The Role of Emotional Stability in Twitter Conversations

Fabio Celli
CLIC-CIMeC
University of Trento
fabio.celli@unitn.it

Luca Rossi
LaRiCA
University of Urbino
luca.rossi@uniurb.it

Abstract

In this paper, we address the issue of how different personalities interact in Twitter. In particular we study users' interactions using one trait of the standard model known as the "Big Five": emotional stability. We collected a corpus of about 200000 Twitter posts and we annotated it with an unsupervised personality recognition system. This system exploits linguistic features, such as punctuation and emoticons, and statistical features, such as followers count and retweeted posts. We tested the system on a dataset annotated with personality models produced from human judgements. Network analysis shows that neurotic users post more than secure ones and have the tendency to build longer chains of interacting users. Secure users instead have more mutual connections and simpler networks.

1 Introduction and Background

Twitter is one of the most popular micro-blogging web services. It was founded in 2006, and allows users to post short messages up to 140 characters of text, called "tweets".

Following the definition in Boyd and Ellison (2007), Twitter is a social network site, but is shares some features with blogs. Zhao and Rosson (2009) highlights the fact that people use twitter for a variety of social purposes like keeping in touch with friends and colleagues, raising the visibility of their interests, gathering useful information, seeking for help and relaxing. They also report that the way people use Twitter can be grouped in three broad classes: people updating personal life activities, people doing real-time information and people following other people's RSS feeds, which is a way to keep informed about personal intersts.

According to Boyd et al. (2010), there are many features that affect practices and conversations in Twitter. First of all, connections in Twitter are directed rather than mutual: users follow other users' feeds and are followed by other users. Public messages can be addressed to specific users with the symbol @. According to Honeycutt and Herring (2009) this is used to reply to, to cite or to include someone in a conversation. Messages can be marked and categorized using the "hashtag" symbol #, that works as an aggregator of posts having something in common. Another important feature is that posts can be shared and propagated using the "retweet" option. Boyd et al. (2010) emphasize the fact that retweeting a post is a means of participating in a diffuse conversation. Moreover, posts can be marked as favorites and users can be included into lists. Those practices enhance the visibility of the posts or the users.

In recent years the interest towards Twitter raised in the scientific community, especially in Information Retrieval. For example Pak and Paroubek (2010) developed a sentiment analysis classifier from Twitter data, Finin et al. (2010) performed Named Entity Recognition on Twitter using crowdsourcing services such as Mechanical Turk[1] and CrowdFlower[2], and Zhao et al. (2011) proposed a ranking algorithm for extracting topic keyphrases from tweets. Of course also in the personality recog-

[1] https://www.mturk.com/mturk/welcome

[2] http://crowdflower.com

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 10–17,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

nition field there is a great interest towards the analysis of Twitter. For example Quercia et al. (2011) analyzed the correlations between personality traits and the behaviour of four types of users: listeners, popular, hi-read and influential.

In this paper, we describe a personality recognition tool we developed in order to annotate data from Twitter and we analyze how emotional stability affects interactions in Twitter. In the next section, given an overview of personality recognition and emotional stability, we will describe our personality recognition system in detail and we present the dataset we collected from Twitter. In the last two sections we report and discuss the results of the experiment and we provide some provisional conclusions.

2 Personality Recognition

2.1 Definition of Personality and Emotional Stability

Personality is a complex of attributes that characterise a unique individual. Psychologists, see for example Goldberg (1992), formalize personality along five traits known as the "Big Five", a model introduced by Norman (1963) that has become a standard over the years. The five traits are the following: **Extraversion** (sociable vs shy); **Emotional stability** (calm vs insecure); **Agreeableness** (friendly vs uncooperative); **Conscientiousness** (organized vs careless); **Openness** (insightful vs unimaginative).

Among all the 5 traits, emotional stability plays a crucial role in social networks. Studying offline social networks, Kanfer and Tanaka (1993) report that secure (high emotional stability) subjects had more people interacting with them. Moreover, Van Zalk et al. (2011) reports that youths who are socially anxious (low emotional stability) have fewer friends in their network and tend to choose friends who are socially anxious too. We will test if it is true also in online social networks.

