Contents lists available at ScienceDirect

Poetics

journal homepage: www.elsevier.com/locate/poetic

Separating the wheat from the chaff: A topic and keyword-based procedure for identifying research-relevant text*[☆]

Alicia Eads[a,*], Alexandra Schofield[b], Fauna Mahootian[c,1], David Mimno[d], Rens Wilderom[e]

[a] University of Toronto, Department of Sociology and the Centre for Industrial Relations and Human Resources
[b] Harvey Mudd College, Computer Science
[c] Cornell University, Computer science
[d] Cornell University, Computer Science
[e] University of Amsterdam, Department of Sociology
[1] Currently employed at Datavant

ARTICLE INFO

Keywords:
Topic model
Latent dirichlet allocation
Filtering
Text analysis

ABSTRACT

Social scientists are using computational tools to expand their content research beyond what is humanly readable. This often requires filtering corpora for complex research concepts. The commonly used off-the-shelf filtering techniques are untested at this task. Dictionaries may not recognize language outside of investigators' expectations and thresholding on topic proportions from topic models may fail to identify brief references to concepts. We develop a typology of texts as they relate to a research concept and use this to structure a filtering procedure. We compare our procedure's performance with dictionary-only and topic-proportion-only approaches on two corpora—government speeches and academic articles—and two research concepts—housing crisis and inequality. Our procedure outperforms overall and on each type of relevant text in the typology. An open-source software package is available for implementing the procedure. This provides researchers with a more structured and tested approach for filtering text data. Additionally, the types-of-text typology analysis provides a unique examination of what constitutes a filtered dataset, allowing researchers to consider how conclusions may be affected.

1. Introduction

Social scientists are increasingly taking advantage of the vast amount of searchable text data and using computational tools to expand their research beyond the scale that is humanly readable (Bail, 2012; Golder & Macy, 2011; Grimmer & Stewart, 2013; Marshall, 2013; Mohr, Wagner-Pacifici, Breiger & Bogdanov, 2013). While computational tools make big data collection possible, approaches using computational tools for identifying the "complex, socially constructed, and unsettled theoretical concepts" that social scientists are often interested in are still largely underdeveloped and untested (Nelson, Burk, Knudsen & McCall, 2018: 1).

Currently, social scientists adopt one of two solutions to the problem of filtering their corpora for complex research concepts. One

[☆] The authors would like to thank Anna Keuchenius, Alexandra Ils, and attendees at the 2019 International Conference on Computational Social Science (IC2S2) for helpful comments on earlier versions of this work.

[*] Corresponding author.
E-mail address: alicia.eads@utoronto.ca (A. Eads).

https://doi.org/10.1016/j.poetic.2020.101527
Received 1 October 2019; Received in revised form 30 November 2020; Accepted 2 December 2020

Please cite this article as: Alicia Eads, *Poetics*, https://doi.org/10.1016/j.poetic.2020.101527

solution is traditional human-reading and hand-coding (Benoit, Laver & Mikhaylov, 2009; Grimmer & Stewart, 2013; Martin, Rafail & McCarthy, 2017). Even social scientists who are otherwise using computational text analysis methods often read their texts to structure them according to specific research concepts (e.g. Slapin & Proksch, 2008). The downside to this approach is that the amount of text that can be analyzed has to stay within the limits of what humans can read in a reasonable timeframe.

The second solution is applying an "off-the-shelf", automated filtering procedure (Nelson et al., 2018). The two most common automated approaches are: 1) using a list of keywords—a "dictionary"—to filter texts (Grimmer & Stewart, 2013); or 2) using the inferred "topics" from a topic model (Blei, Ng & Jordan, 2003; Nelson, 2017). In the dictionary approach, researchers use lists of words that reflect their research concept and identify the texts in their corpus that contain those words. However, dictionary filters have difficulty identifying documents with more nuanced discussions of a concept (Nelson et al., 2018). In addition, they rely on an *a priori* set of distinguishing words, which researchers are not very good at developing and which may not reflect the language used by the producers of the text (King, Lam & Margaret, 2017). The inferred topics approach uses machine learning to automatically find groups of words that co-occur together, or "topics." Researchers then select apparently relevant topics and identify texts that have high proportions of those topics. This is an inductive approach that is not limited to *a priori* sets of keywords, but topics are unlikely to map exactly to the concepts that researchers are interested in (Chuang, Gupta, Manning and Heer (2013)). The process of selecting which topics appear relevant alone may lead to the misclassification of texts if relevant topics are overlooked. Another issue with this approach is finding an appropriate threshold of topic proportion in a text to classify it as relevant. If the chosen topic proportion threshold is too high, relevant texts will be left out but if they are too low, non-relevant texts will be included (Nelson et al., 2018). Both of these automated approaches are useful, but imperfect, as they may result in an incomplete and/or polluted subset of text. Further, because these techniques as filtering approaches have largely been untested, we know little about how incomplete or polluted filtered datasets may be (but see Nelson et al., 2018).[1]

In this paper, we structure and test a process of using computational tools to filter text corpora for complex research concepts. Our process incorporates the benefits of the off-the-shelf approaches discussed above and adjusts for their deficiencies with two additional straightforward steps. Furthermore, we begin by better defining the problem of filtering for complex concepts. That is, we develop an ideal-typology of individual texts in a corpus as they relate to a specified research concept. This clarification of the problem is helpful in structuring a process using computational tools to target specific types of texts. In addition to incorporating the advantages of automatic techniques, we also incorporate the advantages of human skill in recognizing representations of a research concept (King, Lam, and Roberts, 2017).

In what follows, we present the ideal-typology of texts and outline our filtering process. Then, we evaluate how well our approach performs by comparing it to the dictionary-only and topic model-only approaches on two different text corpora and two different research concepts. We discuss each approaches' performance with reference to the different types of texts in the typology, providing a more systematic examination of what will and will not constitute a filtered dataset. We show that our process works well for identifying two types of texts. Our process does not perform as well at identifying two other types of texts, but it performs better than the off-the-shelf approaches. This analysis allows researchers to consider how their conclusions may be affected by what types of texts will constitute their data. We provide an open-source Python software package, wheat_filtration, for implementing our filtering process. It is freely available on GitHub (https://github.com/faunam/wheat_filtration).

2. Ideal typology

We inductively develop a typology of texts in a corpus as they relate to a research concept. This typology provides a useful structure for designing a process for semi-automatically identifying relevant texts. We describe six different ways a text can relate to a research concept. Four types of texts are relevant to the concept: 1) texts that are mostly about the research concept; 2) texts that are partially about the concept but are mostly about other concepts; 3) texts that just mention the concept; and 4) texts that are to some extent about the concept but do not use key words. Types 1 through 3 reflect quantitative differences in how relevant texts can vary. Type 4 is qualitatively different in that the research concept is expressed only implicitly. For example, if the research concept is "housing markets", a text could explicitly refer to "the housing market" or more implicitly to "residential starts and sales". With respect to identifying relevant texts, there are degrees of implicitness. Some implicit references would just require that a greater number of words be considered together, i.e. using any one keyword would miss an implicit reference. A greater degree of implicitness requires context, including surrounding text not contained in the focal document or knowledge about the corpus. A researcher armed with context knowledge may recognize such an implicit reference but an automated approach does not have context knowledge, making type 4 texts the most difficult for an automated filtering approach. There are two additional types of texts that are not relevant to the research concept: 5) texts that are entirely about non-relevant concepts; and 6) texts that are about related concepts and thus may contain related key terms.[2] Table 1 summarizes this typology. Appendix Table A.1 provides example texts from one of our corpora for each type of text.

This typology specifies the problems that a process using computational tools to identify relevant texts must address. It also suggests

[1] Nelson et al. 2018 is the only other explicit test of off-the-shelf automated filter approaches that we are aware of. Their "Coding scheme B" is nearest to the distinctions we make between relevant and irrelevant texts. The supervised machine learning approach performed the best with a precision score of .73 and a recall score of .60, yielding a dataset that is 27% polluted and 40% incomplete.

[2] This typology generalizes and expands the distinctions in relevance for inequality that Nelson et al. 2018 identified. For example, their distinction between explicit inequality and implicit inequality corresponds with our types 1, 2 and 3 versus type 4 texts.

Table. 1
Typology of Texts as Related to a Research Concept

Relevant Texts				Non-Relevant Texts	
Type 1 Mostly about concept	Type 2 Partially about concept	Type 3 Just mention concept	Type 4 Implicit references	Type 5 About non-related concepts	Type 6 About related concepts

that the dictionary-only or topic-proportion-only approaches that many researchers currently rely on are likely to fail in specific ways. Type 1 texts, those that are mostly about the concept, are easy for the topic-proportion approaches to recognize, as long as the topic model solution includes a topic or topics that capture the concept and as long as the researcher recognizes them as such. Dictionaries may also work, as long as the language used to refer to the concept is what a researcher anticipates. Recognizing type 2 and type 3 texts, that are only partially about a concept, is more difficult. A dictionary approach may recognize these texts but with the same caveat that the language is what a researcher anticipates. A topic-proportion-only approach, however, may fail to recognize many of these types of texts since the words of relevant topics would occur in small proportions. Texts that mention or are even entirely about the concept but do not use keywords, as with type 4 texts, will likely be missed by a keyword approach but may be picked up by an inductive topic-proportion approach, if they contain enough topic-relevant words.

Excluding texts that are entirely about non-relevant concepts is fairly easy for both off-the-shelf automatic approaches. However, excluding texts that are about related concepts is more difficult. Keyword approaches may erroneously include some of these texts that contain keywords used in a different sense: as an extreme example, someone interested in financial institutions may be disappointed to find documents included that discuss river "banks" and duck "bills." A topic-proportion-only approach could be better at excluding these texts, provided that the topic model solution distinguishes between researchers' concepts and the related concepts.

Our semiautomatic process is designed to incorporate the benefits of both of these off-the-shelf approaches while addressing their deficiencies. We use a combination of three methods—topic proportions, a topic-term derived keyword list, and a "super" keyword list. Each method in the process is based on an inferred Latent Dirichlet Allocation (LDA) topic model for the corpus of interest (Blei et al., 2003). Thus, instead of relying on *a priori* keywords, both the topics and the keywords are inductively developed.

The process is not fully automatic. Three instances in the process involve researcher intervention to identify which topics or terms are relevant to the research concept. Since recognizing is a task that humans are particularly good at performing (Hu and Boyd-Graber, 2013; King, Lam, and Roberts, 2017), we thus also incorporate the benefits of human skill.

3. Filtering relevant texts

3.1. Preprocessing the corpus

To apply our approach, which is based on LDA, we make several preprocessing decisions aimed at improving the effectiveness of LDA for modeling.[3] In determining the size of sections of text, we note that LDA typically works well on text passages that are reasonably short and similar in length (Boyd-Graber, Mimno & Newman, 2014). Some text corpora automatically produce text well suited to that constraint, but, in many cases, natural text corpora produce documents much longer. Although paragraphs are a good natural language delimiter (Algee-Hewitt, Heuser & Moretti, 2015), our corpora did not have consistent divisions between paragraphs. We instead used a strategy of splitting the text into 6-sentence fragments using the Punkt sentence tokenizer (Kiss & Strunk, 2006) as implemented in NLTK (Loper & Bird, 2002). This provides roughly paragraph-length segments of close to the same size without splitting sentences.

