FINANCIAL FORECASTING WITH ALTERNATIVE DATA

Local Information Advantage and Stock Returns — Evidence from Social Media

光华管理学院
Guanghua School of Management

Feng Li

feng.li@gsm.pku.edu.cn

http://feng.li/

Authors

Yuqin Huang
(CUFE School of Finance)

Feng Li
(PKU Guanghua School of Management)

Tong Li
(XMU School of Economics)

Tse-Chun Lin
(HKU Business School)

Outline

Alternative data in finance?

- Alternative data research is transforming finance and investing by providing insights beyond traditional financial metrics.
- This approach leverages non-traditional data sources to deliver a broader understanding of market conditions, consumer behavior, and economic trends.
- Key sources of alternative data
  - Web Scraping and Online Data: Social media sentiment, News articles and blogs, Job postings and company reviews
  - Satellite and Geospatial Data: Traffic patterns and store parking lot analysis, Agricultural and weather data impacting commodities, Shipping and supply chain movements
  - Mobile App and Web Traffic: App downloads and engagement metrics, Website visitor data
  - Sensor & IoT Data: Smart device analytics (e.g., fitness trackers), Industrial production sensors
  - Consumer Reviews & Sentiment Analysis: Amazon, Yelp, and Glassdoor reviews, Survey and polling data

Are alternative data useful in financial forecasting?

- The availability of short-term-oriented data can induce forecasters to optimally shift their attention from the long term to the short term because it reduces the cost of obtaining short-term information. (Dessaint et al., 2024, JoF)
- Alternative data vs Multimodal Data
  - **Multimodal Data** refers to data collected from multiple sources or different types of modalities, such as text, images, video, and structured financial data.
  - While **alternative data focuses on finding new sources of insights**, **multimodal data focuses on combining different data types** to create a more holistic understanding of financial trends.

My research focus

I **tackle large scale forecasting challenges** by developing

- AI driven forecasting methods for large spatial structures,
- detecting for non-structural, noisy and intermittent signals in spatialtemporal data,
- efficient forecast combination and reconciliation methods, and
- open source solutions for large scale data.

Spatial information asymmetry for stock forecasting: a tale of two tastes

BBC Sign in Home News Sport Earth Reel W

NEWS

Home | Israel-Gaza war | War in Ukraine | Climate | Video | World | Asia | UK | Business | Tech

