*Computational Intelligence*, Volume Special issue on Semantic Analysis in Social Media, Number 000, 2013

# Long Chains or Stable Communities? The Role of Emotional Stability in Twitter Conversations (PRE-PRINT version)

FABIO CELLI
*CLIC-CIMeC, University of Trento, Rovereto, Italy.*
fabio.celli@unitn.it

LUCA ROSSI
*LaRiCA, University of Urbino, Urbino, Italy*
luca.rossi@uniurb.it

In this paper, we address the issue of how emotional stability affects social relationships in Twitter. In particular we focus our study on users' communicative interactions, identified by the symbol "@". We collected a corpus of about 200000 Twitter posts and we annotated it with our 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 by human subjects and against a software for the analysis of Twitter data. Social Network Analysis shows that, while secure users have more mutual connections, neurotic users post more than secure ones and have the tendency to build longer chains of interacting users. Clustering coefficient analysis reveals that, while secure users tend to build stronger networks, neurotic users have difficulty in belonging to a stable community, hence they seek for new contacts in online social networks.

*Key words:* Data Mining; Personality Recognition; Social Network Analysis; Twitter.

## 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". The service rapidly gained worldwide popularity, with over 500 million active users as of 2012, generating over 340 million tweets daily and handling over 1.6 billion search queries per day. Since its launch, Twitter has become one of the top 10 most visited websites on the internet, and its use is spreading more and more[1].

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 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 producing real-time information and people following other people's RSS feeds, which is a way to keep informed about personal interests.

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

[1]Twitter Search Team (May 31, 2011). "The Engineering Behind Twitters New Search Experience". Twitter Engineering Blog. Twitter. Retrieved June 10, 2011.

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" practice. Boyd et al. (2010) emphasizes the fact that retweeting a post is a mean of participating in a diffuse conversation. Moreover, posts can be marked as favorites and users can be included into favorite lists. Those practices enhance the visibility of the posts or the users.

In recent years, scientific interest has begun to be focussed on the Twitter 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[2] and CrowdFlower[3], and Zhao et al. (2011) proposed a ranking algorithm for extracting topic keyphrases from tweets. Of course also the computational personality recognition field saw a great interest towards the analysis of Twitter. For example Quercia et al. (2011) analyzed the correlations between the five personality traits described by the Big5 factor model (see section 3) and the behaviour of four types of users: listeners, popular, hi-read and influential.

In this paper, we describe a system, developed by us, for the automatic extraction from text of the personality traits defined by the Big5 model. Among all the five personality traits, we analyze how emotional stability affects communicative interactions between users in Twitter. In the next section we present the dataset we collected from Twitter. Then we give an overview about the Big5 personality traits, emotional stability and automatic personality recognition from text. We provide a detailed description of our system and, in the last two sections, we report the results of the experiments and we draw some conclusions.

## 2. COLLECTION OF THE DATASET

A social network is a set of social entities connected by social relationships, such as friendship, co-working or information exchange. We define our social network as a set of Twitter users connected by communication exchanges.

We collected the corpus, called "Personalitwit2", starting from Twitter's public timeline[4]. The sampling procedure is depicted in figure 1. 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. We collected all the tweets displayed on the page of each public user, sampled from the public timeline, plus the nicknames of the related users, who had a conversation with the users, detected using the `@` symbol. Then we used the nicknames to collect tweets from the pages of the related users. Users with empty pages were discarded.

By doing this we consider conversations as network connections, rather than using the following-follower relationships. We filtered out all the retweeted posts because they are not written by the users themselves and could affect linguistic-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;

[2]https://www.mturk.com/mturk/welcome
[3]http://crowdflower.com
[4]http://twitter.com/public timeline

- 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 $rp = \frac{page\ reply\ count}{page\ post\ count}$, that provides a measure of user's tendency to communicate with others;
- retweet score (*rt*), defined as $rt = \frac{page\ retweet\ count}{page\ post\ count}$, that provides a measure of the tendency of the users to propagate information through the network;

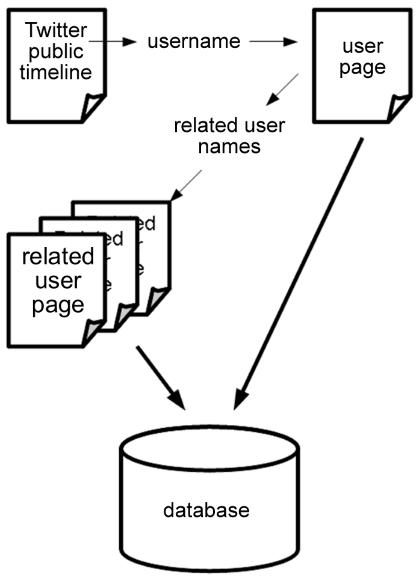


FIGURE 1. Data sampling pipeline.

