Computational Social Science: Methods and Applications

Anjalie Field, anjalief@cs.cmu.edu

Overview

- Defining computational social science
  - Sample problems

- Common Methodology
  - Time series analysis
  - Classification
  - Topic Modeling (LDA)
  - Word Embeddings

Definitions and Examples

What is Computational Social Science?

"The study of social phenomena using digitized information and computational and statistical methods" [Wallach 2018]

Social Science

- When and why do senators deviate from party ideologies?
- Analyze the impact of gender and race on the U.S. hiring system
- Examine to what extent recommendations affect shopping patterns vs. other factors

Explanation

Traditional NLP

- How many senators will vote for a proposed bill?
- Predict which candidates will be hired based on their resumes
- Recommend related products to Amazon shoppers

Prediction

[Wallach 2018]

“Manipulative tactics are the norm in political emails: Evidence from 100K emails from the 2020 U.S. election cycle” [Mathur et al., 2020 Working Paper]

- Assembled corpus of >250,000 political emails from >3,000 political campaigns and organizations sent during the 2020 U.S. election cycle

- Potential for political manipulation, e.g. through micro-targeting, has drawn a lot of attention, but little work has focused on email (or on U.S. campaigns)

We can't hang our heads Inbox ×

Chair Jaime Harrison <info@contact.joebiden.com>
to me

11:11 AM (5 hours ago)

BIDEN

While our programs helped turn out hundreds of thousands of voters -- last night was proof that the GOP is going to fight tooth and nail to win back every seat we've got. We have to be ready. So, I am asking for this team to step up to make a difference as we jump right back into laying the early groundwork for next year.

Rush a $25 donation to the DNC's Democratic Fight Fund today to set Democrats up for success in all 50 states, the territories, and D.C. next year →

If you've saved payment information with ActBlue Express, your donation will go through immediately:

$25 >>	$50 >>
$75 >>	$100 >>
$125 >>	Other >>

Data Collection

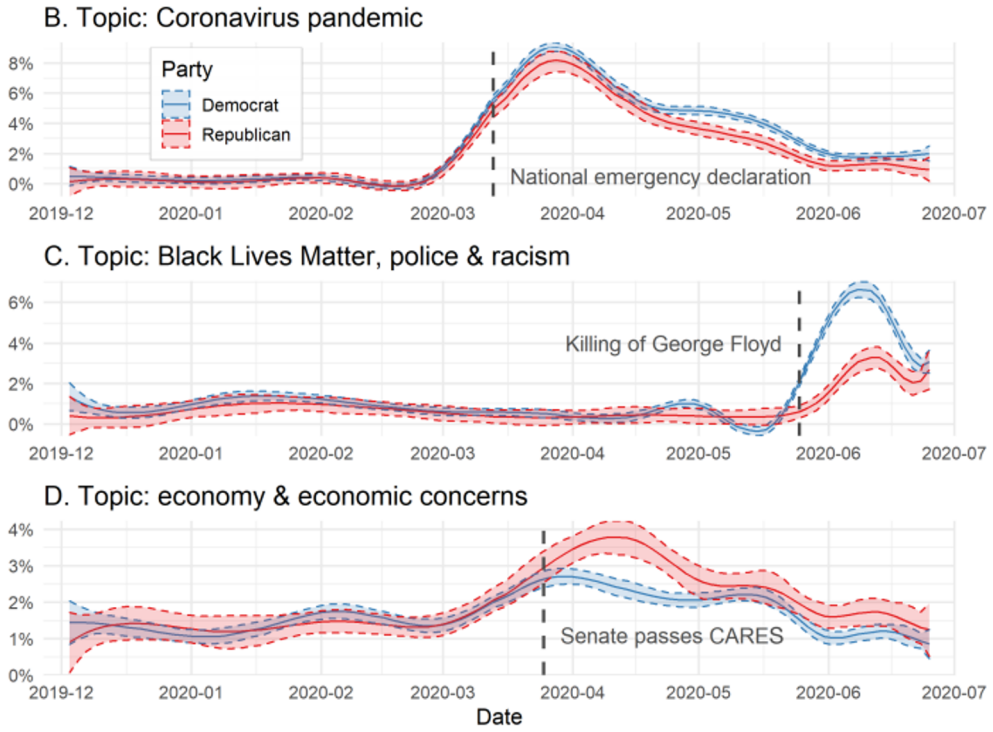
- Gather websites for funding agencies and candidates in U.S. 2020 elections (state and federal elections)
- Build bot to sign up for emails from each website
  - Gender-neutral sign up name
  - Distinct email address for each website
- On receiving emails:
  - Bot opens each message exactly once
  - Clicks on the confirmation link, if one is present
  - Downloads all resources (including tracking cookies) and takes screenshot

