InCorporation: Assessing Corpus Selection for Social Science Applications

Sarah B. Bouchat
bouchat@northwestern.edu

November 2019

Not For Circulation or Citation

1 Introduction

What makes a corpus a corpus? Corpus selection undergirds any inferences drawn from natural language processing projects, yet current practice treats corpora as given and therefore does not account for potential uncertainty about whether the documents under consideration either encompass the universe of relevant documents or serve as a representative selection. While qualitative scholars have contended with the process of document selection and source evaluation rigorously, quantitative scholars leveraging increasingly available text-as-data sources lack inclusion and exclusion criteria for documents broadly, as well as frameworks for assessing error and uncertainty around whether text selection meets the theoretical needs and aims of a given project.

Text-as-data projects in the social sciences rarely account for measurement uncertainty or construct uncertainty arising from corpus selection. For example, topic modeling assumes a common data-generating process for topics among a set of texts, yet no method currently exists to validate that texts do share those commonalities ex ante.

1.1 Motivation

Corpus selection can encompass three separate problems. First is a potential detection problem: can we automatically distinguish between documents that principally belong to a corpus and those that do not? For example, a text processing task seeking to evaluate the collected writings of a single author may encounter unsigned or anonymously written documents, requiring an analysis of authorship to determine inclusion. Likewise, documents might reasonably be selected on the basis of other metadata (e.g., date, language) that are detectable but unknown ex ante.

The critical challenge in detection is whether the plausible confounding documents are suspected and/or have a known "type" prior to analysis. In authorship problems, anonymous or unsigned documents already raise suspicion and require validation. Other metadata, such as dates or language, are subject to error that may or may not be suspected. For example, social media posts may contain loan words, mixed words, or bilingual representations of the same text. Failure to detect these discrepancies certainly will impact quantitative text analysis, but need not imply that these posts do not "belong" in the corpus. A researcher with prior knowledge of the subject matter or texts could develop exclusion criteria (e.g., only French-language posts are retained), but absent that knowledge, evaluating the inclusion of documents in a corpus can only be completed with post hoc rationalization when anomalies are uncovered.

Second is a question of corpus cohesion: to what extent do documents within a proposed corpus tightly coalesce around a set of concepts and themes that relate to the task or research question? For example, shifts in concept measurement (e.g., "democracy" since 1800) or changepoints in discourse plausibly suggest the existence of multiple different latent "corpora" rather than a single tightly cohering corpus evidencing a shift in the underlying concept itself. Whether and when corpus cohesion poses a problem for analysis, or is itself the subject of study, depends largely on the research question.

In analyzing discourse across State of the Union addresses since the founding of the United States, for example, a researcher could seek to measure the prevalence of or relative focus around a concept such as federalism or war. At the same time, however, the structure and function of State of the Union addresses over time means that not only do word choice and emphasis evolve as political discourse changes over time, but the data generating process for what types of subjects or themes *could* emerge in those addresses is also a moving target. For a researcher intrinsically interested in the discourse of State of the Union addresses, the latter concern may not bind; for a researcher interested in presidential discourse or national political priorities over time who chooses State of the Union addresses as a primary source of evidence, however, the coherence or cohesion of documents in that corpus could have significant implications for substantive conclusions.

Finally, corpus selection entails a question of universality and representation: how can we determine whether we have selected all, or the correct, "representative" set of documents that pertain to our research question? King, Lam, and Roberts (2017) highlight this challenge with respect to keyword selection of documents. As they note, keyword selection and Boolean search risks arbitrary, inconsistent, and non-replicable results when implemented exclusively by humans. While a computer-assisted approach using their keyword algorithm performs better, relying on keyword selection and classification to identify documents for a corpus is not a universal method of corpus selection, and does not guarantee or measure the theoretical coherence of documents or their relevance to the research question. Their paper leaves unanswered the theoretical question of how to assess corpus selection absent keyword determinants or even among possible keyword sets: should all keywords be treated as equally relevant in determining document inclusion, and to what extent do keywords' presence accurately reflect the relationship of the documents to the research question, or the documents to each other?

All three of these problems suggest the need for evaluating "corpusness." Defining the theoretical coherence of a corpus can allow researchers to effectively weight results of text analysis to encompass uncertainty about the selection of documents into the corpus at the outset. This paper provides an initial method for distinguishing among potential sub-corpora within a selected possible "corpus," with a proposed extension to the selection of documents into possible corpora.

2 What is a Corpus?

Because the goal in natural language processing and natural language understanding is to learn language patterns and usage, the theoretical or subject-matter "coherence" of documents has less significance, providing few restrictions on what constitutes a corpus. de Marneffe and Potts (2014) define a cropus as "any collection of language data (Kilgarriff & Grefenstette, 2003)" which means "[they] leave open the origin of this data, its size, its basic units, and the nature of the data that it encodes, which could come in any medium. [They] even count as corpora things like dictionaries, specialized word lists (Dewey 1923; Zipf 1949; Wierzbicka 1987; Levin 1993; Hoeksema 1997; Michel et al. 2011), and aggregated linguistic events, but rather aim to encode the general features of the linguistic system." "More specialized definitions," they argue, "would only limit the kinds of questions one can address" (Marneffe and Potts 2017).

For applications that are not *only* focused on learning about linguistic constructs, however, content and data origins play a much more important role. Measurement and detection of concepts of theoretical interest for social scientists relies on assumptions about the data generating process for the text under examination. Still, social science research is subject to blind spots and error because no metric for the coherence of corpora or validation of their relationship to the research question at hand.

2.1 Defining Corpora for Social Science

Three broad bodies of literature converge to offer insights into how to appropriately construct corpora for social science research using text-as-data. The first set of literature spans several disciplines but is focused on the narrow investigation of the appropriate size of corpus for quantitative analysis. In part because of the interdisciplinary nature of this literature, these studies fail to address the content of texts in a corpus, or whether and when rules for inclusion or exclusion need apply. Bowman et al. (2015), for example, heralds the development of a much larger annotated corpus for natural language inference, where the underlying text data are derived from the Flickr30k corpus of photo captions. Vlachos and Riedel (2014) likewise analyzes the optimal construction of text datasets for automated fact-checking, again utilizing convenience data from two fact-checking websites, with an emphasis on pre-labeled and available data.

The second set of literature does address the content of texts in a corpus, but is primarily focused on technical concerns, such as training word embeddings. Spirling and Rodriguez (2019) is one such example, attempting to address the question of whether large, general corpora serve these technical aims better than smaller, more specialized corpora. Even in tackling this methodological issue, however, these studies rely on a false, or at best undetermined, distinction between "general" and "specific" corpora—no criteria are ever defined for either the size or cohesion necessary to qualify a corpus in either of these categories.

Finally, a vast set of substantive literature centers on mis- and disinformation or media in constrained political environments. These studies highlight unexamined assumptions in the previous sets of literature, that documents under consideration for inclusion in a corpus could be considered equivalently "informative" or should carry equal weight in determining relevant topics, keywords, or other textual metrics.

