Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Reading Between the Lines: Quantitative Text Analysis of Banking Crises

Emile du Plessis

2022 Annual ASSA Meeting

7-9 January 2022

Motivation

Digital transformation

Growth in digital content, and text-mining programmes

Applications

Results across literature studies emphasize its usefulness

Untapped data sources

Promising data sources and novel methods

Severe ramifications

Negative impact of banking crises on society

- Text data and approach
- Content analysis
- Index 1: Wordscores
- Index 2: Wordfish
- Index 3: Algorithmic text analysis
- Index 4: Sentiment index
- Index 5: Lexicon and sentiment index
- Robustness check: German indices

Quantitative text analysis as leading indicator

Five indicators are constructed to forecast banking crises. Findings from Granger causality highlight leading indicators status of the **Banking Crisis Lexicon Index**, up to two years preceding a crisis. While the aggregated **Sentiment Index** constitutes a coincidental indicator, for developed economies it is a short-term leading indicator. A **combined lexicon and sentiment index** exhibit solid forecasting performance. Statistical models **Wordscores** and **Wordfish** are introduced to study banking crises and underscore crisis classification strength.

Impact of Banking Crises

- Severe ramifications of Global Financial Crisis and ensuing banking difficulties.
- Recurring phenomenon and agnostic to development level (Reinhart and Rogoff, 2009).
- Half result in economic depression (Barro and Ursúa, 2009).
- High resolution cost (Laeven and Valencia, 2010).

Empirical Literature

- Laver et al. (2003) and Benoit and Laver (2003) advance statistical model Wordscores
- Proksch and Slapin (2008, 2009) develop computer-based model Wordfish
- Transparency in the US Federal Reserve minutes (Acosta, 2014; Hansen et al, 2014)
- Financial reports (Kloptchenko et al., 2004; Hoberg and Phillips, 2010)
- Sentiment and stock market returns (Tetlock, 2007; Loughran and Mcdonald, 2011; Heston and Sinha, 2017; Calomiris and Mamaysky, 2018)
- Economic policy uncertainty index (Baker et al., 2013)
- Emotion index for real GDP growth forecast (Nyman et al., 2015)
- Chinese labour market conditions index (Bailliu et al., 2018)
- Co-movement of a sentiment indicator and financial crises (Püttman, 2018)
- Textual data using machine learning to forecast financial crises (Chen et al., 2019)

Methodology, Text Data and Approach

- Five banking crisis indices including algorithmic and sentiment analysis and statistical models
- Similarity of themes between texts and key themes within a volume of texts
- Thomson-Reuters News archive ▷ reuters.com
  - Large volume featuring over 19 million news articles
  - International coverage allows applicability across multiple countries
  - Newsfeeds used by broadcasters and publishers, accessible to analysts and general public
  - Source is bounded to the News archive
  - Date spans the period 1987 to 2019
  - Language is constraint to English
  - Region include an aggregated global configuration and separate selection of 23 individual countries, respectively 18 developed economies and 5 emerging markets
- *Finanz und Wirtschaft* ▷ fuw.ch
  - Featuring as a robustness check, a separate source is introduced in translated format, which sets the language to German and period from 1995 to 2019
- Both sources are accessed through Dow Jones (Factiva) and specialise in business and financial reporting as opposed to opinion pieces and editorial commentary

Countries and Crisis Years

- Identification and dating of banking crises (Reinhart and Rogoff, 2009)

  1. the closure, merger or takeover of a financial institution by the government or

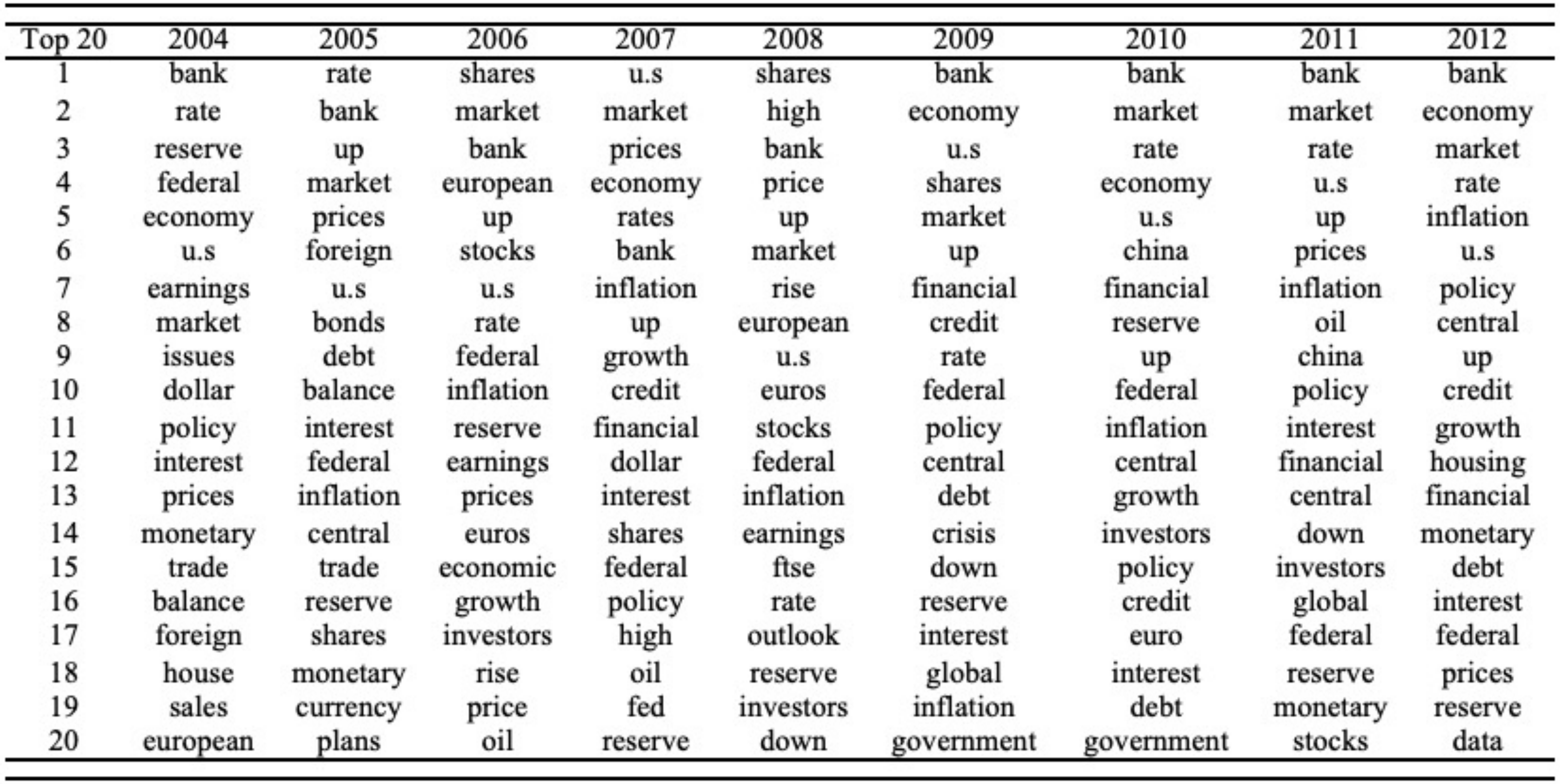
  2. the provision of financial assistance to a financial institution by the government

