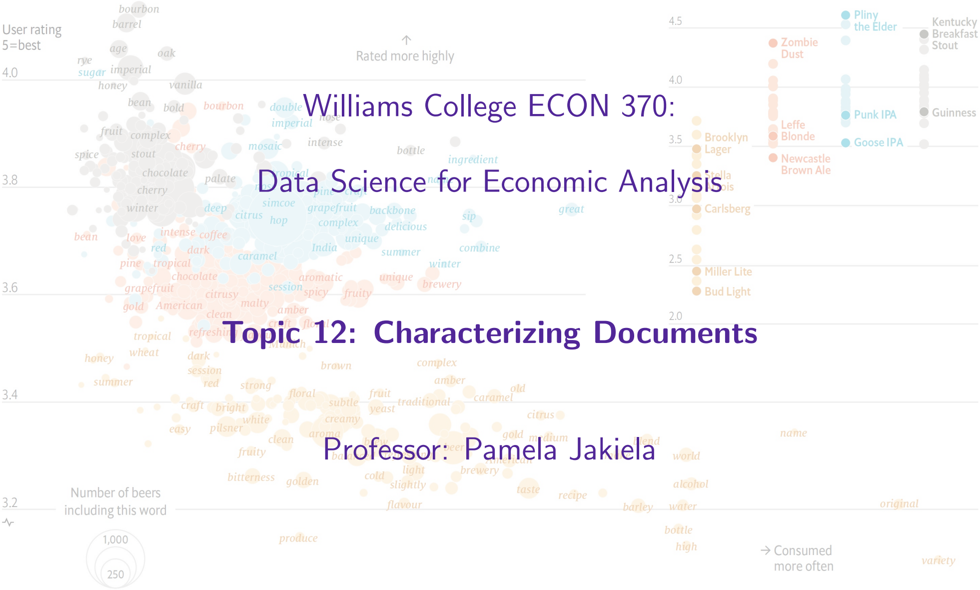


Williams College ECON 370:

Data Science for Economic Analysis

Topic 12: Characterizing Documents

Professor: Pamela Jakiela

Outline

- Sentiment analysis
- Validation
- Dictionary-based methods
- Document distance

Measuring Consumer Sentiment

University of Michigan and Federal reserve Bank of St. Louis

Index of Consumer Sentiment

Based on responses to survey questions:

- We are interested in how people are getting along financially these days. Would you say that you (and your family) are better off or worse off financially than you were a year ago?
- Now looking ahead – do you think that a year from now you (and your family) will be better off financially, or worse off, or just about the same as now?
- Now turning to business conditions in the country as a whole – do you think that during the next twelve months we'll have good times financially, or bad times, or what?
- Looking ahead, which would you say is more likely – that in the country as a whole we'll have continuous good times during the next five years or so, or that we will have periods of widespread unemployment or depression, or what?
- About the big things people buy for their homes – such as furniture, a refrigerator, stove, television, and things like that. Generally speaking, do you think now is a good or bad time for people to buy major household items?

Potential Problems with Survey-Based Measures of Expectations

Survey-based measures may or may not capture consumer expectations about the economy

- Responses may or may not reflect true underlying beliefs
  - Social desirability bias, limited opportunities for reflection
- Sample selection: individuals who respond to surveys may not be representative

Many complimentary sources of text data on expectations about the economy

- Fed Open Markets Committee, Central Bank governors, financial journalists, twitter, etc.

Sentiment Analysis

In **sentiment analysis**, we assign documents to categories reflecting their emotional content

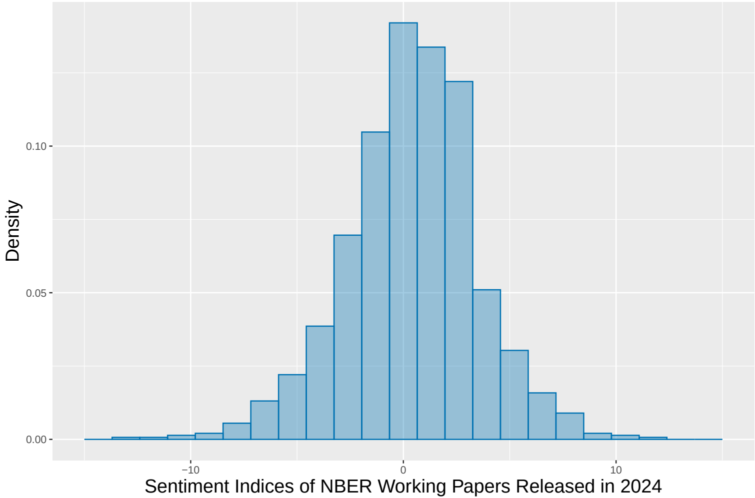
- The simplest forms of sentiment analysis use a **dictionary-based** methodology, counting the number of (for example) positive and negative words contained in each document
- Dictionary-based methods can also be used to assess political slant or capture topics

The bing sentiments data, for example, classifies almost 7,000 words as positive or negative

- Positive: abundance, abundant, accessible, acclaim, accolade, accomplish, achievement...
- Negative: abnormal, abolish, abominable, abrupt, abscond, absentee, absurd, abuse...