3 Method

This paper conceives of "corpus-ness" or corpus coherence as a latent feature of a collection of texts. To provide a measure of uncertainty for quantities in text-as-data analyses, an ideal measure of belonging to or coherence within a corpus will be represented probabilistically.

In particular, while the number or distribution of topics within a proposed corpus might provide one indication of coherence, these topics, and how and when they shift, are often the object of inquiry, so they cannot also be used to define the corpus itself.

3.1 Word Embeddings

In order to achieve these aims, this paper proposes a word embedding-based approach to evaluating corpora. Evaluating the similarity of words and their co-occurrence in this case proxies for the strength of the relationship between underlying concepts present in the text. This quantifies relationships between words with the assumption that words in a shared context (nearby each other in phrases and sentences) share common meanings (Antoniak and Mimno 2018), rather than using simple word counts or other distance metrics.

Bengio et al. (2003), Mikolov, Chen, et al. (2013), and Mikolov, Sutskever, et al. (2013) lay the groundwork for canonical word embedding models such as Continuous Bag of Words (CBOW), skip-gram, and word2vec. While CBOW predicts a target word from context, skip-gram models predict words likely to occur in the context of a target term. word2vec, in turn, learns word embeddings from text by maximizing the similarity between vectors of terms appearing in context. The canonical example of word2vec's capacity to represent semantic patterns in terms of linear relationships is via its capacity to generate "analog" results with vector math, such as: $Madrid - Spain + France = Paris$. By computing a dot product between the target and context words, and minimizing stochastic gradient descent, the algorithm generates expectations about the "distance" between co-occurring terms in vector space; the greater similarity between words (co-occurrence), the smaller their distance.

In standard word embedding models, the initial embedding layer (projection layer) takes a set of training words $w_1, w_2, ..., w_{\mathrm{T}}$ belonging to a vocabulary V and multiplies with a word embedding matrix. Using the embedding of words output from the (hidden) layer(s), the softmax layer ultimately calculates and normalizes the log probabilities of a given word w_{t} to generate a probability distribution over all the words in vocabulary V. Because. the gradient depends on all classes V, softmax can be computationally intensive and require an approximation or hierarchical approach. word2vec improves on this architecture in terms of efficiency and is able to generate the probabilities of words co-occurring. Both word2vec and GloVe offer an approach theoretically similar to counting word frequencies, via a matrix factorization infrastructure (Levy and Goldberg 2014). GloVe goes one step further to provide ratios of co-occurrences, which in turn facilitates evaluating vector distances/differences between words. The empirical component of this paper uses GloVe to take advantage of this property.

Implementing a word embedding approach, broadly speaking, involves:

1. Converting publications to plain text
2. Basic preprocessing (lowercase, remove punctuation, prune vocabulary)
3. Using GloVe
   (a) Tokenize words and create a vocabulary matrix
   (b) Create a term co-occurrence matrix
   (c) Use GloVe to factorize the matrix
4. Visualizing (e.g., with t-SNE)

This paper in particular uses a distributed word embeddings approach modeled on the analysis Rudkowsky et al. (2018). The authors estimate sentiment in Austrian parliamentary speeches using distributed word embeddings: they conduct a supervised sentiment analysis utilizing human-coded sentences. Although some words of interest may not appear in the training set, this approach still allows them to be classified in the test set because of their vectorized similarity to other words. Following Spirling and Rodriguez (2019), furthermore, this paper utilizes pre-trained embeddings rather than requiring a large amount of data to generate bespoke trained word embeddings for a single application.

3.2 Model

Assessing "corpus-ness" demands document-level embeddings that can serve as a measure of the closeness or relatedness between documents in a proposed corpus. This paper evaluates a series of possible applications that utilize labeled training data in order to evaluate how well the proposed method performs in identifying and separating documents that "truly" belong to a corpus, versus those that are more disparately related. Note, however, that "true belonging" in a corpus is a theoretical construct dependent on the research question and application.

For a collected set of documents in each of the following applications, after calculating document-level embeddings, the data are divided into training and test sets. The training set is used to train a support vector machine on document-level embeddings, and model accuracy is assessed on the remaining test data.

The document-level embeddings in this case calculate a mean vector for n documents, with k total words in the vocabulary, and q dimensions specified in the word embeddings. A document is therefore first converted to a document-term matrix $D_{n \times k}$, which is in turn used to calculate the average vector using word embeddings *W*:

$$A = \frac{D_{n \times k} \times W_{k \times q}}{\sum_{i=1}^{n} D_{i,\cdot}}$$

. This is then used to train the support vector machine in order to classify documents into underlying subcorpora.

More generally, the procedure outlined in this paper for identifying corpora among texts is as follows:

- Evaluate document-level embeddings
- Assess results in clusters and visualize
- Select central documents from clusters to construct plausible corpora, and weight additional documents by distance from these centroids

4 Data & Applications

The datasets used for validation of the method in this paper represent a breadth of potential applications across social and political science, but also importantly illustrate the robustness of the proposed method to potential corpus size.

4.1 Application: Yelp vs. Amazon Reviews

The first application evaluates Yelp and Amazon food reviews. While derived from different sources, these reviews provide a hard test for a word embedding approach to evaluating inclusion in a corpus in part because of their relative brevity. While the types of reviews one might encounter on each of these sites seems distinct, furthermore, the reviews themselves are often less distinguishable because of the details that reviewers choose to include. Consider, for example, this review from Yelp in the dataset:

> "After living near the east side Co-op for years, I was hesitant to migrate

> to the west side when I moved. I'm very happy with the west side location and now find it a better experience than the east side. There's more space in the store making it less hectic and packed. With more space the food bar is better. I'm a regular of the vegetarian hot items. They've now added all the tofu varieties in the deli from the east side location. That was huge for me as a dedicated vegetarian. And the deli staff is always friendly and helpful. Also with the west side location, parking is much better. It can be tight at times with the other businesses in the complex. But I can always find a spot."