Country	Crises Years	Control Years
United States	2007 - 2010	-
United Kingdom	2007 - 2014	-
Austria	2008 - 2011	-
Belgium	2008 - 2014	-
Denmark	2008 - 2014	-
France	2008 - 2014	-
Germany	2008 - 2010	-
Italy	2008 - 2014	-
Netherlands	2008 - 2014	-
Sweden	2008 - 2010	-
Canada		2005 - 2013
Japan		2005 - 2013
Greece	2008 - 2014	-
Ireland		2005 - 2013
Portugal	2008 - 2014	-
Spain	2008 - 2014	-
Australia		2005 - 2013
New Zealand		2005 - 2013
South Africa		2005 - 2013
Mexico		2005 - 2013
Thailand		2005 - 2013
Czech Republic		2005 - 2013
Poland		2005 - 2013

Pre-Processing and Term Document Matrix

- 27,408 individual terms
- Dimensionality reduction
- Pre-processing techniques
  - Stop words
  - Case folding
  - Stemming
- Sparsity (0.5 percent)

▶ Word Cloud



Top 20	2004	2005	2006	2007	2008	2009	2010	2011	2012
1	bank	rate	shares	u.s	shares	bank	bank	bank	bank
2	rate	bank	market	market	high	economy	market	market	economy
3	reserve	up	bank	prices	bank	u.s	rate	rate	market
4	federal	market	european	economy	price	shares	economy	u.s	rate
5	economy	prices	up	rates	up	market	u.s	up	inflation
6	u.s	foreign	stocks	bank	market	up	china	prices	u.s
7	earnings	u.s	u.s	inflation	rise	financial	financial	inflation	policy
8	market	bonds	rate	up	european	credit	reserve	oil	central
9	issues	debt	federal	growth	u.s	rate	up	china	up
10	dollar	balance	inflation	credit	euros	federal	federal	policy	credit
11	policy	interest	reserve	financial	stocks	policy	inflation	interest	growth
12	interest	federal	earnings	dollar	federal	central	central	financial	housing
13	prices	inflation	prices	interest	inflation	debt	growth	central	financial
14	monetary	central	euros	shares	earnings	crisis	investors	down	monetary
15	trade	trade	economic	federal	ftse	down	policy	investors	debt
16	balance	reserve	growth	policy	rate	reserve	credit	global	interest
17	foreign	shares	investors	high	outlook	interest	euro	federal	federal
18	house	monetary	rise	oil	reserve	global	interest	reserve	prices
19	sales	currency	price	fed	investors	inflation	debt	monetary	reserve
20	european	plans	oil	reserve	down	government	government	stocks	data