Research Questions

- What topics are discussed in emails? How do they vary by party affiliation?
- How do senders overcome “fundraising fatigue”?
- [What strategies are used to encourage recipient to open emails?]
- [Examine privacy violations: sharing email addresses across campaigns]

What topics are discussed?

- Methodology: Structured Topic model



How do senders overcome fundraising fatigue?

- Methodology:
  - Hand-code examples
  - Verify trends at a larger scale using more automated methods (e.g. building supervised classifier from hand-annotated samples)
- Selected Findings
  - Subjects often don’t relate to content of email
  - Falsely promise donation matching (but this is impossible since FEC has limits on how much an individual can donate to a campaign)
  - References to imminent fundraising deadlines

How do senders overcome fundraising fatigue?

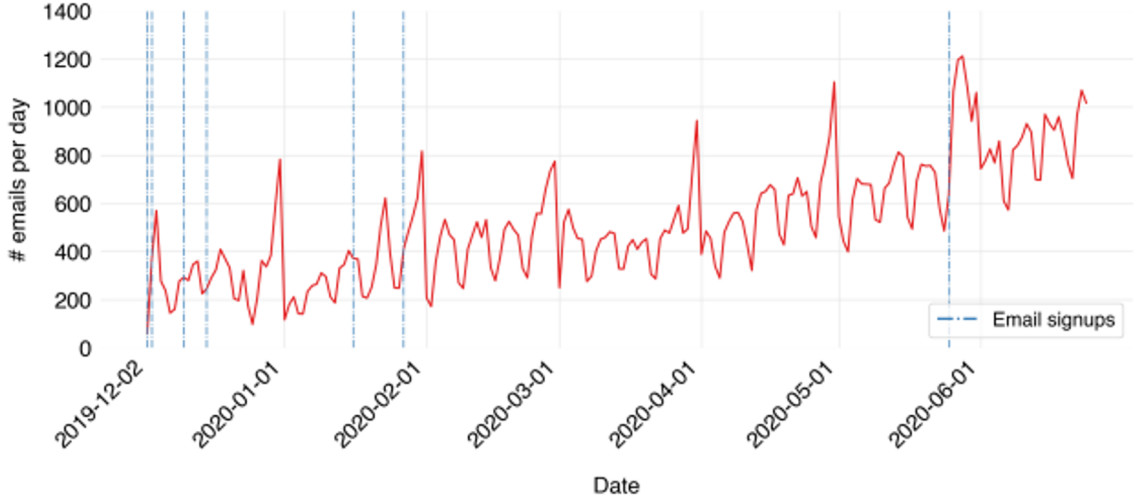


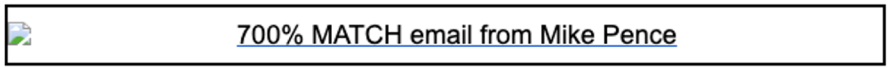
Figure 1: Distribution of the volume of emails over time.

I'll be honest with you. We could send you an obviously fake *700%-MATCH* email like the one from Mike Pence below and fundraise that way. But that's not who we are. **Instead, I want to explain why I hope you'll dig deep and commit to a recurring donation today through Election Day.**

Each and every recurring donation to this campaign represents not just voters we are able to reach, but also a strategic advantage -- knowing that we have recurring donations we can count on allows our advertising and organizing teams to plan out their budgets and targets much more precisely than they would be able to otherwise.

That's why I'm asking you directly: Can you commit to a recurring donation for the last seven weeks of this campaign?

