语言无关的社交网络突发事件发现与情感分析方法

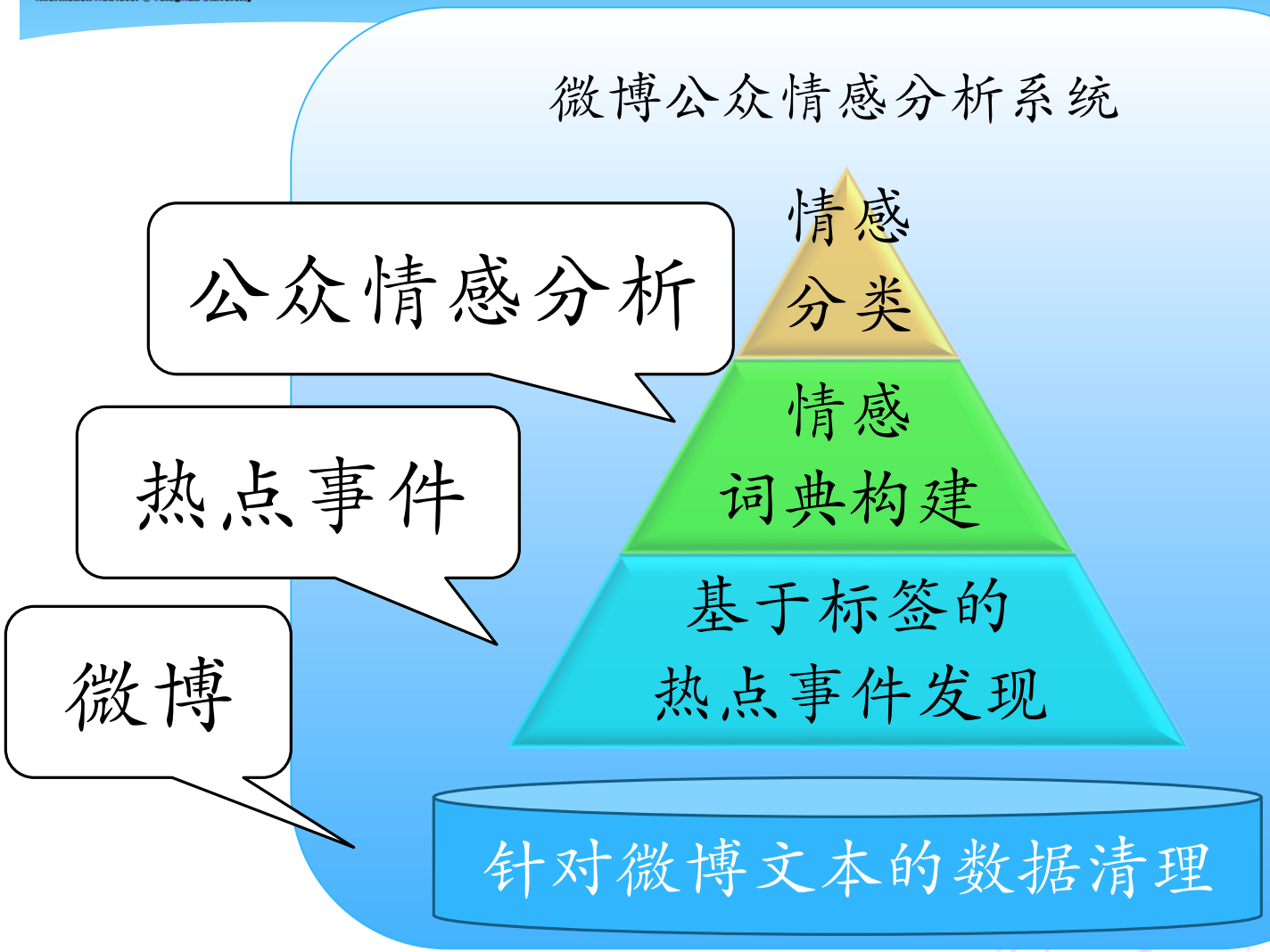
清华大学智能技术与系统国家重点实验室
信息检索组 马少平
http://www.thuir.org/
2013年9月13日 兰州

大数据时代

* Web2.0产生了大量的数据
  * 微博、论坛等
* 大数据金矿
  * 看似杂乱的数据之中，蕴含了大量有用信息

* 微博——用户生成内容，体现群体智慧
* 社会热点、舆论传播的平台
* 在微博中寻找热点事件
  * 商业资讯、社会政治事件
  * 客观事件，如地震
* 公众文本体现的情感
  * 公众情感形成的口碑
  * 个人抒发的情绪

