Psychological Research with Social Media Posts and Computational Text Analysis

Lee-Xieng Yang

National Chengchi University

Outline

- Review some studies designed to use social media data to investigate psychology issues
- Summarize the common procedure of doing a study with social media data
- Demonstrate how the computational text analysis can be applied to dealing with social media data to answer psychology questions, with the gender differences as the example

Social media as new source of human data

- Psychologists have been trying to understand the human mind via observable behaviors, including responses in the laboratory experiments, for the items in questionnaires, physical activities in a task, etc.

- Nowadays, people are used to sharing their lives on social media, directly or indirectly exposing their feelings, emotions, attitudes, or opinions to public events, etc.

- Social media become a new source to observe human behaviors

Emotional contagion through social networks I

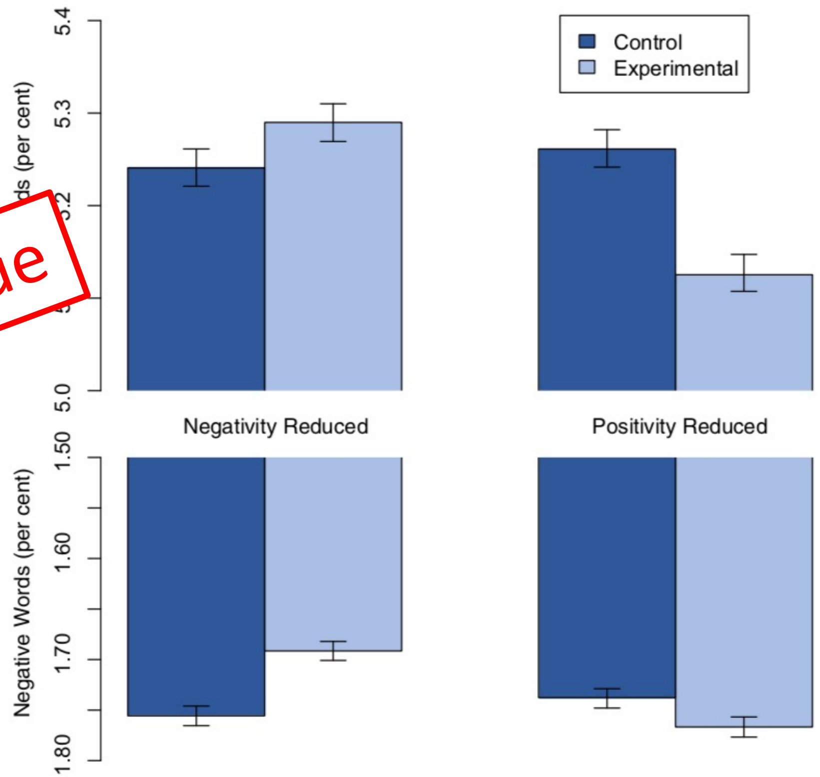
- Kramer, Guillory, & Hancock (2014) (Facebook data-science team)

- Subjects: 689,003 Facebook users

- Manipulate the exte[illegible] which people (experimental group) were exposed to emotional expressions in their News Feeds

Research Ethics Issue

Results



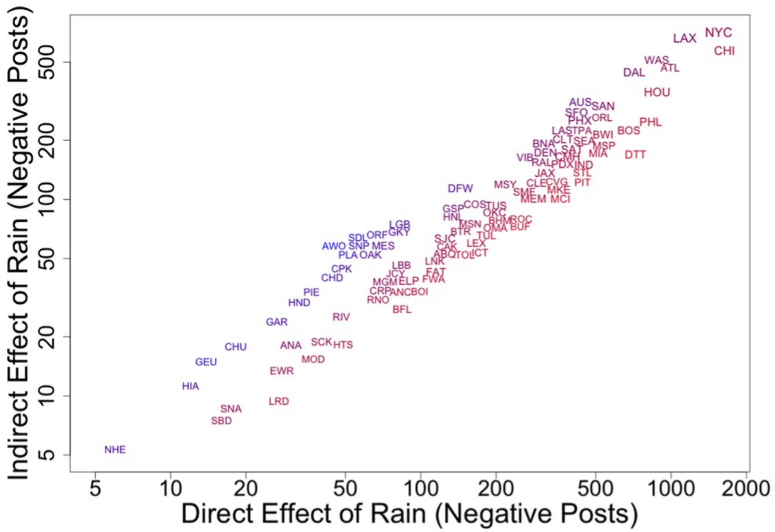
Emotional contagion through social networks II

- Coviello et al. (2014)

- Collected posts of Facebook users in the 100 most popular US cities from January 2009 to March 2012

- The positive rate and negative rate of their posts as DVs

Result



Direct effect: negative posts due to rainfall
Indirect effect: negative posts due to the negative posts of friends

Facebook likes predict personal attributes I

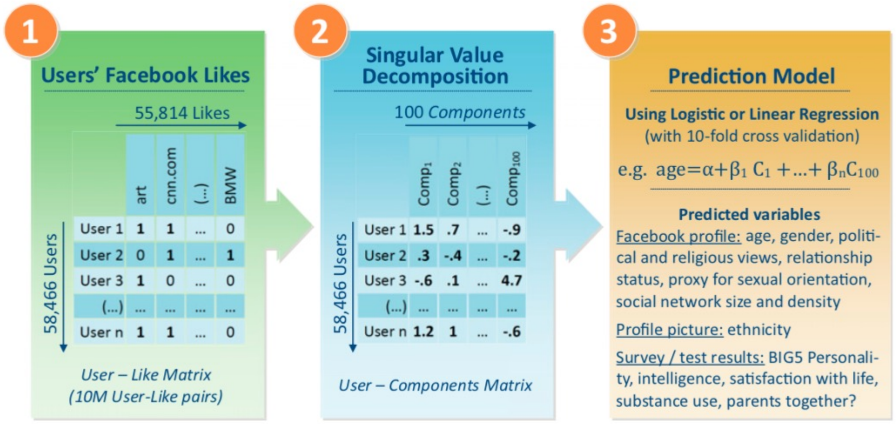
- Kosinski, Stillwell, & Graepel. (2013)

