alternative to surveying experts for the generation of debate themes and keywords may be to start with a set of keywords defined by the researcher and then refine this set using computer-aided methods such as that proposed by King et al. (2017). This approach may work well in cases in which the master topic of interest is signified by a set of keywords that has a high likelihood of occurring in each of the initially unknown sub-debates of interest. It may work less well when the master topic of interest is very broad and encompasses debates that may not share a "common denominator" of keywords. For example, the master topic in our use case (the public role of religion in societal life) encompassed some clearly relevant debates (e.g., the "Handschlag-Debatte" in Switzerland, and the debates on same sex marriage and on adoption rights for same-sex couples in the USA), and some articles on these debates did not include generic references to the master topic. In such a constellation an expert-driven approach may provide better coverage of all relevant debates compared to the researcher-driven, computer-assisted approach by King and colleagues. In short, emic domain knowledge simply may be superior to computational and researcher inputs in some situations that are commonly encountered by communication researchers, especially in cross-national studies of text-as-data.

More recently, word embeddings and neural networks have been proposed for text classification (Rudkowsky et al., 2018; Turney & Pantel, 2010). Even though there has been recent work on the interpretability of word embeddings (Jang & Myaeng, 2017; Şenel et al., 2018), one benefit of using LDA for the initial clustering of the search space is that the results returned by LDA are more easily interpreted by humans. This allows a manual selection of topics for subsequent classification. LDA outputs thus can serve as inductive starting points to identifying relevant and irrelevant topics and documents without requiring training data for supervised machine learning.

Moreover, its LDA-based approach makes EITM transparent from start to end. The LDA clustering results help researchers gain insights into the specific full document set at hand. We regard this as an advantage over modern pre-trained models like BERT, which get trained on massive general text corpora (Devlin et al., 2019; Kaliyar, 2020). However, the EITM framework also, in principle, allows for the use of different clustering algorithms that may have distinct advantages in different research situations.

Finally, once a researcher has selected a set of documents using EITM, they may use this set together with the topic distribution generated by LDA to produce training data for machine learning and iteratively increase the classification performance even further.

In projects where researchers aim for a *reduction* of error in automated text subset discovery, rather than its *elimination*, an EITM-approach alone can achieve this goal. However, as shown above, in projects where the researchers aim at achieving a maximally error-free topic-based identification of document subsets, pre-classification of documents based on EITM-generated relevance scores improves the efficiency of the process by reducing the average time needed for manual coders in a follow-up classification task to identify thematically relevant items within the larger search set of documents. The topics identified as relevant based on EITM are used to estimate for each document in the search set the continuous probability of it belonging to the master topic. Researchers can then define a probability threshold that must be met for any given document to be classified as relevant for further analysis of any form. As we have shown, the EITM approach provides a substantively valid, efficient, and flexible framework for document-set discovery under imperfect domain knowledge of researchers. Here, we have validated it with corpora in three different languages (English, German, and Turkish).

In future applications, the EITM process can easily be expanded to other languages and is open to further extensions based on researcher needs and resources. For example, future applications may use a two-stage expert survey design, in which experts not only provide open lists of topics and keywords but are surveyed a second time to validate the final list of relevant topics to improve the resulting estimates of document relevance probabilities. It is also possible to apply EITM not only to estimate a single topic model per country as we did here but to make more fine-grained distinctions, for example, by estimating one model per media type in each country, in order to allow for an even more context-sensitive detection of topics and debates.

The proliferation of large, heterogeneous sets of media content and other politically relevant text data will make the specific classification task that EITM solves more common. First, primary researchers will not be able to know the discourses, media contexts or countries they aim to study deeply enough in advance to be able to make valid document set choices upfront. This is a pressing issue in comparative studies that reach beyond a small set of countries well known to the primary researchers. It is also important in studies of specialized subcultures and professional milieus for which few prior insights exist. In this sense, EITM opens the door to uncharted territory. Second, EITM's continuous relevance estimates allow for researcher-defined cutoffs and enhanced efficiency in selecting documents and thus increase researcher flexibility. Finally, continuous relevance scores can also be used as inputs for more refined substantive analyses, for example, by comparing item-level content features in various brackets of topical relevance. In sum, EITM provides a new hybrid framework with multiple possible extensions to enable further analysis in text-as-data studies. Importantly, it not only provides a way for social scientists to improve the validity of their document discovery procedures, but also supports their learning about social contexts they may want to study but have incomplete prior knowledge about.

