武漢大學经济与管理学院
Economics and Management School of Wuhan University

情绪分析于在线评论有用性探究中的应用

—— 基于文献：

ANXIOUS OR ANGRY - EFFECTS OF DISCRETE EMOTIONS ON THE PERCEIVED HELPFULNESS OF ONLINE REVIEWS

汇报小组 ： 第一组
汇 报 人 ： 刘郅哲，谢燊
指导老师 ： 高宝俊
汇报时间 ： 2022-6-27

目录

Question: Anxious or Angry?

MIS Quarterly

RESEARCH ARTICLE

ANXIOUS OR ANGRY? EFFECTS OF DISCRETE EMOTIONS ON THE PERCEIVED HELPFULNESS OF ONLINE REVIEWS[1]

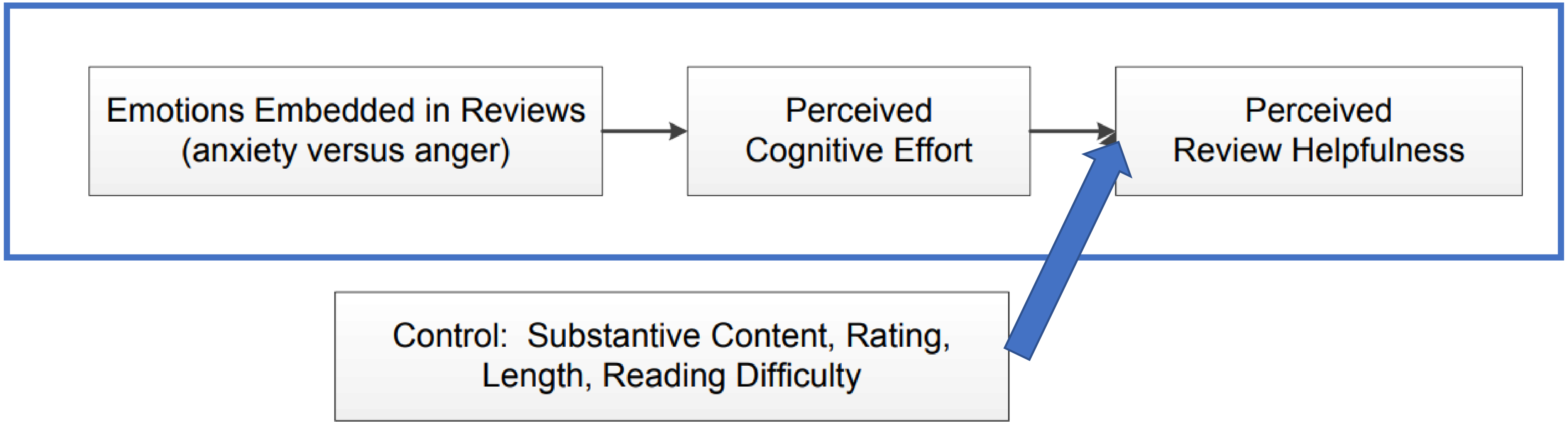
Dezhi Yin

Trulaske College of Business, University of Missouri, Columbia, MO 65211 U.S.A. {yind@missouri.edu}

Samuel D. Bond and Han Zhang

Scheller College of Business, Georgia Institute of Technology, Atlanta, GA 30308 U.S.A.
{sam.bond@scheller.gatech.edu} {han.zhang@scheller.gatech.edu}

Theoretical Framework



Outcome:

同样是表达负面情绪的在线评论，表达焦虑情绪的评论比表达愤怒情绪的评论对读者而言更加有帮助性

Major Goal：复现结果——Retell the Story(Outcome)