You'll *match yourself* each week until Election Day and help power us to victory. And I promise we will never send you an email like this:


700% MATCH email from Mike Pence

$25 →	$50 →
$75 →	$100 →
$125 →	Other →

Social Science

- Defining the research question is half the battle
- Data can be messy and unstructured
- Careful experimental setup means controlling confounds -- make sure you are measure the correct value
- Prioritize interpretability (plurality of methods)

Traditional NLP

- Well-defined tasks
- Often using well-constructed data sets
- Careful experimental setup means constructing a good test set -- usually sufficient to get good results on the test set
- Prioritize high performing models

Methodology

Four principles of quantitative text analysis [Grimmer & Stewart, 2013]

1. All quantitative models of language are wrong—but some are useful
2. Quantitative methods for text amplify resources and augment humans
3. There is no globally best method for automated text analysis
4. Validate, Validate, Validate.

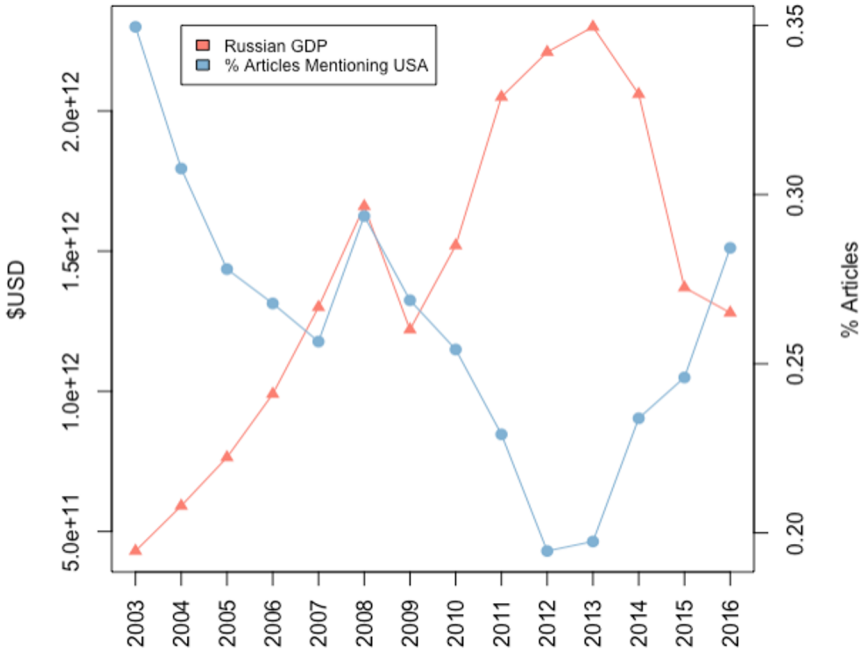
An incomplete sample of common methodology

- Time series / frequency analysis
- Classification
  - Hand-coding + supervised methods
  - Dictionary Methods
- Clustering (when classes are unknown)
  - Single-membership (ex. K-means)
  - Mixed membership models (ex. LDA)
- Word Embeddings

Agenda Setting in Russian News Articles

- Data set: choose a corpus where we expect to see manipulation strategies
  - 100,000+ articles from Russian newspaper *Izvestia* (2003 - 2016)
  - Known to be heavily influenced by Russian government
- Can hypothesize that we will see more manipulation strategies during when the country is “doing poorly”
  - Government wants to distract public or deflect blame
- [Objective] measure of “doing poorly”
  - State of the economy (GDP and stock market)

Benchmark against economic indicators



State of the economy is **negatively correlated** with the amount of news focused on the U.S.

	Article	Word
RTSI (Monthly, rubles)	-0.54	-0.52
GDP (Quarterly, USD)	-0.69	-0.65
GDP (Yearly, USD)	-0.83	-0.79

Granger Causality

$$C(w_t) = \sum_{i=1}^{m} \alpha_i (C(w_{t-i})) + \sum_{j=1}^{n} \beta_j (C(r_{t-j}))$$

- Use last month's economic state to predict this month's amount of U.S. news coverage
- Can show correlations are directed: first economy crashes, then U.S. news coverage increases

w_t frequency of U.S. mentions
r_t economic indicators
α, **β** coefficients learned by regression model

Granger Causality

$$C(w_t) = \sum_{i=1}^{m} \alpha_i (C(w_{t-i})) + \sum_{j=1}^{n} \beta_j (C(r_{t-j}))$$

