Introduction to Text Mining and Natural Language Processing

Session 5: Applications

Hannes Mueller

February 2026

Barcelona School of Economics

Orientation

Course Overview

Part 1: Project Design and Getting the Text (Session 2)

Part 2: Text Mining Basics (Sessions 3 and 4)

Part 3: Dimensionality Reduction (Sessions 5 to 7)

In-class assignment in session 6

Part 4: Supervised Learning with Text (Sessions 8 and 9)

In-class assignment in session 9

Write me your term-paper ideas throughout.

Term paper presentations: 16th of March (feedback!)

Recap

Recap

- We now know how to get to vector representations of documents.
- We have seen three types:
  - Absolute and relative term frequency counts
  - Tf-idf counts
  - Dictionary-based counts
- Until the arrival of LLMs, dictionary methods dominated social sciences. Today we will see some of these applications.

Dictionary Methods

Dictionary Methods

- Remember that the basic problem we are trying to solve is that the number of terms, V, is relatively high.
- Challenge in using text data for decision making is to therefore to reduce its dimensionality down from V.
- Dictionary methods are typically used in two cases:
  - 1) Human has a strong prior on what to use, i.e. specific keywords.
  - 2) Human knows what "kind" of text they want to capture, i.e. text which talks about finance.
- 1) is trivial. Key difficulty is how to come up with a dictionary for 2). We will do mostly literature discussion today.

Dictionary Method: Overview

- What is it?
- How has research derived dictionaries and used the resulting counts to capture specific concepts?
- Implementation of one example - partially left for small homework.

Dictionary Based Method

- Dictionary is a list of terms. Call this set of terms $\mathfrak{D}$.
- Boolean search provides a count of the number of times specific terms appear in a document, $x_{d,v}$.
- Important advantage: often you can query a database that you don't own to give you $x_{d,v}$
- In most methods, this is then aggregated to deliver some sort of score or index at document level (which is then typically further aggregated).
- In applications both the dictionaries vary widely and the ways to score documents vary.

Aggregation Methods

- Aggregation is as important as the dictionaries themselves.
- Simple sum: score d with $x_d = \sum_{v \in \mathfrak{D}} x_{d,v}$
- Why do we typically not use (aggregates of) these raw counts?
- Normalized sum: score d with $s_d = \sum_{v \in \mathfrak{D}} x_{d,v} / \sum_v x_{d,v}$.
- Indicator: score d with $I\left(\sum_{v \in \mathfrak{D}} x_{d,v} > 0\right)$
- Interaction: score d with $\prod\limits_{v \in \mathfrak{D}} I\left(x_{d,v} > 0\right)$

Applications

Cage et al (2019, Review of Economic Studies)

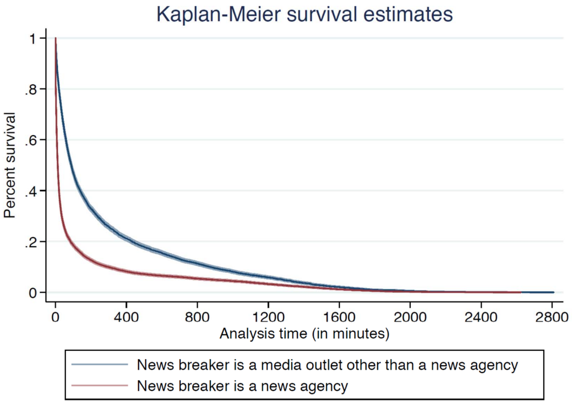
Cage et al (2019) Production of Information in an Online World

Study of news french news content across print media and online sources.

Question is whether old media can survive competition by online sources if the latter can copy.

- document high reactivity of online media: one quarter of the news stories are reproduced online in under 4 min
- substantial copying, both at the extensive and at the intensive margins
- estimate the returns to originality in online news production
  - original content producers tend to receive more viewers, thereby mitigating the newsgathering incentive problem raised by copying

Study findings



Notes: The figure plots the Kaplan-Meier survivor functions when the news breaker is a news agency (the AFP or Reuters) (red line) and when the breaker is a media outlet other than a news agency (blue line). The confidence level for the pointwise confidence bands is 95%.

