Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Understanding Cognitive Styles from User-Generated Social Media Content

Yi Wang, Jalal Mahmud, Taikun Liu
IBM Research-Almaden, 650 Harry Road, San Jose, CA 95120, USA
{wangyi, jumahmud, tliu}@us.ibm.com

Abstract

The linguistic analyses on easily accessible, user-generated social media content offers great opportunities to identify individual characteristics such as their cognitive styles. In this paper, We explore the potential to use social media content to identify individuals' cognitive styles. We first employed crowdsourcing to collect Twitter users' cognitive styles using standard psychometric instruments. Then, we extracted the linguistic features of their social media postings. Leveraging these features, we build prediction models that provide estimates of cognitive styles through statistical regression and classification. We find that user generated content in social media provide useful information for characterizing people's cognitive styles. The models' performance indicates that the cognitive styles automatically inferred from social media are good proxies for the ground truth, and hence provides a promising and scalable way to automatically identify a large number of people's cognitive styles without reaching them individually.

Introduction

Cognitive style refers an individual's preferred and habitual approach to organizing and representing information (Riding and Rayner 2013). Cognitive style is an important personality dimension relevant to people's information processing and decision making. Moreover, it is also relevant to organizational behaviors because individuals of different cognitive styles are continuously interacting with each other in their organizations (Chan 1996). Identifying individuals' cognitive styles has great potentials in various domains such as personalized education (Deborah, Baskaran, and Kannan 2014), marketing (Joseph and Vyas 1984), management (Armstrong, Cools, and Sadler-Smith 2012), and so on.

The conventional method to identify a person's cognitive style is using standard psychometric inventories (Kozhevnikov 2007). These methods are not scalable for they require to reach subjects individually. It is very hard and costly to contact a large number of users to learn their cognitive styles through questionnaire surveys. To gain the benefit of understanding individual's cognitive styles at a large scale, we need to find an approach to automatically identify the cognitive styles without reaching each individual painstakingly.

The popularity of social media has led a huge amount of online user-generated content to be easily accessible. The penetration of social media into individuals' everyday activities provides unprecedented opportunities for researchers to identify individual characteristics from the linguistic and social features of the user-generated data, e.g., Chen et al. (2014) and Yarkoni (2010). It is reasonable to assume that cognitive styles could be automatically inferred from social media content. Indeed, there is preliminary evidence such as (Pennebaker, Slatcher, and Chung 2005) that claimed the feasibility of inferring cognitive styles from the text. However, they have not presented the precise quantitative relationships between cognitive styles and linguistic features.

In this work, we present the first analysis of associations between people's cognitive styles and their word use in social media. We recruited active twitter users through crowdsourcing, and measured their cognitive styles with two established psychometric models of cognitive styles. We collected their tweets, and measured the word use in a number of word categories as defined by the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker, Francis, and Booth 2001). Then, we built regression and classification models to predict the cognitive styles from LIWC features. The models achieve high accuracy in inferring people's cognitive styles from the their social media text.

Our main contributions in this paper are as follows:

- We use crowdsourcing to collect (ground truth) cognitive style measurements from several hundred Twitter users through standard cognitive styles surveys.
- We build both regression and classification models to predict people's cognitive styles from their tweets. We also investigate to what extent people's cognitive styles can be predicted from textual information in social media.

Cognitive Styles

Researchers have proposed dozens of models to conceptualize and measure the cognitive styles (Kozhevnikov 2007). Researchers have concluded that most of these various models are merely divergent conceptions of a superior-order dimension, the poles of which are commonly associated with the specialist functions performed by each hemisphere of

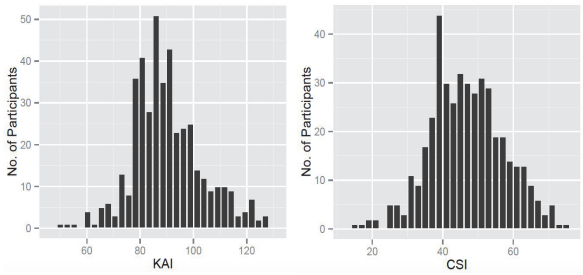

Figure 1: The distribution of ground truth KAI and CSI.

the human brain (Rayner and Riding 1997). In this study, we select two cognitive styles models: Kirton's Adaptor-Innovator model (KAI) (Kirton 1976) and Cognitive Styles Index (CSI) (Allinson and Hayes 1996). Both of them are very stable measurements and independent to people's cognitive ability. The reliability has been validated by hundreds of studies in various areas. Moreover, measuring KAI and CSI does not require the direct interactions between researchers and subjects, which means we can reach a very diverse subjects through online surveys without real-time experiment administration.