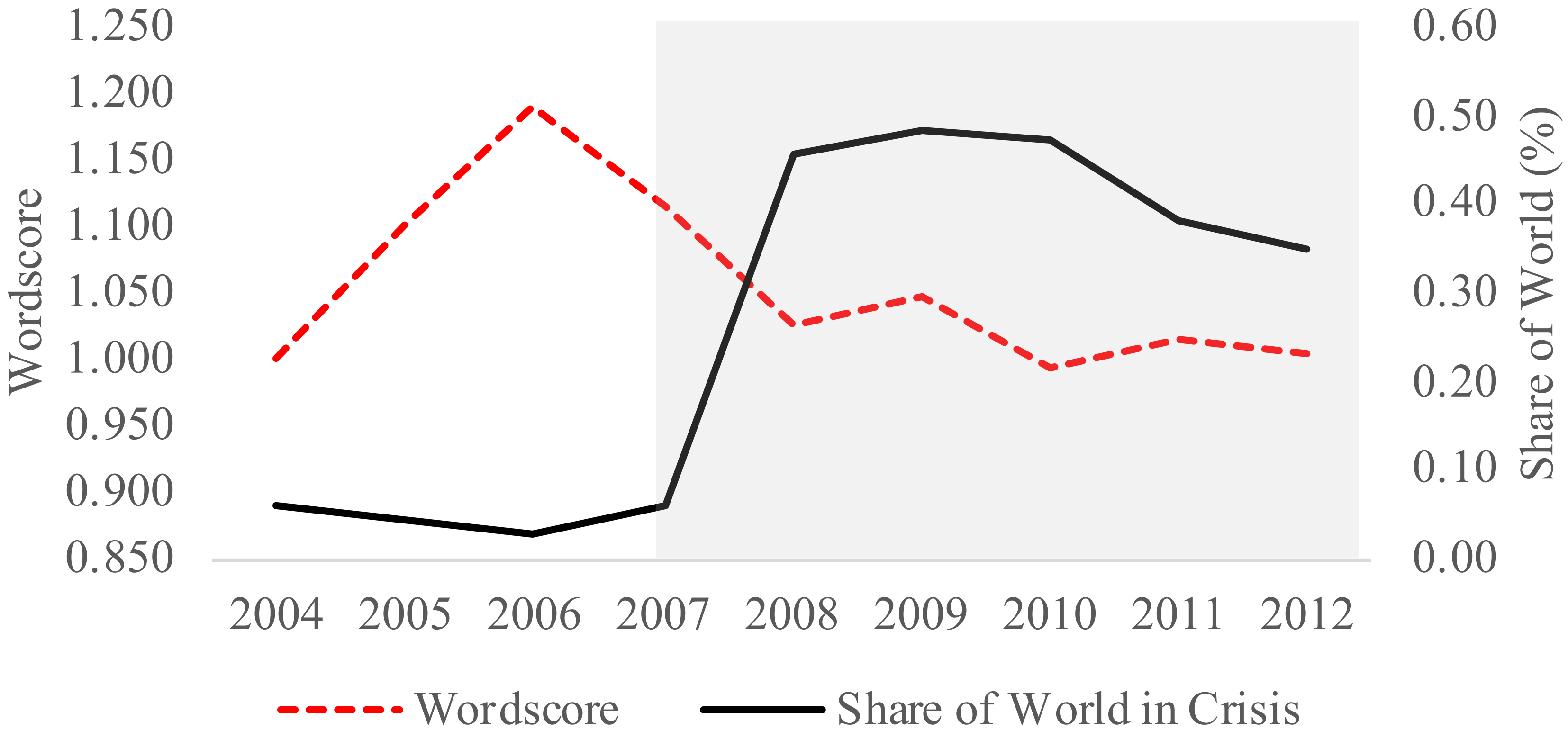
Index 1: Wordscores

- A priori policy positions
- Compare virgin texts to reference texts
- Positions and unique words

1. A_{rd} > reference text *R*, with a priori policy position on dimension *d*
2. F_{wr} > compute reflective frequency of word *w* as proportion of total words in reference text *r*
3. $P_{wr} = \frac{F_{wr}}{\sum_r F_{wr}}$ > use matrix of relative word frequencies to estimate conditional probabilities
4. $S_{wd} = \sum_r (P_{wr} A_{rd})$ > we can then use this matrix P_{wr} to produce a score for each word *w* on dimension *d*
5. $S_{vd} = \sum_w (F_{wv} S_{wd})$ > then we must compute the relative frequency of each virgin text word, as a proportion of the total number of words in the virgin text
6. $S^*{}_{vd} = (S_{vd} - S_{\tilde{v}d}) \left(\frac{SD_{rd}}{SD_{vd}}\right) + S_{\tilde{v}d}$ > where S_{vd} is the average score of the virgin texts, SD_{rd} and SD_{vd} are added as standard deviations of reference and virgin texts

Index 1: Wordscores

- To construct a Wordscore index, reference texts are a priori assigned as 1.00 for 2004 and 1.10 for 2005.
- Scores of the virgin texts are computed from 2006 to 2012.



Index 2: Wordfish

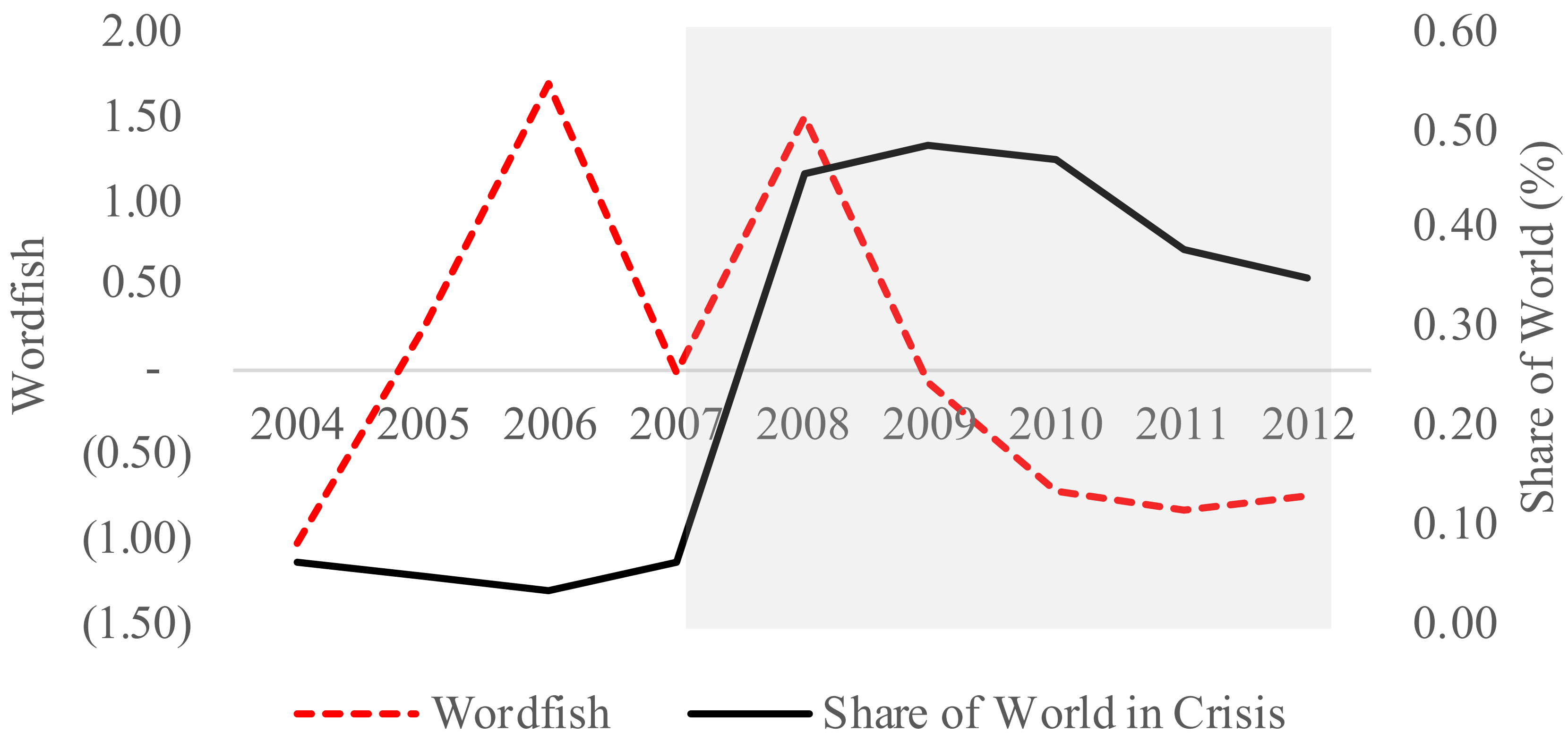
- The model is assumed to follow a Poisson distribution, and formally stated as:

$$Wordcount_{ij} \sim Poisson(\lambda_{ij})$$
$$\lambda_{ij} = \exp(\alpha_i + \psi_j + \beta_j \omega_i)$$

