

Boston University

OpenBU

<http://open.bu.edu>

Boston University Theses & Dissertations

Boston University Theses & Dissertations

2021

Language modeling for personality prediction

<https://hdl.handle.net/2144/41942>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
COLLEGE OF ENGINEERING

Dissertation

LANGUAGE MODELING FOR PERSONALITY PREDICTION

by

ANDREW CUTLER

B.S., Brigham Young University, 2015

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2021

© 2021 by
ANDREW CUTLER
All rights reserved

Approved by

First Reader

Brian Kulis, PhD
Associate Professor of Electrical and Computer Engineering
Associate Professor of Systems Engineering
Associate Professor of Computer Science

Second Reader

Kate Saenko, PhD
Associate Professor of Computer Science

Third Reader

Gianluca Stringhini, PhD
Assistant Professor of Electrical and Computer Engineering

Fourth Reader

Francesco Orabona, PhD
Assistant Professor of Electrical and Computer Engineering
Assistant Professor of Systems Engineering
Associate Professor of Computer Science

*I almost wish I hadn't gone down that rabbit-hole
—and yet—and yet—it's rather curious, you know,
this sort of life!*

Alice in Wonderland

Acknowledgments

We stand on the shoulders of giants, but lean on our friends. I am grateful to all those that helped me along the way. First Jazmin for supporting me through hurricanes, moves and a pandemic. You are a tremendous motivating and steadying force. Brian for his advice about kernels, career, and the food scene in the greater Boston area. Grateful for the space and encouragement he gave to develop my own interests.

To my family. David and Pamela who decades ago answered an endless stream of “why?”. Michelle for always lending an ear. Isaac for a bond only terminally nerdy brothers understand. Nicole for letting me steal her style. Carmen for shared academic interest (good luck on your doctorate!).

Nan, where a drunken argument over the sovereignty of Taiwan led to a lasting friendship; proof-reading papers, traveling across the country, and living together. To the wonderful socials put on my Systems Engineering and the friends I made there: Artin, Arian and Andres. To my brilliant labmates Kubra, Tayler and Xide for filling my head with ideas during group meetings. To AIR for an enlightening series of lectures, lunches, and ski trips. To the risk BU took on me, and the funding they and the NSF provided. To squash and frisbee friends who kept me balanced and took me under their wing: Yaser, Amie, Simon, Nishit, Gaurav and Paul. I learned much about the natural and cultural world from all those around me. Many thanks to all those who helped me write this chapter of my life.

Andrew Cutler

PhD Candidate

Department of Electrical and Computer Engineering

LANGUAGE MODELING FOR PERSONALITY PREDICTION

ANDREW CUTLER

Boston University, College of Engineering, 2021

Major Professor: Brian Kulis, PhD

Associate Professor of Electrical and
Computer Engineering

Associate Professor of Systems Engineering

Associate Professor of Computer Science

ABSTRACT

This dissertation can be divided into two large questions. The first is a supervised learning problem: given text from an individual, how much can be said about their personality? The second is more fundamental: what personality structure is embedded in modern language models?

To address the first question, three language models are used to predict many traits from Facebook Statuses. Traits include: gender, religion, politics, Big5 personality, sensational interests, impulsiveness, IQ, fair-mindedness, and self-disclosure. Linguistic Inquiry Word Count (Pennebaker et al., 2015), the dominant model used in psychology, explains close to zero variance on many labels. Bag of Words performs well and the model weights provide valuable insight about why predictions are made. Neural Nets perform the best by a wide margin on personality traits especially when few training samples are available. A pretrained personality model is made available online that can explain 10% of the variance of a trait with as little as 400 samples, within the range of normal psychology studies. This is a good replacement for Linguistic Inquiry Word Count in predictive settings.

In psychology, personality structure is defined by dimensionality reduction of word

vectors (Goldberg, 1993). To address the second question, factor analysis is performed on embeddings of personality words produced by the language model RoBERTa (Liu et al., 2019). This recovers two factors that look like Digman's α and β (Digman, 1997) and not the more popular Big Five. The structure is shown to be robust to choice of context around an embedded word, language model, factorization method, word set and English vs Spanish. This is a flexible tool for exploring personality structure that can easily be applied to other languages.

Contents

1	Introduction	1
1.1	Contributions	2
2	Background and Framework	5
2.1	Language models	5
2.1.1	Linguistic Inquiry Word Count	5
2.1.2	Bag of Words	7
2.1.3	Deep Learning	10
2.2	Psychometrics	13
2.2.1	Latent Factor Modeling	13
2.2.2	Factors as Psychological Constructs	18
2.2.3	Proliferation of Scales	21
2.3	MyPersonality Dataset	23
3	Inferring Personality from Text	27
3.1	Introduction	27
3.2	Traits	30
3.3	Bag of Words	32
3.3.1	Experimental Setup	32
3.3.2	Word Lists	33
3.3.3	Performance	37

3.3.4	Cambridge Analytica	43
3.3.5	Gender Bias in Atheist vs Agnostic Classifier	45
3.3.6	BoW Conclusion	47
3.4	Deep Learning	47
3.4.1	Experimental Setup	47
3.4.2	Multi-Task Learning	48
3.4.3	Results	49
3.4.4	Gender	52
3.4.5	Restricting Training Samples	54
3.4.6	Deep Learning Conclusion	54
4	ML vs LIWC: a case study in predicting grandiose narcissism	56
4.1	Method	59
4.1.1	Participants	59
4.1.2	Materials	59
4.1.3	Procedure	60
4.1.4	Quantitative Approach	60
4.1.5	LIWC Text Processing	61
4.1.6	ML Text Processing	61
4.1.7	Personality Embedding	62
4.1.8	Statistical Modeling	64
4.2	Results	64
4.2.1	Preregistered Correlation Prediction	67
4.3	Discussion	68
5	The Lexical Hypothesis	72
5.1	Deep Lexical Hypothesis	74
5.1.1	Embedding Context	75

5.1.2	Other Personality Words	83
5.1.3	Embedding IPIP Questions	89
5.1.4	Factorization Choices	93
5.1.5	Multilingual Embedding	94
5.2	Discussion and Conclusion	97
6	Conclusion	100
6.1	Summary	100
6.2	Future Work	102
	Appendix	104
	Bibliography	118
	Curriculum Vitae	135

List of Tables

2.1	Term Counts of Documents	8
2.2	Documents as Normalized <i>tf-idf</i> 's	9
2.3	Adjectives Used in Thurstone's Factor Study	17
2.4	Studies Predicting Big Five from Social Media	23
3.1	Top 15 Words	35
3.2	Top 15 Words	35
3.3	Top 15 Words	36
3.4	Top 15 Words	36
3.5	Prediction Accuracy on Continuous Data	37
3.6	Prediction Accuracy on Categorical Data	37
3.7	Gender Prediction	38
3.8	Pairwise Religion Words	38
3.9	Religion Confusion Matrix	38
3.10	Agnostic vs Atheist Confusion Matrix	44
3.11	Fair Agnostic vs Atheist Confusion Matrix	44
3.12	Multitask Learning	49
3.13	RoBERTa: Explained Variance	49
3.14	RoBERTa Gender Prediction	52
3.15	RoBERTa Politics Confusion Matrix	53
3.16	RoBERTa Politics AUC	53

3.17	RoBERTa Religion AUC	54
3.18	RoBERTa Religion Confusion Matrix	54
4.1	Narcissism Prediction Performance (as measured by R^2)	66
4.2	Correlation Values	68
5.1	Core Merriam-Webster Personality Words	79
5.2	ESL Personality Words	83
6.1	Pairwise Politics Words	105
6.2	Politics Confusion Matrix	105
6.3	Personality Words	106
6.4	Personality Words Continued	107
6.5	Sensational Interest Words	108
6.6	Sensational Interest Words Continued	109
6.7	Psychographic Words	110
6.8	Religion and Politics Words	111
6.9	Race Words	112

List of Figures

3.1	IPIP105 Models Trained on Less Data	55
4.1	Text Embedding Schematic	65
5.1	Factor Analysis of Thurstone Words	76
5.2	Google Ngrams of Outlier Words	77
5.3	Factor Analysis of Thurstone Words	78
5.4	Factor Analysis of Thurstone Words	80
5.5	Factor Analysis of Thurstone Words	81
5.6	Factor Analysis of Merriam-Webster Words	82
5.7	Factor Analysis of ESL Words	84
5.8	Factor Analysis of 1,005 Words	86
5.9	Factor Analysis of 1,005 Words	87
5.10	Factor Analysis of 1,005 Words	88
5.11	Absolute Value of Pairwise Pearson’s Correlations	90
5.12	Eigenvalues of RoBERTa Embedding	91
5.13	Mini-IPIP Questions Mapped to Lexical Factors	92
5.14	PCA vs Factor Analysis	93
5.15	XLMr Embedding in English and Spanish	95
5.16	XLMr Embedding in English and Spanish	96
6.1	Factor Analysis of Thurstone Words	113

6.2	Factor Analysis of Thurstone Words	114
6.3	Factor Analysis of Thurstone Words	115
6.4	Factor Analysis of Thurstone Words	116
6.5	Factor Analysis of 1,005 Words	117

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BIS	Barrat Impulsiveness Scale
BoW	Bag of Words
IPIP	International Personality Item Pool
IQ	Intelligence Quotient
LIWC	Linguistic Inquiry Word Count
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Network
RoBERTa	A Robustly Optimized BERT Pretraining Approach
SIQ	Sensational Interest Questionnaire
SoTA	State of The Art
SWL	Satisfaction With Life

Chapter 1

Introduction

One theory for how humans became intelligent is that the invention of language produced strong incentives to be able to think and communicate abstractions. Those that could rally words could rally friends and live to reproduce another day (Dunbar and Dunbar, 1998). Language functions not only as an impetus towards human intelligence, as a map of our own psychology. The lexical hypothesis predicts that most of the socially relevant personality information is embedded in natural language (Goldberg, 1993). This insight was used to define personality factors such as the Big Five (John and Srivastava, 1999). Dimensionality reduction is performed on the co-occurrence statistics of thousands of personality adjectives to find latent general factors. This process is described in the Chapter 2.

Recently, internet advertising and artificial intelligence companies have built language models to accomplish increasingly human tasks such as translating Wikipedia (Vrandečić, 2020), censoring hate speech (Kielar et al., 2020; Conneau et al., 2019), and automating customer service (Xu et al., 2017). Before suggesting a specific translation or classification, these models vectorize text in a highly informative, low dimensional space. This work uses those models to answer two questions. First, given text from an individual, how much can be said about their personality? This supervised learning problem has received much attention as results inform psychology, marketing, and political science. Second, what personality structure is embedded in language? This has been asked for over a century using

surveys to vectorize words—each dimension representing whether a particular respondent believes an adjective described them (Thurstone, 1934). Here modern language models are used to vectorize personality words. The latent personality structure there is remarkably similar to two meta-traits described by Digman: socialization and self-actualization (Digman, 1997).

Chapter 3 focuses on predicting traits available in the MyPersonality dataset from users' Facebook Statuses (Stillwell and Kosinski, 2012). Three language models of increasing complexity are compared: Linguistic Inquiry Word Count (LIWC), Bag of Words (BoW), and RoBERTa (Pennebaker and King, 1999; Zhang et al., 2010; Liu et al., 2019). Chapter 4 applies these BoW models to a dataset of student essays to predict grandiose narcissism. Chapter 5 explores the factor structure in personality adjectives vectorized by RoBERTa. Chapters 3 and 4 can be viewed as asking how much personality information is explained by language in Facebook Statuses or student essays. Chapter 5 demonstrates a deep method to explore personality structure embedded in language.

1.1 Contributions

The extent to which language models extract psychological information has not yet been fully realized. This work aims to bring advances in language modeling to psychometrics. Contributions include:

1. BoW models obtain state of the art performance predicting many traits from Facebook Status updates. These traits include: Big Five personality, gender, political identification, religion, race, satisfaction with life, IQ, sensational interests, self-disclosure, fair-mindedness, and belief in astrology. There is a theoretical limit to how much information about each trait is in social media text. Previously, this was often severely underestimated. In the case of life satisfaction, previous research showed no relation to Facebook Statuses. This observation led to incorrect theories

about how much people hide depression on social media. Previous politics classifiers treated the problem as binary: liberal or conservative. This work extends that to a dozen classes and maintains accuracy. The mostly highly weighted words in each model are also included as a way to generate hypothesis about language, personality and identity.

2. BoW outputs from the above personality models are used to predict grandiose narcissism on a separate sample of 471 student essays. Text from the essay is first embedded as a vector of 61 predictions. This embedding is compared to the standard language model in social psychology, LIWC, which counts how many words appear in 84 different categories. Our method obtains modest results ($EV = 0.03$), and LIWC obtains negative EV. Subsets of those two feature sets are also compared. From the BoW embedding, four traits theory connects to narcissism: Extroversion, Agreeableness, Openness, and gender. And from LIWC the four most highly correlated word groups: Anxiety/Fear, Tentative, Sensory/Perceptual Processes, and Home. The former obtains the best results ($EV = 0.04$) and the latter fails to extract narcissism information ($EV = 0.00$). The four-variable models outperform, demonstrating the usefulness of designing parsimonious models using domain knowledge. All hypothesis are pre-registered on the Open Science Foundation website [<https://osf.io/8uard>]. Code can also be found there.
3. The narcissism model takes advantage of relationships between text and personality learned from Facebook data. Deep language models can transfer more complicated relationships at much greater scale. To this end a general personality embedding is produced using the pretrained RoBERTa model (Liu et al., 2019). Each of the 100 dimensions relates to an item from the Big Five questionnaire IPIP100 (Goldberg et al., 2006). This embedding outperforms the BoW model by a large margin on labels where samples are scarce. Impulsivity is raised from 0.03 to 0.25 EV and

satisfaction with life goes from 0.03 to 0.19. This is important considering there are scalable interventions for those with depression (Fitzpatrick et al., 2017). LIWC is designed to extract psychometric information, but largely fails to do so. This embedding is a good replacement for prediction tasks in social psychology. It is made available on github.

4. Flexible embeddings of personality descriptions offer a powerful new way to explore the structure of personality. Different sets of personality adjectives are embedded using RoBERTa and the multilingual XLM-R (Conneau et al., 2019). The first factor loads on socialization: *affable, easygoing, appreciative, tolerant, genuine, gracious*, and *polite* vs. *contemptible, vindictive, deranged, narcissistic, callous*, and *prejudiced*. The second factor loads on self actualization: *outrageous, animated, boisterous, zany, salacious, captivating, insolent* and *exuberant* vs. *methodical, inhibited, conformist, aloof, formal, circumspect, restrained*. This matches the higher order personality structure found in many Big Five studies (Digman, 1997). The first two factors are shown to be robust to adjective set, language model, decomposition method, and English vs. Spanish.

Chapter 2

Background and Framework

2.1 Language models

In order to predict personality from text both the input and output must be quantized. There are many methods to transform language to vectors. This work uses Linguistic Inquiry Word Count, Bag of Words, and neural networks. Personality constructs are historically based on a restricted language model. That connection and their development is described in the Section 2.2.

2.1.1 Linguistic Inquiry Word Count

Linguistic Inquiry Word Count (LIWC) is a program that takes text as input and counts the number of words which are in each of 85 of different categories. It was developed in the early 1990s by social psychologists and has since received updates in 1997, 2007, and 2015 (Pennebaker et al., 2015). The 2015 version supports a dictionary of 6,400 words, word stems, punctuation, and emoticons. 85 “subdictionaries” define the word categories including: linguistic dimensions (eg, 3rd person singular pronouns, auxiliary verbs, numbers, etc.), psychological processes (eg. anxiety, family, certainty, ingestion, risk, future focus, motion, etc.), Personal concerns (eg. work, leisure, religion), and informal language (eg, swear words, netspeak, fillers). Words can be counted in multiple categories. For ex-

ample, “cried” is part of *sadness, negative emotion, overall affect, verbs, and past focus*. A full list of word categories can be found in (Pennebaker et al., 2015). Full subdictionaries are not published but the LIWC counting software can be purchased.

LIWC started out as a sentiment analysis project to calculate the number of positive and negative emotion words within a text. As recounted in (Tausczik and Pennebaker, 2010), over the course of several weeks this expanded to 80 categories for more general use. The laborious development of subdictionaries follows these steps (Pennebaker et al., 2015).

1. **Word Collection.** Given a description of a word category 2-6 judges individually generated a list of words, then a group brain-storming session was held among 4-8 judges.
2. **Judge Rating Phase** Each prospective word in a subdictionary was then examined by 4-8 judges for goodness of fit. Words not obtaining a majority vote were excluded.
3. **Base Rate Analysis.** Words not appearing at least once in more than one of Pennebaker’s linguistic corpora are excluded. The corpora include: blog posts, spoken language studies, Twitter, Facebook, novels and student writings.
4. **Candidate Word List Generation.** In order to expand the dictionary with common words that may have been missed, words with high frequencies in the English language were correlated to each of the word lists. Those with a high correlation were examined by 4-8 judges who used majority vote to include them in the subdictionaries.
5. **Psychometric Evaluation.** Each word is now represented as its percentage of each text in 181,000 texts from the aforementioned corpora. Correlation matrices are made for each subdictionary. Words with low correlations to the rest of the group are once again voted on by 4-8 judges.

6. **Refinement Phase.** Steps 1-5 are repeated in their entirety to catch any oversights. Two judges then review the resulting dictionary for mistakes.
7. **Addition of Summary Variables.** These include: total word count, words per sentence, number of words with more than six characters, percent of words in text spanned by the dictionary. Additionally four derivative categories developed by Pennebaker’s lab are included: analytical thinking, clout, authenticity, emotional tone. These are combinations of other categories.

The abstract of the 2007 update starts with the claim: “We are in the midst of a technological revolution whereby, for the first time, researchers can link daily word use to a broad array of real-world behaviors...Empirical results using LIWC demonstrate its ability to detect meaning in a wide variety of experimental settings, including to show attentional focus, emotionality, social relationships, thinking styles, and individual differences.” (Tausczik and Pennebaker, 2010). And indeed this is the dominant language model in social psychology; LIWC2015 has already been cited over 2000 times.

2.1.2 Bag of Words

The Bag of Words (BoW) method takes a text string as input and represents it as a vector of word counts, discarding all syntactic information. Like LIWC, BoW faces the challenge of putting common and uncommon words on the same scale. Uncommon words are more informative, but are drowned out by other counts. One solution is using the term-frequency inverse-document-frequency (*tf-idf*) statistic (Spark Jones, 1972). The calculation of *tf-idf* is described below with a subset of terms from a small body of documents. Consider the following body of documents and three of the key terms.

The number of documents is $N = 3$, and n_t is the number of documents term t appears in. Then n_t will be 2, 1, and 3 respectively for the terms “statistics”, “helpful”, and “class”.

Table 2.1: Term Counts of Documents

Documents	Terms and term counts		
	“statistics”	“helpful”	“class”
“Psychology is a popular class.”	0	0	1
“Statistics suggest statistics is an unpopular class.”	2	0	1
“Statistics is helpful in a psychology class.”	1	1	1

The next step is calculating inverse document frequencies for each term t . The most basic implementation would be the simple formula $\frac{N}{n_t}$, but Python’s scikit-learn library modifies the formula three ways for mathematical convenience (Pedregosa et al., 2011). First, one is added to the numerator and denominator, then a natural logarithm is taken, and finally one is added. These modifications have a smoothing effect by lessening the difference in idf ’s for small values of N and n_t , preventing division by zero if attempting to calculate $tf-idf$ for a term not occurring in the documents, and preventing idf ’s from being unreasonably large if N is large and n_t is small. Therefore the implemented version is

$$idf_t = \ln\left(\frac{N+1}{n_t+1}\right) + 1.$$

Applied to the terms this yields 1.29, 1.69 and 1 for “statistics”, “helpful” and “class” respectively. Notice, that “class”, which shows up in every document, gets the lowest idf , “statistics” is in the middle, and the least common “helpful” gets the highest idf .

To calculate an un-normalized $tf-idf$ for a term t and document d , simply multiply the term frequencies $tf_{t,d}$ in Table A by the corresponding inverse document frequency for term t .

At this stage these values are suitable for comparing term importance within a document, but not across documents. For example, a long document with repeated terms may have a large $tf-idf$, as happens here with “statistics” in the second document. To make values comparable across documents, normalize so that the sum of squared $tf-idf$ values within a document would be one. That is,

$$tf\text{-}idf_{t,d} = \frac{tf_{t,d} \cdot idf_t}{\sqrt{\sum_t (tf_{t,d} \cdot idf_t)^2}}$$

Notice this normalization would be across all terms in a document, not only the three used here for illustration. The final normalized *tf-idf*'s are in Table 2.2

Table 2.2: Documents as Normalized *tf-idf*'s

Documents	Terms and Normalized <i>tf-idf</i> 's		
	“statistics”	“helpful”	“class”
“Psychology is a popular class.”	0	0	0.39
“Statistics suggest statistics is an unpopular class.”	0.62	0	0.24
“Statistics is helpful in a psychology class.”	0.39	0.51	0.3

One way to preserve syntactic information is to expand the dictionary to include *n*-grams. *n*-grams are *n* subsequent words that have a significant meaning beyond their constituent parts. Using tri-grams “New York Times” would include a count for “new”, “york”, “times”, “new_york”, “york_times” and “new_york_times”. The dictionary size increases exponentially with *n* and quickly becomes prohibitively large, particularly if *idf*'s are being calculated. This can be solved by only including common *n*-grams. Another solution is to use the hashing trick to hash each dictionary element to a random position in the document vector (Karp et al., 2003). Some words will be hashed to the same position and therefore hashing cannot be used with the *idf* formulation. Nevertheless, in practice this is a powerful way to vectorize documents.

Latent Semantic Analysis

One extension of BoW is to factorize the term by document matrix using singular-value decomposition. Called Latent Semantic Analysis (LSA), a document can then be represented as hundreds of factors rather than thousands of word counts (Furnas et al., 1988). More will be said about factor analysis and dimensionality reduction in the Section 2.2. However, briefly, the assumption is that what we observe (word counts) are noisy instantiations of a latent structure. For example, the word “large” may be replaced with “big” without

much loss of meaning. Finding the factor structure of word choice allows shared meaning; an author choosing “large” or “big” will both load on to the same factor rather than be counted as separate variables. Like LIWC word categories, factors pool related words into higher order concepts, albeit completely empirically. In a supervised learning setting this is a useful when labels are sparse; there is much unlabeled text to be had on the internet on which to learn the factor structure. Words are represented as vectors by their loading on each of the factors. Documents are represented by the mean of all their word vectors.

2.1.3 Deep Learning

In recent years there has been extraordinary progress in the field of natural language processing, largely driven by information technology companies. Google, whose service depends on understanding text queries, trained the 11 billion parameter language model T5 (Raffel et al., 2020). In the 2018 Facebook’s platform was being used to fan the flames of the Rohingya Muslim genocide in Myanmar. At the time, their auto-translate feature read “Kill all the kalars that you see in Myanmar; none of them should be left alive.” in Burmese as “I shouldn’t have a rainbow in Myanmar.” in English (Stecklow, 2018). As a result, their hate speech filters were useless in the region. The next year Facebook Research introduced language-agnostic sentence representations in 93 languages (Artetxe and Schwenk, 2019). In OpenAI’s stated pursuit of artificial general intelligence they trained the 175 billion parameter GPT-3 which can produce remarkable natural sounding text (Brown et al., 2020).

Following the pattern in computer vision, neural nets (NNs) have rapidly replaced BoW models. Initially, neural nets were used to learn better word vectors. Rather than factorizing a matrix of word usage statistics, the problem was formulated as one of language modeling: given the context of a masked word predict the value of the masked word. For example, “Down one point, The Lakers need to make the next basket to win the [MASK]” should rate “game” and “championship” as likely values for “[MASK]”. The initial layer

of the network projects a one-hot representation of each word into vector space. This layer can later be used to define word vectors like the factor loadings of LSA. Trained over enormous corpora, the network learns an informative projection for each word in the dictionary (Pennington et al., 2014).

Word vectors can be analyzed by using them to complete analogies. For example, “husband is to wife as man is to woman” can be encoded by the vectors of their respective words: $husband - wife \approx man - woman$. To complete “Buenos Aires is to Argentina as Paris is to ___” simply find the vector most similar to vectors of $Argentina - Buenos_Aires + Paris$. On a set of 19,544 such analogies, the Global Vectors (GloVe) model achieved 75% accuracy (Pennington et al., 2014), better than the LSA baseline trained on the same amount of data.

These pretrained embeddings have been used on a host of downstream tasks such as sentiment analysis of product reviews or query retrieval from large databases of text. However, word vectors struggle to represent negation (“not” can completely change the meaning of a sentence) and words with multiple meanings. Indeed, it’s surprising BoW and word vectors perform so well considering they discard all syntax which is vital to human understanding of language.

Transformers

A mechanism called attention now dominates language modeling. Text is represented sequentially as a series of tokens. The first layer embeds each token as a vector, and subsequent transformer layers allow other tokens to *attend* to the updates of any other token. This global attention allows the network to find tokens that are relevant to one another, and perform updates preferentially between those tokens wherever they occur in the sequence. This is in contrast to vanilla NNs where a node combines features from the previous layer regardless of the values being passed to it. An excellent explanation of transformers can be found in (Katharopoulos et al., 2020).

RoBERTa

Google AI's Bidirectional Encoder Representations from Transformers (BERT) set the state of the art on many language tasks, often by a large margin (Devlin et al., 2018). This dissertation uses an optimized version of that model trained by Facebook AI (Liu et al., 2019). Robustly Optimized BERT Pretraining Approach (RoBERTa) is trained after a parameter search over many training decisions including:

- **Static vs. Dynamic Masking.** BERT uses a language modeling loss function. This means that some tokens from the input will be masked and the model will be required to predict those missing values. RoBERTa found that it was not very important to dynamically mask examples so that they are new each epoch.
- **Model Input Format.** BERT can be trained on sequences of up to 512 tokens. Because there is great variation in sentence length one can concatenate sentences to fill the sequence length, or zero pad.
- **Auxiliary Loss** In addition to masking, BERT is trained to predict whether a pair of sentences follow one another. Theoretically, this helps learn long range dependencies, though empirical results have been mixed.
- **Data.** BERT was trained on 13GB of text from books and wikipedia. This is increased to 160GB from books, news and sites linked on reddit.
- **Training Time** With an order of magnitude more training data, RoBERTa benefits from as much as 16x training.
- **Batch Size** BERT used batch sizes of 256. RoBERTa gets better performance with much larger batch sizes of 2k and 8k.

Training these models is an engineering challenge. BERT required enormous amounts of data, compute and expertise to build. RoBERTa introduces no new modeling technique

but achieves much better results by optimizing somewhat mundane engineering decisions. Currently only online advertising companies have the will and ability to produce and share competitive language models.

2.2 Psychometrics

The lexical hypothesis posits that most of the socially relevant personality characteristics are encoded in natural language. For over a century psychometrics has wrestled with how to condense language information into a low dimensional vector that can describe the normal range of variation and is robust to demographic variables such as age, gender, or region.

2.2.1 Latent Factor Modeling

In 1904 Spearman wrote, “Whenever branches of intellectual activity are at all dissimilar, then their correlations with one another appear wholly due to their being all variants wholly saturated with some common fundamental Function (or group of Functions)” (Spearman, 1904). That is, a few latent traits predict performance in many seemingly disparate areas. Spearman found great success in a theoretical single underlying factor of general intelligence (g), more colloquially known as the Intelligence Quotient (IQ). This construct is built out of correlations between myriad tasks such as vocabulary, object assembly, pictorial pattern completion. Defining and analyzing latent traits requires statistical modeling. This section describes three related models in ascending order of complexity.

Principle Component Analysis

Principal Component Analysis (PCA) is a method to reduce the dimensionality of data that maximizes the variance of the projected data (Wold et al., 1987). By virtue of maximizing the variance an informative factor model of the data is learned. Given a $D \times N$ data matrix X , where D is the dimensionality of the data (eg. number of questions on a test) and N is

the number of samples, can the data be projected into a lower dimensional space $M < D$ without losing too much information? This can be written as a reconstruction loss

$$\begin{aligned} \min_{\mathbf{W}} \quad & \|X - \mathbf{W}^T \mathbf{W} X\|_2^2 \\ \text{s.t.} \quad & \mathbf{W} \mathbf{W}^T = \mathbf{I} \end{aligned}$$

where \mathbf{W} is a $M \times D$ matrix that projects the data down to M dimensions. This can be solved by finding the M leading eigenvectors of XX^T and using them to form the columns of \mathbf{W} . This requires $O(N^3)$ computations. It can be solved in $O(D^3)$ which is preferable in psychometrics given D is often much smaller than N .

Probabilistic PCA

PCA is a method to project observed data down to a low dimensional space. Probabilistic PCA formulates the problem in the other direction: given observations find the low dimensional latent space from which they were projected (Tipping and Bishop, 1999). Consider

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mu + \varepsilon,$$

where \mathbf{z} is a latent random variable with M dimensions, \mathbf{W} is a $D \times M$ matrix, μ is the mean of the data, and ε is a Gaussian random variable with zero mean and $\sigma^2 \mathbf{I}$ variance. In this view data is generated by sampling from \mathbf{z} , projecting up to D dimensions, and adding noise. To the extent the observed data is explained by the latent variables ε is small.

As this is a generative model, we define a prior distribution for \mathbf{z}

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} | \mathbf{0}, \mathbf{I}).$$

The conditional distribution is

$$p(\mathbf{x} | \mathbf{z}) = \mathcal{N}(\mathbf{x} | \mathbf{W}\mathbf{z} + \mu, \sigma^2 \mathbf{I})$$

where \mathbf{W} , μ and σ^2 are described above. The assumption of a zero-mean and unit variance for the prior Gaussian does not limit the generality of $p(\mathbf{x}|\mathbf{z})$ as any arbitrary mean and variance can be represented by W , μ and σ^2 . The variance $\sigma^2\mathbf{I}$ is the portion of the observed distribution that cannot be explained by the latent variable.

There exists a closed form maximum likelihood solution for W , μ and σ^2 given the data X .

$$\mathbf{W}_{ML} = \mathbf{U}_{ML}(\mathbf{L}_{ML} - \sigma^2\mathbf{I})^{1/2}\mathbf{R},$$

where \mathbf{U}_{ML} is constructed from the eigenvectors corresponding to the M largest eigenvalues of the data covariance matrix, and \mathbf{L}_{ML} is a diagonal matrix of the eigenvalues. \mathbf{R} is any arbitrary $M \times M$ orthogonal matrix. Consider fixing this as \mathbf{I} ; any orthogonal matrix is a valid rotation in \mathbf{z} space. μ is given by the mean of the data \mathbf{X} , and σ^2 is the average of the discarded eigenvalues. If $\sigma^2 = 0$ (the model is a perfect fit) W is equivalent to PCA. Deviation from a perfect fit will be represented a decrease in magnitude of \mathbf{W}_{ML} , as well as uniform noise. \mathbf{W}_{ML} can be obtained in $O(M^3)$ time. When this is infeasible, or when there is missing data, an iterative Expectation Maximization algorithm can be used to fit the model (Moon, 1996). For a derivation consult the excellent textbook (Bishop, 2006).

