Online Social Capital: Mood, Topical and Psycholinguistic Analysis

Thin Nguyen, Bo Dao, Dinh Phung, Svetha Venkatesh and Michael Berk
Deakin University
Geelong, VIC 3220, Australia
{thin.nguyen,dbdao,dinh.phung,svetha.venkatesh}@deakin.edu.au, mikebe@barwonhealth.org.au

Abstract

Social media provides rich sources of personal information and community interaction which can be linked to aspect of mental health. In this paper we investigate manifest properties of textual messages, including latent topics, psycholinguistic features, and authors' mood, of a large corpus of blog posts, to analyze the aspect of social capital in social media communities. Using data collected from Live Journal, we find that bloggers with lower social capital have fewer positive moods and more negative moods than those with higher social capital. It is also found that people with low social capital have more random mood swings over time than the people with high social capital. Significant differences are found between low and high social capital groups when characterized by a set of latent topics and psycholinguistic features derived from blogposts, suggesting discriminative features, proved to be useful for classification tasks. Good prediction is achieved when classifying among social capital groups using topic and linguistic features, with linguistic features are found to have greater predictive power than latent topics. The significance of our work lies in the importance of online social capital to potential construction of automatic healthcare monitoring systems. We further establish the link between mood and social capital in online communities, suggesting the foundation of new systems to monitor online mental well-being.

Introduction

The concept of social capital was introduced in sociology (Bourdieu 1986; Coleman 1988; Putnam 2000). Like other forms of capitals, social capital can be defined by the capacity of facilitating productive activity. In this view, physical capital and human capital can be determined through the value of tools/machines or skills/capabilities respectively (Coleman 1988); meanwhile, social capital is valued by the resources associated with interpersonal relations that facilitate actions (Coleman 1988). For example, (Coleman 1988) noted that a group with trustworthiness, a measure of social capital, has higher chance to accomplish than another without that. Similarly, (Bourdieu 1986) defined social capital as the actual or potential resources a network of relationship possesses. (Putnam 2000) further divided social capital into bonding and bridging based on whether emotional support is provided (bonding), e.g., between family members and close friends (strong ties), or not (bridging), e.g., among acquaintances (weak ties). Generally, social capital includes social inclusion, social participation and social support (VicHealth 2010). Social inclusion implies access - for example, older people may have less physical access to social communities. Social participation describes how much a person engages with a community. And, social support implies how much support one can get from the community.

The theory of social capital has been spread into several fields, including economics, education and health. For example, according to (Putnam 2001), joining a community might increase one's social capital and subsequently reduce his chance of fatality in the next year by a half; (Coleman 1988) found that the value of social capital students gained from home and outside is positively correlated with the probability of dropping out of high school.

In the age of the Internet, social media genres, such as blogs, social networking sites and forums, have become a popular venue for people to gather, emerging a new form of social capital – digital social capital. This concept has attracted considerable attention, such as investigations into the impact of the new media for the production of social capital (Beaudoin and Tao 2007; Shah, Kwak, and Holbert 2001; Wellman et al. 2001). Insights from such analysis have wide applications, from sociology where the web is viewed as a very large community sensor, to business where e-commerce is now the norm.

This study examines a large corpus of blog posts to analyze the effect of online social capital a user potentially possess on mood, topics and psycholinguistic styles conveyed in the messages he creates. Continuing with the core components of social participation and support, we define two categories of LOW and HIGH social capital. We have 12 cohorts to study - 2 categories of social capital (LOW and HIGH), and in each, 3 measures of social participation (number of communities joined, number of posts written and number of comments made) and 3 of social support (number of friends, number of comments received and number of followers). Using posts made by 60,000 users in Live Journal over 10 years, we explore across all measures of social participation and social support: (a) distribution of mood and mood swings over LOW and HIGH social capital; (b) distribution of latent topics of interest in LOW and HIGH social capital; and (c) distribution of psycholinguistic features expressed in posts made by users in LOW and HIGH social capital.

We assume that there exists a link between online capital derived from social media and mood, latent topics of discussion and language styles. To test the assumption that bloggers with different degrees of social capital exhibit different moods, we will predict the mood of a post using three methods: a) the prior mood transition matrix of the cohort (baseline); b) naive Bayes classifier learned using the post content c) a combination of post content and mood swing. For the hypotheses that bloggers with different degrees of social capital have different interest on topics and writing styles, significant differences will be detected using statistical tests. In addition, the predictive power of the psycholinguistic processes and content topics extracted from the posts will be discovered through classifying blog posts as made by either low or high social capital groups.

Our contribution lies in producing the first work to formulate computable online social capital using social connectivity. Secondly, unlike previous work we view online social capital as a sensor to understand the users in online communities. Besides, to our knowledge, it is the first time the connection between social capital in online social networks and the aspects that differentiate groups of social capital, including mood, topics of interest, and language styles, is formally established.