研究问题概述


微博公众情感分析系统
公众情感分析
热点事件
微博
情感
分类
情感
词典构建
基于标签的
热点事件发现
针对微博文本的数据清理

基于标签的热点事件发现

* 短文本，缺乏上下文
* 同一事件，多角度描述

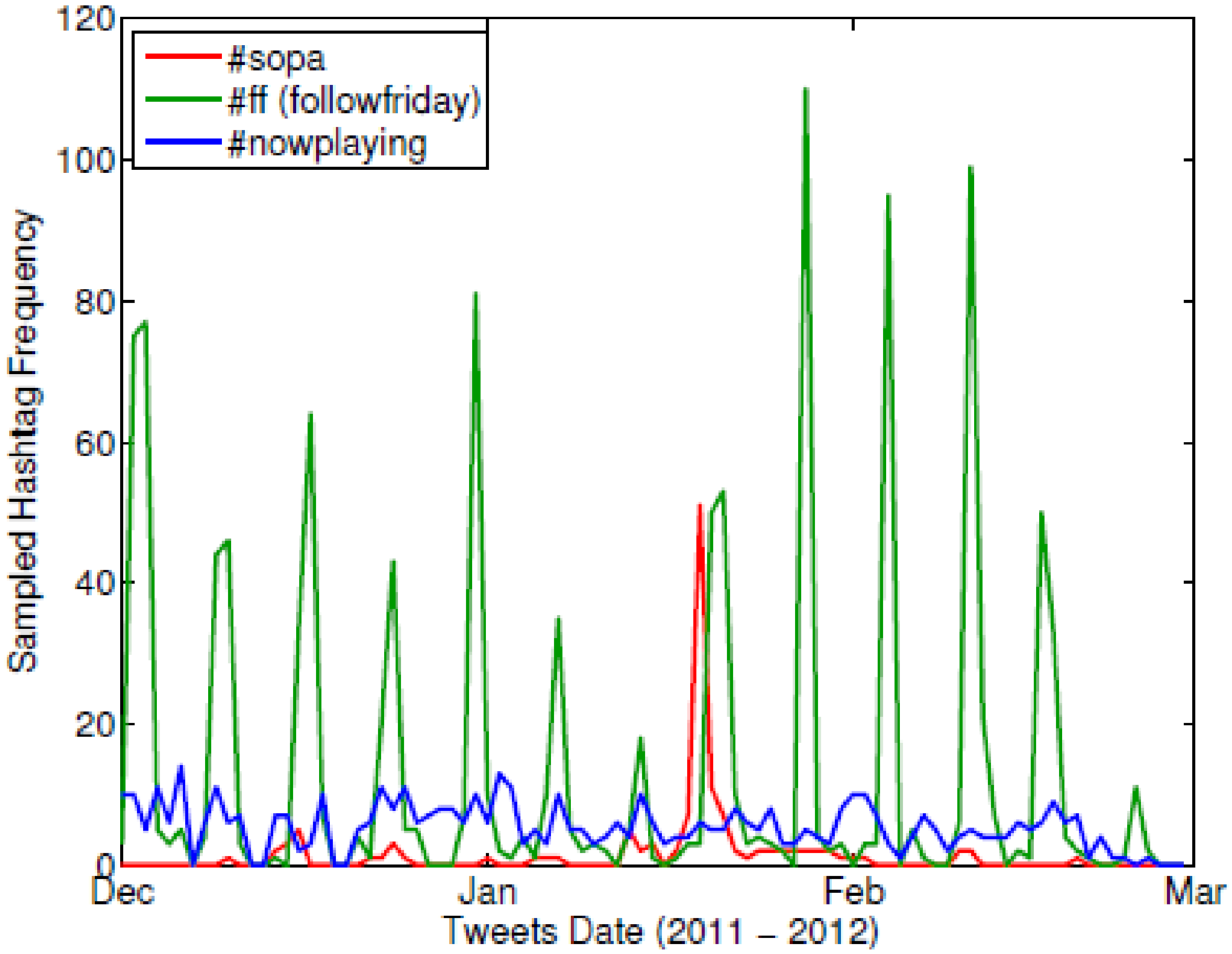
@脑残一孤独之路：科普一下……有些人总说广东不是处于地震带中，以为没有地震威胁，这样想被人压死不关我事……广东处于华南震区东南沿海外带地震带，尽管近300百年都没有发生过大型地震，但是历史上广东还是有过最高8级的地震记录。很难说在未来会否发生大型地震，大家切忌掉以轻心**#广东地震#**

马连奴奥兰迪品牌旗舰店：**#广东地震#** 人在遭遇突发事件时，若能保持良好的心理状态，及时采取自救行为或逃离现场，获救机会会更大。爱生命爱亲友，地震常识知多点，顺手转给你周围的亲朋好友吧 详情:http://t.cn/zYCBBot

@飘o燕：**#广东地震#**中午在饭堂吃饭的时候，天花板突然裂开一角掉下一块砖头，还好没砸到人，之后才知有地震，虽感觉不到震感但豆腐渣工程令人堪忧

* 解决方案：基于标签的热点事件发现

* 热门标签一定揭示热点事件吗？


#sopa
#ff (followfriday)
#nowplaying
Sampled Hashtag Frequency
120
100
80
60
40
20
0
Dec
Jan
Feb
Mar
Tweets Date (2011 – 2012)

Information Retriever @ Tsinghua University

* 热门标签一定来自网民的群体智慧吗？
  * 被营销账号“污染”

No.	Rank	Hashtag	Frequency	No. Authors	Most contributed author / Contribution
1	66	#nsfw	70,610	2,185	@porngus / 38,614 (54.7%)
2	90	#nieuws	57,273	249	@nieuwslogr / 29,936 (52.3%)
3	141	#sport	37,627	2,198	@sportlogr / 20,700 (55.0%)
4	162	#nl	33,892	286	@sportlogr / 20,700 (61.1%)
5	195	#property	30,116	905	@pflive_announce / 25,292 (84.0%)
6	343	#reddit	18,835	586	@redditspammor / 16,974 (90.1%)
7	379	#adult	17,510	634	@tubebutler / 8,906 (50.9%)
8	394	#rulez	17,092	116	@redditspammor / 16,974 (99.3%)
9	431	#praytweets	15,828	251	@praytweets / 15,466 (97.7%)
10	468	#indonesia	14,817	2,519	@beritaindonesia / 10,811 (73.0%)

$$Ent(hashtag) = -\sum_{i=1}^{k} \frac{c_i}{n} \cdot \log\left(\frac{c_i}{n}\right)$$

* 标签的三个维度：
  * 标签频度稳定性

$$\tilde{P}(x) = Pr(X > x \bigvee X < 2\mu - x) \qquad (1)$$

  * 微话题的可能性

$$p_{\text{pos}} = \frac{|\{\text{tweets starting with } h\}|}{|\{\text{tweets containing } h\}|} \qquad p_{\text{word}} = 1 - N/L$$

$$TMP(hashtag) = p_{\text{word}} \cdot p_{\text{pos}}$$

  * 参与用户的分散程度

$$Ent(hashtag) = -\sum_{i=1}^{k} \frac{c_i}{n} \cdot \log\left(\frac{c_i}{n}\right)$$

话题标签子空间的种类

序号	*Inst*	*TMP*	*Ent*	标签示例	可能的分类
1	较低	较低	较低	#abbeydawn, #bring1dtonyc	广告营销
2	较低	较高	较低	–	
3	较高	较低	较低	#property, #belieber	
4	较高	较高	较低	#praytweets	
5	较低	较高	较高	#nowplaying（语料 *Tweets3* 中），#iaintafraidtosay, #foramilliondollars	在线话题
6	较高	较高	较高	#nowplaying（语料 *Tweets6* 中）	
7	较低	较低	较高	#fb, #followfriday, #ff, #teamfollowback, #musicmonday	传统习语
8	较高	较低	较高	#hcr, #sopa, #halamadrid	微博热点事件