- Use myPersonality application on Facebook (Kosinski & Stillwell, 2011) to collect data

- Got 58,466 Facebook users’ authorization to use their data on Facebook for research purpose

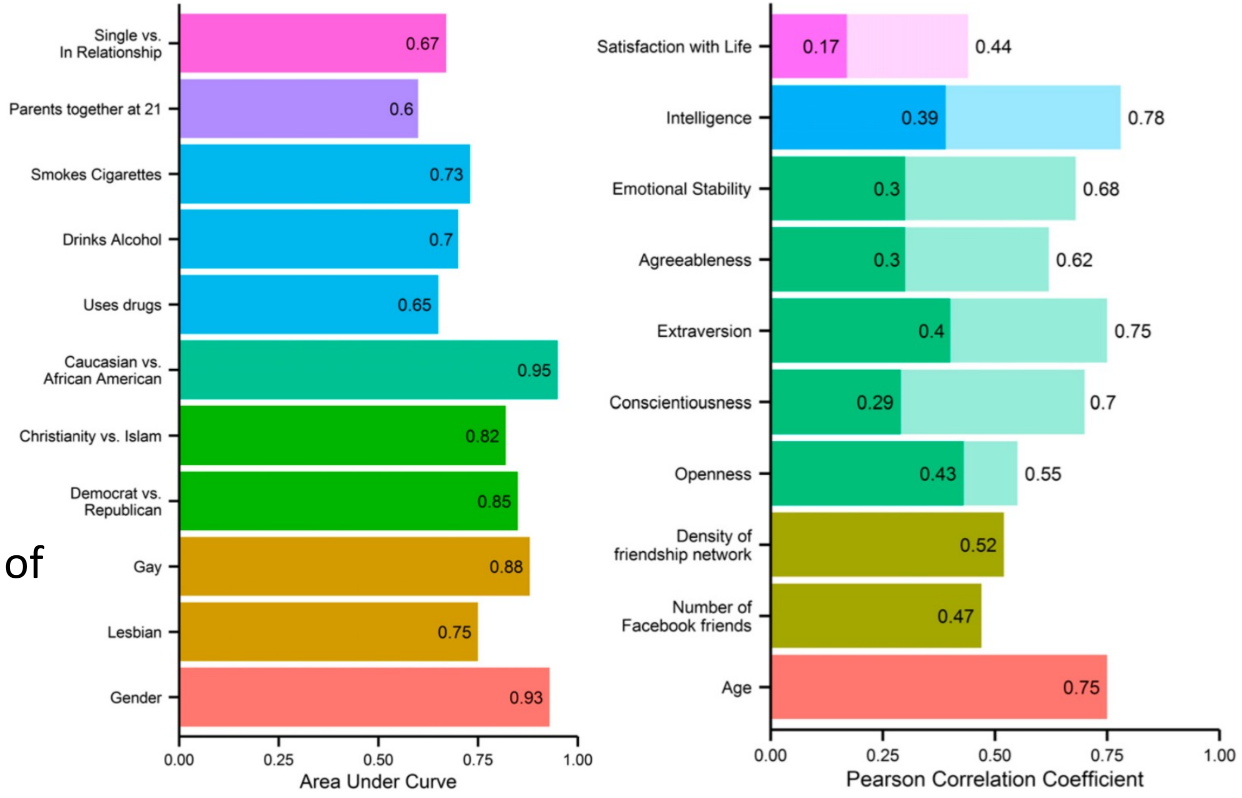
Facebook likes predict personal attributes II

Procedure:



Main Findings: Facebook likes can predict sexual orientation, ethnicity, religious, and political views, personality traits, intelligence, happiness, use of addictive Substances, parental separation, age, and gender.

Results



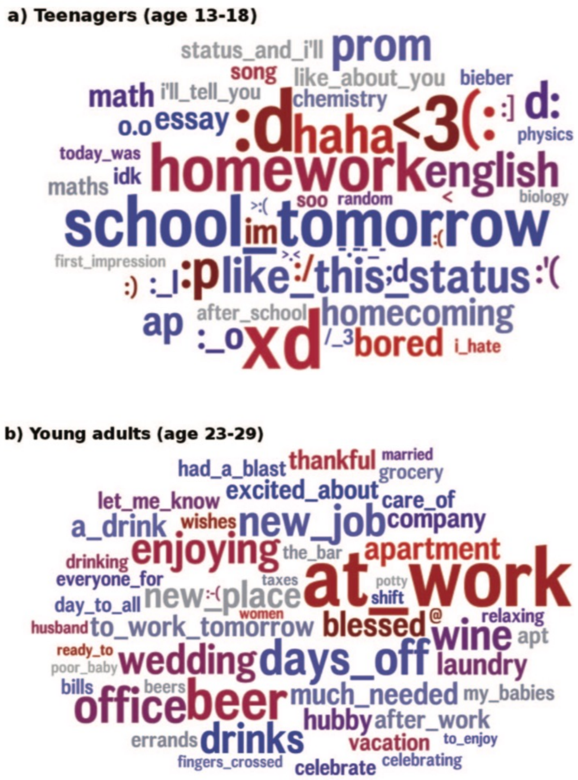
All p values are significant

Aging positive effect I

- In the aging study, it’s often found that the elder people are positive toward lives than the younger people do
- Kern et al. (2014) collected 74,859 Facebook users’ status updates via myPersonality application
- Analyzed the word usage for different age groups

