The Colour of Finance Words

Diego Garcia, Xiaowen Hu, Maximilian Rohrer

JFE

胡震霆 2023/3/08

研究背景

Tetlock(2007)利用Harvard-IV金融词典研究了新闻媒体的文本信息，此后金融和会计领域对文本化数据的研究变得十分火热；

Loughran and McDonald(2011)研究了年度财务报表的文本情绪并在此基础上完善了Harvard-IV词典，提出LM金融词典。当前研究衡量文本情绪的主要方式为计算情感词汇出现的频率，即“词袋法”，依照的字典为LM金融词典。

另外一些研究，如Gentzkow(2019)指出上述方式在一定程度上弱于复杂的机器学习模型方法。

本文核心问题是：通过机器识别构建的字典与人工识别的字典更能准确地衡量情绪并影响股价变动吗？

研究贡献

- 为研究如何衡量文本情绪的文献提供了一种新的方式
- 利用机器学习算法构建了新的词典，既包含常见的一元词汇，也包含多元词汇，对存在情感歧义的一元词汇进行了有效补充

研究主题

- 构造一组新的金融词典(ML词典)，充分衡量金融文本的情绪。
- 比较不同词典在分析情绪文本发布对股价变动的影响时的表现（主要实证结果）
- ML词典的广度、ML词典对于LM词典的优势、二元词在词典中的作用（进一步分析）