Another important preprocessing task is determining what constitutes a single term. We considered whether to combine words from the same root, such as "mortgage" and "mortgages", by using an automatic tool, such as a stemmer or lemmatizer, to rewrite words with irregular endings. While stemming and lemmatizing are common approaches in smaller datasets where too little data risks vocabulary not overlapping across documents, with larger corpora, topic models usually do the work of conflating these terms effectively without changing the terms' endings (Schofield & Mimno, 2016). Because this morphological information may also convey information we care about, we choose not to stem or lemmatize.

Finally, because many of the specialized terms in our dataset are in fact multiword phrases, we artificially joined together words that frequently appear in the same order into single terms. For instance, because the phrase "subprime mortgage" shows up sufficiently many times in one of our corpora, we rewrite it as a single word "subprime_mortgage" joined by an underscore within that corpus. We perform this joining of relatively frequent bigrams and trigrams (or 2- and 3-word phrases) using a tool from the word2vec suite (Mikolov et al., 2013). We also removed stopwords, or words containing syntactic information not relevant to substantive topics. We removed just over 200 unique stopwords.

3.2. Latent dirichlet allocation topic modeling

Although our process includes multiple methods, each part is based on the results of a Latent Dirichlet Allocation (LDA) topic model

[3] There are several implementations of topic models, in addition to LDA. For example, STMs and CTMs. We note that researchers more familiar with these other implementations could most likely use them in place of LDA, though we do not test for comparability in this paper.

(Blei et al., 2003). We base our procedure on LDA because it is the most widely used topic model, and topic modeling is considered the most efficient way to systematically and inductively code the content of a text corpus to date (DiMaggio, 2015; Mohr & Bogdanov, 2013; Nelson, 2017). Though *supervised* machine learning methods, which are designed to allow researchers to classify text according to predetermined concepts, might seem the appropriate tool for our purposes, they rely on hand-coded training datasets. LDA is an unsupervised model, which means that researchers need not fully detail their concepts beforehand and lose out on the benefit of an inductive and automatic exploration of the corpus (DiMaggio, 2015; Evans & Aceves, 2016; Nelson, 2017).

Directly using the topic model results to identify relevant texts, as is done with topic-proportion-only filtering approaches, presents a problem in that topics often do not map onto researchers' concepts (Chuang et al., 2013). This is because the topics will correspond to the dominant structures in the corpus (McAuliffe & Blei, 2008). For example, in one of our corpora, which contains government officials' speeches, discussion based on policies—advocating for them, arguing against them, etc.—is the dominant structuring of the speeches. The topics identified in this corpus therefore tend to reflect specific policy initiatives. If the researcher is interested in the broader concept of housing markets, there may be no topics that apparently reflect this concept. Instead, the concept will be scattered through several topics and may not be recognizable in the most probable terms of any topic, which is usually what researchers use to identify what a topic is about (Mohr & Bogdanov, 2013: 552–554).

One adjustment researchers can make to address this issue is increasing *K*—the number of topics the model solution will identify. We recommend that researchers err on the higher side of *K* in any case to produce more, finer-grained topics, which can facilitate distinguishing between the research concept and related concepts. Choosing *K* is an important step in achieving useful topics, though we stress that there is not a "correct" number for *K*; the goal is "substantively meaningful and analytically useful topics" (DiMaggio, Nag & Blei, 2013, p. 582–583). We further stress that, while quantitative methods exist for choosing *K* (e.g. Lee & Mimno, 2014), LDA solutions with substantively meaningful topics, which researchers should recognize with respect to their research concept, are frequently not the mathematically optimal solutions (Chang et al., 2009; Nelson et al., 2018). Thus, we recommend inspecting several LDA solutions with varying *K* until identifying one with a topic or topics that reflect the research concept. Several additional solutions with still higher *K* should be estimated to see if any additional distinctions emerge between relevant and related but non-relevant concepts. $K = 50$ was sufficient for our government speeches corpus to result in several on-target topics. Our second corpus, consisting of nearly 14,000 full-text sociology journal articles, required $K = 100$ to yield several on-target topics.

3.3. *Identifying relevant topics*

Here, we add an additional straightforward step to aid researchers in identifying relevant topics. While we still examine the top terms in the topics as researchers often do now, we do not rely only on this method. Instead, we recommend additionally using an intuitively derived keyword list and considering the proportion of terms in the topics that are constituted by those keywords. That is, of the distinct terms constituting given topics, what proportion of those terms are in the keyword list? To create the keyword list, the researcher need only do what most researchers are currently doing when creating keyword lists—think of a set of keywords that are specific to their research concept.[4]

The purpose of this intuitive keyword list is not to comprehensively cover relevant terms, but instead to help researchers avoid over-looking topics that may not immediately appear relevant. Table 2 below shows the top terms for the full set of 50 topics for our government speeches corpus and the percent of each topic constituted by our intuitive keyword list. Recall that the research concept of interest in this corpus is housing markets and housing crisis policies. The topics with bold font Fig. 5 in the table are the final set of relevant topics used in the remainder of the process. Consider topic 26 (indicated with the *K*column), which contains terms that reflect discussions of economic growth, with only one term—housing—among the top terms to suggest that it might be a relevant topic. Since 4 percent of all of the terms in the topic are in the intuitive keyword list, this indicates a relevant topic that we might have missed. We do not recommend using an arbitrary percent of terms from the intuitive keyword list for a topic to qualify as relevant. Instead, we recommend using the relative percents as a guide in considering potentially relevant topics, along with examining the top terms of the topic and, if terms are ambiguous, reading documents with a high proportion of that topic. Some topics constituted by a high percent of the intuitive keywords might actually reflect a related but not relevant concept. For example, in our speeches corpus, topic 15 is about a specific Department of Housing and Urban Development (HUD) Indian housing assistance program, as determined by examining the top terms in the topic. It is not about housing markets broadly or a crisis housing policy, so we do not include it as a relevant topic even though it has a relatively high percent of terms in the intuitive keyword list. A similar table displaying the 100 topics for the second corpus is in Table B.1 in Appendix B.

3.4. *Topic proportions, identifying type 1 texts*

To identify texts that are mostly about the researcher's concept—Type 1 texts—we use the distribution over topics, i.e. topic

[4] One approach to augment this list is to train a word embedding model on the corpus such as SGNS or CBOW, popularized in the word2vec package (Mikolov et al. 2013). One can then identify and add the most similar words to those in the researchers' initial list. The effectiveness of this approach will depend on the frequency of the terms of interest and the size of the corpus as the nearest neighbors of words can vary significantly with small changes to the training corpus (Antoniak and Mimno 2018). If taking this approach, we simply recommend inspecting the nearest neighbors before incorporating them. Another approach that we note but do not implement or test here is a "seeded" or "guided" topic model approach in which the researcher's induces some topics by seeding them with particular words (e.g. Hall et al. 2008).

Table. 2
Most Probable Terms of Fifty Topics for Speeches Corpus from LDA Solution and Percent of Terms in Keyword List

K	%	Terms				
1	2	agencies	federal	consumer_protection	financial	consumers
2	0	financial	international	u.s	system	iran
3	0	bonds	tax	bond	credit	state
4	0	service	country	great	life	people
5	1	financial	banks	institutions	services	cra
6	0	cfius	national_security	transaction	committee	transactions
7	0	percent	billion	more_than	year	million
8	0	development	countries	support	resources	economic
9	**3**	**credit**	**loans**	**banks**	**lending**	**markets**
10	0	important	time	policy	public	potential
11	0	investment	funds	sovereign_wealth_funds	hedge_funds	investors
12	1	financial	markets	system	institutions	credit
13	0	risk	risk_management	institutions	risks	supervisory
14	0	tax	health_care	taxes	income	health
15	5	housing	program	section	hud	tribes
16	1	committee	members	opportunity	thank_you	today
17	0	bank	fdic	banks	deposit_insurance	banking
18	0	china	china's	chinese	economic	u.s
19	0	thank_you	today	conference	opportunity	work
20	0	compensation	companies	company	business	government
21	**9**	**people**	**families**	**homeownership**	**housing**	**home**
22	0	capital	basel_ii	banks	u.s	capital_requirements
23	1	market	risk	markets	investors	credit
24	0	energy	oil	production	gas	costs
25	0	countries	global	imf	economies	financial
26	**4**	**growth**	**percent**	**economy**	**year**	**housing**
27	6	housing	hud	communities	local	community
28	0	treasury	process	report	program	information
29	0	rate	federal_reserve	federal_funds	credit	monetary_policy
30	0	financial	system	firms	institutions	crisis
31	2	it's	people	time	make	we're
32	0	banks	capital	institutions	fdic	bank
33	**16**	**loans**	**mortgage**	**borrowers**	**loan**	**subprime**
34	0	liquidity	federal_reserve	banks	market	markets
35	0	economic	work	continue	economy	ensure
36	0	inflation	prices	rate	price	monetary_policy
37	0	u.s	trade	investment	economic	united_states
38	1	markets	financial	demand	prices	increase
39	0	tax	u.s	country	income	treaty
40	1	financial	education	financial_literacy	programs	training
41	1	insurance	state	pension	system	market
42	**3**	**discrimination**	**hud**	**fdic**	**data**	**fair_housing**
43	6	housing	hud	program	families	funding
44	0	treasury	securities	billion	program	debt
45	0	data	payments	system	information	systems
46	**3**	**consumers**	**mortgage**	**disclosures**	**information**	**consumer**
47	0	inflation	monetary_policy	policy	economic	economy
48	0	financial	regulatory	markets	capital	u.s
49	**9**	**fha**	**mortgage**	**gses**	**housing**	**market**
50	**10**	**mortgage**	**borrowers**	**homeowners**	**servicers**	**foreclosure**

Note: Relevant topics have bold font.

proportions from the LDA model. Since the concept may be scattered through several topics, we calculate the sum across all relevant topics. Texts that are constituted, in sum, by 25% percent or more of terms from any of the relevant topics are identified as relevant texts.

There is not yet a standard approach for choosing a threshold like 25% when using topic proportions to filter texts. Researchers tend to use a fairly high threshold to avoid including irrelevant documents, potentially resulting in incomplete datasets (Nelson et al., 2018). A contribution of this paper is that we have tested and devised a starting recommendation for this threshold. We tested a range of thresholds on a labeled subset of the corpora by using Bayesian optimization, a popular strategy used to tune hyperparameters of machine learning models (Snoek, Larochelle & Ryan, 2012). Note that this step is not something that researchers following our process to filter texts need to do. A detailed discussion of this testing procedure is included in Appendix C.

Ultimately, we found that the threshold of 25% for the sum of topic proportions and 15% for the relative entropy keyword list (discussed below) worked well with both of our corpora, which, notably, are constituted by quite different texts. These results are surprising, as they contradict standard intuitions (including those of the authors of this paper prior to this work) that a threshold of 50% or more is necessary for meaningful filtering.