Cage et al (2019) Method

Example of what is called Topic Detection and Tracking (TDT).

They want to detect news reporting on events and how it spreads.

Key is therefore the defintition of an event.

Cage et al (2019)

Their event detection algorithm has clustering at its core. Their method:

- remove stop words and stem
- join headline and the text: apply a multiplicative factor of five to the words of the title as they are supposed to describe the event well
- **apply TFIDF** (why?)
- clustered in a bottom-up fashion to form the events based on their cosine similarity
- iterative agglomerative clustering algorithm is stopped when the distance between documents reaches a given threshold
- cluster is finalized it does not receive any new document for a one-day window

Gentzkow and Shapiro (2010)

Gentzkow and Shapiro (2010, Econometrica)

- They construct similarity measure to get to a measure of media slant.
- Want to distinguish between *Republican* and *Democrat* kind of speech.
- (i) Compute the total number of times that each phrase, v, appeared in newspaper corpus from 2000 to 2005.
  - two-word phrases that appeared in at least 200 but no more than 15 000 newspaper headlines
  - three-word phrases that appeared in at least 5 but no more than 1000 headlines
  - drop any phrase that appeared in the full text of more than 400 000 documents.

Building the Dictionary

- (ii) Call the count of term v in the speeches of congressperson c, $x_{c,v}$.
- Terms v ranked according to statistic:

$$\chi^2_v = \frac{\left(x_{v,R} * x_{\sim v,D} - x_{v,D} * x_{\sim v,R}\right)^2}{\left(x_{v,R} + x_{v,D}\right)\left(x_{v,R} + x_{\sim v,D}\right)\left(x_{\sim v,R} + x_{v,D}\right)\left(x_{\sim v,R} + x_{\sim v,D}\right)}$$

- Where $x_{\sim v,D}$ are all other terms spoken by democrats and χ^2_v captures how special a term is to the Democrats or Republicans.
- They pick 500 two-term phrases and 500 three-term phrases with highest χ^2_v.

TABLE I

MOST PARTISAN PHRASES FROM THE 2005 *CONGRESSIONAL RECORD*[a]

Panel A: Phrases Used More Often by Democrats

Two-Word Phrases		
private accounts	Rosa Parks	workers rights
trade agreement	President budget	poor people
American people	Republican party	Republican leader
tax breaks	change the rules	Arctic refuge
trade deficit	minimum wage	cut funding
oil companies	budget deficit	American workers
credit card	Republican senators	living in poverty
nuclear option	privatization plan	Senate Republicans
war in Iraq	wildlife refuge	fuel efficiency
middle class	card companies	national wildlife
Three-Word Phrases		
veterans health care	corporation for public broadcasting	cut health care
congressional black caucus	additional tax cuts	civil rights movement
VA health care	pay for tax cuts	cuts to child support
billion in tax cuts	tax cuts for people	drilling in the Arctic National
credit card companies	oil and gas companies	victims of gun violence
security trust fund	prescription drug bill	solvency of social security
social security trust	caliber sniper rifles	Voting Rights Act
privatize social security	increase in the minimum wage	war in Iraq and Afghanistan
American free trade	system of checks and balances	civil rights protections
central American free	middle class families	credit card debt

(Continues)

Panel B: Phrases Used More Often by Republicans

Two-Word Phrases		
stem cell	personal accounts	retirement accounts
natural gas	Saddam Hussein	government spending
death tax	pass the bill	national forest
illegal aliens	private property	minority leader
class action	border security	urge support
war on terror	President announces	cell lines
embryonic stem	human life	cord blood
tax relief	Chief Justice	action lawsuits
illegal immigration	human embryos	economic growth
date the time	increase taxes	food program
Three-Word Phrases		
embryonic stem cell	Circuit Court of Appeals	Tongass national forest
hate crimes legislation	death tax repeal	pluripotent stem cells
adult stem cells	housing and urban affairs	Supreme Court of Texas
oil for food program	million jobs created	Justice Priscilla Owen
personal retirement accounts	national flood insurance	Justice Janice Rogers
energy and natural resources	oil for food scandal	American Bar Association
global war on terror	private property rights	growth and job creation
hate crimes law	temporary worker program	natural gas natural
change hearts and minds	class action reform	Grand Ole Opry
global war on terrorism	Chief Justice Rehnquist	reform social security

[a]The top 60 Democratic and Republican phrases, respectively, are shown ranked by χ^2_{pl}. The phrases are classified as two or three word after dropping common "stopwords" such as "for" and "the." See Section 3 for details and see Appendix B (online) for a more extensive phrase list.