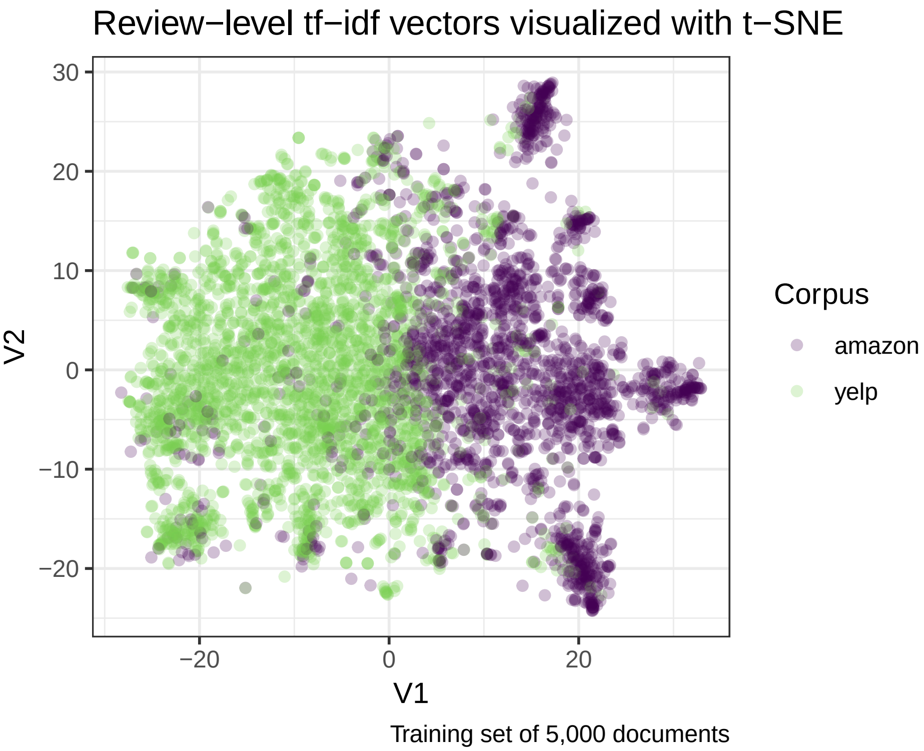
The review does highlight some factors (e.g., location, staff characteristics) that may not appear in an Amazon food review, but its discussion of tofu varieties is more analogous to the type of information likely to be left in an Amazon review of a particular food item.

Consider, for comparison, this Amazon food review from the dataset:

> "The title of this coffee blend pretty much says it all – it's BOLD. I prefer bold coffee, so I thoroughly enjoy this blend. I've tried other brands of so-called bold coffee, and it's surprising how many are wimpy wash-outs. Apparently, any brand that's popular enough to make it onto grocery shelves is blended down so as not to offend the average palate. So I don't waste my time buying supermarket coffee anymore. One more thing I've learned is that you can't get a good cup of coffee if you aren't willing to pay for it. It really does make a difference if a coffee is grown, harvested, and roasted properly. I used to suspect that was a lot of PR razzle-dazzle, but it turns out to be true. If you like good, rich, bold coffee, you'll like this Barista Prima blend."

Again, the emphasis on evaluating the food item (coffee) itself is potentially indicative of the text being an Amazon food review rather than a Yelp review, but mentions of supermarket and grocery store coffee for comparison could be confounding. In terms of evaluating these documents for inclusion in a common corpus, then, a naive approach plausibly would include both, under the assumption that both evaluate grocery offerings but one places relatively more emphasis on structural factors such as location, while the other focuses on food offerings.

From an initial random subset ($n = 50,000$) of a larger dataset ($n = 200,000$) of Yelp reviews combined with the 5-core dataset of reviews for grocery and gourmet food on Amazon ($n = 151,254$), this application utilizes a random selection of 5000 reviews for training. The t-SNE visualization of document-level embeddings for the Amazon and Yelp data reflect the distribution of documents within a compressed two-dimensional format, and demonstrates some degree of overlap in the subject matter across reviews: the Yelp reviews appear to "surround" the Amazon reviews, and the corpora do not have a significant degree of separability (points reflecting individual documents are interspersed).


Review−level tf−idf vectors visualized with t−SNE

Training set of 5,000 documents

In terms of the performance of the model in correctly classifying documents into their respective corpora on the basis of document-level embeddings, however, the predicted vs. true values demonstrate reasonably high performance, with 94-95% accuracy in both cases. This classification performance translates to (area under the curve = 0.986).

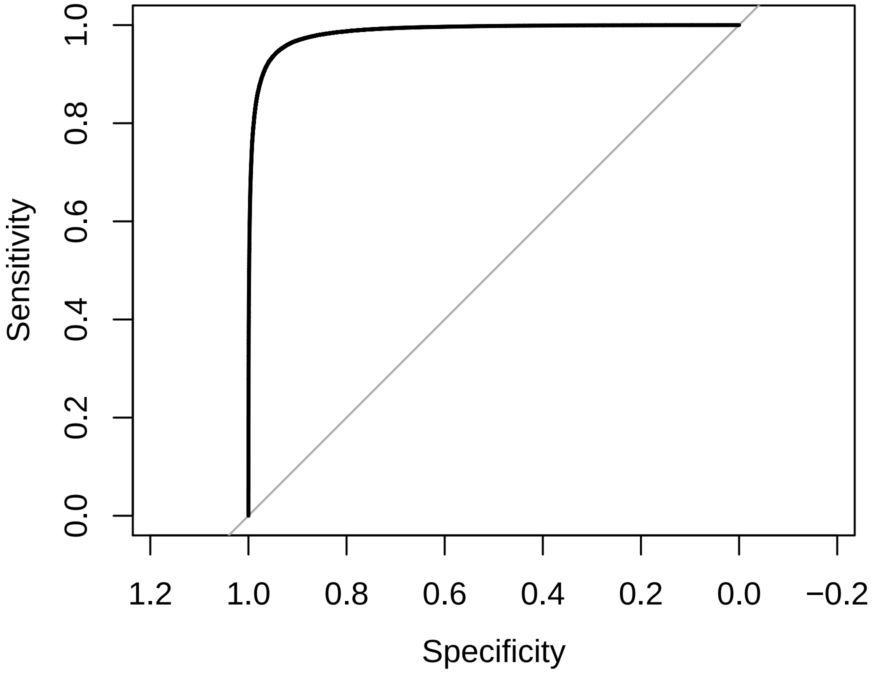
The predicted probabilities of inclusion in either corpus likewise demonstrate the relative success of the model in separating the two corpora, yet the decision value that facilitates this separability is less clear. Practically speaking, this closely aligns with

	Predicted Amazon	Predicted Yelp
True Amazon	139,695	9,373
True Yelp	10,112	187,035

