

A Thousand Words Tell More than just Numbers: Financial Crises and Historical Headlines

Kim Ristolainen, Tomi Roukka and Henri Nyberg

RiskLab/BoF/ESRB Conference on Systemic Risk Analytics
6.5.2022

Motivation

- Historical macroeconomic datasets by Jordà, Schularick and Taylor (2016a) and Baron, Verner and Xiong (2020) have generated new interest in crisis indicators.
 - Much longer time coverage has given credibility to the economic analyses.
 - Large credit growth (Jordà and Taylor, 2012) interacted with rising asset prices Greenwood et al. (Forthcoming), private sector debt (Jordà, Schularick and Taylor, 2016b), and mortgage lending and house prices (Jordà, Schularick and Taylor, 2015).
- As the predictability of crisis is still modest, a possible source for new insights could be text data as it can match the country and time coverage of the mentioned datasets and hold information on issues that current statistical datasets do not include.

The purpose of the paper

1. We assemble a **panel dataset that aggregates historical newspaper information** covering economy related issues. The country coverage of 17 developed countries and the time interval 1870-2016 matches the ones of the macrohistory datasets so that it can be used together with this already existing economic information. [Collection of topics](#) [Topic frequency series](#)
2. We show that **newspaper texts have important additional predictive power to historical crisis** episodes on top of usual indicators. In addition, the text information alone outperforms the usual macroeconomic and financial variables in prediction accuracy. [Crisis prediction improvement](#) [Crisis case studies](#)

Text data from ProQuest Historical Newspapers

- Extensive collection of historical newspaper articles from *the New York Times*, *The Washington Post*, *The Globe and Mail*, *The Times of India*, *The Guardian*, and *The Wall Street Journal* for the **time interval 1870 – 2016**.
- The most relevant 200 economy related articles for each newspaper-year-country combination that mention at least one economic keyword and country name.
- The titles are used in the following analysis as titles efficiently summarise the content of the news articles whereas the full texts include additional content that can in many cases be useless.
- We make the corpus **time - and country invariant** by keeping the words that are mentioned at least once in every 50 year period in the sample and at least once in half of the countries.

Quantifying newspaper data

- We use the Latent Dirichlet Allocation (LDA) topic model originally developed by Blei, Ng and Jordan (2003), which is an unsupervised learning model that uncovers structures in the text data without having any knowledge of the data labels.
- Each article is assumed to be generated by a generative process and to be a combination of multiple topics k . Each topic is a distribution of words from a fixed vocabulary.
- Gibbs sampling is used to estimate the **topic distributions** θ **for each article** and the **word distributions** β **for each topic**. In addition, the number of topics K is optimized via grid search.

What do we get from β ?

Topic	Most frequent and exclusive words
1	gain-total-record-purchases-larger-farm-jump
4	economy-growth-spur-signs-grows-slowness-pace
5	central-bank-governor-institution-raises-eyes-credit
8	confidence-sentiment-game-farmer-mood-caution-spree
9	banks-savings-accounts-cash-lending-institutions-fail
12	investors-briefs-look-wary-feel-yields-interest
16	inquiry-trader-involves-investigating-charges-probe-bribery
24	rate-jobless-unemployment-highest-lowest-discount-drops
25	economic-ties-problems-impact-policies-stability-summit
26	prospects-mail-bright-letter-poor-brighter-good
27	expect-economists-predict-believe-optimistic-forecast-weighted
28	banking-system-giant-scandal-beats-sector-crises
30	curbs-import-restrictions-barriers-retaliation-quota-eased
34	taxes-income-fixed-taxation-affects-levies-billions
37	labor-party-unions-leader-strikes-reds-vote
38	consumer-inflation-check-leaving-wage-restraint-perils
47	wine-wines-liquor-grievances-spirits-beer-whiskey

What do we get from θ ?

$$TF_{k,t,c} = \frac{1}{L_{t,c}} \sum_{l=1}^{L_{t,c}} \theta_{l,k} \quad (1)$$

- We form the annual topic frequency series ($TF_{k,t,c}$), by averaging the proportions of each topic k across the $L_{t,c}$ articles that are published in year t and mention country c .
- Do the topics make sense? Can the most common words be seen to form a coherent understandable topic?
- Do the topic frequencies make sense? Do the periods of large attention to specific topic correlate with historical events?