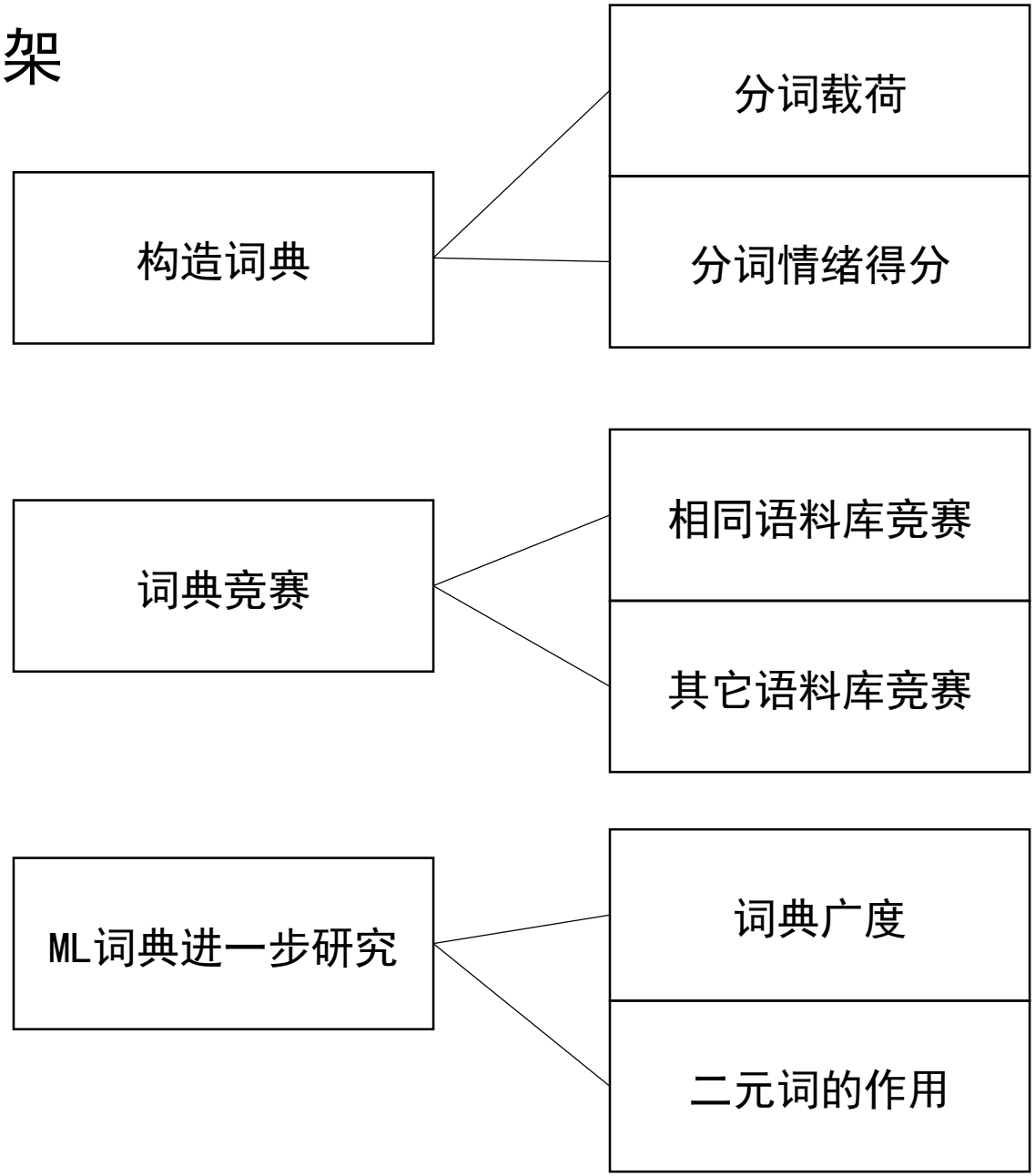
研究结论

- 本文采用多项式逆回归对样本进行有监督学习，最终构造出的ML词典包含121个一元词汇和725个二元词汇，该词典满足“plain money English ”的目标。
- 在利用文本情绪对股价变动的样本外预测中，ML词典构造的情绪指标在系数大小，显著性以及R^2上都优于LM词典（主要实证结果）
- （1）仅包含一元词和二元词的ML词典比LM词典能够很好地覆盖整个语料库；

  （2）ML词典对词汇赋予情绪色彩的方式与LM有很大区别；

  （3）二元词有助于消除一元词的情感歧义

研究框架


构造词典
分词载荷
分词情绪得分
词典竞赛
相同语料库竞赛
其它语料库竞赛
ML词典进一步研究
词典广度
二元词的作用

数据

数据源：

电话会议：Seeking Alpha & Wall Street Horizons

年度报表：Bill McDonald’s webpage

WSJ报道：Factiva(人工下载)

数据选取：

选择电话会议文稿作为主要研究对象，主要原因为其更高的信噪比，且电话会议在时间上优先于年度报表和WSJ报道。

数据特点：

电话会议文本包含更多的口语词汇；年报文本包含更多专业术语；WSJ表达方式更具故事性。

数据

描述性统计：

三种语料中，年报文本长度远超过另外两类语料，电话会议文本对应的上市公司在3000家左右，年报文本在10000家左右，WSJ文本对应189家。

Earnings calls	
Start	13.10.2005
End	07.10.2020
Unique firms	3229
Observations	85,530
Average words per document	3130
Annual reports (10-K)	
Start	02.01.1996
End	27.12.2018
Unique firms	10,076
Observations	76,922
Average words per document	17,294
Wall Street Journal (WSJ)	
Start	03.01.2000
End	31.12.2021
Unique firms	189
Unique articles	144,383
Observations (firm-days)	87,198
Average words per document	457

分词载荷

- 主要思想：

将发布日前后4天的收益率作为标签，对文本数据进行回归，得到文本中每一个分词的情感载荷，根据情感载荷将其分为积极与消极词汇，选取情感更强烈的词汇进入ML词典。

- 文档-分词矩阵：

	Term1	**T2**	**…**	**Tp**
Document1				
D2				
…				
Dn				

行向量表示单个文档内所有高频词出现的次数；列向量表示单个高频词在所有文档中出现的次数；对于一元词，高频词数目选取为p = 16384；对于二元词，p = 65536。

分词载荷

- 核心输出：

每个高频词的对收益率回归的载荷，根据载荷大小和符号将高频词均分为积极、中性、消极，赋值分别为1，0，-1.

- 创新：

上述方法会导致过拟合，产生过量情感词汇；现加入一层额外的“交叉验证层”。

假设原始训练集容量为m，将m分为k个容量为q的子训练集，对k个子集进行随机无放回抽样，在被抽样的子集上对模型进行训练，得到个模型，和若干对应的载荷以及赋值。抽样的随机性控制了行业和季节效性，让训练出来的词汇更具普遍性。

情绪测度

- 分词情绪得分：

每次抽样训练都会得到一组分词情绪得分，将单个分词被标记为积极的次数减去被标记为消极的次数得到积极得分D+，消极得分D-同理。选取D+和D-排名前20%的一元分词进入词典。二元分词为55%。