- Wordcount is the count of word *j* in text *i*, with α a set of document fixed effects, ψ as word fixed effects, β represents an estimated word specific weight to highlight the importance of word *j* in distinguishing between positions, and ω is an estimate of document *i*'s position

- Step 1: Calculate starting values $(\alpha_i + \psi_j + \beta_j \omega_i)$
- Step 2: Estimate document parameters $(\alpha_i + \omega_i)$ & maximise log-likelihood for documents $\sum_{j=1}^{m} (-\lambda_{ijt} + ln(\lambda_{ijt}) * y_{ijt})$
- Step 3: Estimate word parameters $(\psi_j + \beta_j)$ & maximise log-likelihood for each word $\sum_{it=1}^{n} (-\lambda_{ijt} + ln(\lambda_{ijt}) * y_{ijt})$
- Step 4: Calculate log-likelihood $\sum_{j}^{m} \sum_{it=1}^{n} (-\lambda_{ijt} + ln(\lambda_{ijt}) * y_{ijt})$
- Step 5: Repeat steps 2-4 until convergence

Index 2: Wordfish



Index 3: Banking Crisis Lexicon Index (BCLI)

$$\text{BCLI} = \frac{([B_1+B_2+\cdots,B_M]_t.[R_1+R_2+\cdots,R_M]_t.[E_1+E_2+\cdots,E_M]_t)}{(\sum_{i=1}^{w} A_t)}$$

Banking Sector

bank or banking
banking
deposit or credit OR debt
interest rate
inflation or cpi
reserve or gold
liquid or contract or eas or tight or monetary or boom or bust or crisis
fraud or earning or hous

