INFERRING HUMAN PERSONALITY FROM WRITTEN MEDIA

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTORATE OF PHILOSOPHY

IN

COMPUTER SCIENCE

MAY 2020

By

William R. Wright

Dissertation Committee:

David Chin, Chairperson

Kentaro Hayashi, Anthony Kuh, Michael-Brian Ogawa, William O'Grady, Scott Robertson

Table of Contents

List of Figures

List of Tables

Abstract

This work explores the association between human personality and language features consisting of sequences of tokens. My work reveals that there are such features that are predictive of personality over multiple corpora taken from different populations of English speakers. I gathered written text authored by 50 individuals who participated on a bodybuilding web forum (the Forum corpus). Also I administered a personality questionnaire following the protocol provided by the International Personality Item Pool (IPIP). For comparison across other populations I also obtained text corpora from three other research groups, along with the results of personality assessments: the EAR corpora consisting of transcripts of the speech of 96 participants as they go about their daily lives, Essays written by 2,588 undergraduates at the University of Texas and posts by 244 Facebook users. After performing part-of-speech (POS) tagging on the text for all the participants in these corpora, I extracted unigrams, bigrams and trigrams (n-grams) of tokens and their POS tags, and counted every word/tag permutation that appeared.

I considered only features appearing one or more times per 1000 words in the Forum corpus because there was not enough data to consider sparser features. I found 766 such features. From among those features I explored which were relevant across both my Forum corpus and at least one of the borrowed corpora, since those are the most promising, robust features that illustrate the possibility of building models across various corpora using the same language features. 75 of the features were associated with one or more personality dimensions across both the Forum corpus and at least one additional corpus. I devised explanations as to why some of the features are correlated with a given personality dimension. That task establishes that although some of the features may have arisen randomly, one can confidently proceed with the conclusion that English speakers consistently express their personalities through their language usage. In addition, to show that it is possible to use these features for prediction, I generated multiple linear regression models for each corpora-personality

dimension combination; in the best case (Openness with the Forum corpus) I obtained R^2 of 0.686 and S (standard error of the estimate) of 0.561. My work sets a foundation for more robust, accurate models of personality. I hope that others will find additional principled explanations of why the features I found are associated with personality. In the future I anticipate that suitable language-analytical techniques will deepen insight both in the case of English speakers and speakers of additional world languages.

Chapter 1

Introduction

Given similar circumstances, individuals behave in ways that differ significantly between them, and the differences persist over time. Verbal behavior, specifically, communicates something about author personality, age, gender, geography[1], and social relationship to the audience [2]. Of immediate interest is personality: persistent patterns that can be observed and used to anticipate future behavior.

Personality assessment can also be used to predict subjects' use of products and services, thriving in academic programs, and performance in work environments. Already, companies and medical schools have used personality questionnaires to predict the performance of employees and students. Dating websites have also used personality assessment for purposes of assessing compatibility between users.

The costs and practical challenge of administering personality questionnaires, including motivating participants, limit their usage. Also, the possibility of dissimulation threatens the validity of self-report questionnaires, e.g. in the case of medical school applicants [4] or others [5]. To expand the understanding of personality, and to overcome these limitations of administering questionnaires, researchers have developed techniques to infer personality from digitized records of individual behavior—predominantly digitized text [6–8].

Whether well intentioned or not, the knowledge this field provides could be misused in a way that harms individuals or entire classes of people, or that violates laws that protect their rights. For example personality-based educational technology could be deployed to sort individuals according to their presumed future roles in society, while denying them

[1] E.g. amongst 75,000 Facebook volunteers those living in high elevations spoke about mountains [2].

choice or equal opportunity to pursue their potential. Personality inference can also be transferred to other conditions that fall outside normal functioning: personality disorders and other persistent disturbances. Mistakes could affect other areas important to personal freedom such as authorization to travel between regions of the world, criminal sentencing and rehabilitation, or perhaps even a decision whether to award the death penalty in areas where that is legal.

The goal of this area of research is to locate new cues of personality from the language usage of individuals. 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 most studied kind of personality-expressive human behavior is possibly linguistic in nature.

In this dissertation, I present a method for locating linguistic cues, and the outcome of applying that method to several different bodies of text created by distinct populations (corpora). The focus is on language cues that are predictive of personality across several of the corpora, because such language features are more promising for general applications in the future.

Within a single corpora the usage of specific words is likely to be influenced by context. For example, I found [9] a correlation between "hurricane" and conscientiousness in Pennebaker's essays corpus [10] that is likely due to the proximity in time and geography to Hurricane Katrina. Drawing features from several corpora covering different periods of time and geographical areas helps me to exclude correlations that are due to current events. Also, because multiple different word sequences increment the count of a more general Part Of Speech (POS) based features, my POS features can include phrases describing a wider variety of circumstances and are thus less susceptible to the influence of specific time-and-place context. Another advantage of the POS features is that they are more frequent than specific word features which allows me to overcome the limitations of sparser word n-grams.

The research questions are as follows:

Does authors' usage of English grammatical structures reflect their personalities? What methodology extracts and predicts personality from grammar usage?

The anticipated contributions are:

I. Identify grammatical structures whose presence predicts author personality. Report the features identified, and investigate whether there are any that are predictive across different corpora drawn from diverse populations.

II. When a reason is evident, explain why a person high or low on a personality dimension would tend to use such features, contributing new insights regarding the nuances of personality expression.

III. Offer a rigorous, maximally language independent method that others can use to carry out similar research tasks.

This work offers a way to discern how participants express their personalities within a confined social group, for example students at a given university. Besides the corpora I gathered myself, I included multiple data sets from other researchers which allowed me to identify generalizable features with high confidence.

Chapter 2

Literature Review

2.1 Personality theory and assessment

We do not delve deeply into 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 [11] 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 [12]. 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 [13].

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 [13, 14].)

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 Extraverts can accomplish much in

concert with others. Low trait Extraversion is called Introversion; whereas the Extravert 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***. An agreeable person gets along well with other people, is cooperative, unpretentious, trusting, tactful and sensitive in their speech. Those on the opposite end of the continuum are perceived as hostile and unduly judgmental, cold and inconsiderate. An agreeable person tends to account for others' perspectives and needs although when there is a competition for scarce resources a person may behave in a more self-oriented fashion.

Factor III: ***Conscientiousness***. This factor describes effectiveness in performing prescribed rote, repeated activities. Also careful following of rules. However assessments on this trait do not generally ask questions about a person's ethics [15].

Factor IV: ***Neuroticism***. Also called Emotional Stability, reversing the measure. A person high on this dimension tends to persistently experience anxiety, envy, guilt and general uneasiness. Neurotic individuals tend to perceive minor everyday events negatively, and to be very sensitive to the discomfort those events cause them. Also they are affected by a lack confidence, and are apt to cease action in the face of difficulty or to refuse action in anticipation of obstacles.

Factor V: ***Openness***. This dimension is related to qualities that lead to general intellectual growth (in fact the dimension is sometimes called Intellect), i.e. openness to unfamiliar ideas and new experiences. An open person is likely to explore by reading, traveling, engaging others in conversation about ideas, and actively developing their interests in fresh directions.

2.2 Natural language processing contribution

Stable individual differences in human speech and writing behaviors are included in some analyses 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 observations in text that are stable over time but differences 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 [16]. However when automating using Natural Language Processing (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 [17] and [18]), 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. I address only written corpora for three reasons: the first is that there is a substantial 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). My 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 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.

2.3 Excluded discussions

I exclude in depth discussion of the various self-assessment personality questionnaires because the discussion would take us too 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 over-committed to their positions and unaware of the significant trade-offs involved in increasing or decreasing the number of factors. We exclude a significant area of study of evidences of mood in text. If I capture mood at a point in time, future mood may differ considerably, resulting in behavior that does not match the mood at the captured time. Personality (besides perhaps the neuroticism dimension) describes behavior that is fairly consistent over time and varying situations rather than episodic and is unrelated to the external environment.

Although I note it when I see it in the studies of interest, I do not focus on statistical prediction methods (classification and regression techniques) because the results are at best specific to the studies and have not been shown to be generalizable. There is so much to be done in terms of finding better training data.

2.4 Literature

2.4.1 Pioneering studies

The initial study of this area of research, Pennebaker and King's early work in this area [10] 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 to personality. The tool, Linguistic Inquiry and Word Count (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 [19].

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 an earlier sentiment analysis tool, the General Inquirer system [20]. Ultimately they identified 17 features predictive of the personality scores, such as that Neuroticism scores are significantly correlated with more frequent use of the first person singular pronoun and that the frequency of words entailing causation correlated negatively with Openness.

Helpfully, some attention was given to the issue of cross-situational behavior and how it affects 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 affect 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 [16] 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 [21]). 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 Functional 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 generalization that Neurotics tend to express 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. They use two SVM binary classifiers to first divide the High (top third) scoring personalities from the rest, and then to divide the rest into middle and Low (bottom third). One omission is that they do not mention the 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 discrete categories when they could have predicted the actual scores, along with an error analysis, and then at the end if their audience so desires, introduce the classification results with the accuracies computed.

As in the original Argamon et al. 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 [22]. 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 IPIP NEO-PI-R, revised for administration via the Internet [23].

Their features are simply n-grams, but before extracting the n-grams, they identified proper nouns using the corpus analysis tool WMatrix [24] and replaced them all with a common sequence of characters (the marker "NP1"). A major contribution of their work is the use of well a defined, principled manual feature selection process as well as automated feature selection to decide what features to train on, which deserves notice given that others often leave out an explanation of their feature selection process. They trained on five different sets of features: four manual selections and the fifth an automated selection, as follows.

I. The first, least restricted set was a manual selection of n-gram features excluding uncommonly appearing n-gram features.

II. Then they further narrowed the set to include only features that appeared at least five times in the corpus that had significant correlations ($p < 0.01$) with personality test scores categorized into their "High" or "Low" categories (which exclude scorers in the approximate middle 1/3 of the sample) in at least one of the personality traits.

III. Concerned that some of the features selected in (II) might have have arisen for consideration because just a few authors used them a lot, they created another feature set by adding to (II) the test that at least 50% of texts include the feature.

IV. Their last manually selected set consisted of the features that meet the criterion in (II) above, replacing the p constraint with $p < 0.001$, and were included in (III), and exhibit a significant ($p<0.05$) correlation with the actual trait score (not to be confused with the High/Low categories of (II).

V. Finally they selected a set of features in an automated way (using the WEKA tool [25, 26], a BFS search using the CfsSubsetEval evaluator option in the software).

They mention that automatic feature selection gave them the best results, but they expressed a concern about over-fitting (without giving a justification), and suppressed 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 sometimes challenging to follow when they helpfully label what is being discussed but use the labels later in their writing without a word of explanation, producing excessive back-references. 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 [22], built on that work [27]. They called upon their original classifiers, which they had trained on a carefully collected set of < 100 blogs and their authors' personality scores. In the current

work 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 responders 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 [22] convinced them that simply naive Bayes actually outperformed SVM's for personality classification, so they employed the former. The best classification accuracy attained was 66.4%, for Conscientiousness.

The group continued to produce additional studies such as [28], published in 2009 by the Association for the Advancement of Artificial Intelligence. There, they added a discussion of a flow of causation from an individual's essence as Neurotic, Extravert or Open, to the way that individual writes; the former is described as a *motivation* or *desire*.

Mairesse, et al. [29] built an ambitious automated personality recognizing tool, Personality Recognizer; it uses LIWC and a comparable tool MRC [30] 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 trained the model on known personality scores (via self and observer reports) of 2575 students. The usefulness of their model in analyzing dialog 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). It might have been helpful to measure error in terms of distances between the predicted (where the low category is interpreted as the lowest possible score and the high is interpreted as the the highest possible score), and actual personality score. The author thinks that non-extreme scores are "just noise", which is debatable since since most personality scores fall in the middle region. 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. (correspondence, F. Mairesse.) That approach forces the median score to be the divider between the two classes, which is challenging to justify unless the classifier is intended to drive a specific binary decision that needs to be made.

It is helpful that they revealed a much broader range 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 so. 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.

Roshchina et al. [7] 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 to present to the user hotel reviews written by people with similar personalities to that user, a somewhat novel approach introducing psychology to opinion mining [31]. 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 are many publications in this area). 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 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 indicate that they ran the M5 regression tree algorithm on their data. However the tool requires actual personality measurements for training, so it is unclear how they applied that algorithm without actually administering a personality test and furnishing the measurements to the algorithm. The authors offer the hypothesis that the best algorithm is that which predicts personality scores that differ the least amongst various works by the same author, yet work remains to be done to build evidence to support that view. Despite its issues this paper is an example of a well documented application.

Besides expanding their sample, their work could be improved upon by following the prior work of another group studying online reviews: Picazo-Vela, et al. [32], who investigated the possibility that there is a unique user personality that dominates on-line 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. Their results [32] indicate that those high on Neuroticism or Conscientiousness were more likely to write a review.

2.4.2 POS inquiry

In [33] 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, cover-

ing Extraversion, Neuroticism, and Psychoticism (which is commonly believed to inversely correlate with both Agreeableness and Conscientiousness). 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 me wonder why they did not simply ask authors to submit real emails written recently. The study design might have been intended to account for the confidentiality concerns of users or their organizations. I note that the 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 further, they might have adopted similar practices as a later study we cover here [34] 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. 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 collocations (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.

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

Another relatively early group, Estival et al. [35] 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; there were 1033 participants who met the desired criteria: at least five messages written by the participant, which when combined add up to at least 1000 words, and a valid personality questionnaire received (they do not mention how many participants were eliminated by these criteria). 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 do not 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 whether they included longer POS n-grams as did the study of Luyckx and Daelemans, discussed below.

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 WEKA [25, 26], 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 do not 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 list or even say how many features they ultimately used for learning.

It is good 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 in they were in [33]), with no strong constraints on the text being produced. Combining a wide variety

of potential topics (they include work and home email) with no restriction on the subject might have made things more challenging, because allowing the authors so much flexibility increases the variety of language constructs available to them, which in turn increases the sparsity of features which makes it more difficult infer anything from them. However they attribute some of their issues to the small size of their corpus relative to that of Oberlander and Nowson [22] and the fact that they studied blogs rather than emails. Yet, other studies with comparable sample sizes did seem to make some contributions, some of which we cover here e.g. [29], and Nowson later published results on the blog texts [27].

