Towards Discovery of Influence and Personality Traits Through Social Link Prediction

Thin Nguyen*, **Dinh Phung, Brett Adams and Svetha Venkatesh**
Curtin University of Technology
Bentley, WA 6102, Australia
**thin.nguyen@postgrad.curtin.edu.au*
{d.phung,b.adams,s.venkatesh}@curtin.edu.au

Abstract

Estimation of a person's influence and personality traits from social media data has many applications. We use social linkage criteria, such as number of followers and friends, as proxies to form corpora, from popular blogging site Livejournal, for examining two two-class classification problems: influential vs. non-influential, and extraversion vs. introversion. Classification is performed using automatically-derived psycholinguistic and mood-based features of a user's textual messages. We experiment with three sub-corpora of 10000 users each, and present the most effective predictors for each category. The best classification result, at 80%, is achieved using psycholinguistic features; e.g., influentials are found to use more complex language, than non-influentials, and use more leisure-related terms.

Introduction

The notion of a person's *influence* has received much attention with the rise of online social networks, and the ready data they provide. Intuitively, someone who is potentially influential should have many followers (i.e., others who monitor their actions or pronouncements). (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.

While influence is concerned with the effect of a person on a network or its information flow, *personality* has to do with the users themselves, and might also be inferred from social-linkage. Schrammel et al. (2009) found that individuals scoring high on traits of extraversion have more friends than those with low scores. Similarly, Ross et al. (2009) found that extraverts are members in significantly more groups than introverts.

The content of social media messages provides additional evidence for inferring about the attributes of influence and personality. Users can be characterized by bags of n-grams accumulated from their messages, and analyzed for linguistic properties, such as part-of-speech. Further insight can be obtained using richer models of language, such as the Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, and Booth 2007), which categorizes words according to a number of linguistic and psychological processes. Mood is also known to correlate with personality type. E.g., extraverts have been found to experience more pleasant and less unpleasant moods than do introverts (Rusting and Larsen 1995); extraverts in a good mood have been found to be more creative than introverts (Stafford et al. 2010). Mood can be estimated directly from language using special lexicons, such as Affective Norms for English Words (ANEW)(Bradley and Lang 1999). In addition, some social media forums, e.g. LiveJournal, allow the user to formally tag messages with a current mood.

We investigate the possibility of inferring social-linkage from message content. Specifically, we take a user's in-degree (henceforth termed *followers*) as a proxy for influence, and out-degree (henceforth termed *friend*s) and community membership as proxies for extraversion-introversion, and examine the predictive power of psycholinguistic (LIWC) and mood-related (ANEW) properties of the messages users create. Support vector machines (SVM) are used to examine the efficiency of these two feature sets on prediction, and regression models are used to evaluate the role of each feature in the predictions.

Experimental Setup

Dataset

The corpus for these experiments is drawn from blogging site Livejournal. Livejournal is notable for its *current mood* feature, which allows a blogger to attach a mood tag from a vocabulary of 132 predefined moods to a blog post. These mood tags can be viewed as groundtruth for a user's mood at the time of writing. Examples of positive moods include cheerful, happy, and grateful; negative moods include discontent, sad, and uncomfortable.

Livejournal supports one directed person-person link type. For a given user, we term incoming links *followers*, and outgoing links *friends*. Livejournal also allows users to join communities that discuss topics of interest.

The original Livejournal corpus is composed of nearly

19 million blog posts having current mood tags, made by over 1.6 million bloggers, during the years 2000 to 2010. On average, each blogger has 10 followers, 23 friends, and joins 7 communities. In order to explore the question of inferring personality traits and influence, we construct sub-corpora containing users with extreme numbers of followers and friends and community membership. The precise criteria for these sub-corpora are discussed below. To avoid spam and noisy data, the upper bound for link number is 150 (according to Dunbar's number (Dunbar 1993), a quasi-limit on the number of meaningful relationships a person can have). In addition, only users posting not less than 20 posts are taken into account.