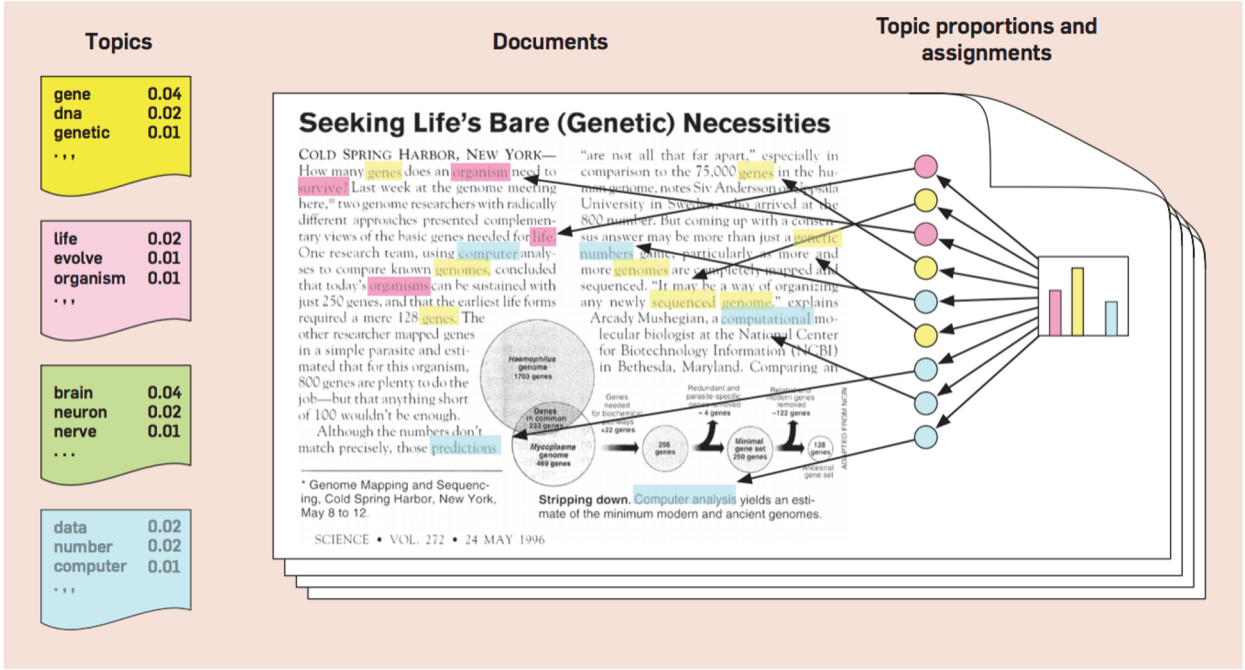
	α;β	**p-value**
w_{t-1}	-0.320	0.00005
w_{t-2}	-0.301	0.0001
r_{t-1}	-0.369	0.024
r_{t-2}	-0.122	0.458

w_t frequency of U.S. mentions
r_t economic indicators
α, **β** coefficients learned by regression model

Challenges in Classification

- What are *Izvestia* articles saying about the U.S.?
  - Hand-code articles according to how they portray U.S., Russia, and other countries
  - Train a classifier to predict portrayals in uncoded articles
- Problems:
  - Annotators need to be fluent in Russian
  - Annotators need to read full-length documents
  - Annotation scheme is potentially subjective and complex
  - Work has the potential to be critical of the Russian government
- What we did instead:
  - Use pre-annotated English data annotated for *media frames* and project them into Russian

Topic Modeling: Latent Dirichlet Allocation (LDA)



- Assume each document contains a mixture of "topics"
- Each topic uses mixtures of vocabulary words
- Goal: recover topic and vocabulary distributions