The authors offer their corpus for others to study. Incidentally the same authors have also studied Arabic emails [36]. 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 [34] extracted 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 [37].

They chose to use as their personality assessment the Myers-Briggs Type Indicator (MBTI) [38]. Some challenges [39, 40] have been brought as to the MBTI's predictive power, the basis of its Type constructs, and consistency issues when an individual re-takes the test. For these reasons most studies employ questionnaires that assess personality as described by the Five Factor model (FFM) [13].

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. The features they extracted include simple lexical n-grams and Part of Speech (POS) n-grams which are based on the parts of speech present in the text.

To extract the syntactic POS features, they employ Memory-Based Shallow Parsing (MBSP) [41]. They use TiMBL, a memory-based technique for training. Feature selection technique is unspecified; they divide features into the three types described above and offer their

results for each group of features separately. They were able to predict Extraversion with an binary classification accuracy of 65.5% with lexical 3-grams, which as they note is better than a number of other studies such as [16], [29] and [27], which we cover in this review. Their prediction of intuition, a factor related to Openness, was from an unspecified set of POS 3-grams, was 62% accurate, similar to the accuracy of [29].

The main contribution of this paper is to demonstrate that it is possible use POS n-grams to predict personality scores of an additional population. They do not report what specific n-grams they used, or the correlation coefficients. Regardless, after seeing this study I wonder if other subsequent efforts had incorporated such features, they might have achieved improved results.

2.4.3 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 is 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 [42]. 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 partially described as a cluster of human personality subtraits in their extremes, involving a often exaggerated (but not necessarily stable) positive beliefs about 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 them. 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, as well as with the sheer quantity of social interaction. 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; if it can be detected early in life it could be addressed in a supportive setting.

2.4.4 LIWC studies of personal communications

In 2012 F. Celli and L. Rossi [43] assembled a classifier of Twitter users according to personality by employing features associated with Neuroticism as published by the creators of the Mairesse tool LIWC [29], 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 made use of the correlations published in [29], which they cite [29, 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. My concern with this very creative approach is that possibly two different features might share so much mutual information that when they both appear simultaneously (or both do not) in a document,

counting them both in this way overestimates their contribution to the personality score, thus introducing error in the estimate.

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, Sumner et al. [44], studied 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 assessing Dark Triad dimensions, such as the obvious ones related to deceptive participants (e.g. the Machiavellians in the sample) who may distort their 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 [44], they use Pearson correlations, instead of Spearman correlations, which are less sensitive to extreme values. For the Narcissism trait they found that words such as “buddy” and “friend” were strongly correlated with a relatively large effect size (the Spearman correlation is 0.073). Also the “other punctuation” category that includes “@” and “#” characters were associated with Narcissism. Those characters are used to bring a tweet to the attention of another

user, and to facilitate the presence of the tweet in others search results, respectively. Both of these usages draw extra attention to the Narcissist.

They sought to classify participants according to personality traits, and to predict personality scores. Their classification results are from a public competition that they held, offering their data to others for analysis. Quality evaluation is an issue that these sorts of events tend to raise, although I find no discussion of it in the paper. Their contest participants attempted a large variety of techniques to address the classification problem; they tried many machine learning techniques. No mention is made of feature selection.

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 a little better than chance, and the latter did not predict true positives, identifying only 2 of the 125 individuals in the top 10%.

A helpful contribution is their example in 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. It would help to add the personality score prediction task to the error analysis. 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.

A significant contribution of this paper is the ethical discussion. They highlight the need for restraint in employing classification tools of quite limited accuracy to make decisions about people. They also discuss how 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.

Given the those limits of LIWC they could perhaps have done more with lexical unigrams, for example, it is commonly known that Extraverts freely use invented or malformed words; they count them and treat them as a single feature. They do note one potential source of bias in their population: 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 suggest that the proper analysis would reveal that the effect of selection bias is negligible in this case.

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. [45] 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?).

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 do an error analysis (beyond just 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. They do try classification with only $n = 80$, which is difficult to do in a statistically valid way with so few participants, because they have to divide the data into sub-samples for cross validation and testing.

A unique 2012 paper from Bai et al. [46] 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 (I) linguistic and social factors present in Renren, and (II) personality as measured by self-report assessments.

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 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 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 training 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 *angry* category that their tool produces) were correlated 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. The study by Buffardi and Campbell looking at cues of Narcissism [42]

did look at Facebook users, but without automated information gathering. Golbeck, et al. have spearheaded the effort by reporting a number of 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 [47]. It is unclear why the word counts were low on average.

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). Interestingly they report that the last name length in characters is positively correlated with Neuroticism (the inverse of emotional stability).

The significant correlations make me optimistic (cautiously, due to the sparse average word counts from each author) that their features have some predictive power rendering them worthy of consideration by anyone trying to predict personality scores of Facebook users. Predicting actual personality scores proved challenging in this case. They performed regression using M5 and GP. They also report the regression correlation coefficient, but additional studies or repeating this one with cross validation on a larger sample would strengthen my confidence in the validity of their model.

Golbeck et al. [48] continued the project by studying 50 Twitter users who each had written an average of 2000 words, and took the BFI [49] to measure their personalities. 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 [30], which categorizes words in a way comparable to LIWC, and General Inquirer [20], 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.

Reporting the features is a helpful contribution for those hoping to build a classifier or to investigate promising features. Although they mention their attempt to build a classifier, they present no results. Instead they report the MAE on a regression task (the normalized MAE varies from 0.12 to 0.18 in different scenarios), but without a comparison of their MAE with a baseline, is it difficult to tell how much of an improvement their features offered. Their approach in trying a variety of tools for feature extraction is worth considering.

Tal Yarkoni conducted a study in 2010 [50] wherein he was able to examine 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. They discuss when and why certain lexical unigrams exhibit an unexpected correlation with a trait. It would be helpful to explore how to identify such situations and how to exclude the unigrams that do not fit.

Participants chose which personality assessment to complete: the IPIP-50 (50 questions) or IPIP-300 (300 questions) [51]. His goal was to explore many features provided by LIWC categories that were skipped over by others previous efforts and well as to consider bag-of-words features (word unigrams) that his large corpus has the benefit of offering in sufficient quantities for inferences to be made. 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 individually 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 addresses this by looking at combined features, which can be helpful if the combined feature is actually more strongly associated with the target variable. He selected features of high enough incidence so as not to be confused with noise; he reports the minimum count of each batch of features, but never tells us whether that is a per-participant frequency, or a count for the entire corpus.

I find fascinating 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 could be more descriptive of people and predictive of their behavior (actually, there is a famous study confirming this [52], which was confirmed by others). When there is enough data from a

sample to support such multiplicity of facets, it would be useful to seek to predict them. This is especially true when some specific trait is sought, e.g. his example of a marketer trying to find people who are body conscious.

Soon after, Holtgraves stated the goal of investigating the use of text messaging as a function of personality [53]. 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. They asked participants to report the answers to those on 5-point Likert scales, which severely limits the resolution of the responses to a few categories (vs., say, asking for ratings on a 1-100 numerical scale would do). Although LIWC includes emoticons in its features, it cannot handle non-standard English. Therefore, they augmented 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, shortening words by dropping letters and 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 the 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 for them), 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 reasonable 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 just relying on whatever words their chosen tool can identify. There is no classification model offered, as is common practice, but I do not think their study suffers from the lack.

In quite an unusual study, Yee et al. present a study [54] 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 40,000 to 50,000 users [55] 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 a 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 that may be explained by an

unwillingness to use non-standard English; the authors note, for example, that some users substitute *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 interactions, that would be absent in monologues like essays or blog entries.

Another recent study of 142 Twitter users by Qiu et al. [56] 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. They administered a written self-report personality assessment, in this case the BFI [49], to the participants. Also, they formed a group of human judges to assess author personality by looking 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 obtained 2 sets of BFI scores for each user.

Introducing observer reports (formal personality assessments are usually conducted by self-report) raises some questions, issues that have been explored in a meta-study [57]. Personality is assessed not on a purely behavioral basis; there are questions in the assessments that relate to a person's subjective internal state of mind. For example the following items, from the BFI [49] are straightforward for a third party to observe:

> Is persistent, works until the task is finished
>
> Keeps things neat and tidy
>
> Avoids intellectual, philosophical discussions

Whereas the following, although they may be accompanied by behavioral cues, are attempts to describe internal states thought to be common in humanity but only observed directly by individuals within themselves:

Feels little sympathy for others

Often feels sad

Is suspicious of others' intentions

The meta-study [57] finds strong evidence supporting the idea that personality is more consistently described by observers, while noting that the internal phenomena (called "internal dynamics" in the paper) may be best assessed by subjects themselves. Of course some observers are more qualified than others (the meta-study describes this as response distortion, and mentions influential factors), and individuals can distort their self-assessment, especially when their self perception in the area of interest disagrees with the normative attitude towards a person similar to them, or when they are motivated to do so.

After comparing the self vs. observer report scores, Qiu et al. [56] concluded that the observers were better able to predict Agreeableness and Neuroticism, although they do not explain in detail how they reached the conclusion. This is an important consideration because the the observer reports were based solely upon Tweets from the subjects rather than in-person acquaintance. They publish a large table (Table 3 in the paper) 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 of personality.

Another goal of theirs, of interest to me, was the discovery of specific linguistic features correlated with personality self-assessments. They located 26 significant ($p < 0.05$) correlations between the self-report BFI [49] 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 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.

2.5 Conclusion

2.5.1 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 table 2.5.1. 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 detailed consideration of possible higher order structures that these simple features may be describing, and such discussion has not led to any significant results yet.

Investigator(s)	Personality Assessment	Corpora	Features	Extraction tool	Stats	Error analysis	# of participants	Population	Pop. mean age (y)	Outcome
[10], Pennebaker and King (1999)	Various	Essays	LIWC	LIWC	Pearson		841	University students		17 significant features
[16], Argamon, et al. (2005)	NEO-FFI	Freewriting	LIWC (selected)		SMO (in WEKA)	Proportion accurate	1157	Students		Bin. Class., max 58.2% accuracy
[22], Oberlander and Nowson (2006)	IPIP NEO-PI-R [23]	Blogs	n-grams	WMatrix	Naive Bayes, SVN	Percentage correct	71 (34% M)	Bloggers	28.3	83.6% accuracy
[27], Nowson (2007)	IPIP-50	Blogs	n-grams	Custom	Naive Bayes	Proportion accurate	1672	bloggers		Bin. class, best 66.4% accuracy
[29], Mairesse et al. (2007)	Self, observer	Freewriting, (EAR- [58])	LIWC (selected), MRC, manual	Various	Bin Class, Regression / WEKA	Proportion accurate	2575	Students		Trained models, bin. class. 62.52% max. accuracy
[42], Buffardi and Campbell (2008)	NPI	Facebook pages	Quantity of text; human judgments	Manual	Simple correlations		129	Undergraduates	18.97	Significant feature correlates
[33], Oberlander and Gill (2006)	EPQ-R, 3 factor	Emails	unigram POS tags, n-grams	Various			105	Students	24.34	Useful feature list.
[35], Estival and Hutchinson (2007)	IPIP-41	emails	Unigrams: lexical, POS	Custom	WEKA: NN (IBk), SVM	Proportion accurate	1033	English speakers		56.73% max. accuracy
[34], 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% accuracy
[7], Roshchina et al. (2011)	N/A	hotel reviews	Mairesse	Mairesse	st. dev. of inferred scores		1030	TripAdvisor.com users who wrote ≥ 5 reviews		M5 regression tree preferred
[45], 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% accuracy
[43], Celli and Rossi (2012)	None	Twitter public timeline	[29], social	Custom, Gephi			13,000	Twitter users > 1 post		Applied prev. features, suggested new social features
[46], Bai et al. (2012)	BFI	Renren, soc.	Word affect	Custom	Various	Reported precision, Recall, F-stat	209	Chinese students	23.8	Chinese language features
[44], Sumner, et al. (2012)	TIPI, SD3	Twitter	LIWC	LIWC	Spearman corr.	AUC, TPR, TNR	2927	Twitter users		Significant corr., weak classification
[47], Golbeck, et al. (2011)	BFI	Facebook text	LIWC	LIWC	Pearson	MAE	167	Facebook users	31.2	16 features correlated, 1 surprising
[48], Golbeck, et al. (2011)	BFI	tweets (max 2000 per user)	Sentiment, punctuation	LIWC, MRC, GI [20]	ZeroR and GP	MAE	50	Twitter users		English language features, score regression
[53], Holtgraves (2011)	Personality	SMS text messages	LIWC, custom	LIWC	Pearson corr.		224 (46.4%M)	University students	19.08	18 significant features
[54], 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
[56], Qiu et al. (2012)	BFI	tweets	LIWC	LIWC	Pearson corr.		142	English language Twitter users worldwide		26 statistically significant features
[50], Yarkoni (2010)	IPIP-50, IPIP-300	Blogs	LIWC, unigrams	Custom, LIWC	Pearson corr.		694 (24.5%M)	Bloggers	36.2	Hundreds of lexical features

This is a budding area of research; in terms of quantity, most of the work in this area was published from 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 and presented a promising result. A researcher in Belgium took him up on the invitation in an exploratory study [34]. While they acknowledge a need to explore for more predictive features to drive modeling of personality, others have not often emulated hers and Oberlander's work, perhaps not being aware of it.

This body of work raises a question about the general statistical validity of features introduced by individual studies; perhaps more meta studies linking the results and evaluating the studies in terms of which results are mutually supporting (and which are not). In some cases, authors publish many pages of statistically significant text feature correlates to personality scores; naturally some of those may be valid and others will be noise. Authors do mention each others work occasionally, I am not dismissing that, I am only saying this area has matured enough to invite a broader view.