- **使用的工具** —— R
- **复现使用的方法**

- Last Term：

Tidy —— 数据的清洗与变量的构建

- Summer School：

Text Analysis ——构建语料库 & TDM；词频分析；绘制词云；情感分析和可读性分析

- According to Our Demand：

用其它包更便捷地达成目的，如生成语料库用quanteda包而不是tm包；生成词云用wordcloud2包而不是wordcloud包

2 What We Did & Learn——Results

	Dependent variable:				
	NumHelpful				
	Base	Base_FE	Text	Text_FE	Text_FE_C
AvgRatingStarsThisUser	-0.101***	-0.175***			-0.184***
	(0.019)	(0.018)			(0.020)
Rating_Deviation	0.296***	0.239***			0.238***
	(0.032)	(0.027)			(0.027)
log(WC)	0.642***	0.378***			0.359***
	(0.033)	(0.029)			(0.033)
Not_Disclosure	-0.477***	-0.305**			-0.307**
	(0.159)	(0.126)			(0.126)
women	-0.232	-0.247*			-0.248*
	(0.154)	(0.127)			(0.127)
MidAge	-0.196	-0.228			-0.228
	(0.179)	(0.145)			(0.145)
OldAge	-0.377*	-0.271*			-0.275*
	(0.199)	(0.164)			(0.164)
angry			0.295***	0.230***	-0.001
			(0.071)	(0.060)	(0.064)
anxious			0.230***	0.200***	0.092**
			(0.059)	(0.042)	(0.043)
readability			0.024***	0.011***	0.001
			(0.005)	(0.004)	(0.004)
sentiment			0.026***	-0.0005	0.007
			(0.006)	(0.006)	(0.006)
Constant	-3.465***	-0.088	-1.351***	1.392***	-0.006
	(0.245)	(0.455)	(0.049)	(0.459)	(0.462)

五个模型回归（左为小组得到的结果）

	Dependent variable:				
	NumHelpful				
	Base	Base_FE	Text	Text_FE	Text_FE_C
AvgRatingStarsThisUser	-0.101***	-0.175***			-0.170***
	(0.019)	(0.018)			(0.022)
Rating_Deviation	0.296***	0.239***			0.237***
	(0.032)	(0.027)			(0.027)
log(WC)	0.642***	0.376***			0.380***
	(0.033)	(0.029)			(0.033)
Not_Disclosure	-0.479***	-0.305**			-0.309**
	(0.158)	(0.126)			(0.126)
women	-0.235	-0.248*			-0.251*
	(0.154)	(0.127)			(0.127)
MidAge	-0.196	-0.228			-0.233
	(0.179)	(0.145)			(0.145)
OldAge	-0.376*	-0.270*			-0.273*
	(0.199)	(0.164)			(0.164)
angry			0.291***	0.199***	-0.009
			(0.069)	(0.059)	(0.063)
anxious			0.206***	0.149***	0.074*
			(0.049)	(0.037)	(0.038)
readability			0.025***	0.011***	0.001
			(0.005)	(0.004)	(0.004)
sentiment			0.014***	-0.019**	-0.001
			(0.007)	(0.007)	(0.007)
Constant	-3.463***	-0.076	-1.324***	1.474***	-0.104
	(0.245)	(0.455)	(0.050)	(0.458)	(0.462)

2 What We Did & Learn——Results

	Dependent variable:			
	AvgRatingStarsThisUser		sentiment	
	ordered logistic		*OLS*	
	Base	Base_C	Base	Base_C
sentiment	0.243***	0.296***		
	(0.006)	(0.007)		
log(WC)		-1.353***		2.295***
		(0.033)		(0.045)
Not_Disclosure		0.267*		0.279
		(0.145)		(0.227)
women		0.424***		0.203
		(0.136)		(0.210)
MidAge		0.021		0.189
		(0.162)		(0.252)
OldAge		0.243		0.377
		(0.172)		(0.266)
factor(AvgRatingStarsThisUser)2			1.081***	1.384***
			(0.142)	(0.126)
factor(AvgRatingStarsThisUser)3			2.890***	3.294***
			(0.129)	(0.117)
factor(AvgRatingStarsThisUser)4			4.331***	4.812***
			(0.120)	(0.114)
factor(AvgRatingStarsThisUser)5			4.741***	5.523***
			(0.117)	(0.115)
Constant			-0.009	-9.992***
			(0.102)	(1.362)