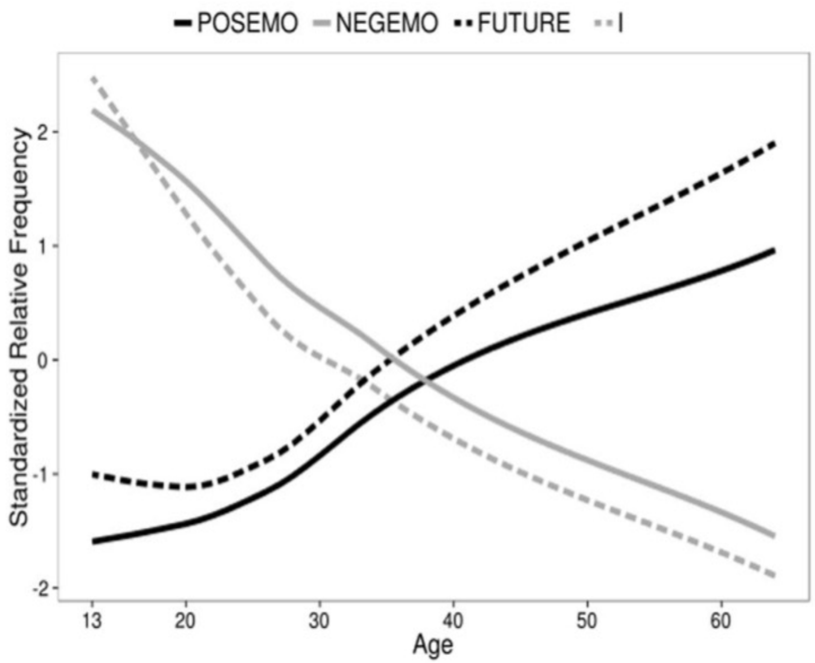
Aging positive effect II

1. Preferred words differed among age groups
2. So did the topics
3. The elders preferred using positive emotional words



Words

Topics



Summary

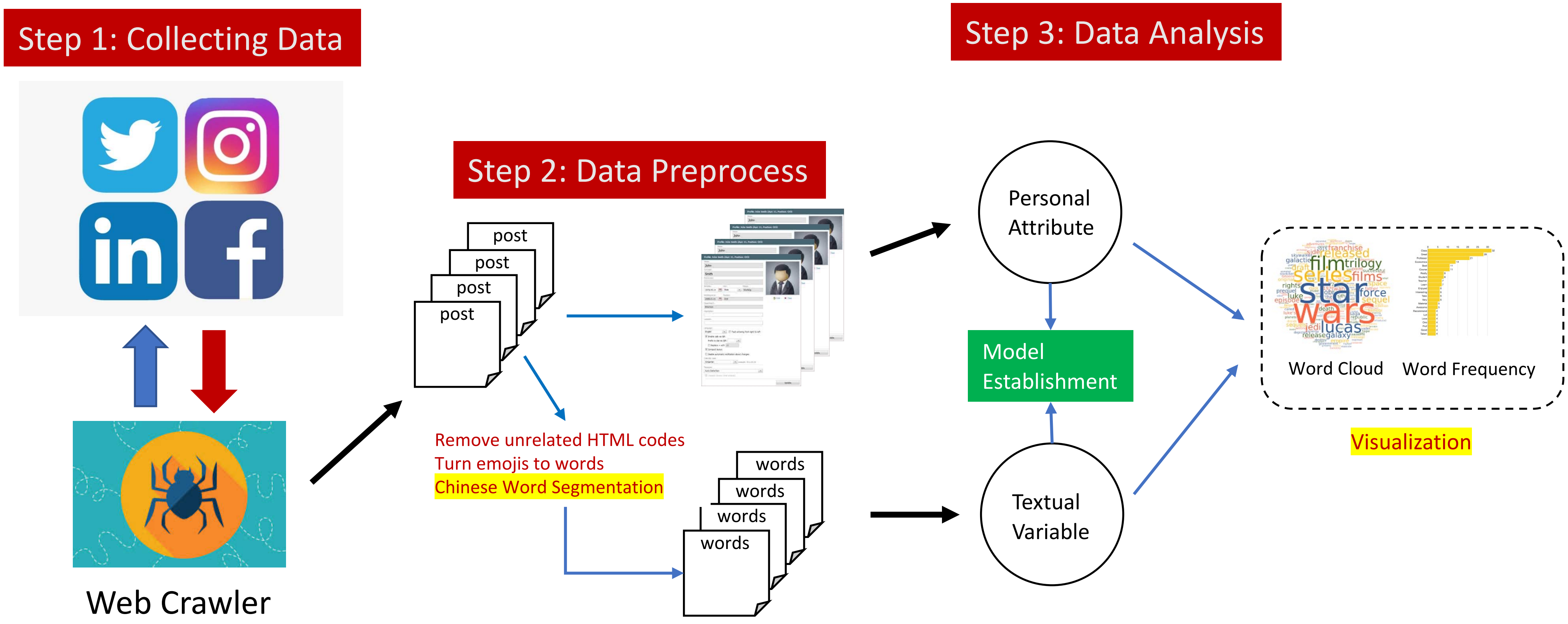
- Social media as a huge data base contain various behavior data for psychological research