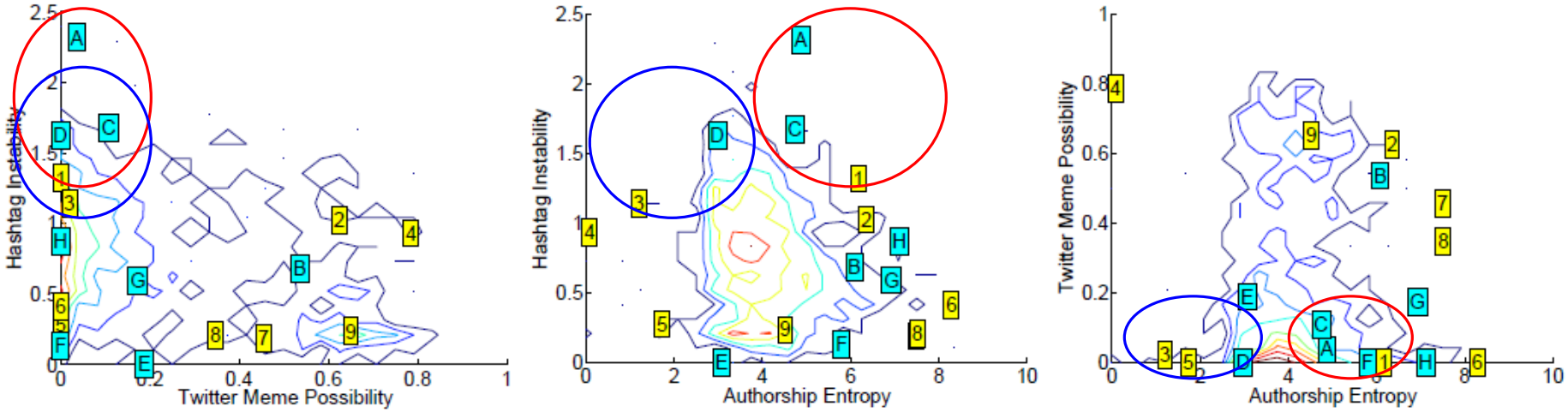


Figure 5: Contour of hashtag distributions. *Tweets6*: 1–#hcr, 2–#nowplaying, 3–#property, 4–#praytweets, 5–#abbeydawn, 6–#fb, 7–#musicmonday, 8–#followfriday, 9–#iaintafraidtosay. *Tweets3*: A–#sopa, B–#nowplaying, C–#halamadrid, D–#belieber, E–#bring1dtonyc, F–#fb, G–#teamfollowback, H–#ff

实验：热点事件分类 (2)

* 实验数据集：
  * 2009年下半年
  * 2011年12月-2012年2月
* 三个维度：标签频度稳定性、微话题的可能性、参与用户的分散程度

Table 4: Experiment Results of Hashtag Categories

Dataset	*Tweets6*						*Tweets3*					
Algorithm	Popularity Pattern			Subspace			Popularity Pattern			Subspace		
Accuracy	17.8%			**40.0%**			31.5%			**38.0%**		
Breaking events[a]	0.250	**0.231**	0.240	**0.333**	0.205	**0.254**	**0.192**	**0.192**	**0.192**	0.167	0.154	0.160
Twitter memes[a]	0.000	0.000	0.000	**0.681**	**0.595**	**0.635**	**1.000**	0.060	0.113	0.725	**0.248**	**0.369**
Advertisements[a]	–			**0.258**	**0.370**	**0.304**	–			**0.053**	**0.385**	**0.093**
Miscellaneous[a]	**0.162**	**0.926**	**0.276**	0.125	0.148	0.136	**0.240**	**0.909**	**0.379**	0.220	0.205	0.212

[a] In each category, precision, recall and F-measure are listed in order.

微博情感分析

- 微博文本短，表达语法不规范
  - 传统语言学方法（如句法分析等）准确度降低
  - 解决方案：构造情感词典进行情感分析
- 新词出现多
  - 自造词，如“给力”
  - 转义词，如“鸭梨”