We recommend that researchers use these thresholds as starting points. While these specific thresholds may not be optimal for every project, we recommend using them as the starting points for two main reasons. First, starting with a lower threshold is better because researchers will (more or less systematically) examine their filtered dataset and will recognize if there are many texts that do not belong. However, they are less likely to examine what is left out and will be unaware if a too-high threshold has resulted in an unacceptably incomplete dataset. Second, a high threshold is only sensible if a researcher requires texts that are only about their concept, such that texts that include other concepts alongside their concept should be left out of the filtered dataset. Even short texts, like tweets, can be constituted by several topics, particularly if researchers are working with LDA solutions with a high K in order to distinguish texts with relevant concepts from those with related concepts. To summarize, we recommend using the optimized thresholds we use, followed by an inspection of the filtered dataset. If there are non-relevant texts, raise the thresholds.

3.5. Relative entropy-"refined" topic, identifying type 2 and type 4 texts

Some texts in a corpus are only partially about a concept of interest and/or refer to the concept without keywords—Type 2 and Type 4 texts, respectively. Automatically identifying these texts is less straightforward. Using a topic proportion approach would require lowering the threshold and possibly increasing the false positives. Instead, using the relevant topics, we generate an inductive keyword list that is more targeted than the relevant topics themselves—a "refined" topic. We note that in this case, we are seeking not necessarily words particularly unique to individual topics, as metrics such as topic exclusivity (Bischof & Airoldi, 2012) might recover, but instead a property of being more prominent across relevant topics and not in non-relevant topics.

We considered several scoring functions for selecting words from the relevant topics to generate this list including, using the log-term frequency (Wiedemann, 2016), tf-idf (Manning, Raghavan & Schütze, 2008), and the relative entropy (also known as the KL divergence) of each word's probability in relevant topics with respect to the full topic model. While these three scoring functions generate fairly similar lists, the relative entropy scoring function worked most consistently across both corpora. For each word w, the relative entropy scoring function is defined as

$$c = -p(w)q(w)$$

where $p(w)$ is the probability that a word sampled from the words assigned to our relevant topics is word w, and $q(w)$ is the probability of sampling that word from any word in the corpus. Conceptually, this results in a ranking of words based on their salience in the relevant topics compared to the corpus overall. We include 200 words with the highest relative-entropy contributions c in our resulting refined topic. Researchers can inspect their list of words sorted by relative entropy and raise the threshold from 200 to include additional words that they recognize as relevant or lower the threshold to exclude words that are not.[5] We identify texts as relevant if they are constituted by 15% of the 200 words in the refined topic. We used the same Bayesian optimization procedure from the previous section to arrive at this 15% threshold. See the above discussion on the thresholds for guidelines in determining if it may be necessary to adjust it.

The intuition behind using this relative entropy-refined topic method is that, while the LDA topics themselves may not map neatly to researchers' concepts of interest, high-quality terms specific to relevant topics do to a larger extent. A more targeted set of terms drawn from relevant topics can be used to identify texts that are constituted in smaller proportions by the research concept—Type 2 texts—without running such a high risk of adding false positives.

Searching for a relatively low percentage of highly relevant terms is also a good approach for targeting Type 4 texts, which have only implicit references to the concept. For example, a Type 4 text might contain the words "credit", "market", "home", "loan", "borrower", or "adjustable-rate".[6] None of these are keywords on their own, but together they suggest a relevant implicit reference. The benefit of using LDA topics to create the refined topic is that topic modeling will identify terms constituting researchers' concepts as they are actually referred to by the producers of the texts.

3.6. "Super" keyword list, identifying type 3 and type 4 texts

In this step, we create a "super" keyword list that is even more targeted than the relative entropy-refined topic used to identify Type 2 and Type 4 texts above. Additionally, this method incorporates researchers' knowledge and abilities to recognize relevant terms. To create the super keyword list, we examine an expanded list—the top 1000 words—of high-relative-entropy-contribution words from the last step and select those words that are unambiguously related to the concept of interest, i.e. likely to be used when referring to the concept of interest and no other concepts. Since researchers are using context knowledge and judgment for each word that is included, there is little trade-off in terms of a complete versus polluted dataset. However, there is a trade-off in terms of researcher time spent reviewing a list consisting of 200 terms compared to one consisting of 1000. We test the effects of the three thresholds (50, 200, and 1000) and share the results in Appendix D. Fig. D.1 shows, for the sociology articles corpus, that expanding the list under consideration for the super keywords from the top 50 relative entropy words, to 200, then to 1000 increases recall from 0.36, to 0.46, to 0.53,

[5] We tested three thresholds (50, 200, and 1000). We share a plot in Appendix D showing the varying performance across these three thresholds. This shows a clear tradeoff between completeness and pollution in raising/lowering this threshold. Figure D.1 shows that while recall increases from .49 (with a threshold of 50) to .74 (with a threshold of 1000), precision decreases from .87 to .61.

[6] These words are among the top 15 words from the relative entropy list for our speeches corpus.

respectively; all while also increasing precision from 0.78, to 0.81, to 0.82, respectively. This is a strong improvement, and we furthermore note that, relative to the time necessary for human reading and hand coding, the additional time it takes to consider 1000 words versus 200 is not much.

In designing this process of creating the super keyword list, we considered two factors. The first is that a topic model-derived list will be more comprehensive since topic models are better than humans at clustering terms in a data-driven manner. The second factor is that humans are good at recognizing words that reflect their concept (King, Lam, and Roberts, 2017). Thus, a researcher examining a comprehensive list of candidate words will be able to easily select those "super" keywords that unambiguously reflect the concept.

In order to improve the likelihood of identifying the implicit references in Type 4 texts, we select not only single terms from the relative entropy list, but also consider any pairs of terms that together create a super keyword. For example, the relative entropy list included the words "loan" and "modification". Neither of these words on their own would identify relevant texts, but any text containing the phrase "loan_modification" would be a relevant text. We note that researchers should use their context knowledge in creating this superkey word list. Out of context, "loan_modication" is not necessarily an unambiguous phrase. In the context of this corpus, however, it is (i.e. a corpus containing government speeches during a housing crisis in which loan modifications were a tool used to prevent foreclosures).

We use these super keyword lists by identifying as relevant any text that contains any one of the super keywords or super keyword pairs. Since some of these super keywords are multi-term, there are multiple ways to implement this step, ranging from more to less conservative. The most conservative approach requires that all of the terms of multi-term words appear in the text next to each other. For example, for "subprime_mortgage", "mortgage" should immediately follow "subprime". In a second, less conservative implementation, we consider any multi-word term that contains all of the words within a given super keyword (or keyphrase) regardless of order and allowing some separation of the words. For example, if "gender_gap" was a super keyword, and "gender_pay_gap" was another multi-term vocabulary element, any text containing "gender_gap" or "gender_pay_gap" would be included. Finally, in the least conservative implementation, the terms just had to appear, in some order, somewhere in the text. We tested each approach, and the latter two less-conservative options performed similarly, but the middle approach had fewer false positives. The results below are based on the second implementation. Table 3 shows a summary of all the steps for the complete procedure.

4. Creating a subset of texts

We test the performance of this procedure on two separate corpora and two research concepts. To perform the tests, we extract a random subset of texts from each corpus and manually label the texts according to the typology (see Table 1). We then use these manually labeled texts to analyze how the procedure performs at identifying each type of relevant text and excluding both types of non-relevant texts. We run the procedure on the full corpora as researchers implementing the procedure on their own data should do. This approach may cause some readers concern about overfitting, which is often an issue with classifying text, since we are not running the procedure on a training set and then testing on a held-out set of data. The purpose of this procedure, however, is to provide structure to a currently unstructured and untested approach to filtering for a complex research concept. The procedure is meant to be based on a full corpus of text, yielding the best topics. It is also semiautomatic with important researcher interventions and can be iterative if researchers discover useful adjustments along the way. We make recommendations for thinking through adjustments where adjustments can be made.

One of the corpora consists of all public addresses (speeches and congressional testimonies) given by officials from four U.S. federal agencies—the Federal Deposit Insurance Corporation (FDIC), the Federal Reserve, the Department of Housing and Urban Development (HUD), and the Treasury Department. These speeches span the timeframe of the U.S. housing market crisis—from January 1, 2006, through December 31, 2009. The concept of interest is the combined concept of the housing markets and housing crisis policies. Two related concepts that are excluded are homeownership and non-crisis housing policies.[7] The obviously overlapping boundaries of these concepts makes identifying the texts a researcher would want in a subset a realistically complex research problem.

After the initial preprocessing step of separating the speeches into (roughly) paragraphs, we randomly extracted 500 of the resulting 25,981 texts. The first author manually labeled each of the 500 texts twice and resolved the discrepancies between the two sets of labels. The first set of labels resulted in 121 (24%) relevant texts. The second set resulted in 141 (28%) relevant texts. All 20 of the texts that received a different coding the second time were changed from non-relevant to relevant. The reason for the change for all but 6 of the 20 was because of a coding clarification in which references to pre-crisis housing policies that were adapted and used as a response to the crisis should be coded as relevant; for example, those related to the Community Reinvestment Act (CRA)—a pre-crisis policy that was adapted during the crisis.

The second corpus consists of all articles from six sociology journals, spanning a timeframe beginning in the year 1895 and going through 2015.[8] The research concept for this corpus is inequality; including economic, racial/ethnic, and gender inequality. We exclude the related concepts of power and general discussions of the economy, race, or gender. We followed the same preprocessing steps for both corpora. Separating this corpus into (roughly) paragraphs resulted in 777,402 texts. Because this was both a larger and a more diverse corpus (in terms of both timespan and topics), we randomly extracted 1000 texts to manually code. The first author manually coded the 1000 articles twice and resolved the discrepancies, yielding 328 (33%) relevant texts. Fig. 1 shows how these

[7] This corpus was collected and the research concept defined by the first author for a separate research project.

[8] The 6 journals are: American Journal of Sociology, American Sociological Review, European Journal of Social Theory, Sociological Theory, Theory and Society, and Theory, Culture, and Society. This corpus was collected by the fifth author for a separate research project.

Table. 3
Summary of Procedure Steps

Pre-procedure steps	
Step 1	Create short, intuitive keyword list encapsulating concept
Step 2	Infer an LDA topic model on the full text corpus
Step 3	Identify LDA topics relevant to concept using the keywords identified in Step 1 and domain expertise
Multi-Part Filtering Methods	
Topic Proportion	Measure the inferred topic proportion of each document. A document is relevant if the sum of the proportions of relevant topics for a document ≥ 0.25. This method is aimed at **Type-1 texts**.
Relative Entropy	Create a list of 200 key terms with the highest *relative entropy* of the probability of each word in the relevant topics $p(w)$ with respect to the probability of the term in the full corpus $\mathrm{q}(w)$. A document is relevant if the proportion of words in the document matching those in this refined topic is $\geq .15$. This method is aimed at **Type-2** and **Type-4 texts**. $relent = p(w)\log\frac{p(w)}{q(w)}$
Super Keywords	Using the intuitive keywords from Step 1 and the 100 relative entropy keywords, select words that unambiguously indicate relevance to the research concept and not related but excluded concepts. A document is relevant if it contains any single super keyword or any combination of multi-term super keywords. This method is aimed at T**ype-3 texts** and **Type-4 texts**.

manually labeled texts from both corpora are distributed across the typology. The bars in the figure show the percent of each type of text contained in the corpus. The black bars represent the sociology articles corpus and blue bars represent the government speeches corpus. This figure shows that the majority of texts in both corpora are non-relevant texts (types 5 and 6), suggesting the usefulness of accurate, efficient procedures that can filter large corpora down to what is relevant for researchers. This figure also suggests that corpora vary in their constitutions of the types of relevant texts. For example, the speeches corpus is made up of a larger percent of texts that are mostly about the research concept (Type 1) than the sociology corpus is. The sociology corpus is constituted in the largest percent by Type 4 texts, those that refer only implicitly to the research concept. A filtering procedure applied to the sociology corpus not geared toward recognizing Type 4 texts would result in an incomplete dataset.