- The digital records (e.g., likes, words in posts) can predict many human attributes, such as personality, age, gender, etc

- However, causality between variables cannot be established via social media research

- Also, research ethics issues need to be addressed

Generic procedure for social media research



Computational text analysis I

- Closed vocabulary analysis (mostly used in Psychology)
  - LIWC (Language Inquiry and Word Count)
  - CLIWC (Chinese version of LIWC)
  - A dictionary sorting out words by psychological and linguistic attributes
  - More than 71 word catalogs, including positive emotional words, negative emotional words, cognitive words, etc.
  - Normally, researchers recode texts as the distribution over (C)LIWC words
  - Top-down style

Computational text analysis II

- Closed vocabulary analysis (mostly used in Machine Learning)
  - Bottom-up style
  - Extract out the keywords/topics from texts by algorithm
  - TF-IDF (Term Frequency-Reversed Document Frequency): Extract out the words best representing the texts controlled by word frequencies in corpus
  - Stylometric analysis: Extract out the words best representing the style of the author
  - LDA (Latent Dirichlet Allocation): Summarize the topics (a bag of words) underneath the texts

DOI:10.6129/CJP.201909_61(3).0003
中華心理學刊 民 108，61 卷，3 期，209-230
Chinese Journal of Psychology 2019, Vol.61, No.3, 209-230

五南

A Text Analysis Approach to Analyzing Gender Differences in Breakup Posts on Social Media

Lee-Xieng Yang[1, 2] and Ching-Fan Sheu[3]
Department of Psychology, National Chengchi University[1]
Research Center for Mind, Brain, and Learning, National Chengchi University[2]
Institute of Education, National Cheng-Kung University[3]

A Test Analysis Approach to Analyzing Gender Differences in Breakup Posts on Social Media

Lee-Xieng Yang
National Chengchi University

Ching-Fan Sheu
National Cheng-Kung University

Main goals of this study

- Methodological goal:
  - Compare the closed and open vocabulary analysis on the accuracy of predicting the gender of authors via their breakup posts
- Theoretical goal:
  - Understand the differences between male and female on their posts about their romantic breakup

Target Social Media: Dcard

- The biggest anonymous social media in Taiwan
  - More than 4 million registered users (about 60% male and 40% female)
- Anonymity
  - No user name, ID, or IP is accessible
  - But user gender is public information and school information can be chosen to be public
  - None of the users could be identified by researchers
- Easy to identify breakup stories
  - Breakup is a category of the posts on the relationship forum
  - PTT or Facebook provides no such a tag for each article


Dcard
搜尋 最浪漫的小事
註冊 / 登入
所有看板
即時熱門看板
好物研究室
遊戲專區
即時熱門看板
感情
有趣
心情
女孩
閒聊
美食
梗圖
工作
更多
分手
追蹤
46338 篇文章 · 4140 人追蹤
熱門文章
最新文章
感情 · 匿名
#圖 woo到男友約炮
首先，我要在這裡鄭重的感謝woo，讓我看清了我的男友，噢，我是說前男友...
62684
3934
收藏
感情 · 匿名
再不分手我們就老了
「你在哪裡？」「家裡。」「喝點什麼嗎？」「不用。」「你餓嗎？」「隨便買。」不到一分...
60740
1454
收藏
感情 · C
分手之後
分手之後我掉進一個坑裡。我做了些事讓自己爬出坑洞，例如剪短頭髮，聽重...
47777
396
收藏

Some examples

是的 ，我們分手，你頭也不回的走了。留著還愛你
的我。我還是放不下你，還在說服自己你還在我身
旁吧你卻再
怕你的冷言
你就是狠下
是會滑落每
怕清楚的告
剩下我一個
走走時，也
此牽著手走

曾經有個讓我到現在遲遲難以放下的感情這是場姐弟
戀而前男友也對我很好他願意為了我改變 很聽我的話
他在我不開心時 會想
都很尊重我對我而言
也許是因為我比較大
分手 是因為一些原因
左右了我一直在想 是
不好 或許我能去陪伴
不是就該陪伴他成長
事情是會給我意見、

昨天你跟我提分手了我問你為什麼你告訴我你沒那麼
喜歡我了習慣我對你好可是已經沒有愛了感情早就變
得太平凡可是親愛的我真的不知道怎麼辦因為我很愛
你所以我才願意為你
理由寧可是我做錯了
在這樣我找不到出口
捨不得離開我們的回
相簿因為全部都是你
要怎樣眼淚才會停住

失戀分手不可怕可怕的是自己有沒有那個勇氣那份勇敢
走出情傷在無限的鬼打牆中求助無門那又能如何呢……
不管是分手後想復合也好想找下一段也好最重要的是先
從➡️尊敬一事無成的自己開始也才有籌碼說復合 或是開
始下一段感情因為自己都不自己了誰還有辦法幫你妳 自己
才是自己心中最大的心魔那道牆也只有自己才能翻得過去
最後送上 -的這首-自己都不自己-兩年感情分手後第天的想
法一起加油吧🙂

Demographic Data of Posts

- Data collection time period: February 1 – June 8, 2017
- Base rate of genders (N = 1,311)
  - Female vs. Male = 71% : 29%
- Mean length of article
  - Female vs. Male = 257.29 words : 255.53 words ($t = 0.10$, $p = .92$)
- Mean received comments
  - Female vs. male = 18.98 comments : 18.18 comments ($t = 0.23$, $p = .82$)

Preprocessing: Chinese Word Segmentation

- In English, a word can be easily identified in a sentence by spaces

A **relationship breakup**, often referred to simply as a **breakup**,[1] is the termination of an intimate relationship by any means other than death. The act is commonly termed "dumping [someone]" in slang when it is initiated by one partner.[*citation needed*] The term is less likely to be applied to a married couple, where a breakup is typically called a separation or divorce. When a couple engaged to be married breaks up, it is typically called a "broken engagement".