Factor Analysis

Factor analysis is equivalent to probabilistic PCA with the single relaxation that the variance is not isotropic (Lawley, 1953). The conditional probability is now given by

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \Psi),$$

where Ψ is a $D \times D$ diagonal matrix. This more general model has advantages for psychometrics. If every dimension of X is a participant's response to a question, what happens if a question is unrelated to other questions and the latent variables? This will increase the

variance explained by discarded dimensions. Probabilistic PCA assumes isotropic noise; the unexplained variance will be distributed evenly on every dimension of X . \mathbf{W}_{ML} will likewise be uniformly distorted. With Factor Analysis an unrelated question can be largely sequestered from the shared latent space by allowing a large entry in the variance matrix Ψ . The model is still encouraged to use \mathbf{z} to explain \mathbf{x} as long as the data's correlation matrix is not diagonal (eg. uncorrelated questions) like Ψ . There is no closed form solution, however the maximum likelihood solution can be found via Expectation Maximisation.

Historical Computation

These models have computational trade offs to this day. Practitioners must take care to select an algorithm that is in $O(D^3)$ vs $O(N^3)$ time depending on their data, or use an iterative method when that is not feasible. There were considerably more restrictions in the 1930s when correlations were found by hand. What follows is a description of the factorization in Thurstone's seminal work *Vectors of the Mind* (Thurstone, 1934).

Participants ($n = 1300$) were asked to think of someone they know well and mark whether 60 common adjectives describe them, reported in Table 2.3. As the data is binary, relationships between variables can be described succinctly by their co-occurrence frequency (eg. the percentage of the time someone is described as both 'earnest' and 'systematic'). This yields a triangular matrix with 1770 entries. It is assumed that the actual value of 'earnestness' a person has is continuous and the value is normally distributed in the sample. The binary response applies the rater's threshold to their noisy reading of that value. The tetrach correlation was designed for such cases. A triangular matrix \mathbf{T} is defined. Each cell is $t_{ij} = \cos(\frac{c_{ij}}{N})$ where $N = 1300$ and c_{ij} is the number of people described as both adjective i and j . The cosine function is not computed for each value. Rather, lookup tables are used (Chesire et al., 1933) which in turn were calculated using Taylor expansion and contained significant deviation from true values (Brown and Benedetti, 1977). This process took several minutes for each cell.

Table 2.3: Adjectives Used in Thurstone's Factor Study

calm	capable	friendly	cheerful	courteous	domineering
tidy	peevish	stubborn	grasping	determined	pessimistic
frank	refined	tolerant	sarcastic	submissive	hard-working
quiet	bashful	spiteful	congenial	suspicious	disagreeable
stern	haughty	reserved	religious	courageous	self-reliant
crafty	jealous	generous	impetuous	headstrong	broad-minded
fickle	earnest	faithful	unnatural	dependable	accomodating
solemn	tactful	reserved	frivolous	ingeniuos	conscientious
awkward	precise	talented	eccentric	systematic	self-important
patient	cynical	careless	satisfied	persevering	unconventional
					quick-tempered

Note: 61 words are included in Thurstone's original table, while his text says there are 60 words total.

The correlations in \mathbf{T} are assumed to be due to a low dimensional factor structure of the data:

$$\mathbf{T} = \mathbf{F}\mathbf{F}^T + \mathbf{E}$$

where \mathbf{F} is $D \times M$, and \mathbf{E} is a $D \times D$ error matrix. M was increased until $\|\mathbf{E}\|$ was sufficiently small, which happened at $M = 5$. Given this was the first experiment of its kind, there was no way of knowing how many factors would be required. Thurstone notes that 5 is less than he expected which bodes well for the search for universal latent variables. The matrix factorization algorithm is not reported, but considering $D = 60$ it was a labor intensive process.

The theory that latent factors generate the psychometric data goes back more than a century. Since then there have been many improvements on the modeling side to extract those factors. It is confusing that myriad methods to find latent factors as well as a specific model are called factor analysis. Empirically, the results are very similar so modeling distinctions are not always made in this paper when summarizing historical work. For an overview of the modeling and history see the book *Factor Analysis at 100: Historical Developments and Future Directions* (Cudeck and MacCallum, 2007)

2.2.2 Factors as Psychological Constructs

Finding a useful factor is not simply a matter of modeling. The set of items must be curated, and the factors validated internally and externally. There is a long ongoing debate about the best way to do this and how much one can expect from the whole project (Jackson, 1970). First, a psychometrician theorizes some scale exists (eg. impulsiveness, narcissism). Then a set of questions that interrogate that concept are written (eg. "I like to look at myself in the mirror"). Jackson recommends two editors, one male and one female, look over this list for: concordance to theorized factor, spread across all possible facets, ambiguity in wording, and potential for bias between groups (eg. knitting as a past time meaning one thing for men and another for women). Quantitative tests are applied as well: items may be removed if they are too rare, load too much on group membership (eg. college students, males), or have too high or low a correlation with other questions.

Factors should have both *reliability* and *validity*. Reliable factors are stable over time when a test is re-administered. A natural advantage of broad factors is that they are more robust due to summing many items with uncorrelated noise. Validity is external; how does the factor correlate with theoretically related constructs or outcomes? This comes in two flavors: *convergent* and *divergent* (Campbell and Fiske, 1959). Convergent validity requires large correlations with related constructs. Divergent validity requires absolutely small correlations with unrelated concepts. As an example of the latter a scale for voice quality failed when it was correlated 0.63 with ratings of perceived intelligence (Thorndike, 1920). As an indicator of the scale of the debate, the paper describing convergent vs divergent validation has over twenty thousand citations.

The Big Five

As described in modeling Section 2.2.1, a century ago Thurstone used factor analysis to decompose personality into five general dimensions. Of course a different set of people and

adjectives may change the factors. There may be a different number of axes that are both interpretable and significantly reduce reconstruction loss. They may be in a different order or represent different propensities. Many researchers have tried their hand at developing both general and narrow personality constructs. Narrow constructs being discovered by limiting the adjectives to certain themes. Allport summed up the situation in the 1950s as “each assessor has his own pet units and uses a pet battery of diagnostic devices” (Allport, 1958).

By the 1990s, after decades of experiments and debate, psychologists were approaching consensus on five general factors of personality (John and Srivastava, 1999). In descending order of typical eigenvalue size the Big Five factors are: Extroversion, Agreeableness, Conscientiousness, Neuroticism and Intellect. The final factor is sometimes exchanged for Openness to experience. Their name is in reference to their broadness rather than their singular importance. As a general personality space combinations of these five dimensions can describe many other factors. For example, Responsibility is defined as Conscientiousness and a bit of Agreeableness, and Cooperation is defined as Agreeableness and a bit of Conscientiousness (Hofstee et al., 1997).

Empirical work still produces some variation from these factors. For example one Dutch study finds factors similar to the first common four, but the fifth is associated with rebelliousness rather than Openness or Intellect (Hofstee et al., 1997). Both have to do with orthodoxy, one with more of an edge. A more recent Dutch study asked 1,466 to rate themselves on 2,356 adjectives and found eight factors. These recapitulated the Big Five and added Virtue, Competence and Hedonism (De Raad and Barelds, 2008). The language used when reporting these new factors was “discover” implying confidence that this is a general rule (at least among dutch speakers) and not some artifact of the population gathered, adjectives administered, or modeling choices.

From a clinical and research perspective it is cumbersome to have a scale that re-

lies on thousands of adjectives. Researchers have developed pools of questions that are more informative than single adjectives. Phrases such as: “Do my best to avoid arguments”, “Get along well with people I have just met.” or, “Shoot my mouth off.”. There are many pools: Minnesota Multiphasic Personality Inventory, California Psychology Inventory, Neuroticism-Extroversion-Openness Inventory, and International Personality Item Pool (IPIP). The last has the advantage of being open to the public, translated in many languages, supporting scoring for many scales, and containing thousands of questions to choose from when constructing a new scale.

To approximate factors found from natural language, researchers must understand those factors, and select a set of suitable phrases that cover different elements of the trait. When finding factors words can load on multiple factors. However to make scoring easier scales only allow an item to count for one factor. This increases distortion from the original space and the total required amount of questions for a good approximation. After a scale has been constructed factor analysis is again performed to verify it maintains a five factor structure. For example Goldberg validated a 50-item scale from IPIP on a sample of 906 Scotts (Gow et al., 2005)s. Twin studies show that the Big Five are quite genetic (Smeland et al., 2017), adding evidence that this process produced axes that describe stable parts of individuals’ lived experience.

Advances in computation and data collection have also been a boon for the search of general factors. Traditional sample sizes for factor analysis studies range from a few hundred to a couple thousand. This makes it difficult to consistently recover factors with small eigenvalues even if they are robust. The sample size problem has been solved with internet questionnaires which hundreds of thousands of people will volunteer to take. The other limiting factor is the number of questions one can ask a study participant. Too many and participants lose interest or must be payed prohibitive amounts of money to participate. One clever solution has been to serve random sets of questions from a IPIP6000 (Condon,

2018). The resulting matrix is sparse but can be easily decomposed with tools from astronomy. The structure has five, seventeen, or twenty seven factors depending on the threshold used.

2.2.3 Proliferation of Scales

Linguistic analysis gave evidence there are five overarching traits English speakers use to describe one another. However, greater granularity or combinations of sub-traits may be desired. Narcissism, for example, is characterized by aspects of high extraversion, openness and conscientiousness along with low neuroticism. It is useful to be able to interrogate this trait directly. Constructing a test now follows the pattern a) develop a set of roughly 60 questions b) give questionnaire to several hundred participants c) perform factor analysis to find a scoring and trim uncorrelated questions d) the academy debates the test's predictive power and usefulness. The factor method is some variant of reconstruction loss $X \approx uu'X$. The basis u is usually restricted to a rank of 5 or less, depending on the data. For easy scoring a hard threshold is used to assign values of u to be $\{0,1\}$. Three examples follow.

Barrett's Impulsiveness Scale

Barrett's Impulsivity Scale was first administered to 412 college undergraduates, 248 psychiatric inpatients, and 73 male prison inmates. They were asked questioned such as "I squirm at plays or lectures", "I often have extraneous thoughts" and "I am future oriented". The first three principle components were labeled attentional, motor, and nonplanning impulsiveness respectively (Patton et al., 1995). This measure has been cited thousands of times in fields as diverse as criminal justice, addiction genetics and developmental neuroscience (Stetler et al., 2014; Verdejo-García et al., 2008; Steinberg, 2008).

Sensational Interests Questionnaire

The Sensational Interests Questionnaire (SIQ) asked 301 community members their interest level in 60 different categories including: “Vampires”, “Knives”, and “Camping” (Egan et al., 1999). The participants included 100 mentally disordered offenders. Control participants were fishermen, security guards, teachers, and nurses. The five first principle components were named: Militarism, Violent-Occult, Intellectual Recreation, Occult Credulousness, and Wholesome Activities. SIQ was shown to be a poor predictor of criminal behavior and has been cited only several dozen times (Charles and Egan, 2009).

Emotional Intelligence

346 participants (218 female, 111 male) were asked 62 questions based on a model of emotional intelligence developed by Salovey and Mayer (Salovey and Mayer, 1990). These included, “I know when to speak about my personal problems to others”, “I have control over my emotions” and “It is difficult for me to understand why people feel the way they do”. After factor analysis, these questions were trimmed down to a pool of 33 that were represented in the first four principle components. As noted in the paper’s abstract: “Validation studies showed that scores on the 33-item measure (a) correlated with eight of nine theoretically related constructs, including alexithymia, attention to feelings, clarity of feelings, mood repair, optimism and impulse control; (b) predicted first-year college grades; (c) were significantly higher for therapists than for therapy clients or for prisoners; (d) were significantly higher for females than males, consistent with prior findings in studies of emotional skills; (e) were not related to cognitive ability and (f) were associated with the openness to experience trait of the big five personality dimensions” (Schutte et al., 1998). This is instructive as to how a measure can be established as useful. Like BIS, this research has been cited thousands of times, indicating staying power in the academy.

2.3 MyPersonality Dataset

Table 2.4: Studies Predicting Big Five from Social Media

Study	O	C	E	A	N	n	Val.	Data
Independent datasets								
(Baik et al., 2016)	-	-	0.42	-	-	565	k-fold	Demographics, Usage stats, Likes
(Wald et al., 2012)	0.77	0.61	0.68	0.7	0.61	537	None	— —
(Golbeck et al., 2011)	0.65	0.6	0.55	0.48	0.53	167	k-fold	— —
(Celli et al., 2014)	0.07	0.06	0.18	0.26	0.19	89	Holdout	Pictures
(Kleanthous et al., 2016)	0.26	0.03	0.28	0	0	62	None	Usage Stats
(Golbeck, 2016)	0	0	0.24	0	0	69	None	Language
MyPersonality dataset								
(Farnadi et al., 2016)	0.19	0.24	0.27	0.16	0.24	3731	k-fold	Demographics, Usage stats, Lang.
(Markovikj et al., 2013)	0.71	0.71	0.7	0.6	0.59	250	None	— —
(Bachrach et al., 2012)	0.33	0.41	0.57	0.1	0.51	5000	k-fold	Usage Stats
(Cutler and Kulis, 2018)	0.41	0.35	0.38	0.3	0.32	84451	Holdout	Language
(Golbeck, 2016)	0.36	0.25	0.37	0.41	0.38	127	None	Language
(Golbeck, 2016)	0.2	0.2	0.22	0.24	0.18	8569	None	Language
(Kosinski et al., 2013)	0.43	0.29	0.4	0.3	0.3	54373	k-fold	Likes
(Kosinski, 2014b)	0.11	0.16	0.31	0.05	0.23	45565	k-fold	Usage Stats
(Laleh and Shahram, 2017)	0.38	0.29	0.34	0.22	0.27	92225	Holdout	Likes
(Nave et al., 2018)	0.3	0.19	0.21	0.17	0.18	21929	k-fold	Likes
(Park et al., 2015)	0.43	0.37	0.42	0.35	0.35	4824	Holdout	Language
(Schwartz et al., 2013a)	0.42	0.35	0.38	0.31	0.31	18177	Holdout	Language
(Thilakaratne et al., 2016)	0.36	0.4	0.44	0.3	0.39	387	k-fold	Language
(Youyou et al., 2015)	0.51	0.42	0.45	0.38	0.4	1919	k-fold	Likes
(Zhang et al., 2018)	0.4	0.35	0.36	0.29	0.32	55835	Holdout	Language
(Farnadi et al., 2018)	0.26	0.19	0.16	0.11	0.14	5670	k-fold	Likes, Language, Pictures

From 2008 to 2012, over 7 million Facebook users took surveys on the myPersonality app developed by David Stillwell (Kosinski et al., 2015). The main survey consisted of 50 questions from the International Personality Item Pool (IPIP50) (Goldberg, 1992). After completing those questions users were scored on the Big Five and given the chance to complete a longer survey of 100 or 300 IPIP questions. Over 3 million of those users agreed to give researchers access to their extant Facebook profile—profile pictures, status updates, Liked Pages and demographic information: gender, birthdate, relationship status, religion, political identity—and their personality responses. A much smaller subset of users answered additional questionnaires about their sensational interests ($n = 4074$), Friends’ personality ($n = 17,622$), belief in astrology ($n = 7115$), satisfaction with life ($n = 2502$), and other personal information. The research community has added to the dataset by providing race labels for several hundred thousand users using computer vision software

from faceplusplus.com (Megvii, 2015), and representing the text of statuses in terms of the LIWC model (Pennebaker et al., 2001).

Dozens of studies have used the myPersonality dataset. Initially the work focused on correlations between easily interpretable usage statistics such as number of friends or posts, and Extraversion (Quercia et al., 2012). In another paper network analysis was used to show how introverts vs extroverts interact with communities (Friggeri et al., 2012). After clustering friend groups, Extroverts were seen to more often act as bridges between communities that tended to be smaller and more distant from one another.

Beyond interesting correlations, the data is a natural supervised learning problem. Rich features such as profile pictures, Liked Facebook pages (eg. a musician’s official page or viral joke communities such as “I like hugging boys who smell nice”), and Status updates can be used to train models which are judged primarily by how well they predict labels rather than on the weights of the model. Table 2.4 provides an overview of the papers that predict Big Five from Facebook data. Most of the studies, and all those with $n > 1000$, use MyPersonality. There was a learning curve moving to this new paradigm; notice how only six of the twenty two studies use a hold out sample to validate the results. That is, the other papers fit either parameters (eg. feature weights in regression) or hyperparameters (eg. penalty term in normalized regression) on the same data they use to report accuracy. This overstates general performance, often by a wide margin. The most widely cited such paper is the provocatively titled “Computer-based personality judgments are more accurate than those made by humans” (Youyou et al., 2015). Friends of myPersonality participants answered a 10 question mini-IPIP. These assessments were farther from the users’ IPIP50 scores than a model that predicted Big Five based on what Facebook Pages someone Liked. While the results are impressive, the human baseline that was surpassed is not a direct comparison. Friends of participants were asked to describe the participant on a 10 question survey the mini-IPIP (Donnellan et al., 2006). Due to brevity, this prediction is noisy. Fur-

ther, there is an established disparity between self and peer reported personality appraisals (Clifton et al., 2005). The “true” personality label may be a combination of the peer and self description. Training the Likes model to guess how people will describe themselves is a great comparative advantage.

In 2013, Schwartz et al introduced the open vocabulary approach (or bag of words) to personality, gender, and age prediction (Schwartz et al., 2013a). This significantly outperforms closed-vocabulary approaches such as LIWC that rely on domain knowledge to assign each word to one or more of dozens of categories. For an excellent overview of related work, we direct readers to that paper’s introduction.

The dataset has also generated some controversial research. Wang and Kosinski claim showed that a convolutional neural network (CNN) can predict sexual orientation from images far better than humans—81% accuracy for men, and 71% accuracy for women (Wang and Kosinski, 2018). Given that this is a straightforward supervised learning problem this is not surprising, but the paper also claims the CNN was using facial features such as nose length and jaw width. Establishing somatic differences is an active area of research, particularly when they correlate with androgen levels (Skorska et al., 2015; Valentova et al., 2014). Results have been mixed due to small sample size and difficulty of defining feminine or masculine facial structure in a few statistics. For those interested in this question, CNNs offer a new path that can solve both the feature creation and sample size problems. However, it is notoriously difficult to understand why a CNN makes a prediction. Attention maps produced by a CNN are noisy and even if one trusts an indicated region is important it’s still not clear why the region is important. One detractor made a strong case that Wang and Kosinski’s heat maps were concentrating on shadows produced by gay men holding the camera at a higher angle (heterosexual men have a stronger preference to appeal taller), rather than facial structure (Agüera y Arcas et al., 2018). Further, they constructed a classifier based off survey responses to seven questions (“Do you use eyeshadow?”, “Do you

ever use makeup?”, “Do you have long hair?”, “Do you have short hair?”, “Do you ever use colored lipstick?”, “Do you like how you look in glasses?”, and “Do you work outdoors?”) that matched the CNN’s classification accuracy on women. Controversial research is not bad in and of itself. The backlash to this paper was due in part to discomfort with research into somatic correlations with sexual orientation as well as the belief that one shouldn’t write a how-to guide for oppressive regimes that may want to infer sexual orientation. But the causal claims are also stronger than attention maps on a CNN could support. This pattern of hyped claims and public backlash eventually led to myPersonality being shut down, as will be discussed in more detail in the Chapter 3.

Chapter 3

Inferring Personality from Text

3.1 Introduction

This chapter focuses on the following problem: given Facebook Statuses, how well can personality and demographics be predicted? The answer to this question provides a lower bound for how much personality information is in the text. The better the model, the closer this is to the true amount. This bound is important for research in social linguistics, marketing, and privacy. The models themselves are also interesting. BoW and LIWC use regression from word counts to target variables. The weights assigned to each count explain why the model makes predictions. This can be used as a sanity check for models, as well as a way to generate hypothesis about language, social media, and the trait. Better performance implies better weights; the models presented here are SoTA. Deep learning models have millions of parameters and are not amenable to such analysis. However, the transfer learning system developed here has far superior performance on many personality traits. It also shares important psychometric qualities with LIWC making it an attractive replacement for researchers.

The prediction of individual traits is important. Table 2.4 lists over a dozen studies that do so using MyPersonality among which the work by (Schwartz et al., 2013a) stands out. It uses a BoW approach to predict age, gender, and personality. Care is taken to choose hyper-

parameters and learn parameters using a training set, and report performance on a holdout set. Other studies on the table report better performance using less data, but no holdout set. This is, in effect, overfitting a model and reporting it as generalized performance. Such easily gamed validation strategies are common in the field. The work here began as an extension of Schwartz et al's fine work to other traits in myPersonality. Later the performance was eclipsed using methods from deep learning. We achieve marked improvement on several traits. For example, previous research failed to find a relationship between Satisfaction With Life (SWL) and Facebook Statuses (Wang et al., 2014). Better modeling presented here shows the statuses can explain a sizeable 19.4% of the variance in SWL. This has real world implications given the existence of cheap, scalable interventions for depression (Fitzpatrick et al., 2017). There is significant interest in political advertising on social media. Previous research on predicting political ideology from text treated the problem as a binary label: conservative or liberal. This work expands that to 13 self-reported classes: IPA, anarchist, centrist, conservative, democrat, doesn't care, hates politics, independent, liberal, libertarian, republican, and very liberal. This allows more fine grained analysis of political identity in the future.

Of particular interest is the role of psychographic models in Cambridge Analytica's (CA) marketing strategy. From leaked internal communications, in 2014 CA amassed a dataset of Facebook profiles and traits almost identical to those in the myPersonality dataset (Rosenberg et al., 2018). The week after CA's project became public, Facebook's stock plummeted \$75 billion (Cherney, 2018). One factor in that drop was the belief that Facebook had allowed a third party to create a powerful marketing tool that could manipulate elections (Cadwalladr, 2018; Rosenberg et al., 2018). There are dozens of publications on the myPersonality dataset. However, this work also predicts SIQ, fair-mindedness, and self-disclosure, which CA discussed in relation to building user models (Rosenberg et al., 2018).

Highly weighted features are also an important way to analyze models. We argue in section 3.3.4 that a militarism predictor CA may have built is accurate, but extracts obvious features. Additionally, by inspecting the features in an Atheist vs. Agnostic classifier we find many gendered words. We demonstrate the bias empirically, then fix the classifier to be more fair. This approach is instructive for interrogating more critical models built on social media data.

This work is unique in the number of traits it analyzes at once. This allows stronger comparisons between models which may do well on one trait by chance or because of some real advantage on the problem. Further, predicting many traits serves as a natural regularizer. Hyperparameters and preprocessing decisions cannot be made to maximize performance on single traits. It also opens the door to transfer learning where patterns learned predicting one trait can be applied to others either concurrently or sequentially. This is particularly important for traits with few training samples such as SWL ($n = 2,500$). Transfer learning is also related to Cambridge Analytica's claim that Big Five prediction can help serve political advertisements. Using status updates, Big Five and political labels we show that a transfer learning system that includes Big Five does not help predict political identity. This holds when using a more sophisticated language model than Cambridge Analytica would have had access to.

Like computer vision before, deep learning has rapidly come to dominate natural language processing (NLP). Since the completion of the BoW experiments new more powerful language models have become available. This chapter starts with BoW results, then moves on to a set of better results from deep learning with BoW and LIWC as a baseline. Contributions include:

1. State of the art prediction of: gender, Big Five personality, race, political identity, religion, sensational interests, intelligence, satisfaction with life, impulsivity, self-disclosure and fair-mindedness.

2. Deep transfer learning system with a personality embedding as an intermediate feature. This model explains five times more variance than LIWC and BoW when predicting life satisfaction and impulsivity. It can be used to predict arbitrary new traits with few samples and is available online.
3. Highly weighted features of the BoW models for researchers to explore connections between social media language and traits.

3.2 Traits

Gender is the binary label users supplied when setting up their Facebook account. Offering this information was common before 2008, and mandatory from 2008-2014. In 2014, (after the collection of this dataset) Facebook added 56 more gender options but still uses a binary representation to monetize users (Bivens, 2017).

Race labels provided in the dataset are inferred from profile pictures using the Faceplusplus.com algorithm which can identify races termed White, Black, and Asian. A noisy measure of visual phenotype is not the gold standard for the study of race, however, our results indicate it is related to social media use.

Political identity is limited to the twelve most common responses: IPA, anarchist, centrist, conservative, democrat, doesn't care, hates politics, independent, liberal, libertarian, republican, and very liberal. These are heterogenous categories from an open-ended question. No work was done to limit labels to political parties (eg. remove "doesn't care"), disambiguate misspelled or similar responses (eg. combine "anarchy" and "anarchist" or "liberal" and "very liberal"), or limit responses to one country. To produce the word list for Liberals and Conservatives in Table 6.8, we combine "liberal", "very liberal", and "democrat" as well as "conservative", "very conservative", and "republican". The individual classes are self explanatory save for IPA which most likely refers to the Independence Party of America, which was in its nascence during this survey. The party is most popular

among young people disaffected by the two party system, a sentiment reflected by the users who report IPA.

Religion categories are limited to the nine most common responses, and similar labels are combined. “Catholic”, “christian-catholic”, and “romancatholic” are combined. Likewise, Christian refers to “christian”, “christian-baptist” and “christian-evangelical”. The entire list includes: Atheist, Agnostic, Catholic, Christian and None.

Belief in star sign is the user’s response to “Horoscopes provide useful information to help guide my decisions?” Options include: Strongly Agree, Slightly Agree, No Opinion, Slightly Disagree, and Strongly Disagree.

Personality is determined on five axes—Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism—by a survey. Users answer 20-300 questions which are used to score each personality component on a scale of 1-5. There is a large body of research showing that five factor analysis is explanatory for behavior (Digman, 1990), and its measurement is reproducible (McCrae and Costa, 1987). That work is now adapting to larger datasets collected online (Kosinski et al., 2015).

Sensational Interests include Militarism, Violent-Occult, Intellectual Recreation, Occult Credulousness, and Wholesome activities. Users can indicate “Great Dislike”, “Slight Dislike”, “No Opinion”, “Slight Interest”, and “Great Interest” for 28 different items including: “Drugs”, “Paganism”, “Philosophy”, “Survivalism”, and “Vampires and Wolves”. Interest levels are calculated by summing responses from relevant items. The full calculation can be found in (Egan et al., 1999).

IQ in this dataset is determined by 20 questions based on Raven’s Standard Progressive Matrices. The development and validation of these questions are explained in (Kosinski, 2014a) and (Kosinski, 2014b). Because performance on IQ tests has been rising at roughly 0.3 points a year over the past century and IQ is defined as mean 100, the scoring of a test is properly defined over an age cohort (Flynn, 1987). These scores do not take age into

account and the mean is 114.

Satisfaction with life (SWL) is a measure of global well being somewhat robust to short term mood fluctuations (Diener et al., 1985).

Self-disclosure is indicative of psychological adjustment and self-actualization (Snell et al., 1988).

3.3 Bag of Words

The BoW language model and its relation to LIWC and deep learning word vectors are described in Chapter 2. Briefly, the vocabulary is first limited to the k most common words in a given training set. Then a matrix of word counts, N , is constructed, where N_{ij} refers to how often word j is used by subject i . Each row is normalized to sum to one, moved to a log scale, and divided by d_j , the ratio of documents in which word j appears. Each element of the *tf-idf* matrix is defined by

$$W_{ij} = \frac{1 + \log\left(\frac{N_{ij}}{\sum_{i=1}^k N_{ij}}\right)}{d_j}.$$

W is then normalized so each row lies on the unit sphere. W can now be used for linear classification or regression with ℓ_2 regularization on the parameters. This is commonly called Ridge Regression. For binary classification problems, labels are assigned values of $\{-1, 1\}$ and a threshold of 0 determines the predicted label. For categorical data with more than two labels, a classifier is trained on each pair of labels. Predicted label is decided by majority vote of the $\frac{c(c-1)}{2}$ classifiers, where c is the number of classes.

3.3.1 Experimental Setup

The point of these experiments is to make claims about how well one can predict traits from Statuses. Therefore, the validation strategy is important. We report explained variance (EV) on a random holdout of data. EV is $1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}$, where \hat{y} is the predicted label. A perfect

score is one, and guessing the test set mean for each sample is zero. Negative scores are possible; indeed predicting the mean of the training labels for each test sample will result in negative EV if there is any random shift between the two sets.

All BoW experiments employ the same preprocessing. Users must have over 500 words in the sum of all their statuses. 80% of the users are randomly assigned to the training set; the remaining samples constitute the test set. A seed of zero is used for the random assignment for replicability. This also takes away one researcher degree of freedom as experiments are not selected for favorable random splits. The vocabulary is limited to the 40,000 most common words in each training set. Words must be used by at least 10 users but no more than 60% of users in the training set. The regularization parameter is tuned via efficient leave one out cross validation (Vehtari et al., 2015) when $n < 10,000$, and 3-fold cross validation for larger datasets. All BoW models are implemented using the sklearn library (Pedregosa et al., 2011). Table 3.5 reports the number of samples and explained variance (EV) of the predictions on continuous data. Table 3.6 reports the number of classes, ratio of samples in the dominant class, homogeneity, and performance on tasks with categorical data.

3.3.2 Word Lists

Before the word lists are presented, it's important to understand the interpretation of highly weighted features in a high dimension regression problem. Many papers in social science regress from observations to some target then try to say something about the weights (Khandani et al., 2010; Cooke et al., 2004; Peciña et al., 2013; Quilty et al., 2009; Tett et al., 1991; Egan and Campbell, 2009; Park et al., 2015; Cesare et al., 2017; Kleinberg et al., 2016). Our particular problem is extreme as observations are 40,000 *tf-idf* word counts. Colinearity abounds. The problem is also ill-posed with fewer labels than observations.

These problems mirror those faced when clustering data. Clustering does not come with

guarantees it will yield sensible answers in diverse scenarios (Kleinberg, 2003). However, it is broadly useful when exploring large sets of data (Jain et al., 1999; Shamir and Sharan, 2002; Dixon et al., 2003). Similarly, model weights can be viewed as a way of ranking features for exploration. A highly ranked observation is not proof it is important. But several highly ranked observations with functional coherence are suggestive; particularly when coupled with domain knowledge. Regularization also helps find a few (relatively) highly explanatory observations a human can interpret. One may use ℓ_1 regularization to obtain an arbitrary small number of non-zero weights (Meinshausen and Yu, 2009). This encourages weighting common words and produces rankings that are less sensitive to train/test splits or preprocessing decisions. We demonstrate that approach with our IQ model in Section 3.3.3. The 55 most highly weighted features for each label are reported in the Appendix, and the top 15 are included in this section.

There are many well-studied phenomena embedded in the model weights. For example, Sarah Palin is the only politician indicated in the liberal word list in Table 6.8. Likewise, Nancy Pelosi ranks just below Ronald Reagan among conservative words. This accords with literature on the memorability of negative ads (Lau et al., 2007), importance of out-group prejudice for social identity (Huddy, 2003; Branscombe and Wann, 1994), and biases women face in politics (Schneider and Bos, 2014; Dolan, 2010). We hope the many word lists in the appendix will be useful to researchers in the development of new hypotheses.

