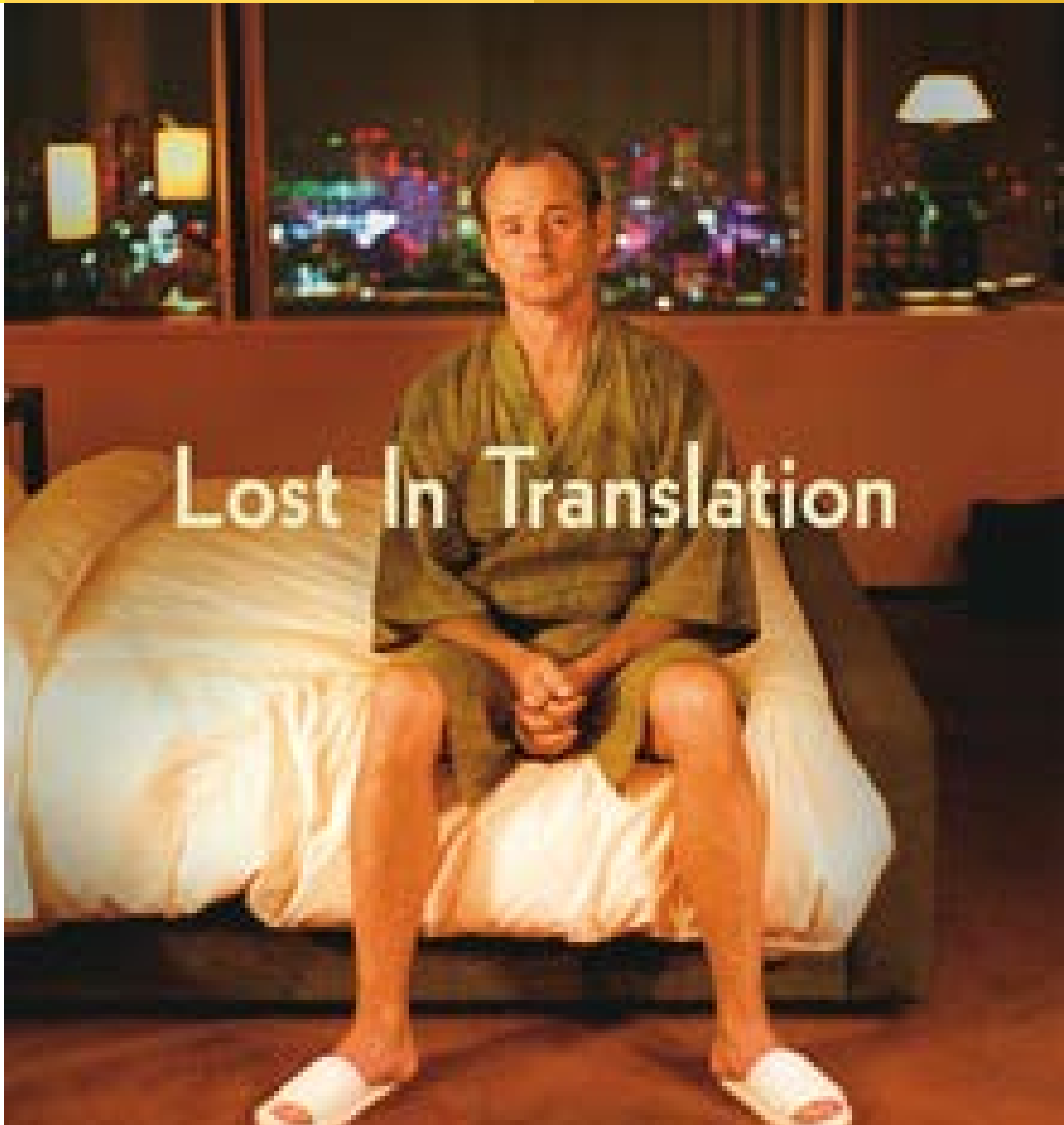
Language and Domain Specificity: A Chinese Financial Sentiment Dictionary

Zijia Du	Shanghai Jiao Tong University
Alan Guoming Huang	University of Waterloo
Russ Wermers	University of Maryland
Wenfeng Wu	Shanghai Jiao Tong University

Disclaimer

- Subtleties in languages
- “Common financial language” evolves into a call to use “applied linguistics” to accommodate for “richness and depth” of languages (Robinson, 2018).