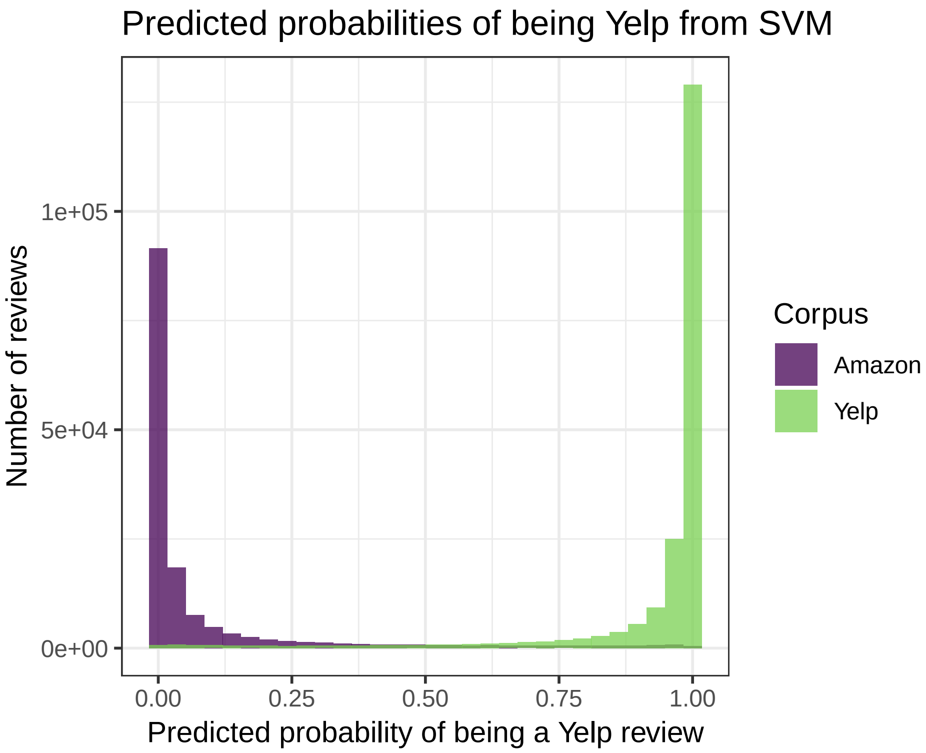
Table 1: Predicted vs. True Values: Number of Documents

	Predicted Amazon	Predicted Yelp
True Amazon	93.7%	6.3%
True Yelp	5.1%	94.9%

Table 2: Predicted vs. True Values: % of Documents



what a researcher is likely to experience in a case where, perhaps, two subcorpora are suspected but not known ex ante.

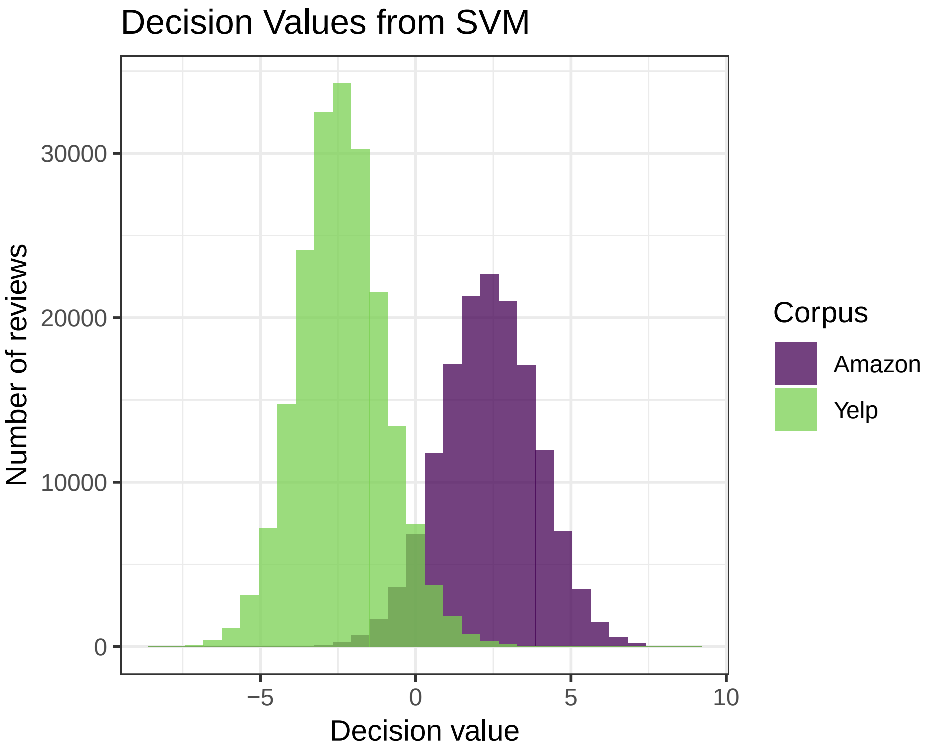


As suggested, the relative emphasis on food evaluations can serve as a confounding factor when predicting whether a document, based on its embeddings, belongs to the Yelp or the Amazon corpus. For example, these texts were predicted as Amazon reviews but were in fact Yelp reviews:

- "Good taste and reasonable price. The size is appropriate. Bubble milk tea is what I like, not too sweet."
- "I don't see what all the fuss is about, the mint chocolate chip is SO minty and not enough chocolate, the vanilla is watery and salty. Not sure why people like it so much!"

Conversely, these Amazon reviews were predicted to be Yelp reviews:

- "We had this tea at the local fish restaurant. Very flavorful tea."
- "Too weak for me. I rather have San Francico Fog Chaser and French Roast. Wolfgang Puck is good but pricey"



Some documents, furthermore, were so ambiguous as to effectively randomize their predicted inclusion in either corpus:

- "The noodles are perfect for making soup with 5 or 6 mixed vegetables and beef/-pork.chicken. Since being made of wheat they absolve the flavor of the broth." (Amazon)
- "Ice cream is good but it is in no way gelato. It's just slightly creamier North American ice cream." (Yelp)

This application illustrates that document-level embeddings provide an initial structure through which to identify potential subcorpora and assess the overall cohesion of a proposed corpus. In this case, with known categories, the slight overlap in predicted probabilities suggests the need for further development of an aggregation principle that accounts for the uncertainty about document inclusion and probabilistic inclusion of documents in a corpus (e.g., including all potential documents in the overlapping space in analyses of each subcorpus; excluding all overlapping documents; or developing a hard cut point for inclusion or exclusion within the overlapping set).