In Section 3.3.5 we use our understanding of the input features to characterize information the model extracts to predict religion. This dataset also includes demographic labels, which show predicted religion labels are more gendered than the ground truth.

Word lists are included to (a) highlight unstudied relationships about these traits (b) illustrate what kind of information is extracted from social media by machine learning systems.

Table 3.1: Top 15 Words

Openness		Conscientiousness		Extraversion	
-	+	-	+	-	+
bored	art	lost	gym	internet	party
boring	poetry	fucking	ready	quiet	guys
husband	beautiful	xd	weekend	bored	amazing
attitude	universe	phone	excited	listening	audition
shopping	peace	im	success	apparently	baby
dinner	poem	bored	finished	computer	haha
tv	writing	fuck	studying	stupid	dance
game	books	gonna	busy	pc	girls
proud	theatre	sick	vacation	hmm	fabulous
ur	dream	procrastination	arm	anime	blast
dentist	mind	internet	officially	tt	ready
daughter	book	computer	family	dark	im
dont	woman	probably	relax	probably	wine
haha	guitar	cousins	tennis	sims	success
stupid	damn	hates	wonderful	didn	lets

Table 3.2: Top 15 Words

Agreeable		Neurotic		SWL	
-	+	-	+	-	+
fucking	wonderful	loving	sick	fucking	bye
stupid	amazing	girlfriend	nervous	bored	haha
kill	awesome	wife	stressed	sick	woot
shopping	haha	awesome	depression	shit	soccer
shit	smile	parties	depressed	hurt	simple
burn	happiness	party	anymore	tired	camping
bitch	phone	weekend	lonely	farmville	weeks
pissed	urself	haha	stress	boredom	hahaha
punch	family	doing	fucking	damn	pie
hates	blessed	game	tired	fuck	camp
death	status	sunday	trying	sleeping	sin
hell	music	kansas	depressing	personality	wow
suck	woop	guy	sims	omg	train
freak	hands	delicious	anxiety	wtf	glory
piss	heart	beach	worst	job	pool

Table 3.3: Top 15 Words

Militaristic		Violent-Occult		Intellectual Recreation	
-	+	-	+	-	+
sleeping	man	lord	hell	im	life
ugh	xbox	pray	zombie	course	jon
sad	gets	cousins	damn	boring	beautiful
excited	gotta	church	fuck	painful	dancing
lovely	good	michael	bitch	decision	yoga
oh	training	allah	ass	hurts	thankful
hair	headed	jesus	drink	bus	peace
shopping	truck	game	blood	game	kinda
husband	guitar	0	lmao	stupid	truly
sick	guys	summer	xd	bak	la
cares	bro	gosh	woot	hero	ich
mum	gun	praise	halloween	problem	miss
boyfriend	boom	sunday	play	yeah	likes
lady	epic	dad	guys	christ	comfort
concert	work	loving	drunk	gona	lol

Table 3.4: Top 15 Words

Occult Credulousness		Wholesome Activities		Star Sign	
-	+	-	+	-	+
church	zombie	coke	woot	minutes	omg
praise	ass	michigan	camping	didn	im
jesus	bitch	stupid	fish	church	ready
lord	halloween	pathetic	life	praise	friend
bible	animal	ops	yesterday	jesus	mind
christ	sign	husband	beautiful	probably	ass
team	omg	didn	rain	physics	butt
quite	xd	hurts	man	jess	stay
loving	job	kurwa	mexico	white	tom
pray	woot	evil	wish	religion	tomorrow
paper	wish	afternoon	river	iv	october
game	cure	problem	love	officially	promise
blessed	street	taylor	path	imagine	lol
salvation	vampire	idea	moon	christ	searching
ops	guys	jess	haha	germany	bitch

Table 3.5: Prediction Accuracy on Continuous Data

Label	N	EV
Personality		
Openness	84451	0.171
Conscientiousness	84451	0.120
Extroversion	84451	0.141
Agreeableness	84451	0.090
Neuroticism	84451	0.100
Sensational Interests		
Militarism	4074	0.165
Violent-Occult	4074	0.192
Intellectual Recreation	4074	0.033
Occult Credulousness	4074	0.144
Wholesome Activities	4074	0.108
Satisfaction With Life	2502	0.034
Self Disclosure	2006	0.092
Fair-Mindedness	2006	0.064
IQ	1807	0.128

Explained Variance (EV) is $1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}$, where \hat{y} is the predicted label.

3.3.3 Performance

Gender

Gender is a well studied variable that can be accurately predicted from even simple tasks like how one divvies up money in a game (Capraro and Sippel, 2017). Table 3.7 compares our gender predictor to several other methods. The BoW model with a vocabulary of 500,000 yields accuracy of 92.8%, 1.4% more accurate than the tri-gram model reported

Table 3.6: Prediction Accuracy on Categorical Data

Label	N	Classes	Mode	Homogeneity	F1-score	Acc
Gender	109104	2	0.598	0.519	0.92	0.903
Race	22059	3	0.682	0.52	0.74	0.766
Political identity	19769	12	0.213	0.133	0.33	0.337
Religious identity	8388	5	0.488	0.318	0.54	0.541
Belief in Star Sign	7115	5	0.331	0.245	0.32	0.334

Mode is the ratio of the dominant class. Homogeneity is the probability two random samples will be of the same class. The F1-Score is the harmonic mean of precision and recall. For non-binary labels, the precision and recall for each class is weighted by its support.

Table 3.7: Gender Prediction

Model	Accuracy
Human Majority Vote	0.840
LIWC	0.784
Tri-grams	0.914
Tri-grams + LIWC	0.916
BoW (40k Vocab)	0.903
BoW (500k Vocab)	0.928
char-CNN	0.901

Human baseline is the majority vote (n=210) in gender prediction on Twitter data (Nguyen et al., 2014). LIWC and Tri-grams are reported in (Schwartz et al., 2013a). char-CNN is reported in (Cutler and Kulis, 2018).

Table 3.8: Pairwise Religion Words

	Atheist	Agnostic	Catholic	Christian	None
Atheist		boyfriend	thank	church	lol
Agnostic	fucking		prayers	church	lol
Catholic	fucking	fucking		lol	lol
Christian	fucking	fucking	mass		xmas
None	fucking	apartment	god	church	

The most highly weighted word from each pairwise classifier. Word implies top label.

by Schwartz et al (Schwartz et al., 2013a). Even though the same dataset is used, the comparison is not direct. The tri-gram model seeks to remove the age information from words, has a larger vocabulary, preserves some temporal relationships in the tri-grams, and draws a different train/test split. Moreover, the preprocessing is more restrictive and only includes users with at least 1000 words. Notwithstanding these discrepancies, which may boost or hinder performance, the results are very similar. When the LIWC representation is added

Table 3.9: Religion Confusion Matrix

	Predicted Label					Total
	Atheist	Agnostic	Catholic	Christian	None	
Atheist	68	29	17	16	21	151
Agnostic	54	69	27	55	11	216
Catholic	27	37	172	130	9	375
Christian	35	48	126	560	26	795
None	22	11	19	50	39	141
Total	206	194	361	811	106	1678

to the tri-grams, there is a slight improvement to 91.6% accuracy. Preprocessing is even less similar for the char-CNN described in (Cutler and Kulis, 2018). The human baseline of 84.0% consists of volunteer judgments based on 20-40 user tweets as reported in (Nguyen et al., 2014). This is less text than is available to the other models, and from a different social media platform. But with 210 volunteer guesses per user, it provides a relevant human baseline.

Big Five

After gender, personality is the most studied trait in this paper. Likewise, Schwartz et al achieve the best results to date (Schwartz et al., 2013a). They report the square root of EV to two significant digits: 0.42, 0.35, 0.38, 0.31, 0.31. In that format, we are just 0.01 beneath the state of the art for openness and agreeableness, 0.01 better for neuroticism, and equivalent for the remaining traits. As with gender, we achieve this with a simpler model.

Political Identity

Prediction accuracy of 33.7% is a gain of 11.7% over the baseline strategy of always predicting the mode, ‘doesn’t care’. As noted in the experiments section, training samples are weighted inversely to their class representation; therefore, ignoring any class will result in an equal loss. This does not provide the highest classification accuracy. However, we believe when some classes are sparsely populated an MSE optimal classifier that is highly biased toward the mode should not be the standard. For reference, equal sample weights and the same training scheme yield classification accuracy of 36.3% and a weighted f1 score of 31.6%. Five classes—IPA, hates politics, independent, libertarian, and very liberal—have no representation in the test set predictions. The weighted classifier predicts each class at least once.

According to Preotiuc-Pietro et al., all previous research on predicting political ideology from social media text has used binary labels such as liberal vs conservative or Demo-

crat vs Republican. They broaden the classification task to include seven gradations on the liberal to conservative spectrum (Preoțiuc-Pietro et al., 2017). When predicting ideological tilt from tweets, they achieve a 2.6% boost over baseline (19.6%) with BoW followed by logistic regression. Word2Vec feature embeddings (Mikolov et al., 2013) and multi-target learning with some hand-crafted labels yield an 8.0% boost. From classification along grades of a single spectrum, we significantly expand the task to twelve diverse identities with varying levels of representation and ideological overlap while maintaining classification accuracy.

In Table 6.1 we report the matrix of highest weighted words for separating users in each pairwise class comparison. As with race, belief in star sign, and religion, we plan on making expanded pairwise lists available online. In Table 3.15 we report the confusion matrix. Note that many errors are between similar labels, such as liberal and democrat. Ease of training, strong performance, and representation of minority classes make a majority vote system of shallow pairwise classifiers a good approach for this task.

For binary comparison, by pooling { ‘very liberal’, ‘liberal’, ‘democrat’ } and { ‘very conservative’, ‘conservative’, ‘republican’ } we achieve 76.4% accuracy; 12.1% above baseline. Table 6.8 shows the top 55 liberal and conservative words.

Religion

Religion seems to be more difficult to glean from statuses than political identity. At 54.1%, accuracy is a modest 5.3% above guessing the mode. The most highly weighted pairwise words are on Table 3.8, and Table 3.9 shows the confusion matrix. The most highly weighted word to distinguish someone who is agnostic from an atheist is ‘boyfriend’. This led us to look deeper at that pairwise classifier in Section 3.3.5. Binary labels were constructed by pooling { ‘catholic’, ‘christian-catholic’, ‘romancatholic’, ‘christian’, ‘christian-baptist’ } and { ‘atheist’, ‘agnostic’, ‘none’ }. We achieve 78.0% accuracy, 5.2% above baseline. Those words are on table 6.8. To our knowledge, there is no other multi class religion

predictor to which our results can be compared.

IQ

In a genome wide association meta study of 78,308 individuals, 336 single nucleotide polymorphisms were found to explain 2.1-4.8% of the IQ variance among the test population (Sniekers et al., 2017). We achieve 12.8% EV with a model trained on less than 2000 users and their statuses. Using ℓ_1 regularization to limit the vocabulary to the ten most informative words—final, physics; ayaw, family, friend, heart, lmao, nite, strong, ur—still yields 5.6% percent EV. The relative accuracy of such a trivial model that leverages intuitive features is a helpful comparison for any project predicting this important trait. To our knowledge, this is the only work to date that infers IQ from social media.

The selected features are also informative. Words suggesting intelligence—‘final’ and ‘physics’—are parsimonious and singularly academic. Whereas the university experience is sufficient to find users with high IQ, features inversely related to IQ are more focused on disposition. From table 6.3, agreeableness is implied by ‘family’ and ‘heart’; conscientiousness is implied by ‘family’ and ‘lmao’; and low openness is implied by ‘ur’. Overall, the list can be characterized as prosocial, or at least concerned with social relationships. Predicting low IQ with prosocial features seems to challenge some previous research.

Gottlieb et al observed that learning disabled children were more likely to engage in solitary play (Gottlieb et al., 1986). Play has also been observed to be more aggressive (Bryan et al., 1976). More directly related to our task, McConaughy and Ritter showed a positive correlation between the IQ of learning disabled boys and social competence scores; and a negative correlation between IQ and behavior problem scores (McConaughy and Ritter, 1986). For further review of the subject see (Bellanti and Bierman, 2000).

A mean square error optimal classifier seeks to generalize information about samples near the average. This can cause bias when classifying minorities, but is instructive when interpreting features. Features should say something about the majority of our sample,

those with IQ near the mean. This explains why antisocial behavior among those with extremely low IQ does not preclude prosocial behavior indicating moderately lower IQ. Reflecting the limitations of this type of study, words like ‘family’, ‘friend’, and ‘heart’ could also be caused by differing norms for social media use or many other factors. Prosocial words predicting lower IQ does however suggest interesting future work.

Sensational Interests

In this study, SIQ is the easiest continuous variable to predict, even with an order of magnitude less training data than personality. The SIQ asks lists 28 discrete interests like ‘black magic’ and ‘the armed forces’. Very similar terms can be recovered from statuses: *zombie, blood, vampire; military, marines, training*. Personality tests, on the other hand, ask more abstract questions like ‘I shirk my duties’ for conscientiousness. Many of these duties seem to be extracted in Table 3.3 such as *studying, busy, obstacles*. But many more training examples are required for similar performance.

This is the first work to demonstrate an automatic system for predicting SIQ. Previous research relied on manually counting the number of sensational interests in statuses. The count was only correlated with militarism among men; the relationship was negative for women (Hagger-Johnson et al., 2011).

Satisfaction With Life

Previous research cast doubt on the relationship between status updates and SWL (Wang et al., 2014). The number of positive words used on Facebook nationwide in a given day, week, or month, is inversely correlated with the SWL of that time period’s myPersonality participants. The interpretation of that result is that it “challenges the assumption that linguistic analysis of internet messages is related to underlying psychological states.” Here we show that a BoW model accounts for 3.4% of the variance in SWL scores. Moreover, the most important words the model finds are intuitive. Lower SWL is implied by “fucking”, “hate”, “bored”, “interview”, “sick”, “hospital”, “insomnia”, “farmville”, and

“video”. The deleterious effects of joblessness, anger, chronic illness, and isolation are well documented. Words positively associated with SWL—“camping”, “imagination”, “epic”, “cleaned”, “success”—make similar sense.

Conversational AI on Facebook Messenger is an efficacious and scalable way to administer cognitive behavioral therapy (Fitzpatrick et al., 2017). Our results show linguistic analysis can shed light on underlying psychological states. This is important to find users that could benefit from such treatment.

Belief in Star Sign

Compared to political identity, BSS has seven fewer classes and a far more homogeneous distribution. Even so, the BSS classifier performs slightly worse than the politics classifier and roughly on par to the baseline of predicting the mode. Unlike our race, gender, politics and sensational interests, we don’t wear belief in astrology on our sleeve.

3.3.4 Cambridge Analytica

With current technology, Facebook statuses are a better predictor of someone’s IQ than the totality of their genetic material (Sniekers et al., 2017). When a marketing firm adds such a tool to their arsenal it is natural to be suspicious. Indeed, The Guardian article that broke the CA story was headlined “‘I made Steve Bannon’s psychological warfare tool’: meet the data war whistleblower” (Cadwalladr, 2018). (Steve Bannon is the former chief executive of the 2016 Trump presidential campaign.) However, closer inspection of psychographic models casts doubt on their ability to add value to an advertising campaign, even when the predictions are accurate. In this paper we show that militarism is one of the most easily inferred traits. At 16.5% explained variance, it is more predictable than any of the big 5 personality traits except openness, even with just 5% of the training data. SIQ is also a much stronger predictor of aggressive behavior than the Big Five (Egan and Campbell, 2009). If this trait was actionable for the Trump campaign, it is interesting that the two most

Table 3.10: Agnostic vs Atheist Confusion Matrix

		Predicted (men)			Predicted (women)		
		Agnostic	Atheist	Total	Agnostic	Atheist	Total
True	Agnostic	40	29	69	85	22	107
	Atheist	31	55	86	31	19	50
	Total	71	84		116	41	

Table 3.11: Fair Agnostic vs Atheist Confusion Matrix

		Predicted (men)			Predicted (women)		
		Agnostic	Atheist	Total	Agnostic	Atheist	Total
True	Agnostic	36	33	69	86	21	107
	Atheist	28	58	86	34	16	50
	Total	64	91		120	37	

highly weighted features are ‘xbox’ and ‘man’. Gaming interest and gender are already available via Facebook’s advertising platform; reaching that demographic does not require an independent model. Additionally, Steve Bannon’s belief in the political power of gamers predates CA’s psychographic model by a decade (Dibbell, 2008).

Readers are encouraged to view the word lists in the Appendix through the lens of task accuracy on Tables 3.5 and 3.6. They may come to the same conclusion as the Trump campaign who, according to CBS News, “never used the psychographic data at the heart of a whistleblower who once worked to help acquire the data’s reporting – principally because it was relatively new and of suspect quality and value.” (Garrett, 2018). Performance results and extracted features allow for more informed discussion; particularly for SIQ, fair-mindedness and self-disclosure on which we report the first accurate prediction model.

There are limitations to this analysis. Our models only use statuses; Likes and network statistics could increase accuracy. Further, other traits beyond militarism may be politically useful but have no obvious demographic stand-in. Finally, we don’t have access to CA’s exact dataset and instead built our models on the myPersonality dataset.

3.3.5 Gender Bias in Atheist vs Agnostic Classifier

Highly weighted atheist words include *fucking*, *bloody*, *maths*, *degrees*, *disease*, *wifey*, and *religion*. Meanwhile, *beautiful*, *santa*, *friggin*, *thank*, *hubby*, *miles*, and *paperwork* imply the user is agnostic. This paints a picture of academic, male, disagreeable and British atheists. Agnostic words are more positive, female, and related to mundane preparation. A more complete list is shown in Table 6.8. What follows is an empirical analysis of our estimator’s gender bias, a discussion of fairness, and results debiasing the model.

In this dataset, atheists and agnostics are 33.5% and 50.3% female respectively. This is a stronger female preference for agnosticism than random surveys across the United States which report 32% and 38%, respectively (Pew Research Center, 2014). Table 3.10 shows the confusion matrices for men and women. The ratio of predicted to true agnostics is 0.945 for men and 1.35 for women. Similarly, the ratio of false atheist to false agnostic predictions is 90.8% larger for men than women. The classification of women, the minority in this dataset, is highly distorted.

Models built to generalize information often amplify biases in training data. Cooking videos elicit female pronouns in machine-generated captions 68% more than male pronouns, even though the training shows only 33% more women cooking (Zhao et al., 2017). Word embeddings used in machine translation (Zou et al., 2013), information retrieval (Clinchant and Perronnin, 2013), and student grade prediction (Luo et al., 2015) produce analogies such as “man is to computer programmer as woman is to homemaker” (Bolukbasi et al., 2016).

There are many notions of fairness defined over an individual (Dwork et al., 2012; Joseph et al., 2016; Kusner et al., 2017), population (Zafar et al., 2017; Hardt et al., 2016), or information available to the model (Grgic-Hlaca et al., 2016). Building a fair estimator often requires domain knowledge to define a similarity metric (Dwork et al., 2012), make corpus-level constraints (Zhao et al., 2017), or construct a causal model that sep-

arates protected information from other latent variables (Kusner et al., 2017). Here, we use the notion of Disparate Mistreatment to measure fairness (Zafar et al., 2017). That is, if protected classes experience disparate rates of false positive, false negative or overall misclassification, the estimator is unfair.

To mitigate Disparate Mistreatment we explicitly encode gender— $\{-1,0,1\}$ for {male, unknown, female}—in the feature vector during train time. At test time the gender of all samples is encoded as unknown. The intuition is that latent variables are amplified when they are easy to extract and correlated with the target. As demonstrated by the accuracy of our race and gender predictors, that is often the case for protected information. There often exist more informative, if more subtle, traits than the protected features. For example, atheists and agnostics report a yawning gap in those that don't believe in God, at 92% and 41% (Pew Research Center, 2014). Additionally, religiosity is shown to be correlated with both Agreeableness and Conscientiousness (Saroglou, 2010). But gender is much easier to extract than belief in God or personality. By explicitly giving the model gender information, we hope that the model will do more to extract those other features.

This approach produces much less Disparate Mistreatment of men and women. The ratio of predicted to true agnostics moves closer to parity at 1.02 for men and 1.22 for women. Additionally, the ratio of false atheist to false agnostic predictions is now only 31.8% larger for men, compared to 90.8% without intervention. The most highly weighted agnostic words for the new fair classifier are also less gendered; *hair*, *wifey*, and *boyfriend* are no longer in the top 55, as reported in Table 6.8. We also saw no decay in classification rate.

The gender bias of the atheism classifier is clear by simply inspecting its most heavily weighted features. More opaque models should be subjected to more rigorous inspection for bias.

3.3.6 BoW Conclusion

Using a similar experimental setup many traits are predicted. Performance is good and the word lists are informative about the models, how people use social media, and personality.

3.4 Deep Learning

The experiments in this section focus on building a system that can learn to accurately predict new psychometric labels with minimal data. BoW learns correlations between word counts and labels from scratch. The toolbox has been expanded with *tf-idf*, a log transformation, regularization, and LSA; but learning patterns from scratch on each new data set is a fundamental limitation. Deep learning is a way to embed information about patterns in language and their correlation to broad psychometric constructs before training on a data set of interest. SoTA performance is achieved on psychometric labels including: IQ, Big Five, SIQ, BIS and SWL. This is accomplished with an intermediate embedding that is interpretable, making the system a good replacement for LIWC in social linguistics experiments.

3.4.1 Experimental Setup

All the the statuses of each user are represented as a 1024 dimensional embedding by the following two steps.

1. **Tokenization.** RoBERTa uses a byte-pair encoding scheme that is a hybrid between character and word-level representations. Following a statistical analysis of the 160GB training corpora, 50k subword tokens were selected that can efficiently represent the entire dataset.
2. **Embedding.** A special <cls> token precedes each status which RoBERTa uses to carry global information about a text. The final output of <cls> functions is a 1024

representation of the entire status. For each user, the mean of all their statuses is taken.

Because the goal is to produce a language model that extracts personality information and not simply to maximize EV, all users with any text are used. This dampens results when users with very little text end up in the test set, but gives the model more data to learn from.

3.4.2 Multi-Task Learning

Deep learning builds intermediate representations rather than modeling the relationship between features and labels directly. When a network is trained to predict multiple labels at once, representations can be shared between tasks. Multi task learning often increases performance on all tasks as multiple objectives encourage extraction of more general features. In practice, this is a function of how similar the multiple tasks are. Completely unrelated tasks don't benefit from sharing representations and instead compete for network capacity. In our case, Big Five scores are a linear combination of item-level scores, ranging from 20 to 100 items in myPersonality. Item level labels are an order of magnitude more inherently pertinent information that can be backpropogated through the network.

Two multitask neural networks are trained. One predicts all Big Five and the other also predicts 100 item level responses from IPIP100. (Hereafter the itemlevel multitask model will be referred to as IPIP105.) There are 25,800 participants who complete the IPIP100 questionnaire. Each model is trained on a random sample of 80%. The input is the 1024 dimensional RoBERTa embedding of statuses from the <cls> token (Liu et al., 2019). The two intermediate layers consist of 256 nodes with ReLU activation and l_2 regularization of magnitude $1e-3$. There is a dramatic improvement over the state of the art BoW model, as shown in Table 3.12. The Big Five multitask model boosts performance by 44% and IPIP105 by a considerable 59%.

Table 3.12: Multitask Learning

Model	O	C	E	A	N	avg
BoW Baseline	0.171	0.120	0.141	0.090	0.100	0.124
Big5 Multitask	0.209	0.166	0.208	0.162	0.154	0.180
IPIP105 Multitask	0.214	0.185	0.222	0.197	0.172	0.198

3.4.3 Results

The IPIP questions are correlated with Big Five scores by construction, but they are also designed to be correlate with any psychological construct. We now use IPIP105 to predict the rest of the labels. LIWC, RoBERTa, IPIP105, and IPIP105 concatenated with RoBERTa are used as inputs into a fully connected NN with two layers of 256 nodes followed by a prediction layer. For BIS, SWL and IQ the final layer is dimension one. For SIQ the five facets of SIQ are predicted at once, providing some benefits of multi-task learning. The AdaMax optimizer is used with a step size of either 1e-3 or 1e-2, depending on performance. Likewise, the two NN layers have an ℓ_2 penalty of 1e-2 or 1e-1.

Table 3.13 shows the results on continuous labels. IPIP105 achieves good performance on all labels, sometimes the best by a wide margin. What follows is an discussion of the performance on each label.

Table 3.13: RoBERTa: Explained Variance

Model	IQ	SWL	BIS	Mili	Viol	Intel	Occult	Whole
BoW	0.128	0.034	0.031	0.165	0.192	0.033	0.144	0.108
LIWC	0.104	0.037	0.035	0.052	0.059	0.038	0.024	0.039
RoBERTa	0.114	0.075	0.116	0.188	0.165	0.083	0.119	0.122
IPIP105	0.140	0.193	0.254	0.183	0.156	0.140	0.116	0.094
RoB + IPIP	0.152	0.174	0.249	0.224	0.173	0.143	0.121	0.126

SWL

SWL is a good test of the model because there are only a few thousand training points, more in line with typical psychometric research than the over 100k Big Five samples. LIWC and BoW both perform slightly better than chance with 0.037 and 0.034 EV respectively. RoBERTa doubles this to 0.075 and IPIP105 doubles that again to 0.193. This narrowly

outperforms the concatenation of the two. Theoretically, if RoBERTa features offer no new information the net should learn to ignore them. The disparity can be explained by noise in the training process as well as the increased modeling difficulty going from 105 to 1129 features.

It's hard to imagine a use case for predicting SWL with an EV of 0.03. Indeed, this is what led previous researchers to posit that SWL could not be predicted from FB statuses (Wang et al., 2014). Other researchers used this same dataset to predict SWL from LIWC, Big Five Scores, and FB attributes (age, network size, relationship status, and number of photos the person is tagged in) (Collins et al., 2015). Following reasoning similar to our work, they also compare models that predict Big Five from those features before predicting SWL—the equivalent of IPIP105. Instead of EV they report mean absolute error, and compare to a baseline of guessing a random label from the distribution (rather than guessing the mean). In terms of EV this baseline would be -1.0, a generous standard to beat. The only models that beat LIWC are predicting from Big Five scores directly or predicting Big Five from FB attributes, then predicting SWL. Had the researchers been using a hold out set and comparing to a baseline with non-negative EV, it would have been clear LIWC and the rest of the rest of the models were producing noise. Our work shows that a NN training program that can predict other psychometric labels like IQ cannot predict SWL from a LIWC embedding. This is the state of much of social linguistics: interesting questions about language and personality, but tools too rudimentary to answer them. IPIP105 offers a solution in line with previous language modeling efforts in the field and an order of magnitude more explanatory power.

BIS

The pattern for BIS matches that of SWL. BoW and LIWC barely manage positive EV, while IPIP105 obtains 0.254. Impulsivity is correlated with many problematic behaviors, including disordered use of social media (Sindermann et al., 2020). Being able to infer

somewhat accurate labels opens up the door to studying that relationship with much larger datasets.

IQ

IQ stands out as the trait most invariant to model selection. The RoBERTa + IPIP105 perform best, but only 50% better than LIWC. Recall from the BoW section, the ten best words to predict IQ are: final, physics; ayaw, family, friend, heart, lmao, nite, strong, ur. LIWC contains categories for both family and friend, which will capture much of this list. LIWC also counts the number of words more than six characters. Considering the vocabulary subtest is the score most correlated with IQ, this is also valuable information (Jensen, 2001). Still, even compared on LIWC's strength it is the poorest performing model on IQ.

SIQ

For both SWL and BIS, the concatenated RoBERTa and IPIP105 embedding have lower accuracy due to the added difficulty of the modeling problem. All five facets of SIQ perform as expected. On some facets RoBERTa did better than IPIP105, on others worse. But the concatenation always did better than either. BoW performed best on two facets: Violent-Occult and Occult Credulousness. Interests, it seems, are more amenable to a BoW model than broad traits. The most highly weighted words listed in Table 3.3 are interests that appear on the questionnaire such as zombie or vampire. RoBERTa, on the other hand, encodes more general information in the [CLS] token and specific word counts will be smoothed over by the embedding and subsequent averaging over statuses. All facets taken together, deep learning was more consistent as LIWC and BoW only achieved EV of 0.033 and 0.38 on Intellectual Recreation compared to 0.143 for IPIP105.

3.4.4 Gender

Deep learning results predicting gender are less impressive. There are over 100 thousand training samples with gender labels therefore transfer learning is less important. Any relevant information from the IPIP105 model can be learned from scratch at training time. Further, as described in Chapter 2, psychometric questions are designed to not load differently based on gender. RoBERTa obtains 86.7% accuracy, compared to 92.8% with the BoW model that uses a 500k vocabulary.

Table 3.14: RoBERTa Gender Prediction

Model	Accuracy	AUC
RoBERTa	0.867	0.940
IPIP105	0.826	0.905
Rob + IPIP	0.862	0.937

Politics

The large disparity in class sizes complicates training. Unweighted, the model will see more “Liberal” samples and learn to only predict other classes when there is compelling evidence. For both the politics and religion classifiers samples are weighted by $\frac{1}{\sqrt{n_{class}}}$ where n_{class} is the number of samples in that class. This is a compromise between a more extreme reweighting of $\frac{1}{n_{class}}$ which caused the model to rarely predict common classes.

The results differentiating the four most common categories—liberal, conservative, moderate and doesn’t care—are shown on Table 6.2 and Table 3.16. These conform to common sense; for example liberals are more easily differentiated from conservatives than moderates. Interestingly, conservatives are easiest to distinguish from “doesn’t care” while liberals are difficult to distinguish from moderates. As with gender, BoW performs better.

Religion

Like gender and politics, deep learning fails to match BoW predicting religion. What these labels have in common is that they are about group membership rather than a trait that

Table 3.15: RoBERTa Politics Confusion Matrix

		Actual				Total
		Con.	DC	Lib.	Mod.	
Predicted	Conservative	612	174	395	142	1323
	Doesn't Care	203	533	399	140	1275
	Liberal	309	298	977	224	1808
	Moderate	18	9	24	13	64
	Total	1142	1014	1795	519	4470

Table 3.16: RoBERTa Politics AUC

	Moderate	Liberal	Doesn't Care
Conservative	0.722	0.743	0.796
Doesn't Care	0.727	0.727	
Liberal	0.621		

Area Under the Curve (AUC) refers to the receiver operator characteristics of a model. 0.5 means a classifier performs as well as a coin flip and 1.0 is perfect performance.

everyone has. The continuous labels that BoW did well on share characteristic. The IQ predictor with just five words obtained 0.056 EV by leaning heavily on group membership; those that talk about college or finals are likely to be smarter. For interesting in the Violent Occult, BoW counts references to zombies and vampires vs Christianity. Apparently there was no such singular group topic for BIS or SWL.

Theoretically, RoBERTa should yield better results than BoW. It's a much richer model that can handle negation and take context into account. However, much of the modeling power is not used in these experiments because each person is represented as the mean of all their statuses. Therefore, if some rare word (eg. zombie) is very useful for classification it might not be salient even in the status representation, and much less so in the user representation. Given users have on average more than a hundred statuses, it's prohibitive to hold all of them in memory and while fine tuning the whole model. Training on single statuses is one solution however when that was tried training was very inconsistent and often made the model worse. Each status is not very informative on its own.