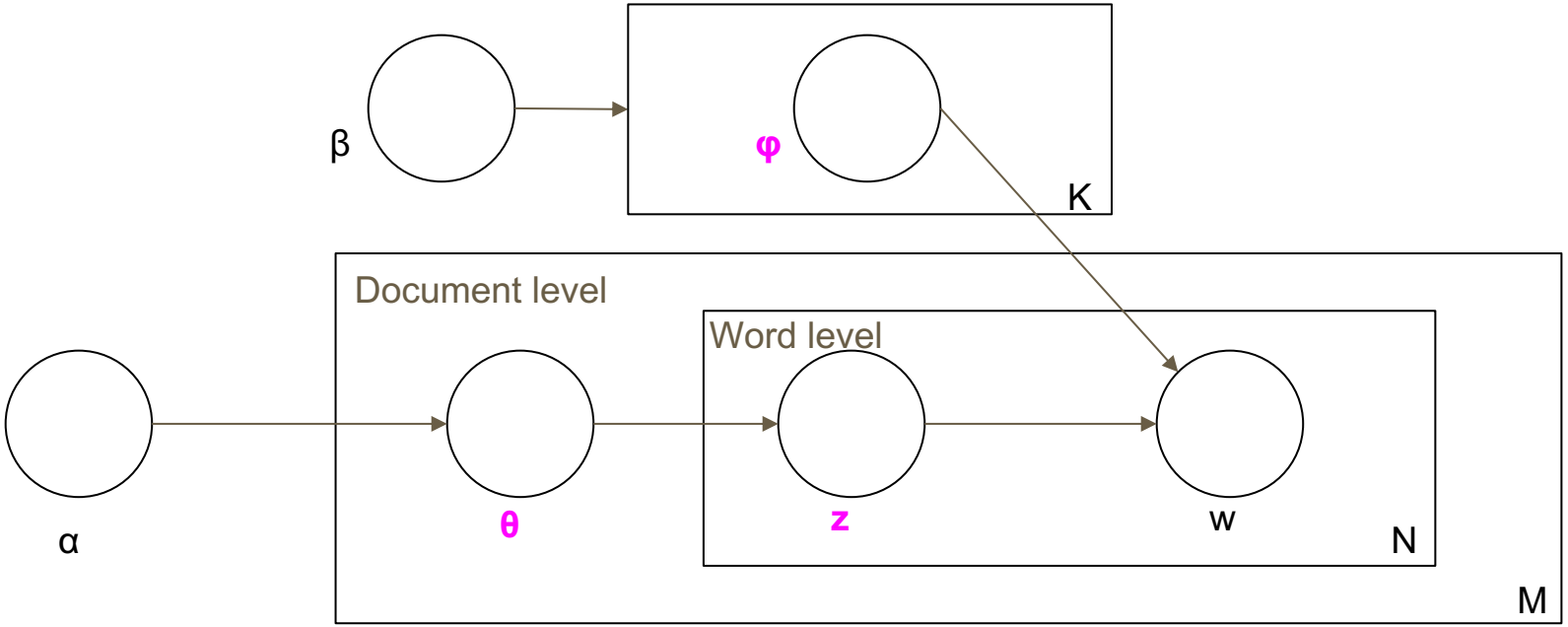
LDA: Generative Story

- For each topic k:
  - Draw $\varphi_k \sim Dir(\beta)$
- For each document D:
  - Draw $\theta_D \sim Dir(\alpha)$
  - For each word in D:
    - Draw topic assignment $z \sim Multinomial(\theta_D)$
    - Draw $w \sim Multinomial(\varphi_z)$

φ is a distribution over your vocabulary (1 for each topic)

θ is a distribution over topics (1 for each document)

Clustering



θ, φ, z are latent variables
α, β are hyperparameters
K = number of topics; M = number of documents; N = number of words per document

Sample Topics from NYT Corpus

#5	#6	#7	#8	#9	#10
10	0	he	court	had	sunday
30	tax	his	law	quarter	saturday
11	year	mr	case	points	friday
12	reports	said	federal	first	van
15	million	him	judge	second	weekend
13	credit	who	mr	year	gallery
14	taxes	had	lawyer	were	iowa
20	income	has	commission	last	duke
sept	included	when	legal	third	fair
16	500	not	lawyers	won	show

LDA: Evaluation

- Held out likelihood
  - Hold out some subset of your corpus
  - Says NOTHING about coherence of topics
- Intruder Detection Tasks [Chang et al. 2009]
  - Give annotators 5 words that are probable under topic A and 1 word that is probable under topic B
  - If topics are coherent, annotators should easily be able to identify the intruder
- Performance on downstream task
  - E.g. document clustering

LDA: Advantages and Drawbacks

- When to use it
  - Initial investigation into unknown corpus
  - Concise description of corpus (dimensionality reduction)
  - [Features in downstream task]
- Limitations
  - Can’t apply to specific questions (completely unsupervised)
  - Simplified word representations
    - BOW model
    - Can’t take advantage of similar words (i.e. distributed representations)
  - Strict assumptions
    - Independence assumptions
    - Topic proportions are drawn from the same distribution for all documents