- 计算情绪：

利用包含m个词汇的词典Di衡量文档j的情绪的计算公式为：

$$S_j = \sum_{i \in D_i} \left(\frac{tf_{ij}}{N_j}\right)$$

分子为情感分词i在文档j中出现的频率，分母为文档j的分词数。

词典竞赛

- 实证模型：

$$R_{jt} = \beta S_{jt} + \gamma X_{jt} + \epsilon_{jt}$$

t为文本发布日，R为公司j在[t-1,t+2]期间的累计收益，S为标准化文本情绪，X为控制变量向量，ϵ为随机误差项。

- 字典分类：

由于ML词典与LM词典存在重合分词，且ML词典中包含一元词和二元词，因此情绪构建时采用以下几种词典：

（1）LM特有词词典

（2）ML特有词词典

（3）LM&ML交叉词典

（4）ML二元词词典

词典竞赛

- 实证结果1：

  ML词典采用2005-2015的观测值拟合得到，实证期为2016-2020.

	Dependent variable:					
	Filing period excess return					
	(1)	(2)	(3)	(4)	(5)	(6)
LM positive	0.41***				−0.14*	0.06
	(6.6)				(−1.9)	(1.0)
LM negative	−0.50***				0.39***	0.24***
	(−4.4)				(6.0)	(3.0)
ML positive		0.98***			0.78***	
		(7.7)			(8.7)	
ML negative		−1.37***			−0.94***	
		(−11.7)			(−9.8)	
LM & ML positive			1.25***		0.90***	0.89***
			(11.4)		(9.5)	(9.7)
LM & ML negative			−1.56***		−1.32***	−1.34***
			(−9.3)		(−9.1)	(−9.4)
ML positive bigrams				1.38***		1.06***
				(10.7)		(12.4)
ML negative bigrams				−1.36***		−0.79***
				(−7.7)		(−5.8)
Adjusted R^2	0.021	0.046	0.054	0.045	0.065	0.064
Observations	39,269	39,269	39,269	39,269	39,269	39,269

词典竞赛

- 实证结论1：
- 用单个词典情绪回归，每个词典都能对文本情绪进行充分总结，对股票收益变动有显著影响，说明ML词典和LM词典均具有有效性。
- 同时采用多个词典情绪回归，LM词典的情绪对股票收益的载荷出现符号反向；ML&LM词典情绪指标预测效果依旧显著，说明ML词典选出了LM字典中的有效分词。

词典竞赛

- 实证结果2：

ML字典采用2005-2020观测值拟合得到，回归数据为年度报表文本。检验LM字典的外部有效性

	Dependent variable:					
	Filing period excess return					
	(1)	(2)	(3)	(4)	(5)	(6)
LM positive	−0.14**				−0.14**	−0.13**
	(−2.2)				(−2.1)	(−2.1)
LM negative	−0.06*				0.03	0.01
	(−1.9)				(1.1)	(0.4)
ML positive		0.11***			0.13***	
		(5.1)			(4.9)	
ML negative		−0.05			−0.01	
		(−1.2)			(−0.3)	
LM & ML positive			0.05**		0.05	0.05*
			(2.4)		(1.5)	(1.8)
LM & ML negative			−0.18**		−0.21***	−0.14**
			(−2.4)		(−2.9)	(−2.2)
ML positive bigrams				0.15***		0.13***
				(3.8)		(3.1)
ML negative bigrams				−0.16**		−0.10**
				(−2.6)		(−2.5)
Adjusted R^2	0.013	0.013	0.013	0.013	0.013	0.014
Observations	76,922	76,922	76,922	76,922	76,922	76,922

词典竞赛

- 实证结论2：

➢ 用单个词典情绪回归，LM特有词词典构造的情绪变量几乎同股价变动无任何关联；LM&ML交叉词典表现依旧稳定，ML二元词词典构造的情绪变量同股价变动之间关联性最显著。

➢ 同时采用多个词典情绪回归，LM词典的情绪对股票收益的载荷出现符号反向；，ML二元词词典构造的情绪变量同股价变动之间关联性最显著。

➢ LM词典在电话会议文本内的表现优于在年度报表内的表现，即便LM词典由后者产生。基于ML词典优于LM词典的表现，本文认为ML词典更多地捕捉了金融词汇的色彩（color of finance words）