There are many known issues with sentiment analysis (accounting for context, phrases, irony), only some of which can be solved through better validation and context-specific dictionaries

Sentiment Analysis in NBER Working Paper Abstracts



Sentiment Analysis in Practice

Most negative paper: **"Diminishing Marginal Utility Revisited"**

How quickly does marginal utility fall with increasing consumption? It depends on the dimension along which we consider concavity of the utility function. This paper estimates the distribution of heterogeneous curvature parameters in individuals' utility functions from hypothetical choice data, while accounting for survey response error. Types of curvature examined include relative risk aversion, intertemporal substitution, the reciprocal of the altruism elasticity, and a new measure of inequality aversion, which queries how much more a dollar means to a poor family than to a rich family...

Second most negative paper: **"Failing Banks"**

Why do banks fail? We create a panel covering most commercial banks from 1863 through 2024 to study the history of failing banks in the United States. Failing banks are characterized by rising asset losses, deteriorating solvency, and an increasing reliance on expensive noncore funding. These commonalities imply that bank failures are highly predictable using simple accounting metrics from publicly available financial statements...

Validation

In **sentiment analysis**, we assign documents to categories reflecting their emotional content

- Whenever researchers use a method designed to measure a construct (e.g. sentiment), they need to establish that the method captures what they think it does (**validation**)
- Many sentiment dictionaries were developed for purposes other than research (in fact, almost all data science tools were not developed for quantitative social science)

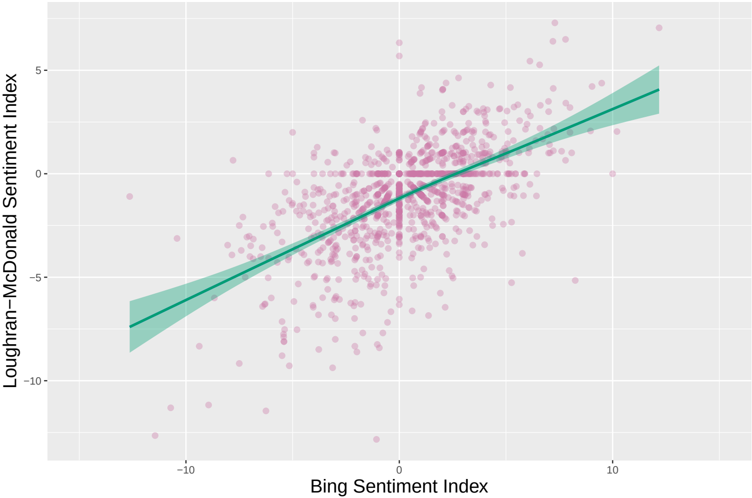
How would you validate a measure of sentiments appropriate for economics?

- ~~Maybe don't look at academic working papers~~
- Be precise about what you are attempting to capture, in relation to your research question
- Hand code a small sample of relevant documents (e.g. articles about the economy)
- Use an off-the-shelf dictionary appropriate to your documents (i.e. economics/finance)
- Check that your results make sense (in a subsample), modify dictionary accordingly

Sentiment Dictionaries Appropriate for Economics and Finance

- Loughran-McDonald Master Dictionary (Loughran and McDonald, 2011)
  - "In a large sample of 10-Ks [filed wtih the SEC] during 1994 to 2008, almost three-fourths of the words identified as negative by the widely used Harvard Dictionary are words typically not considered negative in financial contexts."
- Shapiro, Sudhof, Wilson (2022) build on Loughran-McDonald to create and validate a lexicon specifically designed to capture positive/negative tone in economic news articles
  - Compare dictionary-based sentiment indices to scores by trained human coders

Economics-Specific vs. General Measures of Sentiment



Sentiment Analysis in Economics: Takeaways

- What is your research question?
  - Are papers about wealthier countries more positive?
  - Does research paper sentiment predict future economic growth?
  - Do women/men/tenured professors express a more positive outlook in their writing?
- What construct are you trying to capture?
  - Subject matter vs. emotional content vs. subjective expectations
- How can you validate your proposed measure?