Influentials and non-influentials corpora A corpus of influentials was created by collecting users, beginning with those having 150 followers, and less until a total of 5000 users had been collected. Users in this corpus have been active for less than one year, and have followers ranging from 78 to 150. Similarly, a non-influentials corpus was formed from 5,000 users, active for more than one year, and having only one or two followers.

Extraversion-introversion corpora based on number of friends A corpus of extroverts, based on number of friends, was created with a cut-off of 5,000 users, active for less than one year, having number of friends in the range 108 to 150. The introverts corpus was created from 5,000 users, active for more than one year, having between 1 and 3 friends.

Extraversion-introversion corpora based on number of communities Corpora for extroverts and introverts were created based on the number of communities a user had joined. Each corpus has 5000 users, with extraverts active for less than one year, and introverts active more than one year. The range of number of communities for extroverts is 56 to 150; for introverts, it is one to two communities.

Prediction model

An SVM classifier is used to examine the efficiency of LIWC categories and moods for predicting influence and personality type. A logistic regression model (LR) is also used to examine the predictive effect of each feature once the parameter is estimated.

Suppose we use LIWC features to predict whether a user is an extravert or introvert, the logistic model defines

$$\text{logit}(p) = \log \frac{p}{1-p} = \alpha + \sum_{i=1}^{n} \beta_i x_i$$

where p is the probability of the user being an extravert, $x_1, x_2, ..., x_n$ are independent variables, representing LIWC features, α and βs are regression parameters, and $\frac{p}{1-p}$ is the odds ratio.

The parameters α and βs are estimated through training data using a standard maximum likelihood approach (Wasserman 2004) and the resulting model is used on testing data. If β_i is greater than zero, an increase (or decrease) in x_i value is associated with an increase (or decrease) in the logit(p) (and thus p). Conversely, if β_i is less than zero, an increase (or decrease) in x_i value is associated with a decrease (or increase) in p. If β_i equals zero, x_i has no predictive power for predicting personality traits of a user.

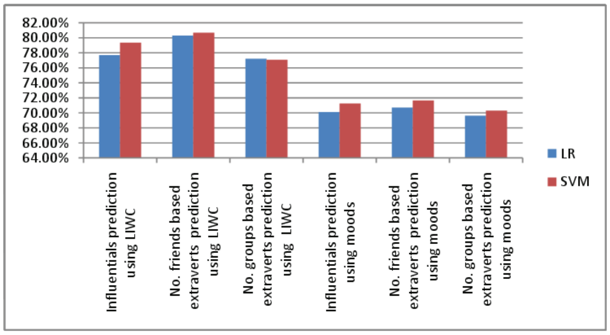


Figure 1: Prediction performance (F-measure).

Certain LIWC groups consist of others. E.g., *affect* (affective processes) consists of *negemo* (negative emotion) and *posemo* (positive emotion), resulting in high correlations among these LIWC groups. To reduce effects of multicollinearity in a multiple regression model, those predictor variables highly correlated are not taken together into the initial model. For example, since *negemo* and *posemo* are selected, *affect* is not taken into the initial model.

The subsequent decision to include or remove a predictor in the regression model is based on the Akaike information criterion (AIC), where AIC is defined as the residual deviance for the model plus twice the number of parameters in the model. Smaller AIC values indicate better models.

Ten-fold cross-validation is conducted for each SVM classifier and each logistic regression. The average accuracy and F-measure of predictions are reported.

Experimental Results

LIWC categories are found to be better than moods for predicting influentials and extraverts. The average accuracy and F-measure of the predictions are approximately 80% when using LIWC and 70% when using moods. We found the SVM results to be comparable to that of logistic regression, as shown in Figure 1. We therefore focus on interpreting parameters estimated from the logistic regression model.

Are you influential?