Asia | China | India

KFC sues Chinese firms over eight-legged chicken rumours

1 June 2015

KFC has over 4,000 restaurants in China

Source: https://www.bbc.co.uk/news/world-asia-china-32964606

Spatial information asymmetry for stock forecasting: a tale of two tastes

东方财富网 股吧首页 基金吧 话题 问董秘 人气榜

上海机场吧(600009) 32.78 ↓ -0.07 -0.21% A股市场人气排名第 995 名 详情>

全部 个人号 机构号 搜索该股票相关信息

阅读	评论	标题	作者	发帖时间
1281	5	上海机场、首都机场最新免税补充协议解读1228	相守湖畔	12-29 12:19
1575	2	中免跟上海机场的协议又重签了	乔令财经 V	12-28 08:55
536	0	上海机场(600009)：7家机构给予"买入"评级——签订免税补充协…	研报快读	12-28 14:16
2690	29	用数据说话	蠡江李二段	12-27 20:11
2119	11	浦东国际机场11月飞机起降量39170架次，同比增长126.26%；…	生意善贾田头草民	12-20 09:59
1440	4	外资成本这么低的吗	yuhun4248	12-20 17:37
1771	4	您还记得"2021年8月11的上海机场"吗?	北京四个石头	12-18 21:59
2112	21	随笔：又一天	看晚霞的无业游民	12-18 15:32
1117	5	民众愤怒：上海机场"区别对待"事件引发网络热议黔S2023-12-1…	股友329uZ99585	12-15 15:01
343	1	上海机场：浦东国际机场11月旅客量同比增长307%	完美的leng	12-14 15:41
1973	6	哈哈，还有故人么?	懒懒的看股	12-09 22:55
3455	12	营收增速翻倍，上海机场摆脱困境，这位置我已经看不懂了	地铁恒道第一人	12-08 17:11
2732	6	明年上机业绩会大幅度增长，按理说机构应该买入做预期，可明…	忠实的海勒	12-04 17:32
976	4	9家机构给予"买入"评级——旺季营收维持高增，盈利水平修复持续	研报快读	11-28 14:46
537	1	申能携手上海机场，绿色能源双丰收!	动态宝	11-24 17:25
677	0	散户要当机立断抛弃流通值几百亿以上的大盘股，抛弃高位高价…	中国悟空	11-22 08:19
1705	3	上海机场(600009)估值分析：4家机构认为"低估"——生产恢复逐…	脱水报告	11-06 13:50
3589	32	周日给大家吃个饼，就yy一下。周一停牌，周二复牌顶10cm板…	留一手吧	11-05 09:38

Source: https://guba.eastmoney.com/list,600009.html

Local investors and information advantage

- Local investors may enjoy an information advantage by gaining access to information earlier that distant investors (Chi & Shanthikumar, 2017).
- After receiving information about a firm, local investors may want to communicate more with others about this particular stock (Hirshleifer, 2020).
- The relative intensity of investors' posting activities likely reflects local investors' information advantage (Ferreira et al., 2017).

The stock posting data

- Our data for message postings come from Guba Eastmoney, China's preeminent stock mes- sage board.
- Guba Eastmoney allows users to read and post messages and it identifies these users through **non-confidential IP addresses**.
- This unique feature enables us to distinguish between local and nonlocal posters and explore the hypothesized local information advantage via message postings.
- Our analysis covers more than **300 million postings** pertaining to **2,239 listed firms** in Chinese A-share markets ranging over 6 year (ca **200 GB raw data**).

Abnormal posting measure

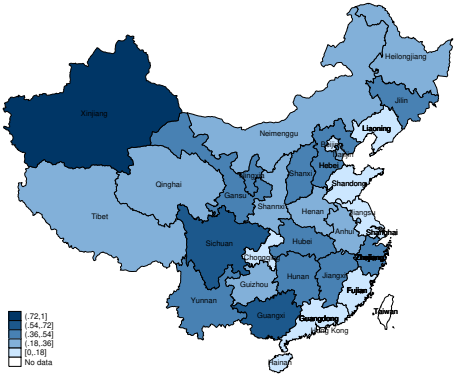
- We define **relative postings** (RP) to measure the relative strength of posting activities by locals and nonlocals. For firm i headquartered in city c, its relative postings measure in week t is calculated as:

$$RP_{i_c,t} = \ln\left(1 + P^{c}_{i_c,t}\right) - \ln\left(1 + P^{-c}_{i_c,t}\right)$$

  where $P^{c}_{i_c,t}$ ($P^{-c}_{i_c,t}$) is local (nonlocal) postings, that is, the total number of messages posted in week t by investors in (outside) city c.
- RP has a conceptional similarity to TF-IDF (Term Frequency-Inverse Document Frequency) in the NLP domain.
- To measure unusual changes in relative postings, we construct **abnormal relative postings** (ARP)

$$ARP_{i_c,t} = RP_{i_c,t} - \text{median}\left(RP_{i_c,t-1}, RP_{i_c,t-2}, \cdots, RP_{i_c,t-10}\right)$$

ARP-based portfolio performance



- Firms in every province are sorted into quintile portfolios based on their ARP measure.
- ARP-based trading strategy is more profitable in under-developed inland regions where firms are relatively opaque.