Topic labels

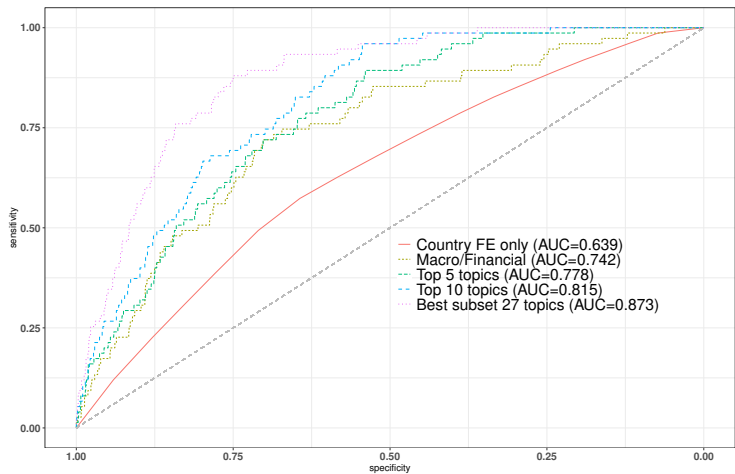
Purpose



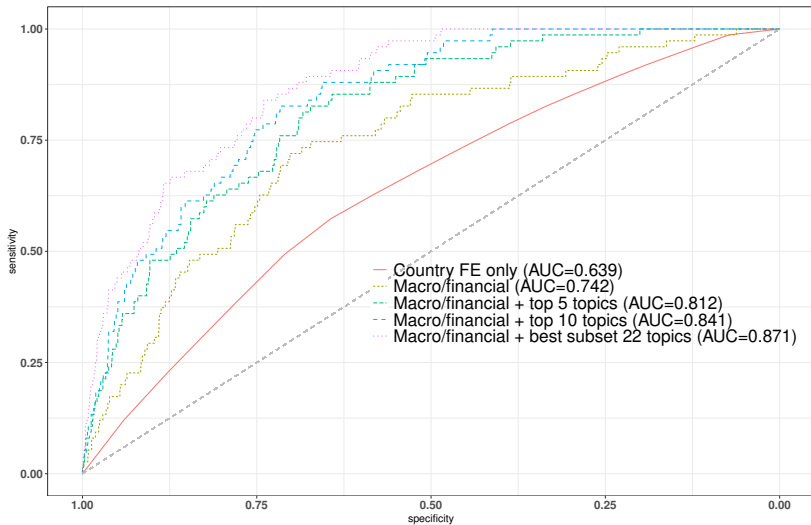
Crisis prediction model

- Annual data for 17 developed countries and 75 crises episodes for the last 150 years.
- A panel logit regression where the explanatory variables are included in lagged 3-year moving averages.
- The baseline model: *change in credit to GDP, return on equity, real GDP growth, inflation.*
- *Change in house prices, bank leverage, non-core funding ratio and liquidity* are included in a robustness analysis as they remove a significant amount of crises from the data set.
- We include country fixed effects and use Driscoll-Kraay (1998) clustered standard errors.
- A forward stepwise selection procedure to select topic frequency variables to the model.

