Building a Subjective Well-being Profile with Social Media Language

Lucia Chen

Colleagues

- Chen Lushi
- Master in Applied Psychology (City University of Hong Kong)
- Master in Linguistics (The University of Hong Kong)

- Tao Gong (Ph.D.)
- Research Scientist in Yale University (computational linguistics)

- Rob Davidson (Ph.D.)
- Open Data Lead at UK National Statistics

Content

- The Power of Big Data and Psychographics
- How to predict psychological traits from social media data.
  - machine learning algorithm
  - Feature selection
    - Sentiment analysis
    - Basics of NLP
      - Document term matrix
      - Word count
      - N gram
    - Latent Dirichlet allocation (LDA) topic analysis

Social media data

Computer-mediated tools that allow people or companies to create, share, or exchange information, career interests, ideas, and pictures/videos in virtual communities and networks.
http://www.hatdex.org/

What can social media data tell us?

Sentiments (positive & negative emotions)

e.g. movie review, writing style

Personality (David Stillwell)

Self-disclosure

Depression

Subjective well-being

Relevant studies

- Predict happiness with Twitter (Mitchell, et al., 2013)
- Predicting historical events and economic trends with sentiment analysis on Google books (Hills et al., 2015)
- Predict personality from FB 'LIKES' (Wu et al., 2015)
- Predict personality from FB language (Park et al., 2015)
- Predict self-disclosure with FB data
- Predict subjective well-being with FB language

Building a Subjective Well-being Profile with Facebook Language

- Subjective well-being profile

  Happiness (affect)

  Satisfaction with Life (SWL) prediction model (RF)

- Two Major Findings

Process

- Collect data
  - participants finish a measurement scale (factors in the scale. E.g. 5 dimensions in personality) (you probably want to have one model for each dimension)
  - collect participant's social media data (twitter API, scrappy) (be aware of the legal issue)
- Data preparation and data pre-processing
  - Clean data (missing data, outlier)
- Feature selection

Selecting features

- Literature review

- Feature reduction (LASSO, ridge regression)

(A **regression** analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.) Minimised the sum of squared error.

http://statweb.stanford.edu/~tibs/lasso.html

  - Facebook functionalities
  - Language feature
    - Linguistic Inquiry and Word Count (LIWC) - WP Engine
    - Latent Diritchlet Allocation topic modeling (LDA) (Python: Gensim, online machine learning)

LDA Topic Probability

	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	...	sum
document1	0.00365785	0.09876196	0.01378728	0.00759707	0.0104108	0.00647158	0.00590884	...	1
document2	0.01515152	0.05411255	0.03246753	0.06277056	0.01082251	0.42640693	0.01082251	...	1
document3	0.30451377	0.003153	0.00979091	0.02704945	0.00149353	0.01277796	0.01344175	...	1
document4	0.01570964	0.00379198	0.00704225	0.04496208	0.00595883	0.00704225	0.00595883	...	1
document5	0.008519	0.0163827	0.0163827	0.01769332	0.008519	0.01245085	0.02293578	...	1
document6	0.01295133	0.01138148	0.0255102	0.00824176	0.35204082	0.01844584	0.01687598	...	1
document7	0.01321752	0.46714502	0.00566465	0.00339879	0.00944109	0.01019637	0.01170695	...	1
document8	0.00099761	0.00139665	0.00179569	0.00099761	0.00099761	0.00139665	0.00139665	...	1
document9	0.01523297	0.02419355	0.03494624	0.02060932	0.0062724	0.0062724	0.06899642	...	1
document10	0.00286092	0.00418134	0.02618838	0.01210387	0.00286092	0.00814261	0.01254401	...	1