TABLE 1. Distribution of Tweets (tw), following (f.ing), followers (f.er), listed (lstd) and favorites (favs) in Personalitwit2 (all) and in the English subset extracted (eng).

| all | **min** | **median** | **mean** | **max** | eng. | **min** | **median** | **mean** | **max** |
|---|---|---|---|---|---|---|---|---|---|
| tw | 3 | 5284 | 12246 | 582057 | tw | 3 | 5438 | 12180 | 254999 |
| f.ing | 0 | 197 | 838 | 320849 | f.ing | 0 | 235 | 890 | 320849 |
| f.rs | 0 | 240 | 34502 | 17286123 | f.rs | 0 | 2970 | 589604 | 17286123 |
| lstd | 0 | 1 | 385 | 539019 | lstd | 0 | 1 | 645 | 539019 |
| favs | 0 | 7 | 157 | 62689 | favs | 0 | 8 | 90 | 29425 |

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 connected to them through the edges. The statistical summary of Personalitwit2, reported in table 1, shows

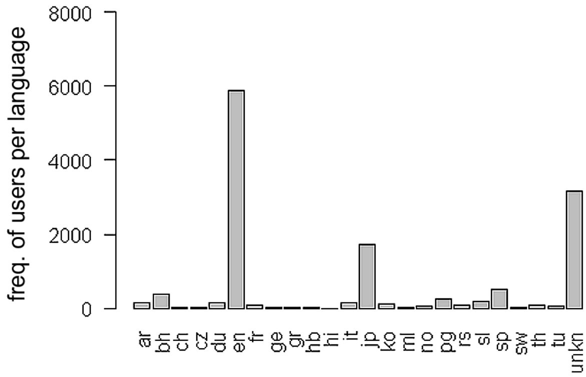


FIGURE 2. Frequency distribution of users per language. From the top: Arabic, Bahasa, Chinese, Czech, Dutch, English, French, German, Greek, Hebrew, Hindi, Italian, Japanese, Korean, Malay, Norwegian, Portuguese, Russian, Slovene, Spanish, Swedish, Thai, Turkish, Unidentified.

that there is a large number of users that do not use favorites and lists, while there are few users that do not have a following and followers. The distribution of users per language is reported in figure 2. We run experiments only on the English subset of the corpus (5392 egonetworks), Leaving the analysis of other languages to future works. Table 1 provides also a comparison of the english subset extracted from Personalitwit2, that shows how English Twitter users have more followers than the general average and less listed users. We extracted a small development dataset from the English subset for feature selection.

## 3. DEFINITION OF PERSONALITY AND EMOTIONAL STABILITY

Personality is seen as a complex of attributes that that uniquely characterise an individual. In psychology it is defined as an affect processing system, see DeYoung (2010), and, according to Adelsteinet al. (2011), it is connected to the behavioral responses of subjects to the environment.

A standard way to describe personality in psychology is the Big5 factor model, that has been introduced by Norman (1963). The Big5 consists of five bipolar personality traits, namely Extraversion, Emotional Stability, Agreeableness, Coscientiousness and Openness, that have been proposed in this form by Costa and MacCrae (1985). The five broad factors emerged from testing several different personality traits by means of factor analysis over self-reports, questionnaire data, peer ratings, and objective measures. It is important to note that different research teams came to very similar conclusions independently (see Digman (1990)). Extraversion describes a person along the two opposite poles of sociability and shyness. Emotional stability, which is sometimes referred by its negative pole (neuroticism), describes the modality of impulse control along a scale that goes from control (a calm and stable person) to instability (an anxious and neurotic person). Agreeableness refers to the tendency to be sympathetic and cooperative towards others, rather than suspicious and antagonistic. Conscientiousness describes a person in terms of self-discipline versus disorganization. Openness to experience refers to the tendency be creative and curious rather than unimaginative. According to Digman (1990), there have been a lot of studies in psychology that independently came to the conclusion that five are the right dimensions to describe personality. Despite a general agreement on the number of traits, there is no full agreement on their meaning, since some traits are vague. For example there is some dis-

TABLE 2. Proposals of Five personality traits Since Norman (1963). Adapted from Digman (1990).