  - 大葱它**肿么**了？
  - 什么都涨，就工资不涨。等工资张了，什么就都又涨了。**鸭梨**，，，
  - 李娜真**给力**，打的太好了。

微博情感分析

* 解决方案：基于表情符号的情感分析

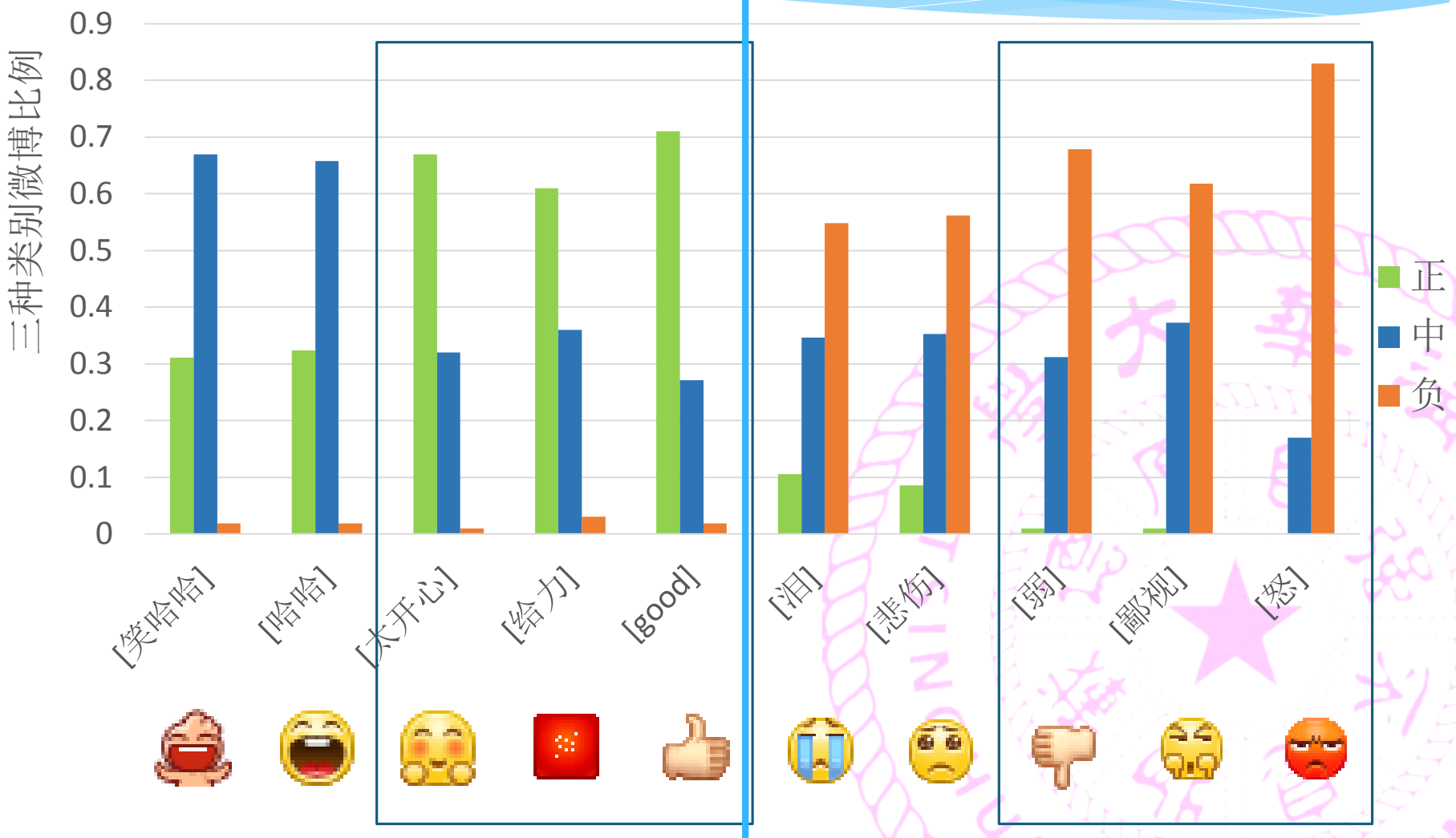
* 传统表情图标
  * 西方式：基于标准ASCII字符，通常需要转头90度

    :-)　:)　=-)　:-D　8)　=P　;-)
  * 东方式：引入Unicode字符，不需转头

    ^_^　\o/　Ơ‿Ơ　。◕‿◕。
  * 卡通式：

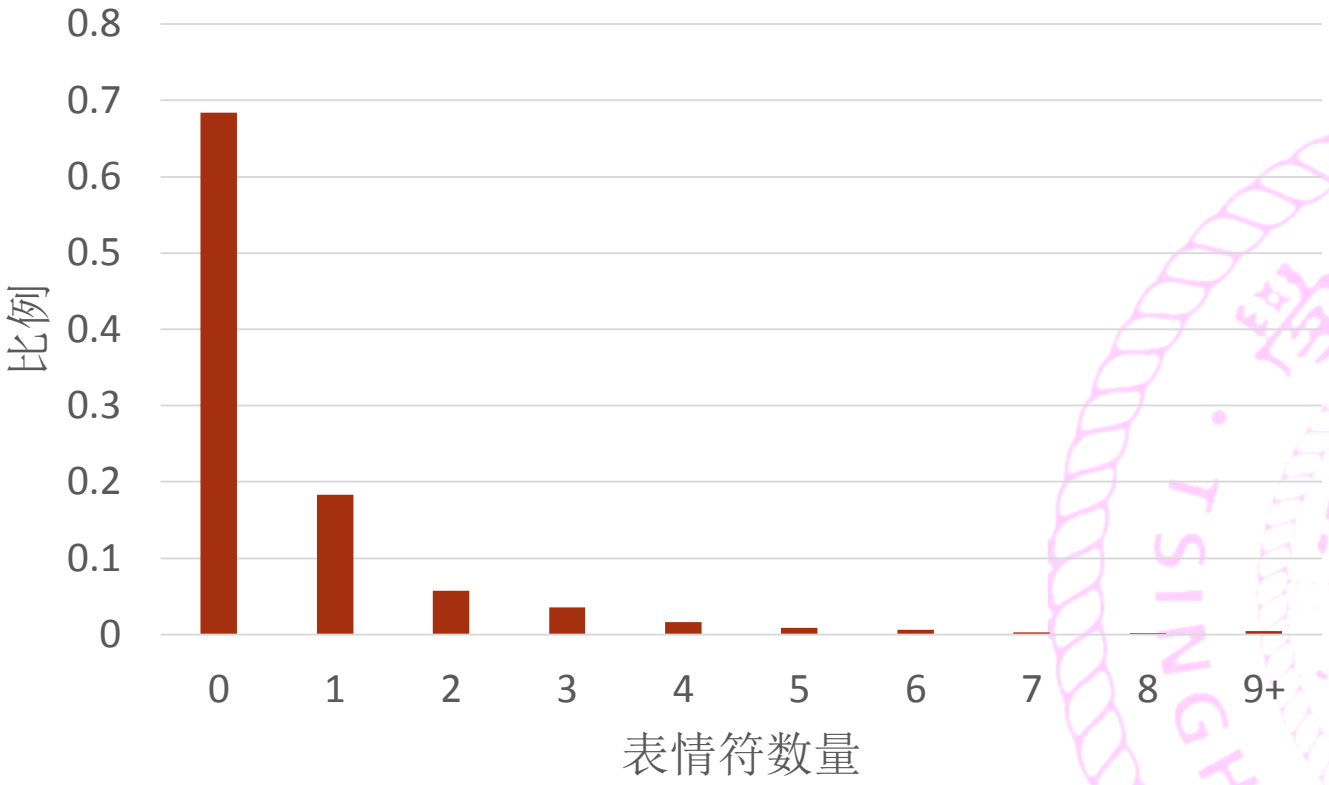
    
    
    
    
    

微博情感分析


三种类别微博比例
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
[笑哈哈]
[哈哈]
[太开心]
[给力]
[good]
[泪]
[悲伤]
[弱]
[鄙视]
[怒]
正
中
负

微博情感分析

* 微博表情符数量分布情况统计



含有表情符：**~32%**

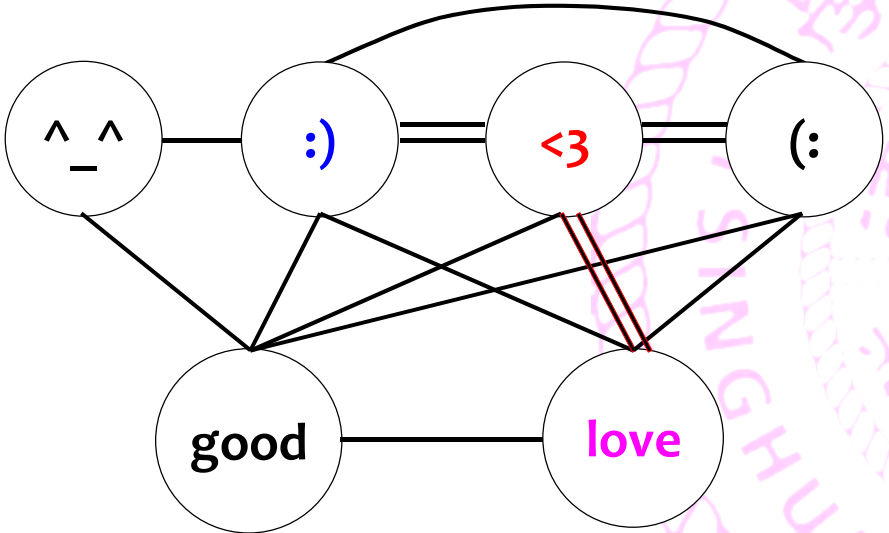
含有一个表情符：**~18%**

含有多个表情符：**~14%**

微博情感分析

* 基本假设
  * 微博内容短，同一条微博中表达的情感是单一的
  * 同一条微博中同现的词语，表达的情感是相似的

Tweets	Emotion tokens	Normal words
^_^ yes we did ! RT @JetLife24_7 Me and @tinyy_tee had some **good** times last summer **:)**	^_^ :)	**good**, …
@JASMINEVILLEGAS Pretty <3 , How are you ? . I wanna see you in the #MyWorldTour on S.America :) . **Love** you	<3 :)	**love**, …
That's A **Good** Boyfriend**(:** RT @CallMeYoshi : I Rather Stay In With My Girlfriend All Night Than Go Out To Party I **Love** Her To Much <3	(: <3	**good**, **love**, …
I can't wait for high-school **(:** gonna be back with my friends <3 on top of that my girls @_BRILove and @_THATSLEX are gonna be there too :)	(: <3 :)	…