2.2 Previous Work and State of the Art

Computational linguistics community started to pay attention to personality recognition only recently. A pioneering work by Argamon et al. (2005) classified neuroticism and extraversion using linguistic features such as function words, deictics, appraisal expressions and modal verbs. Oberlander and Nowson (2006) classified extraversion, emotional stability, agreeableness and conscientiousness of blog authors' using n-grams as features. Mairesse et al. (2007) reported a long list of correlations between big5 personality traits and 2 feature sets, one from linguistics (LIWC, see Pennebaker et al. (2001) for details) and one from psychology (RMC, see Coltheart (1981)). Those sets included features such as punctuation, length and frequency of words used. They obtained those correlations from psychological factor analysis on a corpus of Essays (see Pennebaker and King (1999) for details) annotated with personality, and developed a supervisd system for personality recognition available online as a demo[3]. In a recent work, Iacobelli et al. (2011) tested different feature sets, extracted from a corpus of blogs, and found that bigrams and stop words treated as boolean features yield very good results. As is stated by the authors themselves, their model may overfit the data, since the n-grams extracted are very few in a very large corpus. Quercia et al. (2011) predicted personality scores of Twitter users by means of network statistics like following count and retweet count, but they report root mean squared error, not accuracy. Finally Golbeck et al. (2011) predicted the personality of 279 users from Facebook using either linguistic. such as word and long-word count, and extralinguistic features, such as friend count and the like. The State-of-the-art in personality recognition is reported in table 1. Argamon (Arg05) and Oberlander (Ob06) use naive bayes, Mairesse (Mai07) and Iacobelli (Ia11) use support vector machines and Golbeck (Gol11) uses M5 rules with a mix of linguistic and extralinguistic features.

E.Stab.	**Arg05**	**Ob06**	**Mai07**	**Ia11**	**Gol11**
acc	0.581	0.558	0.573	0.705	0.531

Table 1: State-of-the-Art in Personality Recognition from language for the emotional stability trait.

2.3 Description of the Unsupervised Personality Recognition Tool

Given a set of correlations between personality traits and some linguistic or extralinguistic features, we

[3] http://people.csail.mit.edu/francois/research/personality/demo.html

are able to develop a system that builds models of personality for each user in a social network site whose data are publicly available. In our system personality models can take 3 possible values: secure (s), neurotic (n) and omitted/balanced (o), indicating that a user do not show any feature or shows both the features of a neurotic and a secure user in equal measure. Many scholars provide sets of correlations between some cues and the traits of personality formalized in the big5. In our system we used a feature set taken partly from Mairesse et al. (2007) and partly from Quercia et al. (2011). The former provides a long list of linguistic cues that correlate with personality traits in English. The latter provides the correlations between personality traits and the count of following, followers, listed and retweeted.

We selected the features reported in table 2, since they are the most frequent in the dataset for which we have correlation coefficients with emotional stability.

Features	**Corr. to Em. Stab.**	**from**
exclam. marks	-.05*	Mai07
neg. emot.	-.18**	Mai07
numbers	.05*	Mai07
pos. emot.	.07**	Mai07
quest. marks	-.05*	Mai07
long words	.06**	Mai07
w/t freq.	.10**	Mai07
following	-.17**	Qu11
followers	-.19**	Qu11
retweeted	-.03*	Qu11

Table 2: Features used in the system and their Pearson's correlation coefficients with personality traits as reported in Mairesse et al. (2007) and Quercia et al. (2011). * = p smaller than .05 (weak correlation), ** = p smaller than .01 (strong correlation)

Exclamation marks: the count of ! in a post; **negative emoticons**: the count of emoticons expressing negative feelings in a post; **numbers**: the count of numbers in the post; **positive emoticons**: the count of emoticons expressing positive feelings in a post; **question marks**: the count of ? in a post; **long words**: count of words longer than 6 characters in the post; **word/token frequency**: frequency of repeated words in a post, defined as

$$wt = \frac{\textit{repeated words}}{\textit{post word count}}$$

following count: the count of users followed; **followers count**: the count of followers; **retweeted count**: the amount of user's posts retweeted.