Coding of Newspapers

- 1) Code a dummy (y_c) which takes a value 1 for democrats and then run regression

$$x_{c,v} = \alpha_v + \beta_v * y_c + \varepsilon_{c,v}$$

  and β_v gives you how *democratic* the term is.

- 2) Then the newspaper ideology is taken from regressions of $x_{n,v} - \alpha_v$ on the slope indicators β_v which yields

$$\hat{y}_n = \frac{\sum_{v=1}^{1000} \beta_v \left(x_{n,v} - \alpha_v\right)}{\sum_{v=1}^{1000} \beta_v^2}$$

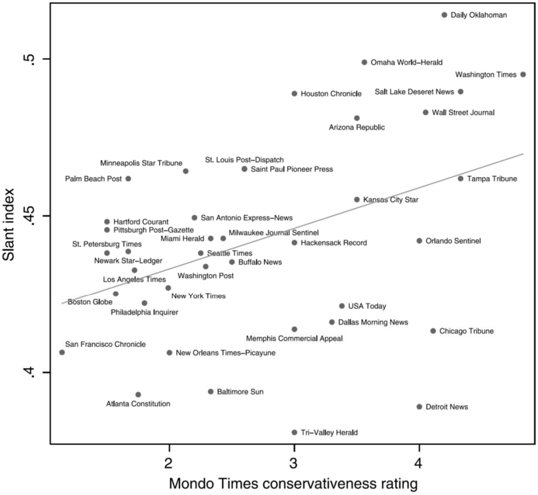


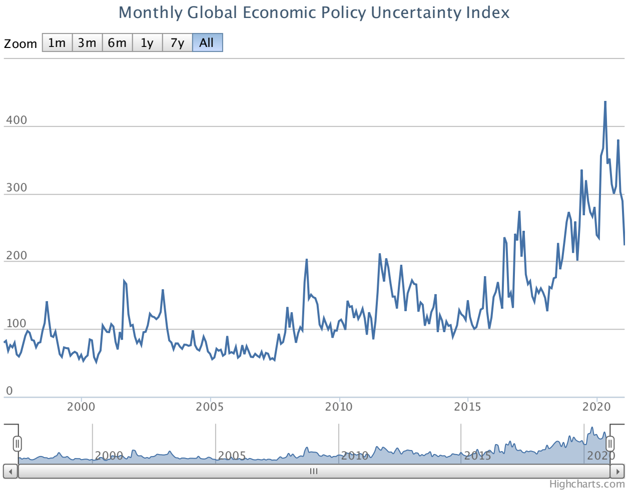
FIGURE 1.—Language-based and reader-submitted ratings of slant. The slant index (y axis) is shown against the average Mondo Times user rating of newspaper conservativeness (x axis), which ranges from 1 (liberal) to 5 (conservative). Included are all papers rated by at least two users on Mondo Times, with at least 25,000 mentions of our 1000 phrases in 2005. The line is predicted slant from an OLS regression of slant on Mondo Times rating. The correlation coefficient is 0.40 ($p = 0.0114$).

Baker, Bloom and Davis (2016)

Example 2: Baker, Bloom, and Davis (2016)

- Baker, Bloom, and Davis (2016) use dictionaries to produce the data behind `http://www.policyuncertainty.com/`
- BBD are interested in measuring economic policy uncertainty.
- Uncertainty about policies might be a key driver of economic activity.

