Recognising Personality Traits Using Facebook Status Updates

Golnoosh Farnadi[1,2], **Susana Zoghbi**[2], **Marie-Francine Moens**[2], **Martine De Cock**[1]
[1]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
{Golnoosh.Farnadi, Martine.DeCock}@UGent.be
[2]Department of Computer Science, Katholieke Universiteit Leuven, Belgium
{Susana.Zoghbi, Sien.Moens}@cs.kuleuven.be

Abstract

Gaining insight in a web user's personality is very valuable for applications that rely on personalisation, such as recommender systems and personalised advertising. In this paper we explore the use of machine learning techniques for inferring a user's personality traits from their Facebook status updates. Even with a small set of training examples we can outperform the majority class baseline algorithm. Furthermore, the results are improved by adding training examples from another source. This is an interesting result because it indicates that personality trait recognition generalises across social media platforms.

Introduction

User generated content (UGC) in online social networking sites provides a potentially very rich source of information for business intelligence applications that leverage this content for personalisation, such as on-line marketing. In this study we contribute to this effort by exploring the use of machine learning (ML) techniques to automatically infer users' personality traits based on their Facebook status updates.

Personality traits are commonly described using five dimensions (known as the Big Five), namely extraversion, neuroticism (the opposite of emotional stability), agreeableness, conscientiousness, and openness to experience. Since more than one trait can be present in the same user, for each trait we train a binary classifier that separates the users displaying the trait from those who do not. We use a variety of features as input for the classifiers, including features related to the text that users use in their status updates, features about the users' social network and time-related factors. We group the features in 4 categories and present the classification results that can be obtained for each of these categories separately as well as for the combinations of them. The results show a clear improvement over the majority class baseline.

To the best of our knowledge, the first work on Big Five personality trait recognition from online UGC was a word based bi-gram and tri-gram approach for the classification of author personality from weblog texts (Oberlander and Nowson 2006). More recently, Qiu et al. identified linguistic markers that are significantly correlated with 4 of the 5 personality dimensions of Twitter users (Qiu et al. 2012). Even though they did not use any ML technique, their study is very relevant to our work, because status updates, like tweets, are a form of microblogs. Compared to the two above mentioned studies, we have substantially less textual data per author. Another major difference is that in addition to linguistic features, in our work we also use social network information and time stamps.

Several authors have looked at the Big Five personality traits of Facebook users (Back et al. 2010). Most closely related to our work is Golbeck et al.'s study on personality prediction based on all publicly available information in a user's Facebook profile (Golbeck, Robles, and Turner 2011). They obtained promising results on a data set of 167 users, which is richer than ours in the sense that they have crawled many more profile features (e.g. gender, religion, list of favorite things,...) that were not available to us. Our experiments are carried out on a 250 user sample of a Facebook data set from the myPersonality project that was released on Feb 1, 2013 (Celli et al. 2013). (Kosinski, Stillwell, and Graepel 2013) used Facebook Likes to recognize personality, age, gender and sexual orientation, among others. More efforts on predicting personality traits using Facebook profile data from the myPersonality project are underway.

In our work we predict personality traits exclusively based on Facebook status updates, network properties and time factors. There have been a few studies that use network properties (Staiano et al. 2012), or temporal features (Chittaranjan, Blom, and Gatica-Perez 2013) to recognize personality. (Staiano et al. 2012) studied network features in the context of mobile phones to recognize personality traits. Their features are similar to the network features that we use in our work, however, the context is different. Recently, (Chittaranjan, Blom, and Gatica-Perez 2013) used temporal and actor-based features (e.g., the duration of out/in-going calls and number of calls), to recognize personality traits based on users' activity on their mobile phone.

In addition, we also train the classifiers on a corpus of 2468 essays labeled with personality traits (Mairesse et al. 2007). These essays are on average much longer than the status updates, and the context is different. Still, our results show that models trained on the essay data set perform well

on the Facebook data, and vice versa. This provides evidence that ML based models for personality trait recognition generalise across different domains. Advantages of this are that training examples from different social media platforms can be used in combination to train more accurate models and that such models are also applicable on social network sites for which no training data is available.

