LITERATURE REVIEW:

SUBMITTED TO THE INFORMATION AND COMPUTER SCIENCES DEPARTMENT OF THE

UNIVERSITY OF HAWAI‘I IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTORATE OF PHILOSOPHY

IN

INFORMATION AND COMPUTER SCIENCES

DECEMBER 2012

By

William R. Wright

# Abstract

Observing someone's personality can be quite useful in user interface design, recommender systems, marketing, employment decisions, and understanding team interactions. Lately researchers have conducted experiments to infer author personality from text corpora posted by the public on the Internet, such as on social network websites, blogs, and reviews of products or services, as well as other offline sources. They are able to use the information extracted from these texts to predict (both by classification and regression) the personality questionnaire scores of the authors. The focus of this area is on inferring author personality from text; the assumption is that personality predicts behavior important for myriad applications.

# Contents

# Chapter 1

# Introduction

## 1.1 Goals and Trends

So the goal of this research area is to infer personality traits from the writing of individuals, and to do so in an automated way, en masse. You are probably envisioning a million applications: the prediction is fuzzy but could revolutionize advertising, where any gain in prediction is very effective. A service could emerge, with guaranteed anonymity, for people to find out how others probably see them, or to assess their acquaintances given the data freely provided to them. This could augment the judgment of people who have difficulty perceiving others, or who have zero acquaintance with another, such as on dating websites. Corporate recruiters and governments could find those most likely to be evil terrorists or Wall Street psychopaths by examining writing samples. Another application, which someone has already started to explore, is in recommender systems: find internet reviews of products written by people similar to the searcher in terms of human personality. Similar matching occurs on dating websites, but the methods are primitive. Also predicting focused personality sub-facets may be more useful for the most targeted applications, but has not been done yet.

This area of research restricts the analysis of behavior to written text, with the end in mind of predicting the personality scores given by established self-assessment questionnaires. Such a restriction is not debilitating; the kind of personality-expressive human behavior studied the most is possibly linguistic in nature or is well described by participants in speech or writing.

## 1.2 Personality theory and assessment

A person's past behavior informs our intuition about predicted future behavior. For example when talking to a job reference, any reasonable person hearing the statements "Alice is creative and reliable, independent while supportive of others", or "Bob is tardy, slothful and loutish" accepts that there may be some unexpected exceptions to such descriptions of Alice or Bob, yet the generalizations serve the narrow purpose of differentiating them from other job applicants and of predicting future overall job performance given similar circumstances.

We do not delve deeply in to the literature on trait psychology because there are well accepted, stable models used consistently by researchers in this area. Suffice it to say, in 1936 Allport and Odbert citeallport1936trait found 17,953 different adjectives we use to describe each other's behavior. Since then, analyses by other methods have categorized those adjectives into general categories that are, to crucially varying degrees, portable between cultures and languages [31]. Over time trait psychologists have classified the adjectives into 3 to 20 different categories, and often several subcategories (called facets) in a hierarchical fashion. Currently they have settled on five major personality traits, forming the five dimensions in their Five Factor Model (FFM) of personality. Each of the factors or dimensions covers a broad range of behavior, so researchers sometimes evaluate the facets individually [17].

The five traits enumerated by the FFM are as follows. Factor I and Factor II are considered mainly interpersonal dimensions; they describe modes of interaction with others; also, the factors are in approximate ascending order of how much they account for individual differences. (Descriptions below adapted from [8, 17].)

**Factor I:** ***Extraversion***. The first of two highly interpersonal factors, the Extravert approaches the world with energy, enthusiasm, lack of inhibition and a sense of adventure; there is a sort of foraging for stimulation. Those communicating frequently, forcefully and glibly fall in this category. Aggressive Extraverts can quickly isolate themselves despite their craving for interaction; conversely more affiliative Extravert can accomplish much in concert with others. Low trait Extraversion is called Introversion; whereas the Ixtravert tends to express thoughts not yet in their final form ("thinking aloud"), the Introvert may be left behind in discussion while thinking through what if anything to say.

**Factor II:** ***Agreeableness***. This is what it sounds like, and a person can be high on Extraversion yet low on Agreeableness, or any other combination of the dimensions.

**Factor III:** ***Conscientiousness***. This factor describes effectiveness in performing prescribed rote, repeated activities. Also assiduous following of rules. However assessments on this trait do not generally ask questions about a person's ethics [3].

**Factor IV:** ***Neuroticism***. Also called Emotional Stability, reversing the measure. Neurotic individuals tend to perceive events negatively, and to be very sensitive to such events, to lack confidence, and apt to cease action in the face of difficulty or to refuse action in anticipation of obstacles. This trait, like all the others, is about more about behavior than about emotion: in this case, the tendency of emotions to underly stagnation in behavior.

**Factor V:** ***Openness***. This dimension is related to qualities that enlighten the mind, i.e. Openness to unfamiliar ideas and new experiences.

## 1.3 Natural language processing contribution

Stable individual differences in human speech and writing behaviors are included in some analysis of personality. Thus enters the usefulness of natural language processing techniques, which can extract a variety of features present in speech and writing. A series of obersvations in text that are stable over time but different between individuals may then correlate closely with a personality trait or traits. For example researchers have found correlations between the results of thoroughly tested personality assessment questionnaires and a variety of features such as function words (such as prepositions) and other English grammatical structures [1]. However when automating using NLP, the focus has been on word counts, $n$-grams, and sometimes part of speech. A few tools, which we enumerate as they appear in the literature, have arisen to facilitate the extraction of such features.

Even before application to human language, computing tools for lexical analysis have thrived for a long time, as they were needed for compilation of computer programs. The extension to natural language processing (analysis of human language) seems natural, though not at all trivial as statistical methods are needed to optimize the resolution of even

simple ambiguities (see [18] and [38]), and human languages do not have simple grammars capable of recognizing every possible sentence.

Work has been done to relate a variety of observations to personality, for example voice pitch and gesture. We address only written corpora for three reasons; the first is that there is a substantial, the existing mature body of work in this area, the second is that text corpora (which comprise written behavior) are the closest to the original theory of personality (which focused on linguistic descriptions). Our third reason is that there are obvious next steps that ought to be explored first in the context of the analysis of existing text corpora, but with natural applications arising in speech processing or production for example. In most cases the text was written by the author but in a few cases we have transcriptions of vocal conversations.

From such text features, participants in this research effort seek to predict the five factors or traits of human personality, which we have enumerated. They employ statistical learning techniques such as Support Vector Machines (SVM) and occasionally regression. After we have acquainted ourselves with their use in the literature, we will conduct a critical analysis of their use. Finally, we will conclude with a summary of untapped potential in this area.

## 1.4 Excluded discussions

We exclude in depth discussion of the various self-assessment personality questionnaires because the discussion would take us far afield. Also we do not participate in the debate on how many personality factors there should be, as we believe the participants in the debate seem overcommitted to their positions and unaware of the significant tradeoffs involved in increasing or decreasing the number of factors. We exclude a significant area of study of evidences of mood in text because the goal is prediction, and mood is less predictive than personality. Of course mood might give insight to possible features predictive of personality if considered longitudinally. Finally, we spend very little time discussing statistical prediction (classification and regression techniques) because they are rather self contained, the results are still poor, and there is so much to be done in terms of finding better training data. At one point we suggest abandoning or supplementing the use of non-optimal techniques in the search for classification and regression solutions in favor of locating the

solutions by brute force because that is feasible with very low $N$ and would tell us how well the faster, fancier methods are really doing in such cases.

# Chapter 2

# Literature

## 2.1 Pioneering studies

In the beginning of time (1999), there appeared a significant study, Pennebaker and King's early work in this area [27] which is important largely because it introduced both a popular and growing corpus (which Pennebaker allowed others to use) and the use of a tool to identify features significantly correlated personality. The tool, LIWC, is used commonly by researchers of the present topic. For our purposes, we will view the LIWC features as stylistic and sentiment based. Such features may be constructed from almost anything in a text that can be changed without directly modifying the essence of objects under discussion or the simple aspects of relationships between them such as ownership and location in time and space. LIWC employs a dictionary of over 4,000 common words, placing them in 70 categories that are seen as related to emotions or somehow self-expressive, such as positive emotion, negative emotion, sexuality, work, sleeping, and many others. More information about LIWC, which Pennebaker offered to the community, is available in [36].

Over the lifetime of their data gathering effort they administered a variety of self-report personality questionnaires to 841 university students, who also wrote essays for purposes of this study. They trace previous efforts at text analysis somewhat alike to their LIWC tool back to the classic sentiment analysis tool, the General Inquirer system [34] of the early 1960's. Ultimately they located 17 features predictive of the personality scores, such as for example that Neuroticism scores are significantly correlated with the use of the first person

singular pronoun and that the use of words entailing causation correlated negatively with Openness.