The processing pipeline, as shown in figure 1, is divided in three steps: preprocess, process and evaluation.

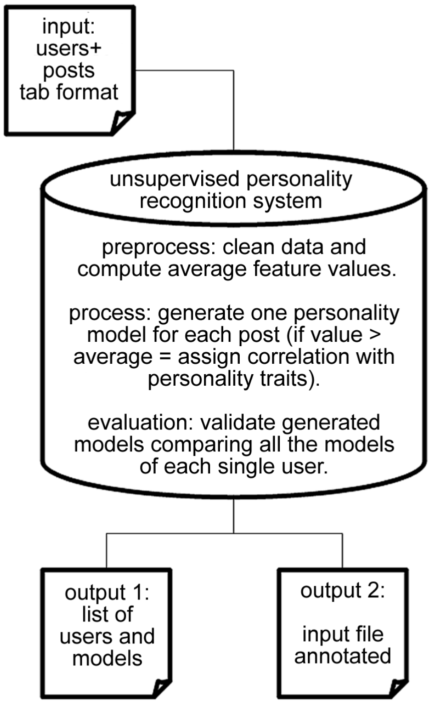
input: users+ posts tab format
unsupervised personality recognition system
preprocess: clean data and compute average feature values.
process: generate one personality model for each post (if value > average = assign correlation with personality traits).
evaluation: validate generated models comparing all the models of each single user.
output 1: list of users and models
output 2: input file annotated

Figure 1: Unsupervised Personality Recognition System pipeline.

In the preprocessing phase the system randomly samples a predefined number of posts (in this case 2000) in order to capture the average occurrence of each feature. In the processing phase the system generates one personality model per post matching features and applying correlations. If the system finds feature values above the average, it increments or decrements the score associated to emotional stability, depending on a positive or negative correlation. The list of all features used and their correlations with personality traits provided by Mairesse et al. (2007) (Mai07) and Quercia et al. (2011) (Qu11), is reported in table 2.

In order to evaluate the personality models generated, the system compares all the models generated for each post of a single user and retrieves one model per user. This is based on the assumption that one user has one and only one complex personality, and that this personality emerges at a various levels from written text, as well as from other extralinguistic cues. The system provides confidence and variability as evaluation measures. Confidence gives a measure of the consistency of the personality model. It is defined as

$$c = \frac{tp}{M}$$

where *tp* is the amount of personality models (for example "s" and"s", "n" and "n"), matching while comparing all posts of a user and *M* is the amount of the models generated for that user. Variability gives information about how much one user tends to write expressing the same personality traits in all the posts. It is defined as

$$v = \frac{c}{P}$$

where *c* is confidence score and *P* is the count of all user's posts. The system can evaluate personality only for users that have more than one post, the other users are discarded.

Our personality recognition system is unsupervised. This means that it exploits correlations in order to build models and does not require previously annotated data to modelize personality. Since the evaluation is performed directly on the dataset we need to test the system before using it. In the following section we describe how we tested system's performance.

2.4 Testing the Unsupervised Personality Recognition Tool

We run two tests, the first one to evaluate the accuracy in predicting human judges on personality, and the second one to evaluate the performance of the system on Twitter data. In the first one, we compared the results of our system on a dataset, called Personage (see Mairesse and Walker (2007)), annotated with personality ratings from human judges. Raters expressed their judgements on a scale from 1 (low) to 7 (high) for each of the Big Five personality traits on English sentences. In order to obtain a gold standard, we converted this scale into our three-values scheme applying the following rules: if value is greater or equal to 5 then we have "s", if value is 4 we have "o" and if value is smaller or equal to 3 we have "n". We used a balanced set of 8 users (20 sentences per user), we generated personality models automatically and we compared them to the gold standard. We obtained an accuracy of 0.625 over a majority baseline of 0.5, which is in line with the state of the art.