In sample prediction accuracy - text vs. macro/financial predictors



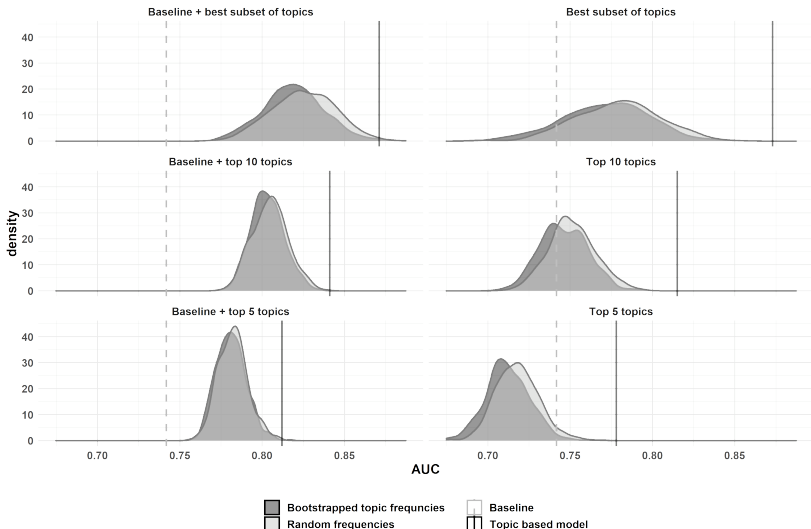
In sample prediction accuracy - text **with** macro/financial predictors



DeLong, DeLong and Clarke-Pearson (1988) test for difference of AUC statistics.

	AUC	Diff.to.baseline	p.value	Diff.upper	p.value
Country FE only	0.639				
Macro/financial	0.742				
Macro/financial + top 5 topics	0.812	0.070	0.000		
Macro/financial + top 10 topics	0.841	0.099	0.000	0.028	0.002
Macro/financial + best subset 22 topics	0.871	0.129	0.000	0.030	0.002
Top 5 topics	0.778	0.036	0.053		
Top 10 topics	0.815	0.074	0.002	0.038	0.010
Best subset 27 topics	0.873	0.131	0.000	0.058	0.000

Robustness analysis - overfitting



Robustness analysis - overfitting

Topic data	Model	$\frac{\#AUC_U > AUC_{BL}}{\#Total}$	$\frac{\#p < 0.05: AUC_U > AUC_{BL}}{\#Total}$	$\overline{AUC_U - AUC_{BL}}$	$\frac{\#AUC_U > AUC_{TR}}{\#Total}$	$\overline{AUC_U - AUC_{TR}}$
Bootstrapped	BL + best subset	1.000	0.151	0.077	0.006	-0.052
Random	BL + best subset	1.000	0.156	0.083	0.004	-0.046
Bootstrapped	BL + top 10	0.932	0.155	0.061	0.000	-0.038
Random	BL + top 10	0.978	0.158	0.063	0.000	-0.036
Bootstrapped	BL + top 5	1.000	0.091	0.039	0.002	-0.031
Random	BL + top 5	1.000	0.087	0.040	0.000	-0.030
Bootstrapped	Best subset	0.876	0.086	0.030	0.000	-0.101
Random	Best subset	0.932	0.090	0.037	0.000	-0.094
Bootstrapped	Top 10	0.556	0.082	0.004	0.000	-0.069
Random	Top 10	0.724	0.076	0.009	0.000	-0.064
Bootstrapped	Top 5	0.016	0.000	-0.030	0.000	-0.066
Random	Top 5	0.042	0.000	-0.024	0.000	-0.060