After implementing our procedure separately on both full corpora, we examined how it performed at categorizing the manually labeled texts. As mentioned previously, the two main failures in selecting a subset of a larger corpus are pollution and/or incompleteness. Two performance measures in particular formalize these issues: precision and recall. A precision score is the proportion of texts identified as relevant that are truly relevant; the higher the precision score, the less polluted the dataset. A recall score is the proportion of all texts that are relevant that are recognized as relevant; the higher the recall score, the more complete the dataset. A dataset may have a high recall score and be fairly complete, while also having a low precision score and be highly polluted. An $F1$ score—the harmonic mean of the precision and recall scores—is a measure of overall accuracy.

Below, we compare our procedure's performance on these measures to the currently common filtering approaches—the dictionary-only and topic-proportion-only approaches. For the dictionary-only approach, we developed standard dictionaries for each research concept. For the speeches corpus, the first author, who collected and analyzed that corpus for another research project, developed the dictionary, which is available in Appendix E. For the sociology corpus, the dictionary consisted of all of the terms from the dictionaries used in Nelson et al., 2018. These dictionaries were focused specifically on economic inequality. Since our research concept included racial/ethnic and gender inequality, we expanded the dictionary to include key terms for these additional dimensions of inequality. The full dictionary is available in Appendix E.[9]

For the topic-proportion-only approach, we used the same relevant topics identified for our procedure and used the same threshold (25%). That is, any text that contained a sum of at least 25% of terms from any of the relevant topics would be identified as relevant through this approach. WWe are unaware of any work that uses a threshold this low to filter, which suggests that published work to date may be based on datasets with high precision scores but low recall scores.

5. Results

Fig. 2 displays the $F1$, precision, and recall performance scores for our procedure compared to the dictionary-only and topic-proportion-only approaches for each corpus. The left-most panel shows the $F1$ scores, the middle panel the precision scores, and the right-most panel shows the recall scores. The scores for our procedure are indicated with the blue bars, those for the dictionary-only approach with the black bars, and those for the topic-proportion-only approach with the gray bars. The scores for the sociology corpus are in the upper panel, and those for the speeches corpus are in the lower panel. Our procedure performs better overall than either the dictionary-only or topic-proportion-only approaches on both corpora, as measured by the $F1$ scores. For the sociology corpus, our procedure scored highest at 0.65, while the dictionary-only and topic-proportion-only approaches scored 0.53 and 0.30, respectively. For the speeches corpus, our procedure scored an $F1$ of 0.77, while the dictionary-only and topic-proportion-only approaches scored 0.35 and 0.61, respectively.

Our procedure also performs better at recall on both corpora than either the dictionary-only or topic-proportion-only approaches, indicating more complete datasets. The right-side panels of Fig. 2 shows that our procedure has a recall score of 0.53 for the sociology

[9] As with our super keyword lists, there are three different ways of using the terms in the dictionaries. We used these dictionaries in the same way that we used our super keyword lists, i.e. by using the middle implementation (see section 3.6 above).

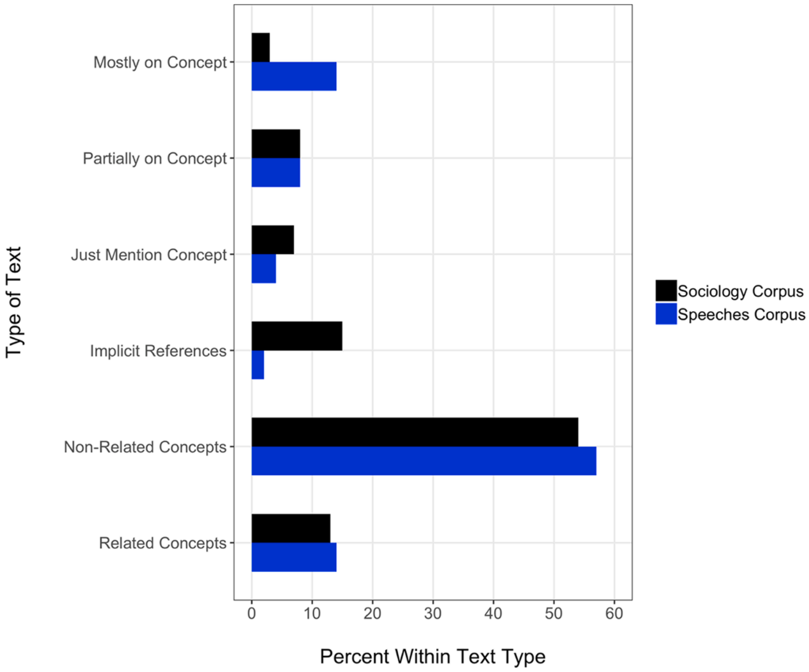

Fig. 1. Descriptive statistics of the percent of each category in the typology for each of the two evaluation corpora.

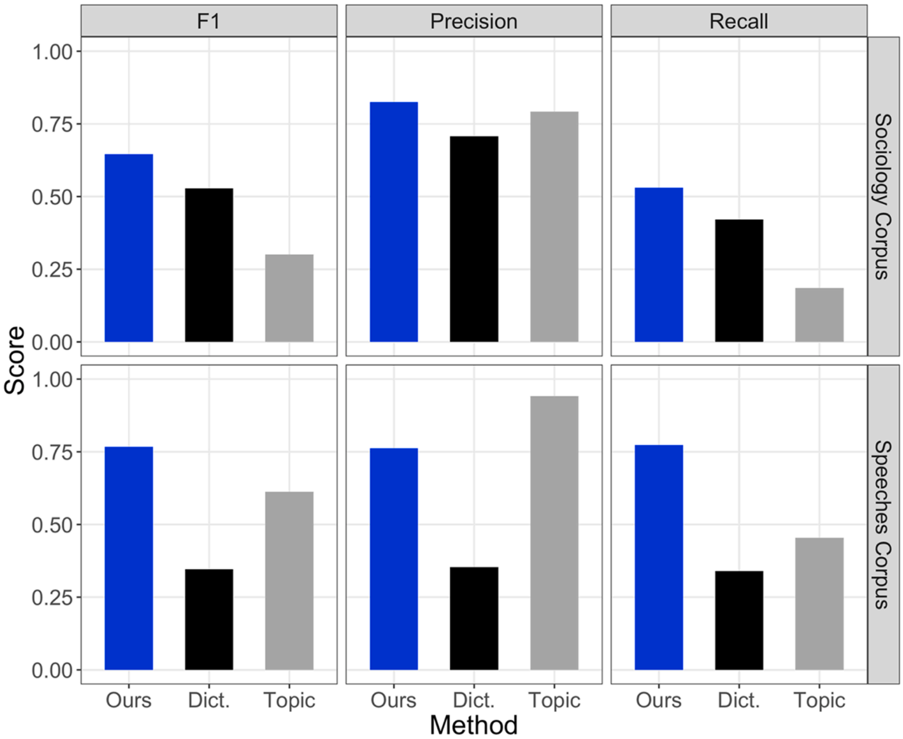

Fig. 2. *F*1, precision, and recall performance measures for our procedure compared to the dictionary-only and topic-proportion-only approaches.

corpus. This means that 53% of all of the relevant texts were recognized as relevant. The dictionary-only and topic-proportion-only approaches achieved recall scores of only 0.42 and 0.19, respectively.

For the speeches corpus, our procedure had a recall score of 0.77, indicating that 77% of all relevant texts were recognized as such. The dictionary-only and topic-proportion-only approaches achieved recall scores of only 0.34 and 0.45, respectively.

Across the three measures—*F*1, precision, and recall—our procedure is weakest in precision compared to the other approaches. However, the absolute precision scores are nevertheless quite high for both corpora. More specifically, our procedure performs about as well on precision (at 0.82) as the topic-proportion-only approach (at 0.79) with the sociology corpus. This means that 82% of the filtered dataset is actually relevant text. The topic-proportion-only approach is more precise than our procedure on the speeches corpus—0.94 compared to 0.76. Better precision scores are often achieved at the cost of recall, which is the case here for the other approaches. In sum, our procedure achieves similar precision scores while also maintaining the highest recall scores. This is why the overall *F*1 scores for our procedure are higher in spite of the similar or lower precision.

In addition to these performance measures, we can also consider how each approach performs across the different types of texts (see Table 1 above). Fig. 3 shows the percent of each type of text that our procedure, the dictionary-only, and the topic-proportion-only approach identified as relevant in each corpora. For example, for the sociology corpus, our procedure identified as relevant 88% of Type 1 texts, which are those that are mostly about the research concept, while the dictionary-only approach identified as relevant 75% of these texts and the topic-proportion-only approach identified 44%. Importantly, this figure shows that our procedure identified as relevant a greater percent of each of the 4 types of relevant texts than either of the other approaches, across both corpora.

Our procedure identified as relevant fewer of the non-relevant Type 5 texts, which are those discussing completely unrelated topics. Our procedure's performance on Type 6 texts, those that are on related but non-relevant topics, is poorest. On both corpora, our procedure identifies as relevant a greater percent of the non-relevant type 6 texts than either of the other two approaches, though our procedure performs only slightly worse.

Given the considerable improvement on recall over the off-the-shelf approaches, some loss of precision is an expected trade-off. Our procedure's somewhat worse performance on Type 6 texts accounts for the similar or lower precision scores discussed above.

Since our complete procedure is constituted by three parts or methods, we can examine how each method separately contributes to the overall performance. The three methods used to identify relevant texts are: topic proportion sum, relative entropy refined topic, and super keywords. Fig. 4 shows the percent of each type of text identified as relevant by each of these methods separately.

The topic proportion method in our procedure is targeted at identifying Type 1 texts—those that are mostly about the research concept. The topic proportion method is represented by the black bar in the figure, and this method did identify a higher percent of Type 1 texts than any of the other types of text across both corpora.

The relative entropy refined topic is targeted at identifying Type 2 and Type 4 texts—those that are only partially about the research concept or refer to the concept implicitly. This method performed underwhelmingly in that it identified about as many Type 2 and Type 4 texts as the topic proportions method did. However, if these are different texts from what the topic proportion method identified, then this would still be an important contribution to the overall performance.