| **Author** | I | II | III | **IV** | V |
|---|---|---|---|---|---|
| Norman | surgency | agreeab. | conscient. | emotion | culture |
| Borgatta | assertiv. | likeability | interest | emotion | intelligence |
| Eysenck | extrav. | - | - | neuroticism | - |
| Guilford | activity | disposition | introversion | em. stability | - |
| Buss&Plomin | activity | sociability | impulsivity | emotionality | - |
| Tellegen | pos. emotion | - | constraint | neg. emotion | - |
| Costa&McCrae | extrav. | agreeab. | conscient. | neuroticism | openness |
| Lorr | involvement | socialization | self-control | em. stability | independent |
| Hogan | sociability | likeability | prudence | adjustment | intellect |
| Digman | extrav. | compliance | will | neuroticism | intellect |

agreement about how to interpret the openness factor, which is sometimes called "intellect" rather than openness to experience. Emotional stability is one of the most robust personality traits. This emerges clearly comparing the proposals of personality traits reported in table 2, adapted from Digman (1990). 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.

## 4. AUTOMATIC PERSONALITY RECOGNITION: RELATED WORK

The Big5 is a formalized model suitable for computational analysis. There are two research fields that showed some interest in automatic personality recognition in recent years: Computational Linguistics and Social Network Analysis.

The Computational Linguistics community started paying attention to personality recognition only recently. In 2005 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. One year later Oberlander and Nowson (2006) classified extraversion, stability, agreableness and conscientiousness of blog authors' using n-grams as features and naive bayes (NB) as learning algorithm. In a very comprehensive work, Mairesse et al. (2007) reported a long list of correlations between big5 personality traits and two feature sets: LIWC (see Pennebaker et al. (2001) for details) and RMC (see Coltheart (1981)), that include word classification, like "positive emotions" or "anger" and scores like age of acquisition and word imageability. They obtained those correlations from psychological factor analysis on a corpus of Essays (see Pennebaker and King (1999) for details) and developed a supervisd system for personality recognition[5]. Luyckx and Daelemans (2008) built a corpus for stylometry and personality prediction from text in Dutch using n-grams of Part-Of-Speech and chunks as features. They used the Myers-Briggs Type Indicator schema, that includes 4 binary personality traits, (see Briggs and Myers (1980)) in place of the Big 5. Unfortunately their results are not comparable to any other because of the different language and schema used. In a recent work, Iacobelli et al. (2011) tested different features, such as stop words or inverse document frequency, and found that bigrams and stop words treated as

[5]demo available online at http://people.csail.mit.edu/francois/research/personality/demo.html

boolean features yield very good results in predicting personality in a large corpus of blogs using Support Vector Machines (SVM) as learning algorithm. 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. In Social Network Analysis, personality recognition has even a shorter history. Golbeck et al. (2011) predicted the personality of 279 users from Facebook using either linguistic (e.g. word counts) or social network features, such as friend count. Quercia et al. (2011) used network features to predict users' personality on Twitter using M5 rules as learning algorithm. In Computational Linguistics there is a tendency to predict classes of personality traits, and the evaluation measure is accuracy (acc), while in SNA the tendency is to predict personality trait scores rather than classes, and therefore related measures are mean absolute error (mae) and root mean squared error (rmse). The State-of-the-art in personality recognition for the emotional stability trait is reported in table 3.

TABLE 3. State-of-the-Art in Textual Personality Recognition. *=Results reported in Luyckx and Daelemans (2008). acc=accuracy; mae=mean absolute error; rmse=root mean squared error.

| **Author** | **Alg.** | **Measure** | **Result for Em. Stab.** |
|---|---|---|---|
| Argamon05 | NB | acc | 0.581* |
| Oberlander06 | NB | acc | 0.558* |
| Mairesse07 | SVM | acc | 0.573 |
| Iacobelli11 | SVM | acc | 0.705 |
| Golbeck11 | M5 | mae | 0.127 |
| Quercia11 | M5 | rmse | 0.85 |

## 5. PERSONALITY RECOGNITION TOOL

### 5.1. Description of the System

We developed a generative system that, given a set of correlations between personality traits and some linguistic or extralinguistic features, generates hypotheses of personality for each user in a social network site for which we have textual data. The system does not make use of any learning algorithm, rather it generates personality hypotheses for each tweet in the dataset, by applying correlations between text and the emotional stability trait, with a threshold filter for scores below the average. In the evaluation phase, the system generates one generalized hypothesis per user by comparing all the hypotheses generated for each user's tweets, computing also a confidence score that can be used for feature weighting, as we did.

In our system, personality can take 3 possible values: secure (s), neurotic (n) and omitted/balanced (o). Those users classified as neurotic are defined to be emotionally reactive and vulnerable to stress. Secure users are less easily upset and are less emotionally reactive. They tend to be calm, stable, and less exposed to negative feelings. The latter class, omitted/balanced, indicates that those users do not show any feature or shows both the features of a neurotic and a secure user in equal measure.