- In Chinese, there is no space within a sentence!

我愛我的前男友，但我知道我們之間有太多太多的不一樣，不一樣的價值觀、不一樣的生活圈，生活中大大小小的摩擦也同時把我的勇氣消磨光了，所以我選擇分開。不過每當想起他的時候，都覺得特別自責，明明我是讓他受傷的人，自己卻每天難受的理直氣壯。

We used Jieba (open source project for text processing) to do word segmentation

"所以" "我" "選擇" "分開"

Closed vocabulary analysis: CLIWC

Regress gender on probabilities of CLIWC categories

$$l = \frac{1}{1+e^{-y}}, y = \sum \beta_i p_i,$$

$$p_{i,post} = \frac{fre_{i,post}}{length(post)},$$

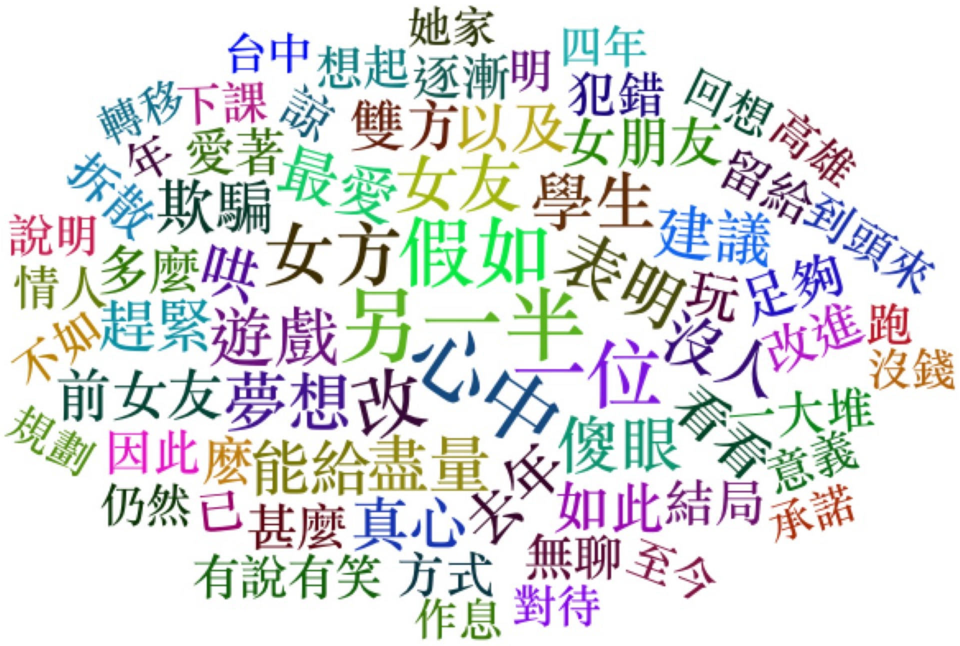
$y = 1$ for female and 0 for male

	Model 1	Model 2	Model 3	Model 4
Predictors	10人稱代名詞：	12非人稱代名詞： 悲傷詞、生氣詞與焦慮詞	Model 1 + Model 2	All 70 word categories
Accuracy	.85	.71	.72	.72
AIC	1067	1559.2	1569.4	1600.6
df	1300	1298	1288	1240

But the personal pronouns are not only used in break-up posts!

Open vocabulary analysis: Stylometric analysis

Female: 695 Words



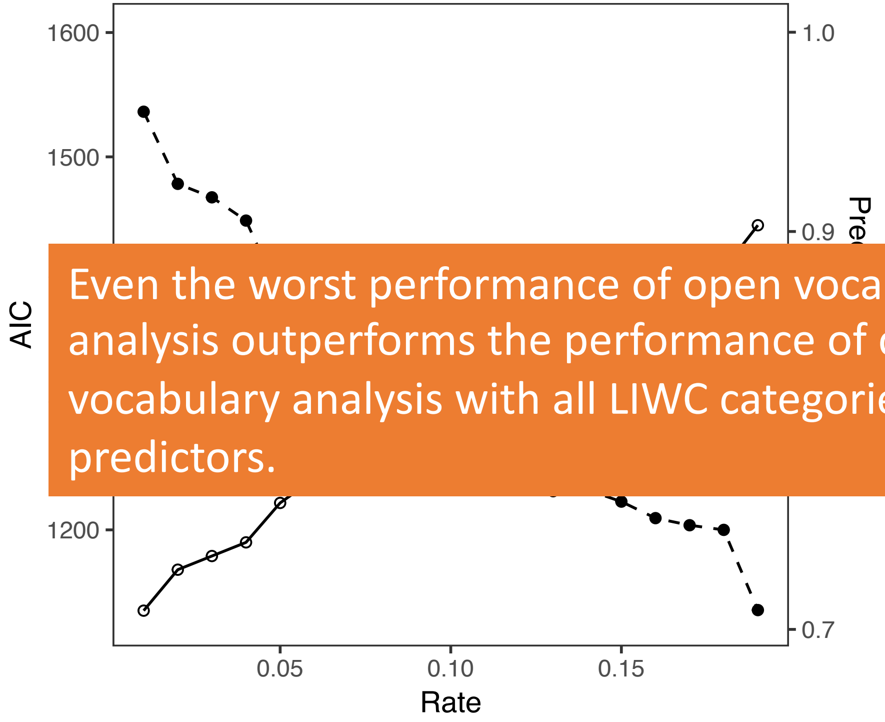
Male: 811 Words

Performance of Open Vocabulary Analysis

Again, gender is regressed on the keywords for the authors.