4.2 Application: Manifestos vs. Constitutions

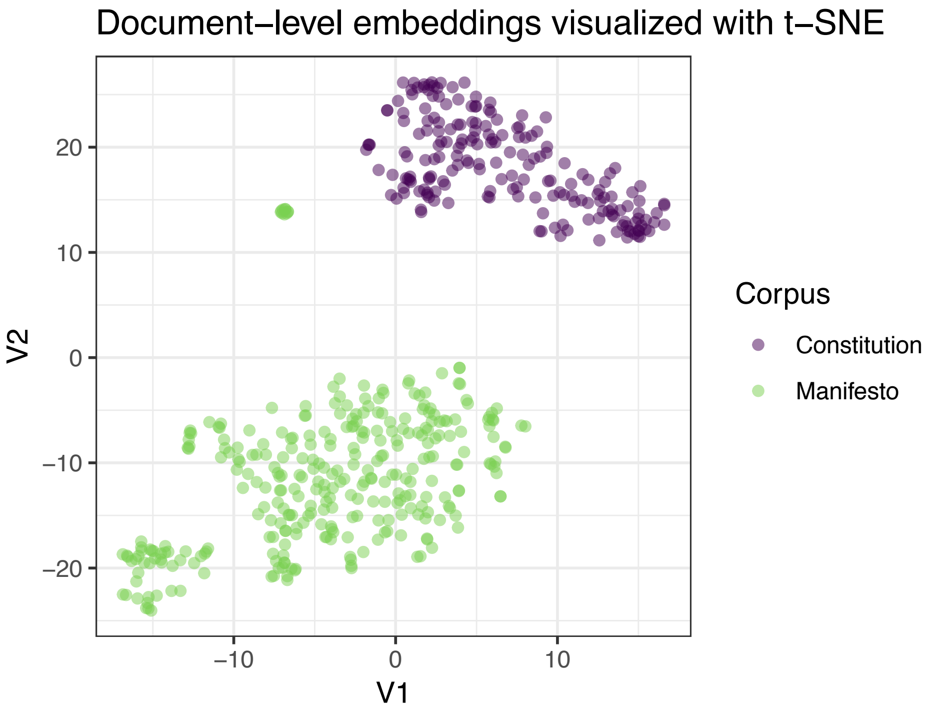
While the Yelp and Amazon reviews present a technical challenge given the relatively short document length, the size of the potential corpus/corpora offers relatively more latitude in selecting documents. More limited corpus sizes are typically binding constraints in social and political science applications, however. The second application, therefore, utilizes a much smaller set of documents from the Comparative Constitutions Project and Manifesto Project (via the R package `manifestoR`). Subsetting to English-language-available documents yields 193 constitutions and 324 manifestos (total of 507 documents with duplicates removed), of which 20% were used as the training set.

In addition to having fewer documents to compose a potential corpus, manifestos and constitutions pose a similarly difficult challenge to the Amazon and Yelp reviews for accurate validation. Both types of documents contain language describing political goals and rights, while manifestos frame these expressions in aspirational terms. For example, consider these passages from the manifesto of South Africa's ANC (2009) and the constitution of South Africa (1996) respectively:

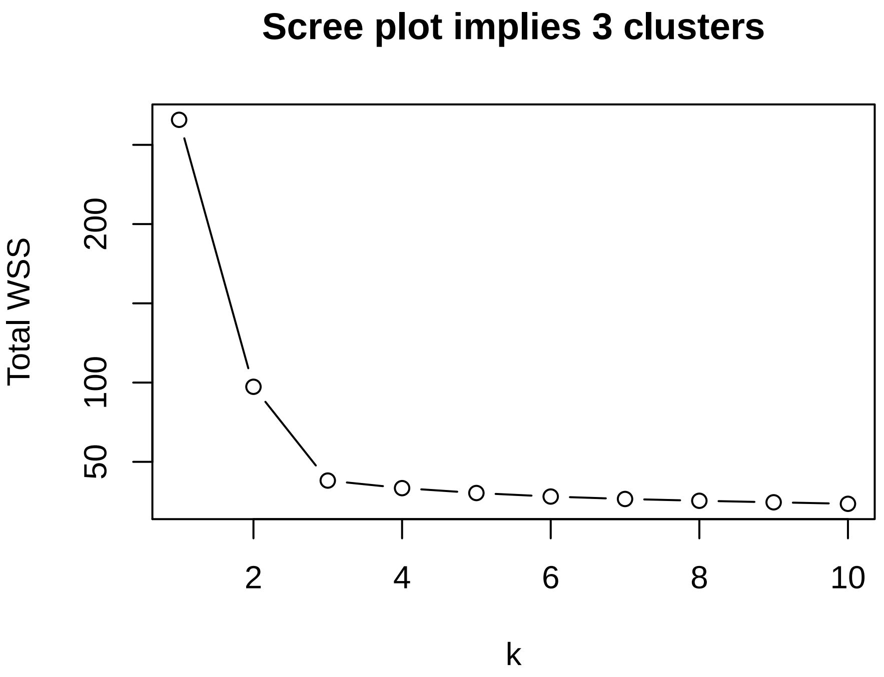
> "Our guiding principle is to live by the motto on our country's coat of arms. We aspire to the creation of a nation united in diversity. It is a goal to which we all aspire and it is the path to achieving our shared goal of a better life for all. Our constitution, inspired by the vision of the Freedom Charter unites a nation of many languages and significant cultural, religious and socio economic diversity. We have to work together to weave the threads that will see us celebrating a nation which is non racial, non-sexist and democratic - a nation that is dedicated to pushing back the frontiers of poverty." – ANC Manifesto
>
> "We, the people of South Africa, Recognise the injustices of our past; Honour those who suffered for justice and freedom in our land; Respect those who have worked to build and develop our country; and Believe that South Africa belongs to all who live in it, united in our diversity. We therefore, through our freely elected representatives, adopt this Constitution as the supreme law of the Republic so as to
>
> Heal the divisions of the past and establish a society based on democratic values, social justice and fundamental human rights;

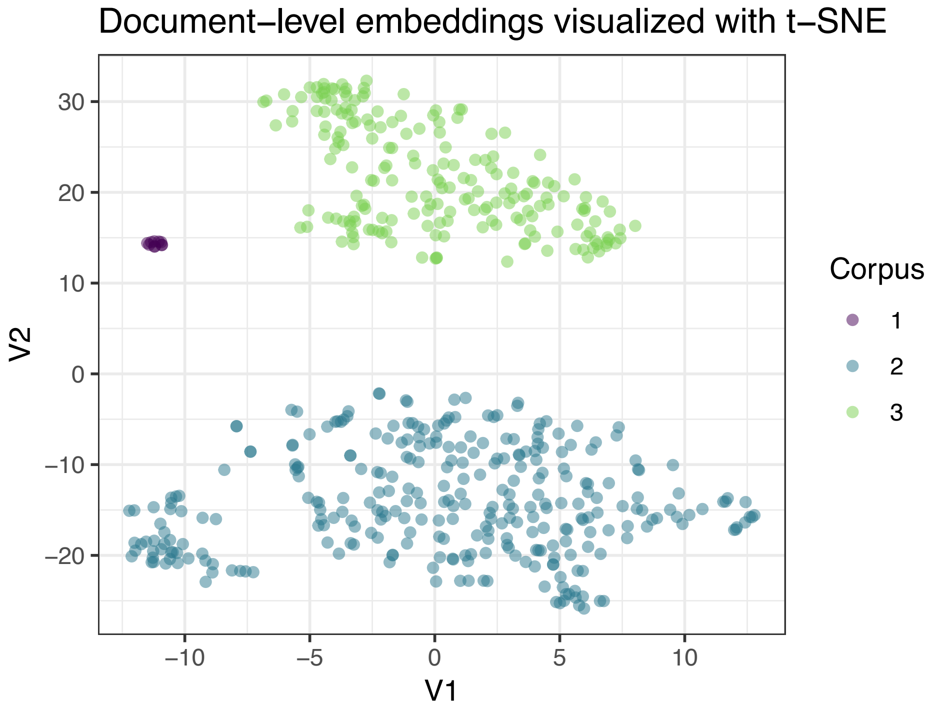
> Lay the foundations for a democratic and open society in which government is based on the will of the people and every citizen is equally protected by law;
>
> Improve the quality of life of all citizens and free the potential of each person; and
>
> Build a united and democratic South Africa able to take its rightful place as a sovereign state in the family of nations." – South Africa Constitution

The significant similarities in word choice and themes (justice, freedom, diversity, rights) suggest that separating or identifying potential subcorpora with the same SVM method as in the Amazon and Yelp review application should encounter analogous challenges. In reviewing the t-SNE visualization for this application, however, constitutions and manifestos appear as distinct clusters within the overall corpus.