Economic Policy Uncertainty

THE

QUARTERLY JOURNAL OF ECONOMICS

Vol. 131 November 2016 Issue 4

MEASURING ECONOMIC POLICY UNCERTAINTY*

SCOTT R. BAKER
NICHOLAS BLOOM
STEVEN J. DAVIS

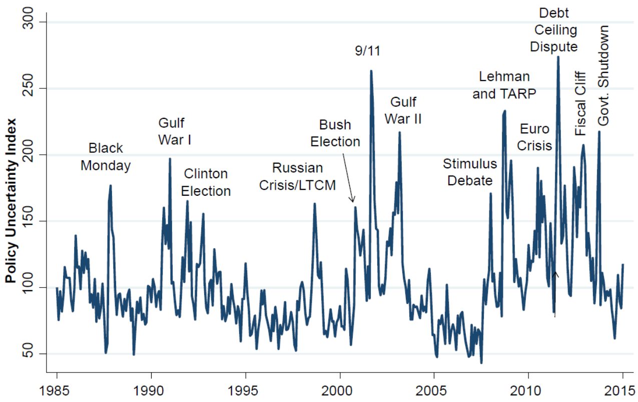
We develop a new index of economic policy uncertainty (EPU) based on newspaper coverage frequency. Several types of evidence—including human readings of 12,000 newspaper articles—indicate that our index proxies for movements in policy-related economic uncertainty. Our U.S. index spikes near tight presidential elections, Gulf Wars I and II, the 9/11 attacks, the failure of Lehman Brothers, the 2011 debt ceiling dispute, and other major battles over fiscal policy. Using firm-level data, we find that policy uncertainty is associated with greater stock price volatility and reduced investment and employment in policy-sensitive sectors like defense, health care, finance, and infrastructure construction. At the macro level, innovations in policy uncertainty foreshadow declines in investment, output, and employment in the United States and, in a panel vector autoregressive setting, for 12 major economies. Extending our U.S. index back to 1900, EPU rose dramatically in the 1930s (from late 1931) and has drifted upward since the 1960s. *JEL Codes:* D80, E22, E66, G18, L50.

Measuring Economic Policy Uncertainty

- **Data:** articles from *USA Today*, *Miami Herald*, *Chicago Tribune*, *Washington Post*, *Los Angeles Times*, *Boston Globe*, *San Francisco Chronicle*, *Dallas Morning New*, *New York Times*, and *Wall Street Journal* published between 1985 and 2015
- **Measure of economic policy uncertainty:**
  - Article contains: "uncertain"
  - Article contains: "economic" OR "economy" OR "economies"
  - Article contains: "regulation" OR "deficit" OR "legislation" OR "congress" OR "white house" OR "Federal Reserve" OR "the Fed" OR "regulations" OR "regulatory" OR "deficits" OR "congressional" OR "legislative" OR "legislature"

⇒ Newspaper-by-month measure of fraction of articles discussing economic policy uncertainty

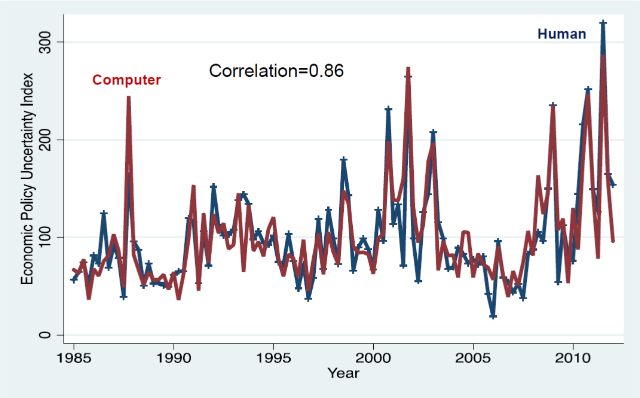
Quantifying Economic Policy Uncertainty



Notes: Index reflects scaled monthly counts of articles containing 'uncertain' or 'uncertainty', 'economic' or 'economy', and one or more policy relevant terms: 'regulation', 'federal reserve', 'deficit', 'congress', 'legislation', or 'white house'. The series is normalized to mean 100 from 1985-2009 and based on queries run on 2 February, 2015 for the USA Today, Miami Herald, Chicago Tribune, Washington Post, LA Times, Boston Globe, SF Chronicle, Dallas Morning News, NY Times, and the Wall Street Journal.

