NLP + Computational Social Science

Andy Halterman

Michigan State University

CS 375: Natural Language Processing
Williams College
Fall 2024

Table of Contents

1. Social science and NLP

2. Applied work

Social science and measurement

Make and test theories about the politics* to learn generalizable knowledge.
*“power” or “who gets what and why?”

Empirical social science rests on measurement:

Real world →

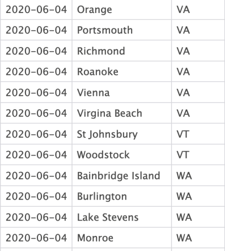
Broad concept →

“Protest"

Systematized concept →

“A crowd gathering to demonstrate their support for a set of political demands or claims to an external audience, typically with things like signs and banners and flags displayed to passers-by in a public space.”

Structured Data



2020-06-04	Orange	VA
2020-06-04	Portsmouth	VA
2020-06-04	Richmond	VA
2020-06-04	Roanoke	VA
2020-06-04	Vienna	VA
2020-06-04	Virgina Beach	VA
2020-06-04	St Johnsbury	VT
2020-06-04	Woodstock	VT
2020-06-04	Bainbridge Island	WA
2020-06-04	Burlington	WA
2020-06-04	Lake Stevens	WA
2020-06-04	Monroe	WA

Methods in political science

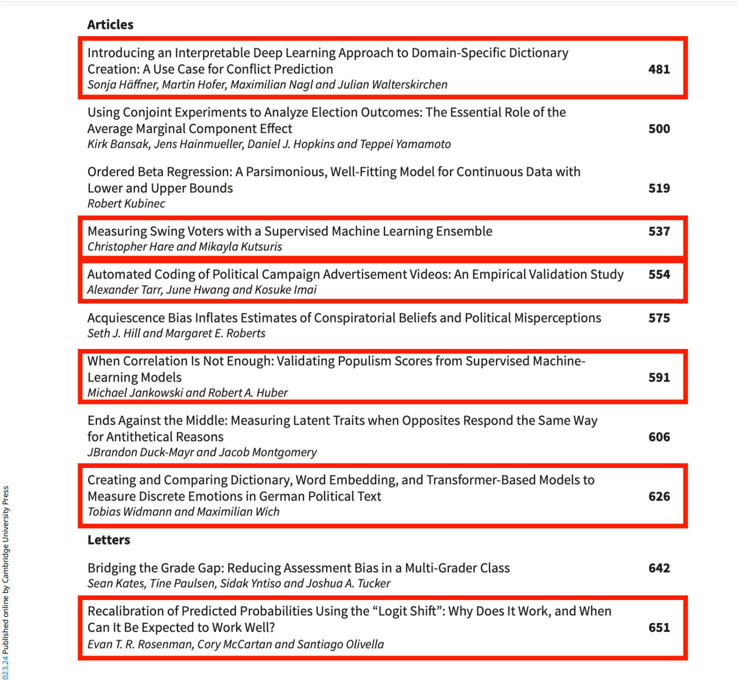
Political science is typically divided into subfields: American politics, comparative politics, international relations, political economy,...and methods. (Economics is similar, with econometrics).

Why do we have a separate methods subfield? We have data and questions that require specific tools.

Previously, this was mostly statistical innovations.

Increasingly, we trade with from computer science, machine learning, and natural language processing.

And it also goes the other way! Political science → statistics and computer science.



Articles

Letters

Text in political science

Text is a valuable source of raw data for political science.

- Text is a source of information about the real world. E.g.:
  - Where did protests take place?
  - Are human rights being respected?
- Text is also an object of study itself. E.g.:
  - How do legislators speak to their constituents?
  - How do Muslim clerics discuss religion + politics?

Text projects require the right combination of question, text source, method, and interpretation.

Why automate?

Why not just read the documents?

A: You should! (Grimmer and Stewart 2013) But you often can't just rely manual analysis.

- Some questions require scale: annotating all documents is infeasible. (E.g., 2 million+ declassified State Department cables from the 1970s)
- Consistent coding: want consistent, repeatable labels. (NB: LLMs change this!)
- Lower cost: important equity consideration.Large, well-funded projects can hire teams to annotate documents by hand. Individual researchers, especially students studying topics without much grant availability, cannot.