In fact, the model perfectly predicts 100% of cases in the dataset in their respective subcorpora. The t-SNE plot does, however, include a small cluster of supposed "manifesto" documents much closer to the constitution documents than the larger manifesto cluster. Using k-means clustering with a scree plot to evaluate diminishing marginal returns on increasing cluster size in fact suggests that the lone "manifesto" entity is itself a distinct, third cluster.





The validation construction contained only two "true" subcorpora, but investigat-

Cluster	Text
1	C'est l'heure. Votez VERT.En 2008, prs d'un milli ...
1	NPD 2011 MON ENGAGEMENT ENVERS VOUSDU LEADERSHIP S ...
1	Table des matires (efface) et Message d'Igniatie ...
1	TABLE DES MATIERES (suite) page PARTAGER LA RICH ...
1	LETTRE OUVERTE DE LA PREMIRE MINISTRE KIM CAMPBEL ...
1	ICI POUR L'EMPLOI ET LA CROISSANCE NOTRE BUT Depui ...
1	Montral, le 20 octobre 2000 Chres amies, Chers ...
1	PLATE-FORME LECTORALE CAMPAGNE 2004. Un parti pr ...
1	Bloc qubcois 2011 INTRODUCTION A bien des gards ...
1	PLATEFORME LECTORALE BLOC QUBCOIS 2015 NATION Q ...

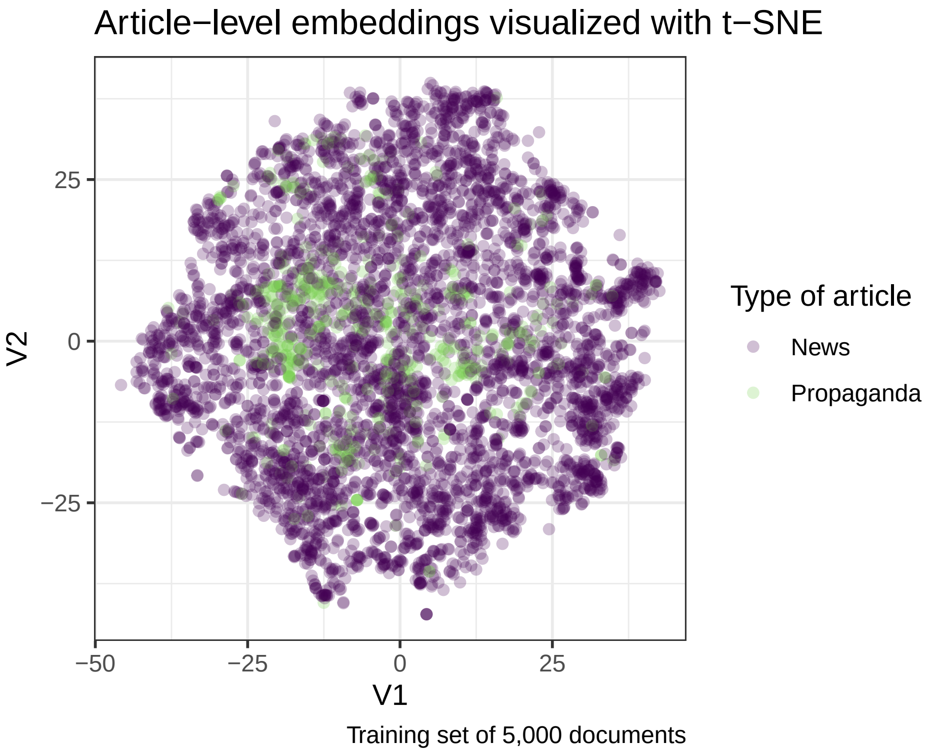
Table 3: Canadian French-language documents in their own cluster

ing this third suspected corpus reveals that in fact, the third subcorpus contains only French-language versions of Canadian documents. While this resulted inadvertantly from specifying English-language *available* (but not exclusive) constitutions and manifestos at the outset of the procedure, its presence usefully illustrates both the value of assessing corpus quality and coherence prior to conducting text-as-data analyses, and the specific utility of using document-level word embeddings to measure plausible corpus membership.

4.3 Application: News vs. Propaganda

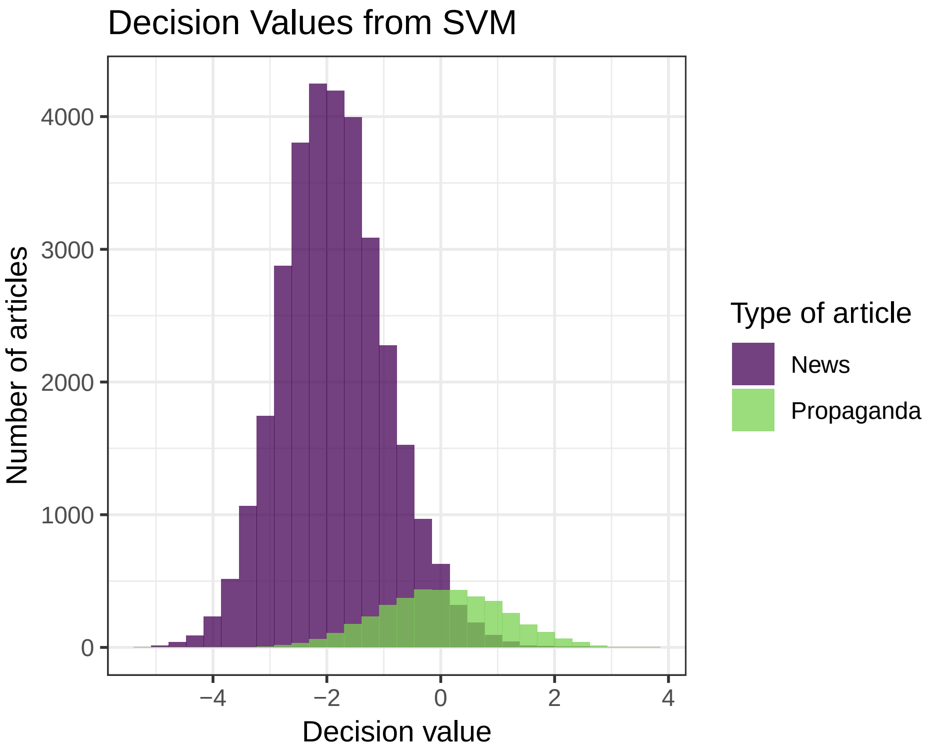
The final application utilizes the Proppy corpus, a 52,000-article corpus of news from over 100 outlets that is manually annotated to identify instances of propaganda (Martino et al. 2019). As in the prior applications, this corpus contains two "true" subcorpora for the purposes of validation, yet this set of texts raises broader theoretical questions about the method and measure of corpus-ness as well. In the Amazon and Yelp reviews application, a researcher would be unlikely to aim to combine both corpora for a research project; manifestos and constitutions could plausibly be combined, but are also unlikely to be ideal as a single corpus for the same reason. In both of these applications, the data generating processes for the texts are arguably distinct across corpora. Where news and propaganda are concerned, however, whether and when these types of text are distinct, or should be combined into a single corpus, is largely determined as a theoretical matter by the research question. Propaganda and "objective reporting" are often considered distinct entities in substantive analyses, but their data generating processes are not entirely independent. Both types of texts arise in response to, and are shaped by, political or social events; both types of texts possibly shape their narratives in response or opposition to the other. If a researcher were seeking to model the information environment in a particular country, for example, there are fewer grounds for excluding propaganda texts from a corpus for analysis. A research question seeking to assess *only* narratives or discourse in propaganda, on the other hand, runs risks of error from misclassification of documents into or out of the corpus in part because of the incomplete distinction between these types of texts.