对用户评分与情感分析两个变量的关系探究

	Dependent variable:			
	AvgRatingStarsThisUser		sentiment	
	ordered logistic		*OLS*	
	Base	Base_C	Base	Base_C
sentiment	0.366***	0.451***		
	(0.007)	(0.009)		
log(WC)		-1.550***		2.085***
		(0.035)		(0.037)
Not_Disclosure		0.279*		0.135
		(0.147)		(0.186)
women		0.438***		0.070
		(0.138)		(0.172)
MidAge		0.088		-0.0001
		(0.164)		(0.207)
OldAge		0.140		0.350
		(0.175)		(0.217)
factor(AvgRatingStarsThisUser)2			1.136***	1.401***
			(0.120)	(0.103)
factor(AvgRatingStarsThisUser)3			2.950***	3.311***
			(0.108)	(0.096)
factor(AvgRatingStarsThisUser)4			4.656***	5.073***
			(0.101)	(0.094)
factor(AvgRatingStarsThisUser)5			5.108***	5.783***
			(0.098)	(0.094)
Constant			-0.221**	-9.015***
			(0.086)	(1.115)

One Step Further – Part of Speech Analysis

- R与Python的“强强联合”
- 基于spacyr，reticulate

Preprocessing Data & Constructing Control Variables

- **筛出UTF-8格式且为英文的评价**

```
ReviewText = iconv(ReviewText, "UTF-8", "UTF-8", sub="")
language = textcat(ReviewText)
```

- **构建变量**

Reviewer specific variables：Age、Gender、Identity Disclosure

Review specific variables：Rating、Rating Deviation、Length of Review、Readability

Hotel FE： `HotelID = as.factor(HotelID)` **#因子化**

Time FE： `year = as.factor(str_extract(year_month, “\\d{2}”))` **#提取年份、因子化**

Construct Variables: Anxious & Angry

- **标准化语料库**

```
review_tokens <- review_corpus %>%
  tokens(
    remove_punct = T,
    remove_symbols = T,
    remove_numbers = T,
    remove_separators = T
  ) #去标点符号、数字、无用空白
```

- **生成DTM矩阵**

```
review_DTM <- review_tokens %>%
  dfm(
    tolower = T,
    stem = T,
    remove = stopwords("english")
  ) #转为小写、词干化、去除停止词
```

quanteda 3.2.0

Construct Variables: Anxious & Angry

- 构建关于Angry与Anxious的自定义词典
- 在DTM矩阵中筛出包含在词典内的词汇
- 依照公式得到Angry与Anxious变量的具体值

$$Angry_i = \frac{\sum Term\ Frequency_{Angry\ Words_i}}{Word\ Count_i}$$

$$Anxious_i = \frac{\sum Term\ Frequency_{Anxious\ Words_i}}{Word\ Count_i}$$

Construct Variables: Readability & Sentiment

- **可读性分析**

```
quanteda::textstat_readability(method = “Flesch”)
```

#得到每条评论的Flesch-Kincaid可读性得分

$$Flesch - Kincaid\ Score = 206.835 - 1.015(Words/Sentences) - 84.6(Syllables/Words)$$

- **情感分析**

```
syuzhet::get_sentiment(method = “nrc”)
```

Based on “Embedding”

```
quanteda.sentiment::textstat_polarity(dictionary = data_dictionary_NRC)
```

Based on “Logit”

Two Methods for Robustness

Negative Binomial Regression

- 负二项回归
  - 不含固定效应的回归
  - 含双向固定效应的回归
  - 加入文本变量的回归
  - 加入控制变量的回归
  - 依据两次情感分析的结果对回归进行稳健性检验

```
MASS::glm.nb(link = log)
```