Helpfully, some attention was given to the issue of cross-situational behavior and how it effects personality assessment. A longstanding criticism of personality psychology is that the behavior it describes may be largely situational rather than an individual difference. One of their goals was to show that the text features they were interested in were fairly stable despite varying situations. The authors were interested in assuring the author's situation did not significantly effect the rate at which they expressed their personality through these text features. For this reason they introduced a phase of their study involving the writing of some additional small groups of study participants: briefly, a sample of 15 residential patients in a substance abuse treatment setting, 34 summer school students, and 40 randomly selected, highly published social psychologists. Ultimately they conclude that although usage varied between topics, writers are consistent in their use of word categories whenever writing about a given topic.

Thus began prediction of human personality from features extracted from their writings. The computer science community took some time to run with this idea, but eventually 3 computer scientists (S. Argamon, S. Dhawle, and M. Koppel) collaborated in a pioneering paper [1] with Pennebaker to do classification on the features his LIWC tool extracts. They were interested in exploring the possibility of *author profiling*, in this case personality prediction, by extracting features from short, informal texts of an author. Their hypothesis is that it is possible to extract text features and ultimately to predict the extrema of author personality in the dimensions of Neuroticism and Extraversion (as measured by the NEO-FFI Five-Factor Personality Inventory [9]). To investigate this idea, they analyzed a corpus consisting of various essays including stream-of-consciousness and deep self-analysis, for totals of 1157 and 1106 essays, respectively. They use machine learning techniques (SVM's, specifically the SMO learning algorithm) for the personality prediction; after training they conducted 10-fold cross-validation to check classification accuracy.

The paper starts with an explanation of some linguistic theory and how it guides feature selection; they advocate a linguistically motivated search for predictive features. They identify Systemic Function Grammar (SFG) as a useful framework for representing non-denotational, stylistic features in which they are interested. Such considerations paid

off: noting that function words are common but limited in their expressive power, they identify features that fall in the following categories according to their stylistic goals: expression, cohesion, assessment, and appraisal (which, beyond assessment, emphasizes the author's overall attitude towards that which is being assessed). This amounted to a thoughtful way of describing the features extracted by LIWC, which they employed in this study.

They make the helpful generalization that Neurotics tend to expression uncertainty with words such as: *perhaps, nobody, uses, try, except, getting, during, hardly*, whereas extroverts express certainty. Ultimately they enumerate the dozen or so words in every LIWC category that most affected their models. The selection of features beyond function words and successful training of classifiers provided an important stepping stone to later work. One omission is that they never mention the obvious possibility of predicting actual personality scores, which left this important step to later researchers. It is hard to imagine why they would reduce scalar personality scores to a classifier when they could have (I) predicted the actual scores, (I) done a more thorough and informative error analysis, and then at the end if their audience demands it introduce the classifications with the accuracies computed.

They use two SVM binary classifiers to divide the High (top third) scoring personalities from the rest, and divide the rest into middle and Low (bottom third). Unsurprisingly the choice of 3 regions instead of simple binary classes naturally results in lower accuracy than that of the later researchers who used a simple binary decision boundary; yet having 3 regions makes more sense than binary in this case because it avoids dividing a normally distributed sample.

As in the original Argamon work, Oberlander and Nowson sought to locate text features correlated to human personality and to demonstrate the possibility of performing reliable binary classification using their features [24]. Building on the work of others studying sentiment extraction on blogs, they examined 71 bloggers (34% male, average age 28.3 years). For personality assessment, they employed IPHP NEO-PI-R, revised for administration via the internet [4].

Their features are simply $n$-grams, but before extracting the $n$-grams, they identified proper nouns using WMatrix and replace them all with a common sequences of characters (the marker "$NP1$"). A major contribution of their work the use of well defined, principled manual feature selection as well as automated feature selection to decide what features to

train on, which deserves notice given that many people never report such efforts if they use them at all. They trained on five different sets of features: four manual selections and the fifth an automated selection, as follows.

- The first was a manual selection of $n$-gram features on a principled basis; they excluded uncommonly appearing $n$-gram features.
- Then they further narrowed the set to include only features with significant correlations ($p < 0.01$) with high and low personality test scores (excluding scorers in the approximate middle 1/3 of the sample) in at least one of the personality traits.
- Since they removed some of the texts, they again tested that a sufficient proportion of texts include each feature of the previous set of feature above. Then they excluded features that appeared in fewer than 50% of the texts corresponding to the authors who caused the feature to be included.
- Their last manually selected set consisted of the features that (a) meet the criterion in (II) above, but with $p < 0.001$, (b) were included in (III), and (c) exhibit a significant (p¡0.05) correlation with the trait score.
- Finally they selected a set of features in an automated way (WEKA, a BFS search using the CfsSubsetEval evaluator).

They mention that automatic feature selection gave them the best results, but the express a concern about overfitting (without justification), and suppress the results. Nevertheless the results with their manually selected features (none of the the manual feature sets always prevails above the others) are truly impressive, even more so considering that they did not use the SFG features from LIWC (from Argamon, LIWC) as so many do. For the binary classification task they correctly classify 75.5% of participants on the Extraversion scale and 83.6% on the Neuroticism dimension.

Their paper is difficult to follow, because they label everything they talk about, then produce the labels later in their writing without a word of explanation, producing excessive back-references. Frustratingly they provide only a few examples of the successful features they trained with. Their innovation is the use of $n$-grams and the identification and counting of proper nouns.

Nowson, a collaborator of Oberlander's who was involved in an earlier project [24], built on that work [22]. They called upon their original classifiers, which they had trained in the earlier project, on a carefully collected set of $< 100$ blogs and their authors' personality scores. They used those classifiers to predict the personalities of a much larger set of bloggers; their subjects are now 1672 bloggers; the average words per blog was about 3000. They went beyond other efforts in this survey attempting to test a classifier on a different corpus than the original training corpus: they were able to check the accuracy of the classification due to the benefit of access to the personality scores of the authors of the large corpus. In addition they offer a secondary, but weakly supported hypothesis that bloggers are unusually high on the Openness trait, unlike what they cite as the usual perception that bloggers are Extraverted, "Exhibitionist Narcissists". Clearly the latter hypothesis is problematic if personality is essentially regarded as the perceptions of others about a person's behavior, whether on a blog or elsewhere.

Amusingly the personality scores that they employ consist of a preexisting internet meme that functioned as a coarse measure of personality, evidently copying a few questions from a personality questionnaire. They were able to trace the questionnaire respondees back to their blogs, thus permitting them to study the content on their blogs. While research on human subjects should not be done primarily with an eye to convenience, perhaps it is appropriate to weigh the interests of obtaining such a large sample of the population that might be quite impractical otherwise. It is not a trivial task to obtain personality questionnaire responses from individuals.

Their features are $n$-grams in the blogs. When they extract the $n$-grams, they employ some hacks to avoid non-author text such as quotations and memes. Their earlier project [24] convinced them that simply naive Bayes actually outperformed SVM's for classification, so they employ the former. The best classification accuracy attained was 66.4%, for Conscientiousness.

The group has continued to churn out studies of this nature such as [13], published in 2009 by the Association for the Advancement of Artificial Intelligence. That paper repeats what they have done before, but with the discussion of a mysterious flow of causation from an individual's supposed essence as Neurotic, Extravert or Open, to the way that individual

writes; the former is described as a *motivation*. Also the word *desire* is used, but never defined. This philosophical discussion is tangential to our interests here.

Mairesse, et al. [20] built an ambitious automated personality recognizing tool, Personality Recognizer; it uses LIWC and MRC to extract linguistic features (principally relative frequencies of words falling in various categories such as emotion, perception, cognition, communication), but also including some others such as punctuation, word length, number of words in a sentence.) Then they train the model on known personality scores (via self and observer reports) of 2575 students. The usefulness of their model in analyzing dialogue or other written artifacts besides the sorts of essays they worked with is worth investigating further; in fact one study we cover later did just that. Users are also allowed to train the tool themselves on their own data.

In evaluating the usefulness of their binary classifiers, they chose to use ultimate classification accuracy (percent of respondents assigned by the classifier to the same class they would be assigned to by the self and observer personality assessments). Their evaluation seems prone to underestimate how well the classifier approximates the underlying distribution. The author thinks that non-extreme scores are"just noise". Also the author of the paper affirms that although it made their task easier in the presence of unbalanced data, it is debatable that the best thing to do was to place their sample personality results into two bins of the same size. (Private correspondence, F. Mairesse.) That decision forces the median score to be the divider between the two classes, which is difficult to justify.