SCHOOL OF
ACCOUNTING
& FINANCE

Chinese vs. English sentiment words in Finance

- There’s no Chinese equivalent to “yes”
- Financial sentiment words
  - Loughran and McDonald (2011)
    - Loughran and McDonald (2011) show that a general dictionary such as the Harvard Psychosociological Dictionary is unfit for sentiment analysis
    - For instance, Harvard-IV-4 misclassifies negative tone roughly 75% of the time when examining annual reports.
  - Our translation to Chinese of Loughran and McDonald (2011) differs significantly from those published by others.
- Both language and domain specificity calls for a unique financial sentiment dictionary for Chinese.

Literature of sentiment words

1) Translation of Loughran and McDonald (2011) (“LM”)—a few Chinese studies
2) Manual reading based: Translation + Manual reading of words from 2,000 news articles: You, Zhang, and Zhang (2018 “YZZ”) and manual reading of 24 IPO prospectuses (Yan et al. 2019)
3) Returns-based machine learning approaches: learning of news from subsequent returns (Mao et al. 2014), and machine learning (long short-term memory) of social media posts based on three-day returns (Yao et al. 2021, in Chinese)
4) General dictionaries (Dalian U. Tech, NTU, Hownet; “generic”)

See Huang, Wu, and Yu (2019) for a literature review

Our approach

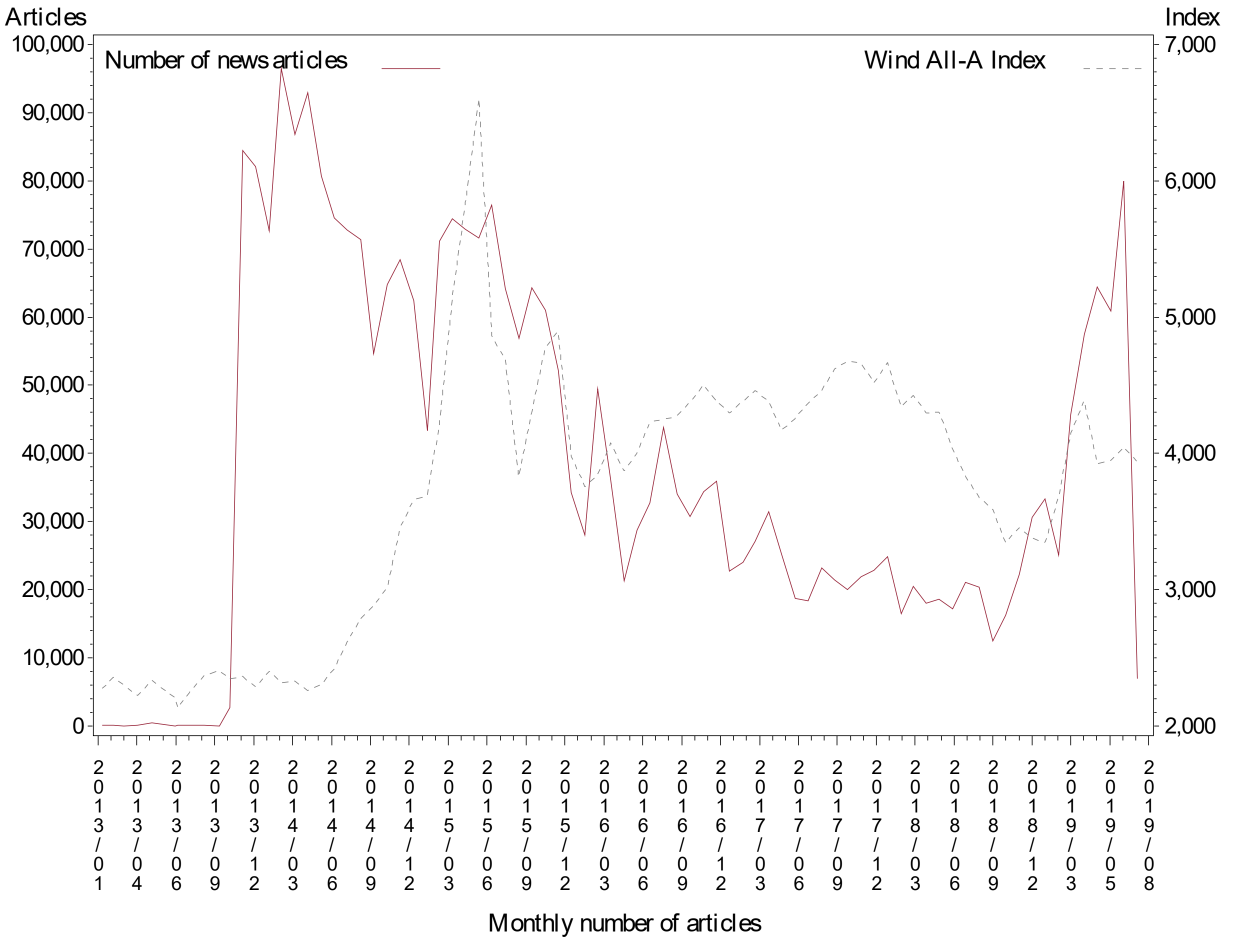
- Principles:
  - Credible sources of input, as large as possible
  - Human expert justification of “sentiment”
  - Avoids inferences of outcomes from returns (return generation is one of the most, if not the most, complicated phenomena in social sciences)
- Supervised machine learning by Word2vec
  - Bags of words –turn words into vectors—calculate the “closeness” of words by cosine similarity (e.g., Mikolov, Corrado, Chen and Dean, 2013; Mikolov et al., 2013; and Jurafsky and Martin, 2019)
  - Example:
    - The startup ‘**burned-cash**’ (烧钱） significantly in 2019.
    - The startup ‘**lost-money**’ （亏损） a lot in 2019.
  - Word2vec iteratively maximizes the similarity of the target word and the context words
  - Seed provided by human being; and iterative outputs supervised by human being (expert opinion on “sentiment”)