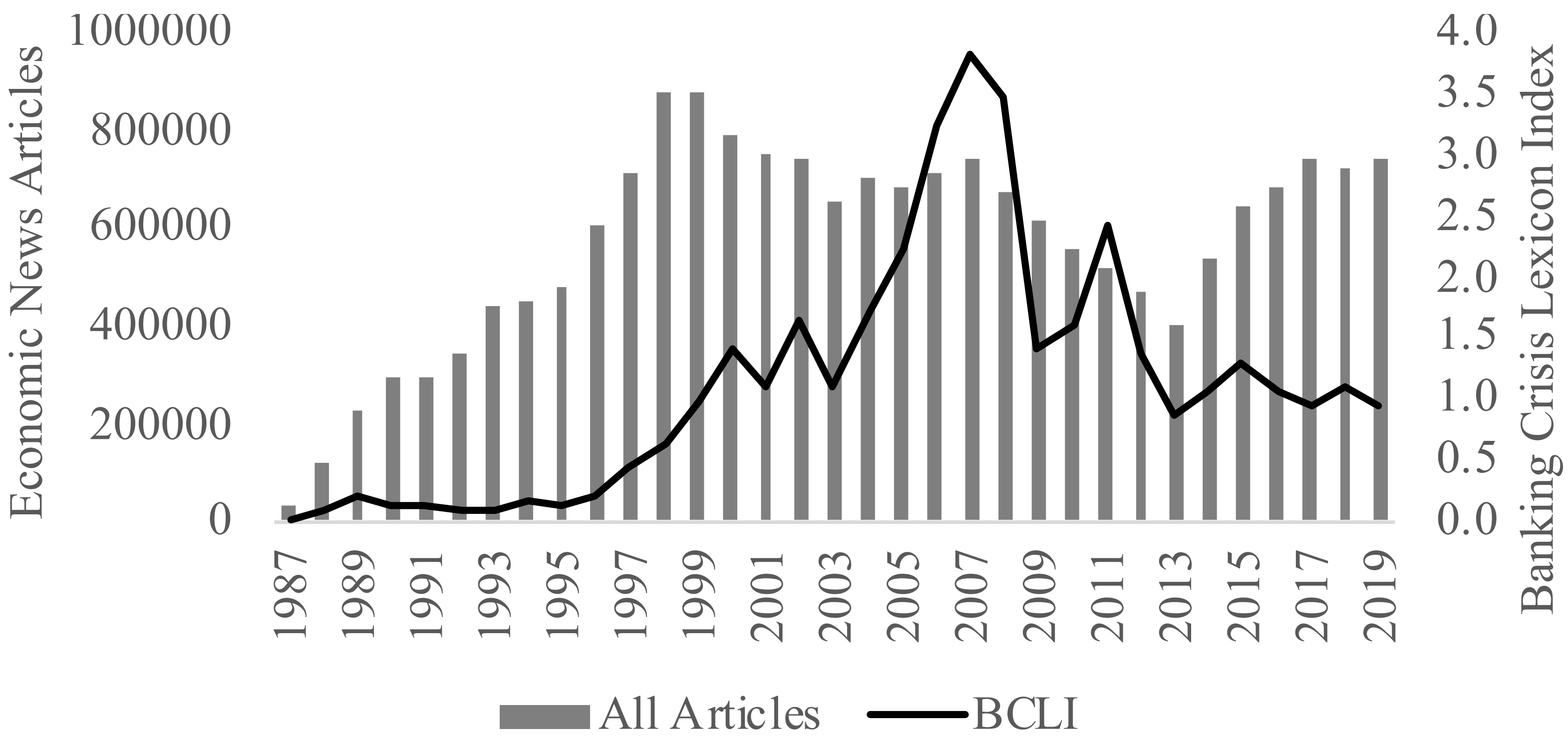
Real Sector

consum or invest or produc

External Sector

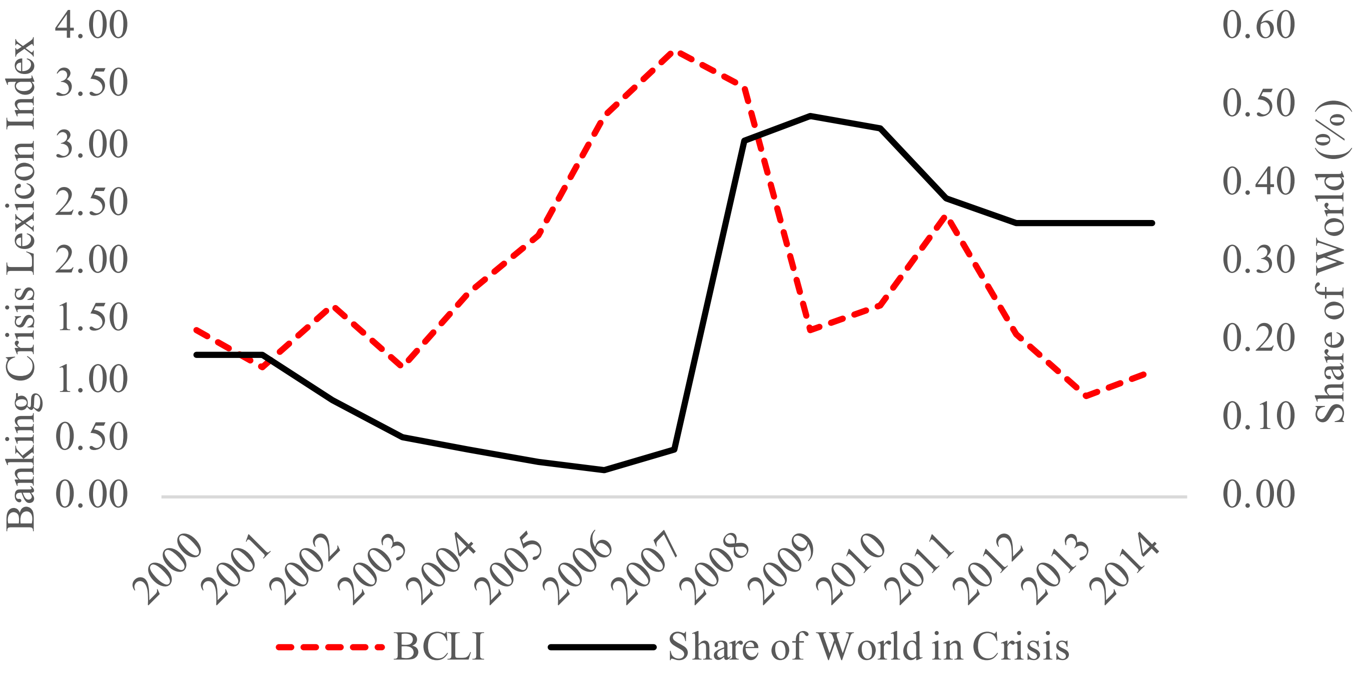
Export or import or trade or terms of trade

▶ Precursors

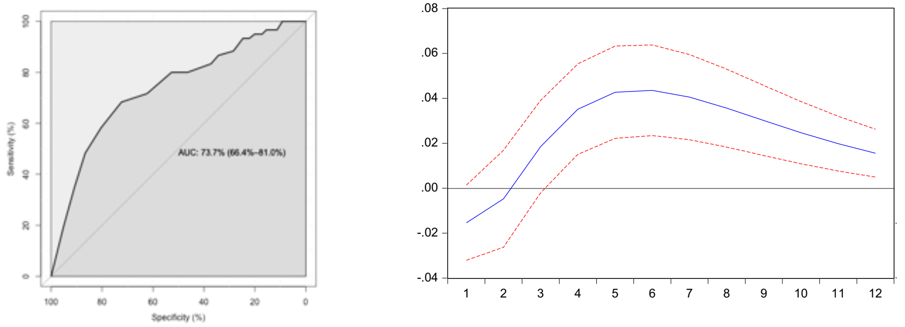


Index 3: Banking Crisis Lexicon Index and Granger Causality

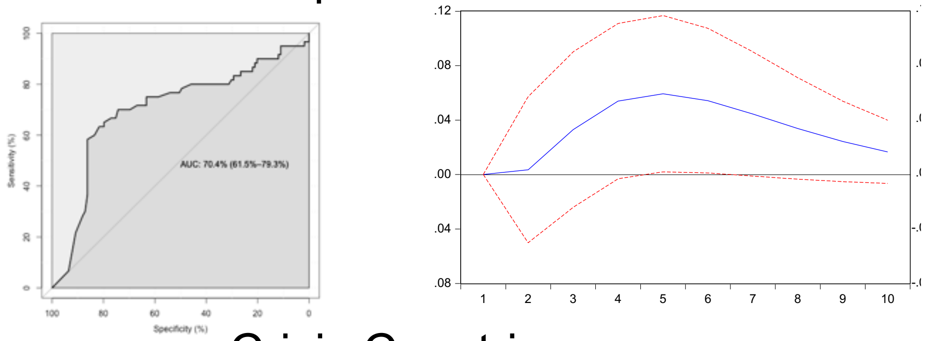
▶ Regional BCLI



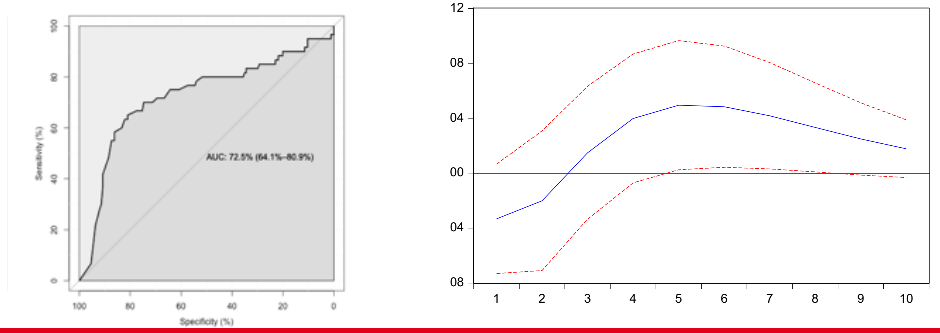
All Countries



Developed Economies



Crisis Countries



Granger Causality

	World		Developed Countries		Emerging Markets	
Lag	BCLI	Crisis	BCLI	Crisis	BCLI	Crisis
1	7.686***	4.685**	8.597***	8.269***	2.267	0.471
2	4.873**	1.851	4.115**	4.983**	2.610	0.089
3	3.987*	3.009	2.764	2.326	2.000	0.277
4	46.074**	3.892	4.112	9.547*	0.866	1.732