It is helpful that they revealed a much broader horizon of new features that are predictive of personality, as well as observing that some of their features observed at the extremes of personality happen to give more accurate classifications than their full set of features, though they do not hypothesize about why that is. They carefully place their work within the context of others' work, and the scope of the their literature review is helpful to the uninitiated reader. Perhaps it is for these reasons that the study is heavily cited by others.

Although there is much redeeming value in this study, the analysis could be improved: for example the decision to impose binary classifiers in a way that shockingly discards scalar predictions in favor of forcing participants into one of two classes. That is unjustified. The data is normally distributed, there is one underlying category for each dimension.

Constructed that way, this classifier is like a police radar gun that conveniently rounds a motorist's speed to 0 or 100 MPH before displaying the speed. They claim that "the midrange ratings are just noise," but they are not. Most people are in the middle. They cite all of this to explain the use of accuracy as their error measure. With the idea that the middle scorers are certainly not noise, as seen in their regression analysis, they could have used a more helpful error measure, such as the quantity the classifier actually seeks to minimize! Speaking of correctness, they arbitrarily, for their convenience in analysis, determined the cutoff for the accuracy measure by splitting their data into two equal sized bins, which is rather peculiar. This is like concluding that a someone has cancer just because they have a slightly higher score on a test than 50% of those who happened to take the test that day.

Finally, we question the need to run non-optimal learning algorithms for prediction. If binary classification is absolutely necessary, why not actually locate the optimal label assignments by checking all $\binom{n}{n/2}$ of the ways to choose them, rather than running non-optimal learning algorithms? Since $n$ is small in their study and they give no indication that their work is intended to be a pilot study whose methods will later be employed with millions of observations, such a method should work well. Better yet, they could do both, and the known correct solution that takes more compute time would serve as a correctness benchmark for the fast non-optimal methods. Finally, like many others, they seem unconcerned that some of the algorithms give drastically different results depending on the initial conditions. They blindly run the algorithms only once instead of many times while varying the parameters in search of a better result. In the end, their main contribution is to provide a tool available for practitioners to use for classification in prototypes of their applications.

Roshchina et al. has tested the Mairesse tool as trained by Mairesse. They are apparently the only ones to do so so far; they designed their recommender system [30] to present to the user hotel reviews written by people with similar personalities, a somewhat novel approach introducing psychology to opinion mining [25].

The stated purpose was to provide better travel service recommendations by pairing individuals with travel services liked by those with similar personalities. Rather than testing personalities conventionally by a questionnaire, they hope to assess personality from reviews of travel services written by system users.

We are enthusiastic about people publishing applications, and also about their idea of classifying people according to some measure of personality similarity over all five personality dimensions. The idea of identifying clusters of users of similar personality is appealing, but they would do well to explain what their clusters represent, present and justify their objective function, and examine the psychological literature on the subject of clustering by personality (there is a universe of papers on such things). They propose to use the k-means algorithm for this task, without justifying the decision and without addressing its well known sensitivity to initial conditions. The paper gives the murky impression that the researchers actually attempted implementing this clustering feature, but they give no results.

For their study, they select just 15 people (from their database of 1030 users) who wrote more than 30 reviews on the TripAdvisor website. The authors clearly believe that the tool actually ran the M5 regression tree algorithm on their data. That is impossible: the Mairesse tool is trained with supervised methods to which one furnishes the actual author personality measurements as well as the text for analysis. The authors apparently never conducted such a personality assessment. Then, the authors offer their *hypothesis*: that the best algorithm is that which produces personality scores that differ the least amongst various works by the same author. Although it seems like a good idea to consider an alternate analysis to that provided by the creators of the tool they are using, they make things much worse by offering this hypothesis. They are trying to find out what models give the best inferred personality scores, but they never measured those scores so as to compare. One is left wondering why they did not accept the conclusion of the authors of their tool as to which model is best for each personality trait, as Mairesse et al. at least collected personality measurements.

Instead, they assume that what they are trying to prove is true, and proceed to compute the personality scores predicted by the various Personality Recognizer models for their 15 participants. Finally, they select the model that best matches this assumption, and announce that they will use that model from now on. In the end, although they have 14,000 reviews produced by 1,030 reviewers, they run the Mairesse tool in 4 different modes on a sampling of just 15 of the reviewers to infer personality scores. Although the authors provide an interesting idea for an application, their presentation makes it evident that they were unable to dedicate much time or expertise to the project thus far.

Besides the paucity of their sample, their work could have been improved by imitating the prior work of Picazo-Vela, et al. [28], who investigated the possibility that there is a unique user personality which dominates online review-writing. In that work, the authors give a robust defense of their use of students as study participants by establishing (including citations to various studies) them as a key group of online shoppers. Yet their population could have a strong self-selection bias. Online and telephone polls are two well known examples drawing biased populations, so the online reviewer population seems likely to exhibit strong bias as well.

Unsurprisingly they found that Neurotics and the Conscientious were more likely to write a review. Despite its issues this paper is a rare example of a well documented application; surely more like it are to come.

## 2.2 POS breakthrough

In [23] Oberlander (again) and Gill investigated a similar set of features with a different corpus and a different personality assessment, a 3-factor personality test, EPQ-R, covering Extraversion, Neuroticism, and Psychoticism (which is commonly believed to inversely correlate with both Agreeableness and Conscientiousness). The presentation of this study is better than their other one, and they report the effective features extensively in several large tables. They extracted features present in emails with the personalities of their 105 distinct authors. Their sampling technique, however, effectively resulted in using only the writings of 71 of their authors. The emails averaged 309 words each. The emails were not actually sent to anyone, and the authors were prompted to write the text for the purpose of the study. Participants wrote two fake emails: one describing what happened to them in the last week, and the other describing their plans for the next week. While such a setting allowed them to control the situation and topics, apparently they had no great desire to do so, which makes us wonder why they did not simply ask authors to submit real emails written recently. These topics they assigned are still very much subject to whatever might be happening at the moment in the lives of the authors. To control the topic, they might have adopted similar practices as a later study we cover here [19] which required participants to write about a particular documentary that they all watched.

The study was designed around sub-samples of their 105 authors, each consisting of those texts whose authors scored at an extreme on one, and only one, trait on the personality test, as well as a control group of texts from authors whose personality scores did not fall in the extremes (outside plus or minus one standard deviation) for any of the traits. This sampling technique enabled them to isolate factors present in their samples that are indicative of individual traits measured by the personality test. Unfortunately, after sampling they were left with only about 20 authors for each personality trait.

The methods they employ go beyond their previous study that employed unigram methods using LIWC. They first introduce features consisting of the POS colocations (mostly 2-grams). Then they proceed to their lexical $n$-gram analysis.[1] Besides the crucial contribution of adding POS features, this time the group offers copious details on their findings, enumerating significant features in several tables. Obviously such results are more useful to a practitioner hoping to predict personality scores.

Another relatively early group, Estival et al. [11] in 2007 extracted text features from emails with the goal of predicting Big Five personality traits as well as the following demographic aspects of an author: gender, age, geographic origin, education and native language. Their participants were from the United States, UK, Australia, New Zealand, and Egypt, and were 1033 in number (after some qualifying criteria such as submitting a minimum of 1000 words were applied). The authors created and trained their own system to annotate each line of author text as signature, reply, quote, and advertisement with accuracy reported at 88.16%. Besides building features from the latter annotations, they assembled a staggering number of English language features that included word case and length, various function words, named entities, and POS features. In all, they report having considered 689 features. They never tell us how they extracted most of the language features (e.g. how the POS tagging was done), although they mention that because most named entity recognizers are based on the news, they had to create their own named entity recognizer using unspecified publicly available resources. It is unclear whether their POS features were simply unigrams or they included larger POS $n$-grams as did the study of Luyckx and Daelemans, below.

[1] an $n$-gram is an ordered sequence of words, so "dog house" and "house dog" are two different $n$-grams with $n = 2$.

Of course when confronted with so many features, feature selection becomes very important. This study presents us with some examples of issues to be aware of in this kind of research. In their training attempt (which employed a variety of ML algorithms in the WEKA, the tool they used), the performance of the resulting binary classifiers on test data is not impressive for human personality; their best prediction of a personality trait was of Extraversion, at 56.73% (although they did better for their demographic predictions). They never mention checking the correlations (and their significance) between individual features and the variables they are trying to predict; if the automated feature selection methods they use offer such an analysis, they never discuss it. Although they report using automated feature selection algorithms, frustratingly they do not enumerate or even say how many features they ultimately used for learning.