Table 3.17: RoBERTa Religion AUC

	None	Christian	Catholic	Atheist
Agnostic	0.744	0.800	0.801	0.602
Atheist	0.754	0.851	0.834	
Catholic	0.825	0.713		
Christian	0.824			

Table 3.18: RoBERTa Religion Confusion Matrix

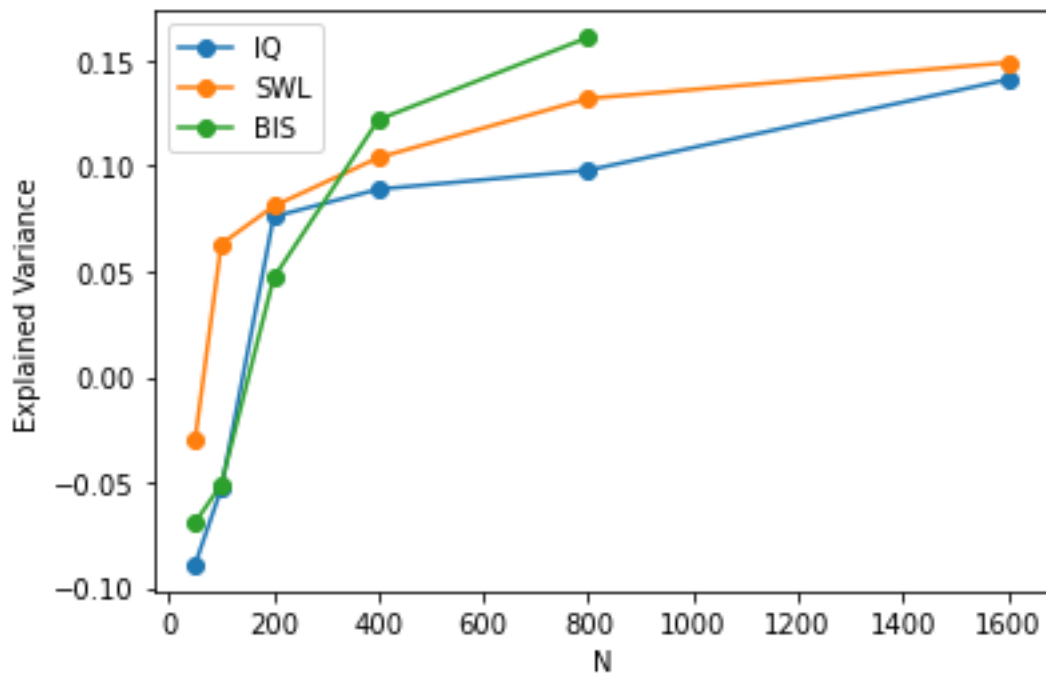
		Actual Label					Total
		Agnostic	Atheist	Catholic	Christian	None	
Predicted	Agnostic	98	57	57	174	24	410
	Atheist	66	81	41	107	18	313
	Catholic	27	24	186	204	11	452
	Christian	46	22	137	756	26	987
	None	27	38	40	113	64	282
	Total	264	222	461	1354	143	2444

3.4.5 Restricting Training Samples

Because this model is useful for social psychology where sample sizes are rarely in the thousands, classifiers for BIS, IQ and SWL are trained on restricted numbers of samples. The architecture is a two layer NN with 128 nodes. Batch size is 50 with twenty epochs. To ensure results are not unduly effected by noise each network is trained ten times from a random initialization and the median is reported. Figure 3-1 shows logarithmic improvement that achieves roughly 10% EV using just 400 samples.

3.4.6 Deep Learning Conclusion

LIWC fails to extract personality information on many traits even when used as input to a neural network. If all LIWC features together cannot predict a trait, then the many studies that claim a single LIWC category is correlated with a trait should be viewed with suspicion. There are many ways random noise looks like signal when strict validation schemes are not used. IPIP105 is a good alternative for extraction of personality information from text.

Figure 3.1: IPIP105 Models Trained on Less Data

Each point is the median test set EV of 10 different runs. BIS only has 1261 samples so the point at $n = 1600$ is missing.

Chapter 4

ML vs LIWC: a case study in predicting grandiose narcissism

Grandiose narcissism involves traits such as leadership, authority, grandiosity, exhibitionism, entitlement, and exploitativeness (Ackerman et al., 2011; Raskin and Terry, 1988). In terms of other major traits in personality and social psychology (Soto and John, 2017), narcissistic individuals tend to be disagreeable extraverts (Paulhus, 2001; Paulhus and Williams, 2002; Vize et al., 2018); they also tend to be slightly more open-minded than average (Vize et al., 2018), and more masculine than feminine (Grijalva et al., 2015). Here, we take a language approach to automatically identifying narcissistic individuals based on previously-determined linguistic profiles for disagreeable, extraverted, open-minded, and masculine people (Cutler and Kulis, 2018).

The general problem of accurately identifying who is narcissistic has perplexed psychologists for decades, in part because identifying who is narcissistic can be quite challenging (Back et al., 2011; Paulhus, 1998). Understanding who is narcissistic is useful for a wide variety of reasons. As a few examples, narcissism is associated with consumer behavior—such as conspicuous consumption (Bagwell and Bernheim, 1996; Griskevicius et al., 2017), so knowing who is narcissistic helps to target potential buyers in a marketplace (Cisek et al., 2014; Sedikides et al., 2007); narcissism is also positively associated

with dangerous sexual behavior (Jonason et al., 2015) and may be linked evolutionarily to short-term mating in general (Holtzman and Strube, 2011; Schmitt et al., 2017), so knowing who is narcissistic may help to identify people who are at risk for contracting sexually transmitted diseases. These two simple examples illustrate that knowing who is narcissistic could be useful in strikingly different domains. So, the need to readily assess narcissism is clear.

An emerging literature has shown how one should assess narcissism in an efficient way when one cannot conduct any formal psychometric testing. For instance, personal appearance cues of narcissism may help one to infer the presence of narcissism. Some evidence points to narcissism being associated with attractiveness (Holtzman and Strube, 2010), but this effect appears to be attributable to the factors that are mostly within the self-regulatory control of the person (Holtzman and Strube, 2013), such as dressing up in fancy clothes or using make-up (Vazire et al., 2008), and the effect is not due to some sort of innate attractiveness. Research on face perception has shown that narcissism may be linked to facial appearance (Giacomin and Rule, 2019; Holtzman, 2011; Shiramizu et al., 2019) and to certain profile picture qualities on social networks (Buffardi and Campbell, 2008). Additionally, non-verbal behaviors may also be useful in pinpointing who is narcissistic (Back et al., 2011). Thus, a wide variety of cues may signal narcissism, even when one does not have formal psychometric test results from a narcissistic person.

Language serves as another potentially useful cue of narcissism, as language has been linked to personality more generally (Fast and Funder, 2008; Kern et al., 2014; Schwartz et al., 2013b). Early work on narcissism and language used word-counting methods to hone in on the putative narcissistic tendency to focus on oneself, such as the tendency to use first-person pronouns (Raskin and Shaw, 1988), however, this turned out not to be a valid cue of grandiose narcissism (Carey et al., 2015). Other research identified several other language cues (Preotiuc-Pietro et al., 2016), including using more sexual words and more swear

words (Holtzman et al., 2019, 2010). Almost all of these language effect sizes are small (i.e., $r < .15$) by modern standards (Gignac and Szodorai, 2016; Hemphill, 2003). One constraint on using language as a cue of narcissism is that the most widely used method (i.e., word counting, using the Linguistic Inquiry and Word Count [LIWC]) categorizes words in a binary fashion (Pennebaker et al., 2001, 2003); that is, words either belong to a LIWC category or they do not. One approach to overcome this limitation is to relax the assumption that words either strictly do or do not belong to a category, which we attempt to do by using machine learning.

Machine Learning (ML) is the study of computational algorithms that use data for prediction, classification, and decision making. The key feature is that instructions are not explicitly programmed by the researcher, but rather obtained from patterns and associations found in the data. The methods are often statistical in nature, but the choice of model usually favors predictive power at the expense of interpretability. ML can be applied to language analysis, and has been used a few times in the study of narcissism (Sumner et al., 2012; Preotiuc-Pietro et al., 2016). The instantiation we use (Cutler and Kulis, 2018) is based fully on multiple regression, and here we demonstrate that it can help researchers identify who is narcissistic. The BoW model in the previous chapter is used to predict four pertinent variables: Extraversion, Agreeableness, Openness, and Masculinity.

In the context of this literature, and based on the effect sizes from two major published articles (Grijalva et al., 2015; Vize et al., 2018), we had four pre-registered point predictions (<https://osf.io/8uard>). We predicted that narcissism would be positively associated with using language like an extravert ($r = .30$), that narcissism would be negatively associated with using language like an agreeable person ($r = -.20$), that narcissism would be positively associated with using language like an open-minded person ($r = .10$), and that narcissism would be associated with using language like a man and not like a woman ($r = .20$). The idea is that we can leverage the information from the BoW models in order to identify

narcissistic individuals based on their language use. In an additional exploratory analysis, we analyzed the possibility that incorporating a pre-trained model based on MyPersonality data would outperform a model based solely on LIWC.

Of note, this chapter is written with a non-engineering audience in mind so some concepts are explained at a more basic level.

4.1 Method

This paper was pre-registered at the Open Science Framework using the aspredicted.org pre-registration form [<https://osf.io/8uard>]. The data described in the Participants subsection were collected and used as part of H. Dorrough's IRB-approved thesis, which did not entail machine learning. So this is old data with new (ML) analyses.

4.1.1 Participants

Participants (total $N = 1,160$; valid $n = 471$) were recruited via Sona Research Software at a large public comprehensive university in the southeastern part of the United States. A large number of the participants failed an attention check question embedded in the survey, leading us to exclude them from all analyses; another exclusionary criterion was that participants had to provide at least 100 words in response to the prompt described below. The final set of participants included people 18 to 25 years of age ($M = 19.57$, $SD = 2.65$). The sample was 67% female; 65% of the respondents were white, while 25% of the respondents were Black.

4.1.2 Materials

Participants completed the 40-item, forced-choice (1 vs. 2) Narcissistic Personality Inventory (Raskin and Terry, 1988), in which a score of 1 is the lowest score possible and a score of 2 is the highest possible score. An example non-narcissistic option is "I prefer to blend

in with the crowd” whereas an example narcissistic option is “I like to be the center of attention.” The measure is a fairly traditional assessment of grandiose narcissism; it is valid (Raskin and Terry, 1988), and it usually yields good reliability. Here, it produced acceptable reliability (Cronbach’s $\alpha = .77$). The mean scores were quite typical for college samples ($M = 1.41$; $SD = 0.18$).

Other measures included in this study were the CESD-R to assess depression (Eaton et al., 2004), the Big Five Aspects Scale self-report of neuroticism (DeYoung et al., 2007), and the Pathological Narcissism Inventory (Pincus et al., 2009). None of these were pre-registered for analysis and so we do not report results about them.

4.1.3 Procedure

Participants completed the study online and did not have to come into the lab. They completed the consent form, the demographics form, the personality measures (blocks and items were randomized). At the end, the participants typed their response to the following prompt:

For the next 20 minutes, write about whatever comes to your mind. Think about what your thoughts, feelings and sensations are at this moment. Write about them as they come to you; follow where your mind naturally goes. Please do not include any identifying information in your writing, like your name. Please write below in the text box.

This is a stream-of-consciousness task that is modeled after (Pennebaker and King, 1999). At the end of the study participants read the debriefing sheet and logged off.

4.1.4 Quantitative Approach

We compare five models built using LIWC representations and ML methods. The LIWC model is constructed in two steps: First, obtain a numerical representation of the text

through LIWC categories, and second, use a statistical model to relate the categories to narcissism. The ML model is constructed in three steps: First, process the raw text to obtain a numerical representation; second, “embed” in a lower-dimensional space by applying an existing language model to reduce the number of variables; third, use a statistical model to relate the embedded variables to narcissism. We describe in greater detail these processes in the following sections.

4.1.5 LIWC Text Processing

LIWC is a program for parsing text, assigning words to categories based on grammatical role or content according to a pre-defined dictionary, and returning the proportion of words in the document for each category (Pennebaker et al., 2015). LIWC can be used to predict (Schwartz et al., 2013b) or understand (Holtzman et al., 2019) personality traits. The output includes around 90 variables representing categories such as personal pronouns, social processes, power, health, along with a few summary variables. Discarding a few non-suitable variables left 84 to potentially be used in modeling.

Two sets of LIWC categories are used to predict narcissism. The smaller set, LIWC4, uses the categories Anxiety/Fear Words, Tentative Words, Sensory and Perceptual Processes, and Home, as these were found by (Holtzman et al., 2019) to be the four categories most strongly correlated with narcissism. The larger set, LIWCFull, includes 84 LIWC categories.

4.1.6 ML Text Processing

We suggest three modifications to improve upon LIWC representation of text.

1. Weight words by importance. In natural language, “cancer” and “fecund” convey information about health, but with different magnitude and direction. LIWC only uses hard assignment (0,1) of a word to a category (meaning that a word either fully

belongs to a category or not), so all of these would be considered ‘health words’.

2. Expand the dictionary. State of the art language models in machine learning, optimized for prediction accuracy, represent hundreds of thousands of words (Pennington et al., 2014) or tens of thousands of sub-words (Devlin et al., 2018). These often include mis-spellings, slang, and different tenses. LIWC2015 has a relatively limited dictionary of approximately 6,400 words (Pennebaker et al., 2015), and so may fail to capture much of the information in text.
3. Extend categories for traits related to narcissism. LIWC is limited to 90 general purpose (often functional) categories. It does not contain categories for speaking like an extrovert, for instance. For our task of understanding narcissism, that would be particularly useful.

Taken together, these frame our approach as an extension of LIWC. We still count words and group them in categories, but choose more relevant categories and use ML to define the extent to which words belong in each category (without the hard (0, 1) assignment). The text is represented as a *tf-idf* BoW (Salton and Buckley, 1988). Categories are defined by personality and politics labels (eg. libertarian vs conservative). The weight that each word is assigned to those categories is defined by the models trained in Section 3.3.

4.1.7 Personality Embedding

The sheer number of words encountered in everyday language can overwhelm multiple regression models. Regression directly from *tf-idf* values to narcissism would produce a model that is difficult to interpret and poorly fit because noise in so many dimensions would drown out the signal. Therefore, language models often make use of a lower-dimensional embedding: a limited set of variables that can summarize a text. LIWC serves as such an embedding, which is interpretable, but much information is discarded as shown by poor performance predicting gender and personality (Schwartz et al., 2013b).

To reduce the number of variables, we use the models discussed in (Cutler and Kulis, 2018), trained on myPersonality data (Stillwell and Kosinski, 2012), to create embeddings. This dataset contains Big Five personality data for over 3 million participants. Of those, approximately 100k also had enough Facebook status updates (greater than 1000) to use in a tf-idf language model. From these 100k observations, there is enough information to model a relationship between text (from concatenated status updates) and personality, as well as what Facebook pages they have liked. A subset of those users also had profile information about religious beliefs (12,000), political affiliation (20,000), and responses to a sensational interests questionnaire (4,000) (Egan et al., 2003). The modeling was done using ridge regression (Hoerl and Kennard, 2000), which is the same method used for other modeling in this study, and is described in the next section.

Two ML embeddings are included in the analysis. The smaller embedding, **Personality4**, relates text to four variables: Gender, Openness, Extraversion, and Agreeableness. The larger embedding, **PersonalityFull**, relates these four and an additional 57 variables representing religious views, political identity, sensational interests, and Facebook “likes”. Note that we do not directly observe any of these variables on the participants in our study, but it is not necessary to. Using the pre-trained Cutler & Kulis model to obtain their predicted values from the participant’s text serve as lower-dimensional, interpretable intermediate summaries which can then be related to the participant’s narcissism score.

From the LIWC and ML embeddings, five models are tested to see which is best at predicting narcissism. See Figure 4-1 for a summary of the process that produces these five models.

1. **PersonalityFull**: All 61 variables from ML processing.
2. **Personality4**: Gender, Openness, Extraversion, Agreeableness.
3. **LIWC4**: Anxiety/Fear, Tentative, Sensory/Perceptual Processes, and Home

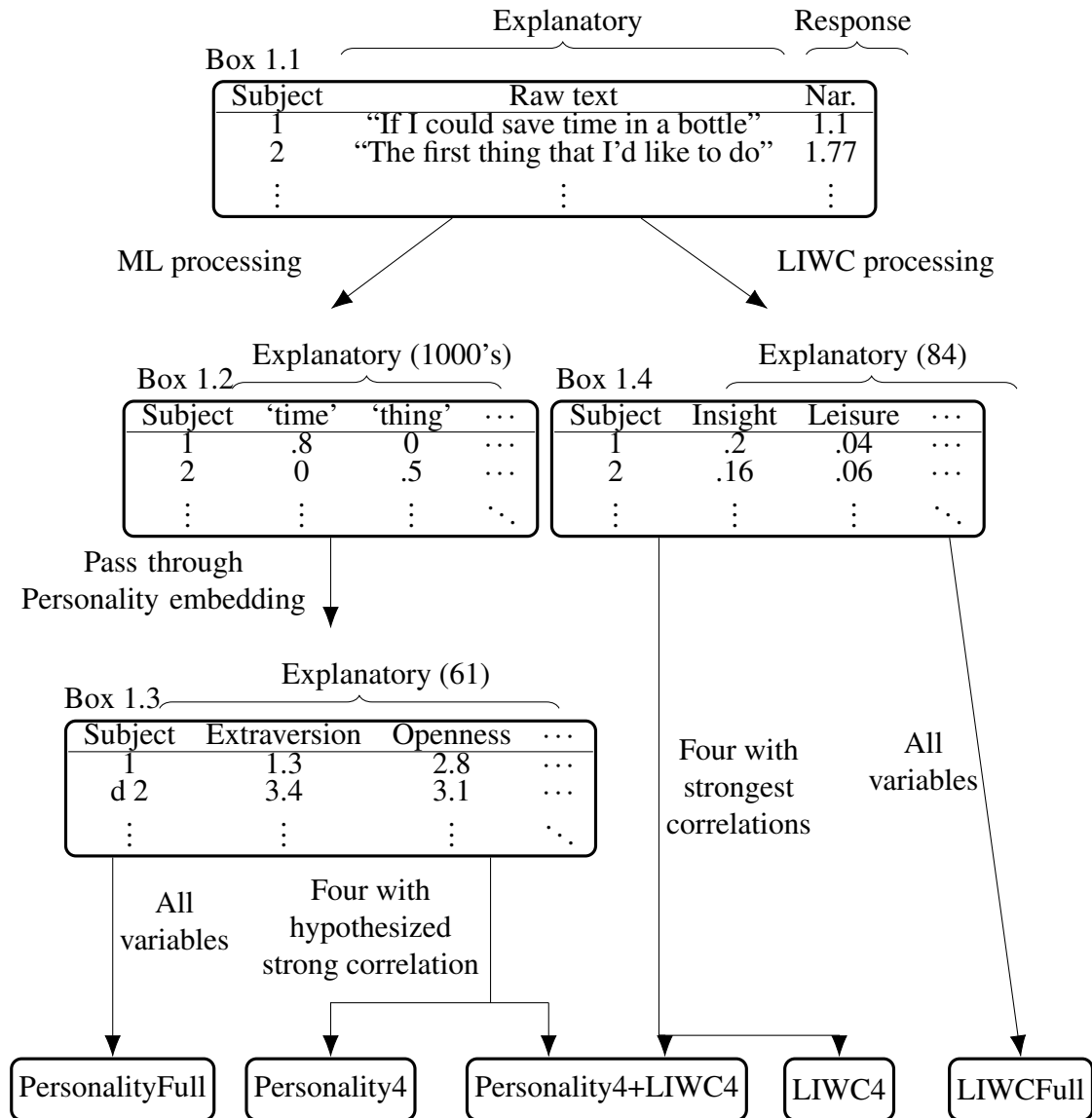
4. **LIWCFull**: All 84 LIWC categories.
5. **Personality4+LIWC4**: Concatenation of Personality4 and LIWC4.

4.1.8 Statistical Modeling

We use the same type of statistical models used in (Cutler and Kulis, 2018), which are ridge regression (on numeric variables such as strength of Big Five characteristics) and ridge classification (on categorical variables, such as political identity) (Hoerl and Kennard, 2000). These are variants of multiple regression with a penalty term scaling with the magnitude of the coefficient estimates, in effect preferring simple models over complex models. This reduces overfitting (Dietterich, 1995), a phenomena in which the model fits the noise and peculiarities in the data rather than the general pattern. An overfit model will have deceptively excellent prediction performance on the original data, but generalize poorly to predictions on new observations. As a further safeguard against overfitting, the penalty term is chosen by cross-validation (Hastie et al., 2009, Chapter 7). Participants were randomly split into training and test groups with probabilities set to .8 and .2, respectively. Using the RidgeCV module in sklearn, a regularization parameter is selected (we search over 100 log-spaced values from 0.1 to 100), and a model is fit. Models were fit on the training group data and evaluated on the test group.

4.2 Results

To find out which of the five language models best captures narcissism information from sample text, we calculate the proportion of variance explained by the model. Specifically, we train the model on a randomly chosen 80% of the data (the training set), and calculate the coefficient of determination, R^2 , on the remaining 20% of the data (the test set). R^2 is $1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$, where N is the number of observations in the test set, y_i and \hat{y}_i are the actual and predicted narcissism scores respectively for the i th observation in the test set, and \bar{y} is

Figure 4-1: Text Embedding Schematic

The workflow transforming text for predicting narcissism. The left path is ML processing, and the right path is LIWC. Both begin with the pairing of respondent's text and narcissism scores as in Box 1.1. In the ML path, tf-idf values for each term and subject are collected in Box 1.2. The large number of variables is reduced using the Personality embedding, with results in Box 1.3. In the LIWC path, proportions of terms in each category are given in Box 1.4. Finally, five ridge regression models predicting narcissism are constructed using explanatory variables from Boxes 1.3 and 1.4.

Table 4.1: Narcissism Prediction Performance (as measured by R^2)

Embedding	<i>Mean</i>	<i>Median</i>	<i>SD</i>
LIWCFull	-5.570	-0.181	11.500
LIWC4	-0.006	0.001	0.030
Personality4	0.037	0.043	0.044
PersonalityFull	0.029	0.029	0.047
Personality4+LIWC4	0.029	0.037	0.047

the average narcissism score in the test set. R^2 can be interpreted as how much better a model does than a naive strategy of guessing the mean.

If R^2 is calculated on the same set of data the model is trained on, then $0 \leq R^2 \leq 1$. However, if calculated on a test set different from the training set (as in our study), then the predicted values can be worse than guessing the mean and R^2 can be negative. This can happen if the test set and training set are sufficiently dissimilar, or if the model fails to capture the trend and is overfitting noise in the training set. See the supplemental material for further details and a simple example of this phenomena.

The R^2 value clearly depends on the random split between training and test sets, so in order to ensure a fair comparison, we repeat this calculation on 100 different random train/test splits and report the mean, median, and standard deviation of R^2 for each tested model in Table 4.1.

This validation strategy is not pre-registered, and differs from the more common approach of simple k-fold validation. However, this enables the selection of a hyperparameter on data that will not be used to report results.

Consider each model in turn, beginning with LIWCFull. Regressing on all 84 LIWC categories embedding produces negative R^2 . The median of the 100 repetitions is -0.181, meaning the fitted model is doing worse than simply using the mean. This is due to the large number of variables containing so much noise that whatever signal is present is hard for the model to find. Also notice there is enormous variance and a much lower mean, indicating a strong left skew and the presence of a few train/test splits with extremely poor performance.

LIWC4 (Sports, Total second person, Swear words, and Optimism/energy) produces practically zero R^2 , with a median of 0.001. Because this embedding is a strict subset of LIWCFull, it contains less narcissism information. However, much of the noise from categories unrelated to narcissism has been removed and the modeling is simpler, thus the higher score. Given the R^2 is essentially zero, any insights drawing on all four of these text categories will not be informative.

Personality4 (openness, extraversion, agreeableness, and masculinity) uses the variables we hypothesized would be correlated with narcissism. This performs the best of the models we tested, with a median R^2 of .043.

PersonalityFull consists of Personality4 plus 57 other language features. These features likely add relevant information, but the results are considerably worse than when regressing on Personality4. As seen with LIWCFull, separating the signal from the noise in so many variables is difficult with the relatively small number of individuals in this analysis.

Finally, Personality4+LIWC4 performs worse than Personality4 alone, showing that the LIWC categories have such a low signal-to-noise ratio that including them is actually harmful.

A Mann-Whitney test on the R^2 scores of LIWC4 and Personality4 shows that the median of the Personality4 model is higher, with a p-value of $< .001$ (Mann and Whitney, 1947).

4.2.1 Preregistered Correlation Prediction

In addition to exploring the general predictive capabilities of these different approaches, we pre-registered specific hypotheses that writing in a manner reflecting Openness, Extraversion, Agreeableness, and Masculinity would be correlated with narcissism as follows: .100, .300, -.200, and .200. The observed correlations turned out to be: .092, -.053, .254, and -.050. Summaries of the predicted and observed correlations, along with p-values and con-

Table 4.2: Correlation Values

	Predicted r	Observed r	p -value for r	95% CI
Openness	.100	.092	.046	[0.002, 0.181]
Extraversion	.300	-.053	.218	[-0.143, 0.037]
Agreeableness	-.200	.254	< .001	[0.167, 0.336]
Masculinity	.200	-.050	.274	[-0.140, 0.040]

Comparing observed Personality4 correlations with pre-registered predictions. Predicted correlation values between narcissism and Personality4 are on osf.io. P -value is relative to a null hypothesis of zero correlation between narcissism and the observed variable. These analyses are not corrected for running multiple tests

confidence intervals, are found in Table 4.2.

4.3 Discussion

We predicted that narcissists would write in ways that were disagreeable (Vize et al., 2018), extraverted (Vize et al., 2018), openminded (Vize et al., 2018), and masculine (Grijalva et al., 2015). The data supported our preregistered hypothesis that narcissists use openminded language, but the data for the other hypotheses did not match our predictions. We had the secondary goal of determining if LIWC could capture narcissism (Holtzman et al., 2019), and we found that—at least in this data from 471 people—LIWC profiles did not do so. Unlike in Holtzman and colleagues (2019), LIWC generally failed to capture narcissism, whether we used a small set of empirically-driven LIWC predictors or the full LIWC profile. This suggests that if psychologists are interested in leveraging language to profile narcissistic personality, it will be necessary to use machine learning (e.g., by collaborating with experts or by training the next generation of psychologists in using these methods) as in the Cutler-Kulis model (Cutler and Kulis, 2018). We speculate that these suggestions apply to the literature on language and personality more broadly. In this discussion, we will recap the main results and consider why narcissists write in openminded ways and we reflect on why narcissists tend to use agreeable language; we will also discuss the value of

machine learning and point to deep learning as potentially an even better tool for this task.

Previous research has shown that the Cutler-Kulis model (Cutler and Kulis, 2018) explains less than 20 percent of the variance in personality assessments. Unfortunately, narcissism assessments were not available in the dataset used in the development of the Cutler-Kulis model. So, we used the Cutler-Kulis model to analyze the new data and then compute Agreeableness, Openness, Extraversion, and Masculinity for each participant, based solely on the text provided. It turned out that people who wrote in openminded ways (as identified by Cutler-Kulis using ML) were more narcissistic (as psychometrically assessed in the 471 participants in the new sample).

The association between narcissism and writing in openminded ways turned out as one of the remarkable successes of this project. The prediction (.10) and the observed association (.09) were nearly identical. This finding means that narcissists wrote in ways that are characteristic of openminded people. This result is in line with research showing that narcissists are more openminded than average (Paulhus and Williams, 2002).

To our surprise, we found that narcissism was positively associated with using agreeable language. One possibility is that narcissists find their ideas agreeable (they are their own yea-sayers) and so, one somewhat humorous interpretation is that—within the realm of private (anonymous) talk with oneself—the conversation turns out to be quite agreeable. Perhaps this is a mechanism for overconfidence. Still, this finding is very different from our Bayesian prior, and so we must see a replication before it is interpreted with confidence.

We speculate that the fact that narcissistic individuals did not write in especially extraverted nor masculine ways may have been because they were writing in private rather than in public. The typical narcissist may inflate their extraversion (Paulhus and John, 1998) and perhaps even their masculinity in an effort to appear dominant in the social sphere. However, dominance has little obvious value in private contexts. So, one possibility is that narcissistic individuals would speak in extraverted and masculine ways in public,

even though they do not write (in private, anonymous contexts) in such ways. The lack of an association between narcissism and masculine private language deserves more attention. This finding has implications for inflated masculine tendencies in public (Johnson, 2019), constituting a potentially interesting avenue for future research. It reiterates the need to consider contextual factors (such as public versus private writing) when understanding language usage and personality. Some research has highlighted this importance (e.g., (Rodriguez et al., 2010)), but more work needs to be done. Indeed, it would be fascinating to integrate more granular contextual information with language-based personality inferences, perhaps using the DIAMONDS situational assessment (Rauthmann et al., 2014).

In terms of limitations, the Cutler-Kulis model is a rudimentary machine-learning model of language. Rather than counting words, other language models represent each word by a 50 to 300 dimensional embedding derived by how the word was used in a large language corpus (e.g. wikipedia) (Pennington et al., 2014). This introduces information from a much larger dataset, and allows more fine-grained information to be extracted in an embedding. Longer text is represented by combining the constituent word embeddings, for example by taking their mean. Regression from a text's embedding to a variable of interest may then be performed. IBM has trained one such model on tweets and Big Five personality which is described in (IBM Watson, 2019). The outputs of that model could be used like the outputs of the Cutler-Kulis model as a personality embedding. Another option is to use a general language model and regress directly to narcissism. The current state of the art, used to process Google's search queries, is BERT (Devlin et al., 2018). Like other deep learning models it produces word embeddings, but BERT has the ability to condition a word's representation on the context in which it appears. This allows the model to use syntax and grammar to make sense of language. The model is cumbersome (345 million parameters), and not specifically designed to extract an author's personality. However, it is a rich embedding that has been shown to perform well on myriad downstream tasks. Integration of

these methods with the narcissism literature would assist in more accurately identifying narcissistic individuals. A second limitation is that this study was based on private stream-of-consciousness text writing, and may not generalize to public speech or text written for a public audience. More research is needed to compare public and private language use of narcissistic people more generally.

In conclusion, ML approaches to understand narcissistic personality appear to have some promise, with the Cutler-Kulis model capturing a significant amount of variation in narcissism scores. We found that narcissists have an openminded language profile (as expected), and that they have a (surprisingly) agreeable language profile. In order to understand how narcissistic individuals use language, it will be useful to move beyond word-counting approaches and instead employ more advanced machine learning language models.