Topic modeling(R)

```
statusswl <- read.csv("user_all_status_swl.csv", header = T, fill=TRUE,row.names=NULL)

    #clean data
    ss<- sam$status
    #lower cases
    ss<- tolower(iconv(ss,"ISO-8859-1","UTF-8"))
    ss <- gsub("\\d", "", ss)
    #remove non-English words
    ss <- gsub("\\W", " ", ss)
    #remove punctuation
    ss <- gsub("[[:punct:]]"," ", ss)
    #remove http
    ...
```

https://github.com/luciasalar/Harnessing-social-media-data/blob/master/topicmodel.R

```
#create a dtm
dtm <- create_matrix(ss, language="english", removeNumbers=TRUE,
stemWords=FALSE, weighting=weightTf)
#remove sparse terms
dtm<- removeSparseTerms(dtm, .99)
```

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

LDA topic analysis

```
# Set parameters for Gibbs
sampling(http://leitang.net/presentation/LDA-Gibbs.pdf)
burnin <- 4000
iter <- 2000
thin <- 500
seed <-list(2003,5,63,100001,765)
nstart <- 5
best <- TRUE

#Number of topics
k <- 7
# LDA model
  ldaOut <-LDA(dtm, k, method= "Gibbs", control=list(nstart=nstart, seed =
  seed, best=best, burnin = burnin, iter = iter, thin=thin))
```

LDA topic analysis

```
Print(ldaOut)
Save(ldaOut, file = "LDA_Output.RData”)
```

```
Topic 1   Topic 2    Topic 3       Topic 4   Topic 5    Topic 6    Topic 7
[1,] "seat"    "dialogu"  "websit"      "census"  "northern" "growth"   "hse"
[2,] "resum"   "church"   "partnership" "disabl"  "univers"  "adjust"   "legisl"
[3,] "suspend" "congreg"  "nesc"        "cso"     "peac"     "forecast" "die"
[4,] "adjourn" "school"   "site"        "statist" "unemploy" "bernard"  "legal"
[5,] "fisheri" "survivor" "nesf"        "survey"  "polic"    "burton"   "child"
```

ML Algorithm selection

- Supervised/ unsupervised
- Regression/classification
- statistical regression (naïve Bayesian, LASSO, ridge)
- state of the art (SVM, decision tree, neural network)
- naiveBayes(e1071)　RF (random forest)

- Cheat sheet
- http://scikit-learn.org/stable/tutorial/machine_learning_map/
- https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-choice/

Language feature: LASSO/Ridge regression SVM & Random Forest

- The large number of features can bog down some learning algorithms, making training time unfeasibly long. Support Vector Machines are particularly well suited to this case

- Select features with LASSO/Ridge regression, apply features in Random forest/SVM

    - LASSO performs both variable selection and regularization in order to enhance the prediction accuracy

Model Evaluation

- Classifier's evaluation: most often based on prediction accuracy

$$\textit{Accuracy} = \frac{\text{percentage of correct prediction}}{\text{total number of predictions}}$$

- Regression evaluation: correlate predictions with self-reported values

Evaluation of Accuracy

- There are two major techniques used to calculate a classifier's accuracy.
  - 2/3 training set 1/3 testing set
  - n-fold cross-validation: divide the data up into n chunks and train n times, treating a different chunk as the holdout set each time.

A holdout set is a (usually) small set of input/output examples held back for purposes of tuning the modeling.

Improve ACC

- Feature selection: combine functionality use, language, egocentric network
- Select data according to timeline
- Document number and document length
- Experiment on different machine learning models
- Focus on feature selection, try not to massage the data!

Predictive validity

How well does your machine predicted value predict a set of criteria

- Pearson correlation
  - e.g. satisfaction with life & depression
  - linear relationship