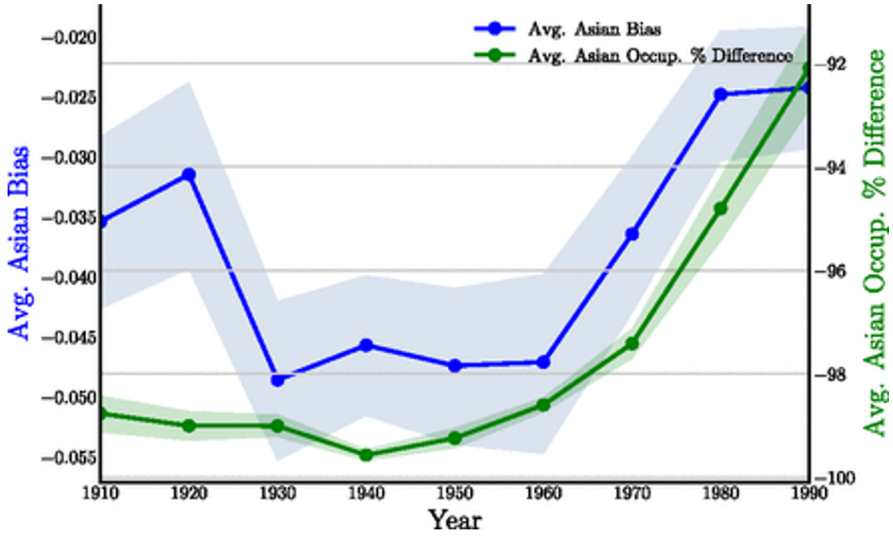
Word Embeddings

- “Man is to computer programmer as woman is to homemaker”
- NLP perspective
  - Seems bad if our models learn gendered associations with occupations
- Social science perspective
  - We can learn social stereotypes from the data

“Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change” [Hamilton et al. 2016]

- Methodology:
  - Construct word embeddings for each time segment of a large corpus and align them across time
  - (Use Word2Vec, but also statistical methods like SVD)
- Evaluation:
  - Examine how well word embeddings capture known shifts in word meanings over time
    - e.g. “gay” moves away from “happy, showy” and toward “homosexual, lesbian”

“Word embeddings quantify 100 years of gender and ethnic stereotypes” [Garg et al. 2018]



1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

- Next: what similar analyses do pre-trained languages models enable?

Summary

- Aspects of social science questions
  - Hard-to-define research questions
  - Messy data
  - “Explainability”
  - Ethics
- Methodology
  - Time series/frequency analysis
  - Classification
  - Clustering
  - Word Embeddings

Why Computational Social Science?

“Despite all the hype, machine learning is not a be-all and end-all solution. We still need social scientists if we are going to use machine learning to study social phenomena in a responsible and ethical manner.” [Wallach 2018]

References

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research*3.Jan (2003): 993-1022.
- Chang, Jonathan, et al. "Reading tea leaves: How humans interpret topic models." *Advances in neural information processing systems*. 2009.
- Darling, William M. "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling." *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011.
- Gregor, Heinrich. "Parameter estimation for text analysis." *Technical report* (2005).
- Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21.3 (2013): 267-297.
- King, Gary, Jennifer Pan, and Margaret E. Roberts. "How the Chinese government fabricates social media posts for strategic distraction, not engaged argument." *American Political Science Review* 111.3 (2017): 484-501.
- Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* 111.515 (2016): 988-1003.
- Roberts, Margaret E., et al. "The structural topic model and applied social science." *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. 2013.
- Wallach, Hanna. “Computational social science ≠ computer science + social data”. Commun. ACM 61, 3 (2018), 42-44. DOI: https://doi.org/10.1145/3132698
- Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou, “Word embeddings quantify 100 years of gender and ethnic stereotypes“, PNAS (2018)
- Hamilton, William L., Jure Leskovec, and Dan Jurafsky, “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change”, ACL (2016)
- Mathur et al., “Manipulative tactics are the norm in political emails: Evidence from 100K emails from the 2020 U.S. election cycle”, Working Paper (2020)