2.5.2 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. By dividing personality at the median or mean score for a dimensions, a binary classifier is like a police radar gun that rounds a motorist's speed to either 0 or 100 MPH before displaying the speed: so much information is lost. Bifurcating the personality trait dimensions through binary classification does does not take full advantage of the predictive power of the features. Also, personality comes in dimensions that are roughly normally distributed, so dividing observations of a personality trait into two classes about the center of the distribution results in a a lot of error since so many people score close to the middle on the personality dimensions. Given the distribution of personality scores, those who provide classifiers placing individuals in Low, Medium, and High categories are contributing something of greater descriptive value. In in fact, naturally, the more categories are used, the lower the error, with some kind of regression or other linear model fitting the data better (predicting actual scores rather than categories) and minimizing the error. This leaves binary classification to situations where binary "yes" or "no" type decisions need to be made. In such cases I am

under the impression that it would be best to predict scores, tuning the model to minimize error and maximize validity on cross validation, and then impose classification at the end to suit the application, rather than making classification the core model.

Often studies use learning algorithms that try to minimize an objective function in a way that is intended to be efficient when working with large quantities of data. However the algorithms settle on a non-optimal solution when they stop iterating since for large data sets the solution space is too large to be exact in the task of minimizing their objective function. But for small studies, to be more precise in describing their data, why not actually locate the optimal label assignments by checking all $\binom{n}{n/2}$ of the ways to choose them, rather than running those algorithms? Nobody makes the case that the classification problem is impossible to solve by brute force due to the size of their dataset. Since n is small most of these studies, such a method should work well. That does not eliminate the value of running those algorithms as a demonstration; it might be helpful to take both approaches and present the results.

After a brief perusal of the solution space by one (or a half dozen) learning algorithms, researchers sometimes become pessimistic about the possibility of training a more accurate classifier. I wonder, do they have a reason to believe that complicating their experiment with many learning techniques is any better than tuning the algorithm or using different initial conditions (as appropriate)? 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 often the practice in other fields of research. One group [44], recognizing these nature of these overall issues, ran a competition for the general public to train the best classifier. More studies like that, with a mutually coordinated, deeper than ever before search of the solution space may offer additional insight into how to model this this data in particular cases.

When it comes to model evaluation, accuracy (error rate) is often the sole measure offered, for classifiers. It would more informative to also report the cost as Mairesse did for his regression results. I am unsure whether researchers are aware that cost can also be looked at in the case of classification models. Caution must be taken when applying a learning algorithm that generates a model that minimizes classification error rate rather than misclassification cost. It is worse to misclassify someone with a score close to the middle of a class than it is to misclassify someone near the extremes; an appropriate cost function

takes this into account. For example, if one is hiring for a job that really favors Extraverts, one would rather misclassify someone near the decision boundary than to misclassify an extremely introverted person as Extraverted, or vice versa; cost is maximized in such cases. If overweighting outliers is a concern, something can be done about that separately.

Finally, I believe future efforts will benefit from considering corpus-relevant feature structure. 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, ad hoc methods might include only the former, or detract from model building by including both.

2.6 Future

2.6.1 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, the studies reviewed here are a snapshot of the behavior of each participant at a given moment or at a few randomly selected moments. That way of sampling language might confound detection of the entire picture of their linguistic behavior. A snapshot in time tends to exclude a lot of important information particularly in the case of participants who draw from a wide pallet of behavior at different times.

Also, it is strongly possible that choice of topic, writing format, time limits, and other such constraints affect expression of personality. Not all researchers signal awareness of this. Some deliberately focus their studies on a single topic while others explicitly welcome a variety. It is possible that imposing restrictions may increase production of certain linguistic patterns while discouraging others. It is possible, but not guaranteed, that such modifications in the relative frequency of those features would be proportional to personality scores. I do not see a detailed discussion of this specific issue by others, as of yet. A couple of researchers consider the analytical challenge represented by a variety of situations in which the participants find themselves, for example in the Bai et al. [46] study. They posit that differences between relatively anonymous, internet based vs. in person behavior

might explain some of their results; they suggest that losing face becomes less important online.

Experiments with online learning might be a welcome application area. Thus far we see static models build on pre-existing data but it would be of interest to see how models fare as they are extended to new incoming data, whether labelled or not.

2.6.2 Prediction of personality subtraits

I wish Yarkoni's direction of 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 may then more narrowly predict specific behaviors that align with those facets.

2.6.3 Conversationalists

I have found no work that takes into account (beyond a simple count) interactions in text with interlocutors writing responses in dialogue with the participant whose personality is being predicted. One group wrote a tool to deliberately cut out email replies; this could be repurposed to label and include the replies, or to aid in discourse analysis and related features that could be constructed from the analysis. 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., [46] hint at the issue but I cannot tell whether they created a feature as a result. For example 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.

2.6.4 Next Steps

Clearly the field is understudied. One future option 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. More thorough 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. That exploration continues in the next chapter.

Chapter 3

Data

Since I was interested in larger collocations of text (n-grams) that appear less often than typical bag-of-words features (word unigrams), I needed larger samples of text from each person to obtain reasonable frequency statistics. Table 3.1 offers back-of-the-envelope calculations illustrating approximately how many words per author are needed to get reasonable statistics. For n-grams that only occur on average once per 1000 words (the threshold for feature consideration that I chose), tens of thousands of words are needed from a participant to get counts in the tens range. The existing available corpora did not offer such large texts, so I collected my own corpus with longer texts per participant, averaging 57,352 tokens (words and punctuation) per person. The corpus is comprised of the writings of 50 authors participating in the Internet forum hosted by www.bodybuilding.com, along with their scores on a 50-question personality questionnaire following the template offered on the website of the International Personality Item Pool (IPIP) protocol [51, 59, 60].

A limitation of this corpus is that the number of participants (50) suggests verification of any conclusions by analyzing additional corpora. Another limitation is that political and economic barriers prevent some populations from accessing an Internet connected device or from using that device to connect to the website www.bodybuilding.com or other information services, from having the leisure to engage in bodybuilding, or the freedom to speak openly in public without impediments to communication (e.g. parts of the internet being rendered inaccessible by authorities) or a credible threat of retribution, which surely affected participant recruitment.

rank	word/POS	frequency (per 1000 words)	Document size necessary to achieve count $n = 10$ unigram	bigram	trigram
1	you	41.8440	239	5,712	136,491
100	ok	1.7104	5,847	$3.419 \cdot 10^6$	$1.999 \cdot 10^9$
1000	worst	0.0779	128,356	$1.648 \cdot 10^9$	$2.115 \cdot 10^{13}$
10000	misjudged	0.0026	$3.794 \cdot 10^6$	$1.440 \cdot 10^{12}$	$5.462 \cdot 10^{17}$
	NN (noun)	106.7377	94	878	8,224
	DT (e.g. the, a, these)	62.2456	161	2,581	41,465
	MD (Model verbs, e.g., can, could, might, may)	17.2558	580	33,584	1,946,248
	WRB (Wh-adverbs, e.g. how, where why)	5.1342	1,948	379,360	73,888,565

Table 3.1: These ranks and calculations of document size required to achieve the $n = 10$ word count threshold are based on observed frequency in TV and movie scripts [1]. Figures for POS features (final four rows) are drawn from observed frequencies in the Forum corpus.

In addition to the bodybuilder forum corpus, which I gathered myself, there are three other corpora shared with me by other researchers, described here below. This allowed me to identify text features that consistently predict personality across several corpora. Such features are likely to be generally useful to practitioners rather than applicable only to bodybuilders. Also looking at multiple corpora helps identify features that are less likely to be just statistical noise.

Essays. The corpus with the most participants consists of 2,588 university students at the University of Texas who each wrote freely ("stream of consciousness") for 20 minutes [16]. The essays were written from 2005 through 2008, and the average word count is 787. Each student also took the Five Factor Inventory, a personality questionnaire [61] that follows the commonly used five factor model of personality (as do questionnaires for each of the corpora). Arbitrarily assigned ID numbers, instead of participants' names, help preserve the confidentiality of participants.

EAR. Another very unique data set I studied was shared with me by a team whose goal was, in their words, to "examine the expression of personality in its natural habitat" [58]. The authors audio-recorded 96 participants as they went about their daily lives over two days. Participants carried a portable recording device, called the Electronically Activated

Recorder (EAR), which samples snippets of sound 30 seconds once every 12 minutes [62]. I worked with transcripts of participants' speech that they produced from the recordings, averaging 1271 tokens per participant. They also collected participants' Big 5 personality scores via a self-report questionnaire, the BFI-44 [63].

Facebook. Finally, we were able to locate Status updates of 244 Facebook users, with average token count of 721 (after combining Status updates from each user) from a repository shared by researchers associated with the myPersonality project [64], along with scores from a personality questionnaire constructed from the IPIP [51]. The myPersonality Facebook software application that they created administered the questionnaire to users and, after they filled out the questionnaire, reported back to the user an individualized personality report. This set of participants is much smaller than their overall database which included 6 million users (they declined to give me full access). A random sample of English speaking Facebook users accesses a somewhat broader population than others that focus only on undergraduates, as the Essays corpus does. However the small document sizes for the Facebook corpus leads to very sparse appearance of features, limiting the conclusions that can be reached from this data.

Despite the smaller text size per participant amongst the shared corpora, they did add value to my work. In the process of looking for features that generally predict personality, it is useful to look at additional corpora sampled from other participant populations and under different conditions. Features that were both predictive of personality in the Forum corpus and in the other corpora are more likely to again predict personality in general applications.

The Essays corpus offers stream-of-consciousness essays, in which participants have no particular audience, writing only whatever comes to mind, as though they are journaling (although a few participants comment about their awareness that someone may sometime read their essays). In this way writers' language usage is not constrained by a particular role or audience. However the corpus is limited to only undergraduate university students.

The Facebook corpus consists of status updates, which are a technology-mediated communication in which a user makes a short announcement either to the public or to their chosen audience members as a group; since these are semi-public announcements, they offer a potentially different type of writing sample than diary-like essays or conversational

transcripts. As social media, Facebook texts may represent some similarities in language usage with other social media venues such as Twitter and Instagram, or others yet to be invented, adding to its worthwhileness as a corpus. The EAR corpus, as it consists of transcripts of speech, offers a valuable look at spoken language usage rather than author-written language.

Chapter 4

Analysis

4.1 Tutorial: constructing the personality data pipeline

This chapter outlines the steps taken to analyze the data sets and serves as a reference for anyone attempting to replicate these experiments, or for practitioners seeking to apply or extend the work. The code samples are not intended to stand on their own, but rather to help the reader envision the coding tasks necessary to apply these methods to their own data. Finally, design trade-offs come with their own advantages and disadvantages; someone else may choose to do it very differently (this is not the only way).

4.1.1 Gathering web data

Before describing the analysis process, a brief note about gathering data from the web. Three of the corpora that I used were shared to me by others; the one that I gathered myself had to be extracted from a website. In an ideal world, an API would be available or site owners would simply send me their data. However sites often lack this, it is offered but not currently working, and/or site owners are unavailable to answer crucial inquiries regarding how to proceed. This is where screen-scraping software becomes useful. After obtaining the necessary Institutional Review Board (IRB) approval and soliciting participants, I created the screen-scraping software. The code consists of scripts that read the website and regular expressions (REGEX) to extract the data from the web markup.

To create such a screen-scraper, one starts by examining and understanding the structure of the website one is scraping, as well as its Terms of Service (TOS) or equivalent document, to avoid doing anything against the valid wishes of the owners. The screen-scraper typically iterates through pages in the website; in the case of the forum it extracted user content, assigned it to the correct user, and iterated to the next page of a given forum topic. To reduce the time taken to conduct the screen scrape, it helps to run multiple processes concurrently, but within reasonable limits. One does not want to disrupt the website's operations or the flow of network traffic by generating too many simultaneous requests to the server(s). After receiving the responses, one needs to organize and review the results. It is possible that missing responses or duplicates may require attention. The process of creating a screen-scraper, for me anyway, was iterative: I reviewed the content that was read in and revised and re-ran the script until I had what I wanted.

My survey was publicly available, and I offered a USD $5.00 reward to participants. This left it vulnerable to those pretending to be valid participants, but falsely filling out the survey many times in order to receive unearned rewards. This in fact occurred twice; in each case the problem was obvious when reviewing the logs: many surveys filled out in quick succession, with suspiciously similar responses e.g. consecutive ZIP codes and email address that did not match the users (female names in email addresses being associated with male bodybuilders).

4.1.2 Extracting and formatting data

For the forum corpus (consisting of a group of participants and their writings, or their utterances reduced to text), the outputs from the screen-scraper were organized into records with unique identifiers for each participant, along with the corresponding text that the participant authored. For the borrowed corpora, some of the data was already offered nearly in this format; others took some basic preprocessing steps to bring it to that point. I could have placed the data into a database, but chose not to as the goal was to go directly to data analysis, not to build a robust system. The data was held in text files and input into data structures as needed for post-processing. Also, none of the external libraries I used employed database connections, so a database would have simply been a distraction.

Before going on it is worth noting that for confidentiality the participant identifier should not be a universally recognized identifier such as a social security number in the US, a person's physical or email address, or their name. In other words, so far as is possible, the records should be purged of any information that an observer could quickly and easily use to personally identify a participant. Applicable laws and regulations sometimes do spell out rules that must be precisely followed. If it is necessary to retain such information for any reason, such identifying information should be stored off-line in a safe location or in a way that complies with applicable law and IRB (Institutional Review Board) or national equivalent. Although it may still be possible for an industrious investigator to eventually identify a participant given a quotation that is publicly available with attribution, the emphasis here is to do one's due diligence to protect participant identities to the point that a casual observer or commonly available hostile toolkits that often crawl through data cannot extract the identities on an automated basis.

Regarding personality scores themselves, after scoring them according to the IPIP instructions, it is necessary to render them into a format comparable to the other corpora. Although the corpora that I study use various personality questionnaires to measure personality, they are all intended to rank participants for each of the dimensions in the above-discussed Five Factor Model (FFM). The challenge becomes to align the rankings between corpora. There is no explicitly specified method of doing so given different personality questionnaires. Sometimes authors standardize the scores about the sample mean and report the standardized scores, which is the practice that I have followed. However sample means most likely do not precisely match between corpora, and the difference between means could potentially be significant when the sampled populations vary dramatically in their personality makeup. So, the standardized scores (also called z-scores) are not directly comparable between corpora. Yet the shape of the score distributions, though, i.e. the way in which the scores spread out from the mean, and the slope of a regression line, should not be affected by this issue.