The results presented here have several applications in areas such as personalized information retrieval, which relies on knowledge of user attributes such as low or high in social capital to re-rank results, and online advertising, where one might make use of estimated user profile attributes to further target advertisements based on the degree of connectivity. Especially, the significance of this work lies in using social media as a barometer for mood. This is critical. With such recognition, medical evaluations and screenings could routinely include variables of social well-being; medical care could recommend if not outright promote enhanced social connections; hospitals and clinics could involve patient support networks in implementing and monitoring treatment regimens and compliance, etc. Health care policies and public health initiatives could likewise benefit from explicitly accounting for social factors in efforts aimed at reducing mortality risk. Individuals do not exist in isolation; social factors influence individuals' health though cognitive, affective and behavioral pathways.

Background

Previous studies have shown associations between social capital and health. For *mental health*, adolescents who have lower social isolation and better peer relationships have lower levels of depression and anxiety (Bosacki et al. 2007). (Rotenberg, Boulton, and Fox 2005) measured social capital using children's trust beliefs in peers. They found that children with extreme levels of the social capital, both very high and very low, were associated with worse mental health. With regards to *health promoting* behaviors, (Wang et al. 2011) found that adolescents with better family communication had better behaviors of health responsibility and physical activities. For *health risk* behaviors, (Morgan and Haglund 2009) found that adolescents from less wealthy families were more likely to eat less fruit and vegetables and engage in less physical activity. (Fulkerson et al. 2006) reported that the rate of family dinners was negatively correlated with the frequency of high risk behaviors, such as substance use. The frequency of religious service attendance has been found to be negatively associated with the odd of binge drinking (Rasic, Kisely, and Langille 2011). For *well-being*, (Zambon et al. 2010) found that young people reporting that they were members of recreational clubs had better satisfaction with life.

The link between the use of Internet and social capital has also been explored. (Wellman et al. 2001) found that Internet could help to increase two forms of social capital: (1) *network capital* – interpersonal connectivity and (2) *participatory capital* – organizational involvement. Similarly, (Shah, Kwak, and Holbert 2001) found that overall Internet use was positively related to social capital, in term of civic engagement and interpersonal trust. However, this is not true to different types of Internet usage. Specifically, the production of social capital was negatively related with the use of the Internet for recreation, e.g., chat rooms or games, but positively related with the use of the Internet for information exchange, such as searching or emailing. This finding is in part in accordance with the result reported in (Beaudoin and Tao 2007) where asynchronous online communication, including e-mail and discussion groups, developed social support among cancer patients, resulting in positive health outcomes.

Social networking sites, such as Twitter, MySpace and Facebook, have become popular media for online communication. These sites allow people freely to participate, mediate their own content and interact with others. As such, social capital is potentially produced in online settings, namely digital social capital (Mandarano, Meenar, and Steins 2010). (Steinfield, Ellison, and Lampe 2008) found that Facebook increased bridging social capital by building and maintaining weak ties among groups of distant friends and acquaintances. For Twitter, (Ye et al. 2012) found that social capital could be transferred from real world to virtual one. On the other way, virtual communities was also found to help increase offline social capital (Kobayashi, Ikeda, and Miyata 2006). To our knowledge, the most generic computational model of social capital in online social networks is the work of (Kazienko and Musiał 2006) which defines social capital for each user as a linear combination of the following functional components of the user: *static component* (does not change with time, e.g., derived his or her profile), *Marched-by-Search* (reflecting his/her openness to new acquaintances), *Activity* (characterizing online activities such as frequency of posting or commenting) and *Social Position* (describing the value of the user in the network derived from social relationships).

Social Capital and Mood

Dataset

We collected data from the Live Journal website. The initial dataset consists of 18,774,223 posts made by 1,616,625

Category	LOW range	HIGH range
#friends	2-3	110-163
#communities	1-1	41-87
#followers	1-1	53-87
#posts written per year	8-12	209-285
#comments made per year	3-6	765-1307
#comments received per year	3-6	615-1038

Table 1: Statistics of the 12 cohorts.

Category	LOW	HIGH
#friends	66,417	82,059
#communities	70,616	78,364
#followers	68,421	79,064
#posts written	60,674	89,224
#comments made	65,581	81,296
#comments received	66,055	80,804

Table 2: The number of posts made by users in each social capital groups across six connectivity categories.

users.

We examine social capital defined based on the degree of social participation (connectivity) and social support (content generation engagement). For connectivity, Live Journal supports one directed person-to-person link type. For a given user, incoming links are followers, and outgoing links are friends. Live Journal also allows users to join communities that discuss topics of interest. Thus we have three indicators of social capital associated with social participation (the degree of networking): *number of friends*, *number of community memberships*, and *number of followers*. From our dataset, on average, each blogger has 10 followers and 23 friends, and joins 6 communities. For social support (the degree of content-generation engagement), we base on three indicators: *number of posts* written, *number of comments made*, and *number of comments received*.