3 How We Made it

	Dependent variable:			
	NumHelpful			
	Text	Text_test	Text_FE_C	Text_FE_C_test
AvgRatingStarsThisUser			-0.184***	-0.176***
			(0.020)	(0.021)
Rating_Deviation			0.238***	0.237***
			(0.027)	(0.027)
log(WC)			0.359***	0.377***
			(0.033)	(0.029)
Not_Disclosure			-0.307**	-0.305**
			(0.126)	(0.126)
women			-0.248*	-0.248*
			(0.127)	(0.127)
MidAge			-0.228	-0.228
			(0.145)	(0.145)
OldAge			-0.275*	-0.272*
			(0.164)	(0.164)
angry	0.295***	0.168**	-0.001	-0.006
	(0.071)	(0.072)	(0.064)	(0.064)
anxious	0.230***	0.111*	0.092**	0.089**
	(0.059)	(0.060)	(0.043)	(0.044)
readability	0.024***	0.026***	0.001	0.001
	(0.005)	(0.005)	(0.004)	(0.004)
sentiment	0.026***		0.007	
	(0.006)		(0.006)	
sentiment_test		-0.131***		0.006
		(0.022)		(0.023)
Constant	-1.351***	-1.123***	-0.006	-0.095
	(0.049)	(0.050)	(0.462)	(0.456)
Observations	10,044	10,044	10,044	10,044
Log Likelihood	-7,581.791	-7,574.484	-6,204.329	-6,204.963
theta	0.337*** (0.015)	0.340*** (0.015)	2.196*** (0.209)	2.197*** (0.209)
Akaike Inf. Crit.	15,173.580	15,158.970	12,512.660	12,513.920

Note: *p<0.1; **p<0.05; ***p<0.01

情感分析变量的稳健性检验

- 可以看到此处系数、符号与显著性均未发生较大改变

WordCloud, One Step Further?

- **What does the wordcloud explain? Can we make it more specific?**

Young

Old

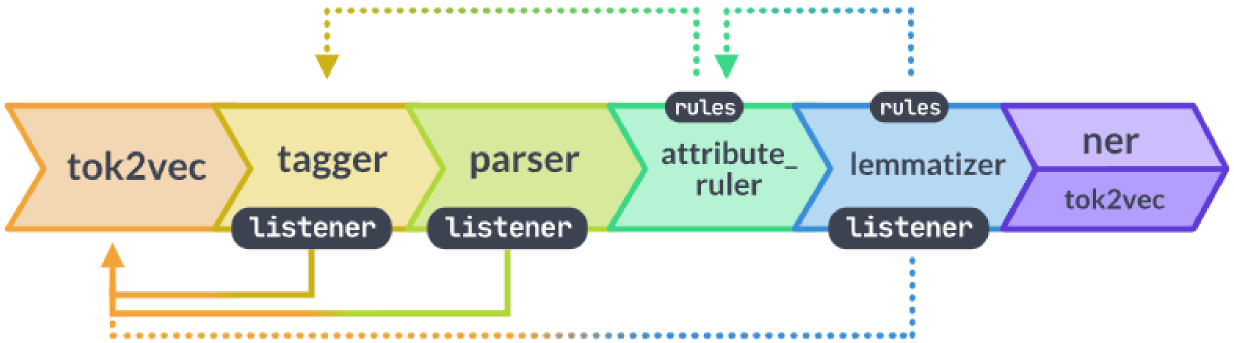
What is spacyr & How does it work?

spaCy

spacyr

- Industrial-Strength Natural Language Processing
- 在R中调用，深度CNN语言模型与简洁处理“强强联合”
- 继承&集成函数