For the reader less familiar with the appearance of raw text data, here is an example from the Facebook corpus (Figure 4.1); the delimiters were tabs. Without any tabs in the author text itself, it is not necessary to "escape" the text (i.e. to introduce unique sequences of characters to further delimit the data fields). It is straightforward to ensure

```
ea890531d3e0547166efde52d843099e    Just wanted to thank everyone for all
    the support (and great tips) yesterday, it meant alot! made it through
    yesterday without smoking at all...and still going strong! :) using
    facebook as procrastination for practicum applications...beautiful lol.
     First day quitting smoking...wish me luck!!! you know when you work
    your \@\$\$ off for a really long time and then finally one day it all
    pays off? today was that day :)
fbe5aa478508d1dc931427ade5d9e1b4    Totalled my car last night. Luckilly no
     one was injured. Thank god I have GAP insurance. Merry Christmas. Oh
    well. Only two things to do when your team gets crushed in the playoffs
    . And I just so happen to have booze handy. And the cabin fever begins.
35efb99775d5ee7e83cf7912591984d5    Facebook me marea. Me hates it long
    time T-T
a764ca41dca158d7a191505dcc8ce47f    Red
deb899e426c1a5c66c24eeb0d7df6257    About mornings and winter,and magic.
    little things give you away.
...[any many additional records until EOF]
```

Figure 4.1: Example, input records.

Ex.																
(1)	I	am	so ADV	excited ADJ	to to	be	done	with	school!							
(2)	It's	so ADV	funny ADJ	to to	me	.										
(2)	After	a	busy	and	funfilled	weekend	it	is	extremely ADV	boring ADJ	to to	sit	in	front	of	a	computer

Table 4.1: POS n-gram examples.

this structure when processing the data by following best practices, i.e. ensuring that no delimiter characters appear within the data itself.

Next I ran a tagger on the data. The Stanford tagger [65] is popular and I find its performance satisfactory [66–68]. In its rawest output format, the tagger I used simply provides a POS tag for each word in a sentence. In my case I was interested in n-grams of consecutive, ordered groupings of the parts of speech that an author is using, sometimes in combination with words, as in 4.1. It is very desirable to configure the tagger's encoding type to match that of the input. Unfortunately data sources do not always declare the encoding style, or it is declared incorrectly. So, it is appropriate to run the tagger on a sample of the data to verify there are no obvious recurring encoding exceptions. The tagger gives warnings at the console. It is important then, to review the entire logger output and perhaps not to log excessive data to the console, which can obscure such warnings.

My code 4.2 observes text through a moving window that iterates forward one token (in this context a token is a word or punctuation mark) at a time. The extractTrigram method extracts each desired feature from the current window, resulting in a single instance of each feature being extracted from the text (the practitioner may alter the features here to restrict or explore other features.) "BOS" is a custom tag that I added to signal the beginning of a sentence. There is no particular tag for the end of a sentence, as endings vary: punctuation signals to us whether a sentence is a question, command, plain assertion, etc. The classes being imported in the code sample are described in the Stanford NLP API documentation [65].

```
// Tagging a corpus (incomplete, reader may adapt and fill in the details)
import edu.stanford.nlp.ling.*
class TagFolder {
public enum TrigramMode {
       P_P_P, P_W_P, P_P, P_W, W_P}
// OUTPUT: Hashtable<String, Integer> featuresExtracted: the $n$-grams for
    a given corpus accumulate in this hashtable over multiple calls to this
     function.
// The TaggedWord structure has two String fields, TaggedWord.tag
    consisting of the string naming the tag of the current token, and
// TaggedWord.value, the the token itself.
static private void extractTrigram(Hashtable<String, Integer>
    featuresExtracted, TaggedWord twoBack, TaggedWord oneBack, TaggedWord
    current, TrigramMode mode){
switch(mode){
case P_W:
       if(oneBack == null){
              trigram = "BOS" + current.value().toLowerCase();
       }else{
              trigram = oneBack.tag() + "" + current.value().toLowerCase()
                  ;
       }}
break;
case P_P:
       if(oneBack == null){
       trigram = "BOS" + current.tag();
}else{
       // This should never happen (should always go to ELSE)
       if(current == null){
       }else{
              trigram = oneBack.tag() + "" + current.tag();
       }
}
break;
// ...[the remaining cases follow this pattern]
featuresExtracted.put(trigram, trigramCount);
```

Figure 4.2: Code snippet illustrating use of the tagging library.

4.1.3 Other bookkeeping

Personality scores for each dimension are read from the flatfiles that identify each record by the same document IDs as the author text. Although it is possible to do normalization of the scores now, I left it to the analysis step so as to keep all the calls to the machine learning library in one place. Along the way, I keep a log detailing actions taken on the data.

4.1.4 Postprocessing

The n-gram tagging process, considering only unigrams bigrams and trigrams produces over 1 million distinct features for a reasonably sized corpus. Filtering out features such that one considers only the features fitting the criterion described at the beginning of this chapter substantially eliminates sparsely appearing features that would cause overfitting of statistically noisy data during the modeling process.

With these considerations in mind, those working on large-scale data will want to ignore irrelevant features as early as possible. For example, if one knows what features are relevant to a predictive model, one may check the feature instances as they are being extracted, and keep only the relevant features – allowing the unwanted features to be discarded as early as possible, avoiding pipeline-clogging network transfers and disk writes.

By suspending the assumption that code runs as designed (whether imported or written by the project team), one can identify potentially crippling logic bugs early in the process, saving the effort of backtracking later. Each step of the way, it is helpful to write and personally generate and review test logs confirming that (1) all the data of interest was processed (i.e. some problem did not stop the process leaving it incomplete) and (2) the planned transformations occurred. To verify the latter, I often do spot-checks on the data – and not only on the first record. I often checked to assure that columns are lined up, and feature counts are correct, especially between corpora if working with multiple datasets. Test driven development automates much of this process but if the tests are incomplete in crucial areas or nobody runs the tests and reviews the results, one can end up with incorrect results and remain unaware of the problem.

4.1.5 Hardware

Large quantities of data to analyze can strain a basic one-machine setup. Using a database that can reduce the memory footprint during key operations or distributing the jobs on a cluster can address this. Otherwise, one can get a little further with this approach by rounding or truncating trailing digits to include only as many as required (1.0339 takes less disk space to store in an Unicode text flat-file than 1.03389482725987 does). If Java is running out of memory, it will also help to verify one is actually running an up to date 64-bit JVM in a mode that coincides with one's CPU, and possibly increase the available memory. Even without memory leaks, Java projects dealing with large enough data structures in memory easily overflow the JVM memory constraints. The default limits vary according to platform but are typically 1/4 of the available memory. In this case one will use Java options that increase the JVM heap size. For example `Java -Xmx32g TagFolder` increases the available memory for use by Java to 32 Gigabytes.

Getting familiar with the state of one's system will help when addressing performance issues. The "Top" command on Linux systems (sample output shown in Figure 4.3) gives both CPU and Memory usage per process, and a summary. The disk may also be a bottleneck; even with file write buffering in place, those developing a prototype following this tutorial will see intensive Input/Output activity. Without my SSD (Solid State Disk), job execution times would have been much longer (days/weeks vs. a few hours).

4.1.6 Programming languages

Python and Java were used; those decisions were dictated by the libraries I chose, with an emphasis on quality and published empirical results showing that the tools work. Some necessary string search and replacement operations are carried out at the command prompt especially during preprocessing.

4.1.7 Identifying features of interest

For statistical analysis I employed SciKit [69], an open source project useful for machine learning. I used it to locate the features of greatest interest (those with a non-trivial number of occurrences and strongest correlation with a personality dimension after the scores are

```
top - 04:46:40 up 10:44,  5 users,  load average: 0.13, 0.14, 0.24
Tasks: 152 total,   1 running, 151 sleeping,   0 stopped,   0 zombie
%CPU(s):  1.3 us,  0.6 sy,  0.0 ni, 98.1 id,  0.0 wa,  0.0 hi,  0.0 si
KiB Mem:  32972460 total,  2049116 used, 30923344 free,   162452 buffers
KiB Swap:  8142844 total,        0 used,  8142844 free.   898272 cached Mem

PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
700 root      20   0  225708 100336  67028 S   3.0  0.3   5:36.69 Xorg
1054 bill      20   0  469168  54344  40080 S   2.0  0.2   0:10.35 konsole
3565 bill      20   0  955772 219440 125220 S   1.3  0.7   6:55.85 Web Con
986 bill      20   0 3089064 100628  69180 S   1.0  0.3   1:06.58 kwin
3529 bill      20   0 1098548 308288 146952 S   1.0  0.9   2:12.18 firefox
3244 bill      20   0 2799492  95404  62304 S   0.3  0.3   0:54.37 texmaker
3528 root      20   0       0      0      0 S   0.3  0.0   0:01.19 kworker
1 root      20   0   28852   5160   3116 S   0.0  0.0   0:01.06 systemd
2 root      20   0       0      0      0 S   0.0  0.0   0:00.00 kthreadd
3 root      20   0       0      0      0 S   0.0  0.0   0:01.34 ksoftir
```

Figure 4.3: Output of "top" command, summarizing system resource allocation.

normalized. Once the data is in the appropriate data structures the essential operations are quite straightforward. Since SciKit is in Python, it is simplest and most direct to write a Python script.

Chapter 5

Features

5.1 Feature analysis

Without any filtering applied the feature extraction in the last section outputs 323,625 features, most of them appearing very sparsely in participants' writing. As already discussed even if a sparsely appearing feature passed some statistical test I would not be confident in the result. I am interested in features appearing at least one time per 1000 tokens (computed within the Forum corpus; frequencies vary in the other corpora), yielding 766 features. Beyond that, I further reduced the feature set to include only features predictive both within the Forum corpus and in at least one of the other corpora: p-value equal to or less than 0.05 in the Forum corpus and less than 0.10 for at least one of the corpora. The p-values are associated with non-parametric Kendall Tau correlations that I computed. The reason that I use a higher p-value threshold of 0.10 for the other corpora besides the Forum is because of the much lower average word counts per participant, that in turn make features sparse, such that a lower reported p-value is less believable and possibly attributable to random sampling affects. This filtering process resulted in 75 features including two exceptions that I happened to notice and believed had reasonable explanations for their appearance: "work", with Forum p-value just over the threshold, at 0.059 and "WRB|i" (table 5.1 explains meanings of POS tags), with Forum p-value also just over the threshold, at 0.0506. Three of the features were applicable to more than one personality dimension, so altogether there are 78 feature-with-personality associations listed in the tables.

POS tag	Description	Examples	POS tag	Description	Examples
BOS	Beginning of sentence	PERIOD	Period (".")		
CC	Coordinating conjunction	e.g. and, but, or.	POS	Possessive Ending	e.g. Nouns ending in 's
CD	Cardinal Number		PRP	Personal Pronoun	e.g. I, me, you, he.
COMMA	Comma (",")	PRP$	Possessive Pronoun	e.g. my, your, mine, yours.	
DT	Determiner	the, a, these	RB	Adverb	Most words that end in -ly as well as degree words like quite, too and very.
EX	Existential there		RBR	Adverb, comparative	Adverbs with the comparative ending -er, with a strictly comparative meaning.
FW	Foreign Word		RBS	Adverb, superlative	
IN	Preposition or subordinating conjunction		RP	Particle	
JJ	Adjective		SYM	Symbol	Should be used for mathematical, scientific or technical symbols.
JJR	Adjective, comparative		TO	to	
ADJS	Adjective, superlative		UH	Interjection	e.g. uh, well, yes, my.
LS	List Item Marker		VB	Verb, base form	Subsumes imperatives, infinitives and subjunctives
MD	Modal	e.g. can, could, might, may.	VBD	Verb, past tense	Includes the conditional form of the verb to be
NN	Noun, singular or mass		VBG	Verb, gerund or present participle	
NNP	Proper Noun, singular		VBN	Verb, past participle	
NNPS	Proper Noun, plural		VBP	Verb, non-3rd person singular present	
NNS	Noun, plural		VBZ	Verb, 3rd person singular present	
LPARENS	Parenthesis	"("	WDT	Wh-determiner	e.g. which, and that when it is used as a relative pronoun
RPARENS	Parenthesis	")"	WP	Wh-pronoun	e.g. what, who, whom.
PDT	Predeterminer	e.g. all, both, when they precede an article.	WPOS	Possessive wh-pronoun	
			WRB	Wh-adverb	e.g. how, where why

Table 5.1: Index of part-of-speech (POS) tags

For reference the tables 5.4, 5.5, 5.6 and 5.7 offer additional statistical details about these features. The p-values fitting the above-mentioned criteria are in bold. Besides the specific examples of features that I discuss in this section, there are many features in the tables for which I did not see any clear explanation as to why they might be associated with personality. Some of them could be spurious noise or they could be beyond current understanding, with explanations of their association with personality to emerge in the future. Altogether there are 30 features associated with Openness, 14 with Agreeableness, 15 with Conscenciousness, 10 with Extraversion, and 9 with Neuroticism.

To interpret the tables, one starts at the leftmost column titled "personality", which indicates the personality dimension being considered, abbreviated "open", "cons", "agree", "neur" and "extra" for Openness, Conscenciousness, Agreeableness, Neuroticism and Extraversion, respectively. The table organizes the data in rows, wherein a row describes a single feature-personality dimension association. Entries in next column, headed "feature" indicate which language feature is considered in that row. Features consist of n-grams of consecutive, ordered groupings of either the parts of speech that an author is using, an actual word, or a combination thereof. Table 5.1 offers an index of the tags being used, with their meanings.