Source: Baker, Bloom, and Davis (2015)

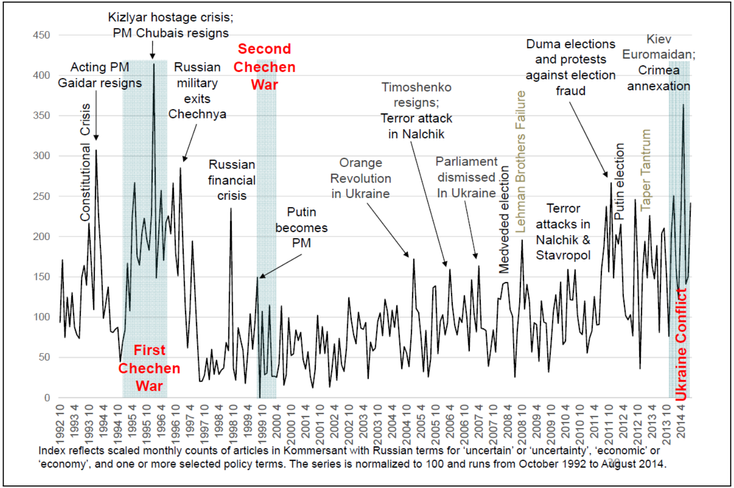
Validating the Measure of Economic Policy Uncertainty



Notes: Index comparison from 1985 Q1 to 2012 Q1 based on 3,723 articles (4,388 audits) in the Chicago Tribune, Dallas Morning News, LA Times, Miami Herald, NY Times, San Francisco Chronicle, Washington Post and Wall Street Journal. Series are plotted quarterly to reduce sampling variability, with an average of 33 articles per quarter. Each series is normalized to 100 from 1985-2009. See text for additional discussion of the audit process and this comparison.

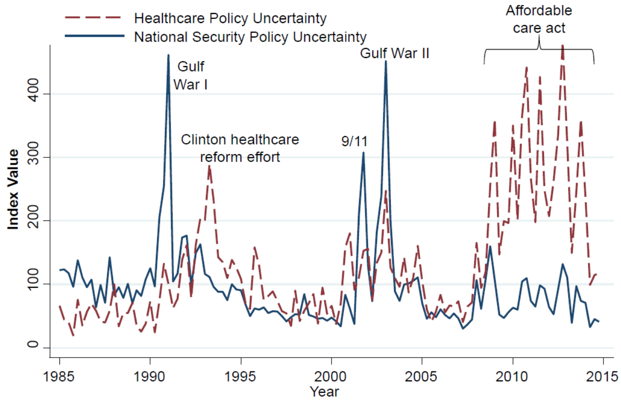
Source: Baker, Bloom, and Davis (2015)

Domain-Specific Economic Policy Uncertainty: Russia



Source: Baker, Bloom, and Davis (2015)

Economic Policy Uncertainty: Healthcare vs. National Security

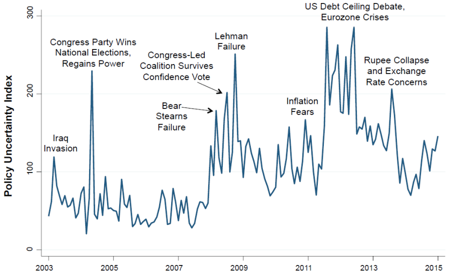


Notes: Indices reflect scaled monthly counts of articles containing the same triple as in Figure 1 ***and*** one or more terms pertaining to national security (e.g., "war", "terrorism", or "department of defense") and healthcare (e.g., "healthcare", "hospital", or "health insurance"), respectively, for the National Security and Healthcare indices. Each series is normalized to mean 100 from 1985-2009 and based on queries run Jan 18, 2015 on Access World News Newsbank newspaper archive, which covers about 1,500 US papers.

Source: Baker, Bloom, and Davis (2015)

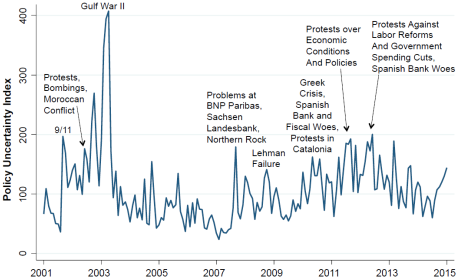
Domain-Specific Economic Policy Uncertainty

News Articles from **India**



Notes: Index reflects scaled monthly counts of articles containing 'uncertain' or 'uncertainty' or 'uncertainties' or 'uncertainties', 'economic' or 'economy', and one or more of policy-relevant terms listed for India in Appendix A. The series is normalized to mean 100 from 2003 to 2010 and based on the following newspapers: The Economic Times, Times of India, Hindustan Times, The Hindu, Financial Express, Indian Express, and the Statesman.