It is nice to see that they made an effort to investigate a wide population using real historical data (the emails were not artificially created by the participants as we saw in another study) with no strong constraints on the text being produced. However those lofty goals may have made the task near impossible. Combining a wide variety of probable topics (they include work and home email) with no restriction on the subject might have made matters even worse. They attribute some of their issues to the small size of their corpus relative to that of Oberlander and Nowson [24] and the fact that they studied blogs rather than emails. The explanation is inadequate; plenty of other studies with comparable sample sizes did better, some of which we cover here e.g. [20] and Nowson later published much better results on the blog texts [22].

The authors offer their corpus for others to study. Incidentally the same authors have also studied Arabic emails [12]. Arabic language NLP presents unique challenges; for example one does not usually see the Arabic word for the present tense of the verb “to be” (David Bean, personal communication, November 16, 2012).

Kim Luyckx and her collaborator Walter Daelemans, both at the university of Antwerp [19] did certainly extract POS $n$-grams as features They note the need to progress to analyzing linguistic features that go beyond lexical unigrams or $n$-grams, and they do so in the context of validating their hypothesis, which is that it is possible to infer the personality of authors from such features. They suspect that syntactic features would be more predictive because they are not controlled so consciously as the use of individual words [33].

Oddly they chose to use as their personality assessment the Myers-Briggs Type Indicator (MBTI), an invalid assessment tool marketed aggressively by consultants to the gullible or unaware, such as corporate clients seeking to refine their techniques of exploiting people. Certainly measuring personality is a controversial topic; for example the validity, orthogonality and other aspects of the Five Factor personality model were challenged by Jack Block to the very end although he used it himself. His last article, which he fired off in 2010 [3], raises concerns such as its empirical validity and the orthogonality of the traits. However this particular instrument, the MBTI has no grounding in personality psychology has been discredited and is dismissed widely by those who formally investigate human personality due to its inferior predictive power, the lack of a basis for its Type constructs, and consistency issues. It forces individuals into Types or binary classes, those falling near the decision boundary can get dramatically different results when taking the test repeatedly. We wish that those who chose to perform binary classification on predicted personality scores might reconsider doing so due to that last point, so as to avoid duplicating the same mistake. Despite all these complaints, the MBTI, if viewed in the scalar dimensions before conversion to Type, does have some well known strong correlations with the Five Factor personality model in every dimension except Neuroticism.

Besides the issue of personality assessment, the study is quite well designed and presented; they had 145 university students write an essay about a documentary they watched on Artificial Life. The essays average 1,400 words. This gave very specific focus to the writing, thus removing many potentially confounding factors while reducing the generality of the study. (Free writing exercises result in greater variety because people may write about different topics on the basis of daily changing moods and circumstances. When the topics vary, the words they use change, as we saw in the CITE study). The features they extract include simple lexical $n$-grams and Part of Speech (POS) $n$-grams which are based on the parts of speech present in text. The POS tags were inferred by a tool, in this case MBSP, which parses the text and gives the POS of the words.

To extract the syntactic features, they employ Memory-Based Shallow Parsing (MBSP) [10]. They use MBL for training, but without explanation of that choice, and no presentation of the feature selection, other than that they present 3 batches of results, one for each of the 3 types of features above. The results were good: they were able to predict Extraversion with

an accuracy of 65.5% with lexical 3-grams, which as they note is better than a number of other studies such as [1], [20] and [22], which we cover in this review. Their prediction of a factor related to Openness was from an unspecified set of POS 3-grams, 62% accurate, like that of [20]. What a pity that subsequent researchers evidence ignorance of their work. One can only imagine how much more accurate their predictions would be along with the POS $n$-grams.

The main contribution of this paper is to demonstrate that it is possible use POS $n$-grams to predict personality scores on a third population. Unfortunately they do not report what specific $n$-grams they used, or the correlation coefficients.

## 2.3 Validation with human judges

In addition to self-assessed personality, a few researchers incorporate human judged personality in their studies. Although not usually in the context of automated analyses of text corpora, there exists a body of work consisting of human-judged personality assessments correlated with expressions of text and other artifacts provided by users of such systems as social networking websites and cell phone networks. These features are sometimes viewed as validating the human judges or showing what the human judges are capable of when somewhat starved for information, while conversely at other times, assuming the accuracy of the observers, viewed as validating the text features. Including human (observer) judged personality in addition to the self-assessments is laudable because it allows one to examine such questions.

We will not extensively review such efforts, but a key example of such a study that should be noticed by computer scientists includes that of Buffardi and Campbell, who examined the possibility of a link between trait narcissism and behavior on social networking websites as evidenced by text and other artifacts [5]. It is not our purpose to give a full account of narcissism here. Briefly, trait narcissism, though not fully described by any one of the traits in the FFM, can be described as a cluster of human personality subtraits in their extremes, involving an exaggerated positive sense of one's abilities, uniqueness, and entitlement, together with a will to assert oneself accordingly. Those around the narcissist report being dissatisfied with their relationship with him (they are predominantly male). Since those with

high trait narcissism show shallowness and low commitment to others, yet often maintain a plethora of contacts and brief goal oriented social interactions, the authors were confident of locating such individuals on social network websites.

The authors obtained the consent of 129 undergraduates to present anonymized versions of the content from their Facebook pages. The authors administered the NPI, a questionnaire measuring trait narcissism, to the participants. Panels of judges, also undergraduates, gave their impressions of various aspects of the content including text and photos. Without knowing the NPI results, the judges assessed the content on a scale devised by the authors, ultimately resulting in a *narcissistic impression* composite score.

Higher narcissism scores were positively correlated with self-promoting quotes and other information, as well as with the sheer quantity of social interaction. Surprisingly they found narcissism scores to be negatively correlated with entertaining quotes, which differs from behavior observed in studies of direct social interactions. Of course the study begs to be developed and eventually automated if possible. More work has been done automating the analysis of personal photos for narcissism than has been done in analyzing author text for narcissism. Progress in this area seems urgent as narcissism is seen as being somewhat antisocial, on a malignant continuum with psychopathy.

## 2.4 LIWC studies of personal communications

Current studies focus on personal communication such as SMS text messages, social networks, and email. Surprisingly current (2010-2013) research employs only lexical $n$-grams and LIWC features that categorize words according to sentiment and style.

In 2012 F. Celli and L. Rossi [6] assembled a classifier of Twitter users according to personality by employing features associated with Neuroticism as published by the creators of the Mairesse tool [20], as well as a few social networking features such as the number of followers a Twitter user has. They extracted textual features from 200,000 Twitter posts of 13,000 users between December 25 and 28th. Although they provide some discussion of their classifier, it is unclear how they put it together, beyond that they apparently copied the correlations published in [20], which they cite [20, Table 2]. They do report that they

incremented their Neuroticism score on each observance of a positively correlated feature, and decremented in the presence of a negatively correlated feature. This seems presumptuous as the presence of individual instances of two different features such as an explanation point and a word of positive emotional affect (or even two instances of an individual feature) might not indicate twice as much Neuroticism as a single feature does. That practice neglects the possibility that two different features might share so much mutual information that their coincidence tells us nothing more than the presence of only one of them.

They noted some social network features associated with Neurotic participants. However since they did not measure the personalities of the Twitter users (e.g. via questionnaires), they were unable to publish helpful correlations. Their Table 1 presents a very nice summary of human personality, with descriptive adjectives.

Also in 2012, another better structured study appeared from Sumner et al. [35], of 2,927 Twitter users, 876 residing in Great Britain, 609 in the United States, and 1,442 in 87 other countries. In addition to measuring personality by the Ten Item Personality Inventory (TIPI), a they also employed the SD3 (Short Dark Triad) to measure the so-called Dark Triad of human personality. Both of these measures are written self-report questionnaires.

Here we will not discuss the Dark Triad at length, but they describe it as incorporating three traits: Narcissism, Machiavellianism, and Psychopathy. Note that the SD3 measures these traits as dimensions, rather than labeling anyone a Narcissist or a Psychopath. It is clearly possible to describe much of what is in these three traits in terms of particular sub-traits of the Big Five, and there are some empirical issues with the work on the Dark Triad, such as the obvious ones related to expecting deceptive participants (such as the Machiavellians in the sample) to give honest responses. Nevertheless it provides a mechanism for focusing on some behaviors that society at large rejects.

Their primary goal was to identify features in text that are correlated with personality. They extracted many features using LIWC and filled an entire page with a table depicting their correlations with personality traits, many of which were statistically significant. The significant features are generally as expected, for example Agreeableness is negatively correlated with negations such as no, not, and never. When they provide correlations between text features and personality scores, with the sole exception of [35], they use Pearson

correlations, instead of Spearman correlations, which are less sensitive to extreme values. For the Narcissism trait they included a feature incorporating words such as buddy and friend was strongly correlated with a relatively large effect size (the Spearman correlation is 0.073). Various punctuation of special social function in Twitter was also indicative of Narcissism.