.1 - .29 no correlation

.30 - .44 moderate

.45 + very strong

- Careful selection with external criteria

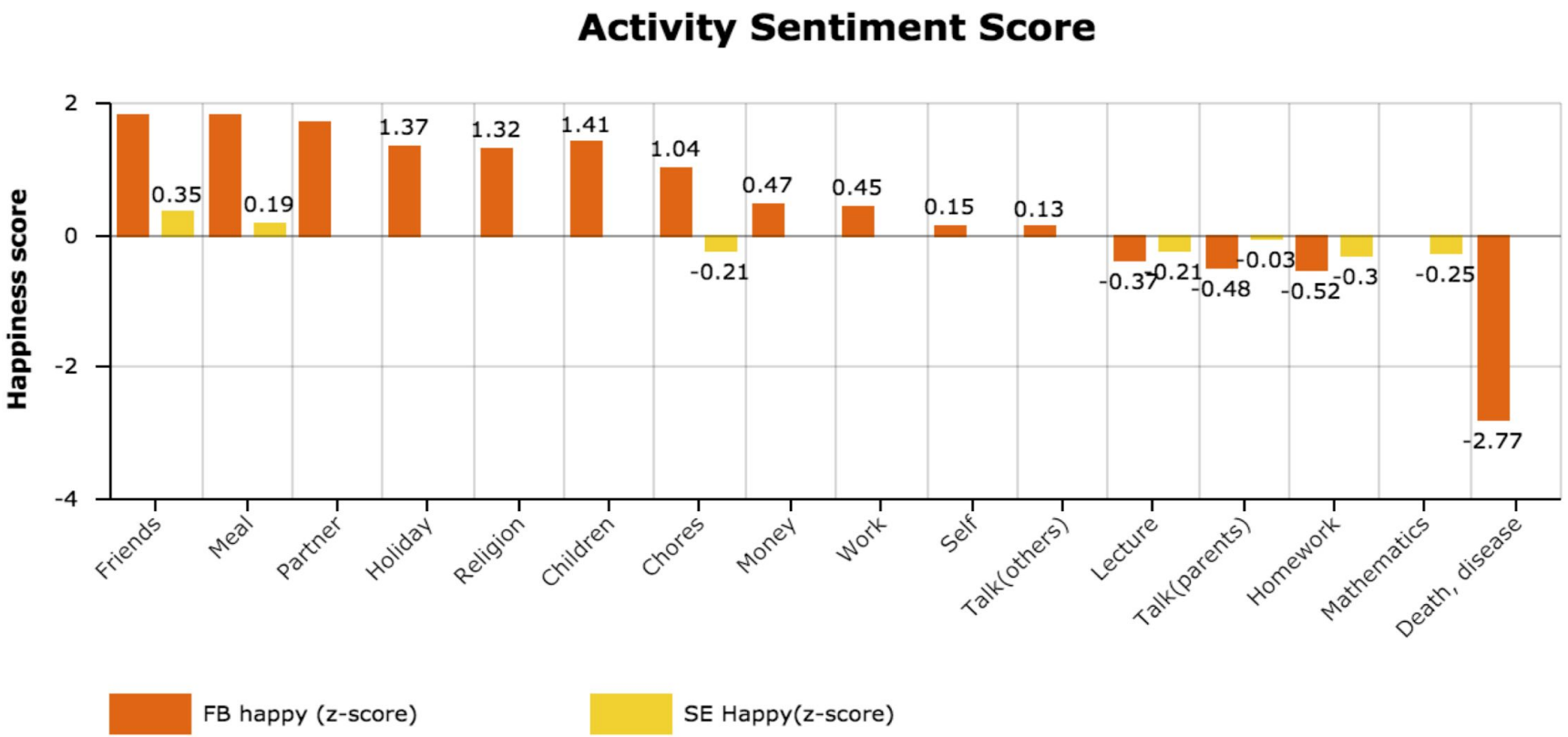

Fig.1 Human activities indicating users' affect scores largely resemble those reported in the previous experimental study

