Language and Domain Specificity: A Chinese Financial Sentiment Dictionary

Zijia Du, Alan Guoming Huang, Russ Wermers, Wenfeng Wu

Review of Finance 2022 June

胡震霆 2023/05/31

Motivation

- The expression of fondness or the lack thereof carries the greatest subtleties in a language, especially Chinese.
- The emphasis on the cultural and societal context of cross-language work has a parallel with the corresponding literature.
- In China, the media is largely controlled by the state with “Mind Politics” pervasive among them.
- Word2Vec is a recent computational linguistic tool which we can rely on.

Problems & Concludes

Problems:

- Dictionary Construction
- Dictionary validation
- Return association of the dictionary
- Politically inclined words and media bias

Concludes:

- We use machine learning built on “word embeddings” to develop our dictionary
- Examinations on firm-day level and article level are consistency with the literature, human reading and SVM prediction

Problems & Concludes

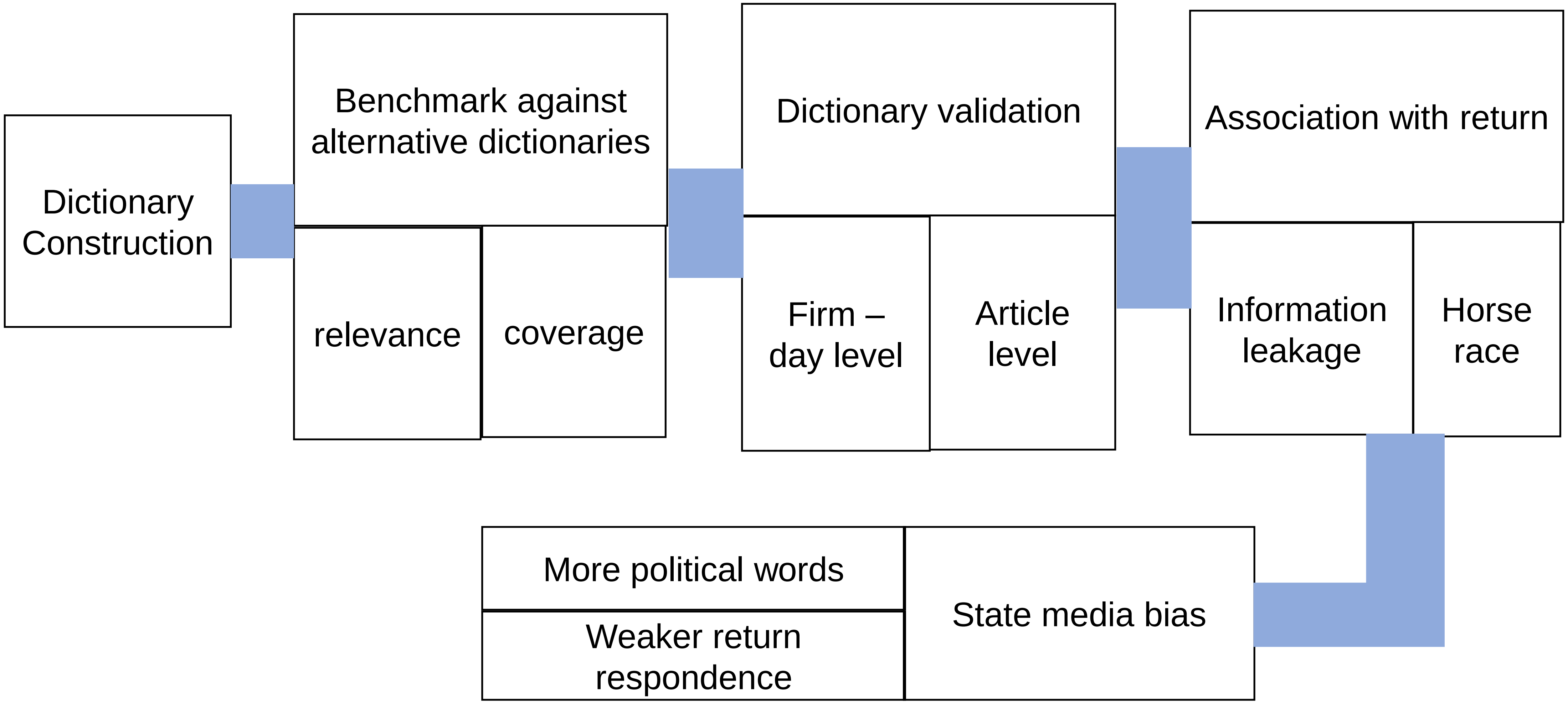
Concludes:

- Sentiment measures based on the dictionary are strongly related to the stock returns.
- The return association of the dictionary dominate those of the three alternative dictionaries.
- The list of political words has the potential to serve as an indicator for sentiment bias in Chinese news.