词典竞赛

- 实证结果3：

ML字典采用2005-2020观测值拟合得到，回归数据为WSJ报道文本。检验LM字典的外部有效性

	Dependent variable:					
	Filing period excess return					
	(1)	(2)	(3)	(4)	(5)	(6)
LM positive	0.10***				0.08***	0.09***
	(7.6)				(6.1)	(6.7)
LM negative	−0.11***				−0.05***	−0.08***
	(−7.5)				(−3.5)	(−5.8)
ML positive		0.12***			0.09***	
		(7.4)			(5.2)	
ML negative		−0.21***			−0.17***	
		(−10.4)			(−9.3)	
LM & ML positive			0.16***		0.12***	0.12***
			(10.7)		(8.8)	(9.2)
LM & ML negative			−0.19***		−0.16***	−0.16***
			(−8.7)		(−7.6)	(−7.5)
ML positive bigrams				0.14***		0.09***
				(7.2)		(4.9)
ML negative bigrams				−0.14***		−0.10***
				(−7.8)		(−6.0)
Adjusted R^2	0.007	0.010	0.009	0.007	0.013	0.012
Observations	87,198	87,198	87,198	87,198	87,198	87,198

词典竞赛

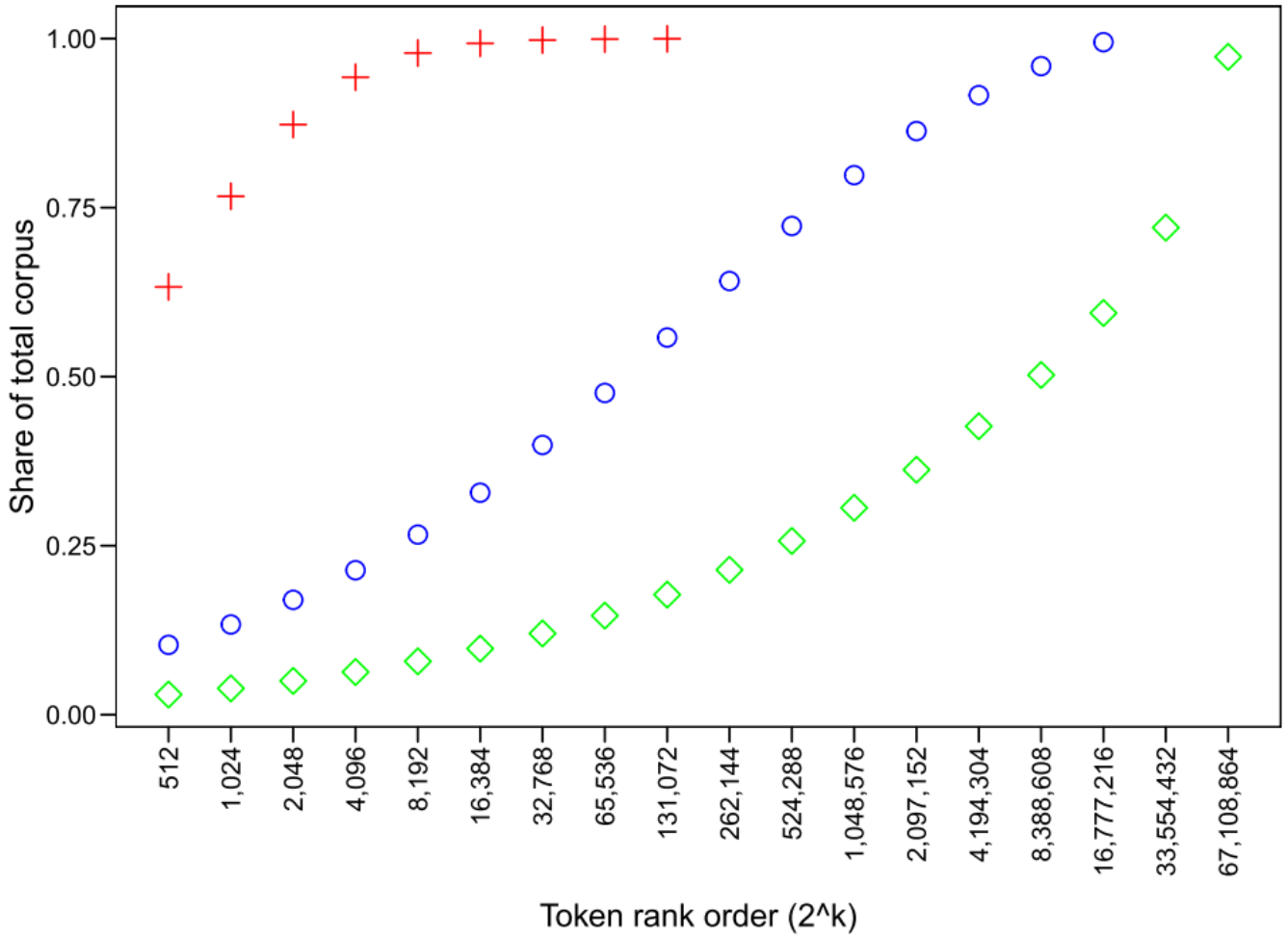
- 实证结论3：
  - 用单个词典情绪回归，所有词典的表现均显著
  - 同时采用多个词典情绪回归，ML词典和LM词典情绪表现均非常显著
  - LM词典在WSJ报道文本内的表现优于在其它语料库内的表现，但ML词典的表现在系数大小、系数显著性以及R2上都有十分明显的优势
  - 实证2和3证明了利用电话会议文本构建的ML词典在其它语料库中也有不错的表现。