Methodology

The corpus contains 250 Facebook users and 9917 status updates, collected in the myPersonality project (Celli et al. 2013). Each user has filled in a questionnaire and, based on his answers, has been assigned one or more personality traits. Our goal is to predict these traits for a given user, where we identify a user with his set of available status updates (treated together as one text per user when extracting linguistic features), their time stamps, and his social network properties. Since more than one trait can be present in the same user, for each trait we train a binary classifier that separates the users displaying the trait from those who do not.

To this end, we use 4 kinds of numeric features:

- **LIWC features**: 81 features extracted using the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker and King 1999), consisting of features related to (1) standard counts (e.g., word count), (2) psychological processes (e.g., the number of anger words such as *hate, annoyed,* ... in the text), (3) relativity (e.g., the number of verbs in the future tense), (4) personal concerns (e.g., the number of words that refer to occupation such as *job, majors,* ...), (5) linguistic dimensions (e.g., the number of swear words). For a complete overview, we refer to (Tausczik and Pennebaker 2010).
- **Social Network features**: 7 features related to the social network of the user: (1) network size, (2) betweenness, (3) nbetweenness, (4) density, (5) brokerage, (6) nbrokerage, and (7) transitivity. For more information about these measures, see (O'Malley and Marsden 2008).
- **Time-related features**: 6 features related to the time of the status updates (we assume that all the times are based on one time zone): (1) frequency of status updates per day, (2) number of statuses posted between 6-11 am, (3) number of statuses posted between 11-16, (4) number of statuses posted between 16-21, (5) number of statuses posted between 21-00, and (6) number of statuses posted between 00-6 am.
- **Other features**: 6 features not yet included in the categories above: (1) total number of statuses per user, (2) number of capitalized words, (3) number of capital letters, (4) number of words that are used more than once, (5) number of urls, and (6) number of occurrences of the string PROPNAME — a string used in the data to replace proper names of persons for anonymisation purposes.

We compare the performance of 3 learning algorithms trained on these features, namely Support Vector Machine with a linear kernel (SVM), Nearest Neighbor with k=1 (kNN) and Naive Bayes (NB).

To score these algorithms, we use a weighted average of the well-known measures of precision, recall and F-measure, where the weights correspond to the relative sizes of the yes- and no-classes. Using the notations from Table

		Predicted class	
		Yes	No
Actual class	Yes	a	b
	No	c	d

Table 1: Binary classification confusing matrix

1, the weights are $w_{\text{yes}} = (a+b)/(a+b+c+d)$ and $w_{\text{no}} = (c+d)/(a+b+c+d)$. The precision and recall for the yes-class are $p_{\text{yes}} = a/(a+c)$ and $r_{\text{yes}} = a/(a+b)$ and similarly for the no-class $p_{\text{no}} = d/(b+d)$ and $r_{\text{no}} = d/(c+d)$. The weighted average precision, recall and F-measure (macro weighted by class-size average) are $P = w_{\text{yes}} \cdot p_{\text{yes}} + w_{\text{no}} \cdot p_{\text{no}}$, $R = w_{\text{yes}} \cdot r_{\text{yes}} + w_{\text{no}} \cdot r_{\text{no}}$ and

$$F = w_{\text{yes}} \cdot 2 \cdot \frac{p_{\text{yes}} \cdot r_{\text{yes}}}{p_{\text{yes}} + r_{\text{yes}}} + w_{\text{no}} \cdot 2 \cdot \frac{p_{\text{no}} \cdot r_{\text{no}}}{p_{\text{no}} + r_{\text{no}}}$$

Note that R coincides with accuracy. In the rest of the paper we simply refer to P, R and F as precision, recall and F-measure.

Classification results

Value	cEXT	cNEU	cAGR	cCON	cOPN
Yes	96	99	134	130	176
No	154	151	116	120	74

Table 2: Distribution of personality types in Facebook data

All results are obtained with Weka (Witten and Frank 2005) and compared against the majority class baseline algorithm (Base). Table 2 indicates that for some of the personality traits the baseline will be easier to beat than for others. Indeed, for Agreeableness (cAGR) and Conscientiousness (cCON) the instances are distributed fairly equally over the yes- and the no-class, but for Extraversion (cEXT), Neuroticism (cNEU) and Openness (cOPN) this is far less so. Especially for the latter trait, where the yes-instances make up almost 70% of the users, the baseline performs well. Oberlander and Nowson found a similar distribution w.r.t. the Openness trait, which even led them to not include this trait in their experiments on author classification of weblogs (Oberlander and Nowson 2006).