For each indicator, two social capital subsets are defined: LOW (containing 5000 people whose the indicator above the cut-off of 2.5 percentile of the corpus), and HIGH (contains 5000 members whose the indicator below the cut-off of the 97.5 percentile of the corpus). This process results in a total of 12 corpora: two social capital groups for each of the six social capital categories. Table 1 shows the statistics obtained within each social capital group. There are 60,000 users, 10,000 in each of the social capital categories being considered. The number of posts made by these users are shown in Table 2.

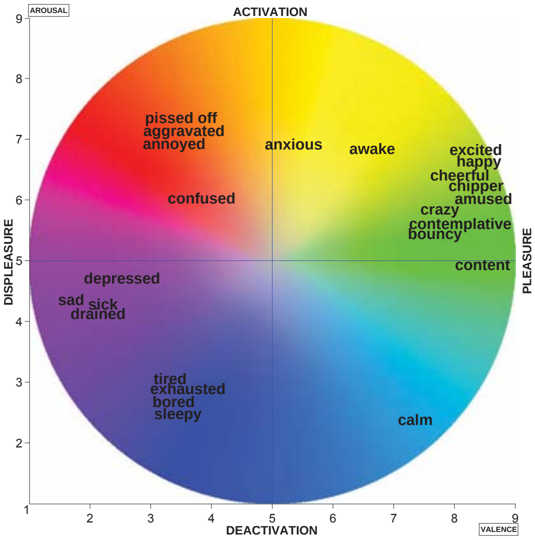


Figure 2: Visualization of 24 primary moods extracted from the data on the core affect circle of emotion structure using valence and arousal proposed in (Russell 2009).

Mood usage

All posts in the corpus are tagged with one of 132 predefined moods, which can be categorized into emotional patterns (Nguyen 2010). Figure 1 shows the histogram of these moods. The exponential decay of the histogram suggests that there is a small set of moods that dominantly represent typical discriminative type of emotions across social capital groups. We therefore extract the moods that constitute 50% of the population by placing a threshold over the histogram and lowering it until more than half of the posts are accounted for, yielding a set of 24 moods, termed the *primary mood set*. Using dimensional representation of the circumplex model of affect for emotion structure proposed in (Russell 2009), this mood set is visualized to show valence and arousal in Figure 2 with the sentiment values are derived from the Affective Norms for English Words (ANEW) (Bradley and Lang 1999). It can be seen that these moods are reasonably diverse, including positive (right hand side) and negative (left hand side) sides of the affect circle.

To empirically inspect the usage of primary moods across social capital groups, Figure 3 shows the difference between low and high social capital groups in valence and arousal of the moods. It is clear that valence and arousal are greater in the HIGH groups vs the LOW groups. It is because cohorts with lower social capital when compared to cohorts higher social capital have more negative moods and less positive moods.

Mood prediction

This section examines the prediction of mood for individuals in a cohort. We consider 24-class (all primary moods) and 4-class, corresponding to the 4 quadrants of the affect circle, predictions. Three following methods of classification are used.

Prediction using mood swing matrix: M1 From training dataset, we derived a mood transition matrix $\mathbf{T}_{K\times K}$, where $T^{ij} \triangleq Pr\{y_t = j \mid y_{t-1} = i\}$, i.e. the probability that the i^{th} mood moves to the j^{th} mood. For a post coming at time t, we aim to predict its class label k^* based on the mood at time $t-1$ the user experienced, say the l^{th} mood. Then