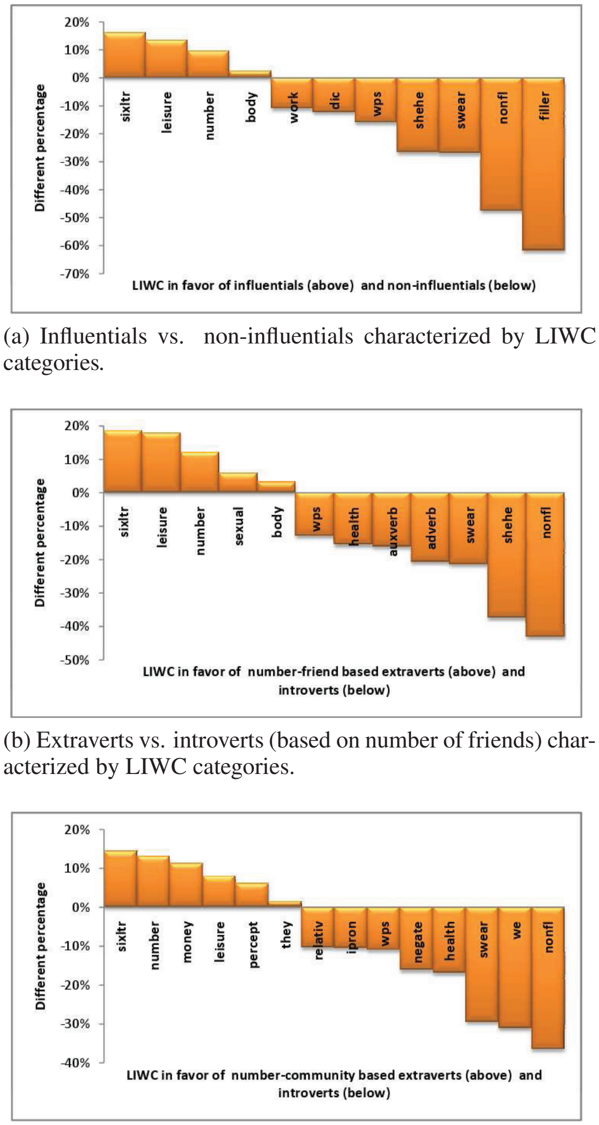
On building a regression model for influentials prediction using LIWC, after removing correlated features–e.g., *affect*, *anx*, *anger*, or *sad* (since *posemo* and *negemo* are used)–39 LIWC features, from total of 68 groups[1], form the independent variables of the initial model. Then predictors are selected according to the AIC criterion. The resulting model for prediction of influentials using LIWC is shown in Table 1.

The difference of language use in terms of LIWC features in the model by influentials and non-influentials can be seen

[1] The description for these LIWC groups shown at http://www.liwc.net/descriptiontable1.php

Features	β_i	Features	β_i
(Intercept–α)**	1.11	dic**	-4.97
sixltr**	30.3	wps**	-0.0289
leisure**	50.8	shehe**	-23.7
number*	30.1	swear**	-21.7
body**	51.2	nonfl**	-120
work**	-61.6	filler*	-89.7

Table 1: Regression model for prediction of influentials using psycholinguistic (LIWC) features (**: ps < .001, *: ps < .01).



(a) Influentials vs. non-influentials characterized by LIWC categories.