To investigate how each feature group contributes to the results, we trained the binary classifiers using all three algorithms. Note that due to the small size of the data, we do not remove the non-English statuses. All results are averaged over a 10-fold cross-validation, and a two-tailed paired t-test is done to evaluate significant differences with respect to the baseline at the $p < .05$ level. The results based on precision are presented in Table 3, those for recall in Table 4 and those for F-measure in Table 5. In the tables, bold values present the most significant improvement compared to

Features	Algorithm	cEXT	cNEU	cAGR	cCON	cOPN
-	Base	0.38	0.36	0.29	0.27	0.50
LIWC	SVM	0.58•	0.48•	0.47•	0.55•	0.60•
	kNN	0.58•	0.54•	0.50•	0.54•	0.54
	NB	0.58•	0.52•	0.52•	0.48 •	0.60 •
Social	SVM	**0.71**•	0.36	**0.60**•	0.45 •	0.50
	kNN	0.62•	0.53•	0.52•	0.47•	0.60•
	NB	0.67•	**0.62**•	0.52•	0.55•	**0.63**•
Time	SVM	0.38	0.36	0.33	**0.59**•	0.50
	kNN	0.63•	0.54•	0.53•	0.50•	0.55•
	NB	0.51•	0.44	0.26	0.26*	0.60
Other	SVM	0.40	0.40	0.35	0.52•	0.50
	kNN	0.45•	0.57•	0.50•	0.51•	0.57•
	NB	0.54•	0.51•	0.46•	0.57•	0.59•

Table 3: Classification results based on precision

Features	Algorithm	cEXT	cNEU	cAGR	cCON	cOPN
-	Base	0.62	0.55	0.54	0.52	0.70
LIWC	SVM	0.61	0.57	0.50	0.54	0.70
	kNN	0.57	0.53	0.50	0.54	0.54*
	NB	0.53*	0.55	0.53	0.47	0.62*
Social	SVM	**0.68**•	0.6•	**0.57**•	0.52	0.70
	kNN	0.62	0.53	0.51	0.47	*0.60
	NB	0.59	0.56	0.50	0.54	0.46*
Time	SVM	0.62	0.60•	0.54	**0.55**	0.70
	kNN	0.62	0.54*	0.53	0.51	0.56*
	NB	0.61	0.43*	0.46*	0.48*	0.38*
Other	SVM	0.62	**0.61**•	0.52	0.53	0.70
	kNN	0.47*	0.56	0.50	0.51	0.57*
	NB	0.60	0.49*	0.49	0.55	0.67

Table 4: Classification results based on recall

The social features often offer the highest improvement over the baseline. All (algorithm,feature set)-pairs, except for NB with time features, perform at least as well as the baseline and often outperform it in terms of the **precision**. Remarkably, NB trained with the time features obtains the most negative results overall.

Recall is harder to improve; however for 4 of the personality traits we have an (algorithm, feature set)-pair that outperforms the baseline in terms of recall. For Openness, the best result is equal to the baseline. The main reason is that in the corpus the distribution for the Openness class is very imbalanced (see Table 2) with the majority (almost 70%) of users being labeled as Open; therefore the majority class baseline is difficult to improve.

In terms of the **F-measure**, most (algorithm,feature set)-pairs perform at least as well as the baseline and often outperform it. There are a few exceptions, including NB with time features.

To improve performance, we iteratively combine different sets of features as presented in Table 6. We start with the (algorithm, feature set)-pair that outperformed the baseline in all 3 measures. If there is more than one pair, we choose the one with better results and fewer features. At each round, we add all the unused feature sets one by one to create new combinations. Among these, we choose the combination that has the highest performance and we compare it to the previous round. This process stops when performance of the current round does not improve w.r.t the previous one.

Features	Algorithm	cEXT	cNEU	cAGR	cCON	cOPN
-	Base	0.47	0.45	0.37	0.36	0.58
LIWC	SVM	0.56•	0.49	0.45•	**0.54**•	**0.61**•
	kNN	0.57•	0.52•	0.50•	0.53•	0.54
	NB	0.53	0.51	0.48•	0.44•	0.60
Social	SVM	**0.62**•	0.45	0.50•	0.41•	0.58
	kNN	**0.62**•	0.53•	0.51•	0.46•	0.59
	NB	0.58•	0.54•	0.47•	0.48 •	0.46*
Time	SVM	0.47	0.45	0.39	0.47•	0.58
	kNN	**0.62**•	0.53•	**0.53**•	•0.5	0.55
	NB	0.50	0.30*	0.31*	0.33*	0.32*
Other	SVM	0.47	0.46	0.38	0.43•	0.58
	kNN	0.46	**0.56** •	0.49•	0.51•	0.57
	NB	0.52	0.45	0.44•	0.49•	0.59