Forecast the excess return

- We forecast the excess return with Fama & MacBeth (1973) models

$$\boldsymbol{R}_{i,c,t+1} = \alpha + \beta \boldsymbol{ARP}_{i,c,t} + \delta \boldsymbol{X}_{i,t} + \epsilon_{i,t+1}$$

  by identifying $\boldsymbol{ARP}_{i,t}$ (**abnormal relative postings**) of firm i in week t related to its headquartered city c; and $\boldsymbol{X}_{i,t}$ is a vector of **firm-level characteristics**.
- Complex forecasting models with similar firm-level variables have been used (Li et al., 2010; Villani et al., 2012; Li & Villani, 2013) but are computationally intensive.

Forecasting at scale: the new normal

- **Large scale data generally require distributed solutions**.
  - IP address, city and firm match for each post is a standard MapReduce task.
  - Both RP and ARP calculations require iterating over all 300 million text data.
  - Without a distributed solution, this work would take weeks to finish (Just reading the 130 GB data into memory takes one hour).
- **Many simple models ensemble a powerful solution** instead of one complex model for everything.
- **Interpretability counts** when choosing appropriate models.

Variable	Definition
Posting Variables	
RP	Relative postings, defined as the logarithm of one plus the number of messages from local posters minus the logarithm of one plus the number of messages from nonlocal posters
ARP	Abnormal relative postings, defined as relative postings for a firm in one week minus the median value of its relative postings in the previous ten weeks
Other Variables	
AG	Asset growth, defined as the annual growth rate of total assets
ALMedia	Abnormal local media coverage, defined as local media coverage on a firm in a given week minus the median value of its local media coverage in the previous ten weeks
BM	Book-to-market ratio, defined as the book value of equity divided by market value of equity
EmpShare	Share of industry employees, defined as the total number of employees in an industry in a given city divided by the total number of employees in the city
ILLIQ	Illiquidity measure, defined as the weekly average of the ratio of absolute daily price change to daily trading volume
IO	Institutional ownership, defined as percentage of shares outstanding owned by institutional investors
IVOL	Idiosyncratic volatility, defined as the standard deviation of residuals from the Carhart (1997) four-factor model
Log(Analysts)	Analyst coverage, defined as logarithm of one plus the number of analysts covering the firm in a given week
Log(GDP)	Logarithm of annual GDP per capita (RMB) of a city
NPR	Net purchase ratio, defined as the number of purchases minus the number of sales divided by the total number of transactions by managers and large shareholders of a firm in a given week
PopDensity	Population density of the firm's headquarters city
$\text{Ret}_{t-4:t-1}$	Cumulative return from week $t-4$ to week $t-1$
$\text{Ret}_{t-52:t-5}$	Cumulative return from week $t-52$ to week $t-5$
ROA	Return on assets, defined as net income divided by total assets
Size	Firm size, defined as the logarithm of market capitalization