The super keywords are targeted at identifying Type 3 and Type 4 texts—those that just mention or implicitly refer to the research concept. The super keyword lists did indeed outperform the other methods in identifying both of these types of texts. In fact, the super

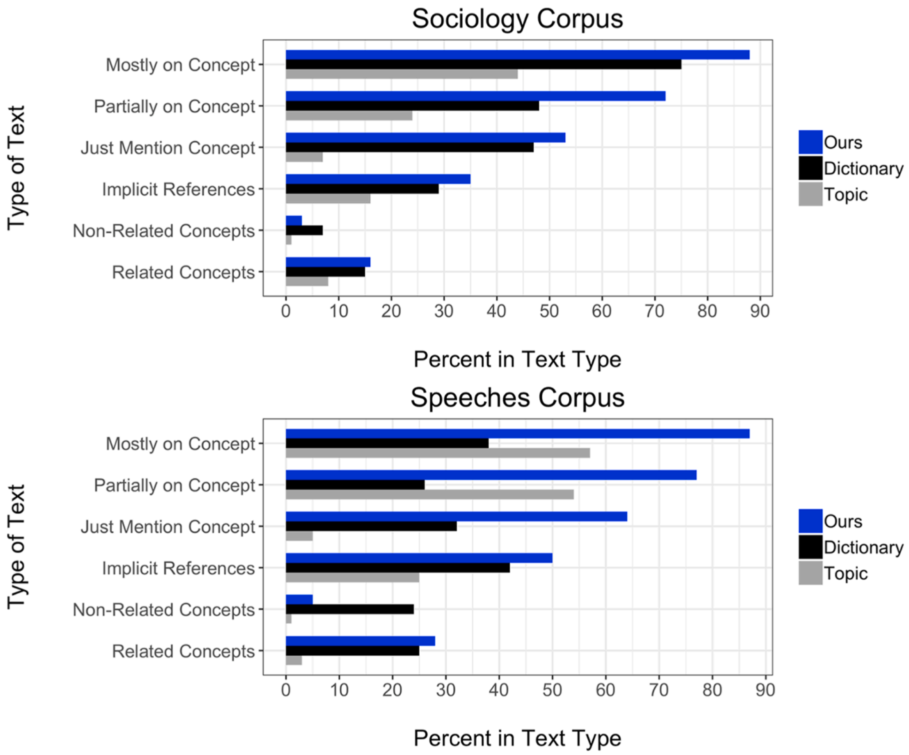

Fig. 3. Percent of each type text that our procedure, the dictionary-only, and the topic proportion only approach identified as relevant.

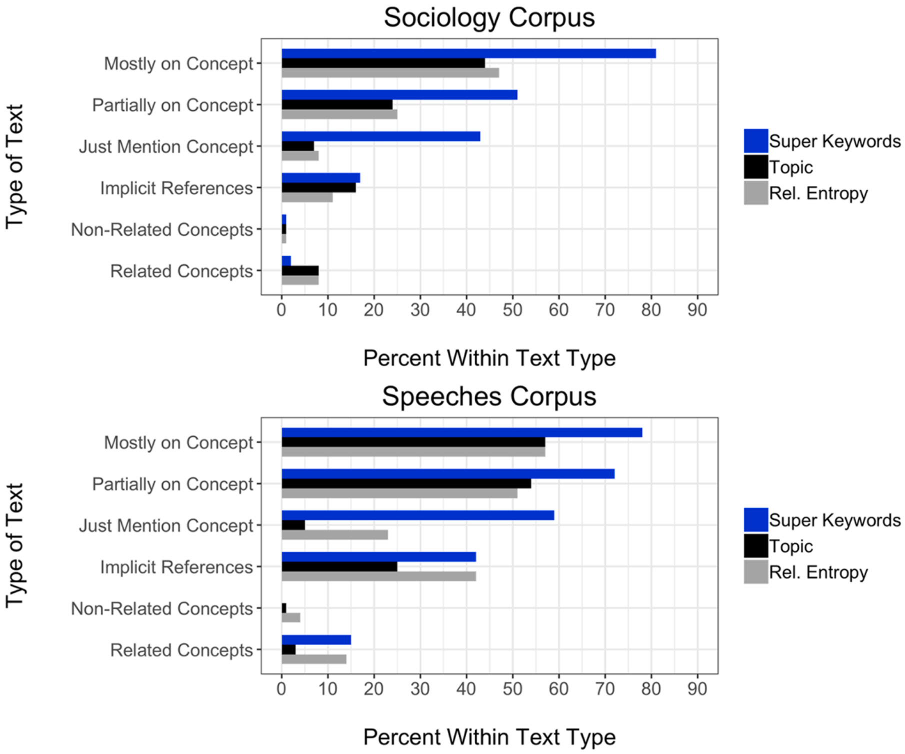

Fig. 4. Percent of each type of text separately identified as relevant by each of the three parts of our procedure. The three parts are: the total topic proportion, relative entropy refined topic, and the super keywords.

keywords outperformed the other two methods across all four types of relevant texts. We discuss this further in the discussion section.

A final consideration regarding the performance of our procedure is whether the texts that each of the methods identify as relevant are non-redundant. It is possible that the super keywords are doing all of the work. There would be little benefit from implementing this multi-part procedure unless each part is uniquely contributing to the overall performance. To check this, we calculate the additional percent of each type of text uniquely added by the topic proportion method, beyond what the super keywords had already identified. Then, we calculate the percent of type of text uniquely added by the relative entropy refined topic, beyond what the super keywords and the topic proportion methods had already identified. We plot these percents in Fig. 5.

The blue bars show the percent of each type of text identified as relevant by the super keyword list. The black bars show the percent of texts additionally identified as relevant by the topic proportion method and the gray bars show the percent of texts additionally identified by the relative entropy refined topic. This indicates that, while the super keywords are doing a lot of the work, they are not doing all of the work; each of the methods is uniquely contributing to the procedure’s performance.

6. Discussion and conclusion

In this article, we develop and test a procedure that provides structure for the currently unstructured process of using computational text analysis tools for coding or filtering a large text corpus for specific research concepts. Although many social scientists are already using these tools, there is currently no standard on how best to apply them in filtering for complex research concepts. Indeed, prior to Nelson et al. (2018), the available off-the-shelf tools social scientists are using were entirely untested at filtering for the sort of complex concepts social scientists are often interested in. Given the seemingly endless expansion of available text data and the still-growing interest in “big” data, an effective procedure should be highly valuable in allowing researchers to work with these data without setting arbitrary limits based on resource constraints, like time for human reading. Human reading is still an important component of many research projects, but such efforts are better spent at more sophisticated interpretation tasks, not filtering.

Our procedure is especially well-suited for social scientists, who tend to work with “complex, socially constructed, and unsettled theoretical concepts, often with ill-defined boundaries”, as it requires neither manual coding nor the *a priori* refinement of a research concept (Nelson et al., 2018: 1). Researchers can inductively refine their research concept as they work through each part of our semiautomatic procedure—while examining the initial topics from the LDA solution, while examining the relative entropy terms, and while developing the super keyword list. Our procedure encourages an iterative analytical approach, which allows researchers to adjust the boundaries of a research concept and apply these to earlier filtering steps. Even in cases where a researcher does not have a specific research concept in mind prior to the inference of a topic model but wishes to first explore the themes in a corpus (McFarland et al., 2013), our procedure is still useful if the researcher intends to ultimately filter on discovered topics.

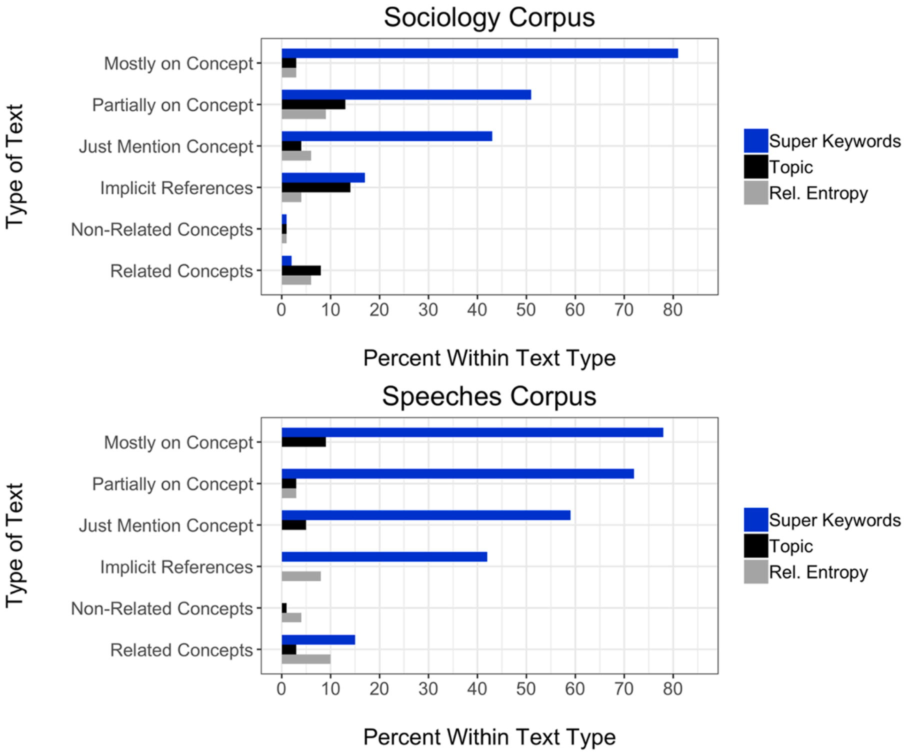

Fig. 5. Percent of each type of text non-redundantly identified by each method. I.e., the blue bars show the percent of each type of text identified as relevant by the super keyword list. The black bars show the percent of texts additionally identified as relevant by the topic proportion method, and the gray bars show the percent of texts additionally identified by the relative entropy refined topic.

Tests presented in this paper show that our procedure out performs two common off-the-shelf approaches at filtering two large corpora for two different complex research concepts. While yielding similar outcomes in terms of precision, our method especially out performs the other methods in terms of recall and, therefore, has higher overall accuracy. Our approach achieves this by incorporating the benefits of the off-the-self approaches and adjusting for their deficiencies by combining the approaches and adding simple steps using the power of human researchers in recognizing terms and topics that reflect their research concepts (King, Lam, Roberts, 2017). One additional step we recommend is considering the proportion of relevant terms (as indicated by an intuitively-derived keyword list, see Section 3.3) in inferred topic modeling topics as another way to identify potentially relevant topics. Usually, researchers examine just a few of the most probable terms in a topic, which risks overlooking potentially useful topics. In a second additional step, we recommend creating a list of high relative entropy terms from the inferred topics (see Section 3.5). Our procedure then uses this list in two ways. One, by using the top 200 of these terms as a refined topic with which to identify relevant texts. Then, by searching through the list and selecting super keywords as a systematic and inductive way to create a highly effective dictionary. This procedure further structures the process of filtering by testing arbitrary thresholds and discussing reasonable guidelines for inspecting results and making adjustments if necessary.