Robustness analysis

Panel A	Reinhart-Rogoff			Jorda-Schularick-Taylor			Excl. panics		
	AUC	p	p	AUC	p	p	AUC	p	p
Country FE only	0.591			0.616			0.616		
Macro/financial	0.673			0.741			0.893		
Macro/financial + top 5 topics	0.778	0.000		0.806	0.002		0.931	0.022	
Macro/financial + top 10 topics	0.790	0.000	0.097	0.819	0.001	0.039	0.973	0.002	0.006
Macro/financial + best subset topics	0.816	0.000	0.008	0.859	0.000	0.002	0.990	0.001	0.020
Top 5 topics	0.732	0.025		0.752	0.356		0.901	0.384	
Top 10 topics	0.789	0.000	0.000	0.802	0.023	0.002	0.963	0.016	0.008
Best subset topics	0.827	0.000	0.005	0.869	0.000	0.000	0.995	0.001	0.018
Crises		77			61			13	
Obs		1680			1751			1587	
Panel B	Excl. GFC			Excl. Great Depression			Out-of-sample		
	AUC	p	p	AUC	p	p	AUC	p	p
Country FE only	0.658			0.661			0.520		
Macro/financial	0.757			0.762			0.622		
Macro/financial + top 5 topics	0.825	0.000		0.822	0.001		0.748	0.000	
Macro/financial + top 10 topics	0.849	0.000	0.007	0.842	0.000	0.032	0.759	0.000	0.176
Macro/financial + best subset topics	0.886	0.000	0.002	0.875	0.000	0.007	0.760	0.000	0.460
Top 5 topics	0.797	0.057		0.790	0.119		0.679	0.068	
Top 10 topics	0.824	0.005	0.035	0.830	0.006	0.014	0.735	0.002	0.003
Best subset topics	0.894	0.000	0.000	0.879	0.000	0.001	0.733	0.005	0.525
Crises		55			65			36	
Obs		1550			1632			1086	

Robustness analysis

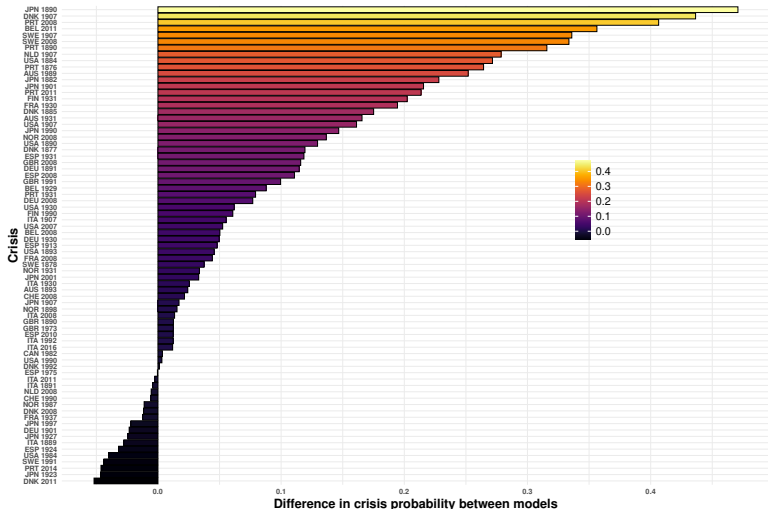
Panel A	R-zone			House			Balance sheet		
	AUC	p	p	AUC	p	p	AUC	p	p
Country FE only	0.646			0.635			0.639		
Macro/financial	0.761			0.742			0.768		
Macro/financial + top 5 topics	0.857	0.000		0.812	0.001		0.829	0.000	
Macro/financial + top 10 topics	0.883	0.000	0.011	0.847	0.000	0.002	0.854	0.000	0.004
Macro/financial + best subset topics	0.916	0.000	0.012	0.873	0.000	0.021	0.876	0.000	0.004
Top 5 topics	0.823	0.070		0.776	0.092		0.789	0.178	
Top 10 topics	0.868	0.009	0.015	0.823	0.002	0.017	0.824	0.012	0.016
Best subset topics	0.897	0.001	0.025	0.872	0.000	0.004	0.874	0.000	0.000
Crises		36			55			68	
Obs		949			1408			1608	
Panel B	Excl. Dec. news			Excl. Q4 news			Min. 10 news per year		
	AUC	p	p	AUC	p	p			
Country FE only	0.638			0.638			0.650		
Macro/financial	0.740			0.739			0.732		
Macro/financial + top 5 topics	0.809	0.000		0.804	0.000		0.802	0.001	
Macro/financial + top 10 topics	0.833	0.000	0.002	0.825	0.000	0.004	0.823	0.000	0.025
Macro/financial + best subset topics	0.857	0.000	0.004	0.852	0.000	0.001	0.866	0.000	0.001
Top 5 topics	0.778	0.050		0.772	0.081		0.773	0.045	
Top 10 topics	0.810	0.003	0.015	0.808	0.005	0.011	0.800	0.007	0.042
Best subset topics	0.856	0.000	0.000	0.852	0.000	0.002	0.865	0.000	0.000
Crises		75			74			58	
Obs		1748			1304				