Table 5: Classification results based on F-measure

Class	Features	Algorithm	Precision	Recall	F-Measure
cEXT	**Social**	SVM	0.71	0.68	0.62
	Social+Time	NB	0.69	0.68	0.65
cNEU	**Other**	SVM	0.40	0.61	0.46
	Other+Social	NB	0.63	0.55	0.53
cAGR	**Social**	SVM	0.60	0.57	0.50
	Social+Time	SVM	0.60	0.57	0.49
cCON	LIWC	SVM	0.55	0.54	0.54
	LIWC+Social	kNN	0.55	0.54	0.54
	LIWC+Social+Time	kNN	0.56	0.55	0.55
	LIWC+Social+Time+Other	kNN	0.54	0.54	0.53
cOPN	LIWC	SVM	0.60	0.70	0.61
	LIWC+Social	SVM	0.62	0.71	0.62
	LIWC+Social+Time	SVM	0.61	0.70	0.62

Table 6: Classification results by combining the feature sets

In terms of accuracy (our recall), the best results for extraversion and agreeableness are obtained using only the social features. Similarly, for neuroticism the highest performance is achieved using only the "other features" set. For conscientiousness and openness the combination of features results in a small improvement.

Features	cEXT	cNEU	cAGR	cCON	cOPN
Network size	0,31•	-0,18•	0,07	0,14•	0,02
Betweenness	0,25•	-0,13•	0,05	0,11	0,04
nBetweenness	0,22•	-0,03	0,11	0,12	-0,06
Density	-0,24•	0,10	-0,08	-0,14•	0,05
Brokerage	0,25•	-0,13•	0,05	0,11	0,04
nBrokerage	0,23•	-0,08	0,09	0,08	-0,01
Transitivity	-0,27•	0,14•	-0,15•	-0,02	-0,06

Table 7: Correlation results between social features and personality traits

To assess which features are important to predict the personality traits, we use Pearson's correlation. Table 7 presents the results of our experiments based on **social features**. Those values with (•) are significantly ($p < .05$) correlated with personality. Interestingly, all the social features are significantly correlated with extraversion. It can be one

of the reasons that social features have the best classification performance for this particular trait. The positive correlation between network size ($\rho = 0.31$) and negative correlation ($\rho = -0.24$) with density for extraversion are in line with the results of Golbeck et al. based on Facebook data (Golbeck, Robles, and Turner 2011). This indicates that extroverts have many friends, but their friends are not likely friends with each other. Similarly, for conscientiousness users, there is a positive correlation ($\rho = 0.14$) with network size and negative correlation ($\rho = -0.14$) with density, so their network is large and sparse. Agreeableness ($\rho = -0.15$) and extraversion ($\rho = -0.27$) have negative correlations with transitivity, however neuroticism presents a positive correlation ($\rho = 0.14$). Another interesting finding is that the correlations of social features with extraversion present opposite signs w.r.t. the correlations of social features with neuroticism.

We also find the correlation of **time features** with personality traits. Concerning correlations w.r.t. the time features, there are only two significant correlations. Both of them are between time-stamps and conscientiousness. These two significant negative correlations are (1) statuses published between 6am to 11am ($\rho = -0.19$) and (2) statuses published between 00am to 6am ($\rho = -0.13$). This indicates that conscientious users are less likely to update statuses between 00am to 11am.

