InCorporation

Assessing Corpus Selection for Social Science Applications

Sarah B. Bouchat

14 March 2019

Northwestern University

Introduction

Roadmap

- Defining corpora
  - Why we need a better definition of “corpus”
  - Existing definitions
- Method/Model
  - Embeddings
  - Examples: Food & Politics
  - Results
- Next steps

Motivation

- Most text analysis projects in social science fail to account for measurement uncertainty and construct uncertainty
- In particular, most results predicated on having selected a or the “right” set of documents/text
    - For example: topic models assume a common data generating process for topics among a set of texts, but no way to validate that the texts share that common DGP ex ante

Theory: What is a corpus, anyway?

Typical Definitions

- **Corpora** in NLP/NLU is merely any collection of text
    - Goal is to learn about language patterns and usage, so coherence of a set of texts around a topic/problem/idea/measurement construct is less important

NLP Corpus Definition

“. . . [We] say that a corpus is any collection of language data (Kilgarriff & Grefenstette, 2003). We 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. We 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 would only limit the kinds of questions one can address, which runs against our goals. . . .”

– Marie-Catherine de Marneffe & Christopher Potts, 2014

Corpora in Social Science

- For applications not *only* focused on learning about linguistic constructs, however, content and data origins play an important role
  - Measurement and detection of concepts relies on assumptions about the data generating process for the text we analyze
- Yet currently no metric for the coherence of our corpora or validation for their relationship to the research question/problem

Theoretical Distinctions

1. Detection: can we automatically distinguish between documents that principally belong to a corpus and those that do not?
   - Authorship problems
2. Cohesion: to what extent do documents within a proposed corpus tightly coalesce around a set of concepts or topics that relate to the task or research question?
   - Concepts shift over time
3. 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) keyword selection

Method

How do we measure corpus-ness?

- The “corpus-ness” or coherence of a corpus is a latent feature
  - To assess uncertainty, want a probabilistic measure of belonging to or cohesion within a corpus
- Topic distributions within a corpus may indicate its coherence, but often these topics and how/when they shift are exactly what we want to assess $\rightarrow$ cannot be used to define the corpus

Word Embeddings

Unlike simple word counts/distance metrics, which are subject to biases in small corpora and with varying document sizes (Antoniak & Mimno), *word embeddings* evaluate words in vector space, assuming that words occurring in common context (nearby) have similar meanings.

- Continuous Bag of Words: predicts target from context
- Skipgram: predicts context from target (good with large datasets)
- word2vec: enables semantic patterns to be represented linearly, e.g., “Madrid” - “Spain” + “France” → “Paris”
- GloVe: provides ratios of word co-occurrence to measure distance/difference

Implementing Word Embedding

1. Converting publications to plain text
2. Lowercase, remove punctuation, prune vocabulary
3. Use GloVe
    - Tokenize words and create a vocabulary matrix
    - Create a term co-occurrence matrix
    - GloVe factorizes the matrix
4. Visualize (e.g., with t-SNE)

Beyond Bag of Words: Distributed Word Embeddings

- Rudkowsky et al. (2018) estimate sentiment in Austrian parliamentary speeches using distributed word embeddings
  - Supervised sentiment analysis via human-coder-labeled sentences
- Advantages:
  - Even though some words may not appear in our training set, we can still classify them in the test set because of their vectorized similarity to other words
- Concerns:
  - Training word embeddings often requires huge amounts of data... what to do?

Spirling & Rodriguez 2019

- Punchline: pre-trained embeddings essentially work as well as or better than human coded data or locally-fit alternatives

Model

- Collect documents
- Calculate document-level embeddings
- Split off labeled training set (in this case, all documents are "labeled" with their corpus)
- Train support vector machine on document-level embeddings in training set
- Assess model accuracy on remaining test data

Document-level embeddings

Calculate mean vector for n documents, with k total words in vocabulary, and q dimensions in the word embeddings

- Convert document to document-term matrix $D_{n \times k}$
- Calculate average vector using word embeddings W:

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

Example: Food/Restaurant reviews on Yelp and Amazon

Data

A random subset of Yelp reviews in comparison to the 5-core dataset of reviews for grocery and gourmet food on Amazon