### 5.2. Feature Selection

As underlined by previous works, such as Mairesse et al. (2007), feature selection is extremely important for the performance of the system. Details are reported in table 4.

TABLE 4. Feature from Mairesse et al. (2007) and Quercia et al. (2011), for which we have correlations with emotional stability.

| **Features** | **freq.** | **conf.** | **ratio** |
|---|---|---|---|
| anxiety words | 0.016 | 0.009 | 1.778 |
| anger words | 0.076 | 0.012 | 6.333 |
| affect words | 0.185 | 0.08 | 2.313 |
| articles | 0.194 | 0.171 | 1.135 |
| exclam. marks | 0.166 | 0.166 | **1** |
| feeling words | 0.005 | 0.001 | 5 |
| family words | 0.028 | 0.019 | 1.474 |
| friend words | 0.000 | 0.000 | - |
| pronoun i | 0.156 | 0.122 | 1.279 |
| leisure words | 0.076 | 0.051 | 1.49 |
| long words | 0.313 | 0.308 | **1.016** |
| negative particles | 0.033 | 0.01 | 3.3 |
| negative emo. | 0.009 | 0.009 | **1** |
| numbers | 0.123 | 0.118 | **1.042** |
| parentheses | 0.019 | 0.018 | **1.008** |
| positive emo. | 0.052 | 0.052 | **1** |
| prepositions | 0.213 | 0.185 | 1.151 |
| pronouns | 0.322 | 0.265 | 1.215 |
| present | 0.194 | 0.172 | 1.128 |
| question marks | 0.175 | 0.171 | **1.023** |
| repeat ratio | 0.019 | 0.019 | **1** |
| sad words | 0.000 | 0.000 | - |
| sight words | 0.014 | 0.006 | 2.333 |
| space words | 0.123 | 0.109 | 1.128 |
| pronoun we | 0.038 | 0.009 | 4.222 |
| word count | 0.336 | 0.012 | 28 |
| n. of chars | 0.223 | 0.199 | 1.121 |
| n. of syllables | 0.336 | 0.301 | 1.116 |
| Kucera-Francis freq. | 0.185 | 0.168 | 1.101 |
| Kucera-Francis categ. | 0.341 | 0.000 | - |
| brown corp. freq. | 0.175 | 0.166 | 1.094 |
| Thorndike-Lorge freq. | 0.175 | 0.156 | 1.122 |
| concreteness | 0.303 | 0.274 | 1.106 |
| familiarity | 0.213 | 0.194 | 1.109 |
| imageability | 0.313 | 0.28 | 11.179 |
| meaninfulness | 0.261 | 0.237 | 1.101 |
| age of acquisition | 0.19 | 0.171 | 1.111 |
| following | 1.000 | 0.959 | **1.043** |
| followers | 1.000 | 0.964 | **1.037** |
| retweeted | 0.251 | 0.251 | **1** |

We exploit correlation between linguistic or extralinguistic cues and emotional stability taken from literature. In particular we used features partly taken 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. They are reported in table 4.

Feature selection algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection iteratively searches the set of possible features for the optimal subset. Good feature selection algorithms are often designed ad hoc on the target dataset, but there is always the risk of overfitting. To avoid this, there are some methodical approaches. From a theoretical perspective, the optimal feature selection, at least for supervised learning problems, requires an exhaustive search of all possible subsets of features of the chosen cardinality (subset selection). In this case this is impractical because of the large numbers of features available, hence we run feature ranking for feature selection.

Exploiting The confidence score, we run feature ranking by computing the ratio between the frequency of each feature in the devset and the confidence obtained with that feature. Best features are the ones with ratio close to 1. We arbitrarily decided to keep only features with ratio below 1.05. The selected features are the following:

- 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
- repeat ratio: ratio between words and 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.

## 5.3. System Pipeline

The processing pipeline, as shown in figure 3, is divided in three steps: preprocess, process and evaluation.