$$k^* = \text{argmax}_k T^{lk} \qquad (1)$$

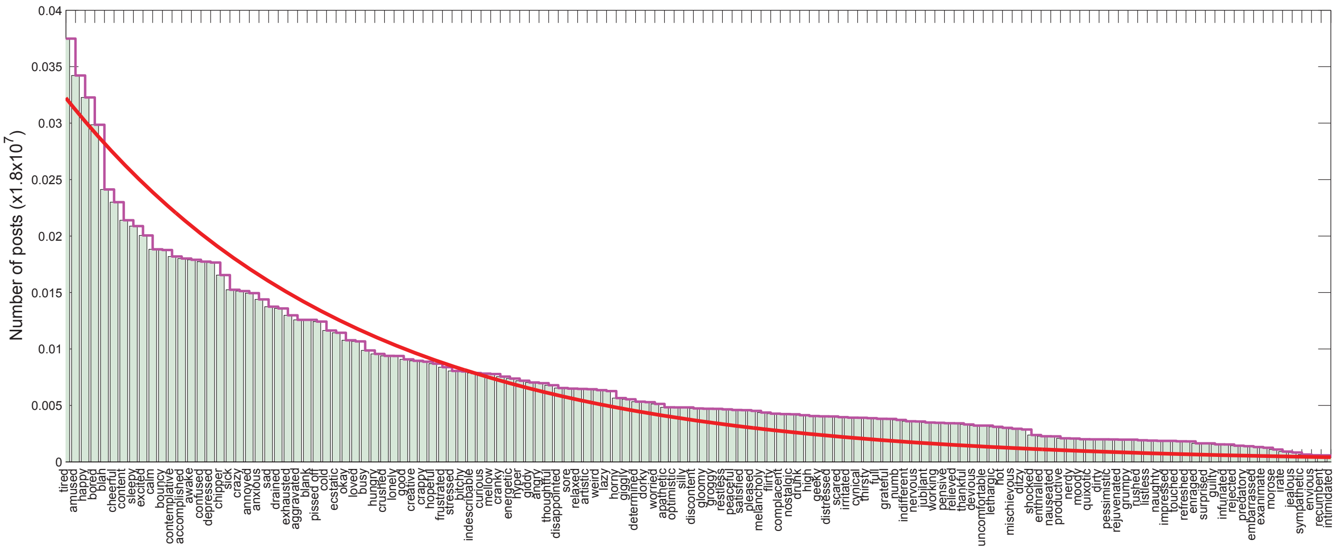
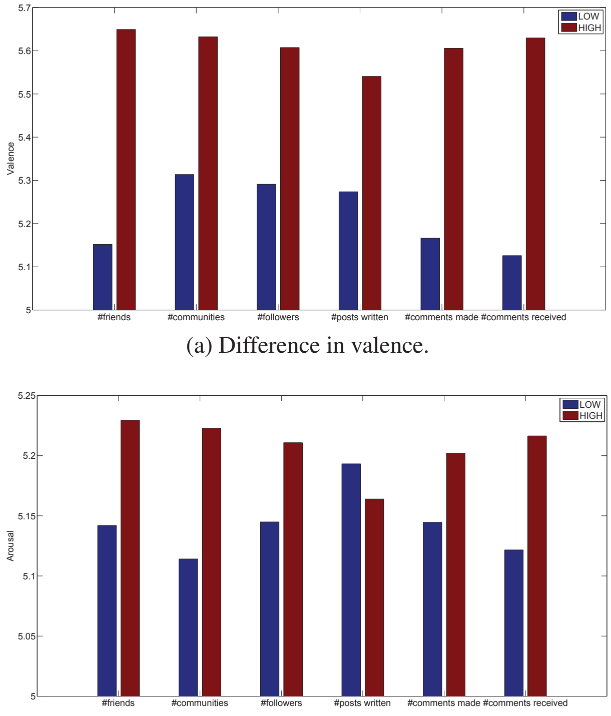


Figure 1: Histogram of moods tagged in a corpus of 18,774,223 posts used in this paper. There is a strong pattern of exponential decay, suggesting the existence of a small subset of primary moods that dominantly explain the data (best viewed in color; pink curve: staircase plot of the true data, red curve: exponential distribution fitting with the mean = 30 after re-numbering the mood label from 1 to 132).



(a) Difference in valence.

(b) Difference in arousal.

Figure 3: Difference between LOW-HIGH in valence and arousal.

This method provides a baseline for the mood prediction.

Prediction using post content (Naive Bayes): M2 Denote by $\mathcal{M}$= {sad, happy, ...} the set of all mood categories and by K the number of moods ($K = |\mathcal{M}|$). Let $\mathcal{D}$ be the training dataset consisting of n data points

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^m, y_i \in \{1..K\}\}_{i=1}^{n}$$

where $\mathbf{x}_i := \{x_i^1, x_i^2, ..., x_i^m\}$ denotes the data vector in the feature space, y_i denotes its class label, and m denotes the number of features chosen to characterize a blog post. The objective of a NB classification (Lewis 1998) is to assign a class label k^* to an unseen post $\mathbf{x}_t$ **:**

$$k^* = \text{argmax}_k Pr(y_t = k|\mathbf{x}_t, \mathcal{D}) \tag{2}$$

where

$$Pr(y_t = k|\mathbf{x}_t, \mathcal{D}) \propto Pr(y_t = k|\mathcal{D}) Pr(\mathbf{x}_t|y_t = k) \tag{3}$$

Prediction using combination of post content and mood swing: M3 Combining the content of the current post with the previous post mood may provide a better prediction of the current mood. When doing that, the posterior for the current mood can be written as

$$Pr\left(y_t = k \mid x_t, y_{t-1}\right) \propto Pr\left(y_t = k \mid y_{t-1}\right) Pr\left(x_t \mid y_t = k\right)$$

and the class label (k^*) for the current post is given by

$$k^* = \text{argmax}_k Pr\left(y_t = k \mid x_t, y_{t-1}\right) \tag{4}$$

When the post content is used in the mood predictions, three feature sets are employed: ANEW (Bradley and Lang 1999), LIWC (Pennebaker et al. 2007), and a lexicon derived

	LIWC	ANEW	IG
24-class	12.87	13.889	17.524
4-class	43.02	46.329	49.457

Table 3: Classification results for feature selection

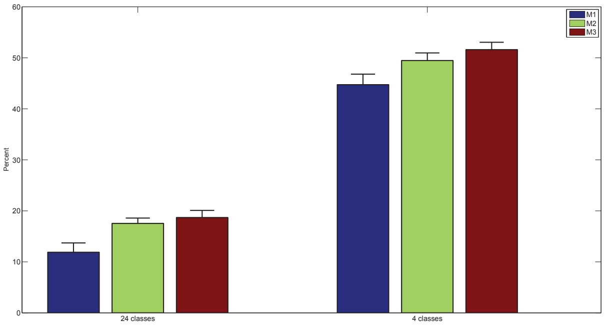


Figure 4: Prediction accuracy (IG features are used in M2 and M3).

by Information Gain (IG) (Sebastiani 2002). While IG was reported as one of the best methods for feature selections (Nguyen et al. 2010), the rest two are chosen because of lower computational costs.