Chapter 5

The Lexical Hypothesis

The lexical hypothesis predicts that most of the socially relevant personality information is embedded in natural language (John and Srivastava, 1999). As the philosopher J. L. Austin put it

Our common stock of words embodies all the distinctions men have found worth drawing, and the connections they have found worth marking, in the lifetimes of many generations: these surely are likely to be more numerous, more sound, since they have stood up to the long test of the survival of the fittest, and more subtle, at least in all ordinary and reasonably practical matters, than any that you or I are likely to think up in our arm-chairs of an afternoon—the most favored alternative method. (Austin, 1961)

Thurstone and psychometricians that followed him mapped these distinctions by asking people how much words describe them. These answers could be arranged in a person by word matrix which was then factorized. As far as the reconstruction loss of those matrices is concerned, personality adjectives (and those they describe) can be represented in roughly five dimensions (Thurstone, 1934; Hofstee et al., 1997; John and Srivastava, 1999).

Later people came to be scored on these Big Five dimensions not by whether single words described them, but by agreement to entire phrases such as “I often feel blue”. This allowed for fewer questions while maintaining sufficiently good orthogonality and test-retest agreement (Goldberg et al., 2006).

Digman collected 14 Big 5 studies and performed factor analysis on the correlations

between factors in each study (Digman, 1997). Given that each factor is designed to be orthogonal one expects noise and distortions to be opportunistically grouped, but for there to remain five essential dimensions. The eigenvalues, however, fell off steeply. Normalized to sum to 1, the average eigenvalues are: 0.41, 0.25, 0.17, 0.10, 0.07. Additionally, the first and second eigenvectors followed the same pattern in each study. The first factor (α) loaded on Agreeableness, Conscientiousness, and Emotional Stability. The second (β) loaded on Extraversion and Openness to Experience. It is instructive to hear Digman's own description of these higher-order traits.

Another possibility...is that Factor α represents the socialization process itself. From Freud (1930) to Kohut (1977), from Watson (1919) to Skinner (1971), personality theorists of various persuasions have been concerned with the development of impulse restraint and conscience, and the reduction of *hostility*, *aggression*, and *neurotic defense*. From this point of view, Factor α is what personality development is all about. Thus, if all proceeds according to society's blueprint, the child develops superego and learns to restrain or redirect id impulses and to discharge aggression in socially approved ways. Failure of socialization is indicated by neurosis, by deficient superego, or by excessive aggressiveness.

Factor β may be interpreted as another very broad concept in personality theory: Personal growth versus personal constriction. Like the socialization interpretation of Factor α , this concept is extremely broad (indeed so broad that it has sometimes been rather difficult to define) and is related to a perspective on personality very different from those that have come from the psychoanalytic or behaviorist traditions: This is the perspective of personal growth theorists, such as Rogers and Maslow. For Rogers (1961) "the organism has one basic tendency and striving—to actualize, maintain, and enhance the experiencing organism" (p. 487). Similarly, Maslow (1950) suggested ways to achieve personal growth: One should "experience things fully, vividly . . . choose risk . . . make the growth choice" and "use your intelligence" (pp. 11-34). For both of these theorists, personal growth or the actualization of self meant an enlargement of self by a venturesome encounter with life and its attendant risks, by being open to all experience, especially new experience, and by the unfettered use of one's intelligence.

Despite the close fit to theory, the Big Two never became a common descriptive framework for psychologists. Ashton et. al argued that α and β were due to Big Five factors sharing subfactors (eg. Politeness contributing to both Agreeableness and Conscientiousness), not higher-order structure. To show this they fit their two candidate models and

found their alternative to the hierarchical model fit better (Ashton et al., 2009). This is not surprising considering the three datasets (481 adults, 480 students, and 230 students) were answers to the Big Five Aspect Scale (DeYoung et al., 2007) designed to produce five factors instead of two.

Half a decade earlier, Ashton, Lee, and Goldberg performed factor analysis on 1,710 English adjectives based on responses from 310 Americans and Australians (Ashton et al., 2004). The eigenvalues for the first 12 components are 88.1, 80.9, 62.9, 52.4, 33.4, 27.2, 25.2, 20.9, 18.7, 17.4, 16.6, and 15.7. Factor analysis was performed with 1-7 factors. Despite the two large leading eigenvalues that loosely map to the Big Two, the Big 5 emerged separately with five factors. Note that in factor analysis individual-level variance not well captured by the overall structure can be represented independently in the diagonal variance matrix Ψ .

What follows is a deep learning approach which yields two factors that look much like α and β . Ultimately, factors produced by different methods are compared by researchers describing them qualitatively. I have no special ability to theorize about factors so will spend more time establishing the stability of this method under different modeling choices.

5.1 Deep Lexical Hypothesis

The word by person matrix approach in psychology is a choice among many word embeddings. In this case words are represented by the personal judgements of a few hundred people (Thurstone, 1934; Ashton et al., 2004). Thurstone used 60 adjectives and 1,300 people to create a 60x60 co-occurrence matrix. People generate statistics (how often two words are used to describe someone) to represent words. In 1988 this same method was introduced in computer science as an information retrieval technique called Latent Semantic Analysis (Furnas et al., 1988). This remained a popular way to find compact representations of words until roughly 2015 when masked language modeling and neural networks became

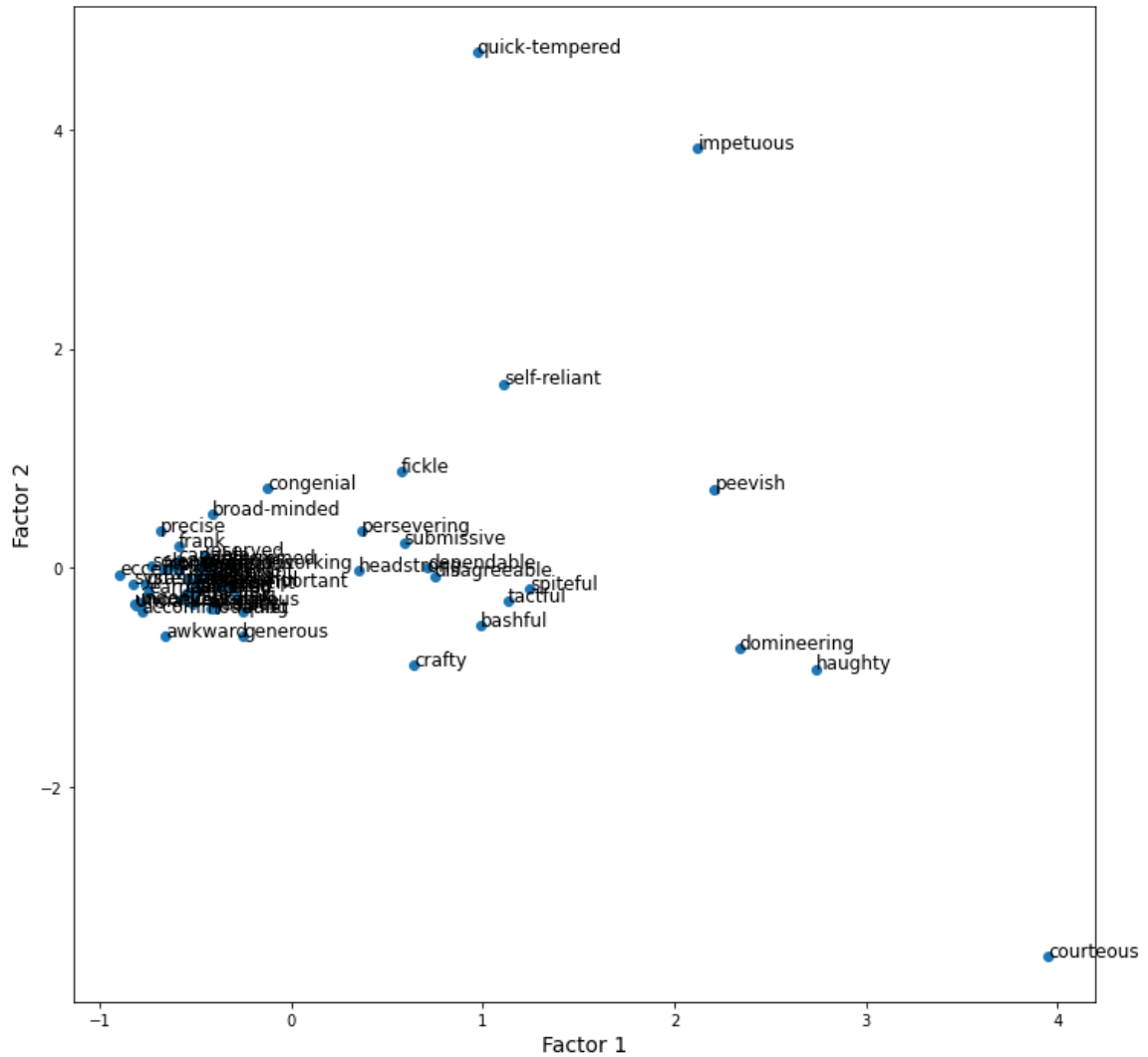
viable (Pennington et al., 2014; Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019). For general word vectors produced by LSA 100-300 dimensions are sufficient for most tasks Pennington et al. (2014). Due to Thurstone’s tightly thematic word and document set he found just five useful dimensions.

As previous studies were factorizing word vectors, mapping personality can be conceptualized as a challenge to create word vectors where personality information is salient. In this Chapter words are embedded using RoBERTa and the multilingual variant XLM-R Liu et al. (2019); Conneau et al. (2019). Two factors much like Digman’s α and β emerge even when varying embedding context, word list, factorization method, descriptive phrases vs adjectives, and English vs Spanish.

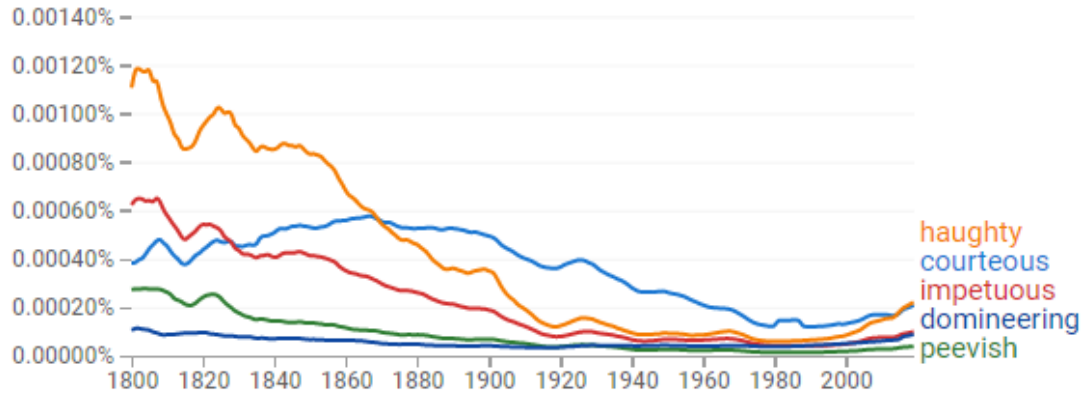
5.1.1 Embedding Context

One advantage of transformer models is the ability to condition word representations based on context. To make it clear the adjective of interest describes a person the sentence “I am a WORD person.” is used, where “WORD” is each of Thurstone’s words. This preserves the flavor of psychological surveys. Only the vector for the personality word is used, the rest of the sentence embeddings are discarded. If a word is tokenized into multiple tokens (eg. *textitbroad-minded*) the average is taken. RoBERTa embeddings are implemented using Facebook AI’s fairseq library (Ott et al., 2019). Factor analysis is done using the FactorAnalysis module in scikit-learn (Pedregosa et al., 2011).

In Figure 5.1 adjectives are plotted on the resulting first two factors. *Quick-tempered*, *impetuous*, *peevish*, *domineering*, *haughty* and *courteous* lay outside the main word cloud. There is no discernable personality factor structure. When a few samples are far away from the rest they come to dominate the resulting factors. It’s unclear what unites these outliers. One possibility is that they are all rare and old fashioned. Figure 5.2 shows how often these words appeared in books from 1800-2019. They were all most popular over a century ago,

Figure 5.1: Factor Analysis of Thurstone Words

RoBERTa embedding of each word token with context of “I am a WORD person”. Factors are dominated by word frequency.

Figure 5.2: Google Ngrams of Outlier Words

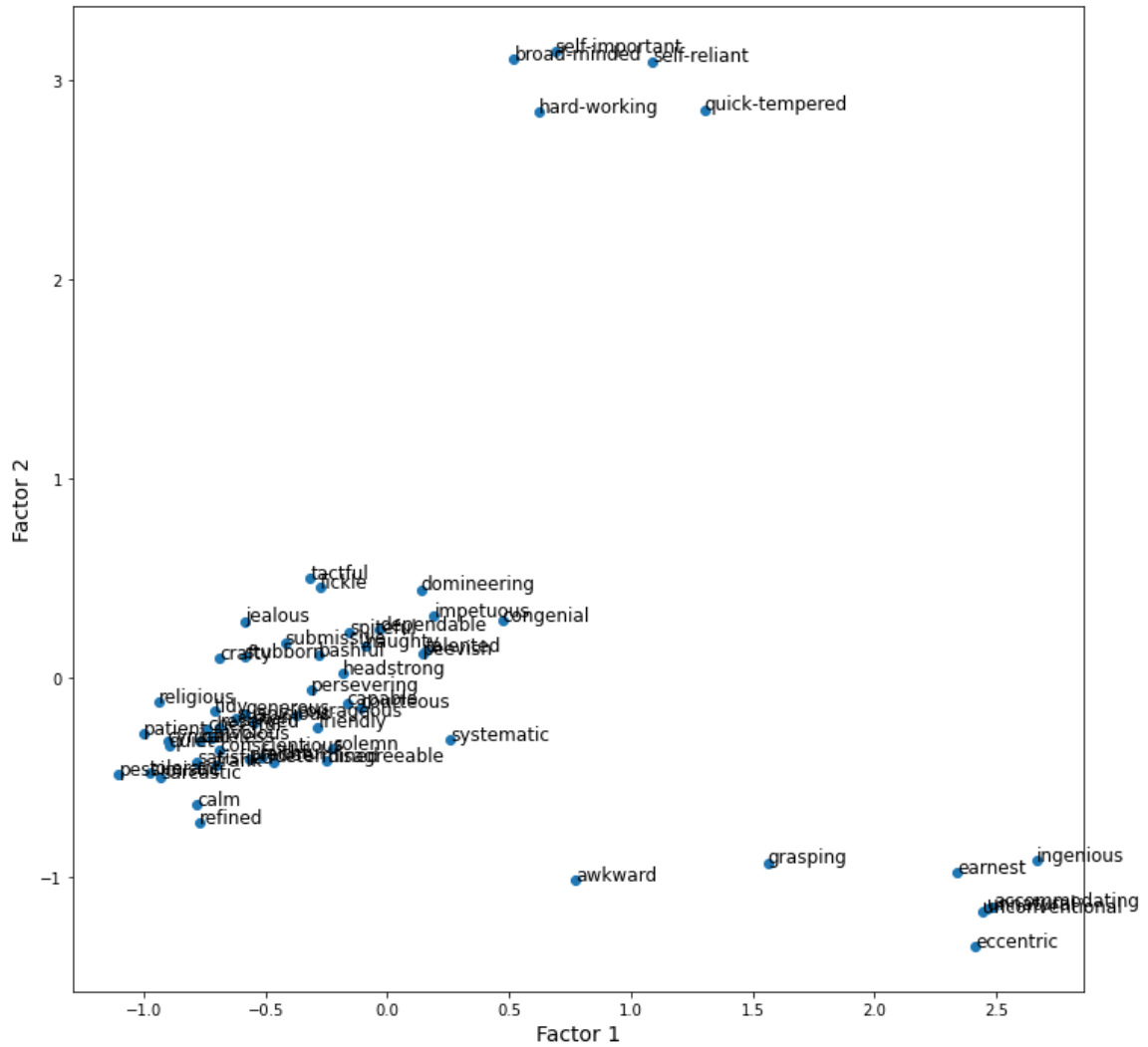
Outlier words in Figure 5.1 are rare and old fashioned.

dated even when Thurstone selected them. As RoBERTa is trained on books the model would know that these words signal old text (or affected speech). Whether or not that is the driving limitation, personality information is not salient.

Another option is to embed the entire sentence as using the special `<cls>` token, as was done for Facebook Statuses in Chapter 3. The `<cls>` token is appended to the beginning of each sentence and lets the model carry information relevant to the entire sentence. Because the only part of these sentences that will change is the adjective, the hope is that factorization of the whole sentence meaning will show personality structure. In Figure 5.3 we can see this is not the case. The structure is dominated by compound words positioned far from everything else. These could be removed, but still the problem remains that the `<cls>` embedding does not bring personality to the fore.

Because RoBERTa is trained on a language modeling task it has the ability to fill in masked tokens in a sentence. A sentence can be constructed that loads personality meaning (and little else) from the Thurstone word onto the masked word. The straightforward sentence “My personality can be described as `<mask>` and WORD.” was selected. Figure 5.4 shows that personality structure emerged with this approach. Factor one represents

Figure 5-3: Factor Analysis of Thurstone Words



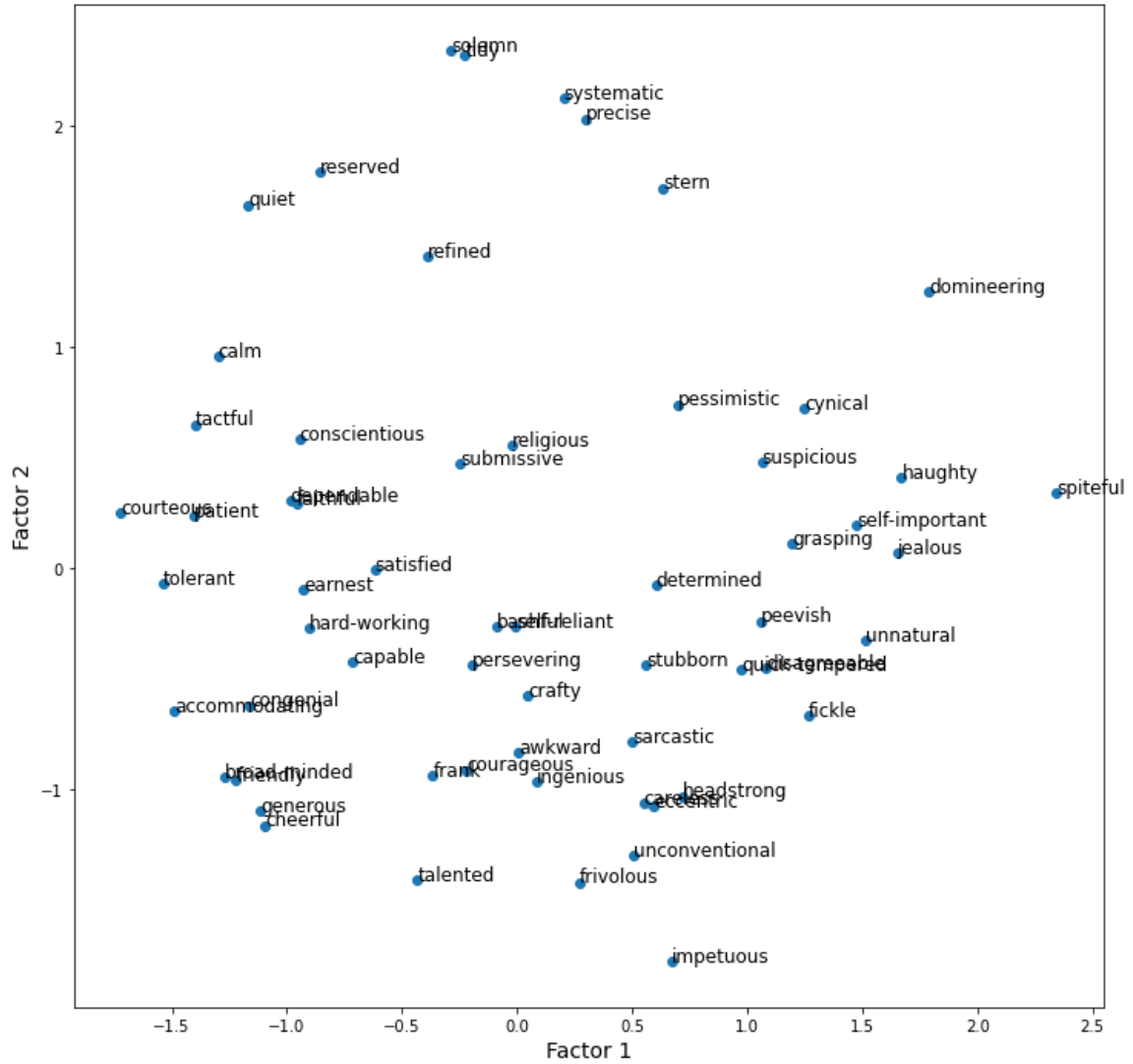
RoBERTa embedding of each <cls> token with context of “I am a WORD person”. Factors separate compound words. No interpretable personality structure.

Table 5.1: Core Merriam-Webster Personality Words

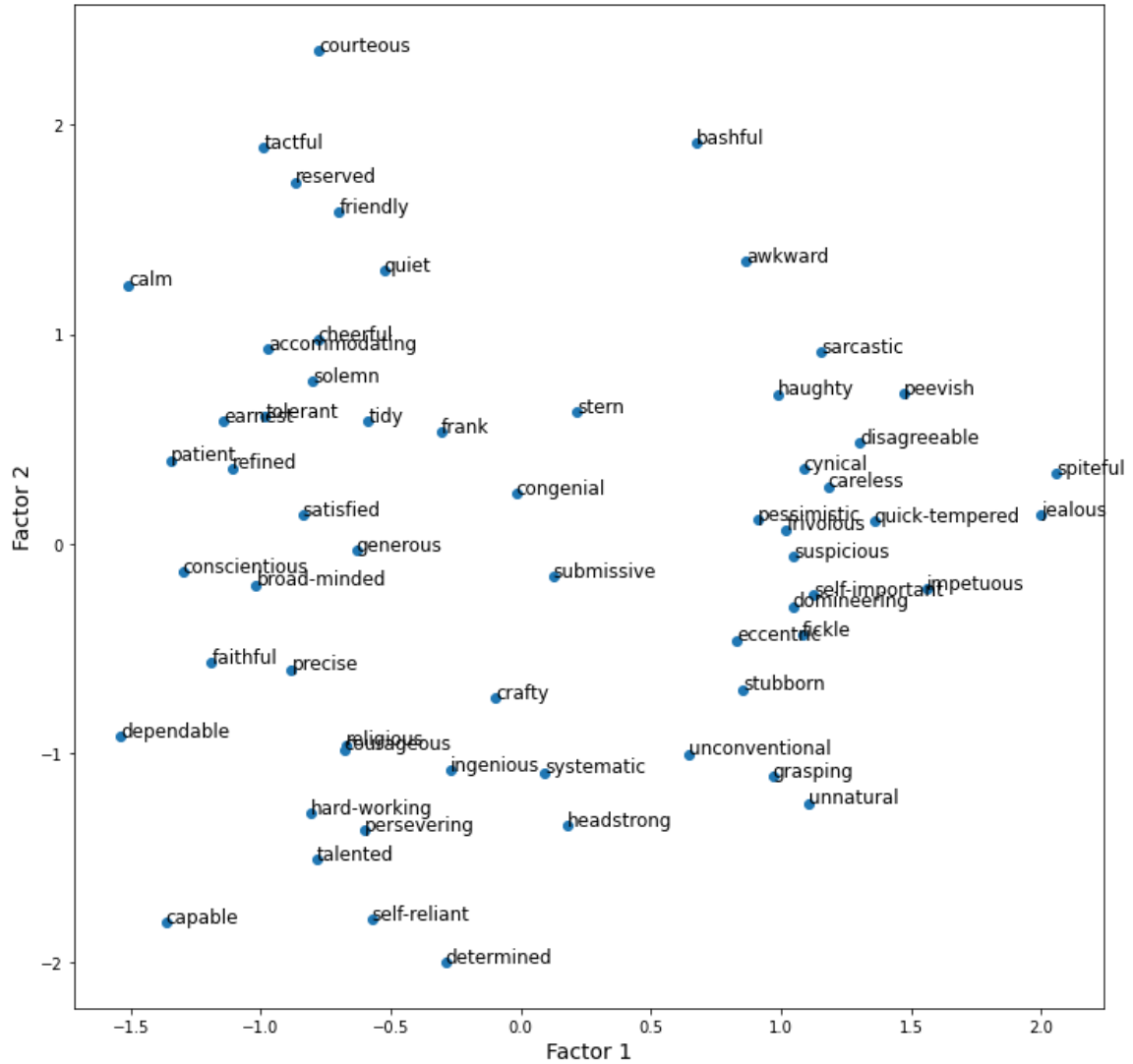
abusive	conservative	funny	modest	secretive
active	courageous	generous	moody	self-centered
adventurous	cowardly	gentle	nervous	selfish
affectionate	creative	greedy	nice	sensible
aggressive	cruel	gregarious	obsessive	sensitive
ambitious	cynical	gullible	optimistic	serious
annoying	decisive	happy	outgoing	shy
anxious	determined	honest	patient	sincere
artistic	direct	imaginative	persistent	sociable
bossy	domineering	impatient	pessimistic	stubborn
brave	easygoing	impulsive	pompous	superficial
calm	emotional	independent	practical	tactful
cautious	enthusiastic	intelligent	rational	tactless
charming	extroverted	introverted	reliable	thoughtful
cheerful	fearful	lazy	reserved	witty
compulsive	frank	loyal	ruthless	
confident	friendly	mean	sarcastic	

socialization, Digman's α . *Courteous, tolerant, accommodating, patient* and *tactful* appear opposite of *spiteful, domineering, haughty, jealous* and *self-important*. Factor two, β , represents self-actualization. *Solemn, tidy, systematic, precise, reserved* and *stern* stand in contrast to *impetuous, frivolous, talented, unconventional, cheerful, eccentric, careless* and *headstrong*.

"My personality can be described as <mask> and WORD" produced a rich personality factor structure. How robust is that to different sentence choices? To answer that, the same words are embedded with the sentence "<mask> is another word for someone who is WORD.". Once again, the first factor loads on socialization: *spiteful, jealous, peevish* and *impetuous* vs. *calm, dependable, capable, conscientious* and *patient*. The second factor loads on self-actualization: *courteous, tactful, bashful* and *reserved* vs. *determined, self-reliant, capable, talented, persevering* and *headstrong*. The results for "Those close to me say I am <mask> and WORD" can be seen Fig 6-4. The first factor remains the same, although more polarized. The second is less ordered.

Figure 5-4: Factor Analysis of Thurstone Words

RoBERTa embedding of each <mask> token with context of “My personality can be described as <mask> and WORD”.

Figure 5-5: Factor Analysis of Thurstone Words

RoBERTa embedding of each <mask> token with context of “<mask> is another word for someone who is WORD”.

Table 5.2: ESL Personality Words

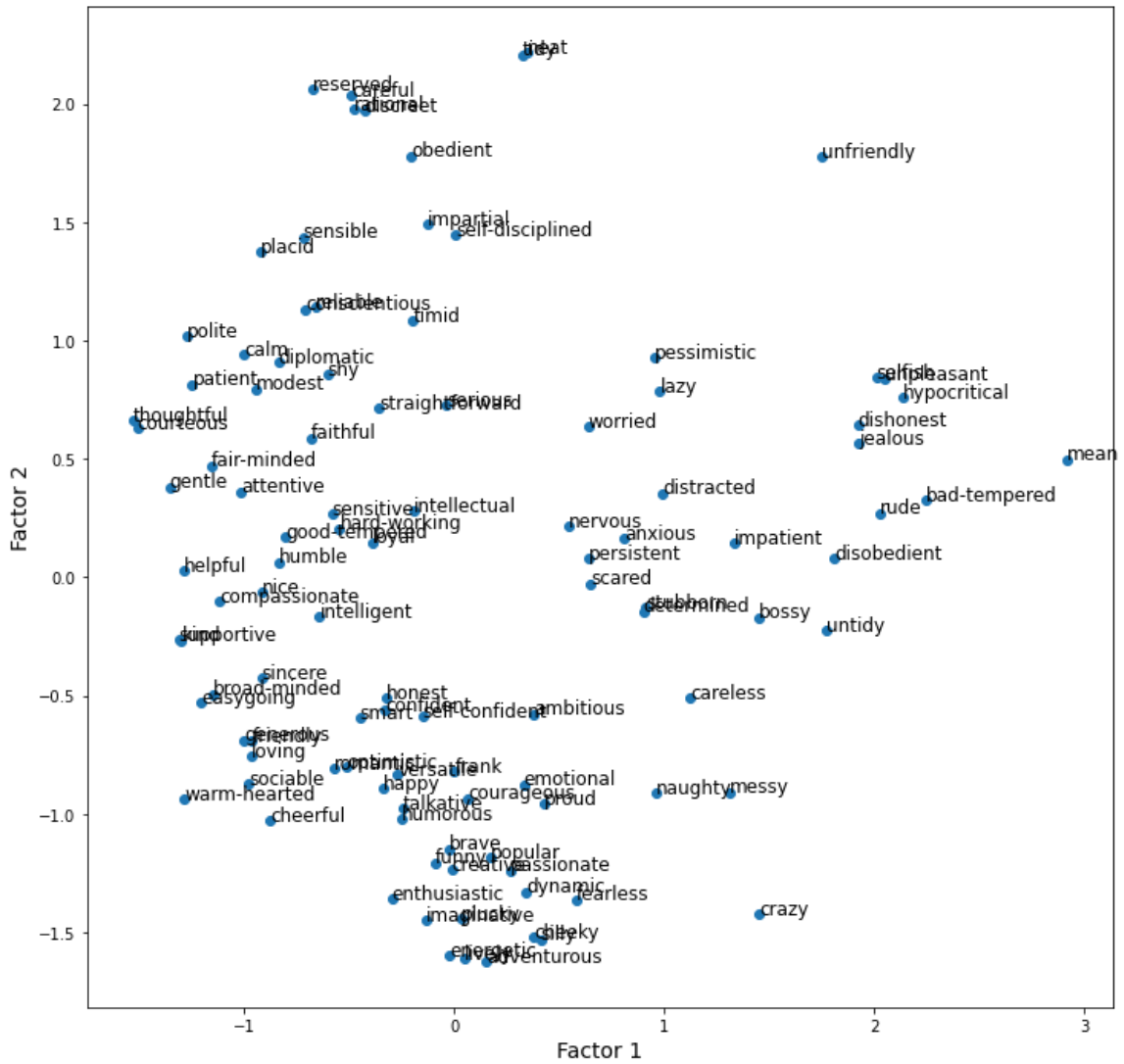
anxious	unpleasant	fearless	impatient	gentle
naughty	talkative	unfriendly	easygoing	neat
stubborn	calm	generous	careless	dynamic
sensitive	passionate	compassionate	messy	fair-minded
intelligent	proud	warm-hearted	hard-working	impartial
nice	sincere	disobedient	creative	supportive
emotional	lazy	straightforward	broad-minded	timid
bad-tempered	lively	selfish	faithful	intellectual
nervous	funny	imaginative	kind	brave
mean	silly	placid	courageous	ambitious
distracted	shy	jealous	loyal	polite
dishonest	determined	helpful	modest	happy
rude	versatile	enthusiastic	tidy	romantic
discreet	sociable	persistent	confident	diplomatic
crazy	worried	sensible	attentive	courteous
cheeky	thoughtful	rational	loving	humorous
cheerful	humble	reserved	reliable	self-disciplined
energetic	friendly	self-confident	scared	popular
untidy	frank	bossy	conscientious	smart
pessimistic	obedient	plucky	good-tempered	serious
optimistic	honest	patient	careful	hypocritical
				adventurous

5.1.2 Other Personality Words

Thurstone’s words were selected 100 years ago and language is always shifting. We proceed with an embedding using “My personality can be described as <mask> and WORD” on words from the Merriam-Webster list of core personality adjectives (Webster, 2014). These can be seen on Table 5.1. The resulting factors in Figure 5-6 are remarkably similar to the ones produced by Thurstone’s words. Factor one loads on socialization: *easygoing*, *thoughtful*, *friendly*, *gentle*, *sociable*, and *affectionate* vs. *abusive*, *mean*, *cruel*, *domineering*, *ruthless* and *self-centered*. Factor two loads on self-actualization: *gregarious*, *enthusiastic* *adventurous*, *impulsive* and *outgoing* vs. *reserved*, *rational*, *cautious*, *introverted*, *sensible* and *conservative*.