Examples

- The Chinese government permits online criticism of the regime, but does not permit attempts to organize online. “Flooding" social media is an effective alternative to direct censorship.
  - Scraped Weibo posts + keywords and topic models (King, Pan, and Roberts 2013; Roberts 2018).
- Local Indian deliberative bodies hold local officials to account; gender quotas reduce gender inequality in who is listened to.
  - Meeting transcripts + structural topic models (Parthasarathy, Rao, and Palaniswamy 2019)
- During the Berlin Crisis (1958-63), public statements were less effective signals than private communication or material actions.
  - 18,000 declassified diplomatic documents + random forest classifiers (Katagiri and Min 2019).

Bag-of-words models

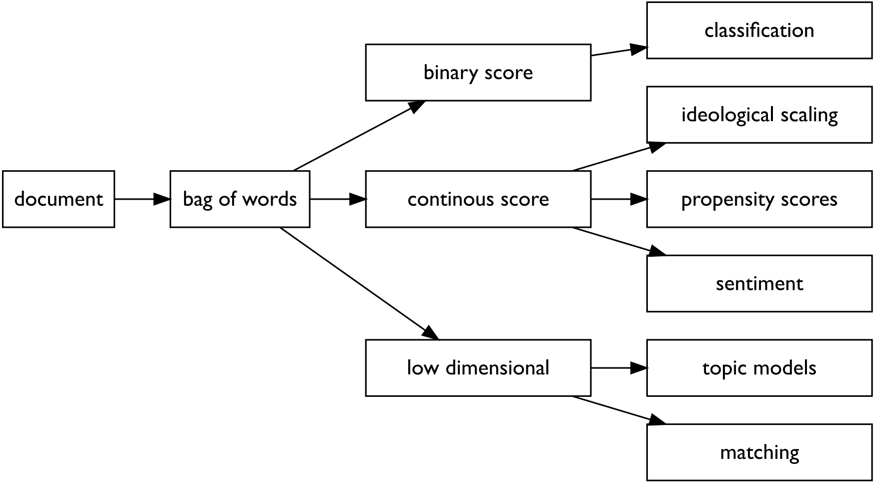


Figure 2. *Second Generation Text Analysis: Document Representation and Tasks*

New NLP techniques → new questions

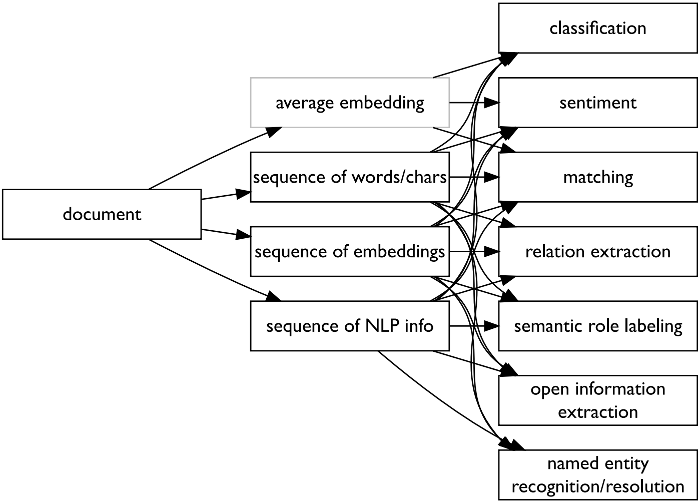


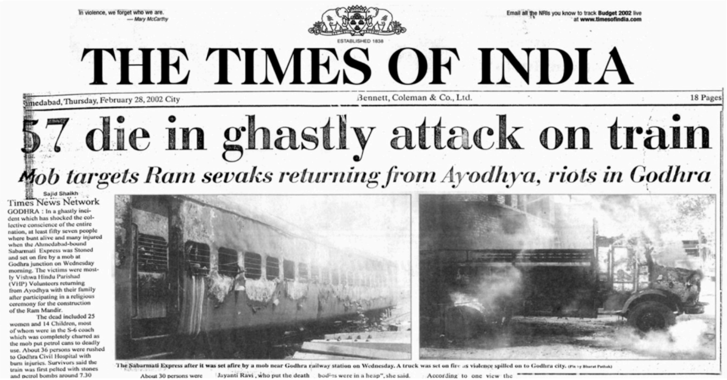
Figure 3. *Word Order-Aware Document Representations and Information Extraction*