KAI can be measured by a standard questionnaire that contains 32 statements such as "*A person who needs the stimulation of frequent change*." A subject is asked to rate how hard (in 5-point Likert scale) to imagine herself to fit each statement. The score of KAI ranges from [32, 160]. The higher it is, the more innovative a person tends to be. CSI is measured by a 38-item standard questionnaire. A typical item is: "I find that to adopt a careful, analytical approach to making decisions takes too long." A subject will rate whether an item describe him well in terms of *false*, *neutral*, and *true*. The raw value of CSI ranges from [0, 76]. The higher it is, the more analytic an individual is.

Data

We leveraged Amazon Mechanical Turk (MTurk) to recruit the participants for this study. MTurk workers who are active twitter user and have more than 200 original tweets in English are eligible to take our study. We sought responses from those who were located in the United States, and had an approval rating on Mturk greater than or equal to 80%. Each Mturk worker was restricted to take the HIT exactly once for $1.00 compensation. We took several strategies to exclude noisy responses and minimize bias. These measures include randomizing survey order, adding attention check questions, setting mandatory minimal time, manually checking after the data collection, and so on.

A total of 529 workers completed our HITs. Using the data cleaning strategies we mentioned above, we have 432 well-cleaned data points. Figure 1 shows the distribution of the two dependent variables: KAI ($mean$: 90.143, $std.$: 13.097) and CSI ($mean$: 46.247, $std.$: 10.539). They are also highly correlated $\beta = -0.473$, $P < 0.001$, which indicates that they are likely to measure the same superordinate concepts (Rayner and Riding 1997). Overall, the distributions of ground truth data well fit the theoretical and prior empirical results of both measurements (Kirton 1976; Allinson and Hayes 2012), which indicates the threats to the validity of using MTurk workers as the data sources should be minimal. The number of participants' original tweets ranges from 252 to 3200, with the average of 1197.63. For each of them, the word count of all original tweets ranges from 1573 to 74098, with the average of 13982.60 and the standard deviation of 5391.27.

Predicting Cognitive Styles

In general, prediction of peoples' cognitive styles can be performed by both regression and classification techniques. In this study, we employed both of them:

- Regression, where the goal is to predict an individuals score for the two measurements of cognitive styles using the linguistic features of their Twitter content.
- Classification, where the goal is to identify individuals with particularly high or low values of a specific cognitive styles according to some predetermined cut-off.

Since both KAI and CSI are continuum measurement and normally distributed, using linear regression is pretty straightforward to develop regression models. We used the stepwise method to select models according to the Akaike Information Criterion (AIC). 10-fold cross validation was adopted in building regression models. For classification, we use the medians of raw KAI and CSI values as cut-off values. Through this way, we created a balanced 50-50 data set for each of them. For KAI, we named the two classes as "Adaptive" and "Innovative". For CSI, the two classes was named as "Intuitive" and "Analytic". We used supervised learning to construct classifiers trained to predict cognitive styles regarding both KAI and CSI. Then, the classifiers were built with a set of different classification algorithms which are all implemented in the "off-the-shelf" WEKA machine learning toolkit (Hall et al. 2009). For our research does not aim to build the best prediction systems for cognitive styles. Rather, we want to demonstrate the feasibility of building cognitive styles prediction systems that have potential to be scalable to millions of users. Hence, all classifiers were developed with the default parameter settings of WEKA.

We use LIWC dictionary to extract the features of the collected tweets. LIWC can automatically detect the links between the words and the psychology-relevant categories. LIWC 2001 (Pennebaker, Francis, and Booth 2001) contains 74 categories. The first 6 categories (e.g., total word count, words per sentence) belong to standard word count information that are not associated with any dictionary. We tested the control models including them, none of them is significant. Hence, we excluded them from our analyses. The rest 68 word categories represent various types of words in linguistic and in psychological processes. We excluded last two categories (*Non-fluencies* and *Fillers*) for the majority values of them are "0" in our dataset.

The use of LIWC has both theoretical and empirical considerations. For the ground truth of cognitive styles is collected from questionnaires that are quite different from social media context, we seek to build models that are highly

context-independent and robust across various content domains. Hence, using well-developed, context-independent and simple LIWC features could be a reasonable choices (Tausczik and Pennebaker 2010). Using LIWC features are also supported by empirical results of our data set. In our dataset, there are statistically significant differences for some LIWC features' values over the two classes for both KAI and CSI measurements even after Bonferroni correction.