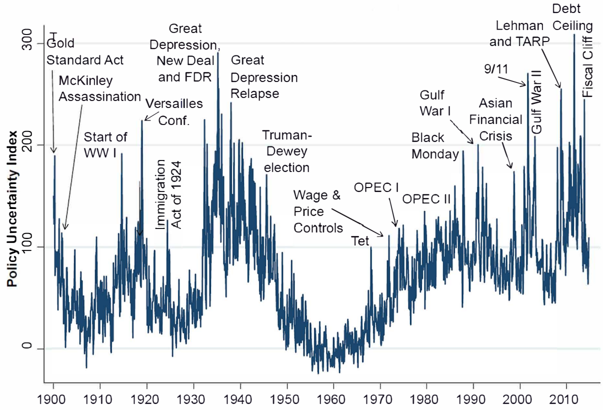
News Articles from **Spain**



Notes: Index reflects scaled monthly counts of articles containing 'uncertain' or 'uncertainty', 'economic' or 'economy', and one or more policy-relevant terms: 'tax', 'policy', 'regulation', 'spending', 'deficit', 'budget', or 'central bank'. The series is normalized to mean 100 from 2001 to 2009 and based on queries in the following newspapers: El Pais and El Mundo

Source: Baker, Bloom, and Davis (2015)

Economic Policy Uncertainty in the Past



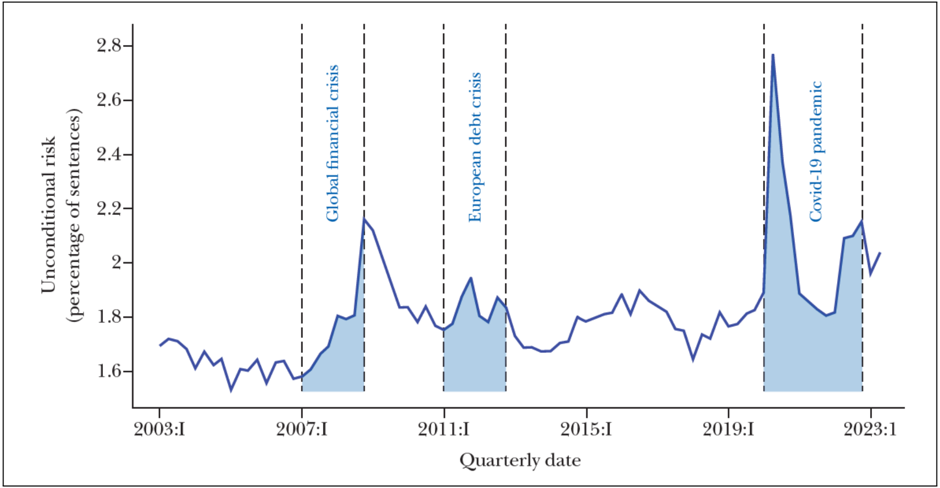
Notes: Index reflects scaled monthly counts of articles in 6 major newspapers (Washington Post, Boston Globe, LA Times, NY Times, Wall Street Journal, and Chicago Tribune) that contain the same triple as in Figure 1, except the economy term set includes "business", "commerce" and "industry" and the policy term set includes "tariffs" and "war". Data normalized to 100 from 1900-2011.

Source: Baker, Bloom, and Davis (2015)

Measuring Risk Exposure Using Text Data

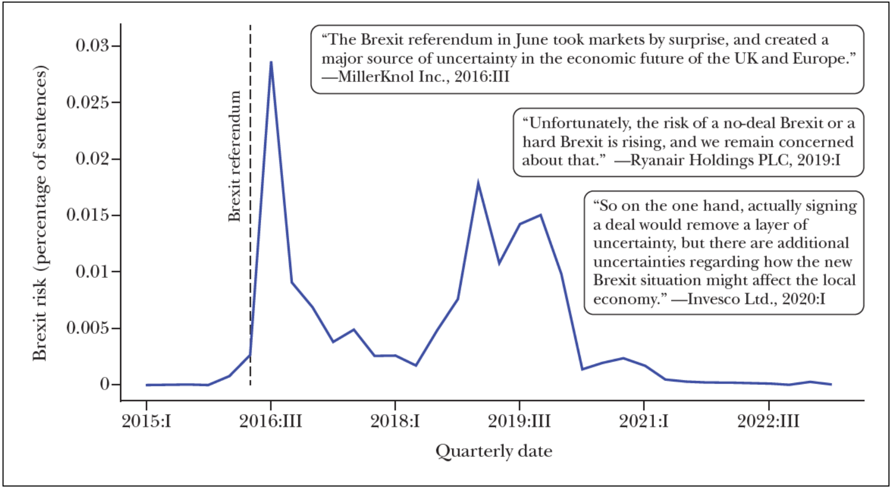
- Hassan, Hollander, Kalyani, van Lent, Schwedeler, and Tahoun (2025) build a data set of transcripts from corporate earning conference calls, mandated by the SEC from 2000
  - 379,227 conference call transcripts
  - 12,805 publicly-traded firms
  - Firms based in 89 different countries
- Analyzed at the level of individual sentences (rather than calls), using the proportion of sentences (within a call) devoted to a topic as a measure of firm attention to that topic
  - Measure a firm's exposure to risk as proportion of sentences mentioning "risk" OR "risky" or any one a set of Oxford English Dictionary synonyms for risk except "question[s]"
  - Use sentiment dictionaries to calculate topic-specific sentiment