In the preprocessing phase the system randomly samples a predefined number of posts (we set this parameter to 2000 for the experiments) in order to capture the average occurrence of each feature. In the processing phase the system generates one personality hypothesis per post matching features and applying correlations. If the system finds feature values above the average computed in the preprocessing phase, 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 5. In the evaluation phase, the system compares all the hypotheses generated for each post of a single user and retrieves one generalized hypothesis 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 hypothesis. It is defined as

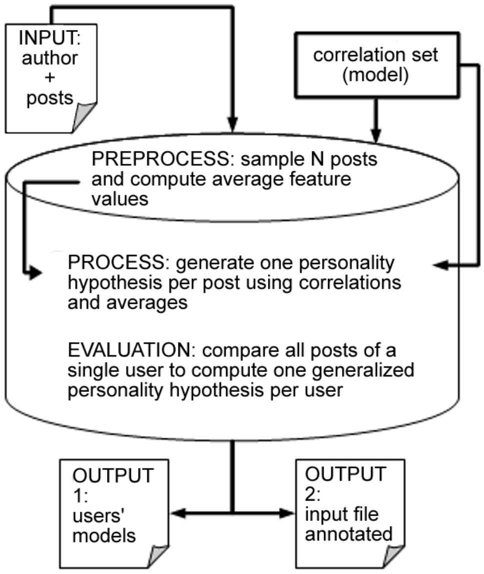


FIGURE 3. Personality Recognition System pipeline.

TABLE 5. 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 0.05 (weak correlation), ** = $p$ smaller than 0.01 (strong correlation).

| 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 |
| repeat ratio | .10** | Mai07 |
| following | -.17** | Qu11 |
| followers | -.19** | Qu11 |
| retweeted | -.03* | Qu11 |

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

where $tp$ is the amount of personality hypotheses (for example "s" and"s", "n" and "n"), matching while comparing all posts of a user and $M$ is the amount of the hypotheses 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 the 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.

We annotated the corpus with that system. The average confidence is 0.601 and the average variability is 0.049. We have seen that our personality recognition system exploits correlations as model in order to generate hypotheses that fit the data. This means that the system does not require previously annotated data, which is very difficult to produce from Social Network Sites, because of privacy issues. The drawback is that it provides only confidence and variability as evaluation measures. This means that we should test the performance of the system on gold-standard data in order to evaluate the real performance. In the following section we describe how we tested the system.

### 5.4. Testing the 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, 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) with an average length of 16 words per sentence. We generated personality hypotheses automatically with our system using only linguistic features (since these is not data from a social network) namely: exclamation marks; negative emoticons; numbers; positive emoticons; question marks; long words and repeat ratio. We compared them to the gold standard and we obtained an accuracy of 0.625 over a majority baseline of 0.5. This results outperforms the one of Mairesse for the Emotional Stability trait, and this is due only to feature selection.

In the second test we compared the output of our system (this time using all the selected features) to the score of Analyzewords[6], an online tool for Twitter analysis based on LIWC features (see Tausczik and Pennebaker (2010)). 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 20 users from our dataset, 10 neurotics, 8 secure and 2 users for which Analyzewords could not provide the analysis, thus discarded. We manually checked whether the classes assigned by our system matched the scores of Analyzewords. Results, reported in table 6, reveal that our system has a good pre-

TABLE 6. Results of test 2.

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

cision in detecting worried/neurotic users. We suggest that the bad results for upbeat/secure users is due to the fact that the class “upbeat” does not fully correspond to the “secure” class. Overall the performance of our system is in line with the state of the art.

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

## 6. 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 ran an 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.

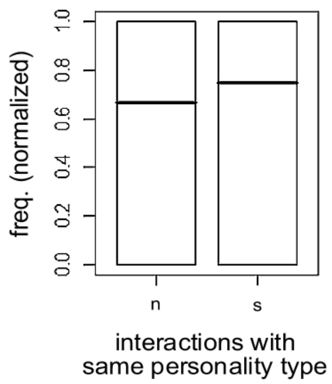


FIGURE 4. Relationships between users with the same personality traits.


FIGURE 5. Relationships between emotional stability and Twitter activity.

We automatically extracted the trait of the personality of the "public-user" (the center of the ego 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" has $x$ 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. In particular 66.7% of the neurotic users and 74.8% of the secure ones have conversations with users of the same personality type.