进一步研究

- 是否需要包含三元词：
  - 平均每个电话会议文本的二元词数目为2785，对应2430个唯一二元词，三元词数目为2455，对应2376个唯一三元词，从单个文本上来看，二元词能够很好地覆盖一个文本。
  - 在p的选择上，如果要尽可能覆盖整个语料库，需要的唯一二元词数目远少于唯一三元词，且两者远少于唯一一元词（如下图）
  - 在未录入表格的检验中，将三元词词典加入字典竞赛，以其构建的情绪的解释能力能够被一元词和二元词词典充分解释。

进一步研究

- 是否需要包含三元词：



进一步研究

- 字典的“广度”比较

下表显示了不同字典的分词数目以及在不同语料库的出现频率。

Dictionary	Number of tokens	Coverage (% of corpus)		
		Earnings calls	10-K	WSJ
Positive words				
LM positive	329	1.9%	0.7%	1.3%
ML positive	57	8.4%	4.1%	3.2%
ML & LM positive	18	1.3%	0.2%	0.4%
ML positive bigram	381	2.3%	0.8%	0.3%
Negative words				
LM negative	2315	1.4%	2.7%	3.2%
ML negative	64	4.5%	4.3%	3.1%
ML & LM negative	30	0.4%	0.4%	0.5%
ML negative bigram	344	1.4%	0.7%	0.4%

整体上来看，ML词典的体量小于LM词典，但ML词典中分词在所有语料库中出现的频率超过LM词典的分词，体现了“plain money English”

进一步研究

- 分词对比1

  右表列出了LM字典中，出现频率前30的词汇，列出了它们在MNIR算法下的情感得分。LM词典和ML词典在一些词汇的情感归类上产生了分歧（improve、confident），两者对一些词汇的情感程度衡量上也有所不同。

Positive words				Negative words			
Token	Cov.	% Pos	% Neg	Token	Cov.	% Pos	% Neg
good^{+}	35.1	99.4	0.0	question	25.6	39.2	7.0
strong^{+}	26.0	100.0	0.0	questions	10.5	21.8	6.4
better^{+}	15.1	92.4	0.0	decline^{-}	8.0	0.0	99.8
opportunities	12.9	58.4	4.6	loss^{-}	6.8	0.0	99.0
able	12.1	63.2	2.2	negative^{-}	4.4	0.2	96.6
opportunity	11.9	68.0	3.8	difficult	3.7	0.0	78.4
positive	10.2	62.6	2.6	against	3.6	7.8	27.4
improvement^{+}	10.0	100.0	0.0	declined^{-}	3.5	0.2	91.4
progress	7.9	56.4	5.0	restructuring	3.2	30.8	30.4
pleased^{+}	7.7	99.8	0.0	losses	2.8	6.0	69.0
improved^{+}	6.9	100.0	0.0	challenges^{-}	2.6	0.0	99.8
improve	6.7	11.0	34.0	challenging^{-}	2.4	0.2	87.0
best	6.5	25.6	10.4	recall	1.8	8.2	25.6
strength^{+}	4.8	100.0	0.0	declines^{-}	1.8	0.0	85.8
success^{+}	4.4	88.8	0.0	volatility	1.7	6.8	42.4
excited	4.4	49.8	4.6	slow	1.6	0.2	66.4
profitability	4.3	63.0	4.8	break	1.5	22.6	6.6
confident^{-}	3.9	0.4	80.4	weakness^{-}	1.4	0.0	99.8
improving^{+}	3.8	82.4	0.0	bad	1.3	6.0	44.4
favorable^{+}	3.6	86.4	0.0	challenge	1.3	0.2	77.4
improvements^{+}	3.5	89.4	0.2	problem	1.3	1.6	71.4
gain	3.4	64.0	1.2	weak	1.2	0.2	78.8
despite	3.3	3.8	33.6	claims	1.2	12.0	61.8
successful	3.2	41.2	2.4	slower^{-}	1.2	0.0	93.0
gains^{+}	3.2	82.4	0.0	negatively^{-}	1.2	0.0	96.8
stronger	3.2	72.0	0.2	lost^{-}	1.2	0.0	96.8
efficiency	3.1	68.6	1.6	cut	1.1	3.4	50.2
advantage	3.0	61.0	1.4	slowdown^{-}	1.1	0.0	96.8
achieve	3.0	32.0	6.0	impairment	1.1	1.2	81.0
innovation	2.8	57.2	6.4	missed	1.0	0.6	49.2