Constructing a Measure of Risk



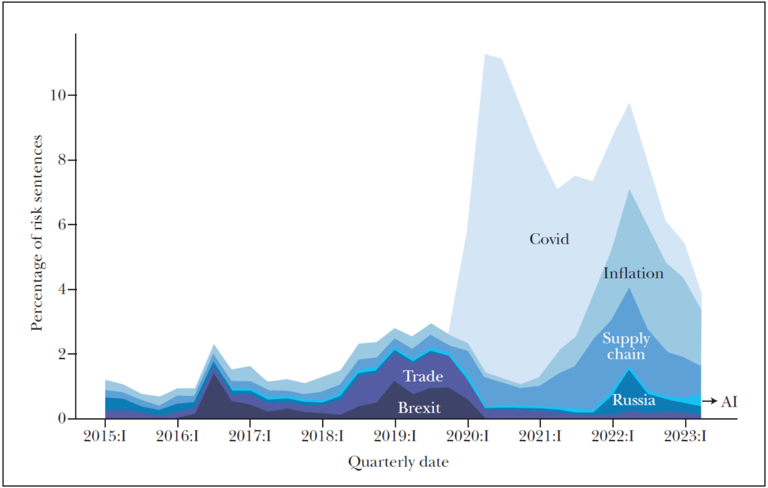
Source: Hassan, Hollander, Kalyani, van Lent, Schwedeler, and Tahoun (2025)

Firm Attention to Specific Risks



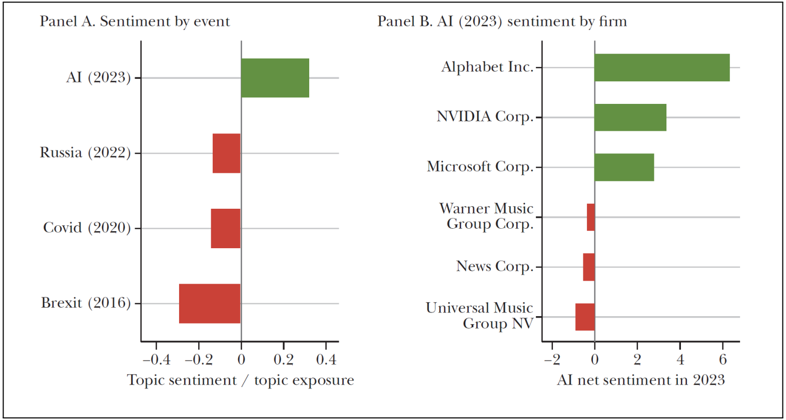
Source: Hassan, Hollander, Kalyani, van Lent, Schwedeler, and Tahoun (2025)

Decomposing Risks



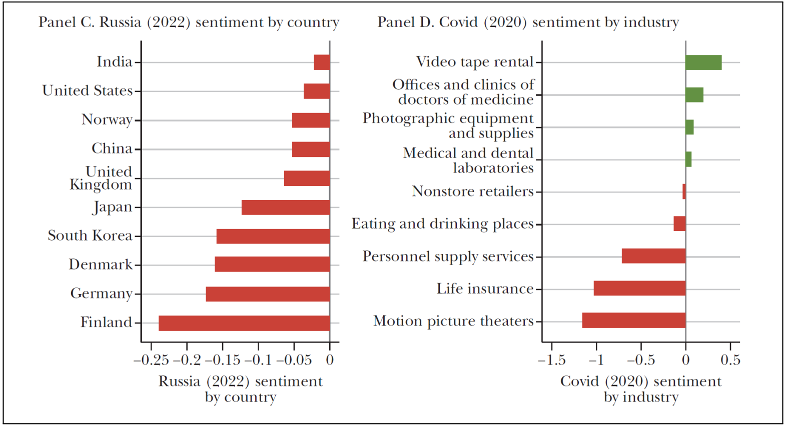
Source: Hassan, Hollander, Kalyani, van Lent, Schwedeler, and Tahoun (2025)