Results

Regression Results

Table 1 presents the regression results. Without Bonferroni correction, the KAI model contains 28 explanatory variables while the CSI model has 34 explanatory variables. The models have good prediction potential. The *adjusted R-squared* of the models are 0.402 (KAI) and 0.397 (CSI) respectively. The correlations between the ground-truth value and predicted value are 0.662 (KAI) and 0.639 (CSI), indicating a good accuracy of the models' predictions. The Mean Absolute Errors (MAEs) are 11.025 and 8.372, which are less than 20% of the average ground truth values. Again, the predicted values are at a promising level of accuracy. . For both models, the highest VIFs are less than 10, which suggests the multicollinearnality should not be a big concern.

Because we have a relative large number of feature variables; hence to counteract the problem of multiple comparisons, we also adopted Bonferroni correction. Since we included 66 LIWC features, the significant level was adjusted to 0.000758. In table 1, the "†" symbol after the original significant level (indicated by "$*$" symbols) indicates that the feature is significant after Bonferroni correction. The Bonferroni correction to the original regression models results a smaller set of significant features. Using these features only, we ran regressions again, the *adjusted R-squared* of the models are 0.338 (KAI) and 0.275 (CSI). These results show that the high *adjusted R-squared* may not results over-fitting since the models using much fewer features also achieves acceptable goodness-of-fit. A possible explanation is that cognitive styles are relatively simple psychological constructs compared with general personality constructs, e.g., Big-Five. Both KAI and CSI are one-dimensional, polarized measurements. Therefore, they should be predicted at higher accuracy levels than personality models whose construct structures are complex. Note that the Bonferroni correction is very conservative. It's high confidence comes at the cost of increasing the probability of producing false negatives, and consequently reducing statistical power. For practical consideration, we report original models in table 1 (Frane 2015).

	Predictive Models	
LIWC Features	KAI (β)	CSI (β)
Intercept	111.20***	36.69***
Linguistic Process		
1st person plural	−1292.54***†	936.11**
Assents	−1719.13*	–
Negations	–	−562.60**
Time	171.32	–
Present Tense	114.07	−181.52**
Future Tense	−343.92	−283.24
Affective Process		
Affect	–	−697.69*
Optimism	–	686.32
Negative affect	322.95	679.22**
Anger	–	654.54
Cognitive Process		
Cognition	−634.94***†	550.64***†
Causation	−742.22*	1058.69***†
Discrepancy	–	413.85
Inhibition	–	−1837.05
Tentativeness	−222.75.	268.71*
Sensation/Perception	594.11***†	–
Certainty	681.04*	–
Perceptual processes		
Hearing	–	1705.31***†
Touching	−1868.00.	1849.22*
Social Process		
Human	−1287.86***†	–
Communication	–	−1325.13***†
Friends	–	−2295.24***†
Relativity		
Space	−391.85***†	148.08
Inclusion	−224.99*	112.82
Personal Concerns		
Occupation	5918.15**	−5868.17***†
School	−5900.80**	5941.68***†
Job	−5286.74**	5051.28***
Achievement	−5837.59***†	6062.18***†
Leisure	−4364.52**	3727.14**
Home	4134.31**	−3796.60**
Sports	3687.01*	−3746.21**
TV, Movie	3410.61*	−2440.38*
Music	4632.38***†	−4185.23**
Money	–	4699.06
Death	1059.08**	−6110.84
Physical states	929.02***†	584.62**
Symptoms sensations	1040.30*	−1059.09***†
Eating	−1857.02***†	1230.77**
Swearing	−802.11	1230.76**
Adjusted R-Squared	0.402	0.397
F-Statistic	11.36***	8.65***
Correlation	0.662***	0.639***
MAE	11.025	8.372

Note 1. .: $p < 0.10$, $*$: $p < 0.05$, $**$: $p < 0.01$. $***$: $p < 0.001$. **Note 2**. We omit the LIWC measures that are excluded from both models.

Table 1: Linear regression models for KAI and CSI.

Classification Results

Now we utilize our proposed classification framework to examine how well we can predict an individual's cognitive styles based on both KAI and CSI measures. As described in the prior section, we compared several different binary classifiers to empirically determine the best suitable classification technique. Given that the dataset is balanced, the baseline classification performances are 50% (by chance) on accuracy. The best performing classifier for KAI (adaptive vs. innovative) was found to be the SMO solver of SVM, while Logistic Regression yields the best performance for

CSI (analytic vs. intuitive).