Corpus	n
Amazon	151,254
Yelp	200,000

5,000 total randomly selected for training set

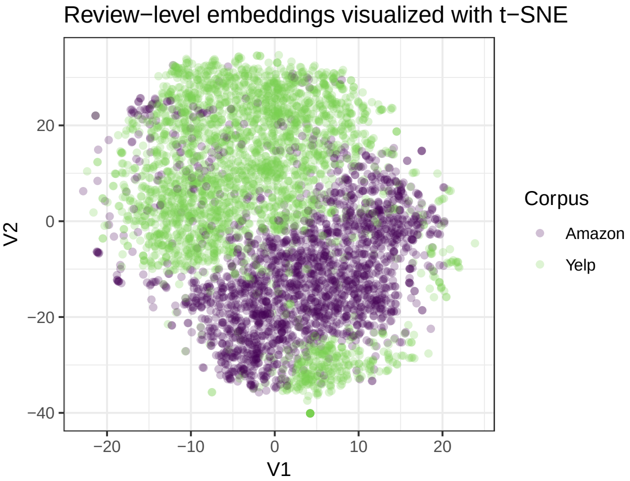
Examples of Yelp reviews

“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.”

Examples of Amazon reviews

“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.”

Distribution of word embeddings



Training set of 5,000 documents

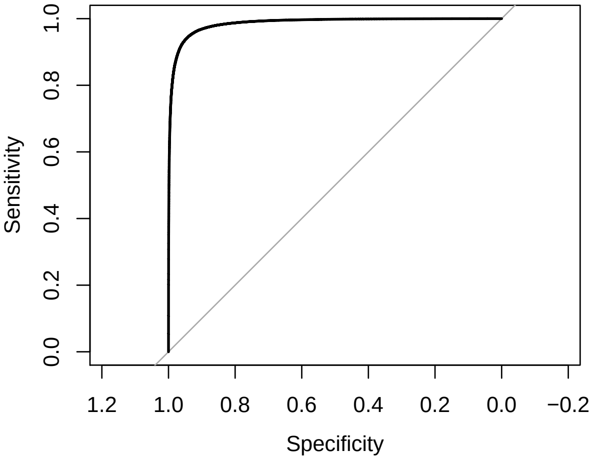
Performance on test set

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

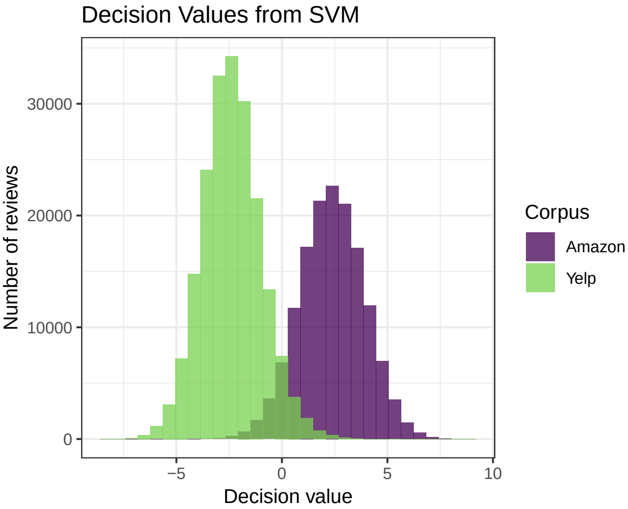
	Predicted Amazon	Predicted Yelp
True Amazon	93.7%	6.3%
True Yelp	5.1%	94.9%

Performance on test set

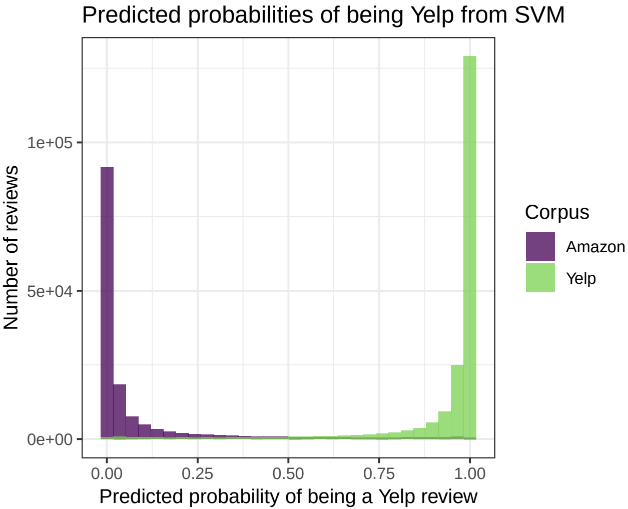
Area under curve = 0.986



Distribution of fitted values



Distribution of predicted probabilities



Borderline examples

Predicted Amazon, was Yelp:

- "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!"

Predicted Yelp, was Amazon:

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

In-between:

- "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."
- "Ice cream is good but it is in no way gelato. It's just slightly creamier North American ice cream."