In the second test we compared the output of our system to the score of Analyzewords[4], an online tool for Twitter analysis based on LIWC features (see Pennebaker et al. (2001)). This tool does not provide big5 traits but, among others, it returns scores for "worried" and "upbeat", and we used those classes to evaluate "n" and "s" respectively. We randomly extracted 18 users from our dataset (see section 3 for details), 10 neurotics and 8 secure, and we manually checked whether the classes assigned by our system matched the scores of Analyzewords. Results, reported in table 3, reveal that our system has a good precision in detecting worried/neurotic users. The bad results for upbeat/secure users could be due to the fact that the class "upbeat" do not correspond perfectly to the "secure" class. Overall the performance of our system is in line with the state of the art.

	p	**r**	**f1**
n	0.8	0.615	0.695
s	0.375	0.6	0.462
avg	0.587	0.607	0.578

Table 3: Results of test 2.

3 Collection of the Dataset

The corpus, called "Personalitwit2", was collected starting from Twitter's public timeline[5]. The sampling procedure is depicted in figure 2.

We sampled data from December 25th to 28th, 2011 but most of the posts have a previous posting date since we also collected data from user pages, where 20 recent tweets are displayed in reverse chronological order. For each public user, sampled from the public timeline, we collected the nicknames of the related users, who had a conversation with the public users, using the @ symbol. We did this in order to capture users that are included in social relationships with the public users.

[4] http://www.analyzewords.com/index.php

[5] http://twitter.com/public timeline

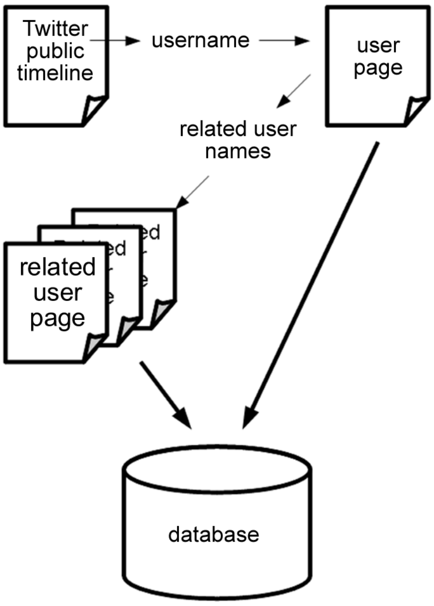

Figure 2: Data sampling pipeline.

We excluded from sampling all the retweeted posts because they are not written by the user themselves and could affect linguisitic-based personality recognition. The dataset contains all the following information for each post: username; text; post date; user type (public user or related user); user retweet count; user following count; user followers count; user listed count; user favorites count; total tweet count; user page creation year; time zone; related users (users who replied to the sampled user); reply score (*rp*), defined as

$$rp = \frac{\textit{page reply count}}{\textit{page post count}}$$

and retweet score (*rt*), defined as

$$rt = \frac{\textit{page retweet count}}{\textit{page post count}}$$

	min	**median**	**mean**	**max**
tweets	3	5284	12246	582057
following	0	197	838	320849
followers	0	240	34502	17286123
listed	0	1	385	539019
favorites	0	7	157	62689

Table 4: Summary of Personalitwit2.

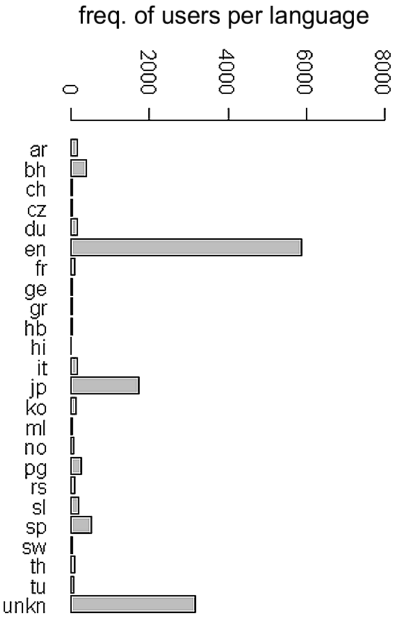

Figure 3: Frequency distribution of users per language. From the top: Arabic, Bahasa, Chinese, Czech, Dutch, English, French, German, Greek, Hebrew, Hindi, Italian, Japanese, Korean, Malay, Norwegian, Portuguese, Russian, Slovene, Spanish, Swedish, Thai, Turkish, Unidentified.