Table 2 shows the best performance of classification models. We can conclude that in this classification task, the classifiers offered a real improvement over random chance.

	Acc.	Pre.	Rec.	AUC
Adaptive/Innovative	0.705	0.705	0.685	0.702
Analytic/Intuitive	0.649	0.649	0.606	0.689

Table 2: Classification performance of cognitive styles based on KAI and CSI.

We examined the importance of the features in classifications through information gain feature evaluation. For "Adaptive vs. Innovative", 12 features have significant influences on the classification. The top five are:*Cognition*, *Causation*, *Present Tense*, *Inclusion*, and *Tentativeness*. For "Analytic vs. Intuitive", 4 features have significant influences on the classification. They are *Cognition*, *Causation*, *Communication*, and *Prepositions*. As our expectation and LIWC's theoretical explanation, *Cognition* and *Causation* are two most important features, which is also consistent with the regression results.

Overall, both the regression and classification suggest that word use on twitter indeed contains predictive information of people's values, and can potentially be used to classify people based on their cognitive styles. The prediction of an individual's cognitive styles in the binary classification setting achieves promising accuracy. The benefit of using our methods in practices will be significant when using it to analyze the cognitive styles of a large sample of individuals.

Conclusion and Future Work

In this paper, we have demonstrated the potential of using social media (Twitter in this paper) in measuring and predicting people's cognitive styles. We developed prediction models for two cognitive measurements (KAI and CSI). The models provide estimates of cognitive styles through statistical regression and classification. The regression models achieve relatively good accuracy and the classifiers yielded promising results with around 65%-70% classification accuracy. There are many promising future directions. On the theoretical side, it would be valuable to explore how cognitive styles influence the word use. Future work is needed to validate these interpretations empirically. Researchers may also extend our work of word use with other cognitive styles measurements. On the practical side, it is promising to use the models to automatically predict a large number of people's cognitive styles. Hence, the immediate application is applying predicted cognitive styles in existing application scenarios in various areas such as education, marketing, management. We also plan to employ more sophisticated modeling approaches to improve the accuracy of the predictions. Since the user-generated content is not restricted to textual information, it is a promising direction to utilize multimedia data such as user-shared pictures to identify cognitive styles.

References

Allinson, C., and Hayes, J. 1996. The cognitive style index. *Journal of Management studies* 33(1):119–135.

Allinson, C., and Hayes, J. 2012. *The Cognitive Style Index-Technical Manual and User Guide*. Pearson Education.

Armstrong, S.; Cools, E.; and Sadler-Smith, E. 2012. Role of cognitive styles in business and management: Reviewing 40 years of research. *International Journal of Management Reviews* 14(3):238–262.

Chan, D. 1996. Cognitive misfit of problem-solving style at work: A facet of person-organization fit. *Organizational Behavior and Human Decision Processes* 68(3):194–207.

Chen, J.; Hsieh, G.; Mahmud, J. U.; and Nichols, J. 2014. Understanding individuals' personal values from social media word use. In *Proc. CSCW*, 405–414. ACM.

Deborah, L. J.; Baskaran, R.; and Kannan, A. 2014. Learning styles assessment and theoretical origin in an e-learning scenario: a survey. *Artificial Intelligence Review* 42(4):801–819.

Frane, A. V. 2015. Are per-family type i error rates relevant in social and behavioral science? *Journal of Modern Applied Statistical Methods* 14(1):5.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 11(1):10–18.

Joseph, B., and Vyas, S. 1984. Concurrent validity of a measure of innovative cognitive style. *Journal of the Academy of Marketing Science* 12(1-2):159–175.

Kirton, M. 1976. Adaptors and innovators: A description and measure. *Journal of applied psychology* 61(5):622.

Kozhevnikov, M. 2007. Cognitive styles in the context of modern psychology: toward an integrated framework of cognitive style. *Psychological bulletin* 133(3):464.

Pennebaker, J.; Francis, M.; and Booth, R. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71:2001.

Pennebaker, J.; Slatcher, R.; and Chung, C. 2005. Linguistic markers of psychological state through media interviews: John kerry and john edwards in 2004, al gore in 2000. *Analyses of Social Issues and Public Policy* 5(1):197–204.

Rayner, S., and Riding, R. 1997. Towards a categorisation of cognitive styles and learning styles. *Educational psychology* 17(1-2):5–27.

Riding, R., and Rayner, S. 2013. *Cognitive styles and learning strategies: Understanding style differences in learning and behavior*. Routledge.

Tausczik, Y., and Pennebaker, J. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54.

Yarkoni, T. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality* 44(3):363–373.