基于另类文本数据的金融预测

光华管理学院
Guanghua School of Management

李丰

feng.li@gsm.pku.edu.cn

http://feng.li/

研究学者团队

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

1. 金融预测中的另类数据研究?
2. 用于股票预测的空间信息不对称
3. 异常发帖指标与可预测性
4. 情绪与主题
5. 其他类型的另类数据?

金融中的另类数据？

- 另类数据研究正在通过提供超越传统财务指标的洞见，改变金融与投资领域。
- 这种方法利用非传统数据源，以更全面地理解市场状况、消费者行为和经济趋势。
- 另类数据的主要来源：
  - 网络爬取与在线数据：社交媒体情绪，新闻文章与博客，招聘信息与公司评价
  - 卫星与地理空间数据：交通模式与商场停车场分析，影响大宗商品的农业与气象数据，船运与供应链动态
  - 移动应用与网站流量：应用下载量与用户参与度，网站访问数据
  - 传感器与物联网数据：智能设备分析（如健身追踪器），工业生产传感器
  - 消费者评论与情感分析：亚马逊、Yelp 与 Glassdoor 评论，调查与问卷数据

另类数据对金融预测有用吗?

- 短期导向数据的可获得性可能促使预测者将注意力从长期转向短期，因为它降低了获取短期信息的成本。(Dessaint et al., 2024, JoF)
- 另类数据 vs 多模态数据
  - 多模态数据指从多个来源或不同模态类型（如文本、图像、视频、结构化财务数据）中收集的数据。
  - 另类数据关注的是寻找新的洞察来源，而多模态数据关注的是将不同类型的数据结合起来，以构建对金融趋势更全面的理解。

我的研究重点

我通过以下研究方向来应对大规模预测挑战：

- 针对大规模空间结构的人工智能驱动预测方法；
- 检测空间–时间数据中的非结构化、噪声性与间歇性信号；
- 高效的预测组合与协调方法；
- 面向大规模数据的开源解决方案。

用于股票预测的空间信息不对称：两种口味的故事

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

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

用于股票预测的空间信息不对称：两种口味的故事

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	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

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

本地投资者与信息优势

- 本地投资者可能因为更早获取信息而拥有信息优势。(Chi & Shanthikumar, 2017)
- 在获得公司相关信息后，本地投资者可能会更积极地与他人交流该股票。(Hirshleifer, 2020)
- 投资者发帖活动的相对强度很可能反映了本地投资者的信息优势。(Ferreira et al., 2017)

股票发帖数据

- 我们的发帖数据来自东方财富股吧，这是中国最具代表性的股票论坛。
- 东方财富股吧允许用户阅读与发布帖子，并通过非保密的 **IP** 地址识别用户。
- 这一独特特征使我们能够区分本地与非本地发帖者，并通过帖子数据探索本地信息优势的假设。
- 我们的分析涵盖超过 **3** 亿条帖子，涉及中国 A 股市场中 **2,239** 家上市公司，时间跨度约 6 年（原始数据约 **200 GB**）。

- 我们定义相对发帖量（RP）以衡量本地与非本地投资者发帖活动的相对强度。对于总部位于城市 c 的公司 i，其在第 t 周的相对发帖指标计算如下：

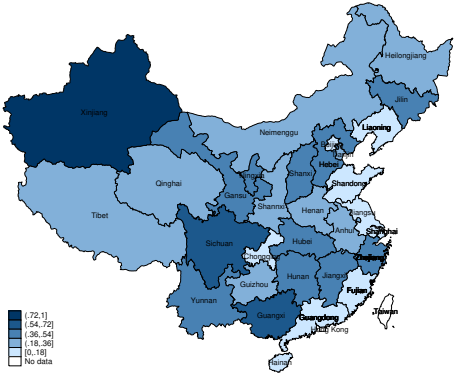
$$RP_{i_c,t} = \ln\left(1 + P^{c}_{i_c,t}\right) - \ln\left(1 + P^{-c}_{i_c,t}\right)$$

其中，$P^{c}_{i_c,t}$（$P^{-c}_{i_c,t}$）分别表示本地（非本地）发帖量，即第 t 周内来自城市 c（城市外）投资者发布的总帖子数。

- RP 的概念与自然语言处理领域中的 TF-IDF（词频–逆文档频率）具有相似性。
- 为衡量相对发帖量的异常变化，我们构造异常相对发帖量（ARP）：

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

基于 ARP 的投资组合表现



- 我们根据公司所在省份的 ARP 指标，将其划分为五分位投资组合。
- 基于 ARP 的交易策略在欠发达的内陆地区更具盈利性，因为这些地区的公司信息相对不透明。