Index 3: Granger Causality and AUROC Results

Granger Causality for Crisis Countries

Granger Causality	BCLI	Crisis	BCLI	Crisis
	Lag Length = 1		Lag Length = 2	
United States	4.838**	0.681	5.273**	1.471
United Kingdom	4.517**	6.911**	4.096*	1.317
Germany	3.659*	0.624	5.097**	0.859
France	2.605	1.891	4.201*	0.765
Sweden	5.252**	0.002	5.187**	2.603
Netherlands	0.552	5.094**	4.404*	2.076
Italy	0.363	1.611	6.171**	1.123
Austria	0.898	11.954***	0.373	7.981**
Belgium	4.804**	4.672**	3.619*	1.760
Denmark	19.164***	0.356	37.473***	4.253*
Greece	0.329	3.167*	0.152	4.941**
Portugal	0.000	5.055**	0.016	1.522
Spain	3.395*	2.624	5.010**	1.367

AUROC results for BCLI forecasts

ROC Results	Mean	CI Lower Bound	CI Upper Bound	Standard Error
United States	0.698	0.342	1.000	0.181
United Kingdom	0.810	0.519	1.000	0.147
Germany	0.833	0.489	1.000	0.175
France	0.950	0.832	1.000	0.059
Sweden	0.533	0.186	0.880	0.177
Netherlands	0.875	0.619	1.000	0.130
Italy	1.000	1.000	1.000	0.000
Austria	0.750	0.374	1.000	0.191
Belgium	0.650	0.373	1.000	0.191
Denmark	0.612	0.278	0.946	0.170
Greece	0.475	0.106	0.843	0.187
Portugal	1.000	1.000	1.000	0.000
Spain	1.000	1.000	1.000	0.000

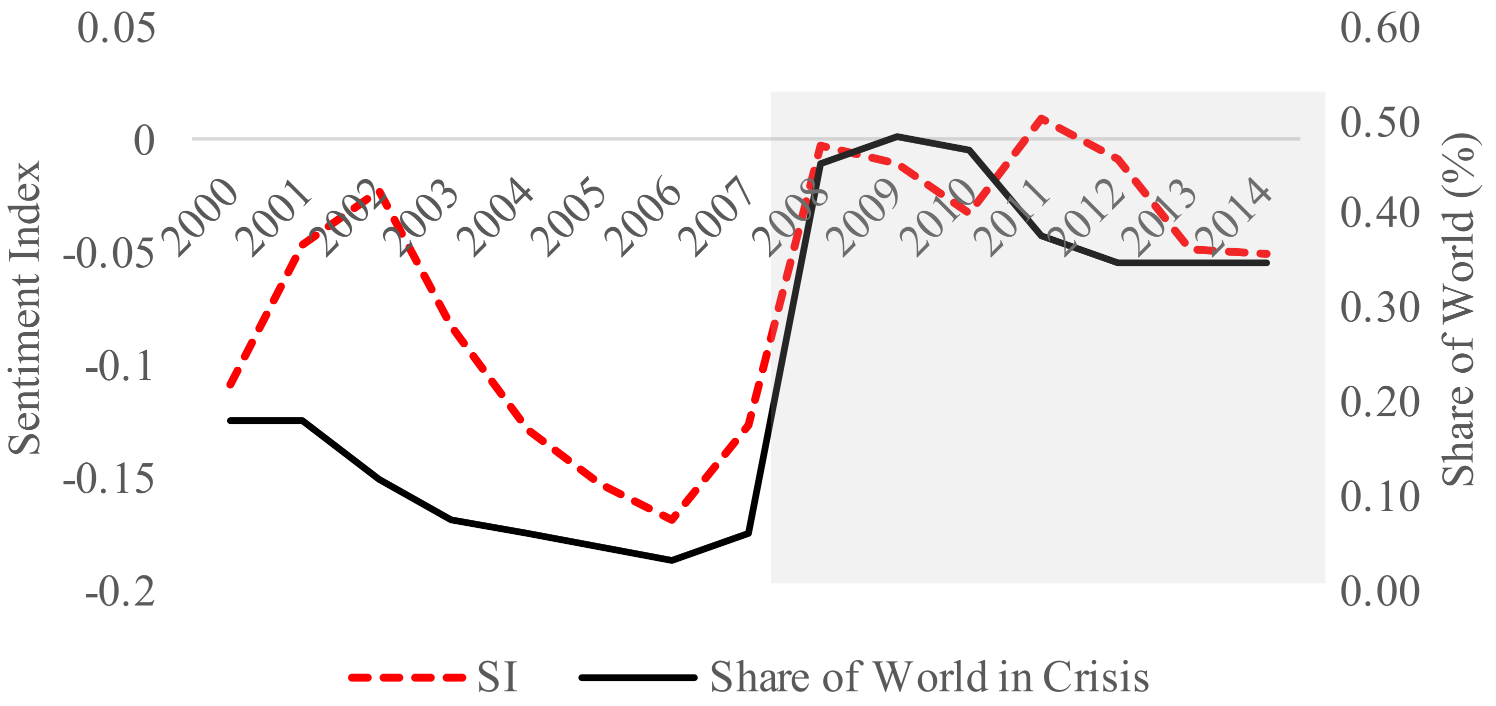
Index 4: Sentiment Index

$$Sentiment\ Index = \frac{(Sentiment_n - Sentiment_p)}{(Sentiment_n + Sentiment_p)}$$

where $Sentiment_n$ = total text units in category: negative sentiment and $Sentiment_p$ = total text units in category: positive sentiment.