Table of Contents

1. Social science and NLP

2. Applied work

India Police Events



In violence, we forget who we are.
— *Mary McCarthy*

ESTABLISHED 1838

Email all the NRIs you know to track **Budget 2002** live
at **www.timesofindia.com**

THE TIMES OF INDIA

Ahmedabad, Thursday, February 28, 2002 City | Bennett, Coleman & Co., Ltd. | 18 Pages

57 die in ghastly attack on train

Mob targets Ram sevaks returning from Ayodhya, riots in Godhra

Sajid Shaikh
Times News Network

GODHRA : In a ghastly incident which has shocked the collective conscience of the entire nation, at least fifty seven people where bunt alive and many injured when the Ahmedabad-bound Sabarmati Express was Stoned and set on fire by a mob at Godhra junction on Wednesday morning. The victims were mostly Vishwa Hindu Parishad (VHP) Volunteers returning from Ayodhya with their family after participating in a religious ceremony for the construction of the Ram Mandir.

The dead included 25 women and 14 Children, most of whom were in the S-6 coach which was completely charred as the mob put petrol cans to deadly use. About 36 persons were rushed to Godhra Civil Hospital with burn injuries. Survivors said the train was first pelted with stones and petrol bombs around 7.30

The Sabarmati Express after it was set afire by a mob near Godhra railway station on Wednesday. A truck was set on fire as violence spilled on to Godhra city. (Pix by Bharat Pathak)

About 30 persons were

Jayanti Ravi , who put the death

bodies were in a heap", she said.

According to one view the

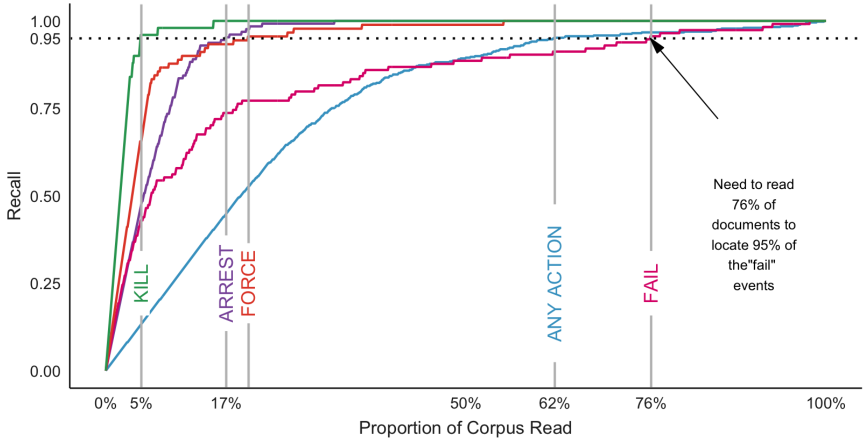
- Substantive: better understand the involvement of police in communal violence in India.
- Methodological: measure recall of event classifiers.

(Halterman, Keith, Sarwar, and O'Connor 2021)

New metric, inspired by applied research

Mixed methods: If a qualitative researchers wanted to read all relevant stories, could a classifier make them more efficient?

Order sentences by $\hat{p}$(label):

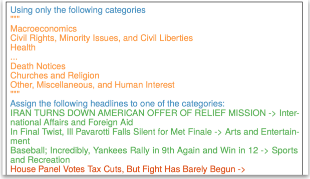


Need to read 76% of documents to locate 95% of the"fail" events

LLM-based document labeling

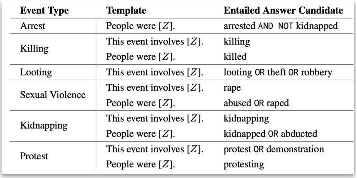
Social scientists are rapidly adopting LLMs to label documents.

- Can be done zero shot–no expensive training process.
- Works pretty well!