进一步研究

- 分词对比2：

  右表列出了ML字典中，出现频率前30的词汇，列出了它们在MNIR算法下的情感得分。与前一张表不同，LM较少地包含了这些高频词。注意到think是情感最积极的词汇，这和一般认知有很大出入。

Positive words				Negative words			
Token	Cov.	% Pos	% Neg	Token	Cov.	% Pos	% Neg
think	97.0	82.4	2.0	not	108.5	0.6	97.0
growth	62.2	96.2	0.4	down	27.9	0.0	94.8
up	54.8	91.4	0.6	back	25.3	0.0	95.8
well	50.0	82.8	0.4	impact	21.8	0.0	99.8
over	47.3	87.8	0.2	believe	18.4	0.4	84.2
really	38.6	97.2	0.0	lower	17.3	0.0	100.0
continue	37.8	96.6	0.0	due	15.2	0.0	91.0
good$^+$	35.1	99.4	0.0	costs	14.1	0.4	89.2
results	27.9	91.2	0.0	expected	11.5	0.0	97.4
share	27.8	91.2	0.8	related	11.2	0.0	99.2
cash	26.9	83.2	1.6	change	10.3	0.0	96.2
increase	26.6	95.8	0.0	need	9.8	0.4	81.4
strong$^+$	26.0	100.0	0.0	offset	8.3	0.0	90.0
basis	25.5	84.0	1.4	expectations	8.0	0.2	81.2
operating	25.3	89.4	0.6	decline$^-$	8.0	0.0	99.8
margin	25.2	93.2	0.2	trying	7.3	0.4	83.2
lot	23.6	87.8	1.0	changes	7.0	0.0	95.0
years	20.9	81.4	0.4	loss$^-$	6.8	0.0	99.0
increased	20.4	96.4	0.0	term	6.7	0.4	83.6
income	17.5	88.6	0.6	certain	6.6	0.0	82.0
performance	17.2	94.0	0.0	factors	6.4	0.0	80.8
better$^+$	15.1	92.4	0.0	taking	6.0	0.0	96.0
pretty	14.3	94.0	0.0	understand	5.9	0.0	100.0
great	13.4	100.0	0.0	timing	5.7	0.0	97.8
across	12.2	91.4	0.2	however	5.2	0.0	99.8
continued	12.2	91.2	0.0	associated	4.9	0.2	91.2
flow	12.1	84.0	1.6	impacted	4.5	0.0	100.0
improvement$^+$	10.0	100.0	0.0	negative$^-$	4.4	0.2	96.6
benefit	8.9	83.6	0.2	decrease	4.4	0.0	96.2
pleased$^+$	7.7	99.8	0.0	issues	4.1	0.0	100.0

进一步研究

- 分词对比2：

  综合两表可以从细节上得出：即便两个词典存在交叉的部分，但是它们在描绘金融词汇的情感色彩时具有很大差别（different color of finance words）