This nonseparability of texts is reflected in the t-SNE plot attempting to classify based on a training set of 5000 documents.


Article−level embeddings visualized with t−SNE

Training set of 5,000 documents

Likewise, while some aspects of the dataset are more clearly propagandistic, the SVM largely predicts news and sees considerable overlap between the propaganda and news sets.

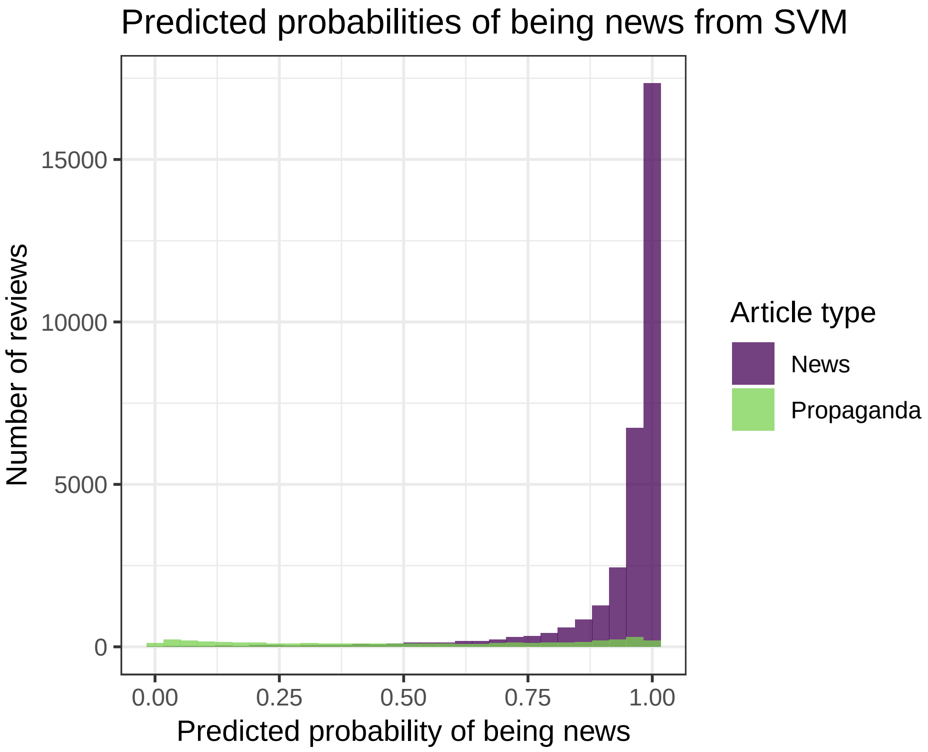


4.3.1 Implications for Text Analysis

Even with the size of the Proppy corpus, only ∼ 6% of documents are labeled as propaganda, meaning that any training set partition will learn relatively less about the markers of propaganda than it will about the markers of news. While this concern can be addressed in future extensions of the validation that oversample propaganda in the training set, the more pressing concern is whether the inconsistencies in inclusion in the corpus have the potential to impact substantive analyses or implications in practice.

In service of investigating this question, I estimate four separate sets of topic models using random sample of the Proppy corpus:

1. A topic model utilizing all documents



2. A topic model using only the predicted news documents, some of which are propaganda
3. A topic model using only the “true” news documents
4. A topic model using only the “true” propaganda documents

Future extensions will more rigorously select the appropriate number of topics for the data, but an initial analysis using 10 topics for investigation already indicates a potential concern with naively implementing text-as-data analyses without evaluating the quality or cohesion of the corpus. The topics in Table 4 reflect an analysis of the full dataset, which includes known propaganda, whereas the topics in Table 5 include only documents that were predicted to be news using the SVM procedure based on document-level word embeddings. Topic 5 in the full-data topic model (related to antisemitism) is notably absent from the topics using only predicted news articles. Whether this topic “should” be present in the analysis depends on the research question under

Topic	Metric	Words
1	Highest Prob. FREX	minister, pakistan, government, party, india, prime, chief punjab, islamabad, modi, nawaz, bjp, kashmir, lahore
2	Highest Prob. FREX	north, china, korea, south, president, trump, korean korea, korean, jong, pyongyang, seoul, korea's
3	Highest Prob. FREX	police, told, family, time, home, news, people markle, goodall, meghan's, burglar, deangelo, thurman, haig
4	Highest Prob. FREX	iran, israel, israeli, syria, deal, military, al syria, iranian, syrian, netanyahu, palestinians, gaza, iran
5	Highest Prob. FREX	people, world, american, political, anti, media, war holocaust, semitism, hitler, farrakhan, communism, jew, navalny
6	Highest Prob. FREX	million, company, business, percent, billion, government, companies earnings, shareholders, renewable, buffett, flipkart, blockchain, cryptocurrenc
7	Highest Prob. FREX	school, students, gun, people, health, children, schools parkland, stoneman, marjory, broward, nikolas, classrooms
8	Highest Prob. FREX	trump, president, house, campaign, white, republican, senate gop, giuliani, nunes, comey, dossier, haspel, steele
9	Highest Prob. FREX	city, water, people, time, day, food, island lava, volcano, volcanic, usgs, surf, leilani, paintings
10	Highest Prob. FREX	court, law, police, federal, judge, justice, charges oxfam, amp, plaintiffs, vekselberg, sereno, duterte's, zwaan

Table 4: Full dataset, highest probability or most frequent/exclusive words by topic

investigation, but at a more fundamental level, and as one would expect, the inclusion or exclusion of documents from the corpus has the potential to significantly impact the types of inferences researchers draw from their text-as-data analyses. Even in this setting, where propaganda comprises a very small proportion of the total amount of text, it still has the potential to interfere and change topic distributions.