Modern Chinese-language specificity

- "*Mind politics*" is a modern Chinese culture that infiltrates ubiquitously into business writing
- Lots of slogan-like words that are different from the "usual" positive words
- We therefore have a separate category of sentiment words for these politically-inclined words
- Three categories of "sentiment": negative, positive, and politically positive words

Sample

- Credible source of input
  - All “Firm News” from finance.sina.com.cn, the most visited Chinese finance website that streams real-time news and stock information for each stock, from 2013 to 2019
  - Firm news there is highly similar in both content and quantity to that on Wind (the Chinese equivalent of Bloomberg Terminal).
  - 3.1 million news articles covering 3,557 stocks.



An example for dictionary construction

- For humanly-identified seed word: “涨停” (verb of “price hitting up-limit”), Word2vec produces the following top seven candidates:
  - “涨停板” (noun of “price hitting up-limit”),
  - “跌停” (verb of “price hitting down-limit”, which is its antonym)
  - “一字板” (another noun of “price hitting up-limit”),
  - “封板” (another verb of “price hitting up-limit”),
  - “大涨” (“stock price soars”),
  - “拉升” (“stock price gap-increases”),
  - “两连板” (“two continuous hits of price up-limit”)
- We proofread these candidate terms into positive, negative, political, or neutral
  - Six terms are labeled as positive-sentiment words and the antonym as a negative-sentiment word.

Construction of Sentiment Words Dictionary

(each round features convergence of opinions by three separate human "experts" + authors)

Panel A: Manually reading 2,500 articles for four rounds and utilizing the YZZ dictionary

Round	# of news	# of unique words	Sentiment words selected incrementally: Negative	Positive	Political
1	50	1,003	41	56	24
2	250	2,980	193	152	65
3	200	1,343	100	41	33
4	2,000	28,245	372	451	184
5	YZZ	--	264	133	16
Total	2,500+YZZ	33,571	970	833	322

Panel B: Synonyms produced by Word2vec and human review

Iter-ation	Seed article words	# of firms	# of news articles	Additional synonyms from Word2vec: Negative	Positive	Political	Additional Valid synonyms: Negativ	Positive	Political
1	500	100	576,153	1,858	1,730	878	594	506	337
2	500	1,000	1,777,178	1,907	1,404	936	579	351	347
3	500	3,557	3,078,175	3,640	3,403	2,563	573	372	319
4	2,000	3,557	3,078,175	782	1,308	735	201	138	108
5	YZZ	3,557	3,078,175	648	1,077	148	69	35	6

* Supervision removes 80% of output;

* A small seed set from reading 500 articles does a good job.