	Ret_{t+1}			Ret_{t+2}	Ret_{t+4}	Ret_{t+6}	Ret_{t+8}	Ret_{t+12}
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ARP	0.91***	0.86***	0.81***	0.39***	0.38**	0.06	0.13	0.05
	(5.51)	(5.61)	(5.39)	(2.88)	(2.57)	(0.36)	(0.94)	(0.32)
Size		−0.14***	−0.05	−0.09*	−0.09*	−0.05	−0.08	−0.06
		(−3.11)	(−1.01)	(−1.72)	(−1.77)	(−1.12)	(−1.65)	(−1.33)
BM		0.06	0.04	−0.02	0.02	−0.00	−0.01	0.00
		(0.92)	(0.75)	(−0.29)	(0.34)	(−0.07)	(−0.14)	(0.04)
$\text{Ret}_{t-4:t-1}$		−0.05***	−0.03***	−0.05***	−0.04***	−0.02**	−0.02**	−0.01
		(−3.99)	(−2.95)	(−4.39)	(−3.87)	(−2.57)	(−2.18)	(−0.89)
$\text{Ret}_{t-52:t-5}$		−0.07***	−0.04	−0.04*	−0.05*	−0.04	−0.06**	−0.06*
		(−2.72)	(−1.49)	(−1.67)	(−1.71)	(−1.53)	(−2.12)	(−1.92)
AG			−0.06	−0.06	−0.11**	−0.15***	−0.16***	−0.14***
			(−1.18)	(−1.51)	(−2.56)	(−3.37)	(−4.13)	(−3.32)
ROA			0.10	−0.06	−0.62	−0.68	−0.40	−0.45
			(0.22)	(−0.13)	(−1.43)	(−1.46)	(−0.91)	(−0.99)
IVOL			−0.13***	−0.08**	−0.03	−0.03	−0.01	−0.03
			(−3.57)	(−2.22)	(−0.99)	(−0.81)	(−0.24)	(−0.85)
ILLIQ			0.38***	0.18***	0.11**	0.16***	0.07**	0.14***
			(8.09)	(4.80)	(2.42)	(3.62)	(1.98)	(3.93)
(Other variables truncated ...)								
Intercept	−0.06	0.91	0.44	0.82	1.06	0.47	0.86	0.80
	(−0.18)	(1.63)	(0.67)	(1.16)	(1.47)	(0.69)	(1.23)	(1.17)
Obs	303,361	303,361	303,361	293,425	279,472	275,838	272,375	265,509
Adj. R^2	0.05%	3.60%	6.39%	5.90%	5.67%	5.27%	5.07%	5.00%

Sentiment in stock markets

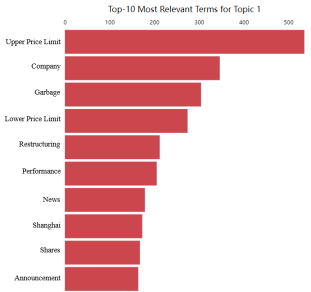
Source: https://markets.businessinsider.com/news/stocks/bullish-stock-market-signal-zweig-breadth-thrust-indicator-just-flashed-2023-4

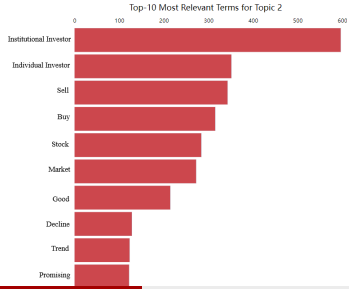
Sentiment of local and nonlocal postings

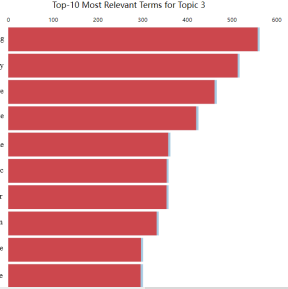
- **The sentiment score is efficiently calculated with the distributed MapReduce framework**.
- We first segment each sentence in a posting into words. Next, we identify sentiment words based on a prespecified sentiment dictionary.
- For words with a positive (negative) tone, we assign a base score of 1(-1). The base score is further weighted according to its modifier words, with weights of 4, 3, 2, and 0.5 for the extreme, strong, moderate, and mild degrees, respectively.
- If a negative word precedes a key sentiment word, we multiply the weighted sentiment score by -1.

Year	**Local Post Sentiment**	**Non-local Post Sentiment**	**Local – Non-local**	*p* value
2007	0.0303	0.0155	0.0148***	0.000
2008	0.0383	0.0216	0.0167***	0.000
2009	0.0319	0.0160	0.0160***	0.000
2010	0.0533	0.0451	0.0083***	0.000
2011	0.1107	0.0646	0.0461***	0.000
2012	0.1556	0.0894	0.0662***	0.000
2013	0.1357	0.1063	0.0294***	0.000
All	0.0902	0.0541	0.0361***	0.000