* 从种子词出发，迭代计算所有词语的情感倾向，直到得分收敛：
  * $x_0 = (1, 0, \ldots, 0)^{\mathrm{T}}, \qquad x_{k+1} = W \cdot x_k + b$
  * x_0为种子向量
  * W为同现图邻接矩阵
  * $b = x_0$ 加重种子权重

* 种子词：对不同情感、情绪分别迭代——每个词被赋予多个得分。
  * 褒义种子1个：笑脸；贬义种子1个：哭脸

一些情感词示例

去打靶	0.012691948697959463	0.987308051302040 5
畜生	0.013437745305673402	0.986562546943267
完没完	0.014545824806600556	0.985454175193399 4
火滚	0.015648995217740 38	0.984351004782259 7
死扑	0.016088173753570 65	0.983911826246429 4
正扑	0.016505167469755 49	0.983494832530244 5
千刀万剐	0.017090877215725463	0.982909122784274 6
败类	0.017282849390511952	0.98271715060948 8
md	0.017691741985087217	0.98230825801491 27
nnd	0.018820096859947587	0.98117990314005 24
艹	0.019448049484870226	0.98055195051512 98
mlgb	0.019589906407834888	0.98041009359216 51

实验：多语言微博情感分类

* 实验数据：斯坦福大学SNAP多语言Twitter数据集
* 用Google翻译API识别英语、葡萄牙语、西班牙语、德语的微博共1213条
* 实验比较基准：
  * SentiWordNet情感词典
  * 在线Twitter情感分析网站Twitrratr与Twittersentiment

实验：英语与非英语的情感分析结果

Information Retriever @ Tsinghua University

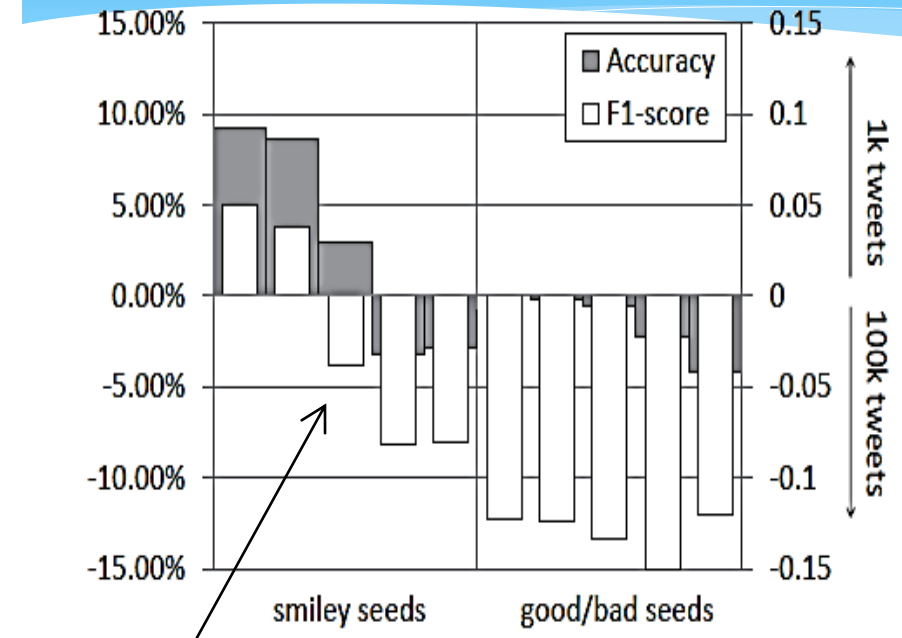
Table 6. Comparative evaluations of algorithms in English tweets

Algorithm	Accuracy	Positive (166)	Negative (62)	Neutral (111)	$\bar{F}_1$
		F_1 / P / R	F_1 / P / R	F_1 / P / R	
1. *SWN*, $\theta = 0.5$	51.3%	0.591 / 59.9% / 58.4%	0.429 / 38.5% / 48.4%	0.448 / 47.5% / 42.3%	0.489
2. *twitrratr*	49.3%	0.557 / 87.2% / 41.0%	0.276 / 27.9% / 27.4%	0.527 / 41.0% / 73.9%	0.454
3. *twittersentiment*	59.0%	0.648 / 78.0% / 55.4%	0.549 / 70.0% / 45.2%	0.548 / 44.2% / 72.1%	0.582
SentiLexicon	57.8%	0.642 / 65.2% / 63.3%	0.149 / 100.0% / 8.1%	0.606 / 49.7% / 77.5%	0.466