Next, there are four groups of columns, a group for each corpus. The Forum corpus represents the data from participants who worked directly with me under my IRB approved study; the other three are the corpora shared with me by other researchers, as previously described. Within each group are three columns headed: rate, τ, and *pval*. "Rate" is the frequency of appearance of the feature per 1000 tokens in the writings considered. For reasons expanded upon at the start of the Data chapter, I only included features appearing at least once per 1000 tokens in the forum corpus. The column headed τ is the Kendall rank correlation coefficient, which can range from -1 to 1, with values farther from zero indicating a stronger association between a feature and the personality dimension indicated in the leftmost column. The column *pval* indicates the likelihood of the null hypothesis: that the given feature has no association with personality. Given that numerous features are under consideration, it is quite likely that there are some features that appear significant, but are not, and arose due to inherent sampling error (sometimes referred to as randomness). I find it of interest to consider for which corpus or corpora, besides the Forum, a feature is significant.

Here I present the outcome of a sometimes tedious but ultimately affirming effort. I read through thousands of examples to identify patterns with for which I can discern plausible explanations as to why they might express the personality of the speaker. It is quite possible, of course, that although I cannot think of an explanation for some of the features, someone else would notice compelling reasons for their presence if they looked through enough examples.

When I describe specific examples of text features hereafter in this section, the format that I that use is first to indicate what personality dimension the feature is associated with, and the direction of the correlation. Within the examples, I indicate the feature under consideration as a token or series of tokens, surrounding its form with brackets. Tokens in all caps indicate POS tags as listed in Table 5.1. Also I report the relative frequency of the feature in the Forum corpus. Then, I give my explanation of why participants might tend to use that feature more or less often given their personality. Examples of notable usage instances follow the explanation. In the examples, I highlight the usage instance with brackets as well, and follow word instances with the POS tag, e.g. "why(WRB)" when the underlying feature template indicates a POS tag.

Participants in the Forum, EAR and Essays corpora scoring high on Agreeableness tend to use the form "Wh-adverb" followed by "I", such as "when I", "how I", "why I". Kendall's tau p-values are 0.0506 for Forum, 0.02708 for EAR and 0.02710 for Essays. The relative frequency is 0.00105 in the Forum corpus. The instances often describe the speaker's own actions, as in the following examples. A possible explanation is that attention to their own behavior enables them to adapt to their environment, including to other people. As a result, others tend to find them agreeable.

> I mean questioning is good and all , but come one , I drive myself insane analyzing and overanalyzing everything ! And it drives me insane . Wow , I really do n't like myself very much ; or I 'm just incredibly self-critical , which is [why(WRB) I] have so many issues . My relationship with myself ... Whatever . I guess I 'll just deal with that ... Somehow . (Essays corpus, [16])

> my questions make my friends uncomfortable sometimes , especially [when(WRB) I] point out inconstencies . I suppose I 'm trying to save them or something . maybe . (Essays corpus, [16])

> I usually do have lots of plans with my other friends , and I always enjoy spending time with them . When David does ask me to do something , however , I almost always accept his proposals very readily . [When(WRB) I] do this , I

> 'm not canceling my plans with my other friends , I just wait to see if he wants to do something with me before I tell my friends that I 'll definitely go with them . This is probably a bad thing , and I should turn him down more often so I 'm not always doing the things he suggests , but the problem is I -LRB- almost -RRB- always end up having a good time with him . Also , I usually feel like if I do n't do things with him when he suggests something that there is a big chance that I wo n't get to see him for a while . (Essays corpus, [16])

In the Forum (with Kendall Tau p-value 0.059) and essays corpora (p-value .046), Conscientious participants tend to use the word "work", speaking of work either as a place or as an activity. The relative frequency in the Forum corpus is 0.00102. Conscientiousness involves planning work and focusing on work, explaining this pattern. Examples follow.

> Uh , who cares about hardware store job ? You are doing what you need to while you finish college . Maybe start looking for some paid internship for [work] experience in your eventual field . Counts a lot after you graduate . The yoga pants are way more comfy than skinny jeans . (Bodybuilding forum corpus, [60])

> I just like good value . Eat at home . Bring my own lunch to [work] . Save as much as possible but spend when it counts . Everyone has their own personal struggles . (Bodybuilding forum corpus, [60])

> I am thinking about work because today was my first day at my new job . I am worried about how I am going to juggle work and school , especially since my new job requires me to [work] weekends as well as weekdays . I am worried that I wo n't be able to keep up my grades or that I will be exhaustively tired everyday . I want to just concentrate on school so I am thinking about changing my hours to only work maybe twice a week . (Essays corpus, [16])

> Its harder to lose weight the more you go down . Also , maybe push it more at the gym ??? Its possible in three weeks . All you need is dedication , discipline , patience , and hard [work] . No cheating either . However , if youre just starting this kind of thing now , I highly doubt you 'd be able to make it , it may sound harsh , but , trying to get " ripped " is pretty hard , especially if you 're just going to decide one day , " oh , i wan na get fit in like 4 weeks . " (Bodybuilding forum corpus, [60])

Extraverts tend to use the word "gain" in all corpora except Facebook. The relative frequency in the Forum corpus is 0.00105. The pattern is that they are speaking of either body weight in terms of personal appearance, either discussing avoiding weight gain or adding muscle (particularly amongst the bodybuilders in the Forum corpus). This way of talking about personal appearance expresses the Extravert's social orientation, in which

they are concerned with others' perceptions about their appearance, as in the following examples:

> I hear its either , lose the fat , or either [gain] the muscle , you ca n't do both . But , you should probably gain the muscle first , then lose the fat =) I know that some people get taller as they bulk up/gain muscle . Like they are fat or short . (Bodybuilding forum corpus, [60])

> I find this phenomenon intriguing , that food is some how more entertaining that some of the things that go on in my life . luckily I 'm pretty active so I do n't tend to [gain] weight even when I overeat . a special blessing from god . (Essays corpus, [16])

> It 's funny in the past few days I 've only had one meal a day . I do n't eat very much . I guess that 's kind of good so I do n't [gain] the freshman fifteen ! (Essays corpus, [16])

Extraverts use the word "with" more frequently across all corpora. The relative frequency in the Forum corpus is 0.005599. Originally I though this feature appeared because extraverts would be speaking of being socially "with" others, but that is apparently not the case after investigating the question. Amongst those scoring above one standard deviation over the mean Extraversion score versus (high scorers) and those scoring below one standard deviation below the mean (low scorers), there was no significant difference amongst them. Looking specifically at the Forum corpus, amongst 448 instances of "with" used by high scorers, 54 could be considered social instances, speaking of others, whereas the low scorers' count was 405 instances, of which 49 were social instances. Some other explanation may exist. Here are some examples of the social usage, followed by the opposite:

> Going to school , while at the University of Texas , its has a lot more stress than going to any other University . Why you ask , because its a very prestious school that any a seldom amount of individuals attend . Hanging out [with] my friends , let me know that I have people that really care about me in my life . I 'm at a place in my life now , where I know what I want out of life , which is nothing but the best . I have many goals set out for myself , because that 's the type of person I 've grown into .

> Ok , so right now I am really hungry because I just went and worked out [with] my pledge sister Andy at Gregorgy . Also I havent really eaten that much but some crap all day . That 's not my fault though because seriously Jester has some nasty food sometimes , and I would rather just eat a cookie and some red bull than eat some of the nasty stuff they serve down in the cafeteria ... I 'm really tired right now , but I think that is because I have n't gotten enough nutrition today . .

is chillin at home [with] my buddy Mike .

... gone to bed with PROPNAME ' Red Dragon . ... makin ' supper then laundry and a little homework . ... heading to the gym [with] PROPNAME , PROPNAME and my sista !! the end of the summer business season is here ... slow days of work ahead ... facebook quiz time hahahaha . Well some of them are fun . is wondering what hurricane Bill is going to be like ... getting ready just in case :o-RRB- ... work , class and gym tonight !

This . I have a very small family , at least who I regularly see , and jeans and a tee shirt going over grandma 's is normal , for us . I usually dress nicely occasionally for work and whenever I 'm out [with] friends , always . Same . Ever since I started working I am pretty strict about getting my 8 hours , and good ones too , and they still appear occasionally .

my dad on the other hand micro sleeps for a grand total of 4hrs a day , is up by 4:30 every morning , has time to read newspapers , books , magazines , workout -LRB- he 's 60 -RRB- , gardening , build a house , consultation projects and teach short courses . He used to be an econometrics professor . I was 17 , lived in a dorm supported by my parents for the first year at 18 , i moved out [with] 3 friends , we split bills , rents , food costs , cooking and cleaning evenly , did n't have many hiccups , except for parties and who would clean up the toilet bowl after Worked as a market research analyst , calling up households with surveys , paid $ 24/hr for the first 6months , then $ 35hr after that . Shifts were flexible , 3 shifts a day , allowed to balance class around it . Blew my savings on a plasma tv -LRB- $ 6K for a 42 " wtf -RRB- when it first came out , at ramen and mcdonalds for a year until I could get my lifestyle back in check .

In other cases "with" is being used to speak, not of joint activities, but in forms like this:

Very little freedom with customizing addressbook entries 6 . Ca n't watch porn I ca n't see what the Iphone5 does better than any other phone . I have n't seen any advancements [with] the OS since the first iphone . I do n't use siri , half the time I have to look up crap myself . What does the iphone5 do better that the Iphone4 or Iphone4s couldn ; t do ?

Such uses of "with" are unnecessary; this sentence could have been more directly formulated as, "I haven't seen the OS advance since the first iphone". One explanation for such instances is that Extraverts, since they are sociable and like to talk, might not think very much about sentence structure, and instead simply speak (or write) without much planning or editing.

More often for Neurotic people in all corpora except Facebook, an adjective (denoted in the examples by JJ) is followed by "and". Sometimes another adjective follows "and", as in the examples below. The relative frequency of this feature in the Forum corpus is 0.00147. Facebook users may be an exception due to the fact that Facebook status updates tend to be very brief, reducing the frequency of long sentences with clauses joined by conjunctions. In general I think that Neurotic people tend to use this form to ascribe a negative adjective to a situation then use "and" to link the adjective a clause giving evidence or reasons why the adjective fits the situation. Kentaro Hayashi also suggested during my defense [70], "Connecting adjectives using 'and' probably has an effect of making the expression more 'accurate'." This explanation seems like a strong one to me, as it stands to reason that a Neurotic person would make efforts to ensure the correctness of their pronouncements.

> Tommy Hilfiger is running a 50 % off sale styles coupon , and their slim chinos were on sale . I have a pair in navy and khaki , and while I do n't love Tommy usually , they 're some of my favorite pants . I was disappointed they only had [gray(JJ) and] burgundy left in stock , but I figured for about $ 25 , I might as well grab a new pair or two . Only size left in both colors ? 40/30 .

> You have to eat in a caloric surplus -LRB- gain weight -RRB- , and eat enough protein and fats to support muscle growth . Drop the freakin ' cardio , eat more and lift . Your diet sounds very [restrictive(JJ) and] inadequate . Are you still working with medical professionals ? You sound like you are still in the disordered thinking of " I 'm fat " .

> However , I still think your concern with gaining too much is hampering your progress , and probably your illness recoveries as well . No , that 's not true . That 115 lb stage weight is completely [dry(JJ) and] depleted . A solid 5 lbs down just in water weight which they gain back immediately after the show . Then , as they compete at 12 % or less , most of them go back up to at least 18 % between shows , and many go much higher than that if they are serious about adding muscle/filling out/bringing up lagging body parts in between shows .

Examples of the form [JJ and] not followed by another adjective:

> I think I just heard the door . nope not the door . god I really am [hungry(JJ) and] I wish she would get here now . I wish I was done typing so that I could call her already this thing really is taking forever ... Why does 20 min pass by so slowly damn thsi is boring I hope that they are not all like this cause that would not be so great I wonder what I 'll make for dinner maybe pizza

> I do n't think it 's about that as much as it is just how he was used in the rumble . His elimination was [lame(JJ) and] he could have just been a bigger

part . Even if they did n't want to have him win , okay , but they should 've had him at least be a bigger part to keep the crowd in it...

Blahhh , I pretty much sound like such a loser , but I really do care about my grades . All through high school , I was in the A/B honor roll and always did my work " overachiever style " ... At the last minute ! Hmm , I 'm so [tired(JJ) and] my elbows are starting to hurt due to the fact they are chilling on my oh-so-uncomfortable desk ! Anyway , I like my outfit today :] its pretty chic . Its a little gothalicious , but only cause its black , like I highly doubt a crazy goth would be wearing my ensemble ever .

Conscientious people tend to use the word "so", often to speak of consequences of a planned, or sometimes past action. Since they are orderly and careful, it stands to reason that Conscientious people would attend to and discuss consequences. The relative frequency of this feature is 0.00418. The correlation holds across the corpora except for Facebook. It could be that Facebook users don't use this form because it is a way of extending a sentence, and status updates tend to be short by default usage.

I like college even though it is a big transition from high school . Outside looks so inviting since the sun is n't beating down on you as soon as u walk outside . It begins to get cold again [so] I turn off the fan . My roommate resumes watching television . It is very dark in my room except for the light of the television and my laptop .

well yes and no . I got my first guitar lessons on classical flamemco guitar at the age of 10 . Got bagged out school for the crap I was forced to learn [so] I gave it up . I went back to acoustic guitar after getting into opeth around 2000 . Try easy songs where there 's easy chord changes and arpegiated open chords .

lol , i couldnt wait bro , im out of myo and i had to grab some putrid whey from walmart i do love your comp tho , invent more products ! non stim pre workout drink . i thought of that , but you need more whey and carbs [so] i just mix creatine with real gains for pre lol , i do too . anything animal is the sh!t ... ... take it , you can take the red pill out if you dont want the stim . got it today , imo , the choc is still the best of the flavors , and xf up2 .0 is still the best choc pb ... ...

I do n't understand why I ca n't just be confident in my abilities . If I am ever going to make it in the business world , I am going to have to build up my confidence . Women already have a disadvantage , [so] I need to really step it up . I hate when I am thinking of a word that I want to use in a paper , but I just can not recall what the word is . It is so frustrating , because I never want to put in a lesser word , but sometimes I am forced to .

> I never used to get really bad headaches ; just sinus headaches . Lately , however , I 've been getting such bad headaches that I even got sick to my stomach and threw up . I also feel like I never get any restful sleep [so] I 'm constantly tired which makes it increasingly hard to study . I 'm trying really hard to keep with all my reading assignments . I am pretty much ordered to read a book ever few days in some of my classes .