In the corpus there are 200000 posts, more than 13000 different users and about 7800 ego-networks, where public users are the central nodes and related users are the edges. We annotated the corpus with our personality recognition system. The average confidence is 0.601 and the average variability is 0.049. A statistical summary of the data we collected is reported in table 4, the distribution of users per language is reported in figure 3. We kept only English users (5392 egonetworks), discarding all the other users.

4 Experiments and Discussion

Frequency distribution of emotional stability trait in the corpus is as follows: 56.1% calm users, 39.2% neurotic users and 4.7% balanced users.

We run a first experiment to check whether neurotic or calm users tend to have conversations with other users with the same personality trait. To this purpose we extracted all the ego-networks annotated with personality. We automatically extracted

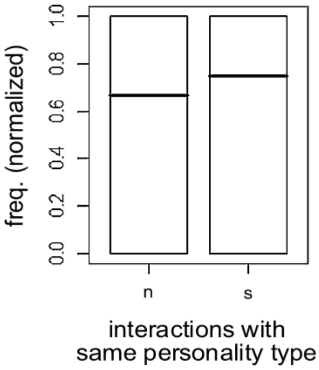

Figure 4: Relationships between users with the same personality traits.

the trait of the personality of the "public-user" (the center of the network) and we counted how many edges of the ego-network have the same personality trait. The users in the ego-network are weighted: this means that if a "public-user" had x conversations with the same "related-user", it is counted x times. The frequency is defined as

$$freq = \frac{\textit{trait count}}{\textit{egonetwork nodes count}}$$

where the same trait is between the public-user and the related users. The experiment, whose results are reported in figure 4, shows that there is a general tendency to have conversations between users that share the same traits.

We run a second experiment to find which personality type is most incline to tweet, to retweet and to reply. Results, reported in figure 5, show that neurotic users tend to post and to retweet more than stable users. Stable users are slightly more inclined to reply with respect to neurotic ones.

In order to study if conversational practices among users with similar personality traits might generate different social structure, we applied a social network analysis to the collected data through the use of the Gephi software[6]. We analysed separately the network of interactions between neurotic users (n) and calm users (s) to point out any personality related aspect of the emerging social structure. Visualisations are shown in figure 6.

Due to the way in which data have been acquired - starting from the users randomly displayed on the Twitter public timeline - there is a large number of scattered networks made of few interactions. Nevertheless the extraction of the ego networks allowed us to detect a rather interesting phenomena: neurotic users seem to have the tendency to build longer chains of interacting users while calm users have the tendency to build mutual connections.

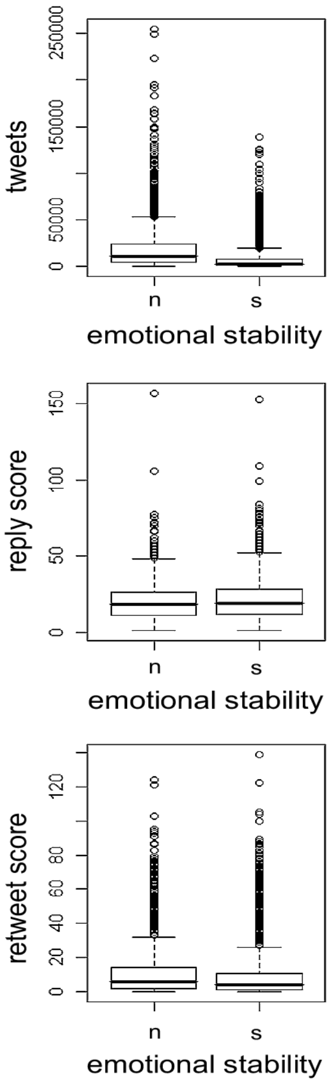

Figure 5: Relationships between emotional stability and Twitter activity.