Contributions

- We create a dictionary of financial sentiment words for China based solely on financial news in Chinese.
- We use human expertise to discipline the approach with language and domain specificity taken into consideration.
- Return served as the driver of earlier study using machine learning while it serves as a manifestation of model(dictionary) efficacy.
- We are the first to create a list of politically inclined positive words that is separate from a generally positive word list.

Framework


Dictionary Construction
Benchmark against alternative dictionaries
relevance
coverage
Dictionary validation
Firm – day level
Article level
Association with return
Information leakage
Horse race
State media bias
More political words
Weaker return respondence

Data

- **Data**: All "Firm News," dated as far back as possible on the website, from 201301 to 201908.
- **Source**: finance.sina.com.cn
- **Filtering**: excludes the firm's public regulatory filings.
- **Plots**: Overall, there is more media coverage of stocks during market upswings and booms.

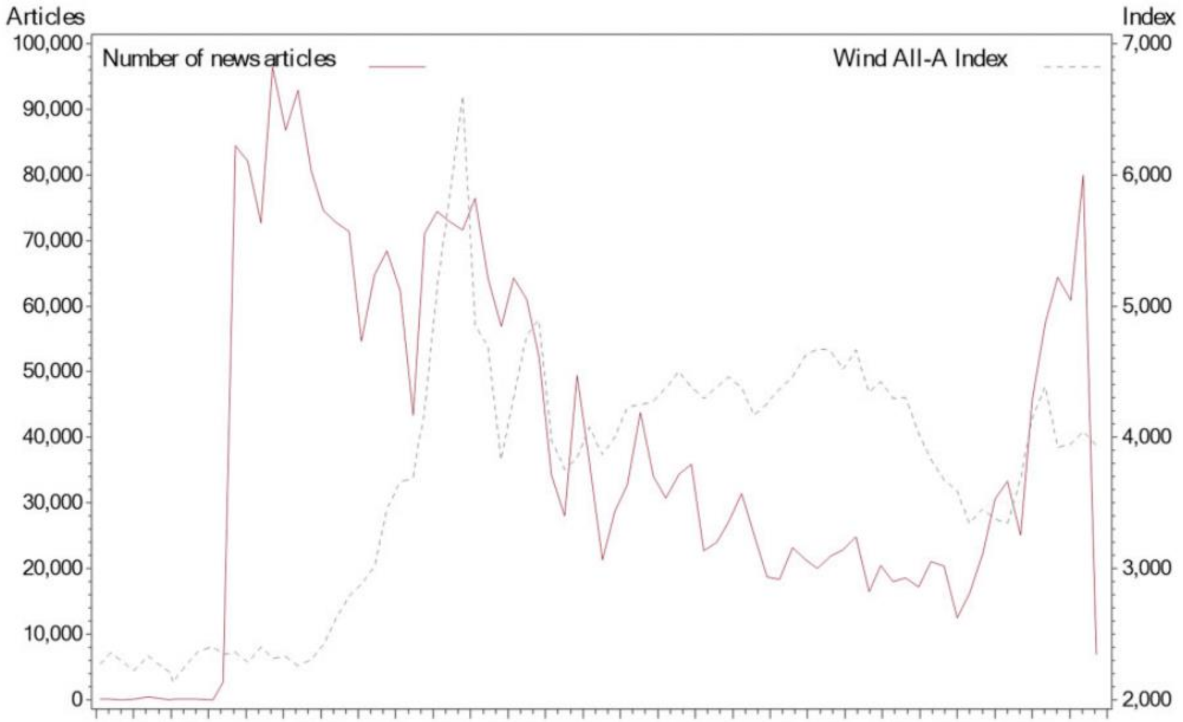


Fig1:Monthly number of articles

Approach

- **Word2Vec**: It employs a neural network algorithm to find semantically similar words to a given target word in a large body of text.
- **Steps**:
- Constructing a starting sentiment list of seed words from reading 500 randomly selected articles, iteratively.
- Using this starting list as an input to Word2Vec with the output of computationally close words.
- Employing human expertise to review the output and filter the semantically close words.
- Add to the starting dictionary the new converged word list before next iteration.