Mood prediction performance Table 3 shows the results with respect to feature selection schemes - IG is observed to be the best feature set for the mood prediction, with an accuracy of 17.5 per cent in 24-class, and 49.5 per cent in 4-class prediction, exceeding that of LIWC and ANEW. We thus choose IG when considering post content.

Figure 4 compares the three prediction methods for 24-class and 4-class mood predictions. Method 3 outperforms all methods, but surprisingly, just using the transition matrix(method 1) does very well. What is means is that the post content is not very discriminative, and mood transition play a much more important role in determining mood.

Table 4 shows the accuracy for low and high social groups across all methods and all mood predictions. The accuracy for HIGH is better in all cases.

Topics and Language Styles

For each cohort, we randomly collect 5,000 posts for LOW social capital people and another 5,000 posts for HIGH people. We performed analysis of two aspects: psycholinguistic processes and topics derived from these posts. For psycholinguistic processes, we use the the LIWC package (Pennebaker et al. 2007). LIWC assigns words into one of four high-level categories – *linguistic* processes, *psychological* processes, *personal concerns* and *spoken* categories – which are further sub-divided into a three-level hierarchy. The taxonomy ranges across topics (e.g., religion and health), emotional response (e.g., positive emotion) and processes not captured by either, such as cognition (e.g., causation and discrepancy).[1]

		HIGH	LOW
Mood swing	24-class	14.228	10.445
	4-class	47.343	42.85
IG	24-class	18.635	17.081
	4-class	51.402	48.106
ANEW	24-class	14.073	13.633
	4-class	47.745	45.355
LIWC	24-class	13.58	12.592
	4-class	44.995	42.071

Table 4: Accuracy in mood prediction for LOW and HIGH.

For topics, we employ a popular Bayesian probabilistic modeling tool – the latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) – to extract the topics from LOW and HIGH groups. LDA is a hierarchical Bayesian mixture models that assigns each word in a document to a topic, where a topic is a distribution over the set of distinct vocabularies found in all documents. Thus, the content of a topic can be roughly interpreted by inspecting the subset of vocabularies corresponding to the highest probabilities in the topic. Since each word in a document is assigned with a topic, each document can be represented by the proportion or mixtures of topics used to generate that document. Due to the size of the data, we have implemented our own version in C# using Gibbs inference detailed in (Griffiths and Steyvers 2004) for the inference part. We set the number of topics to 50, run the Gibbs for 5000 samples and use the last Gibbs sample to interpret the results.

We perform nonparametric Wilcoxon tests on the hypothesis of equal medians in psycholinguistic processes and topics between the LOW and HIGH groups. To further substantiate the difference in the use of these aspects between LOW and HIGH people, topics and LIWC are used as features to classify a post into either LOW and HIGH groups. Modern machine learning supervised classifiers broadly aim to build models from training data and assign a class to a new data point. They differ widely in principle, and we used 15 different classifiers in several classification paradigms, including rule-, tree-, nearest-neighbor-, Naive-Bayes- and support-vector-based. Three meta-learning algorithms are also used in the classification. We briefly describe them below:

- Rule based (e.g., PART, OneR, Jrip) builds rules and they differ in the complexity of the rule construction. OneR builds one rule for each attribute and chooses rules with smallest error rate.
- Tree based (e.g., J48, Random Forest, Random Tree, REPTree): Basic algorithms (J48) construct a decision tree based on training data using information entropy to choose the most effective feature that splits the data into the desired classes, whilst random forest constructs an ensemble of decision trees and outputs the mode of classes output from the individual tress.
- Nearest neighbour (Kstar, IBk) algorithms attempt to classify an instance in terms of its neighbouring points - they

[1] http://www.liwc.net/descriptiontable1.php, retrieved January 2013

Category	Number of rejections
#friends	56
#communities	57
#followers	41
#posts written	38
#comments made	54
#comments received	54

Table 5: Number of rejections in the Wilcoxon tests on the hypothesis of equal medians in the use of 68 LIWC features between LOW and HIGH groups.

differ in the numbers of neighbours considered and the distance measures used.

- Probabilistic methods (Naive Bayes) are methods that construct the conditional probability distributions of underlying features given a class label and classification on unseen case is then done by comparing the class likelihood.
- Support Vector Machines (SMO) is an SVM using sequential minimal optimization for a binary classification that finds the separating plane between two classes with maximal margins.
- Meta-learning algorithms include bagging (Bagging) and boosting (Ada Boost and Logit Boost) techniques. They fall under the class of ensemble learners, that use multiple models to improve performance. Boosting incrementally adds models to concentrate on the errors made in the last training iteration. Bagging is a method to generate a aggregate predictor from multiple versions of predictors generated from the data.