Which is Amazon, which is Yelp?

Borderline examples

Predicted Amazon, was Yelp:

- “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!”

Predicted Yelp, was Amazon:

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

In-between:

- “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**

Manifestos and Constitutions

Manifesto and Constitution data

Data from the Comparative Constitutions Project and Manifesto Project via the `R` package `manifestoR` (subset to English-language only)

Corpus	n
Constitution	193
Manifesto	324

Manifesto Example: ANC, South Africa, 2009

“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.”

Constitution Example: South Africa, 1996

"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

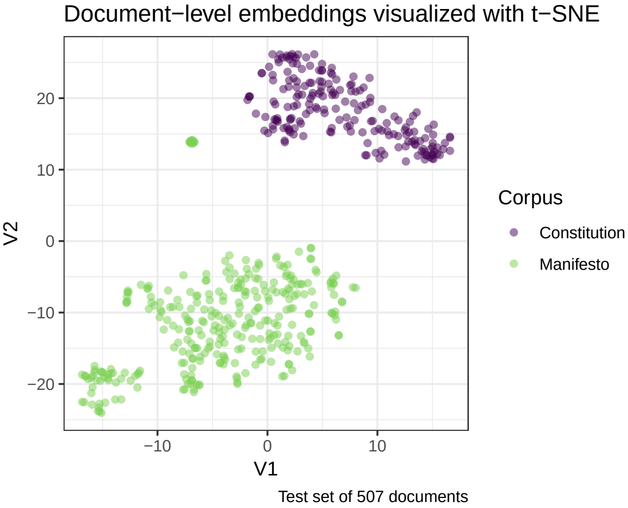
Heal the divisions of the past and establish a society based on democratic values, social justice and fundamental human rights;

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."

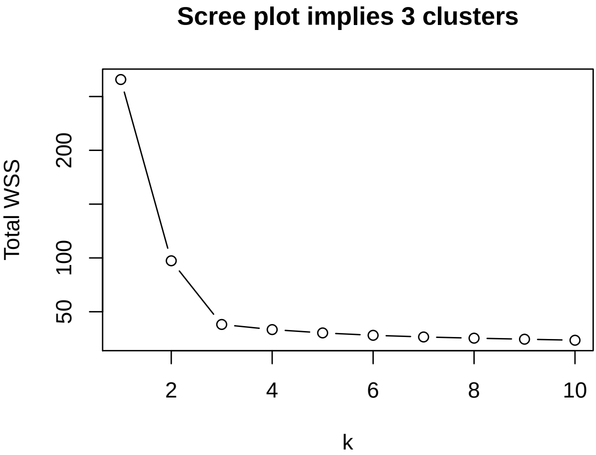
Distribution of word embeddings



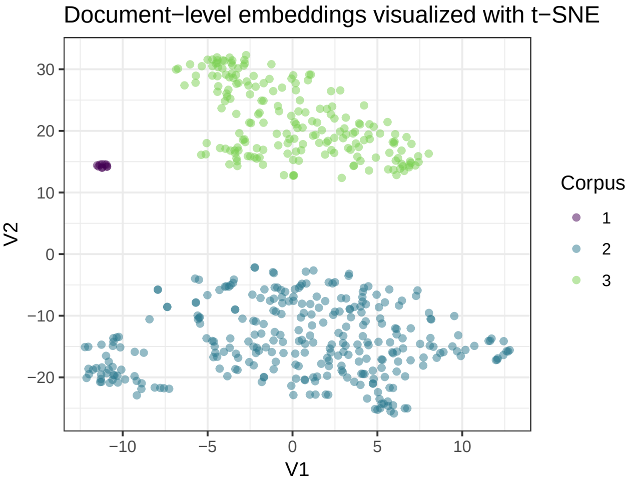
Performance on test set

	Predicted Constitution	Predicted Manifesto
True Constitution	156	0
True Manifesto	0	258

K-means clustering to identify within-corpus variation



Within-corpus variation



Test set of 507 documents

Blame Canada. . .

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 . . .

Questions & Directions

Next Steps

- Metric for individual document inclusion across potential corpora
- Overall metric of "corpusness"
    - Appropriate aggregation across plausible included documents
    - Weighting of possible corpora when assessing research questions
    - Use this weighting to assess existing papers/projects from political and social science → subsets of corpora
- Current setup is a metric for exclusion/classification conditional on collected documents, need a system for identifying plausible documents for inclusion
- Word embeddings
    - As the size of the corpus/corpora increase, the meaning of particular words converges toward their general usage: problematic?