Notes

1. There also are more minor differences concomitant to the more basic differences between the approaches. For example, because it focuses on inductive learning about initially unknown debates in situations where the researchers are lacking necessary domain knowledge, EITM draws on probabilistic topic models for the unsupervised estimation of topics and mixed-membership classification of documents into these topics. In contrast, the KLR approach generalizes after a researcher-selected "seed" search query to identify a reference document set for training purposes and thus relieson supervised machine learning classifiers, which may introduce biasing path dependencies into the process.
2. Experts were given the following instructions: "Please give a brief description of each topic that comes to your mind, using the designated fields. In a second step, please enter for each topic 3–5 keywords that characterize the topic and that are, in your best judgment, helpful in identifying news stories covering this topic."
3. Using only expert-provided keywords is a conservative approach. Another possible approach would be to expand the list of inputted keywords to synonyms. However, such expansion may introduce an additional need to validate the resultant synonyms, which could be done by feeding potential synonyms to experts in a second survey round. It is also important to note that an exhaustive list of keywords will be more important for deterministic approaches based on search-queries. One benefit of using topic modeling for document discovery is that, as a probabilistic approach, it will capture each term in its topic solution that is systematically associated with the expert inputs, which means that we can expect EITM to function well even if the expert inputs are lexically incomplete so long as they are thematically exhaustive. Similarly, the probabilistic process ensures that topics that are well described with only a few keywords are not ranked lower than those with many relevant keywords as such highly discriminant keywords will be assigned greater weight for their respective topic.
4. Australian blogs: 7,503 (round 1) + 274 (round 2) + 141 (round 3) = 7.918 coding acts; Swiss blogs: 1,635 (round 1) + 60 (round 2) + 32 (round 3) = 1,727 coding acts; Turkish blogs: 3,565 (round 1) + 318 (round 2) + 222 (round 3) = 4,105 coding acts.
5. Note that both the speed-up and efficiency gain analysis assess the acceleration of the document discovery process, not a comprehensive cost-benefit analysis, which would include the effort involved in the other steps of the EITM process such as conducting the expert survey.

Disclosure statement

We have no known conflict of interest to disclose.

Funding

This research was funded by the German Research Foundation (DFG) Grant WE 2888/6-1.

ORCID

Eike Mark Rinke http://orcid.org/0000-0002-5330-7634
Timo Dobbrick http://orcid.org/0000-0002-6252-1157
Charlotte Löb http://orcid.org/0000-0002-5874-6986
Cäcilia Zirn http://orcid.org/0000-0002-7183-5684
Hartmut Wessler http://orcid.org/0000-0003-4216-5471

References

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19–42. https://doi.org/10.1017/pan.2020.8

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/v3/blei03a.html

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699. https://doi.org/10.1177/001316448104100307

Brier, A., & Hopp, B. (2010). Computer assisted text analysis in the social sciences. *Quality & Quantity*, *45*(1), 103–128. https://doi.org/10.1007/s11135-010-9350-8

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423http://arxiv.org/abs/1810.04805

Eshbaugh-Soha, M. (2010). The tone of local presidential news coverage. *Political Communication*, *27*(2), 121–140. https://doi.org/10.1080/10584600903502623

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, *78* (1), 35–71. https://doi.org/10.3982/ECTA7195

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

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359. https://doi.org/10.1177/1077699016639231

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271

Jang, K.-R., & Myaeng, S.-H. (2017). Elucidating conceptual properties from word embeddings. *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and Their Applications*, Valencia, Spain, 91–95. https://doi.org/10.18653/v1/W17-1911

Japkowicz, N., & Shah, M. (2014). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.

Kaliyar, R. K. (2020). A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of BERT. *10th International Conference on Cloud Computing, Data Science Engineering (Confluence), Noida, India*, 336–340. https://doi.org/10.1109/Confluence47617.2020.9058044

King, G., Lam, P., & Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, *61*(4), 971–988. https://doi.org/10.1111/ajps.12291

Kopecký, P., Sahling, J.-H. M., Panizza, F., Scherlis, G., Schuster, C., & Spirova, M. (2016). Party patronage in contemporary democracies: Results from an expert survey in 22 countries from five regions. *European Journal of Political Research*, *55*(2), 416–431. https://doi.org/10.1111/1475-6765.12135

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). Sage.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, *21*(4), 475–492. https://doi.org/10.1037/met0000081

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In M.-F. Moens & S. Szpakowicz (Eds.), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain*, (pp. 74–81). Association for Computational Linguistics. https://www.aclweb.org/anthology/W04-1013

Mahdavi, P. (2019). Scraping public co-occurrences for statistical network analysis of political elites. *Political Science Research and Methods*, *7*(2), 385–392. https://doi.org/10.1017/psrm.2017.28

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173

Newman, D., Bonilla, E. V., & Buntine, W. (2011). Improving topic coherence with regularized topic models. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24* (pp. 496–504). Curran Associates, Inc. http://papers.nips.cc/paper/4291-improving-topic-coherence-with-regularized-topic-models.pdf

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA*, 100–108. http://dl.acm.org/citation.cfm?id=1857999.1858011

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA*, 311–318. https://doi.org/10.3115/1073083.1073135

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta*, 45–50.

Rinke, E. M., Löb, C., Dobbrick, T., & Wessler, H. (2019). *Mannheim International News Discourse Data Set (MIND)* [Data set]. https://doi.org/10.7801/305

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. https://doi.org/10.1080/19312458.2018.1455817

Schäfer, M. S., Ivanova, A., & Schmidt, A. (2014). What drives media attention for climate change? Explaining issue attention in Australian, German and Indian print media from 1996 to 2010. *International Communication Gazette*, *76* (2), 152–176. https://doi.org/10.1177/1748048513504169

Şenel, L. K., Utlu, İ., Yücesoy, V., Koç, A., & Çukur, T. (2018). Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1769–1779. https://doi.org/10.1109/TASLP.2018.2837384

Trilling, D., & Jonkman, J. G. F. (2018). Scaling up content analysis. *Communication Methods and Measures*, *12*(2–3), 158–174. https://doi.org/10.1080/19312458.2018.1447655

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Whitaker, E. M. (2017). Emic and etic analysis. In B. S. Turner (Ed.), *The Wiley-Blackwell encyclopedia of social theory* (pp. 1–2). Wiley-Blackwell. https://doi.org/10.1002/9781118430873.est0640