Table 7. Comparative evaluations of algorithms in non-English tweets

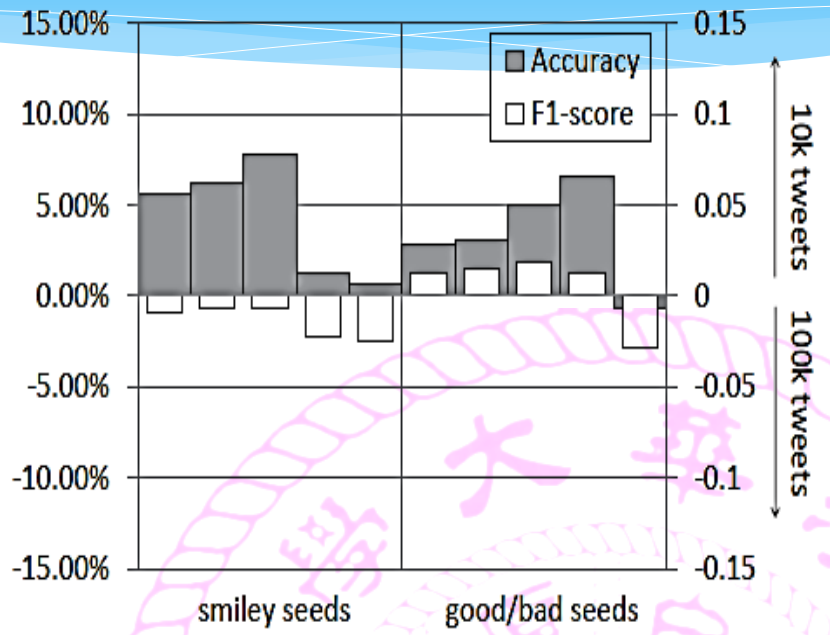
Algorithm	Accuracy	Positive (283)	Negative (149)	Neutral (442)	$\bar{F}_1$
		F_1 / P / R	F_1 / P / R	F_1 / P / R	
1. *SWN*, $\theta = 0.5$	–	–	–	–	–
2. *twitrratr*	52.2%	0.311 / 72.7% / 19.8%	0.130 / 20.9% / 9.4%	0.659 / 52.9% / 87.3%	0.366
3. *twittersentiment*	51.0%	0.218 / 44.1% / 14.5%	0.129 / 52.4% / 7.4%	0.656 / 51.8% / 89.1%	0.334
SentiLexicon	57.4%	0.500 / 51.3% / 48.8%	0.123 / 76.9% / 6.7%	0.685 / 59.8% / 80.1%	0.436

实验：构建情感词典所需数据集规模



不同参数组合

(a) 1k vs. 100k



(b) 10k vs. 100k

Fig. 2. Comparison of different datasizes for building lexicons

扩展

* 加入更多的种子，不仅仅是使用表情符
* 加入语言分析，如“否定”、联接词的处理等
* 特定领域的应用

* 情绪分析
  * 喜悦、愤怒、悲哀、恐惧、惊讶
* 只需给出不同情绪的种子词即可

在数码产品上的应用


机身重量轻，做工不错，电池也比较耐用。只是感觉塑料味儿有些浓！
Its weight is light, quality is good, battery is durable. But Its flavor of plastic is heavy!
FW: Feature Words
OW: Opinion Words
S: sentiment polarity

Information Retriever @ Tsinghua University

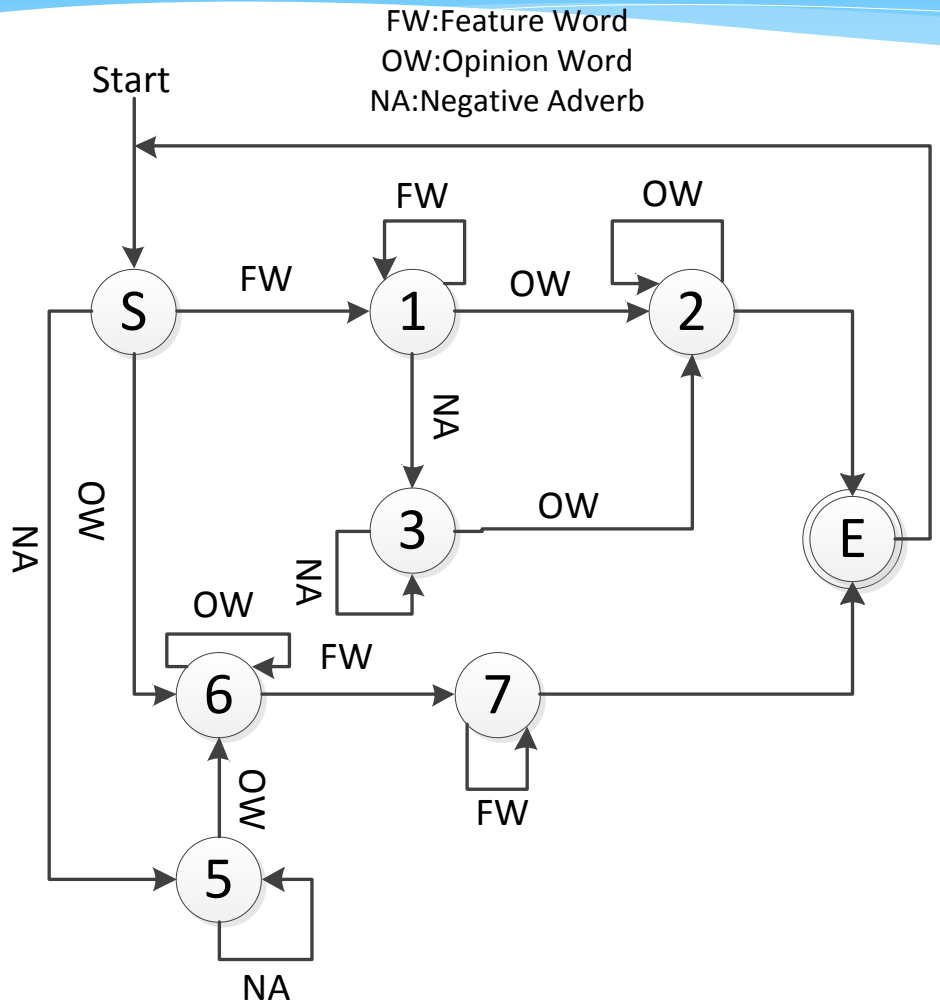
[1]	按键布局|合理	Button Layout | Reasonable
[1]	按键布局|简洁	Button Layout | Concise
[1]	镜头 | 给力	Camera lens | Geili (pretty good)
[1]	按键手感|舒适	Button Touch Feel | Comfortable
[-1]	暗部细节|欠缺	Dark Details | Lack of
[-1]	白天色彩|暗淡	Color in Daylight | Dim
[-1]	变焦杆|硬	Zoom Lever | Stiff
[1]	材质|耐磨	Material | Abrasion Resistant
[1]	长焦|强悍	Telephoto lens | Extreme
[-1]	价格 | 高	Price | high
[1]	性价比 | 高	Cost effectiveness | high
[-1]	电池续航能力|不足	Battery life | Insufficient
[1]	电池续航能力|出众	Battery life | Outstanding
[-1]	电池续航能力|差	Battery life | Bad