Also they sought to classify participants according to personality traits, and to predict personality scores. Their classification results are from a public competition they held, offering their data to others for analysis. The ad hoc nature of these experiments and little mention of how the quality of the results was assured presents an issue. Their contest participants attempted a large variety of techniques to address the classification problem. They tested a whole host of machine learning techniques. No mention is made of feature reduction or selection, which is surely necessary given the huge number of features they offer.

Ultimately they assessed two distinct sets of classifiers: those dividing the the personality dimensions at the middle, and those seeking to predict the top and bottom 10%. In the best cases, the former were little better than chance, and the latter failed to predict true positives, identifying only 2 of the 125 individuals in the top 10%. Although they clearly employed the predicted personality scores in the classification task, no results are given for the specific task of predicting scores.

One commendable practice is their effort in the direction of providing an error analysis; they supply a variety of values such as the Area Under Curve, true positive rate, true negative rate (which are particularly relevant given their emphasis on the top 10% category), and % accuracy. This analysis however omits their performance on the personality score prediction task. At a minimum the RMSE on this task would have been of interest; the omission may be explained by the misgivings they actually express about other researchers who report the RMSE. In sum their concerns, which amount to how extremes are treated, seem to beg the questions of (I) whether their goal is really to minimize the error (which is up to them) and (II) whether the data is not really normally distributed (a worthy research question), in which case one would be justified in seeking a better error estimate upon identifying a truer distribution.

Another refreshing aspect of their paper is the ethical discussion. They highlight the need for restraint in employing poor classification tools to make decisions about people. They also cogently note that LIWC is inadequate to extract information from many of the words used by Twitter users, who are severely limited in the length of their messages; they sensibly suggest future research of the use of language in social media. It is a pity that they, like everyone else right now, seem blissfully unaware of the POS $n$-grams introduced a few years prior by a few researchers at various corners of the globe. Even so, they could have done more with lexical unigrams, for example, it is commonly known that Extraverts freely use invented or malformed words; they could have counted the incidence of such and investigated them as a single feature.

Curiously they note one other potential weakness in their work: their sampling of Twitter users are predominately followers of British celebrity Stephen Fry and US skateboarder Tony Hawk. They offer citations to other studies and claim that the proper analysis would reveal that the effect of selection bias is negligible.

Clearly when attempting to infer personality scores, or to classify people according to personality, it is useful to incorporate non-linguistic features when available. Indeed Chittaranjan et al. [7] combined some well known textual features with data available on the cell phones of participants in their study. Their data is from 83 cell phone users (specifically the Nokia N95) in Lausanne, Switzerland, collected over a period of 8 months in 2009-2010. Although Lausanne is a French-speaking area, it is unclear what language was predominantly used by the participants.

Besides a variety of social features and some very context specific features such as Camera and software use, the authors identified the following textual features as significantly correlated with measured personality: average word length, median word length, number of messages (sent), number of messages (received), and number of interlocutors. They created a binary classifier dividing each personality dimension into two classes; SVN's were used for training. Amusingly they suggest investigating accelerometer measurements (perhaps Extraverts bounce around more?).

The results of this approach were good, and commend the author's practice of combining linguistic features with whatever else is available. Among the binary classifiers their

maximum accuracy was with Extraversion (75.9%). Of course with $n = 83$ there is a great possibility of over fitting the data. There were also very judicious in their selection of features, which is important in training since the task becomes far more difficult as the number of dimensions increases. One wishes that after doing such a thorough job, they would not do an error analysis (beyond providing the % accuracy observed during validation).

We have seen that although many of these groups ultimately present a classifier, their primary contributions are finding new features correlated with human personality. In general the classifiers presented in these papers should not be emulated as models; rather one should use the most significant features available, both textual and non-textual, and then build the classifier. When it is impossible to obtain personality scores for the sampled authors, caution should be used when simply copying the correlations published by others studying a different population or when employing tools built thereon. Furthermore, if the latter approach is used, clearly any new "bootstrapped" features discovered should be tested on a sample whose personality is measured.

The main contribution of this paper is not any particular feature set or classifier. The contribution is the method: combining two different kinds of features sets to infer personality. It is surprising that they attempt classification with only $n = 80$, which is rather paltry especially because they have to divide the data into sub-samples for cross validation and testing; in this case we will pass over those results.

A unique 2012 paper from Bai et al. [2] presents a study of Chinese participants in Renren, a social networking website popular in China. The researchers obtained 209 participants (137 male, 72 female), students of Graduate University of Chinese Academy of Sciences (GUCAS). Their essential hypothesis is that there exist significant correlations between their participants' (I) linguistic and social factors present in Renren, a social networking website popular in China and (II) personality as measured by well accepted self-report assessments. The paper would have benefited from further editing from a native English speaker, but nevertheless it is of interest because of its non-English language focus. It is safe to assume that the participants favor some mutually intelligible Sino-Tibetan language(s).

Their task was a daunting one, as Chinese language tools are scarce. As is often the case in Asian personality studies, the researchers found themselves employing a written self-report personality assessment that was originally created by studying Western subjects and language, a practice accompanied by many well documented issues related to differences in culture and language.

They were able to gather user data from Renren in a similar way as is possible with Facebook: upon obtaining permission from a user, one is able to query the system for virtually everything the user has put on the site. As their features, the authors first created their own rather blunt tool to classify entire texts (presumably Chinese) by overall affect, placing them in categories which they label in English as follows: angry, funny, surprised and moving. They state that they previously trained their classifier using Naive Bayesian methods (with no stated justification) on what they term an emotion dictionary, and apparently some texts, but give few details on this portion of the work. Finally, they include pronouns, emoticons, and volume of text as features.

Besides the obligatory training of a classifier, they indicate that emoticons were correlated with Agreeableness and Extraversion, but do not give the values; likewise what they term *angry blogs* (messages characterized by words in their *angry* category) with Neuroticism. Openness is correlated with the volume of writing.

Surprisingly we have not seen very many Facebook studies correlating text features with five factor personality; recall that the Facebook study by Buffardi and Capbell [5] was not automated in any way. Fortunately Golbeck, et al. have spearheaded the effort by reporting a number of such features; they studied 167 Facebook users of average age offering a rather sparse corpus averaging 42.6 words per author, when combining words from static fields in the user profiles as well as status updates [15]. They offer no explanation of why it might be that their sample was so sparse in the word count.

Their hypothesis is that social media profiles can predict personality traits. They detail 13 LIWC features significantly correlated ($p < 0.05$) with various personality traits; only one is reported for Openness (money words such as “audit”, “cash”, “owe” are negatively correlated). Astonishingly, and to their own surprise, they report that the last name length in characters is positively correlated with Neuroticism (the inverse of emotional stability).

Although they report average personality test scores, they do not offer a comparison of the averages with those of any other population.

Although the significant correlations make us optimistic (a measured optimism due to the rather sparse average word counts) that their features have some predictive power rendering them worthy of consideration by anyone trying to predict personality scores of Facebook users, their attempt at personality score prediction did not go so well. They performed regression using M5 and GP, but it is dangerous to repeat their results due to the methodological issues rampant in this avenue of inquiry, which they painfully repeat here; in this case they suffer from exactly the same deficiencies explained below, related to their otherwise very compelling Twitter analysis. Unlike the latter they also report the regression correlation coefficient, which does not remedy that problem. The contributions of this work seems worthy of consideration; we only hope that the accepted practice of this unnecessary clawing at statistics tools that everyone attempts would either cease or be done so as to make a lasting contribution. Perhaps reviewers at the publishing venues will recognize the problem and start encouraging an allocation of resources to more productive ends.

Again it seems good to welcome such studies due to the sometimes surprising features they offer for further investigation; the sample size of 167 participants lends credibility to the features they offer as candidates for prediction of personality scores.

Golbeck et al. continued the project by studying 50 Twitter users who each provided an average of 2000 words through the available tweets, and took the BFI to measure their personalities [14]. Their hypothesis is unusual; they expected to predict actual personality scores with substantial improvements over a default baseline. They extracted LIWC features as previously, then go beyond their earlier study by adding some features from some tools unconventional in this area of study but widely known elsewhere: MRC, which provides a variety of linguistic measures, and General Inquirer [34], the classic sentiment analysis tool (which ultimately offered a feature correlated with Openness). They supplement the many text features offered by these tools with a few additional features unique to Twitter such as links per tweet and number of Hashtags. Finally, they offer a large table outlining the significant personality-correlated text features they identified. They include details on punctuation, for example reporting question marks as significantly correlated with Extraversion, and commas as negatively correlated with Conscientiousness.