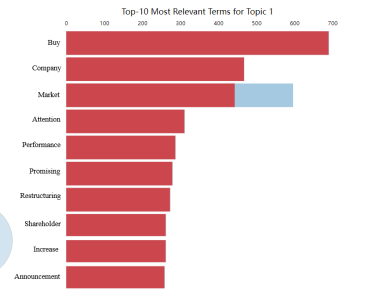
Topical analysis of local posts

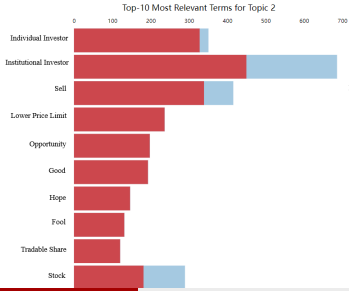


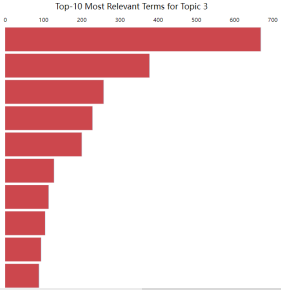




Topical analysis of nonlocal posts







```
# Trains a LDA model with Spark.
from pyspark.ml.clustering import LDA
# Loads data.
dataset = spark.read.format("csv").load("stockdata/*.csv")
lda = LDA(k=4, maxIter=100)
model = lda.fit(dataset)
ll = model.logLikelihood(dataset)
lp = model.logPerplexity(dataset)
# Describe topics.
topics = model.describeTopics(3)
print("The topics described by their top-weighted terms:")
topics.show(truncate=False)
# Shows the result
transformed = model.transform(dataset)
transformed.show(truncate=False)
```

Stock returns: messages with different topics

Topic	Fundamentals (1)	Trading (2)	Noises (3)	Insider (4)
ARP	1.18***	0.68***	0.80	0.33
	(5.96)	(3.08)	(0.19)	(1.47)
Size	-0.05	-0.05	-0.05	-0.05
	(-1.05)	(-1.08)	(-0.99)	(-1.03)
BM	0.05	0.05	0.04	0.05
	(0.84)	(0.82)	(0.79)	(0.83)
$Ret_{t-4:t-1}$	-0.03***	-0.03***	-0.04***	-0.03***
	(-2.84)	(-2.81)	(-3.05)	(-2.95)
$Ret_{t-52:t-5}$	-0.04	-0.04	-0.04	-0.04
	(-1.60)	(-1.57)	(-1.51)	(-1.56)
AG	-0.06	-0.06	-0.06	-0.06
	(-1.21)	(-1.25)	(-1.16)	(-1.12)
ROA	0.12	0.11	0.12	0.10
	(0.26)	(0.25)	(0.25)	(0.23)
IVOL	-0.13***	-0.13***	-0.13***	-0.13***
	(-3.57)	(-3.53)	(-3.56)	(-3.52)
ILLIQ	0.38***	0.38***	0.38***	0.38***
	(8.17)	(8.16)	(8.23)	(8.15)
IO	0.04	0.04	0.04	0.04
	(0.27)	(0.29)	(0.25)	(0.24)
NPR	0.33*	0.35**	0.36**	0.36**
	(1.91)	(2.03)	(2.09)	(2.12)
(Other variables truncated ...)				
Intercept	0.42	0.44	0.43	0.42
	(0.65)	(0.67)	(0.65)	(0.64)
Obs	303,361	303,361	303,361	303,361
Adj. R^2	6.45%	6.45%	6.39%	6.39%

Parallel processing

```
#!/bin/bash -l

#SBATCH -J Stocks
#SBATCH -N 6          # Number of nodes
#SBATCH -p MCMC       # Partition Used.
#SBATCH -t 10-00:00 # Runtime in D-HH:MM
#SBATCH --mail-type=FAIL
#SBATCH --array=1-100%16 # Run a job array

for STOCK in shanghai shenzhen chuangyeban zhongxiaoban
do
    srun python3 main.py ${STOCK} ${STOCK}.csv  $SLURM_ARRAY_TASK_ID
done
```

Working in progress

- **Interpretable video time series forecasting**
- **Forecasting methods**
  - Forecasting reconciliation with large hierarchical structures
  - Multimodal time series forecasting
- **Final thoughts**
  - The future of financial forecasting is **multimodal alternative data** — combining diverse data sources to uncover hidden insights.
  - With advancements in AI, NLP, and deep learning, financial firms can better leverage multimodal alternative data to predict trends, assess risks, and make smarter investment decisions.