Conscientious people people tend to use the word "week", often because they are carefully planning and reviewing their use of future time, as in these examples below. This pattern is apparent in the Forum and Essays corpora. The relative frequency of this feature is 0.00122 in the Forum corpus.

> Already paid it off once . Now I have to do it again . Oh well hopefully by next [week] .

> There 's a lot of horror stories out there is all I 'm saying . A lot of worry over nothing . Follow the good docs advice and get a lot of rest this [week] and you 'll be feeling better then ever soon Best of luck on a speedy recovery ! Beautifully put . She is awesome in so many ways , I just sub 'd her channel .

> Well , it was still worth it , McCombs is a much better business school . Agh ! My head is filling with thoughts for the next [week] ! Too much stuff going on and not enought time ! At least I do n't have any tests this week , which is a good thing .

> And what they are saying like the meetings are going to be on wednesdays was it ? Tuesdays or wednesdays something like that . Idontknow , but I have to make sure that I do n't schedule any classes during that time , because they have it the same time every [week] next semester . so , we already know when they are which is really good so that i do n't schedule classes . it 's like later , like 4 or something .

> I like to run a lot . I do n't know what else to write . Next [week] I have a timed two and a half mile run for rowing . I think it 'll be alright . I 'm excited to try something new with rowing since I have never done it before .

Some of the features that I found confirm the observations of others, for example the word "day", positively correlated with Agreeableness, was also located in [50]. Also the LIWC [19] category "Job/Work" is related to Conscenciousness in the same study, as was the lexicon entry "work" in my study. Another group working with the MyPersonality [64] data found [71] several features same or closely related to mine: in my study, "with" was associated with Extraversion, as were the similar features "wit", "out with" and "time with" in their study; also "day" and "be" associated with Agreeableness; Table 5.2 illustrates

personality dimension	Feature as described here in my work	Similar feature found by [2]
open	VBD|n't, VBP|n't|VB	doesn't, didn't, i can't, isn't, i don't
	BOS|what	what's
	do|RB|VB, do|RB	i don't
cons	work	work, off to work, work tomorrow, day at work, at work, back to work, to work
	so	so ready, so excited
	week	for the week, a great week, weekend, this weekend, the weekend, for the weekend, great weekend, a great weekend
	TO|get	ready to get
extra	with, NN|with	wit, out with, with the girls, night with
agree	MD*|n't|VB	! can't wait

Table 5.2: This table illustrates a subset of my language cues that are the same or very similar to those found by [2] and listed detail on their website [3]. These features are each positively correlated with the personality dimension labelled in the leftmost column. *Part of speech tag "MD" denotes a Modal (usually "can" but also includes: could, might, may). A complete list of the tags and their meanings are in Table 5.1.

some additional features held in common with that study. This confirms that the method is working and motivates further exploration.

5.2 Predictive models

To demonstrate that it is possible to create predictive models from this data, I performed multiple regression on the final set of 78 filtered features with personality score as the labels (i.e., target variable). In Table 5.3, I have reported the R^2 error of the models (closer to 1 indicates a closer fitting model). The MSE (mean squared error) and S, the standard error are in the table so as to give an idea of model fit without the normalization added by R^2. Each model was trained on the all of the available data for the given corpora. Cross validation results would be likely be of some interest but to be fully valid would require a hold-out set from the beginning (before feature selection), which could leave out valuable information especially in the case of the Forum corpus which only had 50 participants.

dimension	corpus	R^2	MSE	S
open	forum	0.6858	0.3142	0.5605
	facebook	0.0836	0.9164	0.9573
	EAR	0.0409	0.9591	0.9793
	essays	0.0070	0.9930	0.9965
cons	forum	0.4833	0.5167	0.7188
	facebook	N/A	N/A	N/A
	EAR	0.1357	0.8643	0.9297
	essays	0.0225	0.9775	0.9887
extra	forum	0.4019	0.5981	0.7734
	facebook	0.0290	0.9710	0.9854
	EAR	0.1143	0.8857	0.9411
	essays	0.0068	0.9932	0.9966
agree	forum	0.3237	0.6763	0.8224
	facebook	0.0236	0.9764	0.9881
	EAR	0.1439	0.8561	0.9253
	essays	0.0120	0.9880	0.9940
neur	forum	0.3045	0.6955	0.8340
	facebook	0.0038	0.9962	0.9981
	EAR	0.0870	0.9130	0.9555
	essays	0.0099	0.9901	0.9950

Table 5.3: Models, multiple regression. S is Standard Error of the estimate.

person-		Essays			Forum			EAR			Facebook		
ality	feature	rate	τ	*pval*	rate	τ	*pval*	rate	τ	*pval*	rate	τ	*pval*
open	DT|CD	0.8832	*0.0230*	**0.0792**	1.3670	*0.1967*	**0.0438**	2.1521	-0.0305	0.6594	0.8694	*0.0976*	**0.0232**
open	VBP|n't|VB	3.7287	0.0023	0.8610	3.0259	*0.1975*	**0.0430**	4.6021	0.0481	0.4876	1.2615	*0.0776*	**0.0710**
open	,|NNS	0.2860	*0.0283*	**0.0311**	1.1232	*0.2070*	**0.0339**	0.2897	0.0010	0.9887	0.7785	0.0423	0.3253
open	do|RB	4.9198	-0.0150	0.2527	3.8174	*0.1988*	**0.0416**	5.9016	0.0502	0.4691	1.6764	*0.1102*	**0.0103**
open	NN|,	8.3050	0.0196	0.1361	14.1002	*0.2348*	**0.0161**	8.2109	0.0228	0.7426	10.1604	*0.1120*	**0.0092**
open	DT|NN|,	2.0176	0.0193	0.1417	2.7992	*0.3122*	**0.0014**	1.7299	0.0265	0.7022	1.7332	*0.1323*	**0.0021**
open	!	2.2830	*0.0309*	**0.0186**	2.0289	*0.2660*	**0.0064**	0.3476	-0.0512	0.4597	15.5418	0.0420	0.3287
open	BOS|what	0.5221	0.0050	0.7054	1.0782	*0.2108*	**0.0308**	3.973	*0.1390*	**0.0449**	0.5683	0.0673	0.1174
open	,	24.7694	*0.0225*	**0.0856**	36.9892	*0.2449*	**0.0121**	37.0649	0.0598	0.3882	27.1115	*0.0898*	**0.0368**
open	NN|,|NN	0.5195	*0.0383*	**0.0035**	2.0676	*0.2387*	**0.0144**	1.5561	0.1083	0.1180	1.9434	*0.1340*	**0.0018**
open	only	1.4466	-0.0144	0.2729	1.5019	*0.2093*	**0.0320**	0.8525	0.0791	0.2534	1.1649	*0.1055*	**0.0141**
open	,|i|VBP	2.7261	0.0000	0.9998	1.5152	*0.2608*	**0.0075**	2.7811	0.0256	0.7121	1.0740	*0.1175*	**0.0063**
open	NN|,|CC	2.3371	0.0056	0.6681	2.9055	*0.2080*	**0.0330**	1.1174	0.0174	0.8016	1.8980	*0.0930*	**0.0304**
open	ca	1.7871	0.0000	0.9982	1.0528	*0.2089*	**0.0323**	1.1754	-0.0669	0.3342	1.3865	*0.0879*	**0.0409**
open	PRP|do	4.4775	*-0.0224*	**0.0880**	2.8128	*0.1978*	**0.0427**	5.6284	0.0000	1.0000	1.1820	*0.1404*	**0.0011**
open	PRP|do|RB	3.7699	-0.0208	0.1133	2.1697	*0.1927*	**0.0483**	4.7262	0.0522	0.4514	0.8694	*0.1201*	**0.0052**
open	,|but	3.3770	-0.0026	0.8438	3.3003	*0.2684*	**0.0060**	1.6968	0.0659	0.3418	1.4206	*0.1210*	**0.0049**
open	do|RB|VB	3.9344	-0.0148	0.2582	3.1608	*0.1984*	**0.0421**	4.8918	0.0392	0.5718	1.3922	*0.0904*	**0.0354**
open	was	6.0087	0.0161	0.2204	4.3367	*0.2306*	**0.0181**	8.7903	-0.0760	0.2728	3.2106	*0.0719*	**0.0944**
open	,|NN	1.0971	*0.0290*	**0.0272**	4.1229	*0.2527*	**0.0096**	4.5441	0.0621	0.3702	3.4720	*0.1287*	**0.0028**
open	ca|RB	1.7832	0.0010	0.9364	1.0462	*0.2123*	**0.0296**	1.1754	-0.0669	0.3342	1.3638	*0.0842*	**0.0503**

Table 5.4: List of features. The Kendall rank correlation coefficient is denoted by τ; rate is frequency per 1000 tokens. Bolded p-values fit the filtering criterion: $p < 0.05$ for the Forum data (the data that I gathered) and $p < 0.10$ for at least one of the borrowed corpora.

person-		Essays			Forum			EAR			Facebook		
ality	feature	rate	τ	*pval*	rate	τ	*pval*	rate	τ	*pval*	rate	τ	*pval*
open	NN|,|and	1.0163	*0.0272*	**0.0378**	1.4852	*0.2066*	**0.0342**	0.5215	-0.0763	0.2709	1.2331	*0.0986*	**0.0218**
open	RB|,	4.0477	0.0005	0.9689	4.3580	*0.2039*	**0.0367**	5.5208	0.0823	0.2346	2.6367	*0.0879*	**0.0409**
open	BOS|RB	5.7043	-0.0071	0.5886	6.5117	*0.2062*	**0.0347**	8.5337	*0.1360*	**0.0497**	6.5861	0.0025	0.9538
open	,|CC	6.4634	0.0111	0.3971	7.6237	*0.2235*	**0.0220**	3.4764	0.0361	0.6026	4.3869	*0.1215*	**0.0047**
open	VBD|n't	1.4221	0.0150	0.2525	1.0493	*0.2218*	**0.0230**	1.8872	-0.0167	0.8099	0.6535	*0.1035*	**0.0160**
open	IN|DT|JJ	3.4702	0.0075	0.5650	4.3594	*0.2323*	**0.0173**	2.0444	0.0559	0.4201	4.2562	*0.0888*	**0.0389**
open	RB|IN|DT	1.4328	-0.0058	0.6559	1.3642	*0.1916*	**0.0496**	0.8525	0.1119	0.1062	1.3127	*0.0885*	**0.0396**
open	how	2.5123	*0.0257*	**0.0497**	2.3619	*0.2376*	**0.0149**	2.3921	0.0461	0.5060	1.6820	*0.0745*	**0.0829**
open	PRP|,	1.3955	*0.0327*	**0.0127**	1.5284	*0.2326*	**0.0172**	1.8292	-0.0199	0.7740	1.0115	*0.1335*	**0.0019**
cons	work	1.0266	*0.0262*	**0.0456**	1.0200	0.1843	0.0589	0.538	*-0.1316*	**0.0576**	2.0628	0.0405	0.3457
cons	BOS|i|MD	1.9442	0.0169	0.1963	1.2414	*0.3116*	**0.0014**	1.3906	*0.1165*	**0.0927**	0.5626	*-0.0766*	**0.0749**
cons	RB|IN|DT	1.4328	*0.0454*	**0.0005**	1.3642	*0.2033*	**0.0373**	0.8525	-0.0352	0.6112	1.3127	*-0.0711*	**0.0982**
cons	so	7.9529	*0.0484*	**0.0002**	4.1777	*0.2048*	**0.0359**	6.3568	*0.1392*	**0.0445**	4.0971	-0.0050	0.9079
cons	VB|TO|VB	2.5784	*0.0219*	**0.0946**	2.0051	*0.2044*	**0.0363**	1.5561	-0.0841	0.2251	1.4434	*-0.1317*	**0.0022**
cons	VBG|TO|VB	3.1241	*0.0334*	**0.0110**	2.0878	*0.2314*	**0.0177**	4.362	0.0128	0.8533	2.5401	*-0.1005*	**0.0194**
cons	VBP|TO|VB	4.7777	*0.0421*	**0.0013**	2.3549	*0.2007*	**0.0397**	2.9467	0.0179	0.7964	1.7559	0.0131	0.7605
cons	week	0.5685	*0.0613*	**0.0000**	1.2216	*0.2656*	**0.0065**	0.2897	0.1045	0.1314	1.0342	*-0.0721*	**0.0935**
cons	NN|WDT	1.5775	-0.0154	0.2412	1.3506	*-0.2875*	**0.0032**	1.2416	*-0.1242*	**0.0731**	0.8922	-0.0226	0.5991
cons	TO|VB|PRP	1.9644	*0.0258*	**0.0492**	1.4660	*0.1914*	**0.0498**	1.9617	0.1100	0.1122	1.4320	-0.0545	0.2046
cons	i|RB	6.4230	*0.0223*	**0.0897**	2.2377	*0.3061*	**0.0017**	2.1189	0.0831	0.2304	1.3809	-0.0574	0.1815
cons	JJ|NN	15.9346	*-0.0486*	**0.0002**	22.6320	*-0.2242*	**0.0216**	11.1658	*-0.1611*	**0.0201**	23.9520	0.0059	0.8908
cons	TO|get	1.2448	*0.0336*	**0.0105**	1.0127	*0.2102*	**0.0312**	0.9933	0.1074	0.1210	0.9944	-0.0703	0.1018
cons	PRP|RB|VBP	3.9584	*0.0362*	**0.0058**	1.8834	*0.1957*	**0.0449**	1.465	0.0699	0.3128	0.8240	-0.0437	0.3092
cons	i|RB|VBP	3.6119	*0.0359*	**0.0062**	1.0542	*0.2672*	**0.0062**	0.8112	0.0247	0.7214	0.5057	-0.0115	0.7883

Table 5.5: List of features (continued). The Kendall rank correlation coefficient is denoted by τ; rate is frequency per 1000 tokens. Bolded p-values fit the filtering criterion: $p < 0.05$ for the Forum data (the data that I gathered) and $p < 0.10$ for at least one of the borrowed corpora.