Ultimately the results were not surprising but reporting the features is a helpful contribution for those hoping to build a classifier or to investigate promising features. Although they lay claim to an attempt on classification, they present no results. Instead they report the MAE on a regression task (the normalized MAE varies from 0.12 to 0.18), which is nearly useless because they never compare that to what they would get with some default baseline. Their example in trying a variety of tools for feature extraction is commendable, although it is surprising to see that since 1999, it has not occurred to anyone to take the trouble of thinking about text features beyond those extracted by the available tools, perhaps building their own. This task does not have to be difficult; perhaps the tools are often too coarse, the categories too broad for the purposes of personality prediction. The energy spent on the summarizing the outputs of classification or regression algorithms (these authors, as some others, repeatedly conflate classification and regression) could surely be productively redirected to the investigation of new features.

Tal Yarkoni conducted a gem of a study in 2010 [37] wherein he had the luxury of examining not only the broad word categories in LIWC as correlates of personality scores but also many individual words. Doing so was feasible because his population sample was relatively large: of the 694 participants (bloggers), 407 had blogs of at least 50,000 words, which he used for that purpose. Such features are surely useful for something, but learning is impractical in the presence of hundreds of features. In this sense he does not move beyond Oberlander and Gill's study [23], introduced 4 years prior. Examining POS $n$-grams as they did might have taken them in a more productive direction. One wonders whether he is aware that a practitioner will need to combine his features somehow in order to predict personality scores, and the fewer, the better. To his credit, he reflects on the significant features found by statistical methods, and infers when and why certain lexical unigrams exhibit an unexpected correlation with a trait. It would be helpful to go a step further by exploring how to identify such situations and exclude the unigrams that do not fit.

Participants chose which of IPIP-50, IPIP-300 self-report personality assessments to complete. His goal was to explore the many . He warns that because he presents many so features significantly correlated ($p < 0.05$) with the personality scores, there is an elevated risk of Type I error (i.e. falsely affirming that a correlation represents a relationship between a feature and a personality score for a trait). What is behind this is that since each of his

features has about a $0.05$ probability of being meaningless, for $k$ features, assuming their independence, the probability that one or more of them is meaningless is estimated by $1 - 0.95^k$, which grows as $k$ gets larger. He tries to address this by looking to combined features, however he risks trivializing the issue by forgetting that $p$ is likely to be larger for the combined features, which works against him. (In fact this plays out in the tables in the appendix; he reports far more single-word features (hundreds) than there are multi-word categories with $p < 0.001$.) It is not hard to find some other glib statements about an undefined phenomenon *chance*, but at least he paid some attention to selecting features of adequate incidence; he reports the minimum incidence of each batch of features, but never tells us whether that is a per-participant frequency or a count for the entire corpus.

Really fascinating, and we celebrate it, is his analysis of personality facets (each of the 5 personality traits has 6 facets, which are like sub-personality traits, for a total of 30 facets). Such facets would obviously be more descriptive and predictive (sure enough, there is a famous study confirming this [26], eagerly replicated by dozens of other groups). When there is enough data to support the multiplicity of facets, it would be commendable to seek to predict them. This is especially true when some specific trait is sought, e.g. a marketer trying to find people who are body conscious (his cardinal example involves that facet; we prefer to leave the specifics to the very thorough reader).

In 2011 Holtgraves arrived on the scene stating the goal of investigating the use of text messaging as a function of personality[16]. 224 volunteers (university students) participated in an experiment called Cell Phone Research. They were asked to bring their cell phones. First they took a five factor personality test, and then provided some social information about those to whom they sent their 20 most recent texts such as ratings of their affiliation, gender, relative age, and duration of acquaintance. Strangely they asked participants to reduce report the answers to those on 5-point Likert scales. Although LIWC includes emoticons in its features, it cannot handle non-standard English. Therefore they augment LIWC features with their own exploration of the idiosyncratic lexicon used by this population in their text messages, aided by a lexicon of common SMS text abbreviations available on webopedia.com. For instance users tend to use slang shorten words by dropping letters or substituting numbers for sounds. Their examples include dunno, doin instead of doing, L8 instead of late. They also note that sometimes participants reverse the prevailing practice of

shorting words, and instead extend a word, e.g. bitchhhhhhhhhhh. Such extensions were most common among females and were correlated with Extraversion.

They offer a valuable contribution by including an extensive description of their features, including their relative frequency of appearance. They supply a table of 18 features correlated with three personality traits (Extraversion, Neuroticism, and Agreeableness; they exclude the other two traits because they did not have enough significant results), and a few descriptions of the most interesting and significant features. The use of acronyms and emoticons correlated with Neuroticism; Extraversion and Agreeableness were both negatively correlated with negative emotion words, and positively with Neuroticism, although the expected correlations with positive emotion words did not emerge. Extraverts use more personal pronouns and fewer impersonal pronouns. Disagreeable people more frequently use words related to health. In general these results are as one might expect given prior research.

No mention was made of avoiding features that are present only for a few participants. Some of the significant features were present at a rate of 1-2%. Among a sample of 224 participants, it seems possible that only a few heavy users of a feature might be responsible for its appearance.

Also commendable was their practice of investigating the above described non-standard words present in their corpus, rather than merely relying on whatever words their chosen tool can identify. Thankfully they do not belabor the reader with a poorly constructed classification project, another practice worthy of imitation.

In quite an unusual study, Yee et al. present a study [39] of 76 university students (undergraduate and graduate), 67.1% male, whom they immersed in the online game Second Life. The students were new to the game and were given an initial 1,000 Linden dollars to spend within the gaming environment as they wished over a period of 6 weeks. A tool gathered data on the activities of the participants within the Second Life world.

Second Life is a virtual reality game that is evidently quite immersive. It has existed since 2003, and on average 40K to 50K users [32] are logged in at any given time. Some have compared the virtual environment to public park. People create intellectual property (such as buildings of their own design) within the system, and interact with other users. There is a

concept of property ownership in the game, where people own plots of land and other objects. The social interactions vary broadly; individuals have even met and begun romantic affairs that eventually resulted in real life meetings.

The goal of the study was to find correlation between user personality and variety of linguistic factors within the text chat feature of the game, as well other behavior. They used LIWC to extract the text features, which averaged only 4 words per chat message. Although they observe that non-standard English words and grammar are employed, they made no effort to extract features from instances of such, as we see in some of the other studies. When compared to other projects, they found fewer (11) text features with significant ($p < 0.05$) correlations with one or more personality traits. The correlations are unsurprising, for example Extraversion and words with more than 6 letters, Extraversion and swear words (negative), and Conscientiousness and tentative words. Conscientiousness is also correlated with the use of words with more than 6 letters, but they may be explained by an unwillingness to use non-standard English; the authors note that some users substitute for example *rly* for *really*.

The authors cogently note that they may have found fewer correlations due to the broad variety of settings in which their participants were immersed, none of which were previously familiar. We hope that others will follow their example constructing studies of participants interacting with each other, as there may be additional clues to extract from the language due, that would be absent in an essay or blog entry.

Another recent study of 142 Twitter users by Qiu et al. [29] also employed LIWC to extract linguistic features for exploration of correlations with author personality. These qualifying participants all had more than 20 and less than 1000 tweets during a predetermined 30-day interval; the average number of tweets was 204.7 consisting of an average of 11.61 words each (after removal of extraneous content). To facilitate LIWC analysis, they replaced emoticons with markers indicating positive or negative emotion. As is usual, they administered a written self-report personality assessment, in this case the BFI, to the participants. Also, they formed a group of human judges to assess author personality by skimming through the tweets and taking the BFI on the behalf of each participant. This enabled them to determine whether it is possible to make zero-acquaintance judgments about

personality on the basis of microblogs (in this case tweets). In this fashion, the researchers were left with 2 sets of BFI scores for each user.

After comparing those two sets of scores, they concluded that the human judges were able to predict Agreeableness and Neuroticism. However they do not explain why they assumed the self-reported personality scores were more accurate assessments of personality than were the human judged personality scores. Although we note that the zero acquaintance judges are deliberately and severely handicapped so as to isolate the Twitter features, there is an appearance of conflict with the possibility that personality consists of others' perceptions of a person's behavior (i.e. the lexical hypothesis which locates clusters of adjectives we use primarily to describe others, not ourselves), rather than self-assessments, which are notoriously different from those of observers. This practice begs the question of how handicapped do judges need to be in order to dismiss scores significantly different from self-assessments? Anyone undertaking such a study needs at a minimum to address these issues. The default position that self-assessments are valid should go out the window in the presence of potentially better observer judged personality.