Leading indicator topics - labels and coefficients

Panel A: Model with macro/financial and the best subset of 22 topics

Topic	Label	Coefficient	Significance	Most common and exclusive words
30	International trade restrictions - Imports	-0.725	**	curbs-import-restrictions-barriers-retaliation-quota-eased
31	International trade restrictions - Exports	-0.778	***	drive-export-prepares-uranium-doubtful-world-wide-approved
92	Commodities	0.623	**	product-domestic-gross-butter-enlarged-dairy-type
3	Industrial production	-0.582	***	output-production-industrial-speed-newsprint-aluminum-copper
83	Change in international trade barriers	0.356	*	duty-materials-items-cloth-bans-autos-wool
97	Relief and help	-1.288	**	relief-busy-unemployed-fund-cross-idle-glad
23	Money markets	-0.328	*	money-current-call-events-tight-developments-speculation
11	Political proposal lifecycle	1.3	***	proposals-representatives-resistance-pledges-restored-freedom-cease
99	Crises and dangers	-0.496	*	coal-fuel-acute-nears-miners-middle-east
88	Foreign investment	0.646	***	like-potential-climate-investment-foreseen-shape-welcome
25	Economic policy agreements and cooperation	-0.215	**	economic-ties-problems-impact-policies-stability-summit
58	National supplies shortage	-1.15	*	needs-food-supply-requirements-adequate-assure-material
77	Promises	0.662	*	promise-given-failing-breaks-authorized-drawn-advice
82	Diplomatic relations	0.462	**	reply-grants-granted-powers-issued-encouragement-recognize
44	Firm level sales and purchases	0.329	***	importer-shut-firm-option-buys-sold-owner
96	Factory problems	-0.555	*	resources-factories-wealth-rich-natural-vast-development
101	Matches and fights	0.439	*	prospect-hill-sports-invasion-cheered-complications-dark
45	High level and government economists	0.193	*	economics-politics-sights-lesson-lessons-search-nobel
40	Analysis and outlook	-1.105	*	improved-avoided-alarm-tumble-hurting-term-response
104	Commodity producers	-1.429	*	producers-pleased-longer-tactics-importers-regret-dealers
21	Plans and ideas	0.512	*	idea-whose-forward-importing-stress-revived-accept
55	Noble metals	0.259	*	gold-metal-standard-silver-reserves-reserve-shipment

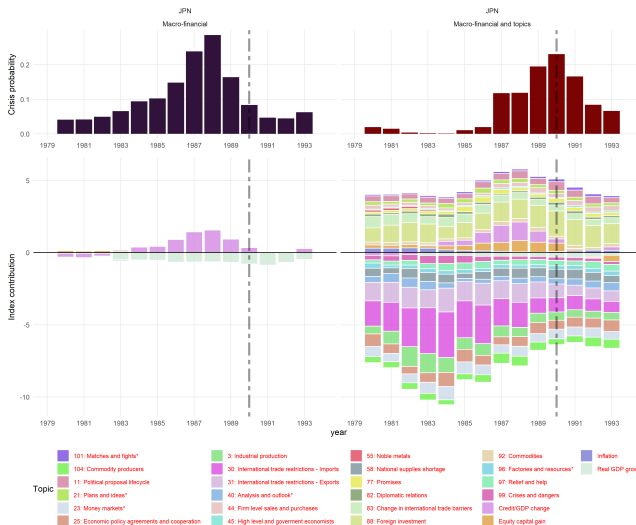
Cases when text improves crisis prediction?