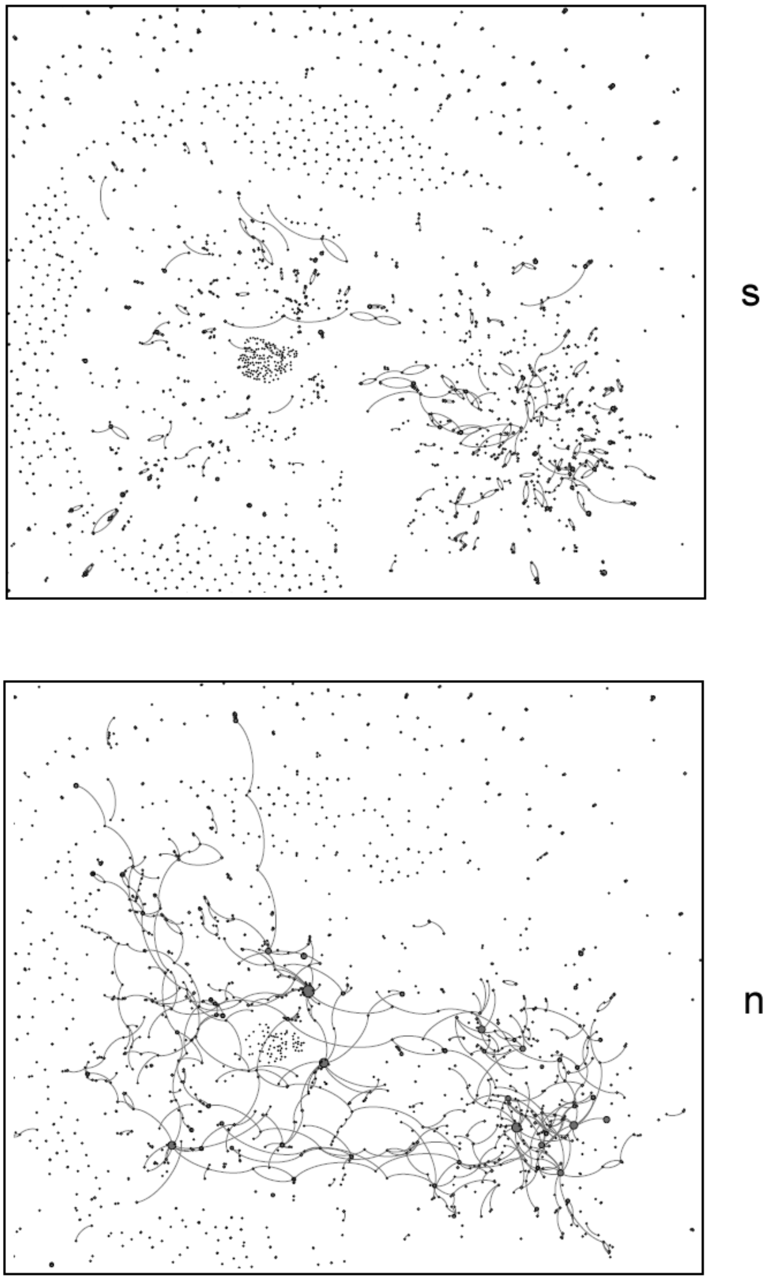


FIGURE 6. Social structures of stable (s) and neurotic (n) users.

We run a second experiment to find which personality type is most inclined 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. A Wilcoxon rank test with continuity correction confirmed that the differences between stable and secure users are significative either for replies ($p$-value = 0.0001238), retweets ($p$-value = 7.727e-14) and posting activity ($p$-value = 2.2e-16).

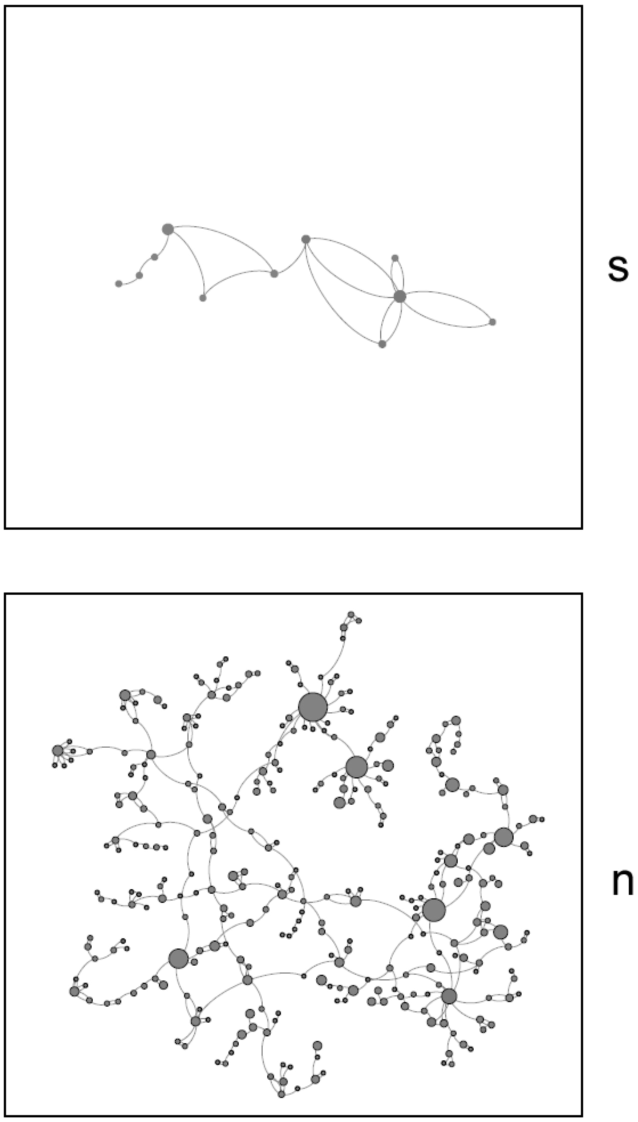


FIGURE 7. Giant components of stable (s) and neurotic (n) users.