Another way to establish core personality words are those that one needs to know when learning a new language. Here we perform factor analysis on a list of 106 introductory words (Table 5.2 from a website that teaches English (Seven Steps to Learn English, 2020). The results in Figure 5-7 show the familiar factors (signed opposite from α and β). It’s

Figure 5.7: Factor Analysis of ESL Words



RoBERTa embedding of each <mask> token with context of “My personality can be described as <mask> and WORD”. Words from an online intro to English guide (Seven Steps to Learn English, 2020). Similar factors to the studies using Thurstone and Merriam-Webster words.

interesting that antonyms don't appear opposite one another. For example, *obedient* is neutral on α , and negatively loads on β . *Disobedient*, on the other hand, loads highly on α and is neutral on β .

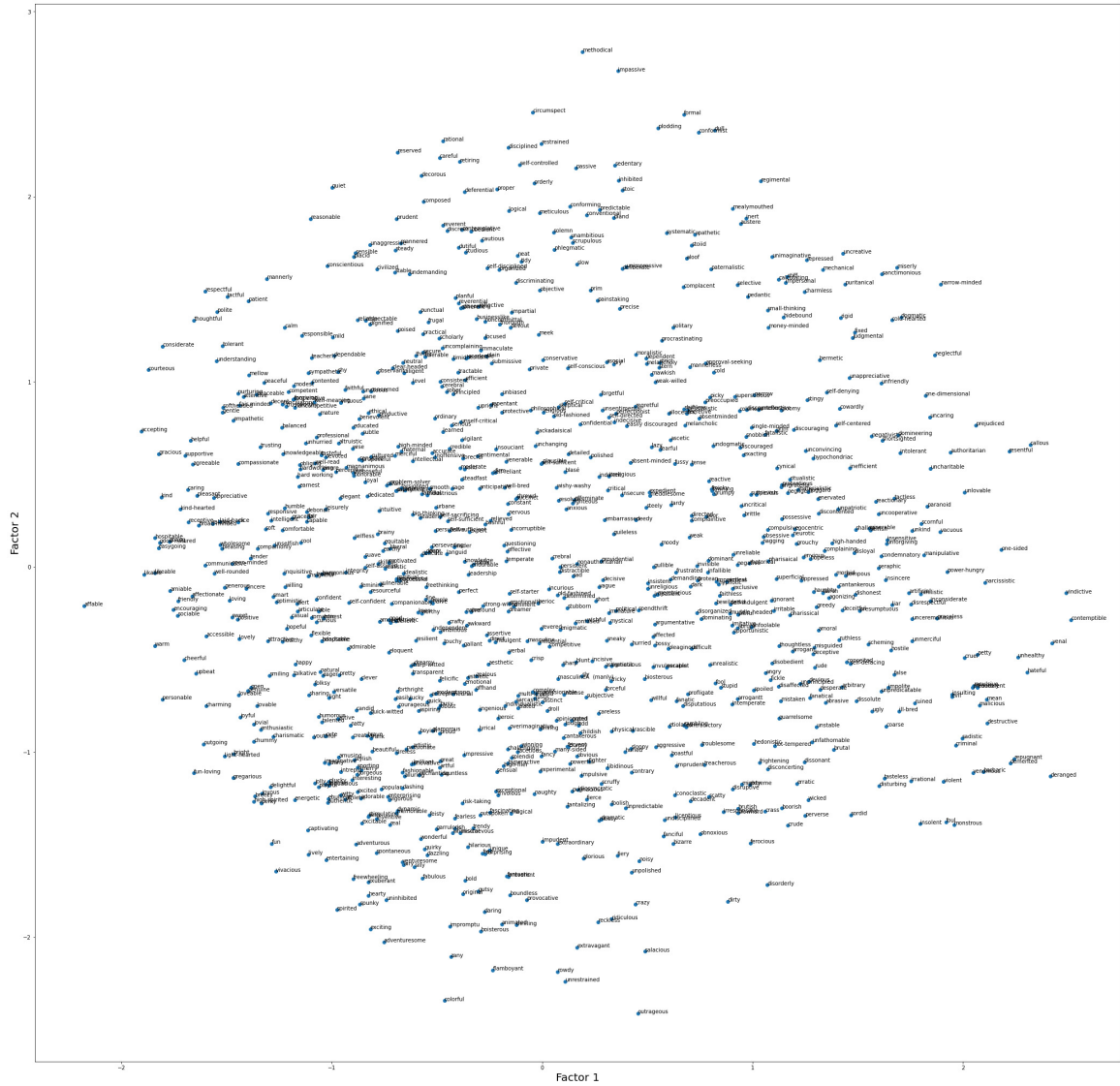
We can extend the method beyond basic adjectives. To do this, the union of two large vocabulary lists are used. The first list was developed to aide authors in their character descriptions (Worsley, 2020). The second was compiled by an eccentric interested in the combinatorics of language (Gunkel, 2013). Separately they contain 800 and 638 words respectively; their union contains 1,005.

In Figure 5-8 α and β come into clearer relief. On factor one *affable, easygoing, appreciative, tolerant, genuine, gracious* and *polite* vs. *contemptible, vindictive, deranged, narcissistic, callous, prejudiced*. On the second factor *outrageous, animated, boisterous, zany, salacious, captivating, insolent* and *exuberant* vs. *methodical, inhibited, conformist, aloof, formal, circumspect* and *restrained*.

Figure 5-9 shows there is also a third factor that loads on willfulness: *steely, competitive, pugnacious, strong-willed, businesslike* and *thrifty* vs. *mealymouthed, mawkish, languid, contemplative, dainty, whimsical, childish, indolent* and *self-denying*. It's interesting that *thrifty* and *self-denying* are on opposite ends of this spectrum. It could be that *thrifty* implies saving for something else whereas denying oneself is simply about not fulfilling desires; a failure to impose one's will on the world.

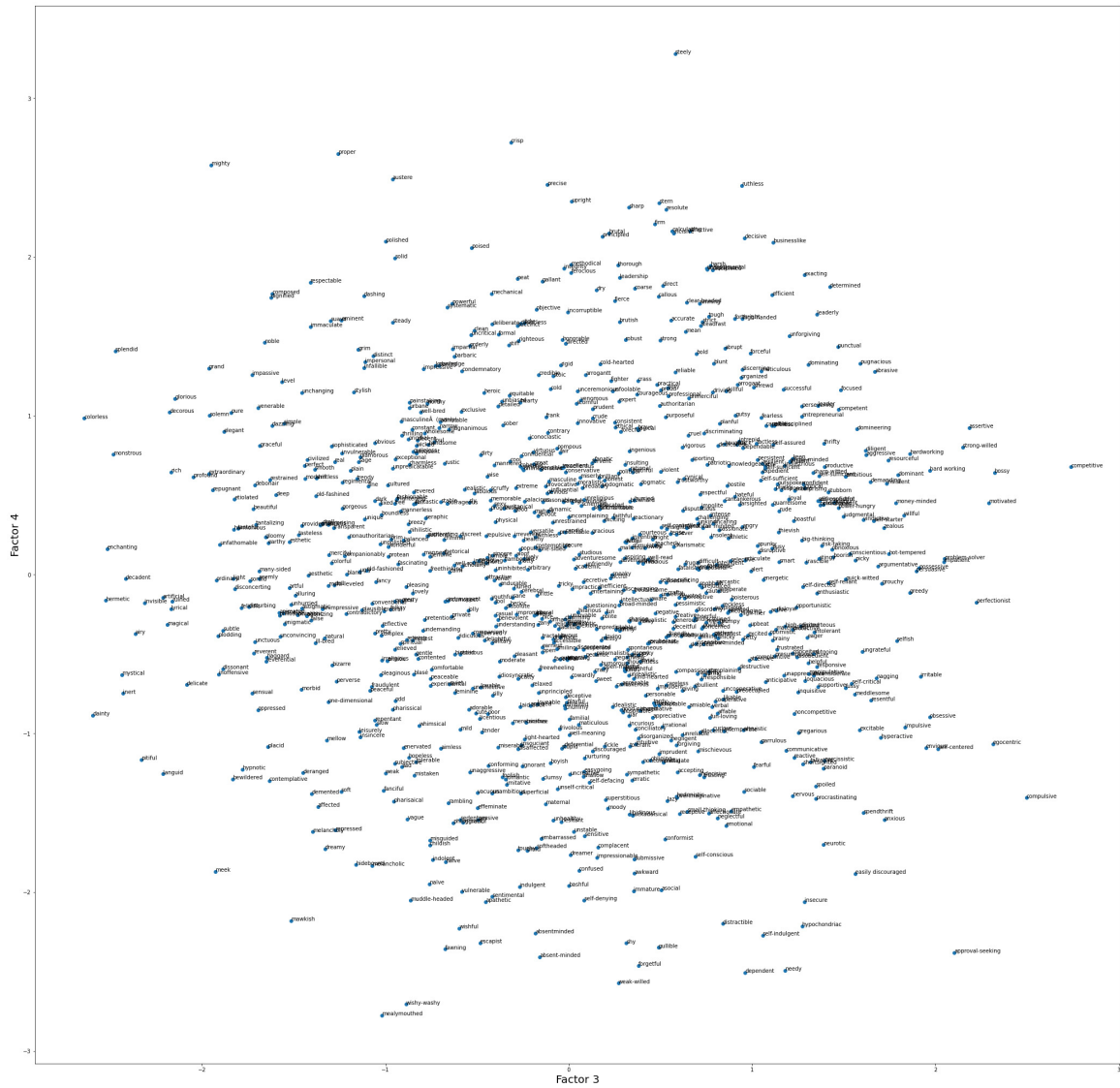
This is the first set of words in which a third factor appears. This complicates the narrative as it includes some aspects of self-actualization. However, note that any rotation of the factors is also a valid model. Combining factor 2 with $-\frac{1}{2}$ factor 3 and $\frac{1}{2}$ factor 4 produces Figure 5-10. This factor is an even stronger candidate for self-actualization. It loads on *mealymouthed, plodding, inert, hidebound, repressed, placid, reverent, retiring,* and *meek* vs. *bold, gutsy, spirited, enterprising, ruthless, crass, winning, insolent* and *incisive*.

Figure 5-8: Factor Analysis of 1,005 Words



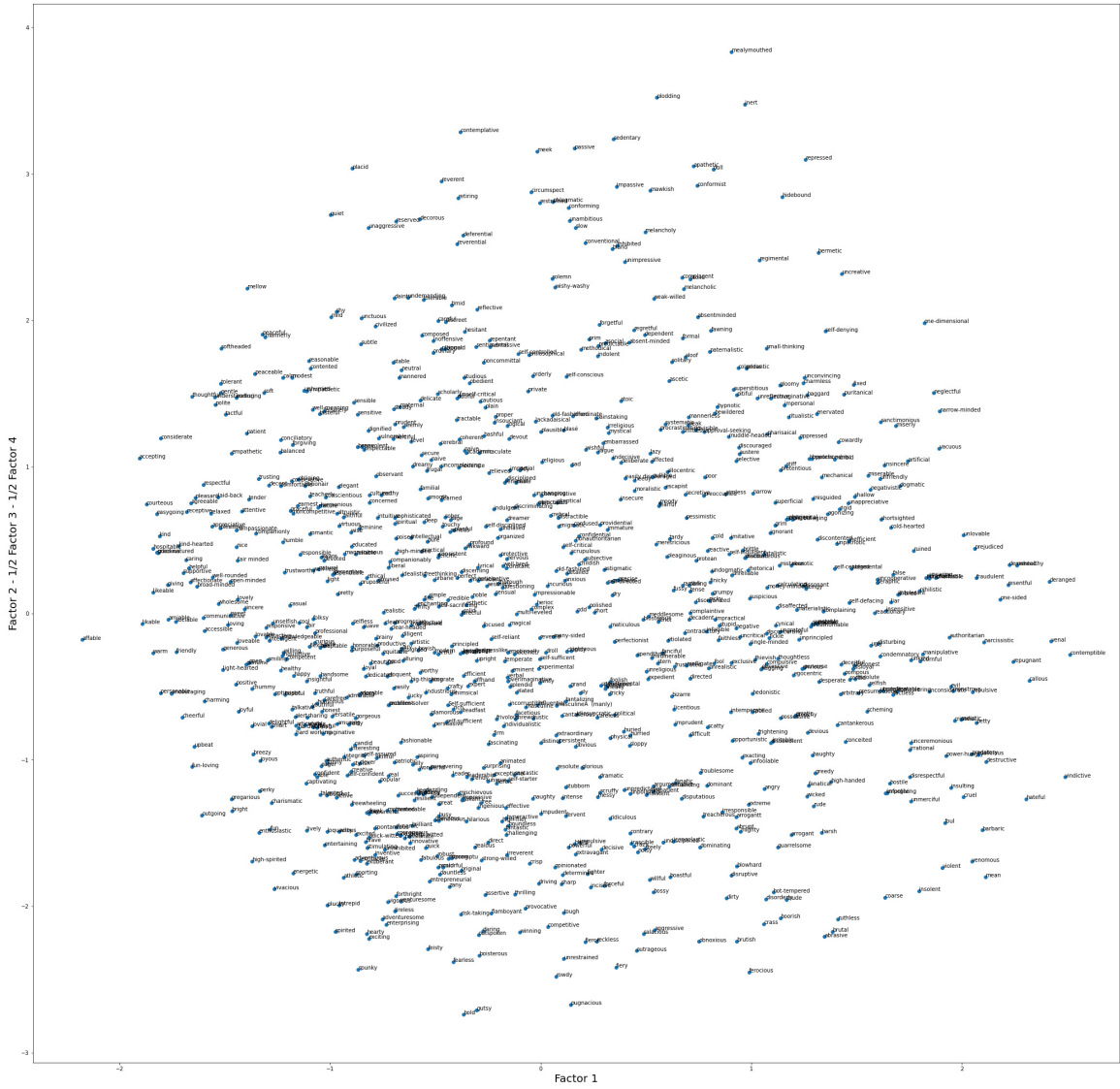
1,005 RoBERTa embeddings of “My personality can be described as <mask> and WORD”. Factor 1 loads on socialization. The second loads on self-actualization. Zoom in to view words.

Figure 5.9: Factor Analysis of 1,005 Words



1,005 RoBERTa embeddings of “My personality can be described as <mask> and WORD”. Personality structure extends to at least a third factor. Diagonal with positive slope loads on strength of will. Zoom in to view words.

Figure 5-10: Factor Analysis of 1,005 Words



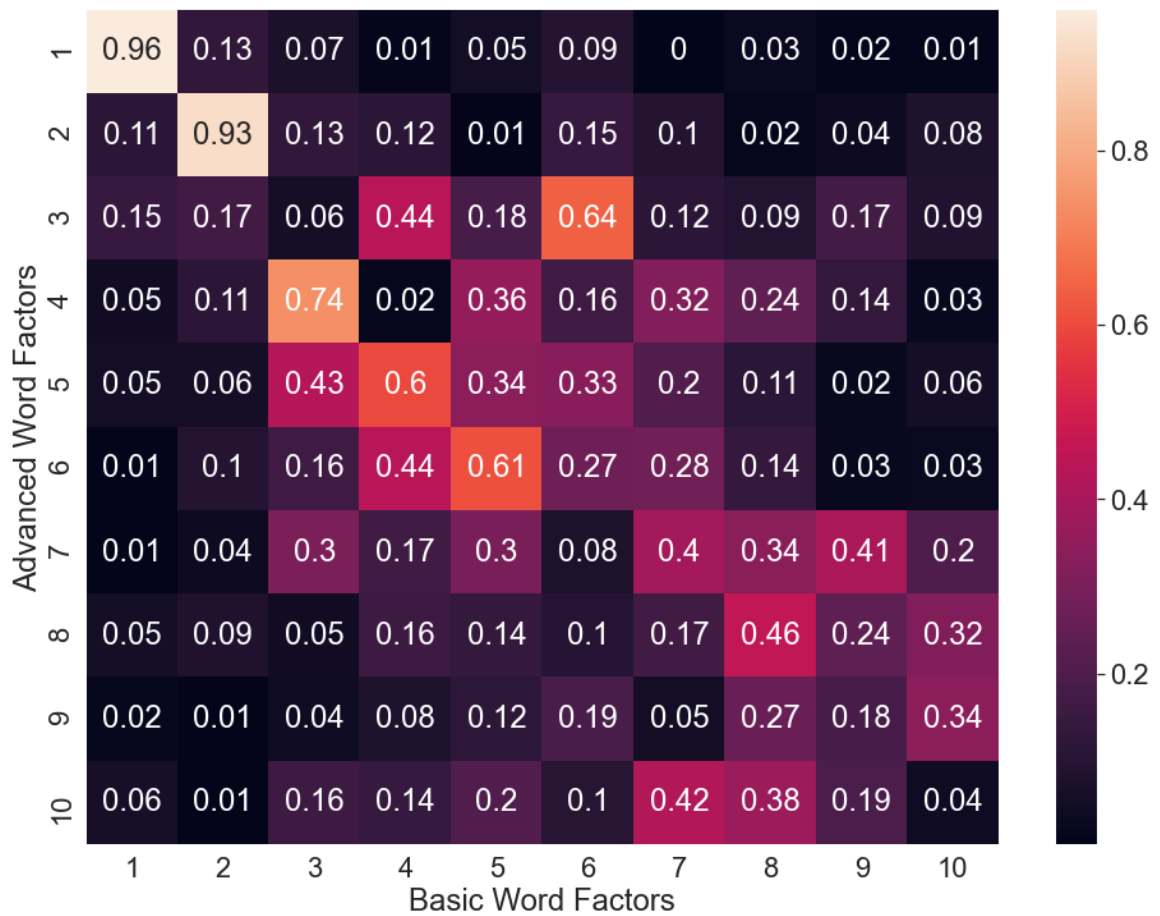
1,005 RoBERTa embeddings of “My personality can be described as <mask> and WORD”. Third (diagonal) factor combined with the second to better represent self actualization. Zoom in to view words.

It is a qualitative debate as to whether the combined factors better capture self actualization. However, no more than three factors are needed to capture the readily interpretable dimensions. The fifth and sixth factors can be seen in Figure 6-5.

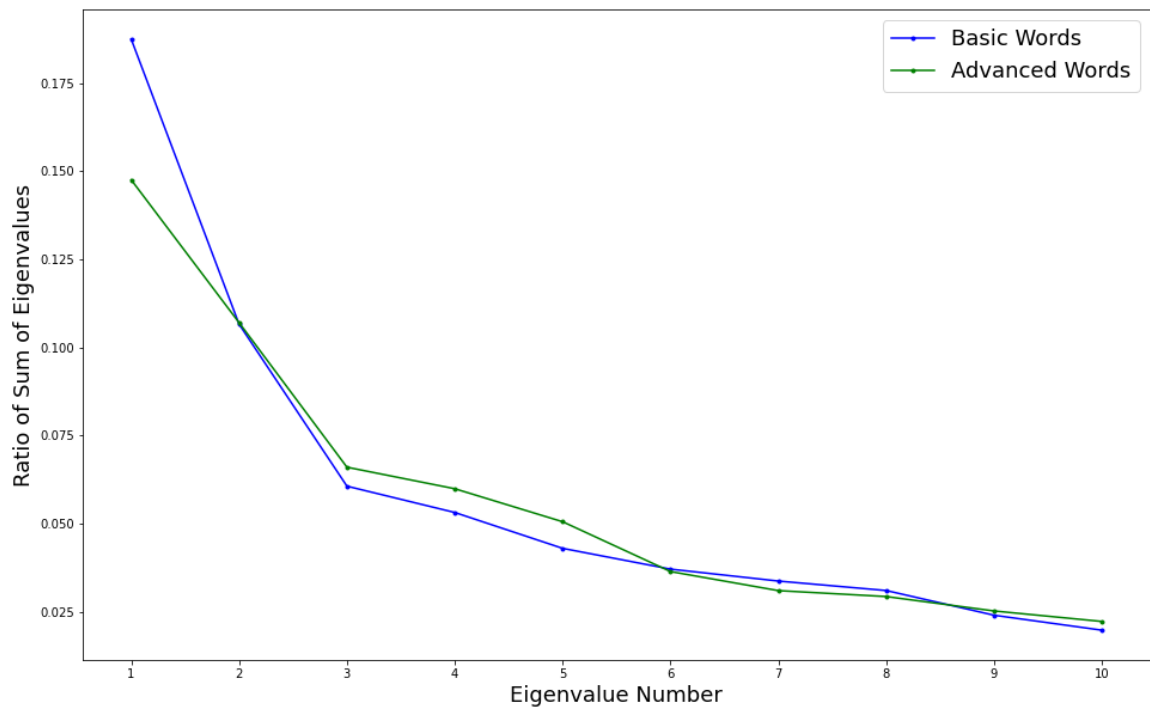
Each factor is a vector in the 1024 space defined by RoBERTa's embedding. Therefore, one can quantitatively compare the resulting vectors in each experiment. Two sets of words are constructed. Basic words are the union of Merriam-Webster and ESL words ($n = 143$). Advanced words are the 1,005 words described above, excluding any basic words. This comes to 881 words. Factors analysis is performed with ten total factors in each experiment. Figure 5-11 displays the pairwise Pearson correlations. Despite factorizing completely different sets of adjectives, the first two factors are correlated at 0.96 and 0.93 respectively. Near the diagonal there are moderately sized correlations of up to 0.74. Similar information is being grouped, but on different factors and sometimes spread over several. Figure 5-12 shows the percentage of variance explained by each factor. The first two explain 29.4%, as much as the next 8 combined.

5.1.3 Embedding IPIP Questions

The <mask> token can also take the value of a phrase, such as those in the mini-IPIP (Donnellan et al., 2006). Questions are embedded with “My personality can be described as <mask>, or in other words, [IPIP item].” where IPIP items include “I like order” and “am the life of the party”. These are projected onto the factors found with 1,005 adjectives using the phrase “My personality can be described as <mask> and WORD”. Figure 5-13 shows that the twenty questions map to the expected areas. Socialization loads on being relaxed and sympathizing with others vs getting upset and making a mess of things. Self-actualization loads on engaging at parties and having an imagination vs liking order and being withdrawn. This indicates sentences can be amended for different types of descriptions and still projected to the same space.

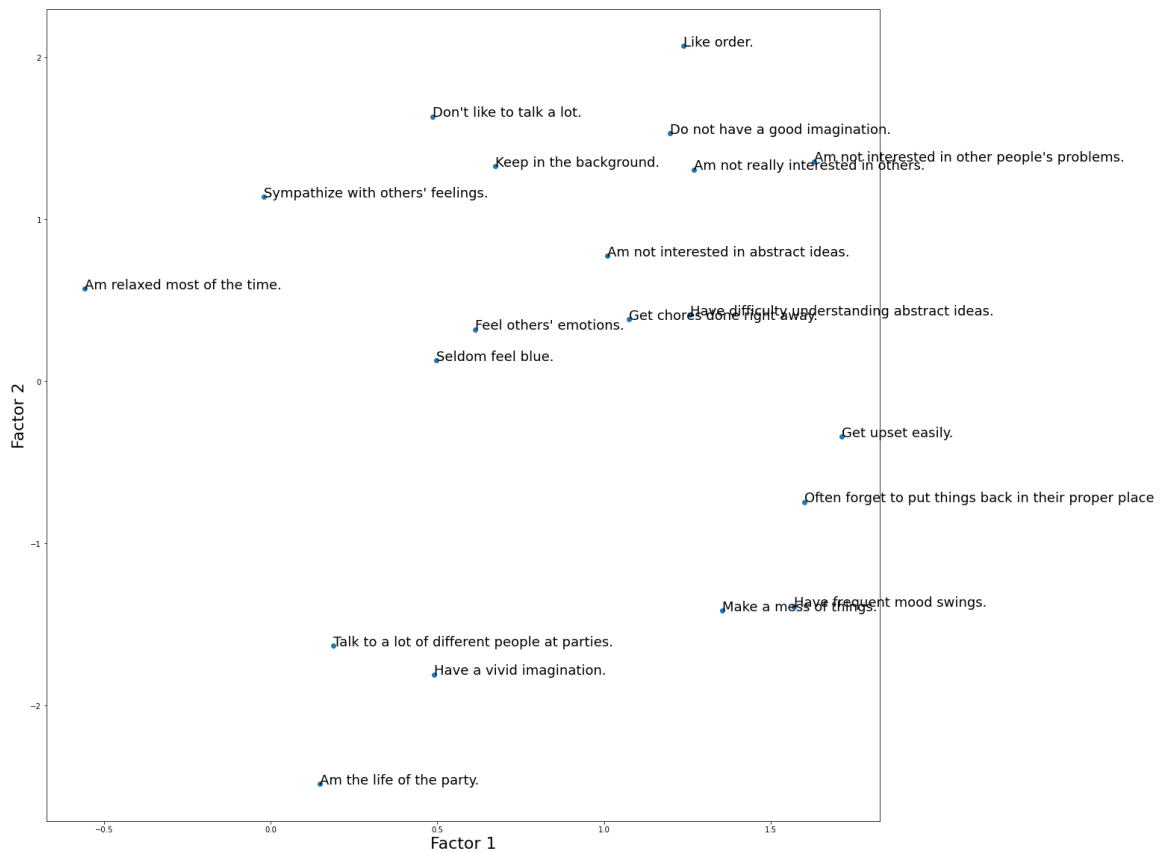
Figure 5-11: Absolute Value of Pairwise Pearson's Correlations

Correlation of axes produced via factor analysis of advanced and basic word sets. The first two factors are highly correlated in each experiment despite there being no overlap between the word sets.

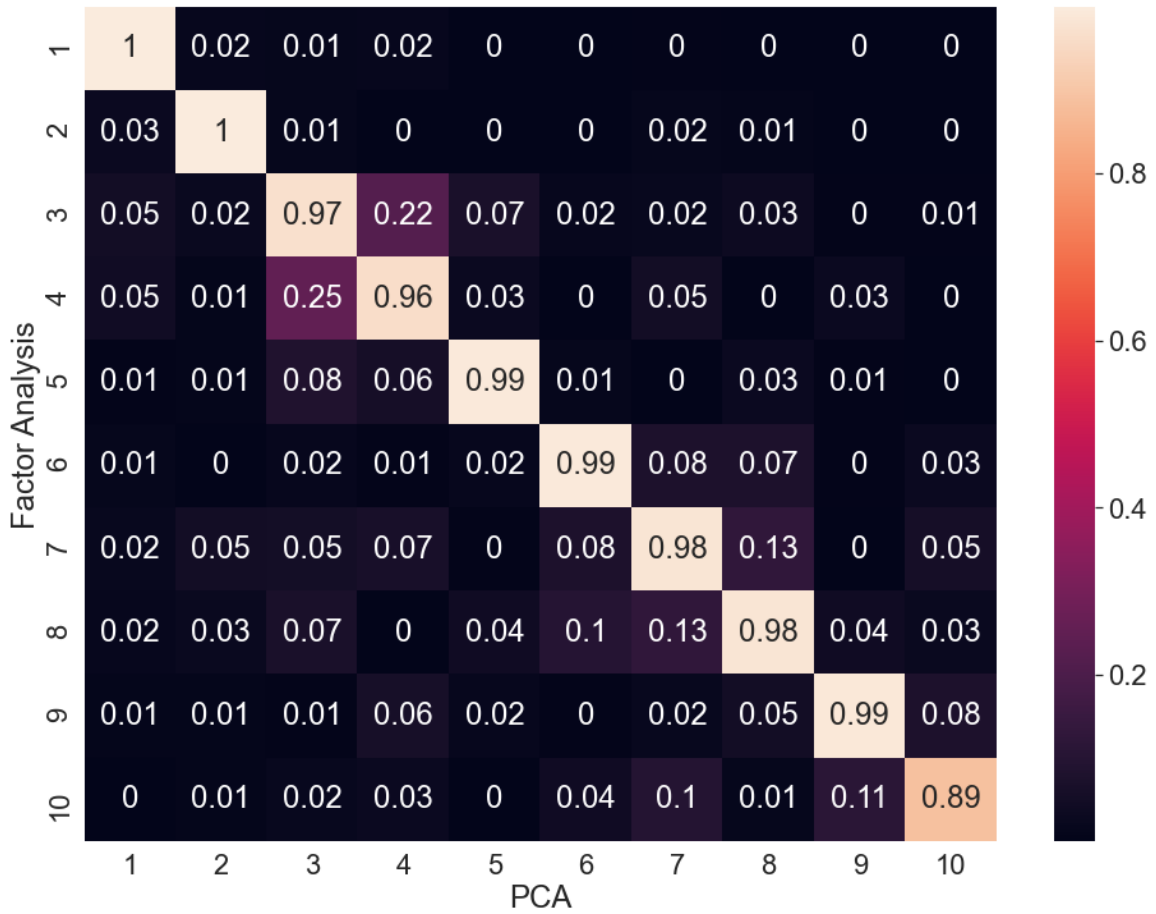
Figure 5-12: Eigenvalues of RoBERTa Embedding

29.4% of total variance explained by first two eigenvalues. As in other studies, values fall off quickly indicating few factors are needed to represent personality (Ashton et al., 2004; Thurstone, 1934).

Figure 5.13: Mini-IPIP Questions Mapped to Lexical Factors



IPIP questions are embedded using RoBERTa and the sentence “My personality can be described as <mask>, or in other words, [IPIP item].” They are then projected down to the previously solved latent space displayed in Figure 5.8.

Figure 5-14: PCA vs Factor Analysis

Absolute value of pairwise Pearson's correlations. Results approximate an identity matrix; factorization method not important in these settings.

5.1.4 Factorization Choices

When Ashton, Lee and Goldberg explored the lexical hypothesis their data were word vectors as well; each dimension defined by one of 310 students (Ashton et al., 2004). With $c = 2$ they found a loose resemblance to α and β . Adding $c = 3$ through 7 that structure collapsed. From the hierarchical schematic on page 716 of their paper one can see that adding additional factors splits and combines factors from the $c - 1$ solution. This indicates the factor analysis solution is far from PCA. In PCA additional orthogonal eigenvectors are

added without disrupting the previous factors.