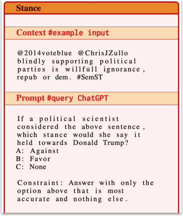
Using only the following categories
""""
Macroeconomics
Civil Rights, Minority Issues, and Civil Liberties
Health
...
Death Notices
Churches and Religion
Other, Miscellaneous, and Human Interest
""""
Assign the following headlines to one of the categories:
IRAN TURNS DOWN AMERICAN OFFER OF RELIEF MISSION -> International Affairs and Foreign Aid
In Final Twist, Ill Pavarotti Falls Silent for Met Finale -> Arts and Entertainment
Baseball; Incredibly, Yankees Rally in 9th Again and Win in 12 -> Sports and Recreation
House Panel Votes Tax Cuts, But Fight Has Barely Begun ->

Rytting et al. (2023)



Event Type	Template	Entailed Answer Candidate
Arrest	People were [Z].	arrested AND NOT kidnapped
Killing	This event involves [Z].	killing
	People were [Z].	killed
Looting	This event involves [Z].	looting OR theft OR robbery
Sexual Violence	This event involves [Z].	rape
	People were [Z].	abused OR raped
Kidnapping	This event involves [Z].	kidnapping
	People were [Z].	kidnapped OR abducted
Protest	This event involves [Z].	protest OR demonstration
	People were [Z].	protesting

Lefebvre and Stoehr (2022)



Stance

Context #example input

@2014voteblue @ChrisJZullo blindly supporting political parties is willfull ignorance, repub or dem. #SemST

Prompt #query ChatGPT

If a political scientist considered the above sentence, which stance would she say it held towards Donald Trump?
A: Against
B: Favor
C: None

Constraint: Answer with only the option above that is most accurate and nothing else.

Ziems et al. (2024)

But are we actually measuring what think we're measuring?

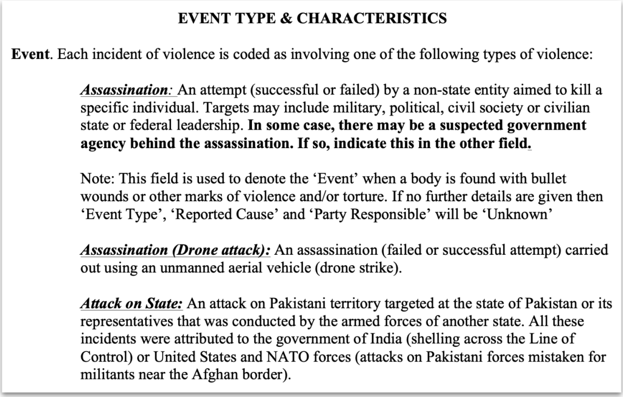
Is the LLM faithfully applying the definition we provide it? Or relying on heuristics and shortcuts?

Valid measurement with LLMs

When social scientists label documents, they rely on codebooks–documents that lay out labels, precise definitions, and coding instructions.

These codebooks are systematized constructs, rather than “background concepts.”

A different codebook should yield different labels.



EVENT TYPE & CHARACTERISTICS

Event. Each incident of violence is coded as involving one of the following types of violence:

***Assassination**:* An attempt (successful or failed) by a non-state entity aimed to kill a specific individual. Targets may include military, political, civil society or civilian state or federal leadership. **In some case, there may be a suspected government agency behind the assassination. If so, indicate this in the other field.**

Note: This field is used to denote the ‘Event’ when a body is found with bullet wounds or other marks of violence and/or torture. If no further details are given then ‘Event Type’, ‘Reported Cause’ and ‘Party Responsible’ will be ‘Unknown’

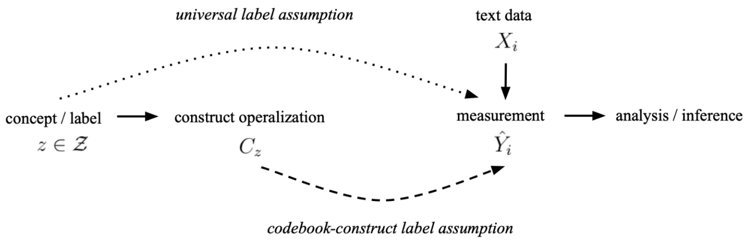
Assassination (Drone attack): An assassination (failed or successful attempt) carried out using an unmanned aerial vehicle (drone strike).

Attack on State: An attack on Pakistani territory targeted at the state of Pakistan or its representatives that was conducted by the armed forces of another state. All these incidents were attributed to the government of India (shelling across the Line of Control) or United States and NATO forces (attacks on Pakistani forces mistaken for militants near the Afghan border).