From the analysis of the "o" class emerged that those users tend to have interaction only with other user types (48.1% with neurotics and 51.9% with secures), and that they have an average reply rate and posting activity. Only retweet score is higher for the omitted/balanced users with respect to secures and neurotics. A manual look at the dataset revealed that there are many retweets that are only links without other text, thus classified as "o". We suggest that omitted/balanced cannot be considered as a real class, but it is just useful for cleaning the noise from "s" and "n" classes.

In order to study if conversational practices among users with similar personality traits might generate a different social structure, we applied a social network analysis to the collected data through the use of the Gephi software[7], an open-source network analysis and visualization software package. The gathered data allowed us to construct a weighted directed network counting 10192 nodes (or Twitter users) and 10479 weighted arcs. The arcs have been weighted according to the number of messages actually exchanged between two users.We analyzed 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 then we analyzed inter-group interactions. Visualizations are shown in figure 6.

Due to the the data acquisition strategy - starting from the users randomly displayed on the Twitter public timeline - there are 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.

The average path length value of neurotic users is 1.535, versus the average path length measured on the calm users of 1.349. This difference results in a network diameter of 7 for the network made of only neurotic users and of 5 for the network made of secure users. A few points 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 visualizations 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 be clear, the neurotic network contains more complex interconnected structures than the calm network even if it has on average smaller social networks, as we claimed before.

In order to explain the biggest network diameter discovered among the neurotic users it could be useful to take into consideration the weight of the connections between the nodes. If, as we declared before, the weight of the edges represents the intensity of the exchange of messages happening between two users we could ask if a larger network is necessary a more active network.

Table 7 displays the average weight of the edges connecting together Neurotic users (Intra-Neutotic Network), Stable Users (Intra-Stable Network) or bridging together Neurotic with Stable users (Inter-personality Network). The results clearly show that intra-stable network has an average edge weight higher that intra-neurotic users. This means that on average stable users seems to have more constant and frequent communications with similar users while neurotic users seem to have a more erratic use of their network with less repetition and less defined preferential connections.

TABLE 7. Avg. weight for Neurotic and Stable users

| | **Av. Weight** |
|---|---|
| Intra-Stable Network | 0.074 |
| Intra-Neurotic Network | 0.030 |
| Inter-personality Network | 0.063 |

This result is consistent with the differences in network diameter that we have detected and, at the same time, it could be a partial explanation of them. While stable users tend to communicate in a stronger way with a limited number of users, neurotic users seems to communicate in a less stable way with a larger number of users ending up, in this way,

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

producing more larger communication structures.

Nevertheless the communication structures built by the neurotic users seems to be more fragile and less solid. The analysis of clustering coefficient, as define by Watts and Strogatz (1998) and implemented in Gephi following Latapy (2008), of neurotic and stable users support this idea: the average cluster coefficient for neurotic users is 0.016 while the average cluster coefficient for stable users is 0.377. The data suggest that stable users contribute to smaller but better connected communicative structures while neurotic users contribute to wider but less tight communicative structures.

## 7. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a generative system for personality recognition that is able to annotate large amounts of data in social network sites, a domain where it is very difficult to produce gold standard annotations with personality traits. We produced a quite large and richly annotated Twitter dataset, that we make available to the scientific community. The system proved to have a high performance in detecting neurotic users, and outperforms the system from which features are taken.

Results of the analysis of neurotic and secure users support non-internet related socio-psychological theories, for example the fact that secure people tend to choose friends who are also secure. Nevertheless we found also that the behaviour of neurotic users is very different online and offline. For instance, while offline neurotic users tend to choose friends who are also neurotic, online they search for relationships appeearently without caring about the fact that the other person is a neurotic or a stable user.