Overlapping with other dictionaries

Panel C: Overlapping of our dictionary with other dictionaries (percentage in parer

	Negative	Positive	Political	Total
Loughran and McDonald Translation	489 (16.38%)	144 (6.44%)	43 (2.99%)	676 (10.15%)
YZZ Dictionary	1,145 (38.35%)	812 (36.33%)	208 (14.45%)	2,165 (32.51%)
Generic Chinese Dictionaries	134 (4.49%)	153 (6.85%)	74 (5.14%)	361 (5.42%)
Total	1,434 (48.02%)	910 (40.72%)	280 (19.46%)	2,624 (39.40%)

- Loughran and McDonald Translation: Our own translation augmented by a computational Synonym package
- Generic: the intersection of Dalian, NTU, & Hownet

“New words” identified

Panel a: Top 50 negative words (8 new words by our dictionary, in bold font)

Panel b: Top 50 positive words (13 new words by our dictionary, in bold font)

Validation

- Internal: Is news sentiment related to common-sense variables, such as firm fundamentals?
- External: Is sentiment from word-counting consistent with overall judgment from reading the entire news article?

- News filtering for "**firm-specific**" tests
  - Remove news articles that are in essence industry and market-wide
    - Half of the news articles are "general" articles such as market commentary covering many firms.
  - Remove duplicate news and news reprints/recombinations
  - Other institutional considerations: news around IPO, trading halts, and news released on the same day and/or intra-day, etc.
  - 3.1 million news to 424,758 news-days.
- Sentiment and tests largely follow the literature (Tetlock et al. 2008, Huang, Tan, Wermers 2020):
  - *Neg* (*Pos*): % of negative (positive) word occurrences. In US markets, *Neg* tends to have a larger impact than *Pos*.
  - *Neg_net: Neg-Pos*.

Internal validation using fundamentals

	(1)	(2)	(3)	(4)
	Neg_net	*Neg*	*Pos*	*PoliticalPos*
beta	0.300***	0.202***	-0.098***	-0.081***
	(6.37)	(7.58)	(-2.89)	(-3.85)
Log market cap.	-0.297***	-0.050*	0.250***	0.084***
	(-5.55)	(-1.76)	(6.52)	(3.36)
Book to market	1.109***	0.616***	-0.496***	-0.266***
	(7.67)	(7.91)	(-5.00)	(-4.32)
Turnover	-3.946***	-0.299	3.681***	0.276
	(-4.78)	(-0.70)	(6.05)	(0.66)
Volatility	-6.985***	4.860***	11.869***	-2.570***
	(-4.72)	(5.86)	(10.09)	(-2.96)
SUE	-0.132***	-0.066***	0.066***	0.045***
	(-11.17)	(-10.02)	(8.26)	(8.65)
Dividend yield	-3.615**	-3.114***	0.543	0.000
	(-1.96)	(-3.38)	(0.39)	(0.00)
Stock age	-0.088	0.150***	0.235***	-0.104***
	(-1.24)	(3.45)	(4.54)	(-2.94)
CSI300 dummy	0.069	0.012	-0.060	-0.041
	(0.81)	(0.25)	(-0.95)	(-1.00)
SOE dummy	0.191	0.024	-0.165	-0.043
	(1.29)	(0.25)	(-1.61)	(-0.50)
Historical articles	0.301***	0.178***	-0.125***	-0.070***
	(8.83)	(9.24)	(-5.17)	(-4.41)
Number of articles$_t$	-0.186***	-0.104***	0.080***	0.022
	(-4.95)	(-5.42)	(3.05)	(1.08)
Excess Return$_{t-1}$	-0.157***	-0.034***	0.122***	0.023***
	(-34.07)	(-15.27)	(36.82)	(13.81)
Excess Return$_{t-2}$	-0.037***	0.002	0.038***	0.003*
	(-10.06)	(0.79)	(14.55)	(1.95)
Excess Return$_{t-5,t-3}$	-0.105***	-0.008**	0.097***	-0.000
	(-16.50)	(-2.49)	(19.67)	(-0.10)
Excess Return$_{t-10,t-6}$	-0.138***	-0.023***	0.116***	0.002
	(-17.66)	(-5.38)	(19.25)	(0.48)
Excess Return$_{m-12,m-2}$	-0.003***	-0.001***	0.002***	0.001***
	(-11.29)	(-6.40)	(11.26)	(4.89)

Article level validation

- Vs. human:

Panel A: Article-level sentiment judgment by sentiment-word counting vs. by human

	# of articles	# of artciles by *Neg_net* value	Accuracy
Human-labeled negative news	2,500	2,147 with *Neg_net* ≥ 0	85.88%
Human-labeled positive news	2,500	2,210 with *Neg_net* < 0	88.40%
Overall	5,000	4,357	87.14%

- Vs. traditional machine-learning approach of support vector machine:
  - SVM classifies an article into positive or negative based on a training set; in our case, we use articles in Panel A.