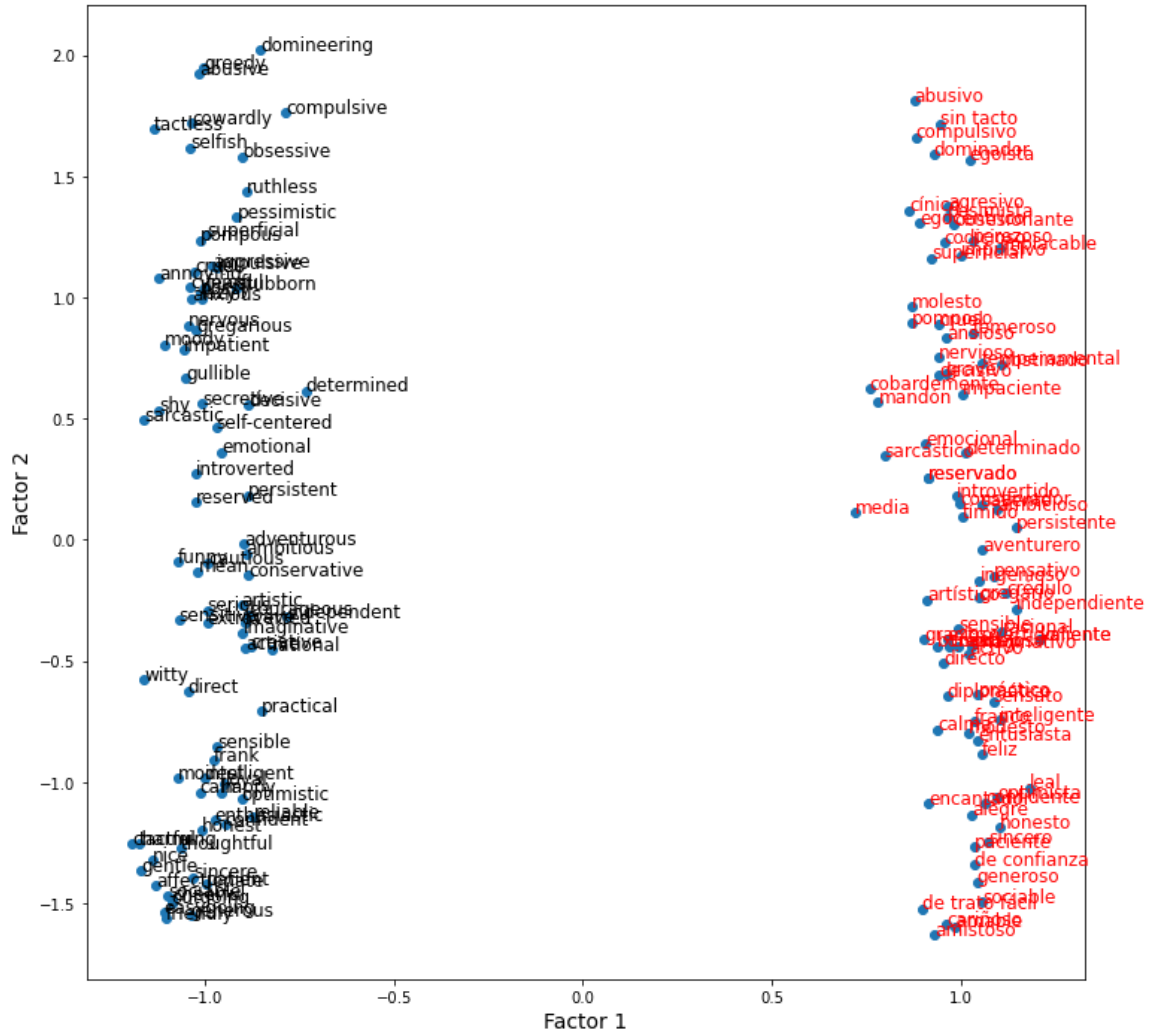
Figure 5.14 shows the correlation of axes found via factor analysis and PCA on the basic word set. The two solutions are very similar; no off-diagonal element is larger than 0.25. Comparing this to Ashton's results, a factor model farther from the PCA solution by necessity includes more item-level variance—more student responses not well explained by the common factor structure. It's not clear if this property of the data makes for worse factors, but it is a substantial difference.

5.1.5 Multilingual Embedding

One problem that has plagued psychology is their available samples, oftentimes a few hundred undergraduate students or soldiers in a psychiatric hospital (Ashton et al., 2004; Eysenck, 1944). It is difficult to make general theories of personality when participants are so singularly Western, Educated, Industrialized, Rich, and Democratic—or WEIRD (Henrich et al., 2010). The problem is even deeper when studying the lexical hypothesis as a single language may embed a biased version of more general human nature. For example, the debate on the number of factors in Dutch vs English samples still continues (De Raad and Barelds, 2008).

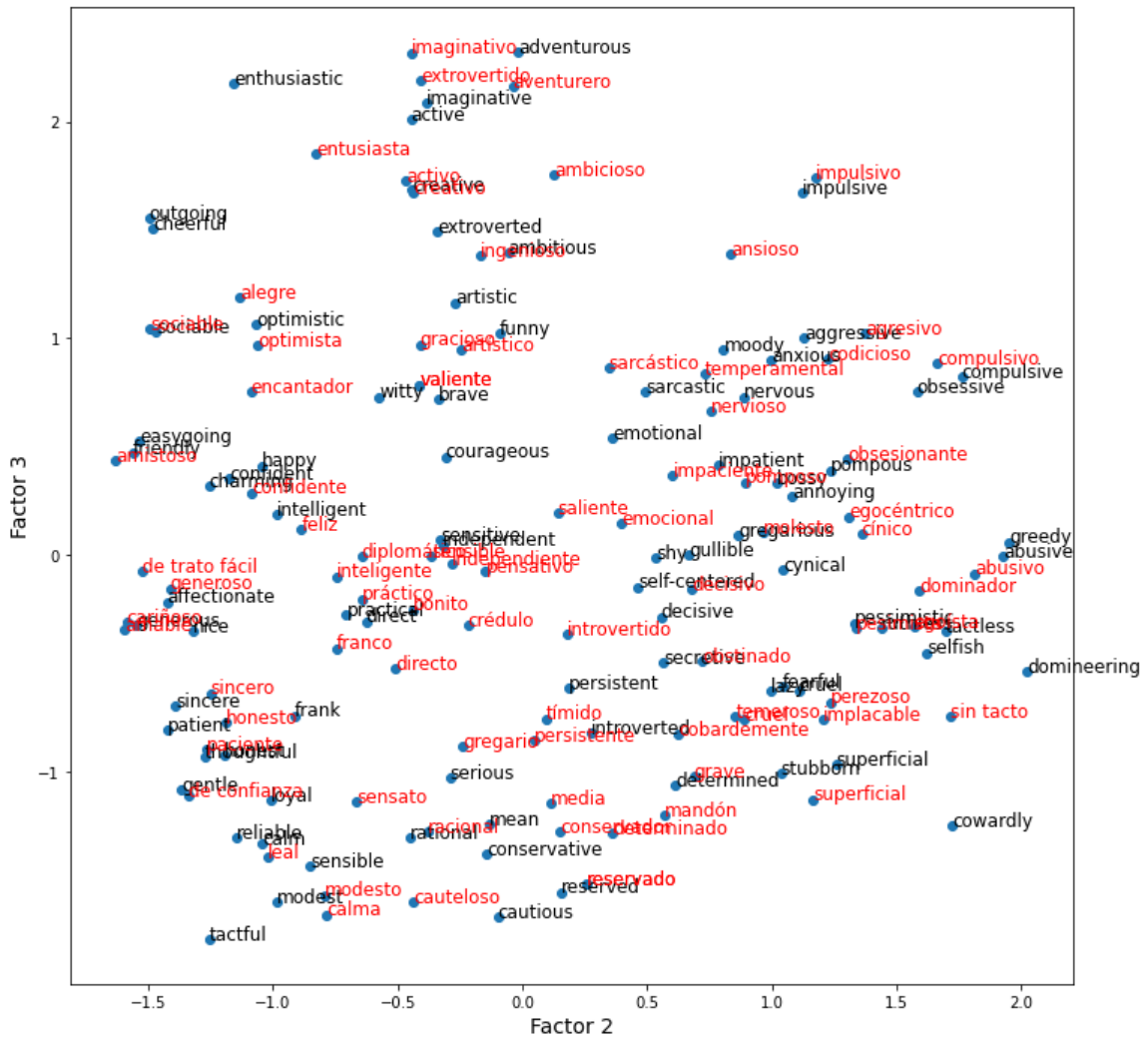
Language models expand sample size by including text written by millions of people from all walks of life (though connected to the internet and inclined to write there). XLM-R is a 550 million parameter transformer based language model that was trained on over two terabytes of internet text spanning 100 languages (Conneau et al., 2019). This is an order of magnitude more text than RoBERTa saw. Training a model to fill in masked words in so many languages forces each language to compete for space in the model. This encourages the model to share representations between languages. When trained a few languages this will sometimes increase performance on all languages, but especially languages with fewer training samples. Certain language structures and ideas are generalized; sharing parameters

Figure 5-15: XLMr Embedding in English and Spanish



XLMr embedding of each <mask> token with context of “My personality can be described as <mask> and WORD” for English adjectives and “Mi personalidad se puede describir como <mask> y WORD” for Spanish adjectives. Words from Webster’s Core Personality adjectives and translated using Google Translate.

Figure 5-16: XLMr Embedding in English and Spanish



XLMr embedding of each <mask> token with context of “My personality can be described as <mask> and WORD” for English adjectives and “Mi personalidad se puede describir como <mask> y WORD” for Spanish adjectives. Words from Webster’s Core Personality adjectives and translated using Google Translate.

acts as a regularizer. However, with more than a few languages competition for space becomes the driving force and performance for each language decreases compared to a monolingual model baseline. Such a multilingual model is perfect for analysing the lexical hypothesis as it is concerned with global personality structure.

Websters core word list is translated into spanish via google translate. English words are embedded using “My personality can be described as <mask> and WORD”. Spanish words are embedded using “Mi personalidad se puede describir como <mask> y WORD”. The first two factors plotted on Figure 5-15 show the English and Spanish words are mapped to different areas. This separation is captured almost completely by the first Factor. Plotting factors two and three in Figure 5-16 places the English and Spanish word clouds in the same neighborhood. A familiar structure emerges. Factor two loads on socialization: *domineering, greedy, compulsive, cowardly, and abusive* vs *friendly, easygoing, sociable, cheerful, and patient*. Factor three loads on self-actualization: *adventurous, enthusiastic, imaginative, active, and creative* vs. *tactful, cautious, reserved, modest, conservative, and cowardly*.

Many word pairs are one another’s closest neighbors. For example, *impulsive* and *impulsivo* and *reserved* and *reservado* appear almost on top of one another. Others are a bit removed from one another such as *tactful* and *diplo^mático*. But for the most part the latent factors align the two word clouds.

5.2 Discussion and Conclusion

For decades psychologists have interrogated the personality structure embedded in language by vectorizing personality words via surveys. In Goldberg’s excellent review of the lexical hypothesis he remarks on a study that used “three large samples” of 583, 521, and 324 (Goldberg, 1993). Instead, this work vectorizes words using two pre-trained language models: RoBERTa and XLM-R. These language models generalize textual information

from terabytes of text written by millions of people in a hundred languages. Given that so many studies of language have produced structure similar to the Big Five, it is surprising that just two factors consistently emerged in this work.

One can't make strong claims about the proper descriptive framework for personality from one result. Further, the author is not well suited to work this into the larger debate about the hierarchy of traits. However, it is worth considering that a search for a descriptive framework for personality yielded a previously described explanatory framework, α and β . This demonstrates that the method finds reasonable structure, which is sufficient to use this as a psychometric tool. As such, it has several advantages.

- **Multilingual.** XLM-R is trained on 100 languages. Future language models may represent even more languages. This work shows that two factors emerge with English and Spanish words. More distant pairs like Hindi and Chinese would be more interesting and are less studied in the psychometric literature.
- **Open Science.** The code to produce these results is available on at <https://github.com/andcut/DeepLexicalHypothesis>. If a mistake has been made, or another researcher would like to change one of these experiments, they can easily be replicated. Consider the question asked in “How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon” (Gurven et al., 2013). In that study researchers translated the 44 item Big Five Inventory into the spoken language of a tribe in the Amazon. The results supported just two factors. If a research disagrees with the translation of an item or the way data was collected there is little that can be done besides going to the Amazon to redo the experiment. A potential future work with our method could be “How universal is are the Big Two? Testing the two-factor model in Hindi and Mandarin speaking internet users”. The code to produce results would be online allowing fast iteration of ideas, error checks, and collaboration.

- Deep Learning. For other tasks requiring word vectorization, deep learning performs better than LSA. One would expect word vectors via RoBERTa to be more informative than those obtained by survey.

The debate about the structure of personality continues (Möttus et al., 2020). This provides a new tool in the psychometric toolkit that is free, flexible, and multilingual. In these experiments results are shown to be stable across many experimental choices: embedding context, language, phrases vs words, adjective lists, and factorization method.

Chapter 6

Conclusion

6.1 Summary

This work starts with a straightforward supervised learning problem. Given Facebook Statuses, how well can one predict 20 different user traits. Three language models are compared: LIWC, BoW, and RoBERTa. LIWC obtains an order of magnitude less EV than the other methods when predicting narcissism, SWL, BIS, and sensational interests. This is a considerable drawback considering it was designed to extract psychological information from text. The IPIP105 transfer learning model is a good replacement in predictive settings and is available on github. Instead of pooling statistical information into hand-crafted features, patterns from text in RoBERTa's training corpus as well as personality information from the myPersonality dataset are generalized in a 105 dimensional embedding. This embedding has the benefit of being interpretable as each dimension corresponds to one of 100 Big Five questions in IPIP, as well as Big Five scores. Where BoW and LIWC failed to explain more than 4% of the variance of SWL and BIS, IPIP105 achieved 19% and 25% respectively. This moves the predictions from mostly noise to a range where decisions about low cost interventions can be made.

On deeper inspection, predicting personality labels is not so straightforward. Chapters 3 and 4 treat labels assigned by survey as ground truth. The ideal model would perfectly pre-

dict personality scores from statuses. Yet administering the same questionnaire weeks later will produce different labels. It's not just that the labels are noisy; they are approximations of personality structure observed in language. Consider the task of mapping personality with just 20 questions. The Mini-IPIP chooses to spend four asking whether someone 1) is the life of the party 2) talks to a lot of different people at parties 3) doesn't like to talk a lot 4) keeps in the background (Donnellan et al., 2006). Given such a small budget and large landscape, these are remarkably similar. The most popular instruments accept harsh constraints to maintain fidelity to personality theory.

Thinking critically about the ground truth of these labels led to an interest in how they are anchored to reality. Curiously, RoBERTa can be used to predict personality scores in a supervised setting as well as define them in an unsupervised setting. The discovery of the Big Five is a century long story of trying to extract and structure information from descriptions of character. A century ago the electrical engineer LL Thurstone brought a form of Latent Semantic Analysis to bear on the problem, finding the variance in personality adjectives "could be accounted for by as few as five factors." In his words, "This fact leads us to surmise that the scientific description of personality may not be quite so hopelessly complex as it is sometimes thought to be" (Thurstone, 1934). As recently as 2015 LSA was a popular way to vectorize words in computer science. Since then recurrent and then transformer networks have dominated performance benchmarks. Models came to outperform human baselines and a more difficult benchmark was made (Wang et al., 2019). This work is a timely re-derivation of personality structure using these more powerful models.

The Big Five are so widely accepted it was assumed they would emerge again. However, experiments here show just two factors: socialization and self-actualization. This is consistent over many modeling choices. Factor analysis of language can answer how people are described, but need not align with explanatory theories of why personality develops. In our case, they do. Digman found these same two traits explain much of the

variance captured in Big Five surveys. He considered this a unifying discovery as it is empirical work that supports theoretical models of personality development (Digman, 1997). Ashton pushed back on Digman’s hierarchical interpretation of traits because they were found by analyzing correlations of constructs, not item level data (Ashton et al., 2009). This work solves that problem by finding α and β in the most basic description of personality, natural language. Structure found via deep learning fitting so neatly into previous theoretical and empirical work also validates that this method is a viable way to explore personality structure, opening up new research possibilities.

6.2 Future Work

IPIP105 is available online as a general personality extractor. With a few hundred samples a predictor can be trained on other arbitrary personality labels such as narcissism or subjective well being. Because dimensions correspond to IPIP items, this is also an interpretable embedding. LIWC is often used to find correlations between specific dimensions and a variable of interest. It is possible IPIP105 could be useful in this explanatory setting as well.

Chapter 5 demonstrates that deep language models can be used to extract reasonable structure from language. In the future, it will be interesting to compare the qualities especially when changing the embedding context or language. Just adjectives are used in these experiments, but the models can also embed descriptive phrases like “mighty as a lion”. Nouns could also be used with a phrase like “<mask> is another way to describe a PRISONER”. Much larger sets of descriptions can be explored than vectorization via surveys allows.

When factorized together, English and Spanish adjectives have two common factors. It will be interesting to see if this pattern continues with more distant language pairs or with more than two languages at once.

These are a few ways these tools may be used. Ultimately, the hope is that they extract more information from language than alternatives and are flexible enough for others to ask questions not considered here.

Appendix

Table 6.1: Pairwise Politics Words

IPA	anar.	centrist	con.	dem.	DC	HP	indep.	lib.	libert.	repub.	v. lib.
excited	fuck	wishes	wishes	smh	yay	rain	congrats	wishes	money	church	damn
xd	fuck	wishes	driving	excited	lol	dont	driving	excited	ready	school	excited
xd	fuck	damn	lord	today	tattoo	shit	surgery	shit	government	school	damn
packers	fuck	wishes	tonight	fb	anymore	shit	damn	damn	art	school	damn
class	fuck	wishes	lord	smh	stupid	fuck	died	wishes	government	church	wishes
hates politics	music	doy	loves	fb	tht	shit	definitely	wishes	government	church	damn
independent	fuck	wishes	lord	valentine	sitting	fuck	movie	wishes	email	camp	damn
liberal	fuck	final	lord	im	xd	im	gonna	wishes	beer	parents	damn
libertarian	fuck	headache	lord	walk	xd	dont	till	packing	government	church	damn
republican	fuck	wishes	lord	smh	mum	fuck	minute	wishes	fucking	girls	vacation
very liberal	xd	boy	lord	im	xd	xd	school	missing	im	im	damn

Table 6.2: Politics Confusion Matrix

IPA	Predicted Label											Total
	anar.	centrist	cons.	dem.	DC	HP	indep.	lib.	lib.	repub.	v. lib.	
0	2	3	3	11	18	2	1	3	1	16	1	61
anarchist	24	4	3	5	21	1	3	15	5	4	3	88
centrist	9	74	40	52	66	3	6	95	7	43	4	401
conservative	2	29	113	26	31	0	7	53	5	62	0	333
democrat	5	17	36	321	101	4	18	80	9	89	3	736
doesn't care	3	39	29	122	373	12	12	105	12	102	9	869
hates politics	0	4	1	6	30	5	3	6	0	2	0	63
independent	8	16	13	35	22	1	8	29	4	25	1	162
liberal	1	18	27	74	51	6	6	223	15	24	13	509
libertarian	0	12	9	17	28	0	6	32	11	12	4	148
republican	1	8	57	67	64	1	8	29	3	179	3	439
very liberal	0	4	2	11	22	2	2	67	1	6	3	145
Total	14	150	333	747	827	37	80	737	73	564	44	3954

Table 6.3: Personality Words

Openness		Conscientious		Extroversion	
-	+	-	+	-	+
bored	art	lost	gym	internet	party
boring	poetry	fucking	ready	quiet	guys
husband	beautiful	xd	weekend	bored	amazing
attitude	universe	phone	excited	listening	audition
shopping	peace	im	success	apparently	baby
dinner	poem	bored	finished	computer	haha
tv	writing	fuck	studying	stupid	dance
game	books	gonna	busy	pc	girls
proud	theatre	sick	vacation	hmm	fabulous
ur	dream	procrastination	arm	anime	blast
dentist	mind	internet	officially	tt	ready
daughter	book	computer	family	dark	im
haha	woman	probably	relax	probably	wine
stupid	guitar	cousins	tennis	sims	success
ni	damn	hates	wonderful	didn	lets
ipod	awesome	sims	special	watching	excited
bed	tea	anybody	win	slow	super
justin	apartment	charger	glad	depressing	text
gift	insomnia	sister	piano	calculus	chill
2nd	xd	playing	scholarship	kind	phone
hurt	adventure	grounded	received	anymore	dear
ohh	cali	poker	lmao	repost	parties
baseball	far	tt	degrees	maybe	support
mum	philosophy	status	state	draw	loves
pray	sigh	momma	tons	yay	pics
school	nature	ftw	motor	trying	hey
repost	maybe	press	obstacles	books	big
booked	music	dead	research	shadow	hit
lord	blues	failed	extremely	bother	met
ops	chill	forgot	circumstances	damned	pirate
nice	fam	depression	workout	suppose	ben
tmr	epic	lazy	paid	reading	rocked
dam	places	youtube	100	cat	gang
idol	rights	420	hit	poor	sex
snowing	dragons	school	surgery	depression	sing
pissed	woot	http	law	sigh	btw
shut	vampire	awsome	university	games	gorgeous
maths	soul	pokemon	anatomy	drawing	musical
msn	eclipse	woke	blessings	odd	cali
aldean	drawing	dammit	hmmmm	10th	girlfriend
vodka	strange	hair	husband	pokemon	stoked
comes	planet	wished	counting	nice	folks
eid	yay	cleaning	calc	essay	ponder
alot	dreams	fine	louis	pointless	wanna
waste	blood	dunno	delhi	managed	hahahaha
worst	sushi	enemy	final	looks	pool
kiero	smoking	social	drive	grr	tanning
soo	contact	yo	lets	darkness	hello
mas	lines	procrastinator	iphone	saw	pumped
staff	deep	black	lunch	crying	chillin
12	genius	magic	yankees	lonely	theatre
piss	novel	wasn	running	laptop	kiss
transformers	smh	fans	weather	shouldn	office
car	worried	kinda	zone	paranoid	cock
	folks	trying	smart	walking	lauren

Table 6.4: Personality Words Continued

Agreeable		Neurotic		Satisfaction With Life	
-	+	-	+	-	+
fucking	wonderful	loving	sick	bored	family
stupid	amazing	girlfriend	nervous	fuck	loving
kill	awesome	wife	stressed	fucking	hope
shopping	haha	awesome	depression	hates	thankful
shit	smile	parties	depressed	bday	india
burn	happiness	party	anymore	apparently	wonderful
bitch	phone	weekend	lonely	damn	busy
pissed	urself	haha	stress	internet	friend
punch	family	doing	fucking	zero	heart
hates	blessed	game	tired	chem	man
death	status	sunday	trying	wat	yum
hell	music	kansas	depressing	supposed	fb
suck	woop	guy	sims	ma	glad
freak	hands	delicious	anxiety	hating	beautiful
piss	heart	beach	worst	spend	lauren
dead	spirit	definitely	hair	la	lord
xmas	smiles	swag	fed	dumb	wine
karma	guy	started	scream	young	swim
fight	moment	ready	fine	british	energy
blood	beautiful	hunting	nightmare	killed	lunch
awful	movie	power	rip	hmm	locked
deal	theres	funniest	tears	france	woot
misery	car	melody	horrible	chances	sons
fuck	dancing	hawaii	flu	simply	special
enemies	lord	action	worse	exams	trust
fake	guitar	hit	issues	mum	wish
pathetic	sore	chillin	scared	main	weeks
irony	sara	workout	stressful	hate	day
dumb	help	flow	fml	edge	father
cunt	walk	portland	care	dnt	tried
care	excited	seat	shes	party	journey
devil	prayers	smart	stressing	kept	hospital
black	knowing	snowboarding	ugh	dat	email
ich	valentines	knowing	sad	didn	business
russian	borrow	sore	gary	months	santa
idiots	laura	greatest	hates	du	walked
cunts	notifications	success	die	rain	lights
wtf	beard	basketball	actually	pass	kingdom
crap	reli	update	scary	bus	work
truck	snowboarding	gf	boyfriend	okay	lol
deleted	sorry	women	pills	australia	mommy
anger	chillin	gotta	crying	shooting	turkey
die	hill	followed	kitty	england	nap
tu	whats	jumping	awful	africa	revenge
nightmare	hearts	fool	hurt	rachel	truly
annoyed	kindness	dancing	bored	fml	son
rip	study	greatness	fair	metal	final
bloody	worry	blast	screaming	uk	reached
drama	clients	woke	dreading	school	survived
bitches	smells	ass	friggin	wtf	dont
stupidity	troops	hitting	suicide	matt	0
hair	sing	cock	miserable	freakin	god
wifi	good	wise	quiet	15	kitchen
fat	holy	kiss	xd	200	normal
rage	faster	toes	sadness	free	blessing

Table 6.5: Sensational Interest Words

Militaristic		Violent-Occult		Intellectual Recreation	
-	+	-	+	-	+
sleeping	man	lord	hell	im	life
ugh	xbox	pray	zombie	course	jon
sad	gets	cousins	damn	boring	beautiful
excited	gotta	church	fuck	painful	dancing
lovely	good	michael	bitch	decision	yoga
oh	training	allah	ass	hurts	thankful
hair	headed	jesus	drink	bus	peace
shopping	truck	game	blood	game	kinda
husband	guitar	0	lmao	stupid	truly
sick	guys	summer	xd	bak	la
cares	bro	gosh	woot	hero	ich
mum	gun	praise	halloween	problem	miss
boyfriend	boom	sunday	play	yeah	likes
lady	epic	dad	guys	christ	comfort
concert	work	loving	drunk	gona	lol
today	weight	mum	thanx	id	wtf
gaga	gym	team	animal	sittin	insomnia
okay	bike	hospital	sanity	die	chicken
pic	dang	10	fucking	horse	children
adorable	game	tv	dragons	yell	tired
sunday	blast	christ	burn	chuck	lovely
ordered	lol	heal	vampires	2day	ap
birth	war	usa	blah	tommorrow	funny
lots	black	personal	man	ow	things
poor	fish	best	loved	bored	man
ben	military	ray	pissed	fukin	simple
fine	woot	nervous	lil	inbox	thank
settings	12	thing	bday	race	period
birthday	till	look	send	basketball	countdown
cousins	ppl	week	body	word	baby
shoes	brave	2morrow	metal	rhys	beach
art	17	quite	head	tell	hey
omg	fight	poor	piss	step	depression
stop	success	brazil	blast	wats	jobs
wear	marines	cup	theyre	coke	cure
prince	hrs	zumba	cause	football	manage
round	sword	account	gun	penguins	sugar
come	make	website	death	won	aware
neighbours	ko	tryna	vampire	facebookers	singing
basement	friend	study	bleh	letters	egg
music	hit	haha	tattoo	awsome	taste
speak	play	soccer	ppl	dont	rains
thoughts	pics	feeling	dead	blah	log
story	hahaha	christmas	woman	till	taught
weird	troops	round	purple	playing	coolest
awful	army	youth	peaceful	dead	yellow
quite	running	story	message	fact	cheers
rachel	mag	bible	shit	learned	small
hear	strong	woah	angel	visit	society
alice	knw	grace	kinda	address	fly
tea	beer	prayers	tongue	14	social
promised	hehehe	plan	sushi	chilling	boo
jesus	comwatch	feat	wolf	win	beauty
actually	xoxo	anybody	poke	pokemon	world
counting	run	stressed	kick	sees	sunshine

Table 6.6: Sensational Interest Words Continued

Occult Credulousness		Wholesome Activities		Belief in Star Sign	
-	+	-	+	No	Yes
church	zombie	coke	woot	minutes	omg
praise	ass	michigan	camping	didn	im
jesus	bitch	stupid	fish	church	ready
lord	halloween	pathetic	life	praise	friend
bible	animal	ops	yesterday	jesus	mind
christ	sign	husband	beautiful	probably	ass
team	omg	didn	rain	physics	butt
quite	xd	hurts	man	jess	stay
loving	job	kurwa	mexico	white	tom
pray	woot	evil	wish	religion	tomorrow
paper	wish	afternoon	river	iv	october
game	cure	problem	love	officially	promise
blessed	street	taylor	path	imagine	lol
salvation	vampire	idea	moon	christ	searching
ops	guys	jess	haha	germany	bitch
summer	send	glee	snow	giants	bleh
michael	lol	mum	bike	saw	eye
spent	thanx	mental	hahaha	wants	cute
youth	luck	meg	ghost	north	family
cousins	wtf	mad	baking	decided	halloween
word	nature	360	grandma	discovered	hanging
god	cancer	pissed	live	11th	haunted
homework	woohoo	club	goin	ouch	japanese
alarm	miss	uni	sky	skin	mother
0	barely	lyrics	cat	doesnt	dinner
haha	moment	head	animal	bacon	card
player	bar	recently	netflix	train	help
sunday	safe	internet	birds	hahaha	bored
college	proud	min	smile	lasts	luv
wedding	woman	lesson	happiness	america	luck
prayer	mom	bus	mom	haven	neighbors
glory	away	rly	yum	burning	yum
forgiveness	dare	debate	fishing	pray	fireworks
ann	inches	kevin	truly	thursday	lmao
mm	boyfriend	inbox	fell	jessica	tt
political	il	jeez	make	prince	tired
fact	nd	official	clean	knew	person
greatest	pls	nite	portland	umm	nd
confused	aware	ms	smells	quero	watch
appreciated	xmas	lack	lake	deserves	ya
algebra	hell	saw	create	heres	prom
brazil	solstice	troy	making	finds	crazy
travel	date	sims	2010	kim	upload
daughter	vampires	school	josh	heard	elf
bacon	copy	thinks	children	punch	hehe
laura	purple	thanking	laughing	groups	crack
personal	haunted	die	sa	car	bell
week	theyre	hates	law	amazing	human
greater	lmao	stuff	jobs	sick	finish
statement	later	band	earth	tape	lnk
messed	interview	thieves	gets	drink	june
tv	peeps	feels	hehehe	morn	change
em	peaceful	elm	swimming	dallas	costume
poor	drunk	germany	wa	cops	shit
trust	dunno	sat	monkeys	waters	decorating