Case study: Japan 1990

- The strategic policies of Japan's government, the Ministry of International Trade and Industry (METI), and the Ministry of Finance together with the so-called *keiretsu* way of forming company networks have been argued to be some of the key factors responsible for the country's economic success.
- The US accused Japan of keeping the yen undervalued relative to the dollar, running a persistent trade surplus with the US, and using unique Japanese organizations (*keirutzu*) that, according to competing U.S. exporting firms, create trade barriers and unfair competition.

Case study: Japan 1990



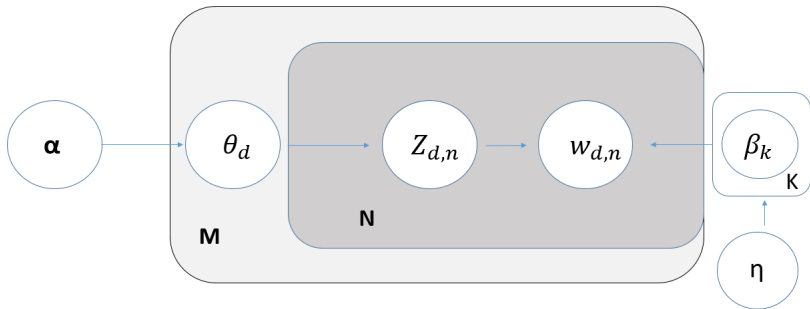
Case study: Japan 1990

- The so-called Plaza Accord was signed in 1985 to depreciate the value of the dollar relative to the currencies of France, West Germany, the United Kingdom and Japan. This agreement led to significant appreciation of the yen with respect to the dollar, which later negatively affected the competitiveness of Japanese exporters.
- The Bank of Japan reacted to the appreciation of yen by raising interest rates, which has been argued to be the trigger of the collapse of asset prices (Fung et al., 2020).
- The biggest contributor to the crisis probability during the peak year relative to the preceding years is the decreased frequency or the absence of the *international trade restrictions on imports* topic, which might capture the U.S.-Japan trade conflicts of the 1980s.

Conclusions

- We extracted time series of different topic frequencies in historical economic news articles for a panel of 17 developed countries and 150 years.
- The estimated topics capture historical events and form coherent entities.
- Models with text topic information improve prediction accuracy significantly on top of usual indicators.
- The topics seem to relate with phenomena and factors that have been discussed in public and in individual studies to have caused or contributed at least partly to a specific historical crisis episode.
- These topics relate to important changes in economic policies, external shocks, detailed information on production or output, and economic atmosphere .

LDA - Graphical representation



- Lets assume that there is a corpus of M news titles with N words in each title from a vocabulary of size V and a total number of K topics.

Dirichlet distribution

Small α Large α

The generative process

- For title d there is a distribution $\theta_d \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_K)$ of topics and for each topic k there is a word distribution $\beta_k \sim \text{Dirichlet}(\eta_1, \eta_2, \dots, \eta_V)$.
 - If we would have $K = 3$ were the corresponding topics would be about **cats**, **dogs** and **fish**, then a title ***Dogs are nice animals, but they do not like cats*** could have a θ equal to $(0.6, 0.4, 0.0)$.
 - If the corpus vocabulary would have the words ***dog, cat, nice, bad, like, fish, car, nature, puppy***, then β_{cat} could be $(0.9, 0.3, 0.5, 0.4, 0.6, 0.1, 0.1, 0.2, 0.9)$
- The actual title d is generated so that for each word position j in the title a topic $z_{d,j}$ is generated from the topic distribution so that $z_{d,j} \sim \text{Multinomial}(\theta_d)$.
- Next a word $w_{d,j}$ is generated for each word position j in the title from the word distribution of assigned topic $z_{d,j}$ so that $w_{d,j} \sim \text{Multinomial}(\beta_z)$.