Core implementation of the learning algorithms used Weka data mining software (Hall et al. 2009). 10-fold cross validations are run on the two feature sets. This essentially averages the results on 10 runs, sequentially using one held-out data fold for testing and other nine folds for training. To evaluate the performance of classification, we use the F-measure score. Given True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), the F-measure for the classification is defined as:

$$\text{F-measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where Precision and Recall are determined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

Analysis of psycholinguistic processes

We perform nonparametric Wilcoxon tests on the hypothesis that low and high social capital groups have equal medians in LIWC features. The null hypothesis H_0 is considered to be rejected at $p \leq .05$. All 68 categories in LIWC are considered. We found rejections of this hypothesis across most LIWC features in all social capital indicators, shown in Table 5. Therefore there is significant difference in the LIWC features between low and high social capital people. Especially, for *number of communities* indicator, only 11 of 68 LIWC features were found failed to be rejected the null hypothesis at the 5% significance level. Therefore it can be stated that people who joins more online communities are greatly different in the language styles from those who join less communities.

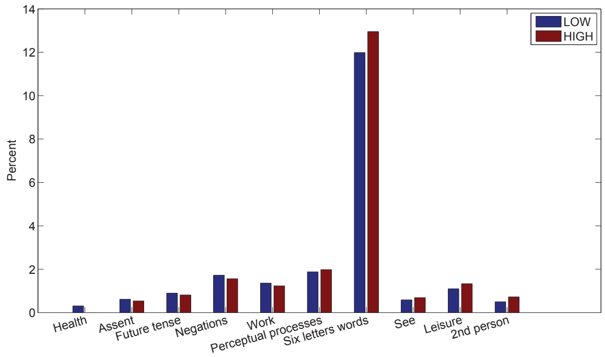


Figure 5: Difference in linguistics between low and high social capitals, based on number of followers.

Category	Number of rejections
#friends	30
#communities	35
#followers	18
#posts written	17
#comments made	30
#comments received	25

Table 6: Number of rejections in the Wilcoxon tests on the hypothesis of equal medians in the use of 50 topics between LOW and HIGH groups.

Based on *number of followers*, a user can be consider an influential or non-influential. (Bakshy et al. 2011) have demonstrated that Twitter users having many followers have a large influence in terms of forcing the diffusion of a URL, hence driving a greater volume of attention to the given site. Finding influentials is a key task in viral marketing, which co-opts social networks into marketing a brand or product (Rayport 1996). Influentials likely cause information about new products to spread more quickly than do non-influentials. What makes a user an influential, in term of the language style she or he uses? As shown in Figure 5, influentials were found to say more about *leisure* and less about *work* and *health* than do non-influentials. Also, influentials tend to use more formal words (six-letter words) while non-influentials prefer using spoken languages (*assent*, for example).