EPU



Monthly Global Economic Policy Uncertainty Index
Zoom 1m 3m 6m 1y 7y All
400
300
200
100
0
2000
2005
2010
2015
2020
Highcharts.com

Research Goal

- Capture uncertainty about
  - who will make economic policy decisions
  - what economic policy actions will be undertaken and when
  - the economic effects of policy actions (or inaction)
- Including uncertainties related to the economic ramifications of "noneconomic" policy matters, for example, military actions

Method Overview

- BBD create an index based on Boolean searches of newspaper articles from major newspapers.
- For each paper they submit the following queries (separately):
  - 1. (E) Article contains "economic" OR "economy"
  - 2. (P) Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"
  - 3. (U) Article contains "uncertain" OR "uncertainty"
- **How would you combine these indicators?**

Aggregation Across Newspapers

- They indicate EPU if there is at least one E AND P AND U.
- Note that the EPU does not capture intensity within articles.
- Count number of EPU articles each month.
- Take resulting article counts, and normalize by total newspaper articles that month.
- Call this the EPU frequency, X_{it}, note we have one for each newspaper i and month t.
- Standardize X_{it} by times-series variance, σ_i, for each newspaper
- Take mean value of standardized values across newspapers.
- Normalize so that the overall mean is at 100.

Hassan et al (2019)

Hassan et al (2019)

- Hassan et al (2019) Firm-Level Political Risk: Measurement and Effects
- They want to build measure of political risk faced by individual US firms.
- Data: 178,173 earnings conference call transcripts
- Idea: measure of the share of the quarterly conversation between call participants and firm management that centres on risks associated with political matters

Building the Dictionary

- training library of political text archetypical of the discussion of politics, $\boldsymbol{P}$
- training library of non-political text, archetypical of the discussion of non-political topics, $\boldsymbol{N}$
- ($\boldsymbol{P}$) William T. Bianco and David T. Canon - American Politics Today
- ($\boldsymbol{N}$) Robert Libby, Patricia A. Libby, and Daniel G. Short's - Financial Accounting
- each library is the set of all adjacent two-word combinations ("bigrams") contained in the text
- $\boldsymbol{P} \setminus \boldsymbol{N}$: terms that are in the political texts but not in the non-political text

Building the Dictionary

- First key statistics for them is $f_{b,\boldsymbol{P}}/B_{\boldsymbol{P}}$
  - $f_{b,\boldsymbol{P}}$ is the frequency of bigram b in the political training library
  - $B_{\boldsymbol{P}}$ is the total number of bigrams in the political training library
  - What is the ratio $f_{b,\boldsymbol{P}}/B_{\boldsymbol{P}}$ therefore?
  - Relative term frequency of bigram b in P. Similar to $tf_{d,v}$.

Building the Dictionary

- Second key statistics is $\mathbf{1}\left[b \in \boldsymbol{P} \backslash \boldsymbol{N}\right]$
- $\boldsymbol{P} \backslash \boldsymbol{N}$: in political texts but not in the non-political text
  - Where $\mathbf{1}\left[\cdot\right]$ is an indicator function.
  - This is a particularly brutal way of doing an idf_b across libraries. **Why?**
  - idf_b would give more weight to terms that are "special" to library $\boldsymbol{P}$, i.e. not as frequent in $\boldsymbol{N}$**.**
  - Here the weight is set to 0 for all terms in $\boldsymbol{P}$ that are also in $\boldsymbol{N}$**.**