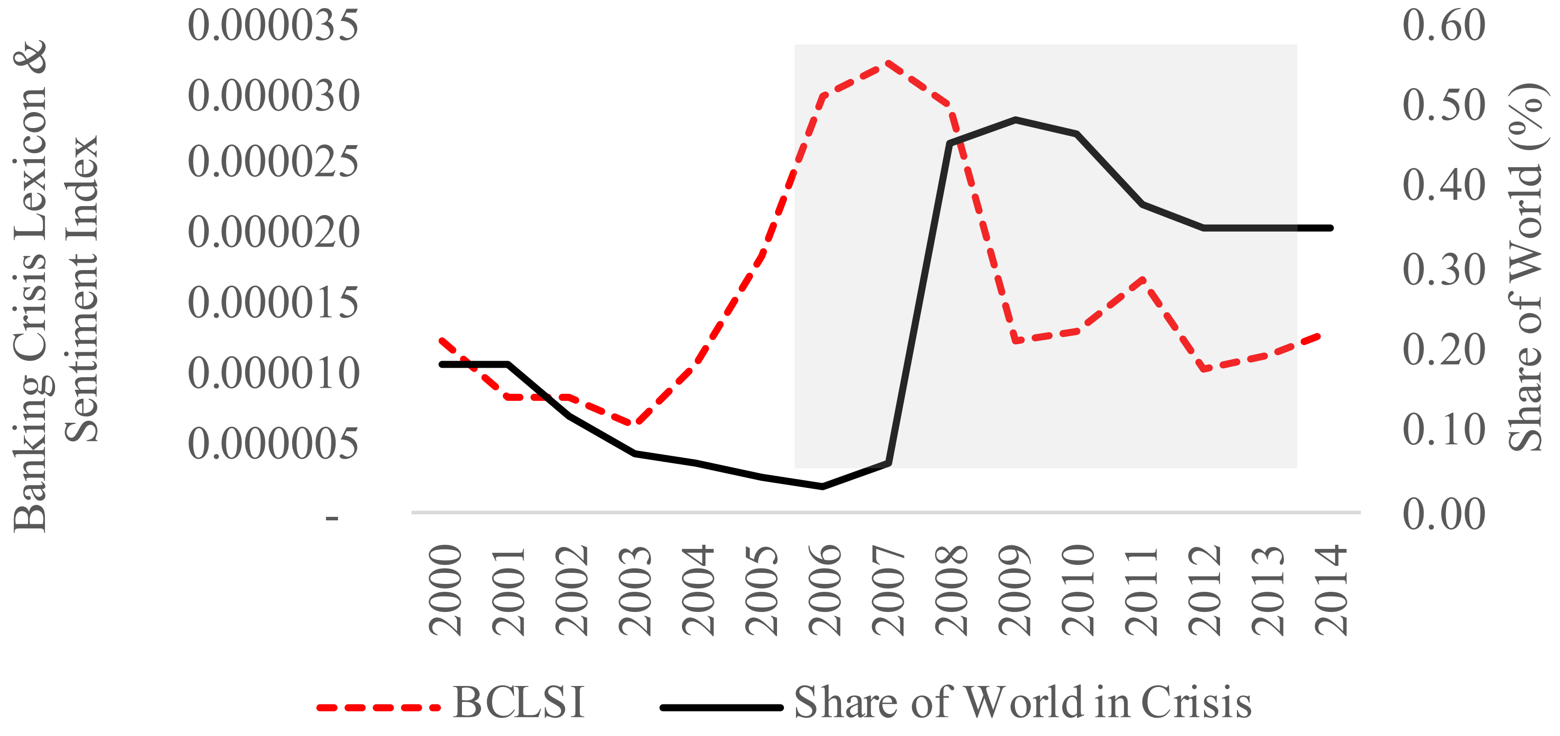
Positive Sentiment
bank *and* certain, expect, clarity, encourage, excit, incredible, pleas, attract, excel, impress, postiv *or* good

Negative Sentiment
bank *and* cncertain, unexpect, concern, discourage, bad, poor, panic, jitter, fail, crisis, distrust, jeopardy, terribl, worr, erod, reduc, warn, complicat, fear, woes, slump *or* low



▶ Regional SI

Index 5: Banking Crisis Lexicon and Sentiment Index



	World	
Lag	BCLSI	Crisis
1	9.246***	4.579**
2	6.640**	1.555
3	6.090**	1.189
4	7.598	3.951

▶ Robustness

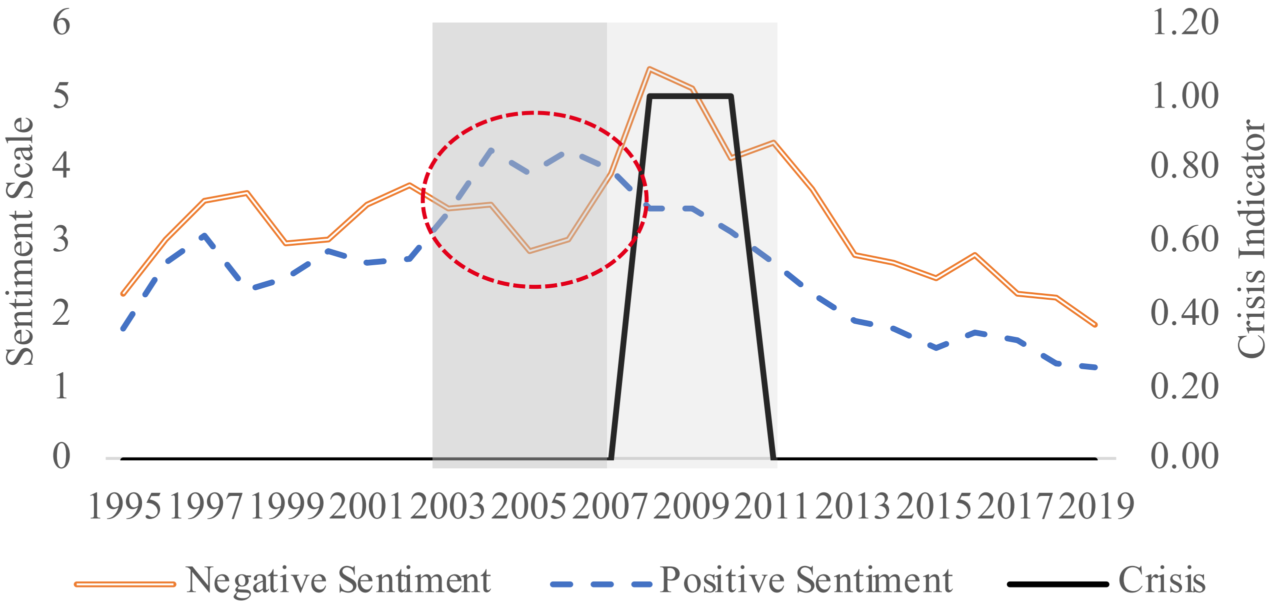
Robustness Test: German Language Index

German Banking Crisis Lexicon Index

Banking Sector	Real Sector	External Sector
Bank Depot or Kredit or Schuld Zins Inflation or VPI Reserv or Gold liquid or reduz or locker or eng or verschärf or monetär or Geldpolitik or Boom or Pleite or brech or Krise Betrug or Verdienst or Haus	verbrauch or investi or produzi	Export or Import or Handel