Panel B: Article-level sentiment judgment by sentiment-word counting vs. SVM evaluated on human training sample

Training vs. test set size ratio	SVM training results							% in test set consistently judged by SVM and *Neg_net*
	Negative human-labeled news			Positive human-labeled news			Weighted F1-score	
	Precision	Recall	F1-score	Precision	Recall	F1-score		
7:3	88.05%	88.40%	88.22%	88.35%	88.00%	88.18%	88.20%	83.60%

- Vs. the third party of Wind Terminal
  - Wind tags a news article into positive, negative or null (with an undisclosed method)
  - We download 50,000 articles from Wind, and find that 86.75% of the articles that are labeled as positive or negative are consistently judged by *Neg_net*

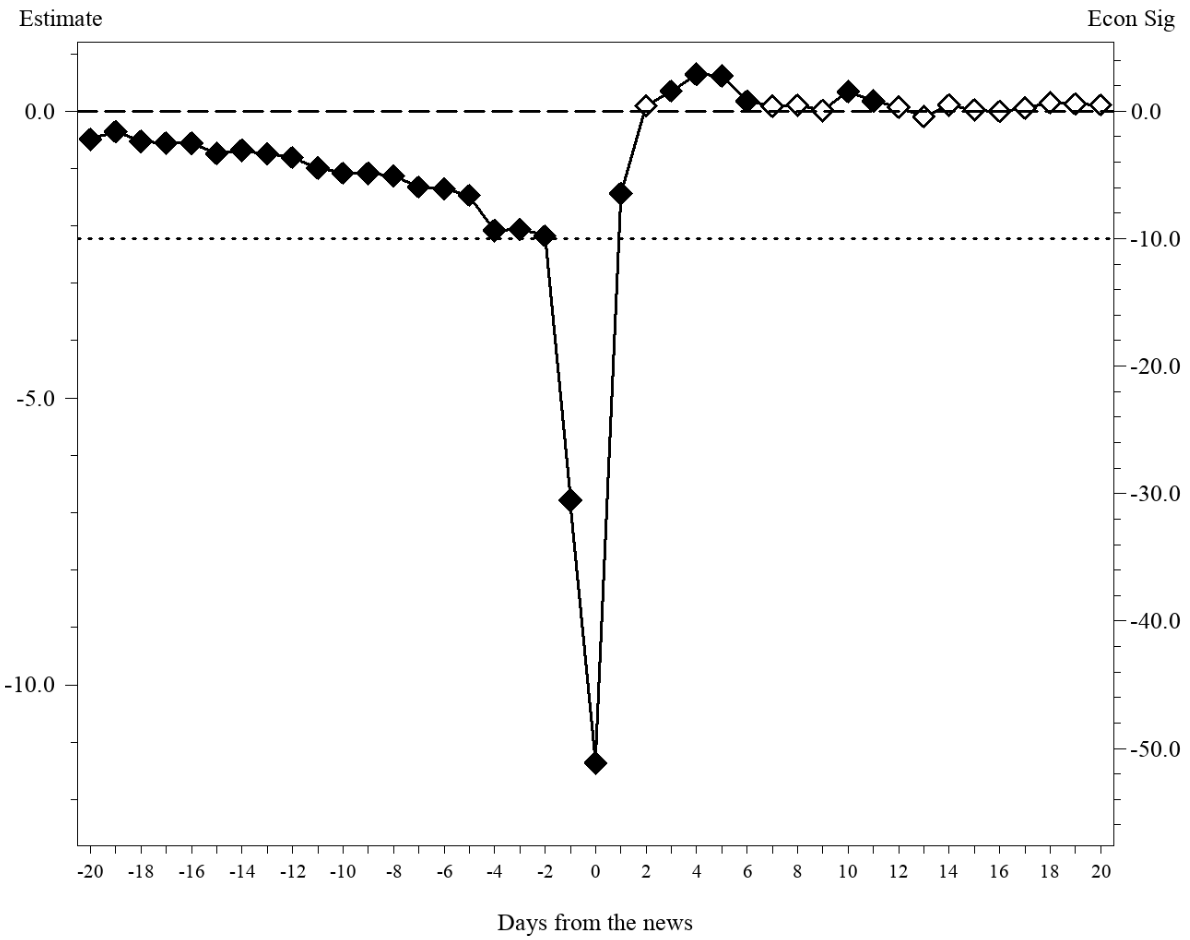
Use cases for our dictionary

- Stock returns
  - News may drive returns, or may be driven by returns
- Media bias
  - Any systematic biases in news sentiment, in particular, in state media?
  - Any peculiarities in *PoliticalPos*?