Estimating the parameters

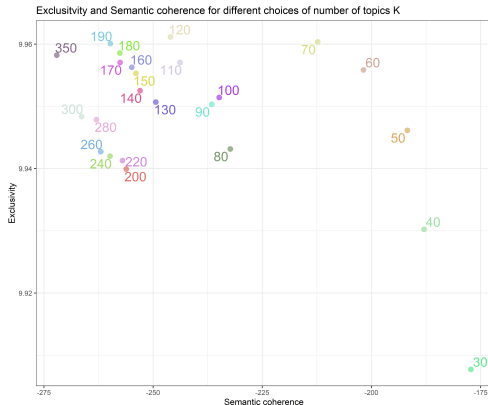
$$P(\theta, \beta, Z, W) = \prod_{k=1}^K P(\beta_k | \eta) \prod_{d=1}^M P(\theta_d | \alpha) \prod_{i=1}^N P(z_{d,n} | \theta_d) P(w_{d,n} | \beta, z_{d,n})$$

- We only know $w_{i,j}$ and the number of topics K as it is predetermined by the researcher. Everything else is unknown.

$$P(\theta, \beta, Z | W) = \frac{P(\theta, \beta, Z, W)}{P(W)}$$

- We can estimate the posterior distribution with Gibbs Sampling. This procedure searches for the parameter values that make the generative process generate newspaper titles that are as close to the ones that we observe.

Number of topics K ?



- Models' semantic coherence (the most probable words in a given topic frequently co-occur together), exclusivity (high probability words of a topic are not likely to be high probability words of another topic) and human judgement to select the optimal K

Dirichlet distribution

- Multivariate generalization of Beta distribution when $K > 2$
- For $K = 2$ the Dirichlet distribution is a beta distribution
- Probability density function for the Dirichlet distribution

$$p(x|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} x_1^{\alpha_1-1} \dots x_k^{\alpha_k-1}$$