Approach

- **Augmentation**: To produce a robust dictionary, augment the process with 2 sets of seed words:
- (1) Another 2000 randomly selected articles (iteration 4)
- (2) Additional sentiment words from YZZ dictionary (iteration 5)

Panel B: Synonyms produced by Word2vec and human review

Iteration	Seed articles/ words	Number of firms	Number of news articles	Additional synonyms from Word2vec			Additional valid synonyms		
				Negative	Positive	Political	Negative	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

Table 1

Comparison with Existing Dictionaries

- **Existing Dictionaries**:
- (1)The direct Chinese translation of LM's dictionary with synonyms of the words.
- (2)YZZ dictionary.(You, Zhang and Zhang, 2018)
- (3)The set of common words shared by 3 general Chinese sentiment-word dictionaries.(not finance-specific)
- **Comparison**:
- (1)The size of our dictionary is much larger.(6600 words in total)
- (2)Our dictionary only has a limited degree of overlap with the existing dictionaries, especially the list of politically inclined positive words.

Comparison with Existing Dictionaries

Panel A: Total number of words in our dictionary				
	Negative	Positive	Political	Total
Total	2,986	2,235	1,439	6,660
Single character	70	12	3	85
Two-character	1,859	1,186	544	3,589
Three- and Four-character	1,039	1,016	877	2,932
Five-character and more	18	21	15	54

Panel B: Words in other dictionaries			
	Negative	Positive	Total
LM translation	1,337	327	1,664
YZZ dictionary	1,583	1,425	3,003
Generic Chinese dictionary	566	639	1,205
Total	3,034	2,108	5,130

Panel C: Overlapping of our dictionary with other dictionaries (percentage in parentheses)				
	Negative	Positive	Political	Total
LM 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 dictionary	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%)

Table 2

Sentiment Measures

- **Variables**:
- Neg_net: The number of negative-word occurrences minus positive-word occurrences, divided by the total number of words.
- Neg(Pos): The number of negative(positive)-word occurrences divided by the total number of words
- PoliticalPos: The ratio of political words to total words.
- **Summary**:
- (1)The mean of Neg_net is negative
- (2)The mean of PoliticalPos is larger than that of Neg.

Dictionary Validation – Firm-Day level

We regress the sentiment measures on a number of firm and news attributes.

- **Results**:
- (1) Firms that are smaller, have higher betas, or have higher BM ratios tend to exhibit a more negative news tone.(riskier and less profitable)
- (2) Firms with greater earnings surprise tend to exhibit a more positive tone.
- (3) Higher volatility and turnover both lead to less negative news.
- (4) More historical media coverage induces a more negative tone
- (5) Past returns of all horizons are significantly negatively related to Neg_net and positively related to Pos.

Dictionary Validation – Firm-Day level

	(1) Neg_net	(2) Neg	(3) Pos	(4) 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)

Table 3

Dictionary Validation – Firm-Day level

	(1) Neg_net	(2) Neg	(3) Pos	(4) PoliticalPos
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)

Table 3 continued

Dictionary Validation – Article level

- **Approaches**:
- (1) Manually reading 5000 news articles(until reaching an equal number of positive and negative articles), compare our dictionary-sentiment measure and human labels
- (2) Compare our dictionary-sentiment measure with traditional supervised machine-learning method of SVM
- (3) Compare our dictionary-sentiment measure with Wind Terminal labels.
- **Results:**
- Our dictionary based sentiment measure matches well with traditional sentiment label, with a more continuous measure for sentiment.

Return Association

We carry out an in-depth analysis of the association between news sentiment and past and future stock returns.

- **Dependent Variable**:
- Historical and future returns from day [–10] to day [10].
- **Control Variables:**
- FF3 of Chinese version and other past stock and news attributes, all measured at or before day [-11].
- **Independent Variables:**
- Sentiment measures mentioned before, all measured at day[0].

Return Association

- **Results**:
- The coefficient estimates of the control variables show a number of well-known patterns in Chinese markets.
- Our sentiment measures are associated with past and future returns:
- (1) Neg_net and Neg are negatively, and Pos is positively associated with excess returns from days[-10] to [1].
- (2) The positive association between PoliticalPos and returns is only positive and significant from days [-2] to [1], which is much weaker.

Return Association

	Industry- and size-adjusted return over day(s)								
	[−10, −6]	[−5, −3]	[−2]	[−1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Neg_net	−1.212***	−1.605***	−2.171***	−6.787***	−11.366***	−1.432***	0.096	0.558***	0.153***
	(−18.40)	(−19.33)	(−16.87)	(−38.43)	(−50.78)	(−11.66)	(0.80)	(8.57)	(3.11)

	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)

Table 4

Return Association

Neg_net is significantly related to returns from day[-10] to day[-1], with day[-1] being the most significant, this suggests that there may exist information leakage of news before it is released.