The inductive aspect of our approach is a critical factor in achieving good filtering performance. The inductive nature of topic modeling is the best argument for using the topic-model-only approach to filtering. However, as our results show (see Fig. 4) the super keyword lists we developed performed the best of the three methods constituting our procedure. The super keyword lists are essentially a refined dictionary and applied in a similar manner to how researchers usually employ traditional dictionaries. Given that these super keyword lists performed the best across the three methods constituting the full procedure, readers would be justified in wondering whether dictionaries are the best approach. We stress that developing a high-performing traditional dictionary usually requires an extended process of careful refinement. For example, for the sociology corpus research concept of inequality in this article, we used a dictionary developed over several research projects by several scholars (see Nelson et al., 2018). Even so, such a high-quality dictionary may not take into account the nuances of inequality in a particular corpus, with the result that, while precision may be high, recall may suffer (Grimmer & Stewart, 2013: 8–9). Developing the super keyword lists inductively, as our procedure does, allows for the development of precise as well as comprehensive dictionaries in a relatively short amount of time. These super keyword lists outperformed even the authors' expectations, and, thus, we encourage researchers implementing our method to focus in on this part of the procedure. We compiled the super keyword lists by taking single terms from the relative entropy list as well as making pairs. The performance of these lists could be improved even further by considering relatively entropy terms beyond the 1000 terms threshold or creating not just bi-grams but also tri-grams, etc. Since researchers are using context knowledge for each word that is included in the super keyword lists, there is little trade-off in terms of a complete versus polluted dataset (see Appendix D).

In addition to the development and testing of the filtering procedure, this paper also contributes the development and analysis of the typology of texts (see Table 1). We are unaware of other work that considers the constitution of textual datasets from this perspective. Researchers already recognize that pre-processing decisions on textual datasets can have important effects on analyses (Denny & Spirling, 2017), just as sample-selection and missing data can affect conclusions drawn from analyses of traditional datasets. We argue that it is also important to consider the constitution of corpora from the perspective of these different types of texts, particularly when working through the process of filtering. As discussed above, the sociology corpus is constituted in the largest percent by Type 4 texts (those that refer only implicitly to the research concept; see Fig. 1). The topic-proportion-only approach performed the worst at identifying those texts with implicit references to inequality. Conclusions about the prevalence of inequality in sociology articles, for example, would likely be inaccurate if the corpus were filtered with this approach. Additionally, it may be older articles that lack contemporary, explicit references to inequality. Researchers analyzing, for example, how scholarship on inequality has changed over time would likely have biased results. We hope that future research examines how filtering processes perform with respect to these types of texts and how differently constituted datasets affect analyses.

This paper contributes by offering an effective procedure that structures the process of using computational tools for coding or filtering for complex research concepts. In addition to the possibility for improving the performance of the super keyword lists just discussed above, there are likely other ways to improve this process. We hope that this paper serves as a beginning framework upon which others will build, improving scholarship using computational content analysis.

Appendix A. Typology of texts

The corpus these example texts are from consists of speeches by officials from one of four U.S. federal agencies from 2006 to 2009. The research concept is the housing markets and housing crisis policies. Table A.1 shows, in the example for type 1 texts, a text that is entirely about the housing markets. The example type 2 text reflects a discussion that is about the Federal Open Market Committee's (FOMC) thinking and policy decision, financial market disturbances, and the housing markets. This text is thus partially about the research concept. The example type 3 text just mentions housing crisis policies at the end but is otherwise about non-related concepts

Table. A1
Typology of Texts as Related to a Research Concept

Types of texts	Example in relation to: housing market crisis & crisis housing policies
Relevant texts	
Mostly about concept	But the current contraction in housing did follow an unusually_large run_up in sales and construction and, even more so, in prices relative to the returns on other financial and real assets. Our uncertainty_about what pushed home_prices and sales to those elevated_levels raises_questions about how the market will adjust now that expectations of the rate of house_price_appreciation are being trimmed…. ~ Federal Reserve Jan., 2007
Partially about concept	The FOMC noted that the downside_risks to growth had increased appreciably. However, to allow time to gather and evaluate incoming information, possible policy action was deferred.... A key issue at that meeting was the extent to which the market disturbances had affected the outlook for the housing sector. Financial markets overall had improved somewhat, but tighter terms and standards in the mortgage market particularly in the nonprime and jumbo segments appeared likely to intensify the correction in housing significantly, with adverse_implications for construction activity and house_prices. ~ Federal Reserve Oct., 2007
Just mention concept	Today, we spend more per_person in health_care than any country in the world. Like tax policies that allow our small_businesses to reinvest their profits into innovation and job_creation, rather_than sending more dollars back to Washington. Small_businesses create two_thirds of the new jobs in our country. We need that entrepreneurial energy to continue. And we need policies to help people keep their homes at a time when so many of them are losing them. ~ HUD Dec., 2008
Refer to concept without keywords	And much is at stake. These are very challenging_times for everybody. And the American_people expect their leaders and our government…to lock arms and work to turn this economy around. Thank_you. ~ FDIC Mar., 2009
Non-relevant texts	
Entirely about other concepts	The administration supports the role of the intelligence_community as an independent advisor to CFIUS, and thus opposes giving the DNI a policy role, rather_than an advisory role. As I stated previously, the DNI has a formal role in the process to coordinate and facilitate the intelligence assessment in each CFIUS investigation. I must also point_out that H.R. 556, as currently drafted, retains a timing conflict that was present in H.R. 5337. ~ Treasury Feb., 2007
About related concepts, contains related terms	The President has delivered increased funding, record levels, to enable our partnership to reach more homeless people, especially with more permanent_housing. Last_year, HUD announced grants of $1.5_billion nationwide to address homelessness, the latest in a commitment that,_since_2001, has totaled_approximately_$10_billion to support housing and services. Together, we have been able to devote more resources to help persons who are homeless. ~ HUD Jul., 2008

like healthcare and tax policies. The example type 4 text shows a speech in which the official says that the “American_people expect their leaders and our government…to lock arms and work to turn this economy around.” This example demonstrates the most difficult classification problem for automatic techniques. A human researcher, armed with expertise on the research topic and the ability to read surrounding texts, can easily identify this text as an official advocating for government action in the housing crisis and therefore a relevant text. A computer, on the other hand, has only the words in this text.

Appendix B

Table. B.1.

Table. B.1
Most Probable Terms of 100 Topics for Speeches Corpus from LDA Solution

K	%	Terms				
1	1.4	work	women	men	family	time
2	0.7	social	life	society	people	men
3	0.1	science	research	scientists	scientific	academic
4	0.1	per_cent	population	number	age	total
5	0.0	actors	model	choice	individual	behavior
6	0.0	study	data	sample	respondents	survey
7	0.0	indian	culture	native	india	eurpoean
8	0.0	network	networks	ties	structure	actors
9	1.2	figure	number	distribution	probability	values
10	0.0	question	fact	does_not	point	case
11	1.2	countries	united_states	country	nations	development
12	0.0	language	meaning	words	word	communication
13	0.4	weber	durkheim	weber’s	durkheim’s	social
14	0.2	social	individual	life	society	culture
15	0.4	social	delinquency	health	age	adolescents
16	0.1	organizations	organizational	organization	competition	institutional
17	0.2	sociology	research	department	chicago	california
18	0.0	modernity	theory	history	philosophy	critique
19	0.5	social	research	analysis	study	important
20	1.1	land	production	agriculture	agricultural	labor
21	**1.6**	**policy**	**state**	**social**	**welfare**	**policies**
22	**8.3**	**income**	**inequality**	**earnings**	**economic**	**income_inequality**
23	0.0	people	time	work	good	things
24	**7.8**	**white**	**black**	**whites**	**race**	**blacks**
25	0.0	game	play	sport	ritual	games
26	0.0	public	media	news	events	television
27	0.3	crime	police	criminal	prison	crimes
28	0.4	occupations	professional	occupation	occupational	workers
29	0.0	talk	interaction	conversation	response	turn
30	0.0	patients	medical	health	patient	hospital
31	0.0	art	esthetic	images	work	image
32	**4.0**	**identity**	**gender**	**race**	**gay**	**identities**
33	0.6	firms	business	firm	industry	corporate
34	0.2	period	change	time	age	years
35	0.2	south	states	state	north	southern
36	2.0	model	age	years	table	education
37	**2.8**	**women**	**men**	**female**	**male**	**sex**
38	0.0	political	party	vote	parties	election
39	0.2	new_york	london	cambridge	eds	politics
40	**2.4**	**class**	**production**	**economic**	**capitalism**	**marx**
41	0.1	exchange	power	actors	trust	relations
42	0.1	social	networks	friends	ties	capital
43	**2.8**	**status**	**class**	**social**	**position**	**classes**
44	0.0	emotions	love	emotional	freud	emotion
45	**1.9**	**education**	**occupational**	**occupation**	**income**	**effects**
46	0.0	social	cultural	culture	practices	actors
47	0.0	foucault	power	body	subject	life
48	0.7	bourdieu	field	cultural	social	bourdieu’s
49	0.3	variables	table	factor	measures	analysis
50	0.5	japanese	japan	soviet	russian	russia
51	0.5	global	world	national	globalization	political
52	0.0	role	behavior	social	values	norms
53	0.3	human	biological	animals	species	environmental
54	1.7	market	money	economic	markets	price
55	0.8	job	workers	jobs	work	employment

(*continued on next page*)

Table. B.1 (*continued*)

K	%	Terms				
56	0.0	technology	information	media	space	time
57	0.0	world	human	nature	reality	life
58	0.0	theory	social	sociology	theoretical	work
59	0.6	table	percent	high	differences	low
60	0.3	social	sociology	science	study	history
61	0.7	ethnic	immigrants	united_states	immigration	immigrant
62	0.1	group	groups	members	individual	individuals
63	0.1	model	models	variables	effects	table
64	0.0	movement	political	protest	movements	social_movements
65	0.2	data	measure	analysis	number	sample
66	0.0	mexico	italian	italy	mexican	latin_america
67	0.1	chinese	china	government	england	political
68	0.0	school	students	schools	education	college
69	0.2	new_york	cit	ibid	chicago	journal
70	0.0	paris	der	und	die	des
71	0.1	items	scale	scores	score	item
72	**0.4**	**mobility**	**class**	**origin**	**table**	**status**
73	0.0	food	street	city	house	home
74	0.6	city	cities	urban	population	community
75	0.1	music	film	musical	source	content
76	0.0	war	military	violence	army	peace
77	0.3	workers	labor	union	unions	strike
78	0.0	work	organization	organizational	organizations	control
79	0.0	law	legal	laws	state	rights
80	0.0	community	organizations	local	members	organization
81	0.0	religious	religion	church	catholic	catholics
82	0.0	simmel	elias	theory	culture_&_society	vol
83	0.3	political	american	movement	revolution	french
84	0.1	effect	effects	model	results	significant
85	0.0	state	work	public	made	committee
86	0.0	data	theory	analysis	research	study
87	0.3	problems	conditions	make	change	resources
88	0.0	man	life	people	world	death
89	0.1	family	children	parents	child	families
90	0.0	political	public	moral	social	habermas
91	2.0	neighborhood	black	segregation	neighborhoods	blacks
92	1.1	state	power	political	economic	institutions
93	0.0	jewish	jews	isreal	islam	islamic
94	0.5	status	task	subjects	characteristics	group
95	0.0	action	social	behavior	actions	act
96	0.7	social	system	society	systems	structure
97	0.0	studies	e.g.	et_al	research	found
98	0.5	marriage	women	married	fertility	age
99	0.0	published	book	books	articles	work
100	0.4	attitudes	political	alienation	attitude	social

Note: Relevant topics are outlined and have bold font.