person-		Essays			Forum			EAR			Facebook		
ality	feature	rate	τ	*pval*	rate	τ	*pval*	rate	τ	*pval*	rate	τ	*pval*
agree	NN|on	1.2306	-0.0050	0.7020	1.6582	*0.2630*	**0.0070**	0.927	*0.1513*	**0.0290**	1.5627	-0.0085	0.8424
agree	MD|n't|VB	2.3577	0.0087	0.5068	1.8015	*0.2158*	**0.0270**	1.4733	*0.1173*	**0.0904**	1.5229	0.0470	0.2740
agree	VB|TO	4.1589	-0.0185	0.1591	2.9345	*0.1933*	**0.0476**	3.0708	*0.1191*	**0.0855**	2.4890	0.0090	0.8342
agree	WRB|i	2.3036	*0.0290*	**0.0271**	1.0469	0.1908	0.0506	0.9519	*0.1531*	**0.0271**	0.6705	0.0675	0.1164
agree	day	1.1190	*0.0357*	**0.0065**	1.5849	*0.2773*	**0.0045**	0.596	-0.0179	0.7963	2.9777	*0.0763*	**0.0758**
agree	BOS|CC	3.1705	*0.0258*	**0.0493**	1.8569	*0.2635*	**0.0069**	5.0573	*0.1976*	**0.0044**	1.6082	0.0410	0.3401
agree	VBP|a	1.6183	-0.0196	0.1346	1.7684	*0.2094*	**0.0319**	1.2085	0.0854	0.2180	1.0570	*0.1513*	**0.0004**
agree	for|PRP	1.1099	*0.0515*	**0.0001**	1.0964	*0.2523*	**0.0097**	0.778	-0.0262	0.7052	0.5626	0.0350	0.4151
agree	ca|RB	1.7832	0.0058	0.6566	1.0462	*0.2025*	**0.0380**	1.1754	*0.1249*	**0.0715**	1.3638	0.0090	0.8350
agree	PRP|VBP|a	1.2881	-0.0168	0.2000	1.1783	*0.2174*	**0.0259**	1.0015	0.0861	0.2139	0.6137	*0.1047*	**0.0148**
agree	ca	1.7871	0.0053	0.6885	1.0528	*0.2042*	**0.0364**	1.1754	*0.1249*	**0.0715**	1.3865	0.0030	0.9450
agree	NNP|VBZ	1.0949	*-0.0246*	**0.0604**	1.0535	*-0.1944*	**0.0464**	0.389	-0.0786	0.2563	2.2276	-0.0383	0.3733
agree	on|DT|NN	1.0558	0.0209	0.1109	1.2795	*0.2026*	**0.0379**	0.8277	*0.1964*	**0.0046**	1.1820	0.0398	0.3544
agree	be	6.2079	*0.0222*	**0.0902**	5.7577	*0.2107*	**0.0308**	3.882	-0.0384	0.5798	5.0745	0.0124	0.7735
neur	a|NN|NN	1.0803	-0.0075	0.5683	1.7293	*-0.1950*	**0.0456**	1.0264	*-0.1543*	**0.0259**	1.7389	-0.0344	0.4230
neur	JJ|CC	3.0323	*0.0316*	**0.0160**	2.0547	*0.1939*	**0.0470**	1.1257	*0.1563*	**0.0241**	1.7616	-0.0416	0.3327
neur	JJ|and	2.2199	*0.0353*	**0.0072**	1.4716	*0.2462*	**0.0117**	0.7449	*0.1216*	**0.0791**	1.3468	*-0.0854*	**0.0470**
neur	VB|IN|PRP	1.7931	*0.0303*	**0.0208**	1.0755	*0.2246*	**0.0214**	1.5561	-0.0067	0.9231	0.8240	*-0.1045*	**0.0151**
neur	PRP|would	1.9245	*0.0258*	**0.0488**	1.4465	*0.2077*	**0.0333**	1.2498	-0.1113	0.1082	0.4319	-0.0105	0.8068
neur	PRP$|NN|NN	1.3834	*-0.0426*	**0.0012**	1.1602	*-0.2061*	**0.0347**	0.8525	0.1050	0.1297	1.5627	-0.0090	0.8340
neur	DT|NN|NN	2.9992	*-0.0224*	**0.0876**	3.8248	*-0.2373*	**0.0150**	2.6238	-0.0048	0.9449	4.5915	-0.0283	0.5103
neur	NN|NN	9.3411	*-0.0477*	**0.0003**	18.7762	*-0.2041*	**0.0365**	9.6677	-0.0507	0.4640	24.9123	-0.0207	0.6299
neur	NNS|VBP	2.8262	-0.0106	0.4208	3.8771	*-0.1917*	**0.0495**	1.2416	*-0.1533*	**0.0269**	2.9208	*-0.0784*	**0.0683**

Table 5.6: List of features (continued). The Kendall rank correlation coefficient is denoted by τ; rate is frequency per 1000 tokens. Bolded p-values fit the filtering criterion: $p < 0.05$ for the Forum data (the data that I gathered) and $p < 0.10$ for at least one of the borrowed corpora.

person-		Essays			Forum			EAR			Facebook		
ality	feature	rate	τ	*pval*	rate	τ	*pval*	rate	τ	*pval*	rate	τ	*pval*
extra	NN|with	1.1786	*0.0420*	**0.0014**	1.4046	*0.2382*	**0.0147**	0.5297	0.0696	0.3153	1.8525	*0.0732*	**0.0884**
extra	IN|NNS	2.8614	-0.0213	0.1048	4.0131	*0.1947*	**0.0460**	1.4568	*0.1169*	**0.0916**	2.5572	-0.0102	0.8123
extra	VBZ|VBG	2.6875	0.0033	0.8027	1.1905	*-0.2989*	**0.0022**	1.7134	*-0.1214*	**0.0796**	5.3814	0.0366	0.3943
extra	with	4.7798	*0.0313*	**0.0169**	5.5994	*0.1925*	**0.0486**	2.3838	*0.1763*	**0.0110**	5.4268	*0.0899*	**0.0364**
extra	JJ|JJ|NN	0.6449	-0.0022	0.8689	1.1738	*0.2168*	**0.0263**	0.4883	*0.1496*	**0.0308**	1.4888	0.0318	0.4594
extra	RQUOTE	1.4135	0.0093	0.4803	2.8989	*-0.2555*	**0.0088**	0.389	*-0.1202*	**0.0828**	5.3416	-0.0481	0.2632
extra	it|.	1.6123	*-0.0290*	**0.0269**	1.4611	*-0.2155*	**0.0272**	3.0046	-0.0108	0.8757	1.1195	-0.0427	0.3209
extra	by	1.1954	-0.0160	0.2223	1.2690	*0.2650*	**0.0066**	0.4552	-0.0067	0.9228	1.5854	*0.1195*	**0.0054**
extra	gain	0.0447	*0.0423*	**0.0013**	1.0479	*0.2175*	**0.0258**	0.0083	*0.1347*	**0.0519**	0.0114	-0.0157	0.7150
extra	RB|JJ|NN	1.4109	0.0055	0.6775	1.2533	*0.2052*	**0.0355**	0.6787	*0.1870*	**0.0069**	1.4945	*0.0782*	**0.0687**

Table 5.7: List of features (continued). The Kendall rank correlation coefficient is denoted by τ; rate is frequency per 1000 tokens. Bolded p-values fit the filtering criterion: $p < 0.05$ for the Forum data (the data that I gathered) and $p < 0.10$ for at least one of the borrowed corpora.

Chapter 6

Conclusion

6.1 Summary of work

This work investigates the question of whether there are grammatical language features that are predictive of personality. To confidently answer the question, I gathered both my own data and extended the investigation to three additional corpora from other research groups. This is the first effort of its kind in terms of looking at multiple corpora. The answer is yes: I found specific instances of features that are associated with personality across a variety of contexts, which I discussed in detail in the previous section. The complete list, with or without explanations found, consists of 78 language feature-with-personality dimension associations (there were 75 features and 3 of them were associated with two personality dimensions), as shown in the tables 5.4, 5.5, 5.6 and 5.7. The language features that I found to be predictive of personality amount to 30 features associated with Openness, 14 with Agreeableness, 15 with Conscenciousness, 10 with Extraversion, and 9 with Neuroticism. 16 of the feature-personality associations that I report in these tables are bag of words features. The remaining 62 are features involving only POS tags, or n-grams of POS tags combined with a word unigram.

Ultimately this work firmly establishes the connection between human personality and such language features, and sets the foundation for future extension of the effort into other features extracted by additional Natural Language Processing (NLP) tools whether new or existing. The work fits well into the broader Artificial Intelligence endeavor when it comes to computer-driven agents that represent, express or detect human personality as

well as recognizing psychopathology in individuals and predicting the behaviors associated therewith. The numerous imaginable future possibilities in terms of adapting systems to their users and even making previously inconceivable applications possible motivate ongoing work in this area.

Also I offer a methodology for extracting and comparing features across corpora, and for model building, and carried out the method to completion. The four corpora that I worked with include a variety of styles of expression, which I believe is important because it reveals the possibility of building models that can predict personality of individuals whose expression differs somewhat from those whose data was used to create a model.

6.2 Limitations

The Forum corpus was drawn from a place where, although many topics are discussed, the participants share a passion for one central interest (bodybuilding), introducing a situational bias that may involve patterns of language usage differing from the broader population. Frequencies of some of the features could be affected, limiting the validity of their correlations with personality dimensions. The cross-corpus analysis addresses this issue by considering only features from the Forum corpus that are also associated with the same personality dimension in at least one of the additional shared corpora. Also discerning reasonable explanations of some of the features' associations with personality helps to address this concern.

The research focuses on English to the exclusion of other languages which may lack or have completely different grammar structures associated with personality. Also it is likely that there are styles of language for which the previously known predictive language features would not be applicable, especially when circumstances require speakers to adapt their self-expression. For example Twitter and SMS text messaging via cell phones require short, concise expressions. Private channels observed only by the participants may incorporate language features absent otherwise, or affect the usage of the known features. There are also more formal styles than the ones I examined, such as academic writing or job interviews. This possible variety of expression is somewhat of a confounding aspect when it comes to building practical models. In practice such situations may require training by a sample of

individuals of known personality who have expressed themselves in a way that is similar in style to the texts that are of interest.

With only 50 participants in my Forum study, I was limited in the amount of data to select features from, so that I had to disregard many potential features that may indicate personality but were too sparse in the data to consider; I examined only features appearing on average one or more times per 1000 words. Another constraint that limited my research is that although it possible that there are larger structures that indicate personality, I considered only unigrams, bigrams and trigrams. In general frequencies of larger structures are lower and require larger amounts of data to provide an adequate volume of data for study.

Although they all conducted assessments of the same basic five dimensions of human personality, the shared corpora that I incorporated into my research employed different personality questionnaires to perform that task. For cross-corpora analysis, I normalized the scores for each corpora about the respective means. However since the samples were from different populations, there is a strong chance that the process introduced inaccuracies that limit the value of comparing the normalized scores. So then, if one were to rank individuals from the various corpora on the basis of those personality scores (even after the attempt at normalization), there would be errors in the ranking. This issue may in part explain the substantial variations that can be seen between the corpora in the correlation slopes fitting the features to the scores.

An additional limitation I faced in my research is that one of the shared corpora I employed for the cross-corpora analysis, the Facebook data, has an average of only 721 tokens per participant, and that was only made possible by considering many separate, smaller messages from each participant. So, notwithstanding the statistical tests, inclusion or exclusion of features on the basis of this very sparse data may rest upon noisily random patterns introduced by sparsity of data rather a more empirically reliable basis. A future research experiment could be to eliminate this corpora from consideration and see what features then arise as significant between the remaining corpora.

6.3 Future

Practical applications include building models optimized to a given population, creating Chatbots (text based language generators that respond to the computer user in a conversational-style interaction) and agents for various purposes that simulate humans of specified personality. Context aware intelligent agents that can detect user personality would be able to tailor their behaviors to better serve the user. Perhaps more challenging is creating an agent that can produce natural language that has a given personality, such as a GPS guidance system whose interactive voice prompts are adjustably Extroverted or Introverted. Tutoring systems can be adapted to students. Marketing messages can be adapted to personality. Customer service teams can practice interactions with agents that simulate challenging customer scenarios.

In the future it would be worthwhile to consider linguistic tools that dig further than a POS tagger does. Although my POS features have some significant associations with personality, POS tagging remains at a more superficial level of analysis than necessary given that tight relationships that often exist between words that are sequentially separated by an arbitrarily large number of less related words in a sentence. POS taggers are focused on locally situated structures such as, in English, adverbs and their verbs, and adjective-noun pairs. POS n-gram feature extraction tends to generate numerous different word-order dependent features whose count is too sparse for analysis, but are mutually equivalent if abstracted in the way that a dependency parser does. For example “I think” might be counted but “I really think” might be excluded as a POS feature due to sparse counts (in a given corpus), whereas a dependency parser could combine them into a single feature, for this purpose, giving them equal consideration. Also when multiple noun phrases are coordinated to form a compound subject, a dependency parser could help to find corresponding relationships within expressions like “my friend thinks" and “my friend and I think”. Many additional such relationships can be extracted by a dependency parser, which, instead of focusing on content within phrases, looks for relationships between pairs of words and entire phrases in an utterance, whether local or not.

A dependency based parser is suitable for finding a noun or entire phrase that serves as the subject of a verb when separated by an arbitrary number of words. The effect is especially strong in languages with free word order. These tools have also begun to reflect the rich

taxonomies that linguists use to describe many additional relationships in addition to the basic subject-predicate relation, e.g. prepositional phrases, and direct and indirect objects, as well as appositional and adjectival modifiers (POS taggers provide less informative labels). [72, pp. 3-4]

Computationally useful dependency-based tools are available; one worthy of attention is that of Danqi Chen and Christopher Manning, who have announced a neural network powered tool [73] that produces dependency parses of sentences along with fully annotated dependency types. Although localized POS tagger based features have contributed to the understanding of how personality is expressed in language, I anticipate that in future analysis, abstractions independent of word order will tell much more.