否定、转折的处理

* 一些例子：

* **显示器**并**不是**想象的那么**好**
* **画面清晰**但是**音效不是**很**好**
* **价格不**得**不**说很**便宜**

自动机：处理否定



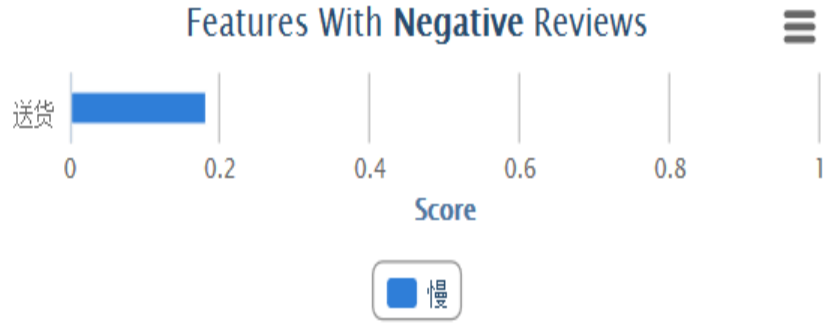
现阶段自动机可以处理的情况：

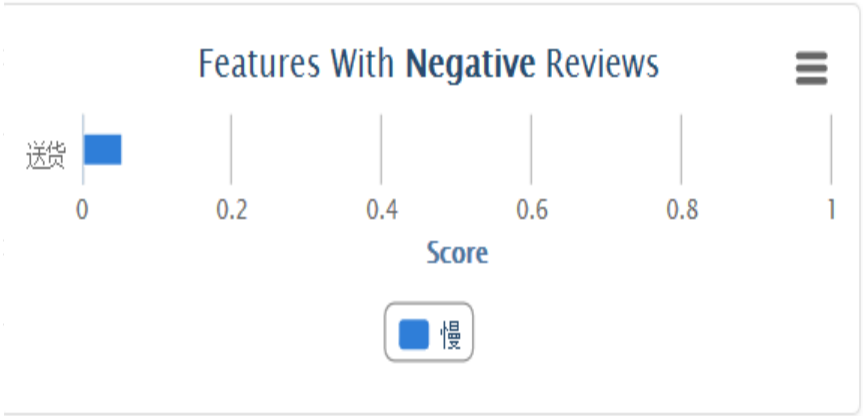
1. 简单的 n 个属性词对应 m 个观点词的情况
2. 包含 p 个否定词的情况

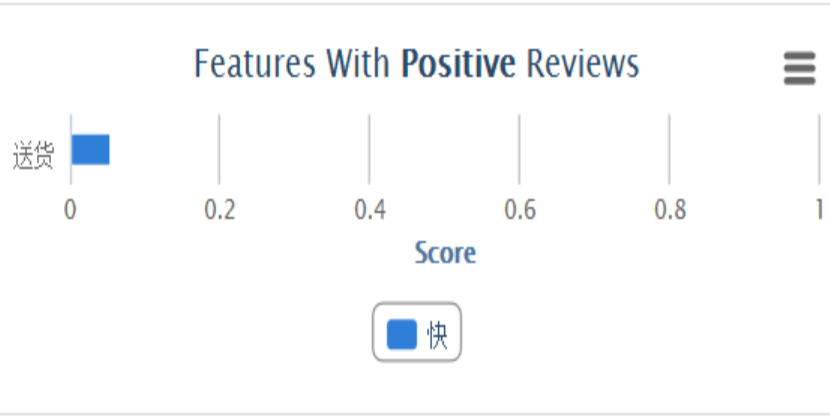
例如：

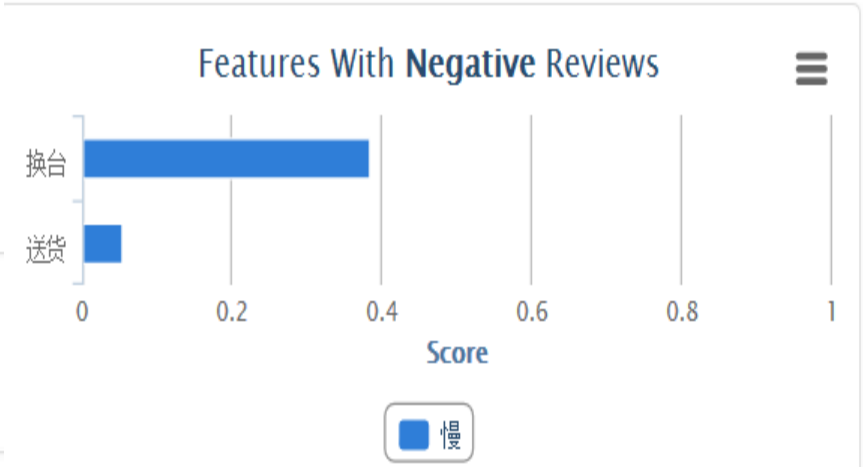
- “画面清晰但是音效不是很好”匹配流程为：
  S→1→2→E→1→3→2→E
- “价格不得不说很便宜”匹配流程为：
  S→1→3→3→2→E