- **Approaches**:
- We control for potential persistence in news sentiment when examining the direction of causality between news releases and stock returns.
- We standardized each sentiment measure by subtracting the past average measure divided by the standard deviation.
- **Results:**
- The abnormal sentiment measures is significantly related to returns only on day[-1] and day[0], suggesting there exists some degree of news leakage.

Return Association

	Industry- and size-adjusted return over day(s)								
	[–10, –6]	[–5, –3]	[–2]	[–1]	[0]	[1]	[2]	[3, 5]	[6, 10]
Ab_Neg_net	0.006***	0.000	–0.006	–0.137***	–0.271***	0.001	0.033***	0.041***	0.020***
	(2.70)	(0.14)	(–1.23)	(–22.06)	(–36.48)	(0.29)	(6.75)	(17.03)	(10.47)
Ab_Neg	0.003	0.016***	0.030***	–0.037***	–0.200***	–0.008*	0.009**	0.019***	0.009***
	(1.39)	(6.03)	(5.89)	(–6.01)	(–30.30)	(–1.77)	(1.97)	(8.16)	(4.79)
Ab_Pos	–0.005**	0.008***	0.023***	0.139***	0.216***	–0.009*	–0.032***	–0.037***	–0.020***
	(–2.13)	(2.74)	(5.01)	(23.44)	(30.98)	(–1.94)	(–6.52)	(–14.81)	(–10.32)
Ab_PoliticalPos	–0.004*	–0.007**	0.003	0.054***	0.095***	0.011**	–0.009*	–0.012***	–0.007***
	(–1.88)	(–2.55)	(0.64)	(10.29)	(16.58)	(2.48)	(–1.74)	(–4.99)	(–4.01)

Table 5

Return Association – Horse Race

- **Approaches**:
- We include all those Neg_net_XXX from all the dictionaries used in the research in the return regression model above, the control variables are the same.
- **Results:**
- The statistic significance of Neg_net is preserved in the regressions.
- Signs of both Neg_net_LM and Neg_net_generic are reversed.
- Neg_net_YZZ remains significantly negative for all windows between days[-10,1] while the magnitude of its estimate is substantially smaller than that of its standalone regression.
- In sum, Neg_net subsumes all of the return relations of Neg_net_LM and Neg_net_generic, as well as much of that of Neg_net_YZZ.

Return Association – Horse Race

	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)

Table 6

Politically Inclined Words

- **Background**:
- Qin, Stromberg and Wu (2018) propose to measure media bias in China based on the coverage of “government mouthpiece” content.
- Piotroski, Wong and Zhang (2017) tag an article’s political bias by the frequency of political phrases in the Dictionary of Scientific Development.
- Both Piotroski, Wong and Zhang (2017) and YZZ report that state media outlets issue fewer negative corporate news articles
- Those authors also report that news by state media have lower value relevance.

Politically Inclined Words

- **Problems**:
- (1) Does state media use more politically inclined words and fewer negative words?
- (2) Do sentiment measures from state media exhibit a lower association with returns?

- **Results**:
- State media uses more politically inclined positive words and fewer negative words.
- Sentiment bias in PoliticalPos and Neg are largely uncorrelated.
- State media prefer to insert PoliticalPos words into its articles, and such insertions are not a simple substitution for Pos words.
- Stock prices respond less to sentiment words from state media as they do to non-state media.
- State media’s sentiment bias is distinct from its bias in mentions of political entities or nouns

Politically Inclined Words

	(1)	(2)	(3)
	PoliticalPos	Neg	MediabiasIndex
State media	0.258***	–0.271***	0.529***
	(9.68)	(–11.85)	(12.28)

(4)	(5)	(6)	(7)
	PoliticalPos	MediabiasIndex	
Pos	+ Pos	+ Pos	PoliticalNouns
–0.783***	–0.524***	–0.253***	0.615***
(–14.71)	(–7.39)	(–3.06)	(18.59)

Table 7

Politically Inclined Words

	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)

Conclusion

- **Conclusion**:
- Our study highlights the importance of language and domain specificity in dictionary-based sentiment analysis.
- The emphasis on language specificity reflects that the fields of economics and finance are related to culture.
- Our human-aided Word2vec method incorporates domain-specific knowledge in the construction of dictionary
- Compared with other machine-learning methods, sentiment judging from our dictionary is straightforward to implement, and is also better able to uncover the true sentiment of the text.