Positive words				Negative words			
Token	Cov.	% Pos	% Neg	Token	Cov.	% Pos	% Neg
think	97.0	82.4	2.0	not	108.5	0.6	97.0
growth	62.2	96.2	0.4	down	27.9	0.0	94.8
up	54.8	91.4	0.6	back	25.3	0.0	95.8
well	50.0	82.8	0.4	impact	21.8	0.0	99.8
over	47.3	87.8	0.2	believe	18.4	0.4	84.2
really	38.6	97.2	0.0	lower	17.3	0.0	100.0
continue	37.8	96.6	0.0	due	15.2	0.0	91.0
good$^+$	35.1	99.4	0.0	costs	14.1	0.4	89.2
results	27.9	91.2	0.0	expected	11.5	0.0	97.4
share	27.8	91.2	0.8	related	11.2	0.0	99.2
cash	26.9	83.2	1.6	change	10.3	0.0	96.2
increase	26.6	95.8	0.0	need	9.8	0.4	81.4
strong$^+$	26.0	100.0	0.0	offset	8.3	0.0	90.0
basis	25.5	84.0	1.4	expectations	8.0	0.2	81.2
operating	25.3	89.4	0.6	decline$^-$	8.0	0.0	99.8
margin	25.2	93.2	0.2	trying	7.3	0.4	83.2
lot	23.6	87.8	1.0	changes	7.0	0.0	95.0
years	20.9	81.4	0.4	loss$^-$	6.8	0.0	99.0
increased	20.4	96.4	0.0	term	6.7	0.4	83.6
income	17.5	88.6	0.6	certain	6.6	0.0	82.0
performance	17.2	94.0	0.0	factors	6.4	0.0	80.8
better$^+$	15.1	92.4	0.0	taking	6.0	0.0	96.0
pretty	14.3	94.0	0.0	understand	5.9	0.0	100.0
great	13.4	100.0	0.0	timing	5.7	0.0	97.8
across	12.2	91.4	0.2	however	5.2	0.0	99.8
continued	12.2	91.2	0.0	associated	4.9	0.2	91.2
flow	12.1	84.0	1.6	impacted	4.5	0.0	100.0
improvement$^+$	10.0	100.0	0.0	negative$^-$	4.4	0.2	96.6
benefit	8.9	83.6	0.2	decrease	4.4	0.0	96.2
pleased$^+$	7.7	99.8	0.0	issues	4.1	0.0	100.0