Return regressions

Panel A: Return association with *Neg_net*

	Industry- and size-adjusted return over day(s)					
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]
Neg_net	-1.212***	-1.605***	-2.171***	-6.787***	-11.366***	-1.432***
	(-18.40)	(-19.33)	(-16.87)	(-38.43)	(-50.78)	(-11.66)
beta	-0.035***	-0.010	-0.027	0.001	0.011	-0.011
	(-2.61)	(-0.65)	(-1.26)	(0.04)	(0.45)	(-0.58)
Log market cap.	-0.231***	-0.211***	-0.207***	-0.237***	-0.225***	-0.128***
	(-15.03)	(-14.52)	(-9.98)	(-11.57)	(-10.57)	(-6.30)
Book to market	0.260***	0.270***	0.266***	0.361***	0.321***	0.306***
	(6.92)	(6.65)	(4.91)	(6.10)	(5.35)	(5.83)
Turnover	-4.964***	-3.328***	-3.310***	-3.400***	-3.682***	-2.483***
	(-16.07)	(-10.16)	(-7.19)	(-6.63)	(-7.62)	(-5.47)
Volatility	4.815***	3.093***	3.225***	4.022***	1.784*	0.636
	(7.09)	(4.34)	(3.33)	(3.87)	(1.82)	(0.68)
SUE	0.008***	0.006*	0.012***	0.005	0.007	0.004
	(2.75)	(1.96)	(2.91)	(1.22)	(1.58)	(1.07)
Dividend yield	0.131	0.136	-0.162	-0.178	-0.362	-0.581
	(0.31)	(0.29)	(-0.23)	(-0.26)	(-0.46)	(-0.90)
Stock age	-0.110***	-0.052**	-0.088**	-0.079*	-0.160***	-0.084**
	(-4.61)	(-1.97)	(-2.16)	(-1.74)	(-3.66)	(-2.23)
CSI300 dummy	-0.028*	-0.021	-0.042*	-0.015	-0.043	-0.058**
	(-1.69)	(-1.14)	(-1.70)	(-0.50)	(-1.50)	(-2.48)
SOE dummy	0.017	-0.005	0.043	0.060	0.052	0.017
	(0.51)	(-0.14)	(1.20)	(1.37)	(1.13)	(0.37)
Excess Return$_{t-5,t-3}$			0.152***	0.062***	-0.045***	-0.037***
			(16.61)	(6.94)	(-5.29)	(-4.85)
Excess Return$_{t-10,t-6}$		0.059***	0.003	-0.016	-0.026***	-0.023***
		(7.44)	(0.26)	(-1.51)	(-2.85)	(-2.58)
Excess Return$_{m-12,m-2}$	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(-5.28)	(-4.86)	(-4.00)	(-3.20)	(-4.96)	(-3.39)
Historical articles	0.060***	0.026***	-0.018	-0.045***	-0.067***	-0.037***
	(7.29)	(2.93)	(-1.57)	(-3.63)	(-5.11)	(-3.24)
Number of articles$_t$	0.056***	0.117***	0.167***	0.320***	0.397***	0.003
	(6.50)	(10.36)	(9.61)	(15.41)	(15.85)	(0.16)
Constant	5.583***	4.905***	5.136***	5.678***	5.793***	3.417***
	(14.80)	(13.44)	(9.86)	(11.01)	(10.66)	(6.80)
Observations	413,156	411,751	410,943	411,751	411,751	410,145
Adj R-squared	0.031	0.027	0.025	0.032	0.050	0.013



- Econ Sig = coefficient estimate times the standard deviation of *Neg_net*, in bps
- Evidence of limited information leakage with econ sig on day [-1] (even after adjusting for news persistence not shown here)