Appendix C. Bayesian optimization testing procedure for identifying thresholds

To determine an optimal value for the topic-proportion threshold, we optimize over the $F1$ scores of a manually labeled subset of the corpora. An $F1$ score is an overall measure of accuracy that combines: 1) a precision measure, which is the proportion of documents identified as relevant that are truly relevant with 2) a recall measure, the proportion of all relevant documents that are identified as relevant. We used a scoring function that calculated the $F1$ scores of the manually labeled subset for a series of topic proportion sum thresholds, using Bayesian optimization with Gaussian processes to determine the threshold that resulted in the best $F1$ score. Bayesian optimization finds a distribution of different functions that describe the relationship between parameters (thresholds, in this case) and an output function ($F1$, in this case). The algorithm starts by picking random parameter inputs within given ranges and uses the function value at that point to update its prediction. Then, it uses the new prediction to choose the next set of inputs to check and continues iterating until it reaches the number of iterations specified.

Using scikit-optimize, a Python implementation of Bayesian optimization, we optimized two thresholds: the minimum sum of topic proportions and the minimum proportion of words from a relative entropy keyword list, discussed below (Head, 2018). We used 500 iterations of optimization across numerous trials. Ultimately, we found that the threshold of 25% for the sum of topic proportions and 15% for the relative entropy keyword list worked well with both of our corpora, which, notably, are constituted by quite different texts.

For detailed information on the manual labeling of documents in the corpora necessary to run the above tests, see Section 4 in the main body of the paper.

Appendix D. Threshold tests

We show the change in performance across varying levels of two different thresholds. One threshold is the number of high relatively entropy contribution terms included in the refined topic. See Section 3.5 of the paper. Plots D.1 and D.2 show the numbers of the actual (rows of the matrices) and predicted (columns of the matrices) relevance/irrelevant labels and the *F1* scores (at the top of the matrices). The y-axis labels at the left indicate the three different relative entropy thresholds (50, 200, and 1000) we tested. The other threshold we examined is expanding the list under consideration for the super keywords (see Section 3.6). The x-axis labels across the top of the plot shows the three thresholds we tested (50, 200, and 1000). The middle rows of each plot show the threshold we ultimately used for the refined topic, and the right-most columns show the threshold we used for the super key list. Fig. D.2

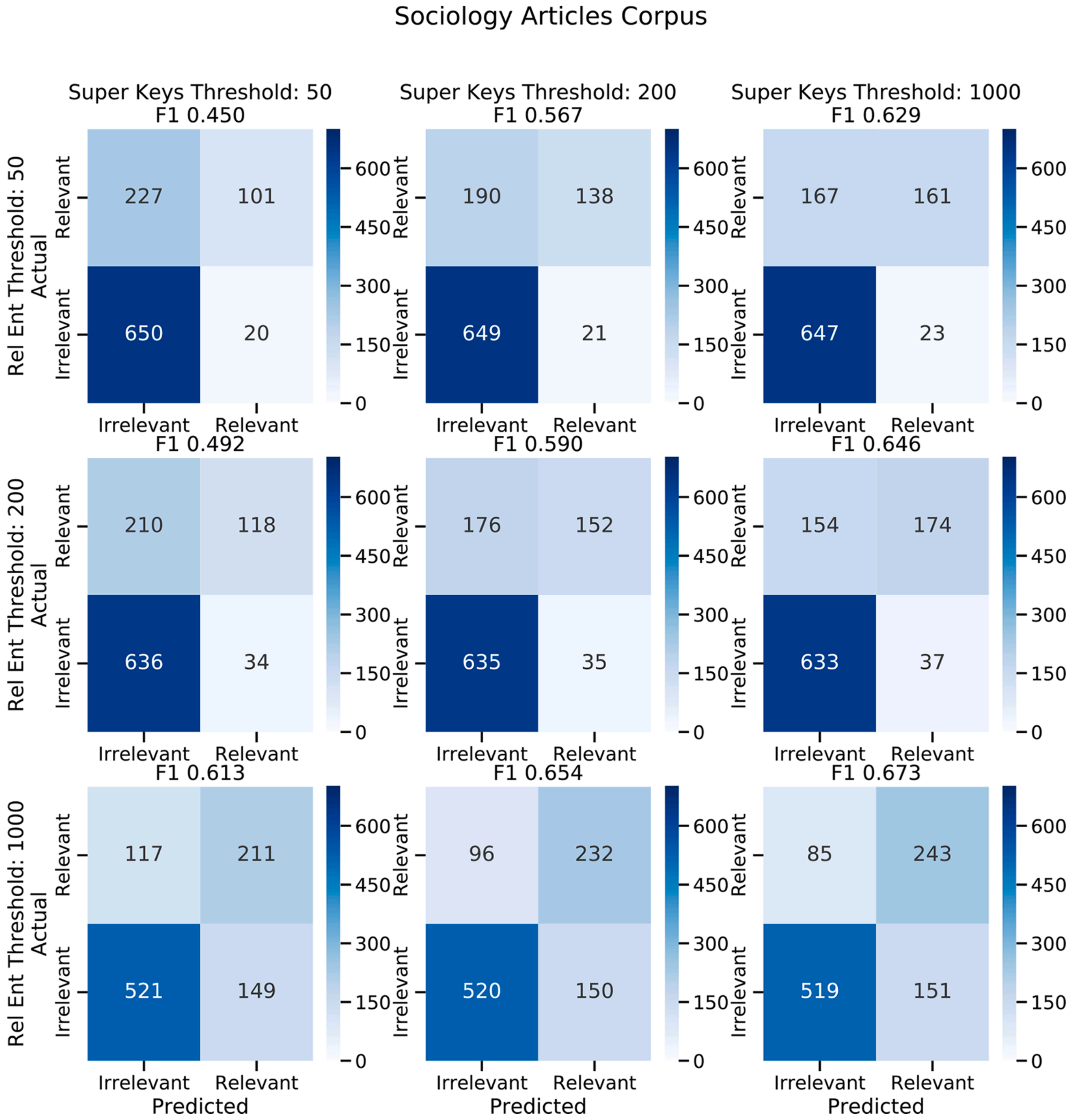

Fig. D.1. (above in blue), shows results for the sociology articles corpus with inequality as the research topic. These results show that increasing the threshold for inclusion in the refined topic (section 3.5) from 50, to 200, to 1000 improves recall but at the cost of precision. This also shows that expanding the list under consideration for super keywords (section 3.6) from the top 50 relative entropy terms to 1000 does not involve the same trade-off. Specifically, considering the right-most column, increasing the threshold for inclusion in the refined topic increases recall from 0.49, to 0.53, to 0.74, but decreases precision from 0.87, to 0.82, to 0.61. Considering the middle row, expanding the list under consideration for super keywords increases recall from 0.36, to 0.46, to 0.53, while precision also increases from 0.78, to 0.81, to 0.82.

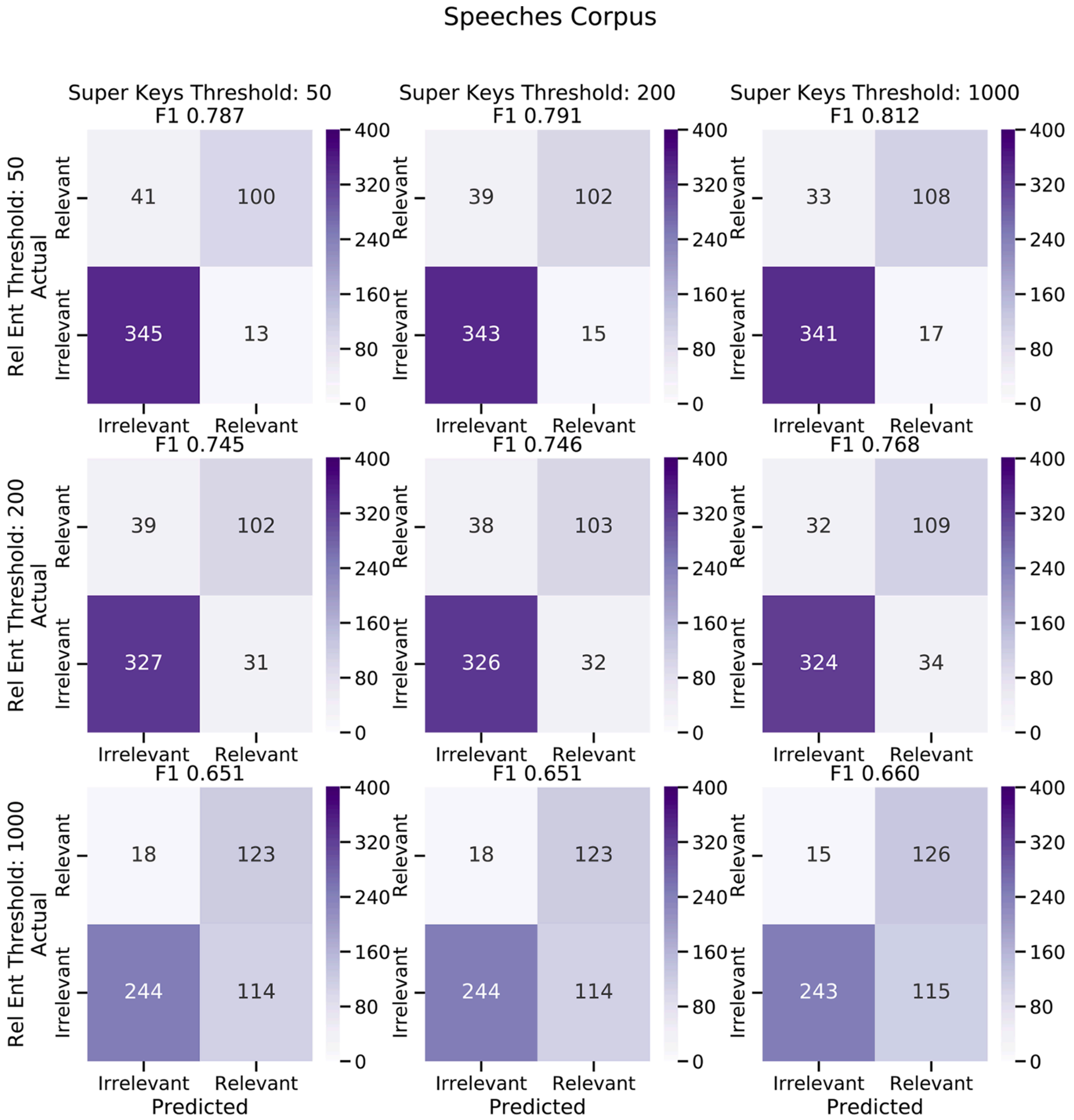

Fig. D.2. (above in purple), shows results for the government speeches corpus with housing markets and housing crisis policies as the research topic. The results for this corpus show similar patterns for both thresholds, i.e. increasing the threshold for inclusion in the refined topic has clear tradeoffs between recall and precision, while increasing the number of terms considered for inclusion in the super keys list does not. Specifically, considering the right-most column, increasing the threshold for inclusion in the refined topic increases recall from 0.77, to 0.77, to 0.89, but decreases precision from 0.86, to 0.76, to 52. Considering the middle row, expanding the list under consideration for super keywords increases recall from 0.72, to 0.73, to 0.83, while precision decreases only slightly from 0.77, to 0.76, to 0.76.