The significant correlations between **LIWC features** (Tausczik and Pennebaker 2010) and personality traits are as follows. Extroverts tend to use dictionary words ($\rho = 0.16$), 2nd person ($\rho = 0.15$) and 3rd person singular ($\rho = 0.16$) pronouns, past tense verbs ($\rho = 0.14$), social interaction words ($\rho = 0.14$), cues associated with the five senses ($\rho = 0.13$), health related words ($\rho = 0.14$) and not swear words ($\rho = -0.14$). Neurotic users tend to update their statuses with anger words ($\rho = 0.20$) and are less likely to use social interaction words ($\rho = -0.13$), positive emotions ($\rho = -0.14$), or and prepositions ($\rho = -0.14$). Agreeable users are more likely to use sexual words ($\rho = 0.19$). Conscientious users often present cues associated with the five senses ($\rho = 0.15$) and prepositions ($\rho = 0.15$) and less likely to use verbs ($\rho = -0.13$), 1st person singular pronouns ($\rho = -0.15$), or negative emotions ($\rho = -0.13$). Open users update their statuses by using dictionary words ($\rho = 0.15$), social interaction words ($\rho = 0.17$), affective processes ($\rho = 0.14$), cues associated with hearing ($\rho = 0.20$), 2nd person singular ($\rho = 0.15$) and 3rd person plural ($\rho = 0.13$) pronouns.

Cross-domain learning

Given the relatively small size of the Facebook corpus, we investigate the use of an additional corpus to train our classifiers and apply them to our original data source. The new corpus contains 2468 essays from psychology students who were told to write whatever comes to their mind for 20 minutes (Mairesse et al. 2007). Each essay was analysed and labeled based on the Big Five personality classes by (Pennebaker and King 1999). From this data source we can only extract linguistic features, as time and social network information is not available. Table 8 shows the distribution of labeled classes in this corpus. Compared to the Facebook data, they appear to be more balanced.

Value	cEXT	cNEU	cAGR	cCON	cOPN
Yes	1277	1233	1310	1254	1272
No	1191	1235	1158	1214	1196

Table 8: Distribution of personality types in Essay corpus

We extract LIWC features on the Essay corpus and train SVM, kNN and NB classifiers. The result shows that SVM ("Essay" in Table 9) always outperforms the majority baseline ("Base-Essay" in Table 9) in Essay corpus. This is inline with the study of (Mairesse et al. 2007) on this corpus. Due to limited space we do not present the corresponding results of kNN and NB.

To evaluate the effect of cross-domain learning, we set up 3 experiments:

1. FB-Essay: Train the classifier based on Facebook data and apply it on Essay corpus.
2. Essay-FB: Train the classifier based on Essay corpus and apply it on Facebook data.
3. Essay+FB-FB: Train the classifier based on the combination of Essay corpus and Facebook data, and apply it on Facebook data. For this task, we manually create 10 folds out of the Facebook data. Each training fold is expanded with the Essay corpus. The results of this experiment are also averaged over 10 folds.

Due to space restrictions, we only present the results of the most successful classifier (among SVM, kNN and NB) in Table 9. As can be expected, the best results for the Essay corpus are achieved with classifiers that are trained on essay data. Still, classifiers that are only trained on the Facebook data always do at least as well as the baseline algorithm, and substantially improve the precision and F-measure.

Similar to the previous experiment, using the Essay corpus as training examples for classification of the Facebook data outperforms the baseline in terms of precision and F-measure. In fact, unlike in the previous experiment, the results of classifying Facebook users trained either by the Essay corpus or the Facebook data do not differ much. Even more, for agreeableness, better results are obtained for the Facebook data when training on the Essay corpus than on Facebook data. This is likely due to the difference in size of the training sets, but is at the same time an indication that personality trait recognition generalises across social media platforms.

Finally, the best results are typically obtained when training on both sources simultaneously. The main exception is agreeableness where, as mentioned above, training only on the essay corpus gives the best results.

Conclusion and next steps

We explored the use of ML techniques (SVM, kNN, NB) for the automatic recognition of personality traits of a Facebook user, based on his social network properties, the text of his status updates and their frequencies and time of posting.