Return regressions, other measures

Panel B: Return association with other sentiment measures

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Neg	-0.632***	-0.433***	-0.418	-4.942***	-12.916***	-1.382***	0.216	0.489***	-0.026
	(-5.12)	(-2.96)	(-1.61)	(-15.82)	(-37.94)	(-6.67)	(1.11)	(4.21)	(-0.28)
Pos	1.715***	2.459***	3.399***	8.874***	12.743***	1.724***	-0.046	-0.700***	-0.270***
	(20.39)	(22.31)	(21.81)	(41.00)	(45.65)	(10.79)	(-0.28)	(-8.69)	(-4.46)
PoliticalPos	0.035	0.058	0.775***	3.267***	6.837***	0.990***	-0.102	-0.306***	0.003
	(0.31)	(0.41)	(4.01)	(13.56)	(24.00)	(5.31)	(-0.36)	(-2.94)	(0.03)

- *Pos* more significant than *Neg*
- *PoliticalPos* not as significant

Return horserace with other dictionaries

- Pooling all four dictionaries:

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Neg_net	-1.010***	-1.394***	-1.725***	-5.591***	-10.645***	-1.197***	-0.144	0.543***	0.108
	(-10.35)	(-11.40)	(-8.95)	(-24.33)	(-37.87)	(-6.73)	(-0.78)	(5.18)	(1.45)
Neg_net_YZZ	-0.345***	-0.591***	-1.658***	-3.508***	-2.683***	-0.565***	0.122	0.171*	0.161**
	(-3.80)	(-5.35)	(-9.00)	(-15.13)	(-12.90)	(-3.36)	(0.71)	(1.73)	(2.21)
Neg_net_LM	0.158	0.694***	2.661***	4.921***	3.910***	0.817***	0.334	-0.343***	-0.220**
	(1.27)	(4.59)	(10.33)	(14.45)	(13.61)	(3.72)	(1.40)	(-2.97)	(-2.32)
Neg_net_generic	1.327***	1.959***	2.752***	5.498***	8.302***	-0.524	0.524	-0.479	-0.570**
	(3.99)	(4.90)	(4.59)	(8.39)	(10.94)	(-0.89)	(0.71)	(-1.44)	(-2.27)

- *Neg_net* based on dictionary from seed words of only 2,500 news articles (instead of also including YZZ seed words)

	Industry- and size-adjusted return over day(s)								
	[-10, -6]	[-5, -3]	[-2]	[-1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Neg_net_2500	-1.006***	-1.331***	-1.472***	-5.132***	-10.341***	-1.113***	-0.100	0.509***	0.090
	(-10.25)	(-11.02)	(-7.50)	(-21.42)	(-37.38)	(-6.31)	(-0.55)	(4.89)	(1.22)
Neg_net_YZZ	-0.237***	-0.312***	-0.800***	-1.898***	-1.197***	-0.366**	0.228	0.055	0.072
	(-2.77)	(-2.92)	(-4.93)	(-10.06)	(-6.27)	(-2.35)	(1.28)	(0.60)	(1.07)

Media bias

- "Party-line" journalism
  - Qin, Strömberg, and Wu (2018)) measure media bias in China based on the coverage of "government mouthpiece" content, such as the number of mentions of party leaders and the number of cites of Xinhua News Agency
  - Piotroski, Wong, and Zhang (2017) tag an article's political bias by the frequency of political phrases in the Dictionary of Scientific Development (Xi, 2007) (*PoliticalNouns*)
  - Piotroski, Wong, and Zhang (2017) and YZZ report that state media outlets (relative to business or market media outlets) issue fewer negative corporate news articles.
- Informativeness
  - The above literature documents that news stories by state media have lower value relevance, including a smaller price impact on corresponding stocks.
- What can our sentiment dictionary add to this literature?

State media's sentiment bias

- State media uses more politically-inclined positive words and fewer negative words

	(1)	(2)	(3)	(4)
	PoliticalPos	*Neg*	*MediabiasIndex*	*PoliticalNouns*
State media	0.258***	-0.271***	0.529***	0.615***
	(9.68)	(-11.85)	(12.28)	(18.59)
beta	-0.129***	0.217***	-0.345***	0.112***
	(-5.08)	(6.44)	(-7.03)	(3.06)
Log market cap.	0.076**	0.003	0.073	-0.179***
	(2.32)	(0.08)	(1.26)	(-5.02)
Book to market	-0.256***	0.728***	-0.984***	-0.594***
	(-3.13)	(6.86)	(-6.06)	(-6.10)
Turnover	-0.062	-0.275	0.213	0.965
	(-0.13)	(-0.51)	(0.25)	(1.60)
Volatility	-1.500	4.835***	-6.335***	-4.605***
	(-1.57)	(4.24)	(-3.78)	(-3.89)
SUE	0.046***	-0.072***	0.118***	0.048***
	(6.50)	(-8.32)	(8.50)	(4.99)
Dividend yield	-1.459	-1.791	0.332	1.538
	(-1.22)	(-1.47)	(0.16)	(1.53)
Stock age	-0.137***	0.163***	-0.300***	0.146***
	(-3.25)	(2.70)	(-3.91)	(2.67)
CSI300 dummy	-0.022	0.022	-0.044	-0.042
	(-0.42)	(0.29)	(-0.41)	(-0.65)
SOE dummy	-0.118	0.055	-0.173	0.037
	(-0.87)	(0.34)	(-0.61)	(0.43)
Histotical articles	-0.073***	0.198***	-0.271***	-0.111***
	(-3.45)	(7.10)	(-6.57)	(-4.12)