All is not lost, though: a genuinely fascinating aspect of this study is that they publish a large table (their Table 3) of correlations between observer-judged personality scores and LIWC features. They propose that the significantly correlated features and scores infer something about what information the human users employ in their judgments. It is this sort of analysis that might be used to challenge the self-assessments! After all, if an author's behavior were physically restricted to providing tweets, or their observers were restricted to reading such (a scenario not entirely fanciful), arguably that author's personality is described solely by behavior consisting of the tweets only, and in such cases the human observed personality may be seen as normative.

Another goal of theirs, of greatest interest to us, was the discovery of linguistic features correlated with personality self-assessments. They located 26 significant ($p < 0.05$) correlations between the self-report BFI personality scores and the LIWC features (again, found in their Table 3). They note particularly that they found a negative correlation between Agreeableness and negation words (which include words such as no, not, never). Their fundamental contribution here consists of some unusual correlations such as Extraversion with assent words, function words (negative), and impersonal pronouns (negative).

Openness was positively correlated with prepositions but negatively correlated with the use of adverbs, non-fluent words, affect, and swear words. It is amazing to see such a quantity of significant features produced from such a diverse population, and such a sparse corpus. Someone needs to check these correlations with a much larger sample of the population.

The inclusion of human judged personality is a critical contribution of this study, although they do not defend their practice of using the BFI, which is intended to be a self-report assessment, for that effort. Although such a practice is not new, it is quite unusual in this area. Others may do well to note their example of grounding personality assessment in the judgment of human observers of behavior in addition to self-report questionnaires.

| Investigator(s) | Personality Assessment | Corpora | Features | Extraction tool | Stats | Error analysis | # of participants | Population | Pop. mean age (y) | Outcome |
|---|---|---|---|---|---|---|---|---|---|---|
| [27], Pennebaker and King (1999) | Various | Essays | LIWC | LIWC | Pearson | | 841 | University students | | 17 significant features |
| [1], Argamon, et al. (2005) | NEO-FFI | Freewriting | LIWC (selected) | | SMO (in Weka) | Proportion accurate | 1157 | Students | | Bin. Class., max 58.2% accuracy |
| [24], Oberlander and Nowson (2006) | IPIP NEO-PI-R [4] | Blogs | $n$-grams | WMatrix | Naive Bayes, SVN | Percentage correct | 71 (34% M) | Bloggers | 28.3 | 83.6% |
| [22], Nowson (2007) | IPIP-50 | Blogs | $n$-grams | Custom | Naive Bayes | Proportion accurate | 1672 | bloggers | | Bin. class, best 66.4% |
| [20], Mairesse et al. (2007) | Self, observer | Freewriting, (EAR-[21]) | LIWC (selected), MRC, manual | Various | Bin Class, Regression / Weka | Proportion accurate | 2575 | Students | | Trained models, bin. class. 62.52% max. |
| [5], Buffardi and Campbell (2008) | NPI | Facebook pages | Quantity of text; human judgments | Manual | Simple correlations | | 129 | Undergraduates | 18.97 | Significant feature correlates |
| [23], Oberlander and Gill (2006) | EPQ-R, 3 factor | Emails | unigram POS tags, $n$-grams | Various | | | 105 | Students | 24.34 | Useful feature list. |
| [11], Estival and Hutchinson (2007) | IPIP-41 | emails | Unigrams: lexical, POS | Custom | Weka: NN (IBk), SVM | Proportion accurate | 1033 | English speakers | | 56.73% max. |
| [19], Luyckx and Daelemans (2008) | MBTI | Personae (Dutch) | lexical, POS, CGP $n$-grams | MBSP | TiMBL (MBL) | F-score, percentage accuracy | 145 | Belgian university students | | 82.07% |
| [30], Roshchina et al. (2011) | N/A | hotel reviews | Mairesse | Mairesse | st. dev. of inferred scores | | 1030 | TripAdvisor.com users who wrote $\geq 5$ reviews | | M5 regression tree preferred |
| [7], Chittaranjan, et al. (2011) | TIPI | Lausanne LDCC | Word length, # of interlocutors, social | | Corr., SVN | Proportion accurate | 83 | Population | 29.7 | Bin. Class., best 75.9% |
| [6], Celli and Rossi (2012) | None | Twitter public timeline | [20], social | Custom, Gephi | | | 13,000 | Twitter users $> 1$ post | | Applied prev. features, suggested new social features |
| [2], Bai et al. (2012) | BFI | Renren, soc. | Word affect | Custom | Various | Reported precision, Recall, F-stat | 209 | Chinese students | 23.8 | Chinese language features |
| [35], Sumner, et al. (2012) | TIPI, SD3 | Twitter | LIWC | LIWC | Spearman corr. | AUC, TPR, TNR | 2927 | Twitter users | | Significant corr., weak classification |
| [15], Golbeck, et al. (2011) | BFI | Facebook text | LIWC | LIWC | Pearson | MAE | 167 | Facebook users | 31.2 | 16 features correlated, 1 surprising |
| [14], Golbeck, et al. (2011) | BFI | tweets (max 2000 per user) | Sentiment, punctuation | LIWC, MRC, GI [34] | ZeroR and GP | MAE | 50 | Twitter users | | English language features, score regression |
| [16], Holtgraves (2011) | Personality | SMS text messages | LIWC, custom | LIWC | Pearson corr. | | 224 (46.4%M) | University students | 19.08 | 18 significant features |
| [39], Yee et al. (2011) | IPIP-50 | Second Life game: text chat and other behavior | LIWC | LIWC | Pearson corr. | | 76 (67.1%M) | University students | 21.07 | 11 significant features |
| [29], Qiu et al. (2012) | BFI | tweets | LIWC | LIWC | Pearson corr. | | 142 | English language Twitter users worldwide | | 26 statistically significant features |
| [37], Yarkoni (2010) | IPIP-50, IPIP-300 | Blogs | LIWC, unigrams | Custom, LIWC | Pearson corr. | | 694 (24.5%M) | Bloggers | 36.2 | Hundreds of lexical features |

# Chapter 3

# Conclusion

## 3.1 State of the art

Presently the state of the art consists of a variety of text features exhibiting strong correlations with self-assessment personality scores. Attributes of current work are depicted in the table. Most of the studies involve unigram word sentiment categories from the LIWC tool, or similar features extracted with custom tools. Just a few studies go deeper into lexical and part of speech $n$-gram analysis. Even rarer is thoughtful consideration of possible higher order structures that these simple features may be describing, and such discussion has not led to any significant results yet.

This is a budding area of research; in terms of quantity most of the work in this area was published in 2011 to the present. However most of the work takes the same approach: extracting stylistic features (restricted to unigrams) using LIWC and perhaps $n$-grams, then correlating with self-assessment personality scores. Early on, Oberlander suggested looking at POS $n$-grams with $n > 1$ and even presented some results, but only lone researcher hiding in Belgium took him up on the invitation [19]. Strangely everyone has ignored both hers and Oberlander's work, often while noting a dire need for more predictive features to drive classification by author personality scores. While they are enjoyable to read, we have to restrain our exuberance about the abundance of features constantly being discovered for a variety of populations due to problems with dimensionality explosion in learning. In extreme

cases, authors publish many pages of statistically significant text feature correlates to personality scores.

For example the single feature consisting of the relative frequency counts of the bigram “I think” may tell us more about personality than a pair of features consisting of the relative frequencies of “I” and “think”, in which case it is better to use the former and discard the latter two. However these ad hoc methods might include only the former, or needlessly complicate learning by including both. Classification results are discussed in the next section.

## 3.2 Critical analysis

Many of these authors use statistical techniques to predict personality scores or to perform classification of writers into two or more classes for each personality trait. Bifurcating the personality trait dimensions through binary classification does more harm than good. There is no theoretical basis for binary personality types (the MBTI has long been rejected as spurious); as we already mentioned personality comes in dimensions that are normally distributed, so dividing observations of a personality trait into two classes is senseless. Given the distribution of personality scores, those who provide classifiers placing individuals in Low, Medium, and High categories are contributing something of greater value. Predicting individual scores by regression would be much better, but only Mairesse tried it [20], with significant results. Often they declare they are using linear kernels; one would wish that they would either justify that decision or try some other kernels too, now that they have been well known for 20 years.