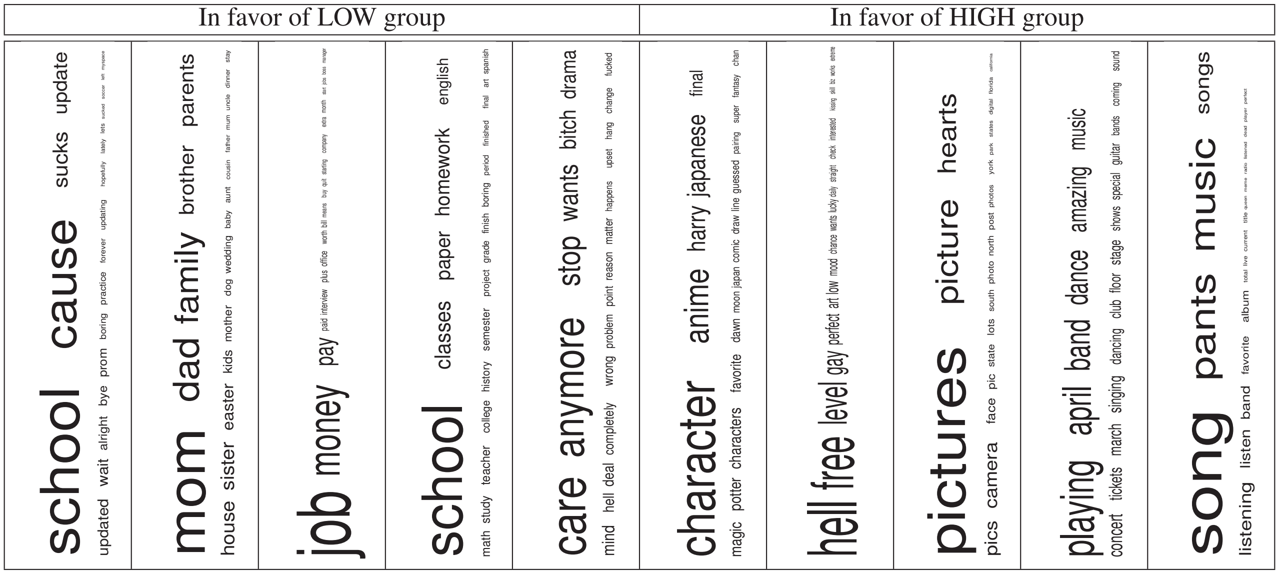

In favor of LOW group
In favor of HIGH group
school cause sucks update
updated wait alright bye prom boring practice forever updating
mom dad family brother parents
house sister easter kids mother dog wedding baby aunt cousin father mum uncle dinner stay
job money pay
school classes paper homework english
math study teacher college history semester project grade finish boring period finished final art spanish
care anymore stop wants bitch drama
mind hell deal completely wrong problem point reason matter happens upset hang change fucked
character anime harry japanese final
magic potter characters favorite dawn moon japan comic draw line guessed pairing super fantasy chan
hell free level gay perfect art low mood chance wants lucky sally straight check interested kissing
pictures picture hearts
pics camera face pic state lots south photo north post photos york park states digital florida
playing april band dance amazing music
concert tickets march singing dancing club floor stage shows special guitar bands coming sound
song pants music songs
listening listen band favorite album

Figure 6: Difference in topic consideration between low and high social capitals, based on number of followers.

Analysis of topics

We also conducted nonparametric Wilcoxon tests on the hypothesis of equal medians in topics between low and high social capital groups. The null hypothesis H_0 is considered to be rejected at $p \leq .05$. About half of topics were found statistically different in the use between the two groups, shown in Table 6. Thus it can be stated that the content discussed by bloggers having low vs. high degrees of social capital exhibits significant differences. This result brings a lot of implications. For example, in viral marketing, how to determine influentials based on their preference of topics? Figure 6 shows discovered topics (in cloud visualization: the size of a word is proportional the probability of that word in the topic) in the preference of influentials and non-influentials. It can be seen that influentials write more about entertainment such as *dance*, *songs*, *movies*, and *animation*, whilst non-influentials focus particularly on school-related topics. In addition, non-influentials were found to discuss more on job-related topics, which is in coincidence with the founding in the linguistic that influentials discuss more on *leisure* and less on *work* than do non-influentials.

Social capital classification

Since our analysis thus far has indicated that contents and language styles are highly discriminative in low and high social capital groups, we further examine if topics and linguistic features are predictive of social capital. Specifically, we study the problem of classifying a *unknown* post into low and high social capital group using topics and psycho-linguistic processes as features. For topics, 50 topics learned from LDA and their mixtures are used as features; for psycho-linguistic processes, all of 68 LIWC types are used as features. We use the popular machine learning package Weka with 10-fold cross validation scheme across 15 different classifiers.

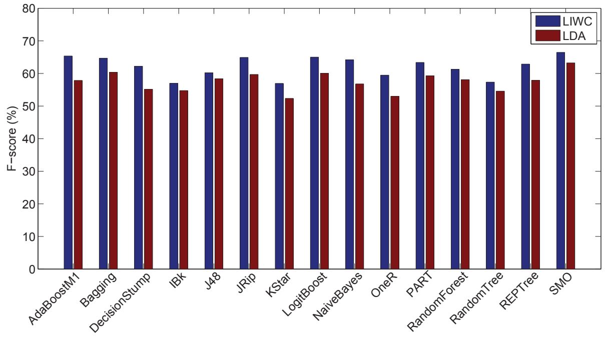


Figure 7: F-score for different algorithms in classifications of low and high social capitals, based on number of friends.

The results for this two-class classification (LOW vs. HIGH social capital) using LIWC and topics as features, in the case of *number of friends* index, are shown in Figure 7. We observe that prediction is better achieved when LIWC is used as features, achieving an accuracy of approximately 70 per cent. Without the need for a feature selection stage, leading to a lighter computational cost, the results for the predictions using psycholinguistic styles (through LIWC) are better to topics. This indicates a potential application of this information, language styles, for the purpose of analyzing the networking properties of social media.

Conclusion

With the popularity of social media, large numbers of people use online communities to connect to society, nurturing the novel concept of online social capital. We aimed to study the characteristics of people in different social capital. We established a framework of how social media can be used as a barometer for mood, and how we can use the underlying statistics to predict future moods. We addressed the formulation of online social capital, and establish the link between online social capital defined from social connectivity and users' mood. Social media, indeed, can be a barometer of mood. And, if used wisely, can play an important part in monitoring well-being.

We have also found significant differences between groups having different social capital when characterized by latent topics of discussion and psycholinguistic features: Wilcoxon tests rejected the hypothesis of equality on psycholinguistic processes and topics between two groups. Linguistic features are found to have greater predictive power than latent topics when classifying blog posts as either low or high social capital. Clear discrimination between writing styles and content, with good predictive power to classify posts is an important step in understanding new social media and its use in mental health. Results in this paper can form the foundation of early warning systems.

References

Bakshy, E.; Hofman, J.; Mason, W.; and Watts, D. 2011. Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the International Conference on Web Search and Data Mining*, 65–74.