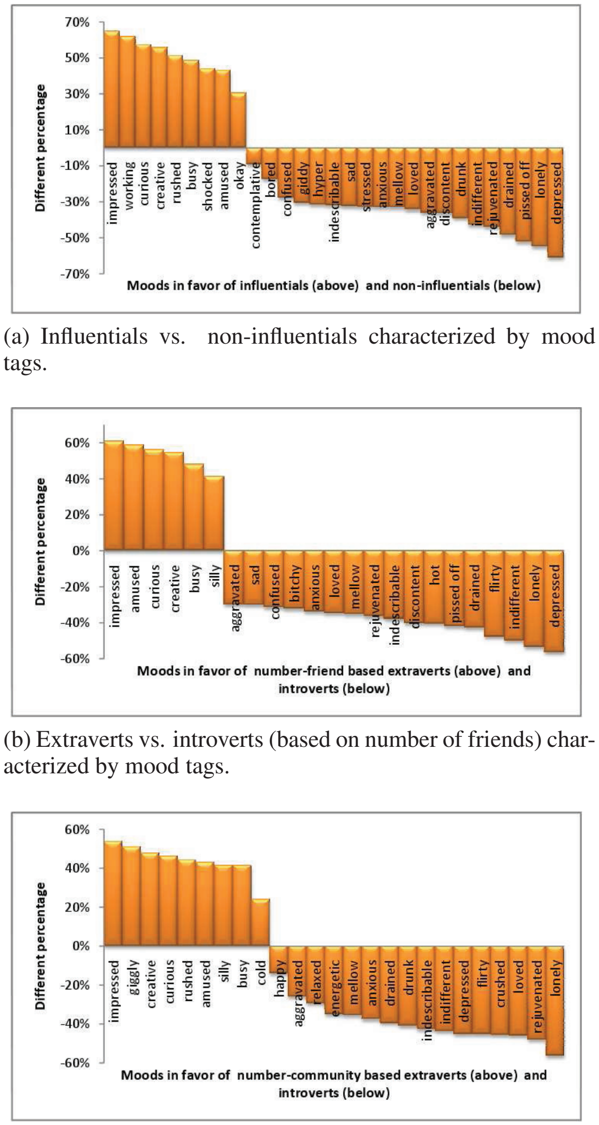
(b) Extraverts vs. introverts (based on number of friends) characterized by LIWC categories.

(c) Extraverts vs. introverts (based on community membership) characterized by LIWC categories.

Figure 2: Differences between influentials and non-influentials, and extraverts and introverts, characterized by LIWC categories.

Features	β_i	Features	β_i	Features	β_i
α	0.6	contemplative	-3.3	loved	-5.1
impressed	20.4	bored	-3.9	aggravated	-7.4
working	8.2	confused	-4.0	discontent	-10.4
curious	13.5	giddy	-7.0	drunk	-8.5
creative	10.4	hyper	-7.4	indifferent	-13.3
rushed	14.0	indescribable	-6.0	rejuvenated	-17.0
busy	8.6	sad	-7.2	drained	-11.6
shocked	11.0	stressed	-5.7	pissed off	-12.4
amused	5.0	anxious	-7.5	lonely	-11.5
okay	5.5	mellow	-9.5	depressed	-14.5

Table 2: Regression model for influentials prediction using mood tags as features (ps<.001).



(a) Influentials vs. non-influentials characterized by mood tags.

(b) Extraverts vs. introverts (based on number of friends) characterized by mood tags.

(c) Extraverts vs. introverts (based on community membership) characterized by mood tags.

Figure 3: Differences between influentials vs. non-influentials and extraverts vs. introverts on tagging moods.

in Figure 2a. Influentials prefer using words in the *sixltr* category, which is referred to as complex language (Tausczik and Pennebaker 2010) and a marker of presidential language (Slatcher et al. 2007). On the contrary, non-influentials prefer using words in the *spoken* category, e.g., nonfluency (*nonfl*, e.g, er, hm, umm) or fillers (e.g., blah). Also, influentials use more *leisure* words (e.g., karaoke, chat), more *number* words (e.g., firstly, billion, twice), fewer *swear* words (e.g., damn, piss), and fewer words per sentence (*wps*) than do non-influentials.

On building a regression model for influentials using mood, all 132 moods are used as predictors in the initial regression model. The features are then selected into the final model according to the AIC criterion. The resulting model for prediction of influentials using moods is shown in Table 2.