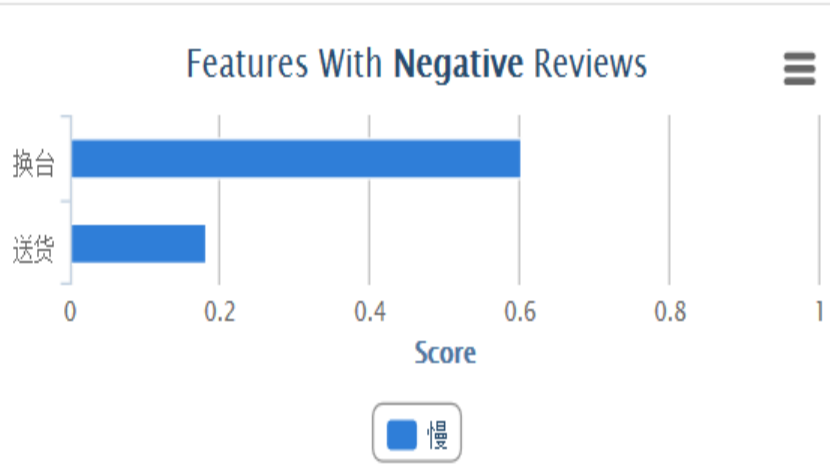

Features With Positive Reviews
送货
0
0.2
0.4
0.6
0.8
1
Score
上门
快
好


Features With Negative Reviews
送货
0
0.2
0.4
0.6
0.8
1
Score
慢


Features With Negative Reviews
送货
0
0.2
0.4
0.6
0.8
1
Score
慢

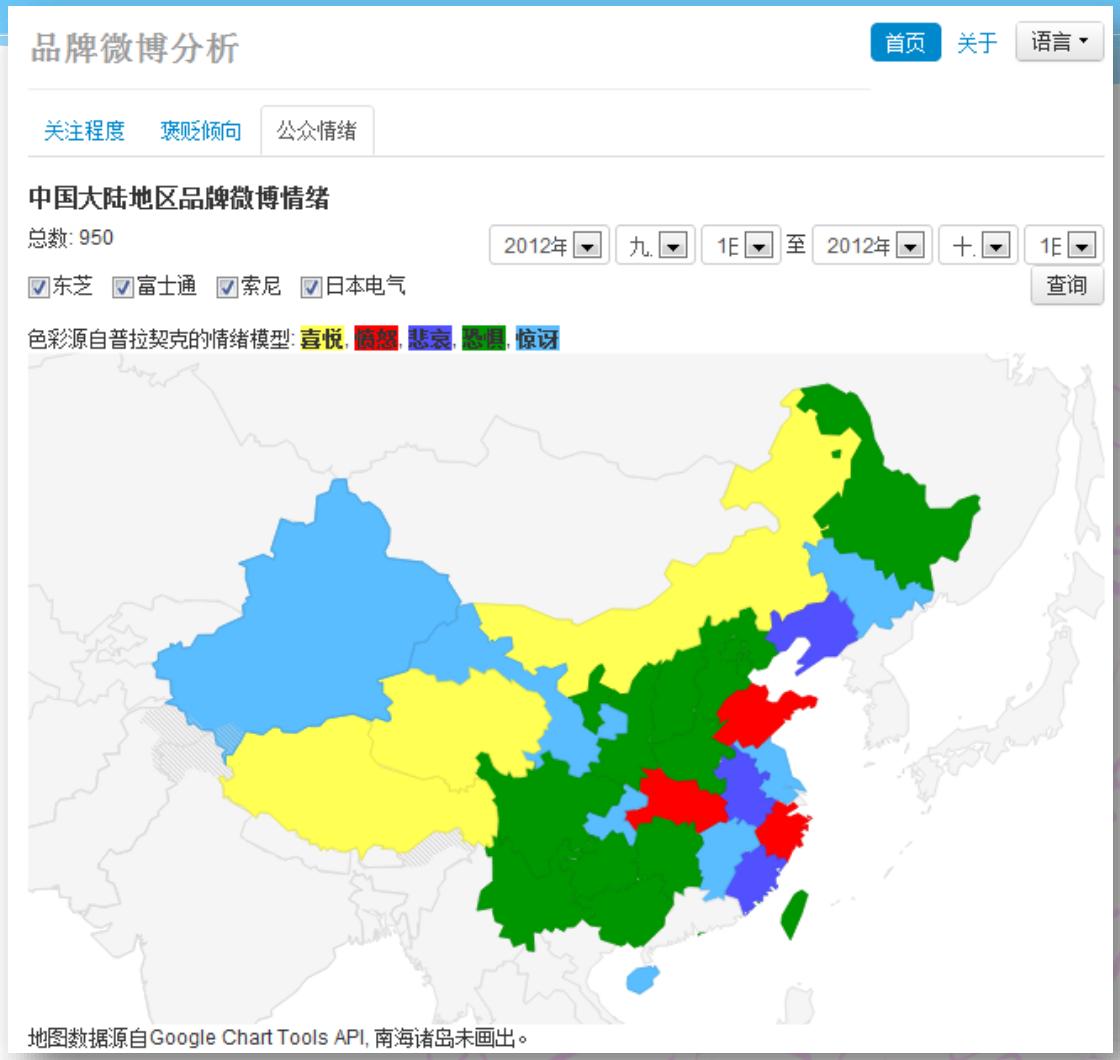

Features With Positive Reviews
送货
0
0.2
0.4
0.6
0.8
1
Score
快


Features With Negative Reviews
换台
送货
0
0.2
0.4
0.6
0.8
1
Score
慢


Features With Negative Reviews
换台
送货
0
0.2
0.4
0.6
0.8
1
Score
慢

微博情感分析示例

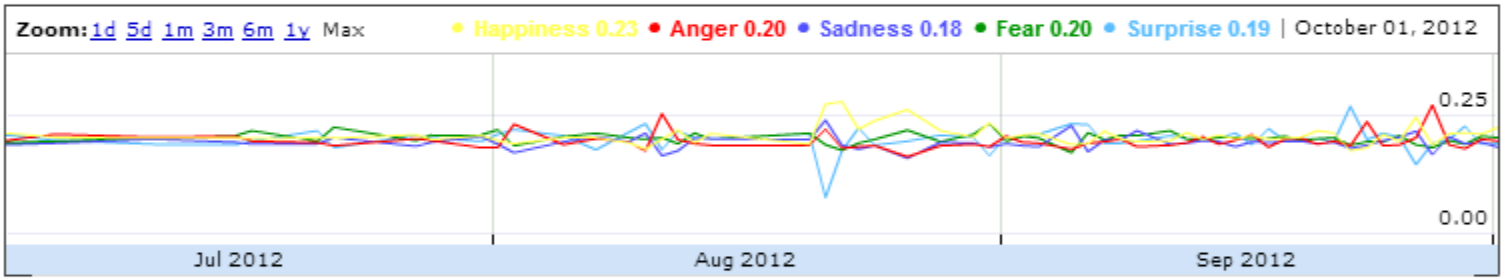
* 采用情绪词典可对每条微博的情绪进行评估，进而衡量其喜悦、愤怒、悲哀、恐惧、惊讶的程度



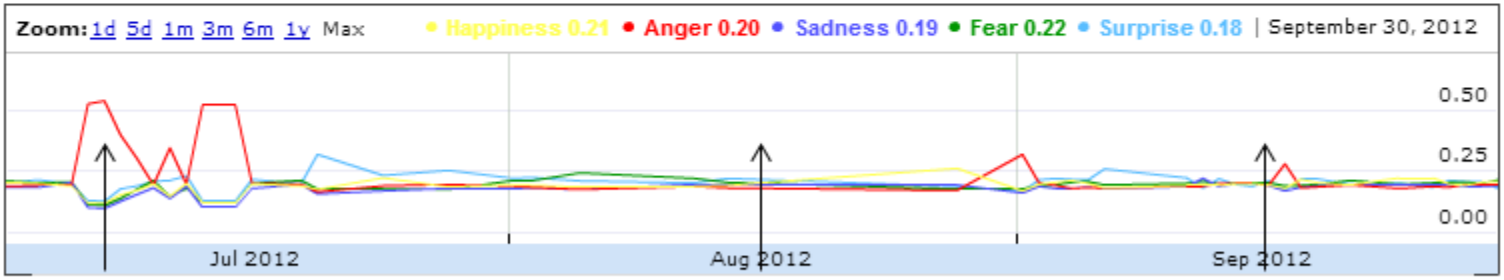
东芝
富士通
索尼
NEC

东芝

索尼



NEC



7月初
“七七”

8月中
香港保钓

9月中
反日示威

新浪微博，北京，2012年7月-9月，微博文本情绪趋势（喜怒哀惧惊）

谢谢！