Our results confirm also the fact that neurotic users have weaker social networks at the level of a single user, but they tend to build longer chains of interactions, searching for new relationships. This means that a tweet propagated in “neurotic networks” has potentially higher visibility, but it will move through less *significant* links, in the sense that there is less affect between these links with respect to the links between secure users. The Neurotic average clustering coefficient is, in fact, significantly lower that the stable users’ clustering coefficient and this suggests a bigger difficulty in belonging to a stable community. We also found, that neurotic users have the highest posting rate and retweet score, and this is once again explained as their will to seek for new relationships, in the hope to find some that can become stable.

In the future we would like to repeat the experiments on other social networks, like Facebook for example, in order to see whether different social practices, like mutual friendship relations influence network building for secure and neurotic users. It would be also very interesting to explore how other personality traits affect user’s behaviour into a social network, and also whether other traits, such as agreeableness, affect secure and neurotic users in communication exchanges.

Another interesting avenue of future research would be to improve the generative personality recognition system by testing new feature sets, finding new ways for testing its performance and examining whether it’s output can be used for unsupervised or semisupervised learning.

## REFERENCES

ADELSTEIN J. S, SHEHZAD Z, MENNES M, DEYOUNG C.G, ZUO X-N, KELLY C, MARGULIES D.S, BLOOMFIELD A, GRAY J.R, CASTELLANOS X.F, MILHAM M.P.. 2011. Personality Is Reflected in the Brain’s Intrinsic Functional Architecture. *In* PLoS ONE. **6**(11): pp. e27633.

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., 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.

BRIGGS, I., and MYERS, P.B.. 1980. Gifts differing: Understanding personality type. Mountain View, CA: Davies-Black Publishing.

CELLI, F., DI LASCIO F. M. L., MAGNANI, M., PACELLI, B., and ROSSI, L. 2010. Social Network Data and Practices: the case of Friendfeed. *In* Advances in Social Computing, Series: Lecture Notes in Computer Science, Springer, Berlin. pp. 346–353.

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

COSTA, P.T., JR. and MCCRAE, R.R.. 1985. The NEO Personality Inventory manual. *In* Psychological Assessment Resources. Odessa, FL.

DEYOUNG, C.G.. 2010. Toward a Theory of the Big Five. *In* Psychological Inquiry. **21**: pp. 26–33.

DIGMAN, J.M.. 1990. Personality structure: Emergence of the five-factor model. *In* Annual Review of Psychology **41**: pp. 417–440.

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 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10). pp. 80–88.

GOLBECK, J., ROBLES, C., and TURNER, K. 2011. Predicting Personality with Social Media. *In* Proceedings of the 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 CHI2011. pp. 1105–1108.

GOLDBERG, L. R.. 1992. The Development of Markers for the Big Five factor Structure. *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.

IACOBELLI, F., GILL, A. J., NOWSON, S. OBERLANDER, J.. 2011. Large scale personality classification of bloggers. *In* Lecture Notes in Computer Science (6975), pp. 568–577.

LATAPY, M., 2008. Main-memory Triangle Computations for Very Large (Sparse (Power-Law)) Graphs. *In* Theoretical Computer Science (TCS) 407 (1-3), pp. 458–473.

LUYCKX K. and DAELEMANS, W.. 2008. Personae: a Corpus for Author and Personality Prediction from Text. *In* Proceedings of LREC2008, the Sixth International Language Resources and Evaluation Conference. pp. 1–7.

JOHN, O. P., DONAHUE, E. M. and KENTLE, R. L.. 1991. The Big Five Inventory–Versions 4a and 54. *In* University of California, Berkeley, Institute of Personality and Social Research, pp. 1–8.

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

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

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. *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.

pp. 627–634.

PAK, A., PAROUBEK P.. 2010. Twitter as a corpus for sentiment analysis and opinion mining. *In* Proceedings of LREC2010. 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. and BOOTH, R. J.. 2001. Inquiry and Word Count: LIWC 2001. Lawrence Erlbaum, Mahwah, NJ.

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

QUERCIA, D., KOSINSKI, M., STILLWELL, D. and CROWCROFT, J.. 2011. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. *In* Proceedings of SocialCom2011. pp. 180–185.

TAUSCZIK, Y. R., and PENNEBAKER, J. W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *In* Journal of Language and Social Psychology, **29**(1): pp. 24–54.

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. *In* Journal of Personality, **79**:pp. 499–526.

WATTS, D.J., STROGATZ, S.H., 1998. Collective dynamics of "small-world" networks. *In* Nature 393, pp. 440–442.

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 GROUP2009. pp. 243–252.

ZHAO, W. X., JIANG, J., HE, J., SONG, Y., ACHANANUPARP, P., LIM, E. P. and LI, X.. 2011. Topical keyphrase extraction from Twitter. *In* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT11). pp. 379–388.