Topic-Specific Sentiments



Source: Hassan, Hollander, Kalyani, van Lent, Schwedeler, and Tahoun (2025)

Topic-Specific Sentiments



Source: Hassan, Hollander, Kalyani, van Lent, Schwedeler, and Tahoun (2025)

Developing a Dictionary

Dictionary development starts (and sometimes ends) with a researcher's list of key words

- Examples: Brexit, Russia
- Hassan et al. (2025) use synonyms in the (actual) dictionary

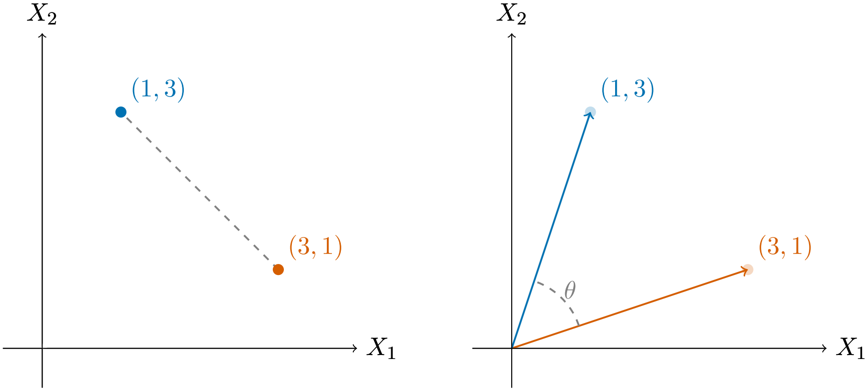
Expanding a dictionary from an initial set of key words (e.g. "climate change")

- Identify pairs of words likely to occur together
  - Climate appears with: change, cost, policy, effect, model, impact, emission, carbon, natural
- Identify highly correlated words
  - Climate most correlated with: emission, carbon, temperature, damage, extreme, natural, scenario, change, gas, disaster, global, wildfire, electricity, energy, grid, heat, will, investment

Alternative approach is to build a dictionary using an initial set of categorized documents

- Last step is always validation, irrespective of method used to construct dictionary

Euclidean Distance vs. Cosine Similarity



Measuring Document Distance Using Cosine Similarity: Example

6 NBER working papers from 2024 mention Japan or Uganda in the title:

- **WP 32142:** "Invoice Currency Choice in Intra-Firm Trade: A Transaction-Level Analysis of Japanese Automobile Exports"
- **WP 32170:** "Do Information Frictions and Corruption Perceptions Kill Competition? A Field Experiment On Public Procurement in Uganda"
- **WP 32645:** "Intrahousehold Welfare: Theory and Application to Japanese Data"
- **WP 32744:** "The Nexus between Long-term Care Insurance, Formal Care, Informal Care, and Bequests: The Case of Japan"
- **WP 32785:** "The Returns to Skills During the Pandemic: Experimental Evidence from Uganda"
- **WP 32910:** "Invoicing Currency and Exchange Rate Pass-Through in Japanese Imports: A Panel VAR Analysis"

Measuring Document Distance Using Cosine Similarity: Example

	32142	32170	32645	32744	32785	32910
32142	1	0	0.021	0	0	0.1019
32170	0	1	0	0	0.0503	0
32645	0.021	0	1	0	0	0.0235
32744	0	0	0	1	0	0
32785	0	0.0503	0	0	1	0
32910	0.1019	0	0.0235	0	0	1

Cosine similarity based on TFIDF after removing stopwords

Cosine Similarity Among NBER Working Papers

1,112 NBER working papers from 2024 $\Rightarrow$ 1,235,432 pairwise combinations of papers

- 44,980 pairs have cosine similarity of 0
- Mean across all pairs of distinct papers: 0.0160
  - No obvious benchmark for what we should expect to occur by chance...
- 90^{th} percentile: 0.035 $\Rightarrow$ most papers are different
- 9 pairs of papers have cosine similarity above 0.5 (indicating very similar text/content)

When Cosine Similarity = 0.773

WP 32101

Title: CEO Compensation and Cash-Flow Shocks: Evidence from Changes in Environmental Regulations

Authors: Seungho Choi, Ross Levine, Raphael Jonghyeon Park, Simon Xu

Abstract: This paper investigates how shocks to expected cash flows influence CEO incentive compensation. Exploiting changes in compliance with environmental regulations as shocks to expected future cash flows, we find that adverse shocks typically prompt corporate boards to re-calibrate CEO compensation to reduce risk-taking incentives. However, this pattern is not uniform. Financially distressed firms exhibit milder reductions in compensation convexity, with some even increasing it, suggesting a "gambling for resurrection" strategy. Moreover, the strength of corporate governance influences shareholders' capacity to align...