The prediction accuracy increases, as the proportion of keywords chosen in the model increases (from top 1% to top 19%).

However, this does not result from the increased number of parameters, as the AIC decreases all the way down.



Even the worst performance of open vocabulary analysis outperforms the performance of closed vocabulary analysis with all LIWC categories as predictors.

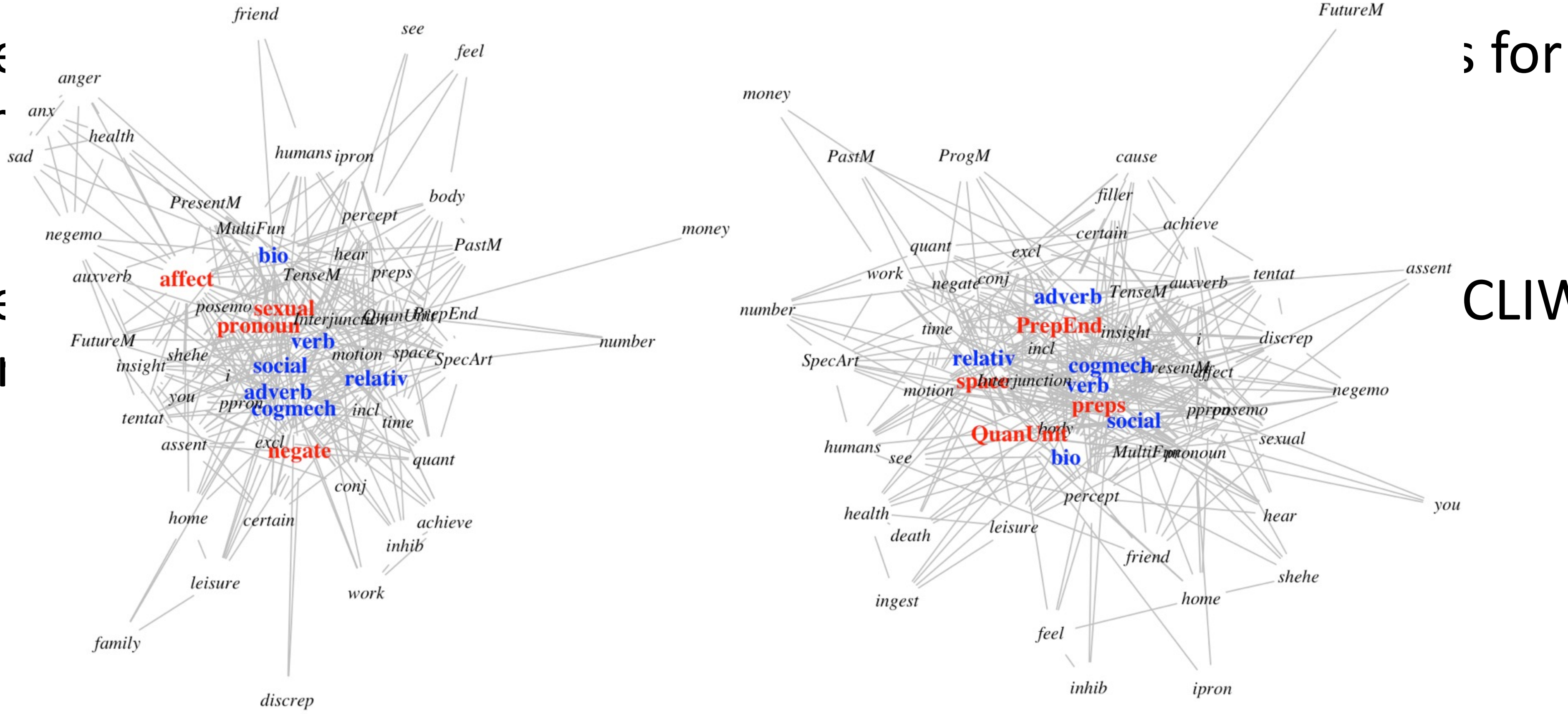
Summary for Comparison between Two Vocabulary Analyses

- Open vocabulary analysis is better at extracting the key features of texts than closed vocabulary analysis

- However, what psychological aspects do those keywords actually reflect cannot be revealed by open vocabulary analysis

Combining open and closed vocabulary analysis

- Che… s for each ger…
- Use… CLIWC attr…



Female Male

Conclusion for now

- Open vocabulary analysis is better than closed vocabulary analysis at extracting key features of text
- However, closed vocabulary analysis can provide psychological explanations to the keywords of text
- The combination of these two types of vocabulary analyses leads to a new research framework for text analysis
- In breakup posts, both genders make self exposure on the aspects of cognitive and social processes. However, only females have the expressions of emotion and affection as the focus

Gender Differences in Topics of Breakup Posts on Social Media with Topic Model

Lee-Xieng Yang
National Chengchi University

Ching-Fan Sheu
National Cheng-Kung University

Goal of this study

- In this follow-up study, we tried to further extract out the contents instead of words from the breakup posts