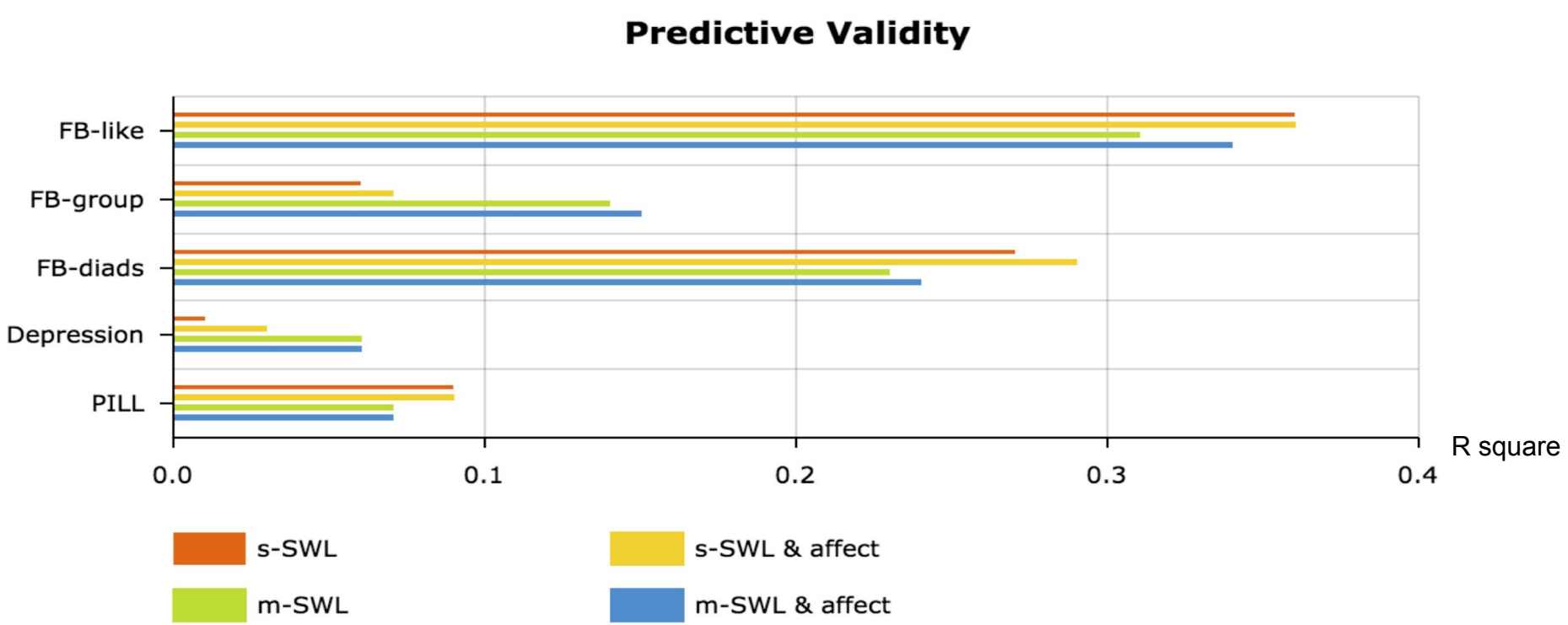

Fig 2. Hierarchical regression model show that subjective well-being profile can predict some of the user's' life outcomes better than or nearly as well as self-reported SWL alone.

S- SWL: self-reported Satisfaction with life

M-swl: machine predicted satisfaction with life

Future

- Combine social media data from different accounts

  http://www.hatdex.org/

  cloud-based information system, your own microserver

- Machine predict psychological profile
- Conduct mega scale study with computer predicted variables (e.g. depression symptoms prediction study with Weibo data)

Reference

1. G. Park, H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, M. Kosinski, D. Stillwell, L.H. Un- gar, and M.E. Seligman, *Automatic personality assessment through social media language*, J. Person. Soc. Psychol., 108 (2015), pp. 934–952.
2. Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*,*8*(5), e64417.
3. T. Hills, E. Proto, and D. Sgroi. (2015). *Historical analysis of national subjective wellbeing using millions of digitizsed books*, IZA Disscussion Paper No. 9195.
4. W. Youyou, M. Kosinski, and D. Stillwell, *Computer-based personality judgments are more accurate than those made by humans*, Proc. Natl. Acad. Sci USA, 112 (2015), pp. 1036–1040.