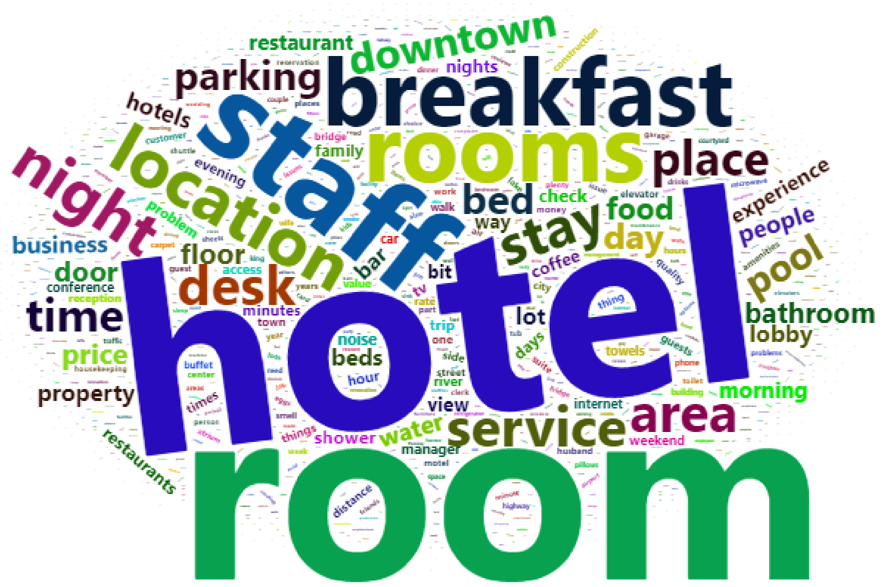
```
spacy_parse(pos = TRUE,
            dependency = TRUE,
            entity = TRUE,
            noun_phrase = TRUE)
#提取词性、依赖关系、实体关系、名词短语
```



spaCy流水线示意图

WordCloud Based on POS

- 利用形容词形成的词云更能反映客户心态
- 利用名词形成的词云能够提取客户关心的特征
- 参考(Nikolay Archak et. al, 2011)



Sentiment Analysis Based on Features

- **从词云中直观提取特征——客户关心的酒店客观条件**
- **编写函数：特征情感提取器**

利用匹配的思想，将至少含有一个上述特征的样本从语料库中筛出

通过依赖关系找到形容该特征的词汇

按照情绪词典匹配得分，平均得到该条评论在该特征维度的得分

$$sentiment_{feature,i} = \frac{\sum sentiment\ score\ on\ selected\ feature}{num\ of\ mentions\ of\ selected\ feature}$$

- **通过函数得出共八个特征（右表）在每个样本上的情感得分**

Extracted Features
room
service
parking
breakfast
bed
price
staff
location

Rating & Sentiment on Features

- DV: Rating - Ordered Discrete Variable (from 1 to 5)
- IV: Sentiment on features
- Control: Age, Gender, Identity Disclosure, Readability, Word Count
- Hotel FE, Year FE √
- Discrete Choice Model – Ordered Logistic Regression

```
DCM <- MASS::polr()
```

4 Further Exploration

武汉大学经济与管理学院
Economics and Management School of Wuhan University

	Dependent Variable
	rating
room	0.491***
service	1.009***
parking	0.139
location	0.316***
breakfast	0.660***
staff	0.554***
bed	1.005***
price	0.603***
women	0.534***
MidAge	0.087
OldAge	0.349**
readability	0.004
Log(WC)	-0.805***
Not_Disclosure	0.342**
Hotel FE	√
year FE	√

Intercepts:

	Value	Std. Error	t-value
1|2	-6.2785	0.2279	-27.5528
2|3	-5.2254	0.2255	-23.1731
3|4	-4.1208	0.2238	-18.4111
4|5	-2.7025	0.2227	-12.1377
Observations:	10044		
Residual Deviance:	26766.53		
AIC:	26882.53		

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

- **提取出的特征大多对评价具有显著影响**
- **关于影响不显著的parking**

Outlook for Future Directions

- 复现基于特征的情感分析对销量/收入的影响（Nikolay Archak et. al, 2011)
- 继续挖掘离散选择模型与文本分析产生数据之间的交融
- 采用分层聚类、关联规则等无监督学习方式深度挖掘文本信息
- 利用更加精确的语言网络模型——预训练模型+针对训练

武汉大学经济与管理学院
Economics and Management School of Wuhan University

谢谢观看

自强，弘毅，求是，拓新

汇报小组 ： 第一组
汇 报 人 ： 刘郅哲，谢燊
汇报时间 ： 2022-6-27