(Bueno de Mesquita et al. 2015)

Codebook measurement process

Codebook measurement task



universal label assumption

text data

X_i

concept / label

$z \in \mathcal{Z}$

construct operalization

C_z

measurement

$\hat{Y}_i$

analysis / inference

codebook-construct label assumption

Example

"protest"	CCC: must be directed toward a specific group or person, in proximity to them. Distinct from rallies or demonstrations.	news story + LLM (Mistral)	Do protests affect legislative votes?
"protest"	CAMEO: any collective action such as protests or demonstrations, carried out by civilians. May be violent. Gatherings supporting a person or policy are excluded.		

Data collection

We collect three codebooks/datasets

- BFRS dataset on political violence in Pakistan
- CCC (Crowd Counting Consortium) dataset on protests in the US
- The Manifesto Corpus dataset on party manifestos and ideology.

We compile the raw text, structured output/labels, and the original codebooks.

We then reformat the codebooks into a universal, semi-structured format.

Behavioral tests for LLM codebook compliance

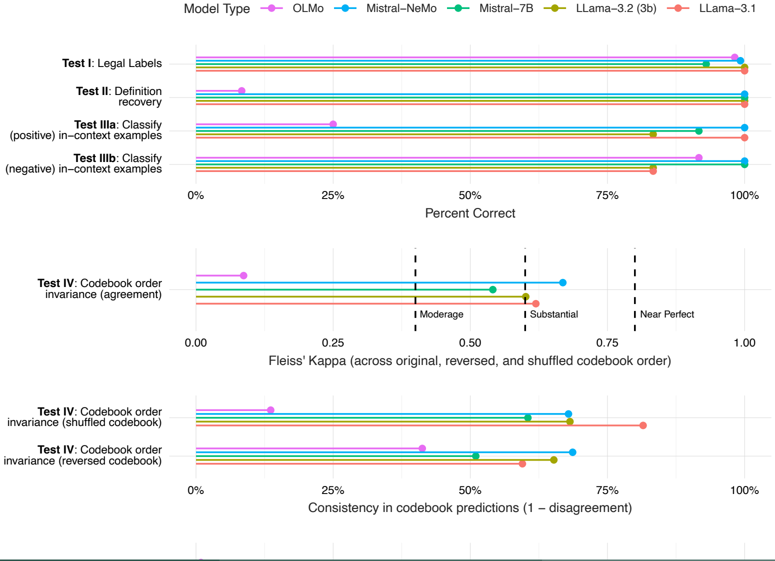
Inspired by the CHECKLIST approach proposed by Ribeiro et al. (2020) (Week 10), we propose basic behavioral tests for LLMs' ability to apply codebooks.

E.g.:

- An LLM should correctly label a verbatim definition or example from a codebook
- An LLM should only return allowed labels
- An LLM's predictions to be invariant to the codebook's order.
- An LLM should follow explicit, minimal instructions.

If an LLM fails these tasks, our confidence in its labels decreases.

Behavioral test results



Zero shot performance

Dataset	**Codebook Type**	**Llama-3.1-8B**	**Mistral-7B-v0.2**
manifestos	new	0.188	0.149
manifestos	original	0.206	0.141
ccc	new	0.609	0.649
ccc	original	0.484	0.511
bfrs	new	0.566	0.533
bfrs	original	0.547	0.436

Table: Performance comparison across datasets and codebook types

Ablation results (BFRS)

We can ablate parts of the codebook and re-run the zero shot pipeline to understand the important components of the dataset.

F1	Output Reminder	Pos. Ex.	Neg. Ex.	Clarif.	Negative Clarif.	Defn
0.28	1	0	0	0	0	0
0.42	0	1	1	0	0	0
0.25	0	1	1	1	1	0
0.09	1	1	1	1	1	0
0.04	1	1	1	1	1	1

1 = component ablated, 0 = component present

Discussion

Think back the social science measurement process shown above:

Real world → Broad concept →
Systematized concept → Structured Data

- Where can NLP improve these steps?
- Where can't it?
- Do LLMs fundamentally change how we can do measurement?
- What are the pitfalls of mis-applying NLP in social science research?