Appendix E. Keywords included in the dictionary-only approaches

Keywords included in the housing market and housing crisis policies dictionary: housing market*, hous* price*, residential starts, housing bubble, housing correction, residential default*, foreclosure*, mortgage*, subprime, fannie, freddie, gse*, housing couns*, fha, homeowner*, borrower*

Keywords included in the inequality dictionary:

List of keywords, retrieved from Nelson et al., 2018: concentration of income, concentration of wealth, distribution of income, distribution of incomes, distribution of wealth, distributional, economic disparities, economic disparity, economic distribution, economic divide, economic equality, economic inequalities, economic inequality, economic insecurity, egalitarian income, egalitarian wealth, employment insecurity, equal economic outcomes, equal income, equal pay, equal wage, equal wealth, equality of economic outcomes, equality of income, equality of incomes, equality of wealth, equalize income, equalize incomes, equalize wealth, equalizing income, equalizing incomes, equalizing wealth, equitable distribution, equitable income, equitable wealth, equity in income, equity in incomes, equity in wealth, equity of income, equity of incomes, equity of wealth, gini, income concentration, income decile, income difference, income differential, income disparities, income disparity, income distribution, income divide, income equality, income equalization, income gap, income inequalities, income inequality, income inequity, income polarization, income quintile, income redistribution, income stratification, inegalitarian income, inegalitarian wealth, inequality of economic outcomes, inequality of income, inequality of incomes, inequality of wealth, inequitable distribution, inequitable income, inequitable wealth, inequity of incomes, inequity of wealth, job insecurity, maldistribution, pay difference, pay differential, pay divide, pay equality, pay gap, pay inequality, redistribution of income, redistribution of incomes, redistribution of wealth, top incomes, unequal economic outcomes, unequal economy, unequal income, unequal incomes, unequal wealth, unequal pay, unequal wage, unequal wealth, uneven distribution, wage difference, wage differential, wage disparities, wage disparity, wage divide, wage equality, wage gap, wage inequalities, wage inequality, wealth concentration, wealth difference, wealth differential, wealth disparities, wealth disparity, wealth distribution, wealth divide, wealth equality, wealth equalization, wealth gap, wealth inequalities, wealth inequality, wealth inequity, wealth polarization, wealth redistribution, wealth stratification

List of keywords and instructions for dictionary retrieved from Nelson et al., 2018:

GROUP I TERMS: EXPLICIT DISTRIBUTIVE LANGUAGE – 1 AND (2 OR 3)

(1) Distribution inequality equality unequal distribution gap divide

(2) Income/wealth (private): economic wage income earning pay

(3) Income/wealth (govt): cash transfer non-cash transfer welfare food stamp unemployment insurance social security differential difference disparity polarization dualism dual society

AND ((compensation benefit wealth asset stock return

OR (Medicaid

Medicare housing assistance public housing earned income tax credit

EITC equity inequity inequitable egalitarian inegalitarian concentration bonus investment tax stock) social spending social program redistribution redistributive))

GROUP II: IMPLICIT DISTRIBUTIVE LANGUAGE – 1 AND (2 OR 3) (1) Social class groups (terms from at least two groups): top rich executive CEO affluent wealthy wealthier wealthiest professional white collar high income high wage high skill investor upper class employer manager

1% middle class blue collar middle income median wage median earner average wage

57 average earner poor worker minimum wage worker

(2) Income/wealth (private): economic wage income earning pay

(3) Income/wealth (govt): cash transfer non cash transfer welfare food stamp unemployment insurance social security

Notes: union low income lower class bottom low wage

AND ((compensation benefit wealth asset stock return

OR (Medicaid

Medicare housing assistance public housing earned income tax credit EITC

List of keywords added for the current analysis to incorporate racial/ethnic and gender inequality: racial | race| ethnic | gender | black and white | male and female | men and women

AND disparit* divide

equal

security gap* stratification

References

Algee-Hewitt, M., Heuser, R., & Moretti, F. (2015). *On paragraphs: Scale, themes, and narrative form*. Stanford Literary Lab.

Antoniak, Maria, & Mimno, David (2018). “Evaluating the Stability of Embedding-based word similarities. *Transactions of the Association for Computational Linguistics, 6*, 107–119.

Bail, Christopher A. (2012). “The fringe effect civil society organizations and the evolution of media discourse about Islam since the September 11th attacks. *American Sociological Review, 77*(6), 855–879.

Benoit, Kenneth, Laver, Michael, & Mikhaylov, Slava (2009). “Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. *American J Political Science, 53*(2), 495–513.

Bischof, Jonathan M., & Airoldi, Edoardo (2012). “Summarizing topical content with word frequency and exclusivity. *International Conference on Machine Learning*, 9–16.

Blei, David M., Ng, Andrew Y., & Jordan, Michael I. (2003). “Latent Dirichlet allocation.” JMLR 993-1022.

Boyd-Graber, Jordan, Mimno, David, & Newman, David (2014). “Care and feeding of topic models: problems, diagnostics, and improvements. *Handbook of Mixed Membership Models and their Applications 225255*.

Chang, Jonathan, Gerrish, Sean, Wang, Chong, Jordan, L., Boyd, Graber, & David M, Blei (2009). Reading Tea leaves: how humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296.

Chuang, Jason, Gupta, Sonal, Manning, Christopher, & Heer, Jeffrey (2013). “Topic model diagnostics: assessing domain relevance via topical alignment. *International Conference on Machine Learning*, 612–620.

Denny, Matthew, & Spirling, Arthur (2017). “Text Preprocessing for Unsupervised Learning: why it matters, when it misleads, and what to do about it. *Political Analysis, 26*(2), 168–189.

DiMaggio, Paul (2015). “Adapting computational text analysis to social science (and vice versa). *Big Data & Society, 2*(2), Article 2053951715602908.

DiMaggio, Paul, Nag, Manish, & Blei, David (2013). "Exploiting Affinities Between Topic Modeling and The Sociological Perspective on Culture: application to newspaper coverage of us government arts funding. *Poetics, 41*(6), 570–606.

Evans, James, & Aceves, Pedro (2016). "Machine Translation: mining Text for Social Theory. *Annual Review of Sociology, 42*, 18–30.

Golder, Scott A., & Macy, Michael W. (2011). "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science (New York, N.Y.), 333* (6051), 1878–1881.

Grimmer, Justin, & Stewart, Brandon (2013). . "Text as Data: the Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis,* 1–31.

Head, Tim et al. (2018). scikit-optimize/scikit-optimize: V0.5.2 (Version v0.5.2). Zenodo. 10.5281/zenodo.1207017.

King, Gary, Lam, Patrick, & Margaret, E. Roberts (2017). Computer-assisted keyword and document set discovery from unstructured text. *American J Political Science, 61*(4), 971–988.

Kiss, Tibor, & Strunk, Jan (2006). "Unsupervised multilingual sentence boundary detection. *Computational Linguistics, 32*(4), 485–525.

Lee, Moontae, & Mimno, David (2014). Low-dimensional embeddings for interpretable anchor-based topic inference. in proceedings of the 2014 conference on empirical Methods in Natural Language Processing. October 25−29 (p. 1319–1328).

Loper, Edward, & Bird, Steven (2002). "NLTK: The Natural Language Toolkit." arXiv preprint cs/0205028.

Manning, Christopher, Raghavan, Prabhakar, & Schütze, Hinrich (2008). *Introduction to information retrieval.* Cambridge University Press.

Marshall, Emily A. (2013). "Defining Population Problems: using Topic Models for Cross-National Comparison of Disciplinary Development. *Poetics, 41*(6), 701–724.

Martin, Andrew W., Rafail, Patrick, & McCarthy, John D. (2017). "What a Story? *Social Forces, 96*(2), 779–802.

McAuliffe, Jon D., & Blei, David M. (2008). "Supervised Topic Models. *Advances in neural information processing systems*, 121–128.

McFarland, Daniel, Ramage, Daniel, Chuang, Jason, Heer, Jeffrey, Christopher, D. Manning, & Daniel, Jurafsky (2013). Differentiating language usage through topic models. *Poetics, 41*(6), 607–625.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Greg, S. Corrado, & Jeff, Dean (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Mohr, John W., & Bogdanov, Petko (2013). Introduction—Topic models: what they are and why they matter. *Poetics, 41*(6), 545–569.

Mohr, John W., Wagner-Pacifici, Robin, Breiger, Ronald L., & Bogdanov, Petko (2013). "Graphing the Grammar of Motives in National Security Strategies: cultural Interpretation. *Automated Text Analysis and the Drama of Global Politics." Poetics, 41*(6), 670–700.

Nelson, Laura K. (2017). "Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, 1–40.

Nelson, Laura K., Burk, Derek, Knudsen, Marcel, & McCall, Leslie (2018). "The future of coding: a comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 1–36.

Schofield, Alexandra, & Mimno, David (2016). "Comparing Apples to Apple: the effect of stemming on topic models. *TACL, 4*, 286–300.

Slapin, Jonathan B., & Proksch, Sven-Oliver (2008). "A scaling model for estimating time-series party positions from texts. *American J Political Science, 52*(3), 705–722.

Snoek, Jasper, Larochelle, Hugo, & Ryan, P. Adams (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 2951–2959.

Wiedemann, Gregor (2016). *Text mining for qualitative data analysis in the social sciences, 1*. Springer Fachmedien Wiesbaden.

Alicia Eads is an Assistant Professor in the Department of Sociology and the center for Industrial Relations and Human Resources at the University of Toronto. Her research focuses on economic inequality. Recently, she has examined the policy response to the housing market collapse in the United States. In ongoing work, she examines the process of financialization and considers the promise of economic advancement as well as the potential for exploitation. One current project focuses on housing finance. She completed her Ph.D. in Sociology in 2017 at Cornell University.

Dr. Alexandra Schofield is an Assistant Professor of Computer Science at Harvey Mudd College. Her work in natural language processing and machine learning focuses on the practical aspects of using distributional semantic models for analysis of real-world datasets, with problems ranging from understanding the consequences of data pre-processing on model inference to enforcing text privacy for these models. She completed her Ph.D. in Computer Science in 2019 at Cornell University.

Fauna Mahootian earned their B.S. from Cornell University in spring of 2019, where they studied Information Science and Social Science. They are interested in applying behavioral insights to improve the outcomes created by social systems and institutions. They are currently applying for data science Masters programs and data analysis and software engineering jobs.

David Mimno is an associate professor in the department of Information Science at Cornell University. He holds a PhD from UMass Amherst and was previously the head programmer at the Perseus Project at Tufts and a researcher at Princeton University. His-work is supported by the Sloan foundation and the NSF.

Rens Wilderom is a PhD candidate in Cultural Sociology at the University of Amsterdam. His-research interest lies with change processes in a wide spectrum of fields, most importantly markets. Combining theory and methods from various subdisciplines, such as social movement studies, organization studies and cultural sociology, his PhD project focuses on the emergence and institutionalization of electronic/dance music in the US, UK, and the Netherlands between 1985 and 2005.