If the Low, Medium, High personality classifications are useful, some adjustments are need to the approaches that typify this body of work. After a brief perusal of the solution space by one (or a half dozen) learning algorithms, researchers often quickly become pessimistic about the possibility of training a more accurate classifier. Do they have a reason to believe that complicating their experiment with many learning techniques is any better than running the same algorithm five times with different initial conditions? It would be better to see more compute time devoted to searching the solution surface with hopes of avoiding being stuck in a local optimum, perhaps varying the input parameters, as is done in many other avenues

of inquiry. Ultimately they are, at best, demonstrating a lower bound on their classification accuracy, since they never do an exhaustive search of the solution space. One group, recognizing these issues, ran a competition for the general public to train the best classifier. But perhaps a coordinated, deeper search of the solution space would have given a better outcome. Nobody makes the case that the classification problem is impossible to solve by brute force due to the size of their dataset; in fact in some cases the samples are so small after subsampling ($N < 60$) that such a solution would be computable in milliseconds. If the latter is at all feasible, even for larger $N$ it would be better to crowd source for compute time and locate the known optimal solution rather than having everyone run their favorite heuristic. When doing so, it should always be noted that such demonstrations will necessarily be inferior to what a practitioner might obtain combining both text and non-text (social, demographic, etc.) features.

People should not use accuracy (error rate) as the sole measure of the usefulness of their classifiers. It would be fine to simply report the cost as Mairesse did for his regression results, perhaps as a multiple of the baseline cost. They could give both. The learning algorithms minimize misclassification cost, not error rate. It is worse to misclassify someone with a score in the middle of a class than it is to misclassify someone on the boundary; cost takes this into account. If overweighting outliers is a concern, something can be done about it.

# Chapter 4

# Future

## 4.1 Studies over time, topic and situation

Longitudinal studies might provide more insight, as personality is defined in terms of a pattern of observed behavior over time; however with the exception of the early Pennebaker study, none of them employ carefully designed samplings of texts from the same group of participants over an extended period of time. Thus many of these studies can be regarded as a snapshot of the mood of each participant at a given moment or at a few randomly selected moments, in a way that confounds detection of their usual linguistic behavior pattern. Also, variety in situation and topic affect personality assessment. Not everyone signals awareness of this, but some deliberately focus their studies on a single topic and others explicitly welcome a variety (in which cases it is important to do a larger study). A couple of researchers consider the challenge represented by a variety of situations in which the participants find themselves, for example in the Bai et al. study. They posit that differences between online and offline behavior might explain some of their results; they suggest that losing face becomes less important online. Also related to chronology, the obvious requirement for online learning is ignored.

## 4.2 Prediction of personality subtraits

We sincerely wish Yarkoni's contribution by studying personality facets (there are 30 in all, each of the 5 personality traits are divided into 6 sub-traits) would catch on; for sufficient large volumes of text and numbers of participants it is surely possible to find reliable predictors of the facets, which are themselves more predictive of behavior. Learning might be easier due to the existence of fewer predictive features.

## 4.3 Third parties

Surprisingly we have found no work that takes into account (beyond a simple count) interactions in text with interlocutors writing responses to the participant whose personality is being predicted. In fact one group wrote a tool to deliberately cut out email replies! Such texts are readily available in the case of text messaging and the many social networking studies. Those texts could be very predictive of personality scores (Bai et al. hint at the issue but it is impossible to tell whether they incorporated it into a feature). If it is hard to discern the sentiment of a statement, the answer may lie in how people react. The inclusion of such features appears well supported by the early lexical theory which posits that personality is described by the words a person's acquaintances use to describe their behavior.

Clearly the field is understudied. The way forward is to take the example of those who dig deep to find more structured and principled features. The resulting dimensionality reduction may unlock the predictive potential of text features. Deeper understanding of the linguistic behavior that forms a basis for prediction of personality scores and the behavior patterns that comprise personality would finally result in solid applications that transcend what is currently possible.

# Bibliography


[1] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker. Lexical predictors of personality type. In *in 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[2] S. Bai, T. Zhu, and L. Cheng. Big-five personality prediction based on user behaviors at social network sites. *arXiv preprint arXiv:1204.4809*, 2012.

[3] J. Block. The five-factor framing of personality and beyond: Some ruminations. *Psychological Inquiry*, 21(1):2–25, 2010.

[4] T. Buchanan, J.A. Johnson, and L.R. Goldberg. Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment*, 21(2):115–127, 2005.

[5] L.E. Buffardi and W.K. Campbell. Narcissism and social networking web sites. *Personality and social psychology bulletin*, 34(10):1303–1314, 2008.

[6] F. Celli and L. Rossi. The role of emotional stability in twitter conversations. *EACL 2012*, page 10, 2012.

[7] G. Chittaranjan, J. Blom, and D. Gatica-Perez. Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *Wearable Computers (ISWC), 2011 15th Annual International Symposium on*, pages 29–36. IEEE, 2011.

[8] P.T. Costa and R.R. McCrae. Neo pi-r professional manual. *Odessa, FL: Psychological Assessment Resources*, 396:653–65, 1992.

[9] PT Costa Jr and RR McCrae. Toward a new generation of personality theories: Theoretical contexts for the five-factor model. *The five factor model of personality: Theoretical perspectives. Hrsg.: JS Wiggins. New York*, pages 51–87, 1996.

[10] W. Daelemans, S. Buchholz, J. Veenstra, et al. Memory-based shallow parsing. In *Proceedings of CoNLL*, volume 99, pages 53–60. Bergen: Association for Computational Linguistics, 1999.

[11] D. Estival, T. Gaustad, S.B. Pham, W. Radford, and B. Hutchinson. Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272, 2007.

[12] D. Estival, T. Gaustad, S.B. Pham, W. Radford, and B. Hutchinson. Tat: an author profiling tool with application to arabic emails. In *Proceedings of the Australasian Language Technology Workshop*, pages 21–30, 2007.

[13] A.J. Gill, S. Nowson, and J. Oberlander. What are they blogging about? personality, topic and motivation in blogs. In *Proceedings of the Third International ICWSM Conference*, 2009.

[14] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 149–156. IEEE, 2011.

[15] J. Golbeck, C. Robles, and K. Turner. Predicting personality with social media. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, pages 253–262. ACM, 2011.

[16] T. Holtgraves. Text messaging, personality, and the social context. *Journal of Research in Personality*, 45(1):92–99, 2011.

[17] O.P. John, R.W. Robins, and L.A. Pervin. *Handbook of personality: theory and research*. The Guilford Press, 2008.

[18] C. Leacock, G. Towell, and E. Voorhees. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 260–265, 1993.

[19] Kim Luyckx and Walter Daelemans. Using syntactic features to predict author personality from text. In *Proceedings of Digital Humanities 2008 (DH 2008)*, pages 146–149, 2008.

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

[21] M.R. Mehl, S.D. Gosling, and J.W. Pennebaker. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862, 2006.

[22] S. Nowson. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *In Proceedings of the International Conference on Weblogs and Social*. Citeseer, 2007.

[23] J. Oberlander and A.J. Gill. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 42(3):239–270, 2006.

[24] J. Oberlander and S. Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634. Association for Computational Linguistics, 2006.

[25] B. Pang and L. Lee. *Opinion mining and sentiment analysis*. Now Pub, 2008.

[26] S.V. Paunonen and M.C. Ashton. Big five factors and facets and the prediction of behavior. *Journal of personality and social psychology*, 81(3):524, 2001.

[27] J.W. Pennebaker and L.A. King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[28] S. Picazo-Vela, S.Y. Chou, A.J. Melcher, and J.M. Pearson. Why provide an online review? an extended theory of planned behavior and the role of big-five personality traits. *Computers in Human Behavior*, 26(4):685–696, 2010.

[29] L. Qiu, H. Lin, J. Ramsay, and F. Yang. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 2012.

[30] A. Roshchina, J. Cardiff, and P. Rosso. User profile construction in the twin personality-based recommender system. *Sentiment Analysis where AI meets Psychology (SAAIP)*, page 73, 2011.

[31] G. Saucier and L.R. Goldberg. Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of personality*, 69(6):847–879, 2002.

[32] Online source taterunino.net. Retrieved, December 2012.

[33] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001.

[34] P.J. Stone, R.F. Bales, J.Z. Namenwirth, and D.M. Ogilvie. The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498, 1962.

[35] C. Sumner, A. Byers, R. Boochever, and G.J. Park. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. *Proceedings of the IEEE 11th International Conference on Machine Learning and Applications ICMLA 2012*, 2012. To appear in December 2012.

[36] Y.R. Tausczik and J.W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[37] T. Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373, 2010.

[38] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational*

*Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.

[39] N. Yee, H. Harris, M. Jabon, and J.N. Bailenson. The expression of personality in virtual worlds. *Social Psychological and Personality Science*, 2(1):5–12, 2011.