State media's sentiment measures are less return-informative

	Industry- and size-adjusted return over day(s)							
	[-1, 1]	[-1, 1]	[-1, 1]	[-1, 1]	[-1, 1]	[-1, 1]	[-1, 1]	[-2, 2]
State media	-0.142***	-0.381***	-0.221***	-0.172***	-0.142***	-0.426***	-0.274***	-0.191***
	(-8.43)	(-17.24)	(-15.76)	(-9.48)	(-7.20)	(-15.86)	(-13.84)	(-12.96)
PoliticalPos	5.154***				5.189***			
	(11.79)				(11.68)			
State media × *PoliticalPos*	-2.187***				-2.217***			
	(-5.77)				(-5.77)			
Neg		-12.041***				-12.150***		
		(-20.10)				(-20.14)		
State media × *Neg*		5.497***				5.595***		
		(11.74)				(11.87)		
MediabiasIndex			7.070***				7.217***	4.820***
			(17.86)				(17.91)	(16.33)
State media × *MediabiasIndex*			-3.108***				-3.248***	-2.350***
			(-10.30)				(-10.55)	(-10.32)
PoliticalNouns				0.495*	-0.204	-0.844***	-1.270***	-0.711***
				(1.84)	(-0.74)	(-2.95)	(-4.36)	(-3.18)
State media × *PoliticalNouns*				-0.579**	0.055	0.948***	1.211***	0.994***
				(-2.17)	(0.20)	(3.31)	(4.17)	(4.38)

Appendix: Zoom in on our dictionary vs. YZZ

- Whether we and YZZ agree or disagree on ordinal ranking of news (2 by 2 quadrants by median values of ours and YZZ)
  - 82% of the time we and YZZ rank news similarly ("agreeing news")

	Agreeing News					Disagreeing News				
	Industry- and size-adjusted return over day(s)									
	[-2]	[-1]	[0]	[1]	[2]	[-2]	[-1]	[0]	[1]	[2]
Net negative tone										
Neg_net	-1.290***	-5.695***	-11.350***	-1.417***	0.102	-1.838***	-1.377**	-9.150***	0.119	0.064
	(-5.58)	(-20.01)	(-35.59)	(-6.87)	(0.47)	(-3.32)	(-2.28)	(-13.09)	(0.22)	(0.12)
Neg_net_YZZ	-0.955***	-1.363***	-0.260	-0.106	0.037	-0.626	-0.960*	-3.833***	-0.333	1.134**
	(-4.92)	(-6.04)	(-1.13)	(-0.57)	(0.17)	(-1.40)	(-1.80)	(-7.39)	(-0.72)	(2.53)
Obs.	352,913	353,628	353,628	352,197	351,193	57,915	58,009	58,009	57,834	57,678

Robustness/Other tests

- Abnormal news sentiment adjusted for news persistence
- News clustering as in Huang, Tan, Wermers (2020)

Long significance of sentiment on returns is not due to news persistence.

- Firm- vs. press-initiated news
- Using news headlines only
- Removing all intra-trading-day news
- Excluding news [-3, 3] days around earnings announcements

Conclusion

- We develop a context-specific financial sentiment dictionary in Chinese.
- We demonstrate that such a dictionary needs to be language and domain specific.
- Evidence suggests that positive sentiment is important in return associations and also a limited degree of information leakage in China.
- We also develop a list of politically inclined words, and show that these words are useful towards constructing a media sentiment bias.
- As China now ranks as the second largest stock market in the world by running two of the ten largest stock exchanges in the world, we believe that a suitable sentiment dictionary for financial texts is of significant economic importance.

thank you