进一步研究

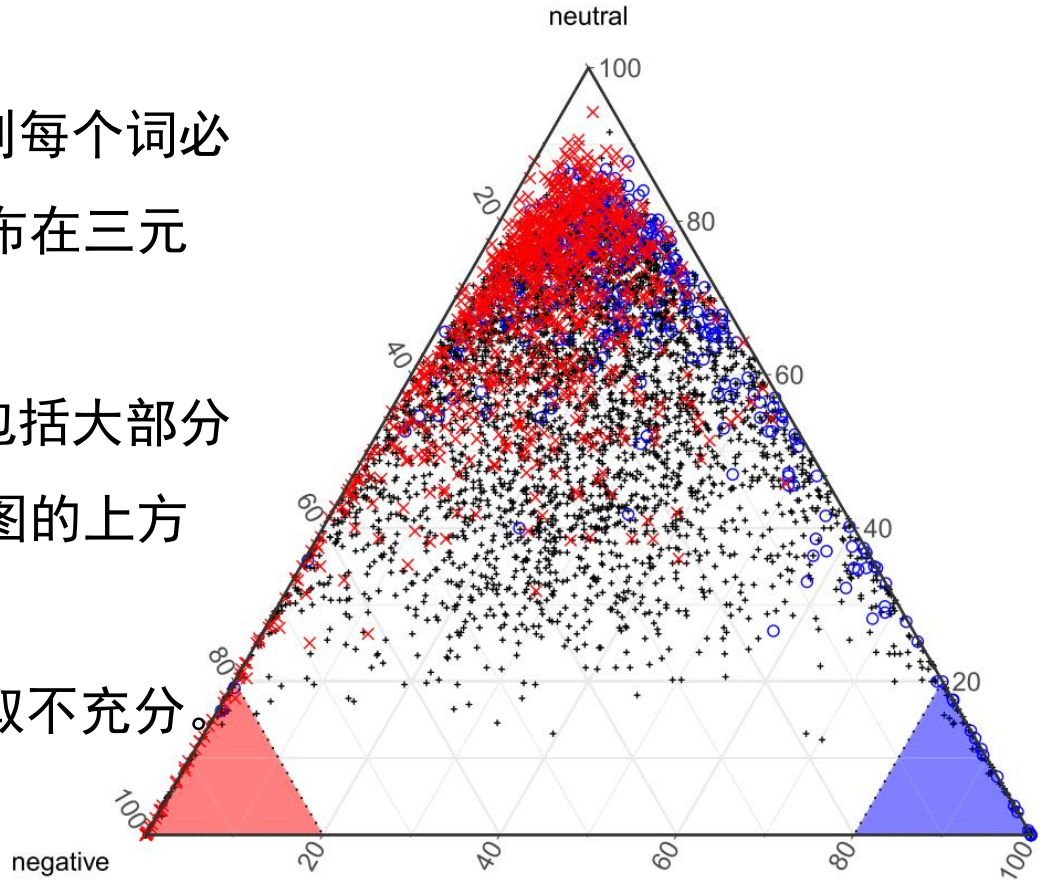
- 一元词的歧义：

下图描绘了MNIR算法下的情绪分数的三元图，红色的点代表LM字典的消极词汇，蓝色的点代表积极词汇，黑色的点代表不在LM字典中的语料库中频率排名前4000的词汇。

假设所有词无需去歧义，则每个词必定有唯一的词性，则所有点分布在三元图的三个角落中。

实际结果为绝大多数词（包括大部分积极词和消极词）分布在三元图的上方角落中，代表中性词。

说明这些词可能存在情感提取不充分。



进一步研究

- 二元词去歧义：

下表选出了二元词中情感得分在对应词族中前20%的分词（共属于6个有歧义的一元词，2个来自LM词典，4个来自ML词典）

Positive bigrams			Negative bigrams		
Bigram	Rel. Freq.	D^+ score	Bigram	Rel. Freq.	D^- score
continue$^+$ improve$^+$	8.88	64.40	improve$^+$ performance$^+$	2.17	39.60
continues$^+$ improve$^+$	2.91	41.60	improve$^+$ over$^+$	1.38	21.40
able$^+$ improve$^+$	0.97	22.80	going improve$^+$	1.11	25.20
			improve$^+$ second	0.80	27.60
			conditions improve$^+$	0.79	22.20
more confident$^+$	5.05	31.60	remain confident$^+$	15.82	75.20
increasingly confident$^+$	0.76	29.40	still confident$^+$	1.74	36.00
			confident$^+$ get	1.53	34.00
			confident$^+$ strategy	1.46	45.40
			confident$^+$ see	1.03	36.20

进一步研究

- 二元词去歧义：

  蓝（红）色词为LM词典的积极（消极）词汇，青（棕）色词为ML词典的积极（消极）词汇

solid+ growth+	7.10	21.20	solid+ tumors	1.19	28.60
solid+ quarter	5.75	37.00			
solid+ results+	3.55	22.40			
quarter solid+	2.53	21.80			
pretty+ solid+	2.51	23.40			
soft− launch	6.48	21.80	bit soft−	10.62	25.00
			soft− demand	9.94	33.60
			soft− quarter	9.43	26.00
cash+ flow+	25.86	77.20	cash+ used	0.54	42.80
free cash+	10.58	63.00	cash+ burn	0.33	27.60
operating+ cash+	2.89	42.40	cash+ cost	0.32	39.20
cash+ balance	1.96	37.60	company cash+	0.26	26.80
strong+ cash+	1.45	83.20	used cash+	0.23	23.40
continue+ see	4.95	47.00	continue+ believe−	1.19	43.60
going continue+	3.15	26.00	continue+ advance	0.19	40.00
continue+ grow	2.96	45.00	continue+ face	0.14	36.80
expect continue+	2.62	51.60	continue+ impact−	0.12	53.00
continue+ focus	1.67	27.60	may continue+	0.11	21.80

进一步研究

- 二元词去歧义：

  上表中得出以下结论：

  - 一些词本身的情感与其下的二元词族的情感可能具有巨大差别。
  - 主观筛选的一元词存在歧义的原因在于与其关联的二元词表现出来的积极和消极情感是差不多的（左右第一列Freq加总差别不大）。
  - 采用算法筛选出来的一元词不存在去歧义的问题，积极词汇下的二元词表现出的积极情感明显多于消极情感。
  - 上述分析其实是不全面的，考虑到分词的数目。但很直观地表现了通过二元词的情感分布来帮助一元词去歧义。

其它

原始数据：

https://leeds-faculty.colorado.edu/garcia/data.html

https://data.mendeley.com

Kaggle数据：

https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

部分关联论文

(1)Gentzkow, et al. 2019. 《Text as data》. J. Econ. Lit. 57 (3),535–574.

(2)Jegadeesh, N., Wu, D., 2013. 《Word power: a new approach for content analysis》. J. Financ. Econ. 110, 712–729.

(3)Ke, et al. 2019. 《Predicting Returns with Text Data》. Technical Report. University of Chicago.

(4)Meursault, et al. 2021. 《PEAD.txt: Post-Earnings-Announcement Drift Using Text》. Technical Report. Federal Reserve Bank of Philadelphia.