- To this end, hierarchical Dirichlet process mixture model (HDPMM) would be applied to extract the topics from the posts

- As a comparison baseline, TF-IDF would be applied to extract the keywords from the posts

Demographic Data

- Posts collected in between March 3 and July 16, 2019
  - In total 25,000 posts, 4,142 posts were tagged with breakup
- Gender ratio
  - There were 1,417 male posts and 2,725 female posts
  - The gender ratio is about 1 : 2 for males vs. females
- Mean number of words per post
  - Male vs. Female = 287.75 : 279.75 ($t = 0.73$, $df = 4,140$)
- Parts of speech
  - No gender differences on the numbers of nouns, adjectives, adverbs, verbs, auxiliaries, and intonations

TF-IDF results: Word cloud in Chinese

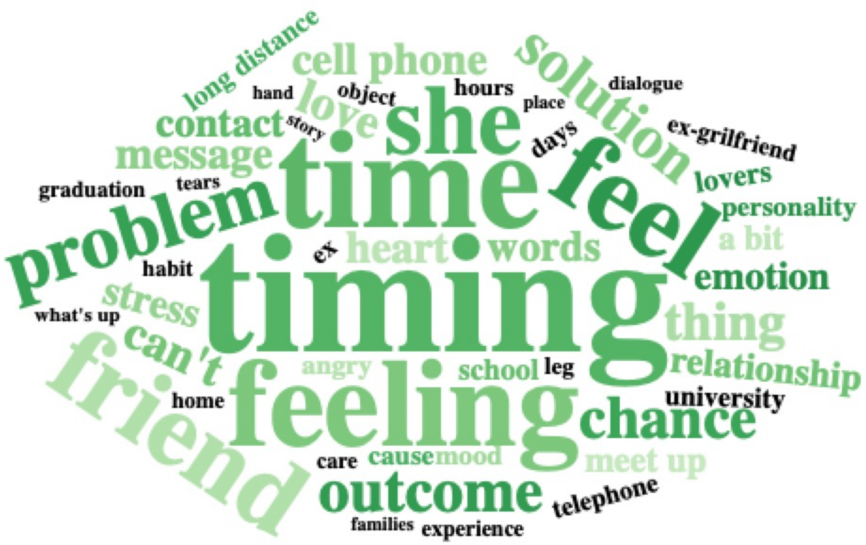
Male

Female

The overlapping rate between two genders is 45/50 (90%)

TF-IDF results: Word cloud in English

Male



Female

Summary

- There are 45 out of the 50 most frequent key nouns shared by two genders
- There is not too much difference between male and female in terms of the most frequent words
- Could it be possible that the gender difference does not result from the words being used, but the way of organizing those words?

Topic Model

- A topic is a bag of words
  - A topic represents a probability distribution of words
  - Could the gender difference be revealed in the level of topic?

- LDA (Latent Dirichlet Allocation) is often used to summarize the topics of texts

- However, how to determine the number of topics is an issue and both genders presumably should have some overlaps on breakup topics
  - Hierarchical Dirichlet process mixture model (HDPMM) is used instead

Method

- Instead of the key nouns, we simply used all 7,345 nouns as our target
- The probability of each noun occurring in each gender's posts was computed as the frequency ratio of it over all nouns
  - There were 7,345 data points for each gender
- Modeling with HDPMM

HDPMM

- Each gender's data points were modeled by the sum of a family of Beta distributions weighted by Dirichlet process
- Those Beta distributions were generated from the base measure G_j
- The base measure G_j itself was also generated by a Dirichlet process

Hierarchical Beta Dirichlet Mixture Model

$y_{ij} \sim F(\theta_{ij}),$

$\theta_{ij} \sim G_j,$

$G_j \sim DP(\alpha_j, G_0),$

$G_0 \sim DP(\gamma, H)$

i: data point

$j = 1$: male and 2: female

α_j: concentration parameter

γ: top−level concentration parameter

F: weighted sum of Beta distributions

Settings for modeling

Prior parameters for γ : 2 and 4
Prior parameters for α : 2 and 4
Hyper prior parameters for G_0: 1 and 0.01
Metropolis Hastings jump size: 0.1 for each parameter
Iterations : 100

Topics Generated by HDPMM

Modeling result: 10 topics generated for male and 9 for female

Parameter	Topics									
					Male					
	1	2	3	4	5	6	7	8	9	10
μ	0.22	0.46	0.56	0.31	0.33	0.29	0.25	0.32	0.42	0.35
υ	1.46	2.02	1.80	2.37	2.06	1.70	1.69	2.31	1.83	1.69
					Female					
	1	2	3	4	5	6	7	8	9	10
μ	0.22	0.46	0.56	0.31	0.33	0.29	0.25	0.32	0.42	
υ	1.46	2.02	1.80	2.37	2.06	1.70	1.69	2.31	1.83	

Distributions of Topics

One common topic for both genders Two specific topics for each gender

	Topics									
	Male									
	1	2	3	4	5	6	7	8	9	10
Word Num	121	77	55	140	155	101	440	4357	1847	52
	Female									
	1	2	3	4	5	6	7	8	9	10
Word Num	31	114	485	1310	24	28	15	5275	63	

	Mean Probabilities of Top 30 Nouns				
	3	4	8	7	9
M			.007	.0006	.001
F	.001	.002	.009		