Beyond personality, these methods are applicable to psychopathology too. The most obvious case: Anxiety Disorders, which are related to Neuroticism [70]. One group recently explored linguistic markers of disordered affective states (depressed or anxious), focusing specifically on "absolutist" thinking as indicated by usage of words in a dictionary that they created, guided by their knowledge and intuition about those who are in such states [74]. Their participants were individuals involved in online discussions of suicide and emotional challenges, as well as others who served as a control group. My approach would be likely to discover numerous features in such data that are not necessarily absolutist: words or even larger grammatical constructs that are associated with depressed or suicidal states. The same would be the case for the Twitter study [44] discussed in the literature review, which examined the association of "Dark Triad" traits (Narcissism, Machiavellianism, and Psychopathy) with word categories extracted by the LIWC tool. There are well known psychopathology tests such as the Levenson Self-Report Psychopathy Scale (LSRP) [75] that measures the degree to which the subject is similar to a diagnosed Psychopath; these and whatever other psychopathology scales are of interest could be employed in prediction from language features.

As already noted models incorporating online learning, whereby the predictive model is adjusted as new data appears might yield productive applications, especially when the data is unlabeled (i.e. subjects of unknown personality). Ongoing examination of unstudied populations promises to both surprise and inform. Although a step of analysis removed from my own work, I hope that as more knowledge of the connection between language and

personality accumulates, we will gain deeper insights about personality and what it is to be human.

Appendix A

Personality Questionnaire

Below is the prompt and entire questionnaire that I administered to Forum participants.

Personality Questionnaire

This survey consists primarily of a personality assessment (50 questions). In addition, you will be asked a few general demographic questions at the end.

The personality assessment portion of this survey consists of the IPIP 50-item questionnaire, and was developed by a psychology researcher, Lewis R. Goldberg, cf. Goldberg, L. R., Johnson, J. A., The International Personality Item Pool and the future of public-domain personality measures. Journal of Research in Personality, 40, 84-96 (2006).

This survey will take approximately 10 minutes to complete.

Instructions: How accurately can you describe yourself? Describe yourself as you honestly see yourself, in relation to your peers. Describe yourself as you usually are now, not as you wish to be in the future. Your test results depend on your accurate response. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence.

I have at least 10 posts on forum.bodybuilding.com (necessary for participation)

True

False

Possible responses to the following questions were:

Very Inaccurate

Moderately Inaccurate

Neither Inaccurate nor Accurate

Moderately Accurate

Very Accurate

The 50 personality questions (items):

I am the life of the party.

I feel little concern for others.

I am always prepared.

I get stressed out easily.

I have a rich vocabulary.

I don't talk a lot.

I am interested in people.

I leave my belongings around.

I am relaxed most of the time.

I have difficulty understanding abstract ideas.

I feel comfortable around people.

I insult people.

I pay attention to details.

I worry about things.

I have a vivid imagination.

I keep in the background.

I sympathize with others' feelings.

I make a mess of things.

I seldom feel blue.

I am not interested in abstract ideas.

I start conversations.

I am not interested in other people's problems.

I get chores done right away.

I am easily disturbed.

I have excellent ideas.

I have little to say.

I have a soft heart.

I often forget to put things back in their proper place.

I get upset easily.

I do not have a good imagination.

I talk to a lot of different people at parties.

I am not really interested in others.

I like order.

I change my mood a lot.

I am quick to understand things.

I don't like to draw attention to myself.

I take time out for others.

I shirk my duties.

I have frequent mood swings.

I use difficult words.

I don't mind being the center of attention.

I feel others' emotions.

I follow a schedule.

I get irritated easily.

I spend time reflecting on things.

I am quiet around strangers.

I make people feel at ease.

I am exacting in my work.

I often feel blue.

I am full of ideas.

Your username on the forum forum.bodybuilding.com (so we can locate your public forum postings)?

Do you identify as:

Male

Female

Other

Decline To Respond

What year were you born?

If you are in the U.S.A., what is your zipcode? If outside the U.S.A., please enter your country. To decline offering this information for any reason whatsoever, type "DECLINE"

What is your maximum current level of educational achievement?

None / Elementary School / Some high school / Completed High School / Some college / Associate Degree / Bachelor's Degree / Master's Degree / Ph.D.

What language are you most fluent in?

Your email address (for us to send you your personality questionnaire results and your Amazon.com Gift Certificate):

Regarding payment, do you require an alternative payment method for your $5.00 reward, instead of a gift card for use on Amazon.com's USA website? (If so, indicate your preferred method in the box below)

Bibliography

[1] "Wiktionary:frequency lists, tv and movie scripts." https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists, https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/TV/2006/1-1000, https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/TV/2006/9001-10000. Online; accessed February 2019.

[2] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, *et al.*, Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[3] "Penn World Well-Being Project." http://www.wwbp.org/data.html, 2013. [Online, last accessed Nov 21, 2019].

[4] B. Griffin and I. G. Wilson, Faking good: self-enhancement in medical school applicants *Medical education*, vol. 46, no. 5, pp. 485–490, 2012.

[5] A. F. Furnham, Knowing and faking one's five-factor personality score *Journal of Personality Assessment*, vol. 69, no. 1, pp. 229–243, 1997.

[6] W. R. Wright, "Literature Review." http://www.williamwright.info/downloads/litreview.pdf, 2012. [Online; accessed 2-March-2013].

[7] A. Roshchina, J. Cardiff, and P. Rosso, User Profile Construction in the TWIN Personality-based Recommender System *Sentiment Analysis where AI meets Psychology (SAAIP)*, p. 73, 2011.

[8] M. Kosinski, D. Stillwell, and T. Graepel, Private traits and attributes are predictable from digital records of human behavior *Proceedings of the National Academy of Sci-*

ences, vol. 110, no. 15, pp. 5802–5805, 2013.

[9] W. R. Wright and D. N. Chin, Personality profiling from text: introducing part-of-speech N-grams in *International Conference on User Modeling, Adaptation, and Personalization*, pp. 243–253, Springer, 2014.

[10] J. Pennebaker and L. King, Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[11] G. Allport and H. Odbert, Trait-names: A psycho-lexical study. *Psychological monographs*, vol. 47, no. 1, p. i, 1936.

[12] G. Saucier and L. Goldberg, Lexical studies of indigenous personality factors: Premises, products, and prospects *Journal of personality*, vol. 69, no. 6, pp. 847–879, 2002.

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

[14] P. Costa and R. McCrae, Neo PI-R professional manual *Odessa, FL: Psychological Assessment Resources*, vol. 396, pp. 653–65, 1992.

[15] J. Block, The five-factor framing of personality and beyond: Some ruminations *Psychological Inquiry*, vol. 21, no. 1, pp. 2–25, 2010.

[16] 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.

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

[18] D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pp. 189–196, Association for Computational Linguistics, 1995.

[19] Y. Tausczik and J. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[20] P. Stone, R. Bales, J. Namenwirth, and D. Ogilvie, The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information *Behavioral Science*, vol. 7, no. 4, pp. 484–498, 1962.

[21] P. Costa Jr and R. 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*, pp. 51–87, 1996.

[22] 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*, pp. 627–634, Association for Computational Linguistics, 2006.

[23] T. Buchanan, J. Johnson, and L. Goldberg, Implementing a five-factor personality inventory for use on the internet *European Journal of Psychological Assessment*, vol. 21, no. 2, pp. 115–127, 2005.

[24] P. Rayson, Wmatrix: a web-based corpus processing environment. 2009.

[25] "Waikato environment for knowledge analysis." http://www.cs.waikato.ac.nz/ml/weka/. Online; accessed October 2013.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, The WEKA data mining software: an update *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[27] 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.

[28] A. 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.

[29] F. Mairesse, M. Walker, M. Mehl, and R. Moore, Using linguistic cues for the automatic recognition of personality in conversation and text *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 457–500, 2007.

[30] M. Wilson, MRC psycholinguistic database: Machine-usable dictionary, version 2.00 *Behavior research methods, instruments, & computers*, vol. 20, no. 1, pp. 6–10, 1988.

[31] B. Pang, L. Lee, *et al.*, Opinion mining and sentiment analysis *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[32] S. Picazo-Vela, S. Chou, A. Melcher, and J. Pearson, Why provide an online review? An extended theory of planned behavior and the role of Big-Five personality traits *Computers in Human Behavior*, vol. 26, no. 4, pp. 685–696, 2010.

[33] J. Oberlander and A. Gill, Language with character: A stratified corpus comparison of individual differences in e-mail communication *Discourse Processes*, vol. 42, no. 3, pp. 239–270, 2006.

[34] K. Luyckx and W. Daelemans, Using syntactic features to predict author personality from text in *Proceedings of Digital Humanities 2008 (DH 2008)*, pp. 146–149, 2008.

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

[36] D. Estival, T. Gaustad, S. Pham, W. Radford, and B. Hutchinson, TAT: an author profiling tool with application to Arabic emails in *Proceedings of the Australasian Language Technology Workshop*, pp. 21–30, 2007.

[37] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, Computer-based authorship attribution without lexical measures *Computers and the Humanities*, vol. 35, no. 2, pp. 193–214, 2001.

[38] I. B. Myers, M. H. McCaulley, N. L. Quenk, and A. L. Hammer, *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*, vol. 3. Consulting Psychologists Press Palo Alto, CA, 1998.

[39] G. J. Boyle, Myers-Briggs Type Indicator (MBTI): Some Psychometric Limitations *Australian Psychologist*, vol. 30, no. 1, pp. 71–74, 1995.

[40] R. R. McCrae and P. T. Costa Jr, Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality *Journal of personality*, vol. 57,

no. 1, pp. 17–40, 1989.

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

[42] L. Buffardi and W. Campbell, Narcissism and social networking web sites *Personality and social psychology bulletin*, vol. 34, no. 10, pp. 1303–1314, 2008.

[43] F. Celli and L. Rossi, The role of emotional stability in Twitter conversations in *Proceedings of the workshop on semantic analysis in social media*, pp. 10–17, Association for Computational Linguistics, 2012.

[44] C. Sumner, A. Byers, R. Boochever, and G. J. Park, Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 2, pp. 386–393, IEEE, 2012.

[45] 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*, pp. 29–36, IEEE, 2011.

[46] 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.

[47] 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*, pp. 253–262, ACM, 2011.

[48] 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)*, pp. 149–156, IEEE, 2011.

[49] O. P. John, L. P. Naumann, and C. J. Soto, Paradigm shift to the integrative big five trait taxonomy *Handbook of personality: Theory and research*, vol. 3, no. 2, pp. 114–158, 2008.

[50] T. Yarkoni, Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers *Journal of research in personality*, vol. 44, no. 3, pp. 363–373, 2010.

[51] "International personality item pool." http:/ipip.ori.org. Online; accessed 11-February-2015.

[52] S. Paunonen and M. Ashton, Big Five factors and facets and the prediction of behavior. *Journal of personality and social psychology*, vol. 81, no. 3, p. 524, 2001.

[53] T. Holtgraves, Text messaging, personality, and the social context *Journal of Research in Personality*, vol. 45, no. 1, pp. 92–99, 2011.

[54] N. Yee, H. Harris, M. Jabon, and J. Bailenson, The expression of personality in virtual worlds *Social Psychological and Personality Science*, vol. 2, no. 1, pp. 5–12, 2011.

[55] "Sample of second life daily usage." http://taterunino.net/statcharts/login-concurrency48.jpg. Online; accessed December 2015.

[56] 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.

[57] I.-S. Oh, G. Wang, and M. K. Mount, Validity of observer ratings of the five-factor model of personality traits: a meta-analysis. *Journal of Applied Psychology*, vol. 96, no. 4, p. 762, 2011.

[58] M. Mehl, S. Gosling, and J. Pennebaker, Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, vol. 90, no. 5, p. 862, 2006.

[59] "International personality item pool, 50 item questionnaire." https://ipip.ori.org/New_IPIP-50-item-scale.htm. Online; accessed 2-October-2018.

[60] W. R. Wright, A Corpus for Modeling Personalities of Web Forum Users in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pp. 209–212, ACM, 2017.

[61] R. R. McCrae, P. T. Costa, Jr, and T. A. Martin, The NEO–PI–3: A more readable revised NEO personality inventory *Journal of personality assessment*, vol. 84, no. 3, pp. 261–270, 2005.

[62] M. R. Mehl, J. W. Pennebaker, D. M. Crow, J. Dabbs, and J. H. Price, The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations *Behavior Research Methods, Instruments, & Computers*, vol. 33, no. 4, pp. 517–523, 2001.

[63] B. Rammstedt and O. P. John, Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[64] D. J. Stillwell and M. Kosinski, myPersonality project: Example of successful utilization of online social networks for large-scale social research *American Psychologist*, vol. 59, no. 2, pp. 93–104, 2004.

[65] "Stanford log-linear part-of-speech tagger." Website, 2016. http://nlp.stanford.edu/software/tagger.shtml.

[66] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 173–180, Association for Computational Linguistics, 2003.

[67] "Pos tagger project page, stanford natural language processing group." https://nlp.stanford.edu/software/tagger.shtml. Online; accessed September 2018.

[68] "Tutorial, pos tagger from stanford natural language processing group." http://new.galalaly.me/index.php/2011/05/tagging-text-with-stanford-pos-tagger-in-java-applications/. Online; accessed September 2018.

[69] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine Learning in Python *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[70] W. R. Wright, "Public dissertation defense." Presentation, 2019.

[71] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, Automatic personality assessment through social media language. *Journal of personality and social psychology*, vol. 108, no. 6, p. 934, 2015.

[72] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. 3 draft, ch. 15: dependency parsing ed., Retrieved on February 13, 2020 from https://web.stanford.edu/ jurafsky/slp3/.

[73] D. Chen and C. D. Manning, A fast and accurate dependency parser using neural networks in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 740–750, 2014.

[74] M. Al-Mosaiwi and T. Johnstone, In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation *Clinical Psychological Science*, vol. 6, no. 4, pp. 529–542, 2018.

[75] M. Sellbom, Elaborating on the construct validity of the Levenson Self-Report Psychopathy Scale in incarcerated and non-incarcerated samples *Law and human behavior*, vol. 35, no. 6, pp. 440–451, 2011.