TABLE II
TOP 120 POLITICAL BIGRAMS USED IN CONSTRUCTION OF $PRisk_{i,t}$

Bigram	$\frac{f_{b,\mathbb{P}}}{B_{\mathbb{P}}} \times 10^5$	Frequency	Bigram	$\frac{f_{b,\mathbb{P}}}{B_{\mathbb{P}}} \times 10^5$	Frequency
the constitution	201.15	9	governor and	26.79	11
the states	134.29	203	government the	26.39	56
public opinion	119.05	4	this election	25.98	26
interest groups	118.46	8	political party	25.80	5
of government	115.53	316	American political	25.80	2
the GOP	102.22	1	politics of	25.80	5
in Congress	78.00	107	White House	25.80	21
national government	68.03	7	the politics	25.80	31
social policy	62.16	1	general election	25.22	30
the civil	60.99	64	and political	25.22	985
elected officials	60.40	3	policy is	25.22	135
politics is	53.95	7	the islamic	25.04	1
political parties	51.61	3	Federal Reserve	24.63	119
office of	51.02	58	judicial review	24.04	6
the political	51.02	1,091	vote for	23.46	6
interest group	48.09	1	limits on	23.46	53
the bureaucracy	48.09	1	the FAA	23.28	22
and Senate	46.33	19	the presidency	22.87	2
government and	44.57	325	shall not	22.87	4
for governor	41.48	2	the nation	22.87	52
executive branch	40.46	3	constitution and	22.87	3
support for	39.88	147	Senate and	22.87	28
the EPA	39.15	139	the VA	22.65	77
in government	38.70	209	of citizens	22.28	12
Congress to	36.95	19	any state	22.28	7
political process	36.36	18	the electoral	22.28	5
care reform	35.77	106	a president	21.70	6
government in	35.19	77	the governments	21.70	201
due process	35.19	6	clause of	21.11	1
President Obama	34.60	7	and Congress	21.11	7
and social	34.60	140	the partys	21.11	1
first amendment	34.01	1	the Taliban	20.64	1
Congress the	34.01	9	a yes	20.64	12
the Republican	33.43	10	other nations	20.53	1
Tea Party	33.43	1	passed by	20.53	13
the legislative	33.43	92	states or	20.53	40
of civil	32.84	14	free market	20.53	29
court has	32.84	30	that Congress	20.53	30
groups and	32.25	109	national and	20.53	194
struck down	31.67	3	most Americans	19.94	2
shall have	31.67	7	of religion	19.94	1
civil war	31.67	8	powers and	19.94	3
the Congress	31.67	50	a government	19.94	92

Use of Dictionary

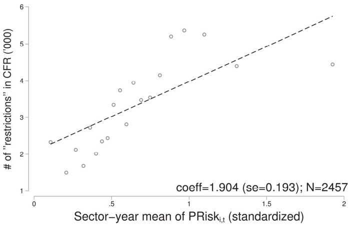
- Count the number of instances where political bigrams are used in conjunction with synonyms for "risk".
- Conference-call transcript of firm i in quarter t into a list of bigrams contained in the transcript $b = 1, ..., B_{it}$.

$$PRisk_{it} = \frac{\sum_{b=1}^{B_{it}} \left(\mathbf{1}\left[b \in \boldsymbol{P} \backslash \boldsymbol{N}\right] \times \mathbf{1}\left[|b - r| < 10\right] \times \frac{f_{b,\boldsymbol{P}}}{B_{\boldsymbol{P}}} \right)}{B_{it}}$$

- r is the position of the nearest synonym for risk or uncertainty]

Some Results

(A) Index of regulatory constraints



(B) Share of revenue from federal government

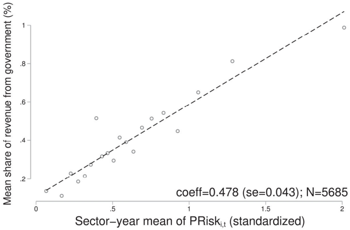


FIGURE III

$PRisk_{i,t}$ and Sector Exposure to Politics

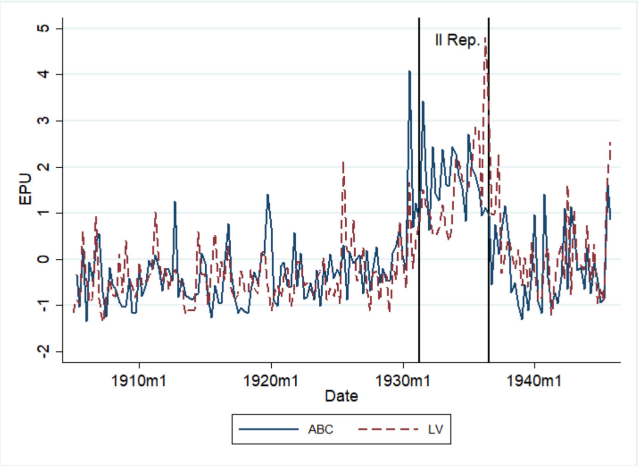
Garcia-Uribe et al (2023, Journal of Economic History)

Garcia-Uribe et al (2023, Journal of Economic History)

- Garcia-Uribe et al (2023) Economic Uncertainty and Divisive Politics: Evidence from the ”dos Españas”
- Construct the EPU index for Spain in 1905-1945.
- Historic data does not provide article split: we simulate this.
- Find shift upward before the civil war broke out.
- Question: did this correlate with political tensions?
- We use a mild version of the Hassan et al (2019) idea.