The average path length value of neurotic users is 1.551, versus the average path length measured on the calm users of 1.334. This difference results in a network diameter of 6 for the network made of only neurotic users and of 5 for the network made

[6] http://www.gephi.org

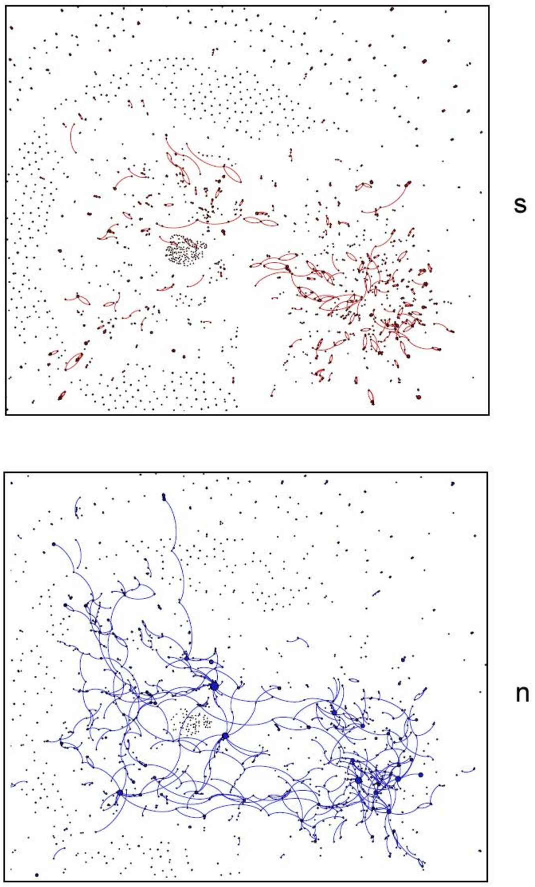

Figure 6: Social structures of stable (s) and neurotic (n) users.

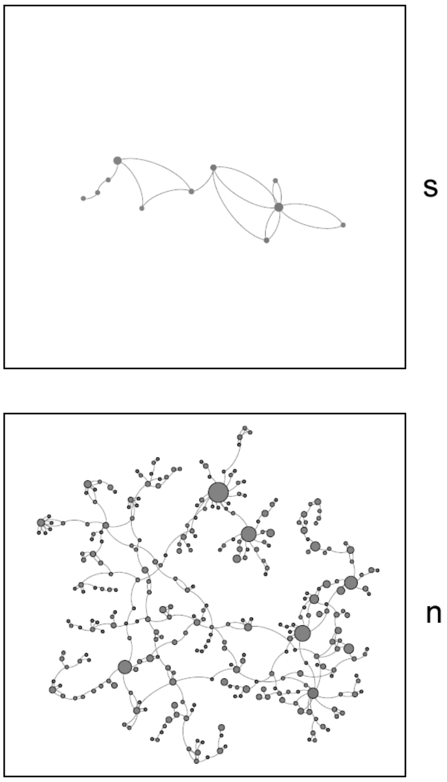

Figure 7: Giant components of stable (s) and neurotic (n) users.

of secure users. A single point of difference in the network diameter produces a neurotic network much more complex than the calm network. While this difference might be overlooked in large visualisations due to the presence of many minor clusters of nodes it becomes evident when we focus only on the giant component of the two networks in figure 7.

The giant components are those counting the major part of nodes and can be used as an example of the most complex structure existing within a network. As it should appear clear neurotic network contains more complex interconnected structures than calm network even if, as we claimed before, have on average smaller social networks.

5 Conclusions and Future Work

In this paper, we presented an unsupervised system for personality recognition and we applied it successfully on a quite large and richly annotated Twitter dataset. Results confirm some offline psychological findings in the social networks online, for example the fact that neurotic people tend to choose friends who are also neurotic.

We also confirm the fact that neurotic users have smaller social networks at the level of a single user, but they tend to build longer chains. This means that a tweet propagated in “neurotic networks” has higher visibility. We also found that neurotic users have the highest posting rate and retweet score.

In the future we should change the sampling settings in order to capture larger networks. It would be also very interesting to explore how other personality traits affect user’s behaviour. To this purpose we need to improve the personality recognition system and we would benefit from topic identification, which is another growing field of research.