Thank you!

https://feng.li
feng.li@gsm.pku.edu.cn

The 45th International Symposium on Forecasting

Beijing, June 29 - July 2, 2025

ISF 2025

The International Symposium on Forecasting (ISF) is the premier forecasting conference, attracting the world's leading forecasting researchers, practitioners, and students. The first ISF took place in 1981 in Quebec City, Canada. Over its 45-year history, the symposium has gathered in Europe, Asia-Pacific, and North and South America and hosted highly respected experts in forecasting, including many Nobel laureates, through keynote speaker presentations, academic sessions, workshops, and networking in social programs and events.

General Chair: Yongmiao Hong
Co-Chairs: Tao Hong, Lean Yu
Local Chairs: Zhengjun Zhang, Feng Li
Program Chairs: Jue Wang, Yanfei Kang

Organizers

The AMSS center for Forecasting Science, Chinese Academy of Sciences
School of Economics and Management, Beihang University
Guanghua School of Management, Peking University
School of Economics and Management, University of Chinese Academy of Sciences

Academic Program

The 45th ISF will be held in Beijing, June 29 - July 2, 2025, Keynote speeches will address interdisciplinary topics from economics, statistics, computer science and data science. There will be a variety of choices of special sessions, workshops and summer schools to facilitate in-depth communication on specific topics and foster exchanges and communications.

Deadlines

March 7, 2025:	Invited session proposals
March 21, 2025:	Abstract submissions
April 4, 2025:	Notification of abstract acceptance/rejection
May 1, 2025:	Registration deadline for accepted abstracts

For further information, please consult:
ISF2025@hotmail.com

Local Organizer
The AMSS Center for Forcasting Science, CAS (CEFS)
http://cefs.amss.ac.cn/

References

Dessaint, O., Foucault, T. & Fresard, L. (2024). "Does Alternative Data Improve Financial Forecasting? The Horizon Effect". *The Journal of Finance* 79.(3), pp. 2237–2287.

Hirshleifer, D. (2020). "Presidential Address: Social Transmission Bias in Economics and Finance". *Journal of Finance* 75.(4), pp. 1779–1831.

Chi, S. S. & Shanthikumar, D. M. (2017). "Local Bias in Google Search and the Market Response around Earnings Announcements". *Accounting Review* 92.(4), pp. 115–143.

Ferreira, M. A., Matos, P., Pereira, J. P. & Pires, P. (2017). "Do Locals Know Better? A Comparison of the Performance of Local and Foreign Institutional Investors". *Journal of Banking and Finance* 82, pp. 151–164.

Li, F. & Villani, M. (2013). "Efficient Bayesian Multivariate Surface Regression". *Scandinavian Journal of Statistics* 40.(4), pp. 706–723.

Villani, M., Kohn, R. & Nott, D. J. (2012). "Generalized Smooth Finite Mixtures". *Journal of Econometrics* 171.(2), pp. 121–133.

Li, F., Villani, M. & Kohn, R. (2010). "Flexible Modeling of Conditional Distributions Using Smooth Mixtures of Asymmetric Student t Densities". *Journal of Statistical Planning and Inference* 140.(12), pp. 3638–3654.

Carhart, M. M. (1997). "On Persistence in Mutual Fund Performance". *Journal of Finance* 52.(1), pp. 57–82.

Fama, E. F. & MacBeth, J. D. (1973). "Risk, Return, and Equilibrium: Empirical Tests". *Journal of Political Economy* 81.(3), pp. 607–636.