- 我们基于 Fama & MacBeth (1973) 模型对超额收益进行预测：

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

  其中，$\boldsymbol{ARP}_{i,t}$（异常相对发帖量）表示公司 i 在第 t 周、与其总部城市 c 相关的指标；$\boldsymbol{X}_{i,t}$ 为公司层面的特征向量。
- 类似的复杂预测模型（使用公司层面变量）曾被用于相关研究 (Li et al., 2010; Villani et al., 2012; Li & Villani, 2013)，但计算代价较高。

大规模预测：新的常态

- 大规模数据通常需要分布式计算解决方案。
  - 对每条帖子进行 IP 地址、城市与公司匹配是一项标准的 MapReduce 任务。
  - RP 与 ARP 的计算都需要遍历全部 3 亿条文本数据。
  - 如果没有分布式解决方案，该过程可能需要数周完成（仅将 130 GB 数据读入内存就需约一小时）。
- 多个简单模型的集成往往优于单一复杂模型。
- 在选择合适模型时，可解释性同样重要。

变量构造

变量	定义
发帖相关变量	
RP	相对发帖量，定义为：本地投资者发帖数加一取对数，减去非本地投资者发帖数加一取对数
ARP	异常相对发帖量，定义为：某公司在某一周的相对发帖量，减去其过去十周相对发帖量的中位数
其他变量	
AG	资产增长率，定义为总资产的年度增长率
ALMedia	异常本地媒体报道，定义为某公司在某周的本地媒体报道量减去其过去十周本地媒体报道量的中位数
BM	账面市值比，定义为股东权益账面价值与市场价值之比
EmpShare	行业雇员占比，定义为某城市中某行业雇员总数除以该城市雇员总数
ILLIQ	非流动性指标，定义为每日价格变动绝对值与每日成交量之比的周平均值
IO	机构持股比例，定义为机构投资者持有的流通股占比
IVOL	特质波动率，定义为基于 Carhart (1997) 四因子模型残差的标准差
Log(Analysts)	分析师覆盖度，定义为一加上在某周跟踪该公司的分析师数量的对数
Log(GDP)	城市人均 GDP（人民币）的年度对数值
NPR	净买入比，定义为公司管理层及大股东在某周买入笔数减去卖出笔数，再除以总交易笔数
PopDensity	公司总部所在城市的人口密度
$\text{Ret}_{t-4:t-1}$	第 $t-4$ 周至第 $t-1$ 周的累计收益率
$\text{Ret}_{t-52:t-5}$	第 $t-52$ 周至第 $t-5$ 周的累计收益率
ROA	资产回报率，定义为净利润与总资产之比
Size	公司规模，定义为市值的对数

预测结果

	未来 1 周收益（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)
$Ret_{t-4:t-1}$		−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)
$Ret_{t-52:t-5}$		−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)
(其他变量省略…)								
截距项	−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)
样本量（Obs）	303,361	303,361	303,361	293,425	279,472	275,838	272,375	265,509
调整后的 R^2	0.05%	3.60%	6.39%	5.90%	5.67%	5.27%	5.07%	5.00%

股票市场中的情绪

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

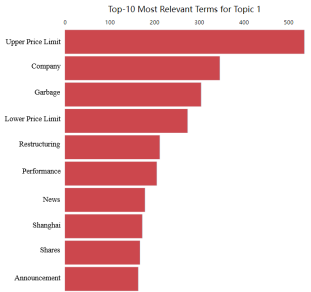
本地与非本地发帖的情绪分析

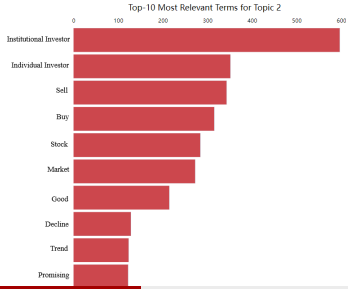
- 情绪得分通过分布式 **MapReduce** 框架高效计算。
- 我们首先将每条发帖的句子进行分词。然后基于预先定义的情感词典识别情绪词。
- 对具有正面（负面）语气的词赋予基础得分 1（-1）。该基础得分会根据修饰词的强度进行加权，极强、强烈、中等、轻微程度的权重分别为 4、3、2 和 0.5。
- 若负面词出现在关键情绪词之前，则将加权情绪得分乘以 -1。