References

Amichai-Hamburger, Y. and Vinitzky, G. 2010. Social network use and personality. In *Computers in Human Behavior*. 26(6). pp. 1289–1295.

Argamon, S., Dhawle S., Koppel, M., Pennebaker J. W. 2005. Lexical Predictors of Personality Type. In *Proceedings of Joint Annual Meeting of the Interface and the Classification Society of North America*. pp. 1–16.

Bastian M., Heymann S., Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of International AAAI Conference on Weblogs and Social Media*. pp. 1–2.

Boyd, D. Golder, S. and Lotan, G. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Proceedings of HICSS-43*. pp. 1–10.

Boyd, D. and Ellison, N. 2007. Social Network Sites: Definition, history, and scholarship. In *Journal of Computer-Mediated Communication* 13(1). pp. 210–230.

Celli, F., Di Lascio F.M.L., Magnani, M., Pacelli, B., and Rossi, L. 2010. *Social Network Data and Practices: the case of Friendfeed*. Advances in Social Computing, pp. 346–353. Series: Lecture Notes in Computer Science, Springer, Berlin.

Coltheart, M. 1981. The MRC psycholinguistic database. In *Quarterly Journal of Experimental Psychology*, 33A, pp. 497–505.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. pp. 80–88.

Golbeck, J. and Robles, C., and Turner, K. 2011. Predicting Personality with Social Media. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, pp. 253–262.

Golbeck, J. and Hansen, D.,L. 2011. Computing political preference among twitter followers. In *Proceedings of CHI 2011*: pp. 1105–1108.

Goldberg, L., R. The Development of Markers for the Big Five factor Structure. 1992. In *Psychological Assessment*, 4(1). pp. 26–42.

Honeycutt, C., and Herring, S. C. 2009. Beyond microblogging: Conversation and collaboration via Twitter. In *Proceedings of the Forty-Second Hawaii International Conference on System Sciences*. pp 1–10.

Kanfer, A., Tanaka, J.S. 1993. *Unraveling the Web of Personality Judgments: The Inuence of Social Networks on Personality Assessment*. Journal of Personality, 61(4) pp. 711–738.

Iacobelli, F., Gill, A.J., Nowson, S. Oberlander, J. Large scale personality classification of bloggers. 2011. In *Lecture Notes in Computer Science (6975)*, pp. 568–577.

Mairesse, F., and Walker, M.. PERSONAGE: Personality Generation for Dialogue. 2007. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.496–503.

Mairesse, F. and Walker, M. A. and Mehl, M. R., and Moore, R, K. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. In *Journal of Artificial intelligence Research*, 30. pp. 457–500.

Norman, W., T. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. In *Journal of Abnormal and Social Psychology*, 66. pp. 574–583.

Oberlander, J., and Nowson, S. 2006. Whose thumb is it anyway? classifying author personality from weblog text. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics ACL*. pp. 627–634.

Pak, A., Paroubek P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC 2010*. pp. 1320–1326.

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

Pennebaker, J. W., Francis, M. E., Booth, R. J. 2001. *Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum, Mahwah, NJ.

Platt, J. 1998. Machines using Sequential Minimal Optimization. In Schoelkopf, B., Burges, C., Smola, A. (ed), *Advances in Kernel Methods, Support Vector Learning*. pp. 37–49.

Quercia, D. and Kosinski, M. and Stillwell, D., and Crowcroft, J. 2011. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *Proceedings of SocialCom2011*. pp. 180–185.

Van Zalk, N., Van Zalk, M., Kerr, M. and Stattin, H. 2011. Social Anxiety as a Basis for Friendship Selection and Socialization in Adolescents' Social Networks. Journal of Personality, 79: pp. 499–526.

Zhao, D., Rosson, M.B. 2009. How and why people Twitter: The role that micro-blogging plays in informal communication at work. In *Proceedings of GROUP 2009* pp. 243–252.

Zhao, W.X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.P., Li, X. 2011. Topical keyphrase extraction from Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. pp. 379–388.