WP 32663

Title: CEO Compensation and Adverse Shocks: Evidence from Changes in Environmental Regulations

Authors: Seungho Choi, Ross Levine, Raphael Park, Simon Xu

Abstract: Although corporate finance theory suggests how adverse shocks influence shareholder preferences toward corporate risk-taking and executive compensation, few researchers explore this relationship empirically. We construct a firm-year measure of unexpected shocks to environmental regulatory stringency. We find that adverse environmental regulatory shocks typically prompt corporate boards to reduce the risk-taking incentives of CEO compensation. However, this pattern is not uniform. Financially distressed firms exhibit milder reductions in compensation convexity, with some even increasing it, suggesting...

When Cosine Similarity = 0.612

WP 32939

Title: The Effect of Inflation Uncertainty on Household Expectations and Spending

Authors: O. Kostyshyna, L. Petersen

Abstract: We use a new Canadian household survey to examine how inflation uncertainty influences inflation expectations and spending. Through randomized information interventions, we provide inflation statistics with or without second moments, creating variations in households' inflation uncertainty. All information types effectively lower inflation expectations and uncertainty. While communicating inflation uncertainty does not affect expectations or uncertainty levels, it increases the probability assigned to expected inflation near communicated ranges. Using Nielsen IQ Homescanner...

WP 33014

Title: The Causal Effects of Inflation Uncertainty on Households' Beliefs and Actions

Authors: D. Georgarakos, Y. Gorodnichenko, O. Coibion, G. Kenny

Abstract: We implement a survey-based randomized information treatment that generates independent variation in the inflation expectations and the uncertainty about future inflation of European households. This variation allows us to assess how both first and second moments of inflation expectations separately affect subsequent household decisions. We document several key findings. First, higher inflation uncertainty leads households to reduce their subsequent durable goods purchases for several months, while a higher expected level of inflation increases them. Second, an increase...

Identifying Novel* Papers

Define a paper-level measure of content uniqueness: mean cosine similarity with other papers

- "Content Moderation with Opaque Policies" by Scott Kominers and Jesse Shapiro
  - **Abstract:** A sender sends a signal about a state to a receiver who takes an action that determines a payoff. A moderator can block some or all of the sender's signal before it reaches the receiver. When the moderator's policy is transparent to the receiver, the moderator can improve the payoff by blocking false or harmful signals. When the moderator's policy is opaque, however, the receiver may not trust the moderator. In that case, the moderator can guarantee an improved outcome only by blocking signals that enable harmful acts. Blocking signals that encourage false beliefs can be counterproductive.
- "Business Failures by Branch of Business in the United States, 1895 to 1935: A Statistical History" by Gary Richardson, Marco Del Angel and Michael Gou
  - **Abstract:** Dun's Review began publishing monthly data on bankruptcies by branch of business during the 1890s. Those series evolved through many iterations. This essay reconstructs the series from 1895 to 1935 and discusses how it can be used for economic analysis.

Identifying Innovation

Papers that are unique in 2024 may not be novel – they might actually be outdated

- Document-based measures of novelty are typically based on long-running data sets
  - All NBER working papers or other journal publications, patents, etc.

Innovations are novel, and they also have substantial impacts on future ideas/technology/etc

- Innovative papers should be distinct from earlier papers, but similar to later ones

Kelly, Papanikolaou, Seru, and Taddy (2021) propose a measure of innovation based on a modified TFIDF, where the IDF weightings are relative to **prior** documents in the series

- Use text data on patents to estimate novelty
- Calculate document similarity using IDF weightings from earlier document

Summary: Economists Using Text as Data

Quantitative social scientists almost always use text in conjunction with document metadata

- NBER programs, dates, author/speaker gender or political party, etc.
- Text data $\rightarrow$ variables measuring new/unobserved constructs

Most common approaches (in economics) either involve dictionaries or document distance

- Identify documents/text associated with specific topics (e.g. Brexit)
- Characterize tone or subtext (which may not be explicitly stated)
  - Positive/negative, uncertainty, and risk are most common
- Document distance proxies for distance between entities associated with documents
  - Group firms, researchers, patents, etc. into (unobserved) categories
  - Generate continuous measures of distance (e.g between firms, syllabi and research)
  - Identifying innovation (e.g. through patents)