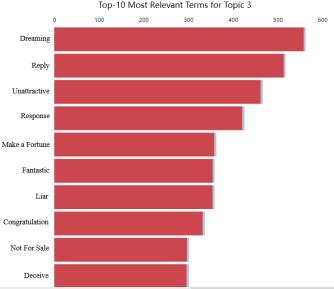
年份	本地发帖情绪	非本地发帖情绪	本地−非本地差值	p 值
2007	0.0303	0.0155	0.0148***	0.000
2008	0.0383	0.0216	0.0167***	0.000
2009	0.0319	0.0160	0.0160***	0.000
2010	0.0533	0.0451	0.0083***	0.000
2011	0.1107	0.0646	0.0461***	0.000
2012	0.1556	0.0894	0.0662***	0.000
2013	0.1357	0.1063	0.0294***	0.000
总体	0.0902	0.0541	0.0361***	0.000

本地发帖的主题分析


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%

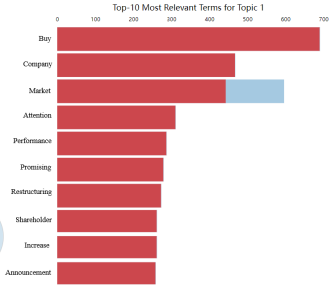

Top-10 Most Relevant Terms for Topic 1
0
100
200
300
400
500
Upper Price Limit
Company
Garbage
Lower Price Limit
Restructuring
Performance
News
Shanghai
Shares
Announcement

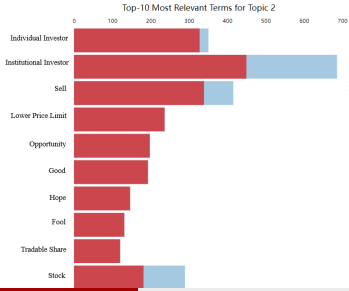

Top-10 Most Relevant Terms for Topic 2
0
100
200
300
400
500
600
Institutional Investor
Individual Investor
Sell
Buy
Stock
Market
Good
Decline
Trend
Promising

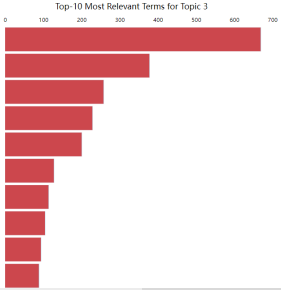

Top-10 Most Relevant Terms for Topic 3
0
100
200
300
400
500
600
Dreaming
Reply
Unattractive
Response
Make a Fortune
Fantastic
Liar
Congratulation
Not For Sale
Deceive

非本地发帖的主题分析


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%


Top-10 Most Relevant Terms for Topic 1
0
100
200
300
400
500
600
700
Buy
Company
Market
Attention
Performance
Promising
Restructuring
Shareholder
Increase
Announcement


Top-10 Most Relevant Terms for Topic 2
0
100
200
300
400
500
600
700
Individual Investor
Institutional Investor
Sell
Lower Price Limit
Opportunity
Good
Hope
Fool
Tradable Share
Stock


Top-10 Most Relevant Terms for Topic 3
0
100
200
300
400
500
600
700
Uptrend
Garbage
Resume Trading
Agree
Reply
No Way
Unwinding
Support
Admire
Reason

```
# 使用 Spark 训练 LDA（主题模型）
from pyspark.ml.clustering import LDA
# 加载数据
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)
# 描述主题
topics = model.describeTopics(3)
print("按最高权重词描述的主题如下：")
topics.show(truncate=False)
# 显示结果
transformed = model.transform(dataset)
transformed.show(truncate=False)
```

股票收益：不同主题下的发帖

主题	基面（Fundamentals）	交易（Trading）	噪声（Noises）	内部信息（Insider）
	(1)	(2)	(3)	(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)
(其他变量省略…)				
截距项	0.42	0.44	0.43	0.42
	(0.65)	(0.67)	(0.65)	(0.64)
样本量（Obs）	303,361	303,361	303,361	303,361
调整后的 R^2	6.45%	6.45%	6.39%	6.39%

并行处理示例

```
#!/bin/bash -l

#SBATCH -J Stocks
#SBATCH -N 6          # 节点数量
#SBATCH -p MCMC       # 使用的分区
#SBATCH -t 10-00:00 # 运行时间（天-小时：分钟）
#SBATCH --mail-type=FAIL
#SBATCH --array=1-100%16 # 作业数组

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

其他类型的另类数据?

- 可解释的视频时间序列预测
- 预测方法
  - 面向大规模层级结构的预测协调（Forecast Reconciliation）
  - 多模态时间序列预测
- 一些思考
  - 金融预测的未来是多模态另类数据——结合多样化数据源以揭示隐藏的信息与洞见。
  - 随着人工智能、自然语言处理和深度学习技术的进步，金融机构将能够更好地利用多模态另类数据来预测趋势、评估风险，并做出更智能的投资决策。

谢谢！

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

参考文献

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.