The differences in manifestation of mood usage in the model by influentials and non-influentials are shown in Figure 3a. Influentials use more high valence moods (all are larger than 6.0, based on ANEW words, where the valence scale ranges from 1–9; the larger the valence value, the happier the word) and more positive emotion moods (based on LIWC, except *shocked*) than do non-influentials. In contrast, non-influentials use more low valence moods (all are smaller than 5.0, except *loved*), and more negative emotion moods than do influentials.

Are you an extravert or introvert?

The process of building regression models for extraverts prediction is similar to that for influentials prediction. The LIWC features selected into the final regression models for predicting extraverts, and the differences in the use of the LIWC features by extraverts and introverts are shown in Figures 2b–2c. Similar to influentials, extraverts use fewer words per sentence (*wps*) than do introverts, in accord with the finding of (Mairesse et al. 2007). Extraverts also use more *sixltr* (complex language), *leisure*, *number*, *money* (e.g., audit, cash, owe), and *percept* (perceptual, e.g., heard, feeling) words than do introverts. In contrast, and similar to non-influentials, introverts use more *swear* and *nonfl* words in their posts than do extraverts. Introverts also use more *health* (e.g., headache, flu) and negations (again, in accord with (Mairesse et al. 2007)) than do extraverts.

The moods selected into the final regression models for predicting extraverts, and the differences in mood usage by extraverts and introverts, are shown in Figures 3b–3c. All moods used by extraverts are positive and high valence (except *cold*). In contrast, introverts use language dominated by negative and low valence emotions. This accords with the finding in (Rusting and Larsen 1995) that extraverts experience more pleasant and less unpleasant moods than do introverts.

Conclusion and Future Work

We have investigated the potential for language style and sentiment information of messages created in a social media site to predict social linkage, as proxy for influence and the personality traits of extraversion–introversion. It was found that psycholinguistic features have more predictive power than sentiment for popularity and personality traits.

Estimation of those who are most influential in a social network is of much use in viral marketing, where marketers are concerned to place the maximum investment in consumers who may drive the diffusion of positive opinion further than non-influentials. There is a role for personality prediction in the implementation of personalized information retrieval applications, or user-targeted advertising, where the most suited advertisement from a set of potential advertisements can be delivered to a particular user.

We have used social linkage as proxy for the complex concepts of influence, and personality traits. Self-report surveys for personality trait groundtruth, and content-based analysis of information diffusion to estimate influence, would augment this analysis.

References

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: Quantifying influence on Twitter. In *Proc. of the Intl. Conf. on Web Search and Data Mining (WSDM)*.

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

Dunbar, R. I. M. 1993. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences* 16(4):681–735.

Mairesse, F.; Walker, M.; Mehl, M.; and Moore, R. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* 30(1):457–500.

Pennebaker, J.; Francis, M.; and Booth, R. 2007. Linguistic inquiry and word count (LIWC) [computer software]. *Austin, Texas: LIWC Inc*.

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

Ross, C.; Orr, E.; Sisic, M.; Arseneault, J.; Simmering, M.; and Orr, R. 2009. Personality and motivations associated with facebook use. *Computers in Human Behavior* 25(2):578–586.

Rusting, C. L., and Larsen, R. J. 1995. Moods as sources of stimulation: Relationships between personality and desired mood states. *Personality and Individual Differences* 18(3):321 – 329.

Schrammel, J.; Köffel, C.; and Tscheligi, M. 2009. Personality traits, usage patterns and information disclosure in online communities. In *Proceedings of the 2009 British Computer Society Conference on Human-Computer Interaction*, 169–174.

Slatcher, R.; Chung, C.; Pennebaker, J.; and Stone, L. 2007. Winning words: Individual differences in linguistic style among us presidential and vice presidential candidates. *Journal of Research in Personality* 41(1):63–75.

Stafford, L.; Ng, W.; Moore, R.; and Bard, K. 2010. Bolder, happier, smarter: The role of extraversion in positive mood and cognition. *Personality and Individual Differences* 48(7):827–832.

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.

Wasserman, L. 2004. *All of statistics: a concise course in statistical inference*. Springer Verlag.