Words of The Common Topic in Chinese

Male | **Female**

Mean Probability = .007

Mean Probability = .009

The overlapping rate is 18/30

Words of The Common Topic in English

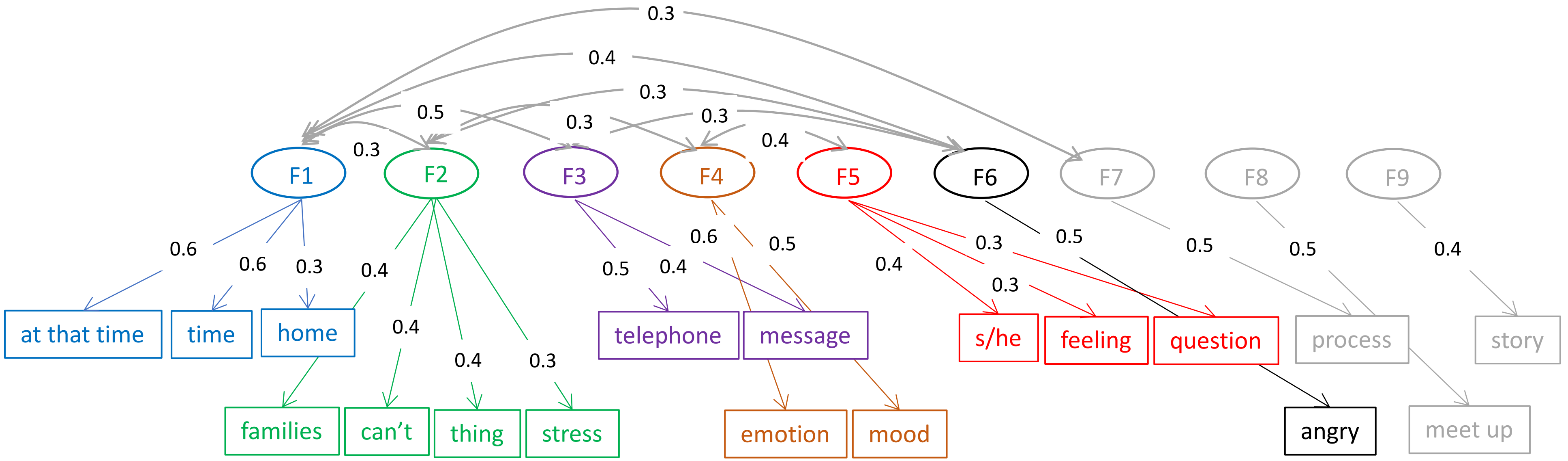
Male

Mean Probability = .007

Female

Mean Probability = .009

EFA for Words in Common Topic



RMSR = .02
RMSEA = .024
TLI = .91

The breakup stories of college students in Taiwan can mainly be described in respects of background (F1), stress from other people (F2), communication (F3), mood (F4), feeling for the partner (F5) and negative emotion (F6).

Words of Specific Topics of Males in Chinese

Topic 7

Mean Probability = .0006

Topic 9

Mean Probability = .001

Words of Specific Topics of Males in English

Topic 7

Mean Probability = .0006

Topic 9

Mean Probability = .001

Words of Specific Topics of Females in Chinese

Topic 3

Mean Probability = .001

Topic 4

Mean Probability = .002

Words of Specific Topics of Females in English

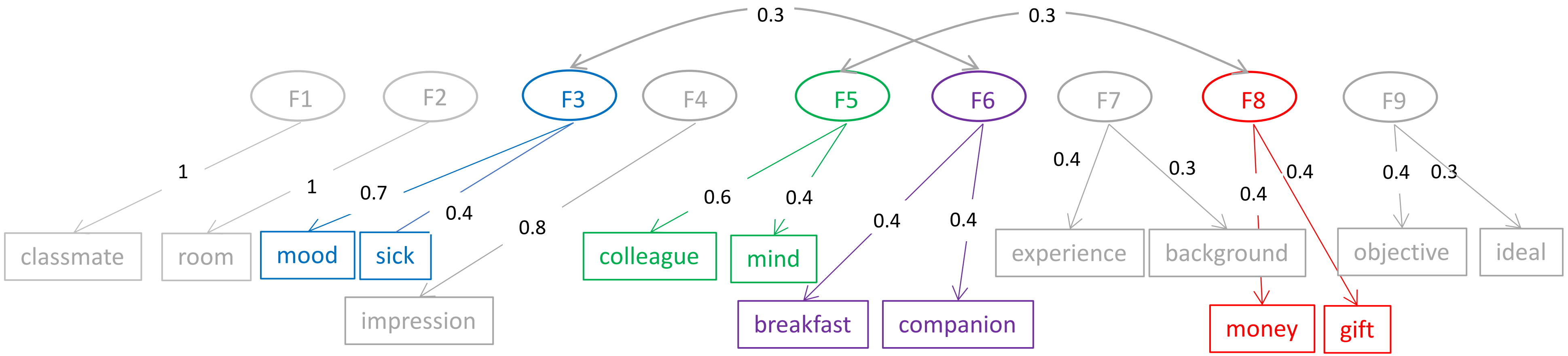
Topic 3

Mean Probability = .001

Topic 4

Mean Probability = .002

EFA for Words in Female-Specific Topic



RMSR = .02
RMSEA = .018
TLI = .88

Two particular aspects are suggested for females.
1. Sentimental aspects: F3 and F6
2. Social comparison: F5 and F8

Conclusions

- Gender differences in breakup posts on social media can be revealed in the topic level not the word level
- The common topic between males and females suggests the factors of breakup stories in Taiwan, including the stress from other people, communication, mood, feeling for partner, and negative emotion
- The female-specific topic particularly shows the factors of sentiment and social comparison