5 Conclusion & Next Steps

As the topic models in the final application illustrate, there are potentially significant implications of corpora selection for substantive research. Researchers engaging with texts for possible inclusion in corpora must contend with at least three types of uncertainty,

Topic	Metric	Words
1	Highest Prob.	school, trump, gun, president, senate, house, republican
	FREX	parkland, nra, blankenship, stoneman, marjory, broward, daca
2	Highest Prob.	pakistan, court, minister, government, india, chief, police
	FREX	punjab, islamabad, bjp, nawaz, kashmir, karachi, lahore
3	Highest Prob.	north, china, korea, trump, south, president, korean
	FREX	korea, korean, jong, pyongyang, seoul, korea's
4	Highest Prob.	police, told, family, time, home, found, news
	FREX	markle, detectives, meghan's, thurman, deangelo, burglar, azy
5	Highest Prob.	iran, israel, israeli, syria, military, deal
	FREX	iranian, syrian, palestinian, netanyahu, palestinians, gaza, iran
6	Highest Prob.	million, company, business, billion, companies, percent, government
	FREX	shareholders, buffett, flipkart, blockchain, shareholder, notley, xerox
7	Highest Prob.	government, party, political, minister, people, eu, country
	FREX	brexit, corbyn, semitism, pashinyan, navalny, may's
8	Highest Prob.	trump, president, house, white, investigation, campaign, fbi
	FREX	memo, mueller, giuliani, nunes, dossier, comey, steele
9	Highest Prob.	people, school, students, women, health, university, children
	FREX	oxfam, ebola, ld, careersource, usf, vaccines, fgm
10	Highest Prob.	water, city, people, time, day, food, island
	FREX	volcano, paintings, kilauea, rockport, rainfall, cyclone, museums

Table 5: Predicted news only, highest probability or most fequent/exclusive words by topic

which the approach described in this paper seeks to address. The first is uncertainty as to whether the texts at hand communicate or contain the information sought according to the research question—are these texts relevant for the research question? Second is uncertainty about the data generating process(es) of the texts—is/are that process(es) shared across all the possible texts? The topic models for the propaganda application in part attempt to address this type of concern specifically, whereas the prior applications are more agnostic as to the data generating process for text. Third, even for documents that *do* share a common data generating process, how much latitude or adherence to the "center" of these types of documents should be allowed? How much variation among documents is tolerable while maintaining adherence to the research question?

For example, if a researcher wants to assess the media or information environment around minority rights in China, they would want to address the first type of uncertainty by identifying sources of media or news on that subject matter in China. The second type of uncertainty relates to the possibility of subcorpora that the method in this paper addresses. Some research projects might benefit from treating all media sources as equal and including as many as possible, whereas others might want to distinguish between, or exclude, state-owned and other media sources to account for bias. The third type of uncertainty, meanwhile, pertains to documents or sources that plausibly share the data generating process but where other error interferes. For example, perhaps some documents among the media sources are "informational" ad content, or are reprints from diasporic sources that do not comprise the media environment specified by the research project. Whether these types of text still represent the information of interest, but were merely not anticipated, rather than diverging from the intended research design, still depends on the project. Without prior knowledge of this type of interference, however, a researcher risks misattributing inferences to the supposed original corpus.

To more directly confront the third concern, additional work with the topic model application above should address the threshhold for propaganda prevalence in the dataset in order to provide a clearer sense of the tolerance of this word embedding method to interference by subcorpora. In general, however, the aspiration of this word embedding approach is to provide researchers with a set of tools to tackle each of these types of uncertainty.

The current paper does not develop a general metric of "corpusness" or a threshhold by which to evaluate whether a given document should be included in a corpus for analysis. Developing this type of metric requires an appropriate aggregation rule across plausible included documents (i.e., how closely related *should* we expect or require documents to be within a corpus, and is that threshhold general or specific to an application?). One plausible alternative to this strict metric is a weighting scheme across possible corpora with respect to the research question at hand: this type of measure would allow researchers to evaluate the robustness of their results in text-as-data analyses to the uncertainty around whether or which corpora were appropriate. Likewise, this approach lends itself to a more nuanced sensitivity test for the presence of multiple possible corpora.

In addition, the current method is conditional on already-collected documents. A more robust procedure for future work could specify a snowballing technique for identifying plausible documents for inclusion based on the parameters of the research project.

Bibliography

Antoniak, Maria, and David Mimno. 2018. "Evaluating the Stability of Embedding-based Word Similarities." *Transactions of the Association for Computational Linguistics* 6:107–119. eprint: https://doi.org/10.1162/tacl_a_00008. https://doi.org/10.1162/tacl_a_00008.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. "A Neural Probabilistic Language Model." *Journal of Machine Learning Research* 3:1137–1155.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. "A large annotated corpus for learning natural language inference." *CoRR* abs/1508.05326. arXiv: 1508.05326. http://arxiv.org/abs/1508.05326.

King, Gary, Patrick Lam, and Margaret E. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61 (4): 971–988.

Levy, Omer, and Yoav Goldberg. 2014. "Neural Word Embedding as Implicit Matrix Factorization." *Proceedings of the Neural Information Processing Systems (NIPS) Conference.*

Marneffe, Marie-Catherine de, and Christopher Potts. 2017. "Developing Linguistic Theories using Annotated Corpora." In *The Handbook of Linguistic Annotation,* edited by Nancy Ide and James Pustejovsky, 411–438. Berlin: Springer.

Martino, Giovanni Da San, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. *Fine-Grained Analysis of Propaganda in News Articles.* arXiv: 1910.02517 [cs.CL].

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint* 1301.3781.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality." *Proceedings of Neural Information Processing Systems (NIPS).*

Rudkowsky, Elena, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Stefan Emrich, and Michael Sedlmair. 2018. "More than Bags of Words: Sentiment Analysis with Word Embeddings." *Communication Methods and Measures* 12 (2-3): 140–157. eprint: https://doi.org/10.1080/19312458.2018.1455817. https://doi.org/10.1080/19312458.2018.1455817.

Spirling, Arthur, and Pedro L. Rodriguez. 2019. "Word Embeddings: What works, what doesn't, and how to tell the difference for applied research." *Working paper.*

Vlachos, Andreas, and Sebastian Riedel. 2014. "Fact Checking: Task definition and dataset construction." In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science,* 18–22. Baltimore, MD, USA: Association for Computational Linguistics, June. https://www.aclweb.org/anthology/W14-2508.