Class	Approach	Algorithm	Precision	Recall	F-Measure
cEXT	Base-Essay	-	0.27	0.52	0.35
	Essay	SVM	**0.54•**	**0.54**	**0.53•**
	FB-Essay	NB	0.52•	0.52	0.52•
	Base-FB	-	0.38	0.62	0.47
	FB	SVM	0.58•	0.61	0.56•
	Essay-FB	kNN	0.58•	0.56	0.56•
	Essay+FB-FB	SVM	**0.64•**	**0.65 •**	**0.60•**
cNEU	Base-Essay	-	0.25	0.50	0.33
	Essay	SVM	**0.58•**	**0.58•**	**0.57•**
	FB-Essay	NB	0.52•	0.55•	0.51•
	Base-FB	-	0.36	0.55	0.45
	FB	kNN	0.54•	0.53	0.52•
	Essay-FB	NB	0.52•	0.55	0.53•
	Essay+FB-FB	SVM	**0.60•**	**0.60•**	**0.55•**
cAGR	Base-Essay	-	0.28	0.53	0.37
	Essay	SVM	**0.56•**	**0.56•**	**0.55•**
	FB-Essay	kNN	0.50•	0.53	0.44•
	Base-FB	-	0.29	0.54	0.37
	FB	kNN	0.50•	0.50	0.50•
	Essay-FB	SVM	**0.58•**	**0.58•**	**0.58•**
	Essay+FB-FB	SVM	0.56 •	0.55	0.54•
cCON	Base-Essay	-	0.26	0.51	0.34
	Essay	SVM	**0.56•**	**0.56•**	**0.56•**
	FB-Essay	kNN	0.54•	0.51	0.41•
	Base-FB	-	0.27	0.52	0.36
	FB	SVM	**0.55•**	**0.54**	**0.54•**
	Essay-FB	kNN	0.54•	**0.54**	**0.54•**
	Essay+FB-FB	kNN	**0.55•**	0.53	0.53•
cOPN	Base-Essay	-	0.27	0.52	0.35
	Essay	SVM	**0.61•**	**0.61•**	**0.61•**
	FB-Essay	NB	0.52•	0.52	0.43•
	Base-FB	-	0.50	0.70	0.58
	FB	SVM	0.60•	0.70	0.61•
	Essay-FB	NB	0.63•	0.70	0.61•
	Essay+FB-FB	SVM	**0.67•**	0.68	**0.63•**

Table 9: Classification results by cross-domain learning

Our most interesting findings on a set of 250 users and 9917 status updates are:

- Even with a fairly small set of training examples we can outperform the majority class baseline algorithm, hence Facebook status updates do contain important cues of their authors' personality types.
- There is no single kind of features that gives the best results for all personality traits.
- The results for linguistic features are improved when using additional training examples from another domain, which provides evidence that ML based approaches for personality trait recognition generalise across domains. Advantages of this are that training examples from different social media platforms can be used in combination to train more accurate models and that such models are also applicable on social network sites for which no training data is available.

Class	Features	Algorithm	Precision	Recall	F-Measure
cEMS	LIWC:Anger	NB	**0.68•**	**0.66•**	**0.59•**

Table 10: Results of applying LIWC:anger on cNEU

Aside from the work we have presented in this paper, there is clear potential in more fine grained feature selection to improve the classification results. As an example, Table 10 presents the results of applying a classifier that is only trained on the LIWC:anger feature for recognising emotional stability. Note that these results are better than any of the results for this personality trait from Table 3, 4 and 5.

References

Back, M. D.; Stopfer, J. M.; Vazire, S.; Gaddis, S.; Schmukle, S.; Egloff, B.; and Gosling, S. D. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science* 21:372–374.

Celli, F.; Pianesi, F.; Stillwell, D.; and Kosinski, M. 2013. Workshop on computational personality recognition (shared task). In *Proc. of WCPR13, in conjuction with ICWSM13. Boston, MA*.

Chittaranjan, G.; Blom, J.; and Gatica-Perez, D. 2013. Mining large-scale smartphone data for personality studies. *Personal Ubiquitous Comput.* 17:433–450.

Golbeck, J.; Robles, C.; and Turner, K. 2011. Predicting personality with social media. In *Proc. of CHI*, 253–262. New York, NY, USA: ACM.

Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proc. of PNAS*.

Mairesse, F.; Walker, M. A.; Mehl, M. R.; and Moore, R. K. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research* 30:457–501.

Oberlander, J., and Nowson, S. 2006. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proc. of ACL*, 627–634. Sydney, Australia: Association for Computational Linguistics.

O'Malley, A. J., and Marsden, P. V. 2008. The analysis of social networks. *Health Services and Outcomes Research Methodology* 8:222–269.

Pennebaker, J. W., and King, L. A. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* 77:1296–1312.

Qiu, L.; Lin, H.; Ramsay, J.; and Yang, F. 2012. You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality* 46:710–718.

Staiano, J.; Lepri, B.; Aharony, N.; Pianesi, F.; Sebe, N.; and Pentland, A. 2012. Friends don't lie: inferring personality traits from social network structure. In *Proc. of UbiComp*, 321–330. New York, NY, USA: ACM.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The Psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29:24–54.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.