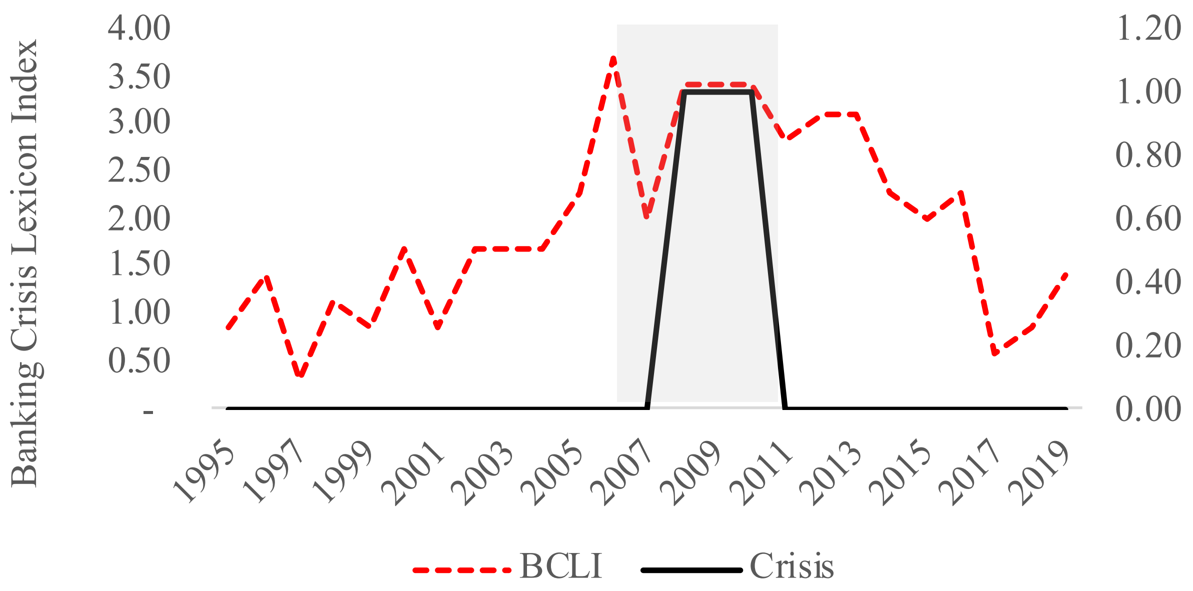
German Sentiment Index

Positive Sentiment	Negative Sentiment
bank *and* sicher, bestimmte, erwart, Klarheit, ermutig, aufreg, unglaublich, erfreulich, anlock, attraktiv, anziehen, übertref, beeindruck, postiv *or* gut	bank *and* unsicher, unerwart, besorgt, entmutigt, schlecht, niedrig, Panik, Jitter, scheiter, Krise, Misstrau, gefährd, schrecklich, Sorg, Erodier *or* Reduzier

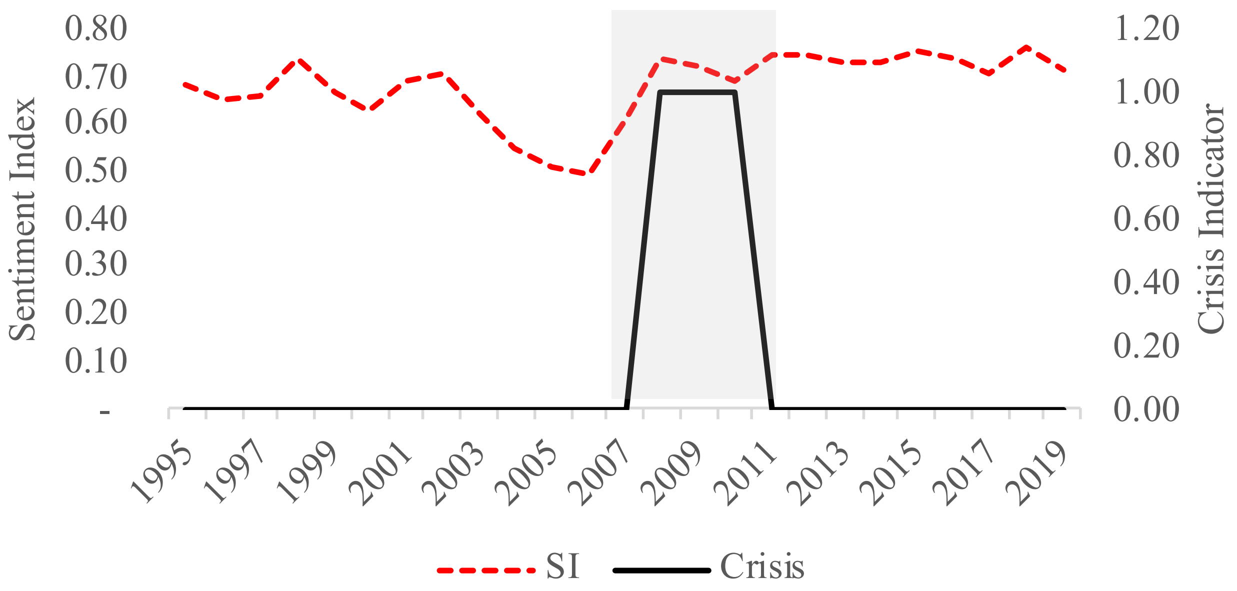


Robustness Test: German Language Index

German Banking Crisis Lexicon Index



German Sentiment Index



Granger Causality

Lag	BCLI	Crisis	Sentiment	Crisis
1	0.018	0.356	0.024	0.276
2	2.588*	0.550	3.533**	0.639
3	3.154*	0.286	4.596**	0.467
4	2.785*	2.323	6.232***	0.834

AUROC Results

ROC Results	Mean	CI Lower Bound	CI Upper Bound	Standard Error
German BCLI	0.941	0.826	1.000	0.058
Sentiment Index	0.725	0.480	0.971	0.125

▶ BCLSI

Conclusion

- Given severe impact of banking crises, the contribution of this paper is the development of **five distinct text-based indices** to enhance forecasting of banking crises, using text as data
- We introduced two **statistical models to study banking crises,** Wordscores using supervised learning and Wordfish as unsupervised approach
- We developed a Banking Crisis Lexicon Index which **signals a crisis, three years in advance**
- Our Sentiment Index serve as both a leading and coincidental indicator to crises and **improves when combining industry and sentiment terms**
- **Future research areas** include language specific indices, frequency extending to real time indices, country specific newspapers and opinion pieces