EPU in Spain before the Civil War

Figure 1: EPU Index for Spain: 1905-1945



Note: The EPU index is calculated using the procedure described in Appendix B. Quarterly data used. Sample period: 1905–1945.

Building the Dictionary for Divisions

- We use *supervision* through the work that historians have done.
- Typically discussions in history books are structured like this: `https://es.wikipedia.org/wiki/Segunda_Rep%C3%BAblica_espa%C3%B1ola`
- Idea: exploit accounts of pre-civil war period talk about four divisive issues - socioeconomic conflict, regional separatism, the power of the military, the role of the church/religious education
- We copy the text of the description of these issues into four different documents and then calculate the tf-idf on the terms in the four documents:
- **What will happen?**
- We then simply take the top terms as our dictionaries for the four issues.

Resulting Dictionary

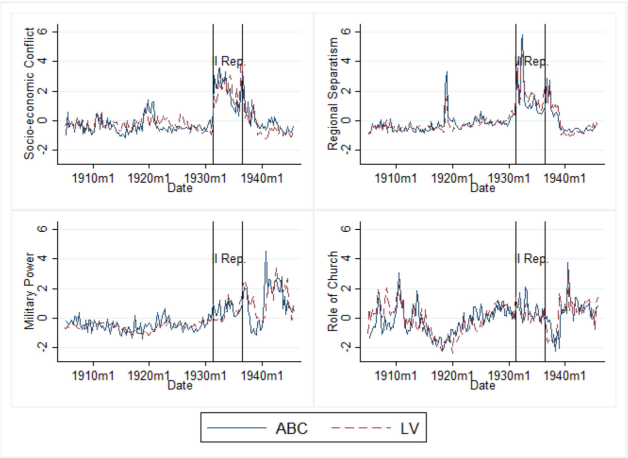
Table 1: Dictionaries of Four Divisive Issues

Socio-Economic Conflict	Regional Separatism	Role of Church	Power of Military
tierras	**estatuto**	iglesia	militar
trabajo	cataluña	**católicos**	ejército
reforma agraria	proyecto	enseñanza	oficiales
reforma	**vasco**	poltica	militares
agraria	catalana	**católica**	guerra
campesinos	**proyecto estatuto**	órdenes	generales
casas	**autonomía**	entonces	**reforma militar**
jurados mixtos	**federal**	**cardenal**	**ascensos**
jurados	**integral**	parte	**orden público**
mixtos	macià	hizo	civil
viejas	**estatuto cataluña**	**conventos**	orden
casas viejas	**república catalana**	segura	reforma
grandes	catalán	madrid	público
largo	barcelona	**iglesia católica**	guardia
largo caballero	izquierda	españoles	decreto
caballero	**catalanes**	**religiosos**	parte
extremadura	consejo	**religiosas**	**mantuvo**
huelgas	**referéndum**	**edificios**	fuerzas
instituto	generalidad	civil	cuerpo
social	navarra	régimen	servicio
contratos	vascos	española	**retiro**
jornaleros	mayoría	intelectuales	seis
propietarios	aprobado	intelectual	seis meses
fincas	nuevo	creía	**armadas**
parte	noviembre	**católico**	**fuerzas armadas**
salarios	regiones	**pastoral**	militar manuel
contratos trabajo	país vasco	órdenes religiosas	**profesional**
hectáreas	votos	marañón	**jurisdicción militar**
obreros	diputados	maestros	**oficialidad**
ministro trabajo	francesc	colegios	armas

Note: The words under each issue are the 30 initial words from the tf-idf model. The bold-faced words are the ones finally used for the indices after removing common and period-specific words. See Appendix C for details. See Table A2 for an English translation.

Resulting Division Measurement

Figure 2: Four Divisive Issues



ote: The four indices are calculated with a tf-idf model. See Appendix C for details. Quarterly data used. Sample period: 905–1945.