Beaudoin, C. E., and Tao, C.-C. 2007. Benefiting from social capital in online support groups: An empirical study of cancer patients. *CyberPsychology & Behavior* 10(4):587–590.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bosacki, S.; Dane, A.; Marini, Z.; and YLC-CURA. 2007. Peer relationships and internalizing problems in adolescents: mediating role of self-esteem. *Emotional and Behavioural Difficulties* 12(4):261–282.

Bourdieu, P. 1986. *The forms of capital*. Greenwood Press. chapter 9, 241–258.

Bradley, M., and Lang, P. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. *University of Florida*.

Coleman, J. S. 1988. Social capital in the creation of human capital. *American Journal of Sociology* 95–120.

Fulkerson, J. A.; Story, M.; Mellin, A.; Leffert, N.; Neumark-Sztainer, D.; and French, S. A. 2006. Family dinner meal frequency and adolescent development: Relationships with developmental assets and high-risk behaviors. *Journal of Adolescent Health* 39(3):337–345.

Griffiths, T., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(90001):5228–5235.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. 2009. The Weka data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11(1):10–18.

Kazienko, P., and Musiał, K. 2006. Social capital in online social networks. In *Knowledge-Based Intelligent Information and Engineering Systems*, 417–424. Springer.

Kobayashi, T.; Ikeda, K.; and Miyata, K. 2006. Social capital online: Collective use of the Internet and reciprocity as lubricants of democracy. *Information, Community & Society* 9(5):582–611.

Lewis, D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the European Conference on Machine Learning*, 4.

Mandarano, L.; Meenar, M.; and Steins, C. 2010. Building social capital in the digital age of civic engagement. *Journal of Planning Literature* 25(2):123–135.

Morgan, A., and Haglund, B. J. 2009. Social capital does matter for adolescent health: evidence from the English HBSC study. *Health Promotion International* 24(4):363–372.

Nguyen, T.; Phung, D.; Adams, B.; Tran, T.; and Venkatesh, S. 2010. Classification and pattern discovery of mood in weblogs. *Advances in Knowledge Discovery and Data Mining* 283–290.

Nguyen, T. 2010. Mood patterns and affective lexicon access in weblogs. In *Proc. of the ACL Student Research Workshop*, 43–48.

Pennebaker, J.; Chung, C.; Ireland, M.; Gonzales, A.; and Booth, R. 2007. The development and psychometric properties of LIWC2007. *Austin, Texas: LIWC Inc*.

Putnam, R. D. 2000. *Bowling alone: The collapse and revival of American community*. Simon & Schuster.

Putnam, R. 2001. Social capital: Measurement and consequences. *Canadian Journal of Policy Research* 2(1):41–51.

Rasic, D.; Kisely, S.; and Langille, D. B. 2011. Protective associations of importance of religion and frequency of service attendance with depression risk, suicidal behaviours and substance use in adolescents in Nova Scotia, Canada. *Journal of Affective Disorders* 132(3):389–395.

Rayport, J. 1996. The virus of marketing. *Fast Company* 6(1996):68.

Rotenberg, K. J.; Boulton, M. J.; and Fox, C. L. 2005. Cross-sectional and longitudinal relations among children's trust beliefs, psychological maladjustment and social relationships: Are very high as well as very low trusting children at risk? *Journal of Abnormal Child Psychology* 33(5):595–610.

Russell, J. 2009. Emotion, core affect, and psychological construction. *Cognition & Emotion* 23(7):1259–1283.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47.

Shah, D. V.; Kwak, N.; and Holbert, R. L. 2001. "Connecting" and "disconnecting" with civic life: Patterns of Internet use and the production of social capital. *Political Communication* 18(2):141–162.

Steinfield, C.; Ellison, N. B.; and Lampe, C. 2008. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology* 29(6):434–445.

VicHealth. 2010. Opportunities for social connection - a determinant of mental health and wellbeing.

Wang, R.-H.; Chen, S.-W.; Tang, S.-M.; Lee, S.-L.; and Jian, S.-Y. 2011. The relationship between selected developmental assets and health-promoting behaviours of adolescents in Southern Taiwan. *Journal of Clinical Nursing* 20(3-4):359–368.

Wellman, B.; Haase, A. Q.; Witte, J.; and Hampton, K. 2001. Does the Internet increase, decrease, or supplement social capital?: Social networks, participation, and community commitment. *American Behavioral Scientist* 45(3):436–455.

Ye, Q.; Fang, B.; He, W.; and Hsieh, J. P.-A. 2012. Can social capital be transferred cross the boundary of the real and virtual worlds? An empirical investigation of Twitter. *Journal of Electronic Commerce Research* 13(2):145–156.

Zambon, A.; Morgan, A.; Vereecken, C.; Colombini, S.; Boyce, W.; Mazur, J.; Lemma, P.; and Cavallo, F. 2010. The contribution of club participation to adolescent health: Evidence from six countries. *Journal of Epidemiology and Community Health* 64(01):89–95.