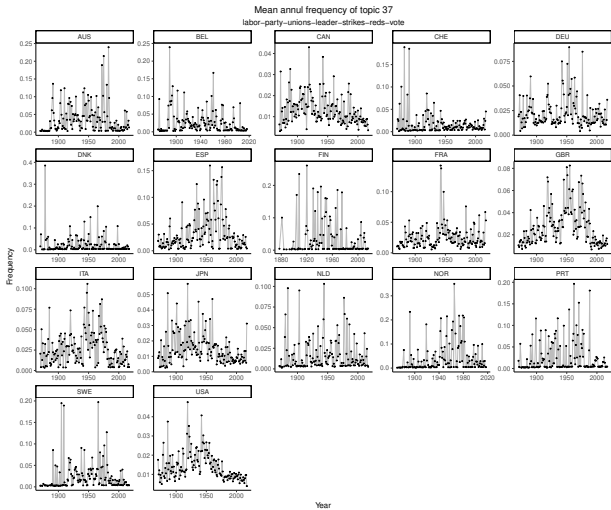
LDA

Most representative titles for specific topics

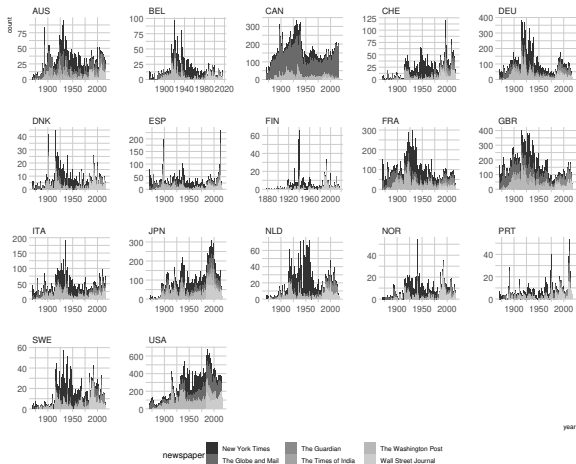
frequency	Title
	Topic 1: gain-total-record-purchases-larger-farm-jump
92 %	canada maintains balance in trade: favorable november figures show \$8,852,377 over imports.
92 %	farm imports from canada soar for july: total of \$9,835,000 more than double figures for last year.
91 %	the british grain trade.: few sales of wheat and diminished deliveries—effect of the american surplus.
	Topic 9: banks-savings-accounts-cash-lending-institutions-fail
90 %	billions of losses at some of japan big banks: billions in losses at banks in japan
91 %	japan savers may find they're paying banks: in japan, savers could find they're paying the banks
92 %	u.s. banks balk at long terms for polish debt: u.s. banks balk at extension of polish debt
	Topic 37: labor-party-unions-leader-strikes-reds-vote
96 %	german code ends old labor rights: annihilates unions, prohibits strikes and does away with collective bargaining...
	Topic 56: moves-currency-falls-devaluation-single-traders-closes
93 %	u.s. dollar slumps in german trading amid mark rumors: dollar slumps in west germany a mid mark revaluation rumors
94 %	dollar plagued by peso: dollar dips as the peso falls again mexico crisis hurts u.s. currency value dollar slips as the peso continues sharp fall
	Topic 52: deficit-billion-widens-payments-budget-surplus-widened
94 %	merchandise trade deficit hits 3.5-year low: trade deficit declines by \$3.6 billion u.s. merchandise trade deficit (surplus) with selected nations
95 %	'85 trade deficit is worst ever: \$148.5 billion gap expected to intensify pressures on congress '85 u.s. trade deficit hits \$148 billion; protectionist pressures seen intensifying
94 %	3d quarter deficit hit \$12.1 billion: u.s. balance of payments deficit reaches \$12.1 billion in 3d quarter

Most common words

What do we get from θ ?



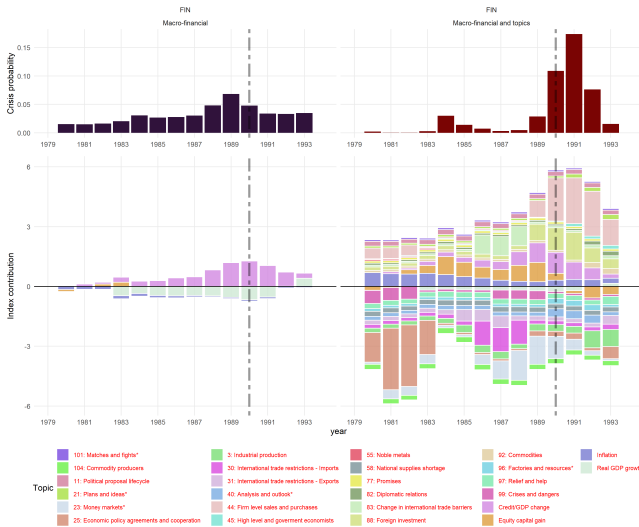
Number of article titles per country-year



Finland 1990

- Financial liberalization in the early 1980s started a large inflow of foreign capital to the country resulting into credit expansion, increased spending, investments, and asset prices.
- Due to the loss of competitiveness and the central bank defending the pegged exchange rate, interest rates increased significantly in 1990. The high rates were unbearable for many households and firms as they had become highly indebted during the boom.
- In addition, trade between the Soviet Union and Finland collapsed when the former fell apart. Finally, as the economic output declined, and asset prices collapsed, a large number of bankruptcies occurred, and loan losses began to accumulate for banks. The government ended up taking control of the Savings Bank Group, and other banks were recapitalized (Kiander and Vartia, 2011).

Finland 1990



Finland 1990

- The baseline model gives an elevated crisis probability for the crisis year and several preceding years, but for the topic-based model, the increase in crisis probability is much larger at the end of the 1980s.
- The increase of *firm-level sales and purchases* and *foreign investments* topics and the change in credit-to-GDP are the largest contributors to the increase in crisis probability for that year.
- Relative to previous years, the *crises and dangers*, *industrial production*, and *international trade restrictions on imports* topics were less frequent in the newspaper article titles. These factors likely capture the collapse of trade with the Soviet Union and problems in the corporate sector that ultimately partly caused the banks to have massive amounts of defaulted loans.