Boehme et al 2020

Boehme et al 2020

- Boehme et al (2020) Searching for a better life: Predicting international migration with online search keywords
- They want to build a dictionary linked to migration.
- Use the website *Semantic Link* `http://semantic-link.com/`
- The page uses text from English language Wikipedia and identifies pairs of keywords which are semantically related.
- Links are built using MI criterion (information theory concept)

Resulting Dictionaries

Table 1
List of main keywords.

English	French	Spanish
applicant	candidat	solicitante
arrival	arrivee	llegada
asylum	asile	asilo
benefit	allocation sociale	beneficio
border control	controle frontiere	control frontera
business	entreprise	negocio
citizenship	citoyennete	ciudadania
compensation	compensation	compensacion
consulate	consulat	consulado
contract	contrat	contrato
customs	douane	aduana
deportation	expulsion	deportacion
diaspora	diaspora	diaspora
discriminate	discriminer	discriminar
earning	revenu	ganancia
economy	economie	economia
embassy	ambassade	embajada
emigrant	emigre	emigrante
emigrate	emigrer	emigrar
emigration	emigration	emigracion
employer	employer	empleador
employment	emploi	empleo
foreigner	etranger	extranjero
GDP	PIB	PIB
hiring	embauche	contratacion
illegal	illegal	ilegal
immigrant	immigre	inmigrante
immigrate	immigrer	inmigrar
immigration	immigration	inmigracion

Improvement in Forecast

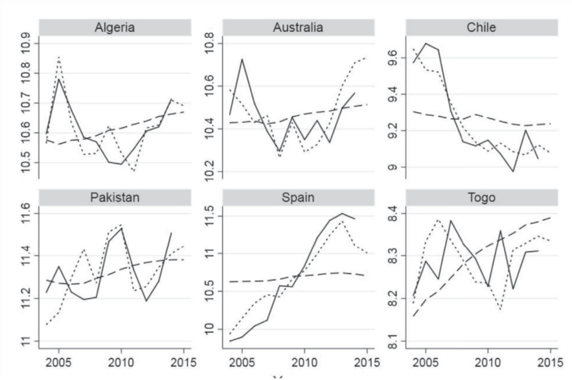


Fig. 1. Descriptive illustration of GTI in predicting migration flows. *Notes:* The figure shows log migration flows (plus one) from six origin countries to the OECD (solid line) and fitted values of two simple regressions that use log GDP, log population size, origin-specific intercepts and fixed effects (dashed line) plus the GTI (dotted line). The regressions are estimated on the full sample including all countries, the model used to fit the data is thus identical across panels. Differences between dotted and dashed lines are thus based on changes in GTI search intensities. As the dashed line shows, GDP and population size change too slowly to explain large short term fluctuations in migration flows.

ICEWS, GDELT, POLECAT

Event Extractors

- There were three event extraction monsters out there called GDELT, ICEWS and POLECAT
- These are massive efforts to extract events from news text.
- GDELT is the only one still running. Go to `https://www.gdeltproject.org/`, to get access.
- Gives you *who does what to whom*?
- GDELT is based on huge dictionaries: `https://www.gdeltproject.org/data/documentation/CAMEO.Manual.1.1b3.pdf`
- Coolest update is POLECAT: `https://arxiv.org/pdf/2304.01331`

ICEWS, GDELT, POLECAT

Event Extractors

- There were three event extraction monsters out there called GDELT, ICEWS and POLECAT
- These are massive efforts to extract events from news text.
- GDELT is the only one still running. Go to `https://www.gdeltproject.org/`, to get access.
- Gives you *who does what to whom*?
- GDELT is based on huge dictionaries: `https://www.gdeltproject.org/data/documentation/CAMEO.Manual.1.1b3.pdf`
- Coolest update is POLECAT: `https://arxiv.org/pdf/2304.01331`

Exam

I asked GenAI something like the following and worked from there.

Help me make an exam for my students. Go through my slides and codes. Give me questions for a 30 minute in-class assignment. Additional explanation: If you could ask just ten questions for an in-class assignment where they just have pen and paper. What would you ask? It needs to be a ”dumb” and essential thing we test. I want this to be basically a test whether people were asleep. Answers need to not be longer essays.

Ask GenAI to develop a training material for this.

Everyone must *understand* the vectorizers and DTM.