Table 6.7: Psychographic Words

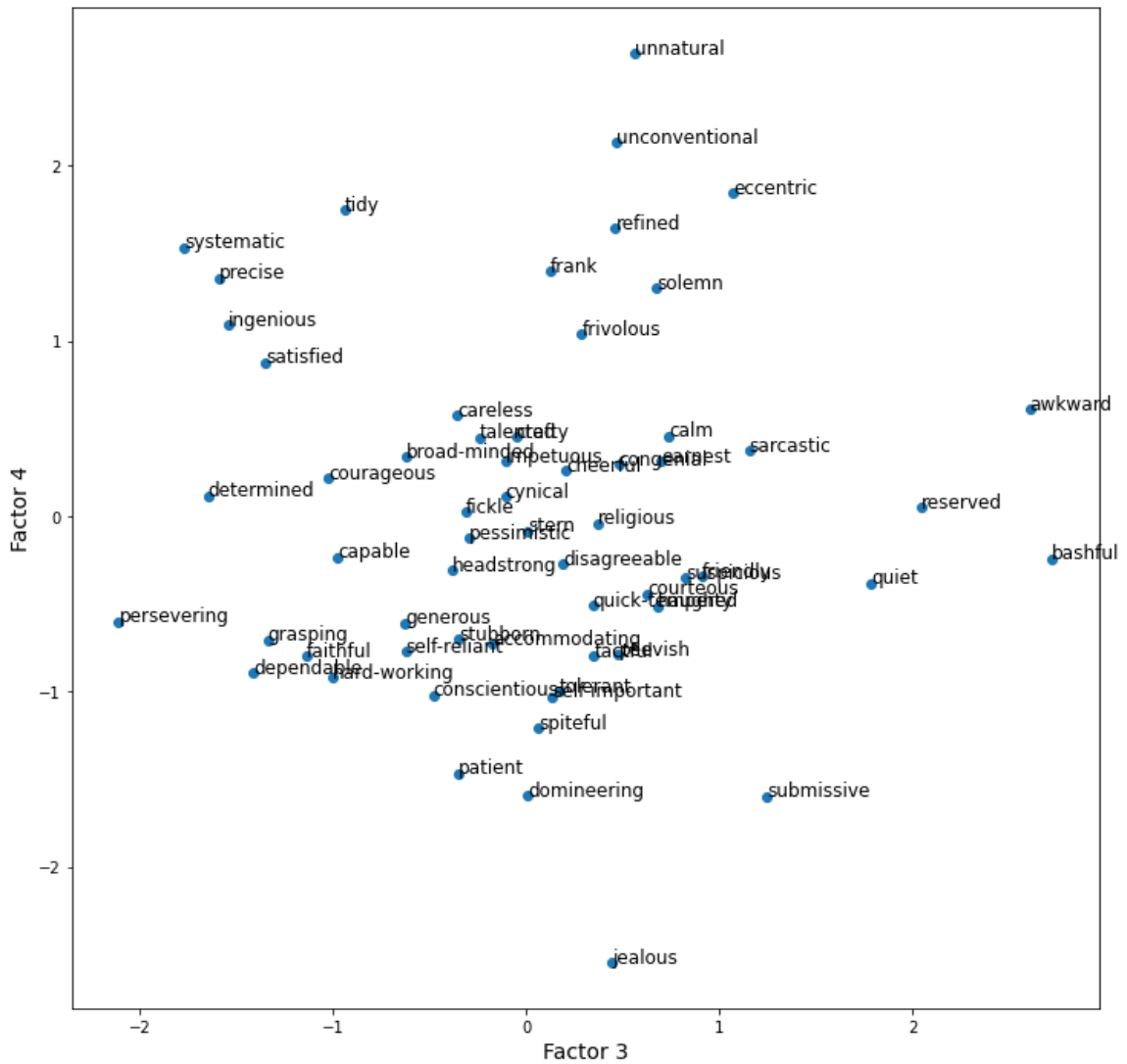
Self-Disclosure		Fair-Mindedness		IQ	
-	+	-	+	-	+
bored	family	bored	excited	nite	exam
fuck	loving	wat	business	ur	hours
fucking	hope	soon	says	lmao	sigh
hates	thankful	dad	apartment	alot	camping
bday	india	xd	great	family	finish
apparently	wonderful	stage	delicious	omg	paper
damn	busy	pass	sure	2011	wtf
internet	friend	moon	needed	city	il
zero	heart	haha	seattle	lol	finds
chem	man	kitty	uni	help	important
wat	yum	tired	airport	wew	read
supposed	fb	mum	thankful	boy	physics
ma	glad	farmville	dallas	heart	google
hating	beautiful	face	learn	com	ra
spend	lauren	drank	weekend	angie	xd
la	lord	fuk	definitely	www	wifi
dumb	wine	fuck	dinner	ha	text
young	swim	ma	card	333	weeks
british	energy	sun	amazing	tom	studying
killed	lunch	crap	tonight	goodnight	training
hmm	locked	bday	exciting	history	course
france	woot	shit	degrees	xxx	student
chances	sons	hopefully	classes	xdd	magic
simply	special	feel	support	friend	kinda
exams	trust	fails	priceless	morning	everytime
mum	wish	va	oh	mum	raining
main	weeks	big	certainly	christmas	yea
hate	day	nd	government	eid	maths
edge	father	smoke	ticket	kay	semester
dnt	tried	yay	food	gives	maybe
party	journey	watchin	january	din	exciting
kept	hospital	sick	couple	beautiful	point
dat	email	wedding	php	folks	kno
didn	business	regret	journey	luv	excited
months	santa	seconds	universe	0	imma
du	walked	im	21	hacked	months
rain	lights	ignore	grateful	secrets	flying
pass	kingdom	tt	pay	iam	final
bus	work	lose	size	forgiveness	nah
okay	lol	marriage	class	strong	library
australia	mommy	lolz	situation	busy	used
shooting	turkey	fukin	duke	jo	chem
england	nap	picture	honesty	hate	brain
africa	revenge	blessing	austin	ti	everybody
rachel	truly	slow	tires	nightmare	awesome
fml	son	anxiety	29	ayaw	groups
metal	final	cy3	sisters	prayer	progress
uk	reached	library	mother	fought	champion
school	survived	tmr	heading	ow	calculus
wtf	dont	fucking	bc	sana	behave
matt	0	epic	piece	tired	den
freakin	god	il	summer	afraid	badly
15	kitchen	marie	breakfast	para	times
200	normal	bunch	answer	sum	mobil
free	blessing	loaded	surgery	movie	fun

Table 6.8: Religion and Politics Words

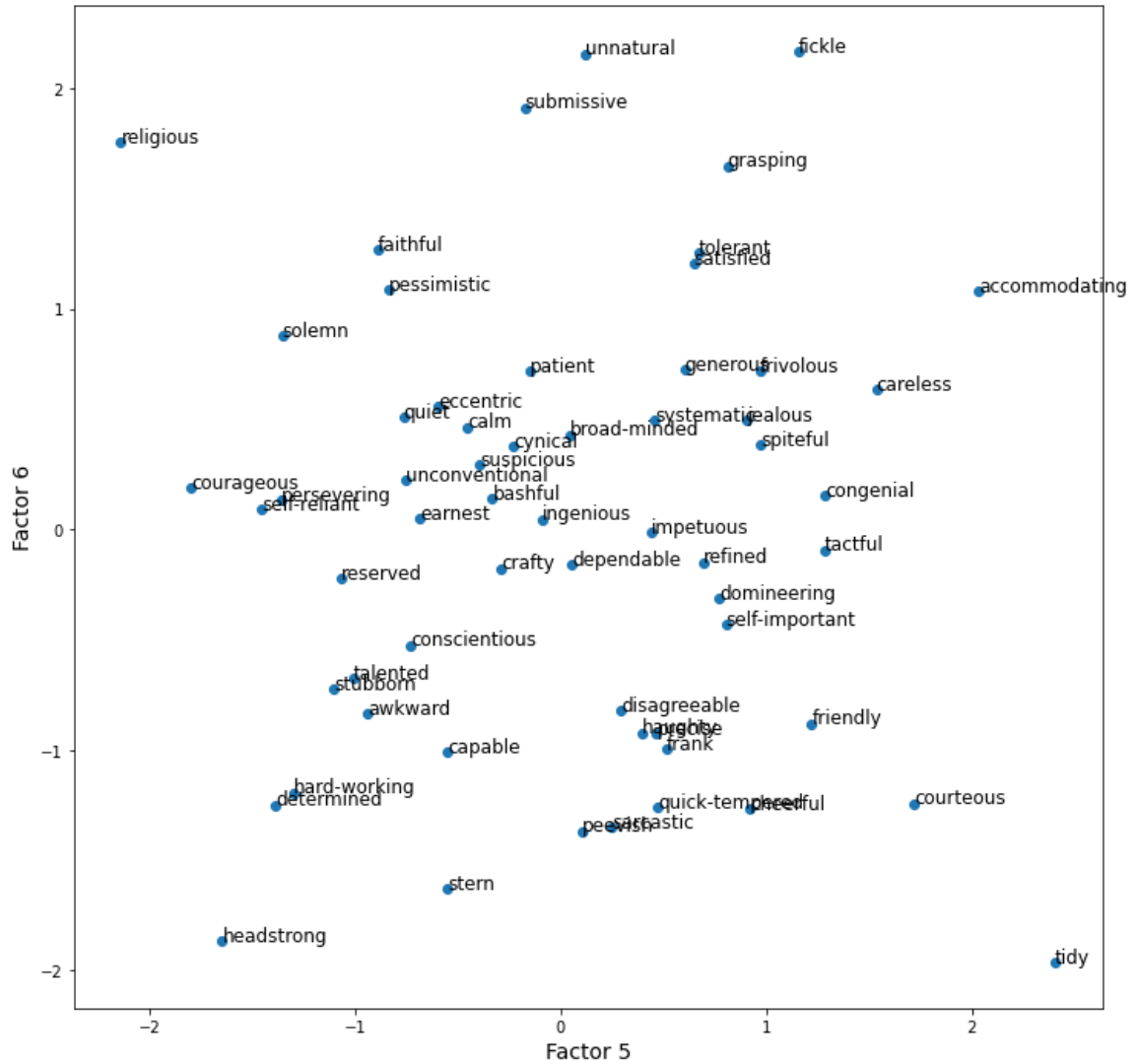
Agnostic vs Atheist	A. vs A. (Fair)	Religious vs Not	Conservative vs Liberal				
extra	physics	miles	fucking	church	fucking	church	damn
miles	fucking	working	physics	pray	fuck	truck	happy
turn	snowing	extra	wat	prayers	xmas	government	fb
hair	shit	awhile	fuck	god	damn	america	smh
packing	wat	packing	bloody	easter	shit	pray	marriage
awhile	write	turn	shit	lord	bloody	haha	xmas
insane	bloody	super	write	blessed	hell	prayers	chicago
working	enter	hubby	maths	christmas	ass	deer	sex
hubby	fuck	chill	xx	ugh	india	christmas	hell
points	sigh	free	snowing	praying	zombie	country	fam
friggin	thinks	sleepy	enter	hw	fuckin	tonight	lovely
santa	talk	santa	thinks	ppl	halloween	17	halloween
heck	weeks	heck	talk	prayer	car	lord	health
wishes	town	ready	science	game	yay	awesome	saw
child	science	friggin	sigh	believe	social	god	yoga
free	maths	vacation	hai	family	xx	military	celebrate
boyfriend	degrees	work	cancer	ready	quite	texas	gay
lady	lolz	thursday	person	fb	religion	freedom	apartment
learn	record	late	coursework	bless	drink	savior	wtf
super	xmas	points	town	im	oh	dad	thoughts
houston	tom	pack	xd	calling	using	bible	shit
service	hai	houston	weeks	dang	shitty	jesus	glee
pack	person	insane	tom	paper	internet	supper	gaga
late	dat	ya	film	jesus	fucked	girls	da
wanting	tyler	relax	dat	school	damned	huge	palin
hasn	cod	join	kill	camp	omfg	praying	2010
mai	afraid	busy	lolz	gosh	meh	camp	help
sleepy	untill	learn	msn	heart	indian	soldiers	mexico
worked	present	child	english	success	post	byu	mother
fly	wifey	headed	xmas	mary	head	christ	indian
chill	movie	favorite	chemistry	strength	cricket	disney	lady
join	xx	beautiful	afraid	butt	anyl	risen	studies
kyle	cancer	season	na	fishing	dragon	beach	social
dun	boring	san	pierced	brother	lovely	tournament	art
thursday	rape	fly	dick	military	body	troops	holiday
taken	month	worked	anatomy	sad	new	schools	shitty
childhood	kill	service	bbc	uncle	boyfriend	leave	ve
mother	welcome	spring	tell	senior	teeth	ill	free
thank	clinton	wanting	untill	fair	nice	blonde	earthquake
headed	nicht	halloween	memory	mom	fml	armed	street
ya	ay	lady	bothered	tan	warped	xbox	phone
london	brother	thank	horse	watching	woke	reagan	lakers
beautiful	tell	childhood	record	em	bleh	utah	ur
jail	hadn	mai	cod	president	wednesday	served	fine
hates	pierced	hair	ki	smh	gods	tide	relationship
paperwork	wild	paperwork	nicht	love	afford	gators	asshole
wanna	use	4th	sheep	haha	japanese	pelosi	worried
clear	perfect	hopefully	chem	future	tongue	husband	purple
san	return	missed	brother	best	robert	stinks	putting
til	needed	peace	fancy	emails	sophie	trial	omg
halloween	paid	hasn	degrees	goin	holy	picked	nature
bring	half	trip	disease	football	eye	beep	prop
kindle	horse	mother	realised	latest	tattoo	gun	black
vida	disease	sunshine	room	thank	decent	trailer	live
powers	chuck	kyle	religion	matthew	odd	ready	eid

Table 6.9: Race Words

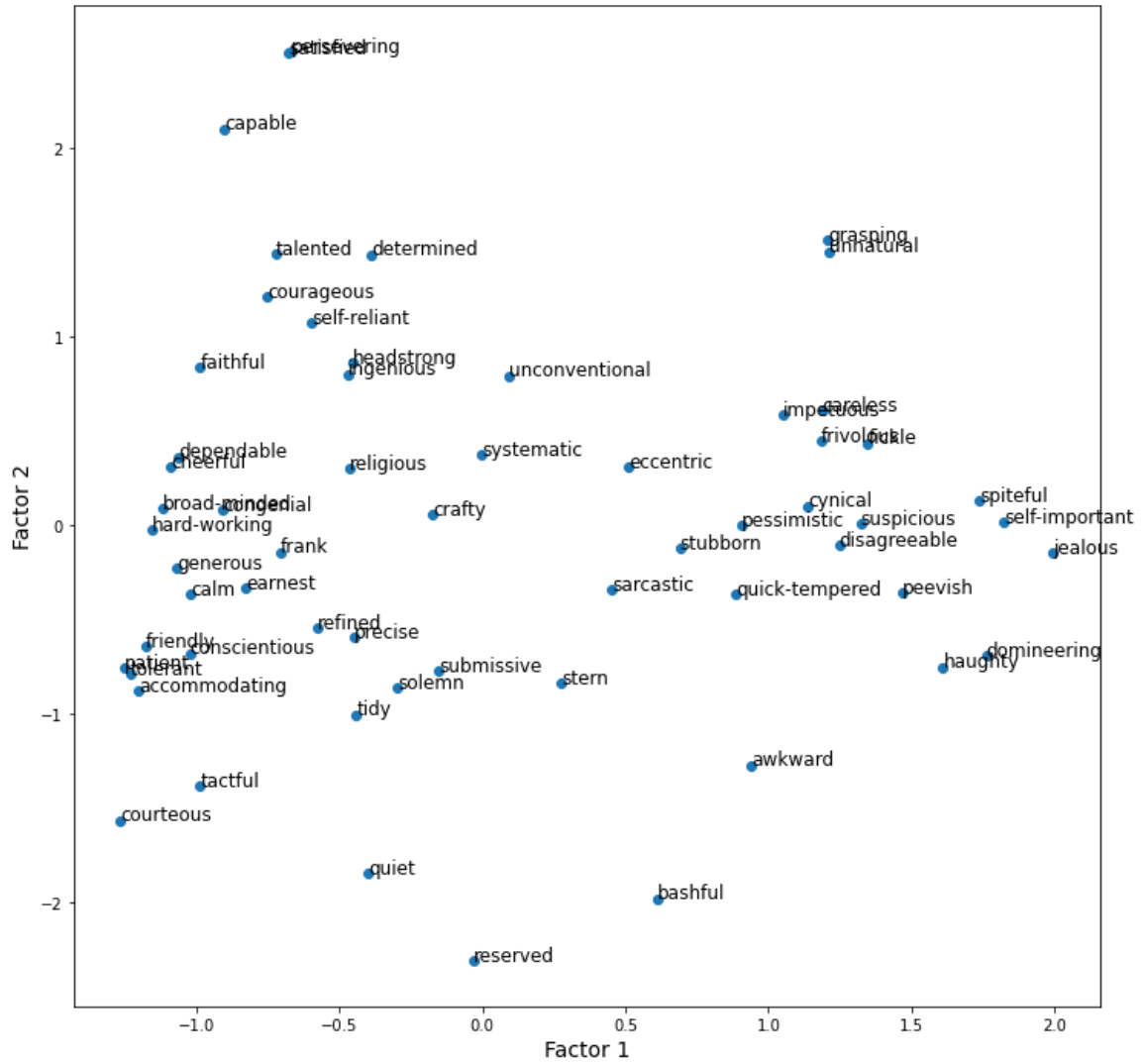
White vs. Black		White vs. Asian		Black vs. Asian	
tonight	smh	tonight	asian	smh	korea
dad	fb	blonde	tt	fb	sa
stupid	lord	town	tmr	lord	na
exited	fam	fuckin	korea	wit	asian
thinks	nigeria	ass	chinese	aint	gay
ends	yall	college	ng	da	chinese
journey	black	gas	na	yall	internet
meet	fathers	dope	korean	lol	korean
hahahahaha	mj	worse	china	say	monday
fun	yuh	night	ang	fam	xd
awesome	gon	men	aq	jackson	tmr
ability	birthday	sons	asians	cos	shooting
night	mad	adult	chen	michael	philippines
mas	lol	pretty	guys	finals	3d
wouldnt	finish	theres	thailand	ass	babe
chargers	dey	idea	taiwan	yuh	heaven
bein	asap	hope	karaoke	black	important
aftr	tryna	ability	sa	ny	tan
pretty	jackson	melissa	chan	sooooo	thailand
eh	came	state	dream	mad	yummy
tom	degrassi	unique	company	mind	completely
exhausted	wat	weekend	craving	season	woot
tough	iz	screaming	zzz	wat	smell
great	hw	mamaya	holiday	birthday	bought
running	ppl	tune	wanna	degrassi	fly
exciting	jus	figure	ms	hell	tt
yankees	braids	inside	nguyen	chelsea	worry
politics	haters	exited	singapore	woman	ruin
mirror	females	wine	yang	figure	passed
pepsi	misfits	5th	hu	african	skating
roll	god	superman	fat	nigeria	english
animal	man	emotionally	ftw	episode	belong
grr	omg	sell	gg	iz	shot
gay	african	sitting	rice	smart	mas
tattoo	desires	february	tttt	saying	grandpa
2nite	chelsea	easter	damnit	asap	lazy
spend	female	months	555	attention	sacrifice
monday	cousin	saying	wong	knowing	grr
sorrow	holla	expecting	achieve	ki	broken
ed	smart	rollin	pa	meeting	yang
healthy	laker	wheres	mode	hw	beer
enjoyable	favour	eminem	lmao	sings	chatting
actually	dis	apparently	pride	india	meet
charity	money	does	bbq	gas	shoulder
delete	happy	status	super	self	ang
iron	mii	legit	1st	ready	funn
blonde	aye	30	long	college	shoes
comforted	hard	wen	skating	mj	wood
standards	wuz	eric	mean	search	dad
shot	ready	yelled	heart	years	apart
chose	nigga	mis	dx	misfits	aj
chatting	jamaica	breaking	faith	blessed	line
damage	bus	homework	expectation	advice	jack
innocent	facebook	actually	research	boys	totally
thnx	cos	wishes	hard	fathers	tomorrow

Figure 6-2: Factor Analysis of Thurstone Words

RoBERTa embedding of each <mask> token with context of “My personality can be described as <mask> and WORD”.

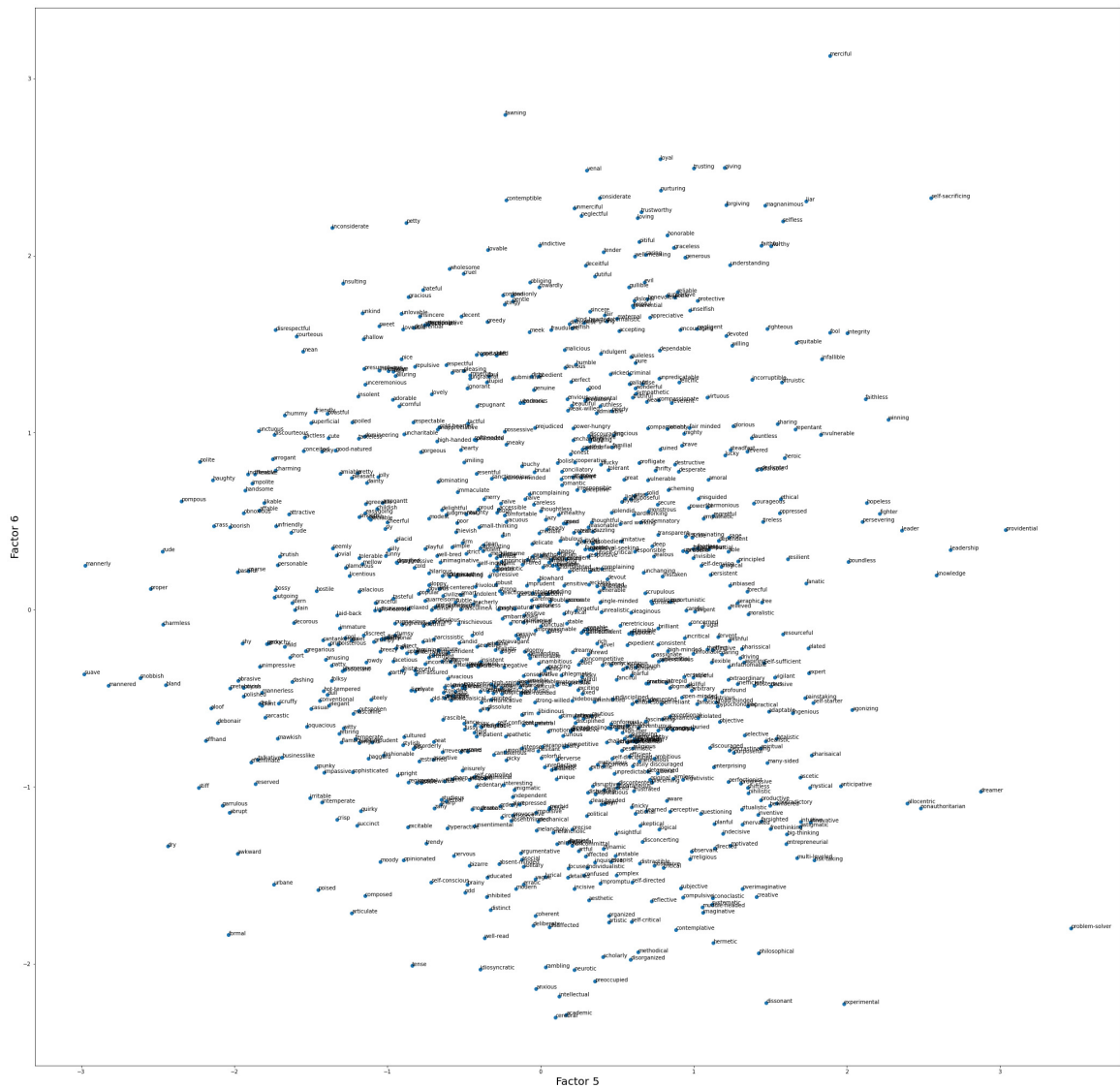
Figure 6-3: Factor Analysis of Thurstone Words

RoBERTa embedding of each <mask> token with context of “My personality can be described as <mask> and WORD”.

Figure 6.4: Factor Analysis of Thurstone Words

RoBERTa embedding of each <mask> token with context of “Those close to me say I am <mask> and WORD”.

Figure 6.5: Factor Analysis of 1,005 Words



1,005 RoBERTa embeddings of “My personality can be described as <mask> and WORD”. Personality structure is inscrutable past the third factor. Zoom in to view words.

Bibliography

- Ackerman, R. A., Witt, E., Donnellan, M., Trzesniewski, K., Robins, R., and Kashy, D. A. (2011). What does the narcissistic personality inventory really measure? *Assessment*, 18(1):67–87.
- Agüera y Arcas, B., Todorov, A., and Mitchell, M. (2018). Do algorithms reveal sexual orientation or just expose our stereotypes? <https://tinyurl.com/y4znkg6d>. Last checked Jan 2018.
- Allport, G. (1958). *What units shall we employ?* In G. Lindzey (Ed.), *Assessment of human motives* (pp. 239-260). New York: Holt, Rinehart and Winston.
- Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Ashton, M. C., Lee, K., and Goldberg, L. R. (2004). A hierarchical analysis of 1,710 english personality-descriptive adjectives. *Journal of Personality and Social Psychology*, 87(5):707.
- Ashton, M. C., Lee, K., Goldberg, L. R., and de Vries, R. E. (2009). Higher order factors of personality: Do they exist? *Personality and Social Psychology Review*, 13(2):79–91.
- Austin, J. (1961). *Philosophical papers*. Oxford: Clarendon Press.
- Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D. (2012). Personality and patterns of facebook usage. In *Proceedings of the 4th annual ACM web science conference*, pages 24–32.
- Back, M. D., Schmukle, S. C., and Egloff, B. (2011). Why are narcissists so charming at first sight? Decoding the narcissism-popularity link at zero acquaintance. *Journal of Personality and Social Psychology*, 98(1):132–145.
- Bagwell, L. S. and Bernheim, B. D. (1996). Veblen effects in a theory of conspicuous consumption. *The American Economic Review*, 86(3):349–373.

- Baik, J., Lee, K., Lee, S., Kim, Y., and Choi, J. (2016). Predicting personality traits related to consumer behavior using sns analysis. *New Review of Hypermedia and Multimedia*, 22(3):189–206.
- Bellanti, C. J. and Bierman, K. L. (2000). Disentangling the impact of low cognitive ability and inattention on social behavior and peer relationships. *Journal of Clinical Child Psychology*, 29(1):66–75.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.
- Bivens, R. (2017). The gender binary will not be deprogrammed: Ten years of coding gender on facebook. *New Media & Society*, 19(6):880–898.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.
- Branscombe, N. R. and Wann, D. L. (1994). Collective self-esteem consequences of out-group derogation when a valued social identity is on trial. *European Journal of Social Psychology*, 24(6):641–657.
- Brown, M. B. and Benedetti, J. K. (1977). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, 42(3):347–355.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Bryan, T., Wheeler, R., Felcan, J., and Henek, T. (1976). “come on, dummy” an observational study of children’s communications. *Journal of Learning Disabilities*, 9(10):661–669.
- Buffardi, L. E. and Campbell, W. K. (2008). Narcissism and social networking web sites. *Personality and Social Psychology Bulletin*, 34(10):1303–1314.
- Cadwalladr, C. (2018). “I made Steve Bannon’s psychological warfare tool”: meet the data war whistleblower. <https://tinyurl.com/yxq3487k>. Last checked July 2019.
- Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81.
- Capraro, V. and Sippel, J. (2017). Gender differences in moral judgment and the evaluation of gender-specified moral agents. *Cognitive processing*, 18(4):399–405.
- Carey, A. L., Brucks, M. S., Küfner, A. C., Holtzman, N. S., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Mehl, M. R., et al. (2015). Narcissism and the use of personal pronouns revisited. *Journal of personality and social psychology*, 109(3):e1.

- Celli, F., Bruni, E., and Lepri, B. (2014). Automatic personality and interaction style recognition from facebook profile pictures. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1101–1104.
- Cesare, N., Grant, C., and Nsoesie, E. O. (2017). Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*.
- Charles, K. E. and Egan, V. (2009). Sensational interests are not a simple predictor of adolescent offending: Evidence from a large normal british sample. *Personality and Individual Differences*, 47(4):235–240.
- Cherney, M. (2018). Facebook valuation drops \$75 billion in week after cambridge analytical scandal. MarketWatch. <https://tinyurl.com/y4xm6nwd>. Last checked April 2018.
- Chesire, L., Saffir, M., and Thurstone, L. L. (1933). *Computing diagrams for the tetrachoric correlation coefficient*. Chicago: University of Chicago Bookstore.
- Cisek, S. Z., Sedikides, C., Hart, C. M., Godwin, H. J., Benson, V., and Liversedge, S. P. (2014). Narcissism and consumer behavior: a review and preliminary findings. *Frontiers in Psychology*, 5(232).
- Clifton, A., Turkheimer, E., and Oltmanns, T. F. (2005). Self-and peer perspectives on pathological personality traits and interpersonal problems. *Psychological assessment*, 17(2):123.
- Clinchant, S. and Perronnin, F. (2013). Aggregating continuous word embeddings for information retrieval. In *Proceedings of the workshop on continuous vector space models and their compositionality*, pages 100–109.
- Collins, S., Sun, Y., Kosinski, M., Stillwell, D., and Markuzon, N. (2015). Are you satisfied with life?: Predicting satisfaction with life from facebook. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 24–33. Springer.
- Condon, D. M. (2018). The sapa personality inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. <https://psyarxiv.com/sc4p9>. Last checked Dec 2020.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Cooke, L., Wardle, J., Gibson, E., Sapochnik, M., Sheiham, A., and Lawson, M. (2004). Demographic, familial and trait predictors of fruit and vegetable consumption by pre-school children. *Public health nutrition*, 7(2):295–302.

- Cudeck, R. and MacCallum, R. C. (2007). *Factor analysis at 100: Historical developments and future directions*. Routledge.
- Cutler, A. and Kulis, B. (2018). Inferring human traits from facebook statuses. In *International Conference on Social Informatics*, pages 167–195. Springer.
- De Raad, B. and Barelds, D. P. (2008). A new taxonomy of dutch personality traits based on a comprehensive and unrestricted list of descriptors. *Journal of personality and social psychology*, 94(2):347.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- DeYoung, C. G., Quilty, L. C., and Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of Personality and Social Psychology*, 93(5):880–896.
- Dibbell, J. (2008). The decline and fall of an ultra rich online gaming empire. *Wired*. <https://www.wired.com/2008/11/ff-ige/>. Last checked April 2018.
- Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75.
- Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, 27(3):326–327.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.
- Digman, J. M. (1997). Higher-order factors of the big five. *Journal of personality and social psychology*, 73(6):1246.
- Dixon, S., Pampalk, E., and Widmer, G. (2003). Classification of dance music by periodicity patterns. *Proceedings of the Fourth International Conference on Music Information Retrieval*.
- Dolan, K. (2010). The impact of gender stereotyped evaluations on support for women candidates. *Political Behavior*, 32(1):69–88.
- Donnellan, M. B., Oswald, F. L., Baird, B. M., and Lucas, R. E. (2006). The mini-ipp scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192.
- Dunbar, R. and Dunbar, R. I. M. (1998). *Grooming, gossip, and the evolution of language*. Harvard University Press.

- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM.
- Eaton, W. W., Smith, C., Ybarra, M., Muntaner, C., and Tien, A. (2004). *Center for Epidemiologic Studies Depression Scale: review and revision (CESD and CESD-R)*, volume 3: Instruments for Adults, pages 363–377. Lawrence Erlbaum, Mahwah, NJ, 3rd edition.
- Egan, V., Austin, E., Elliot, D., Patel, D., and Charlesworth, P. (2003). Personality traits, personality disorders and sensational interests in mentally disordered offenders. *Legal and Criminological Psychology*, 8(1):51–62.
- Egan, V., Auty, J., Miller, R., Ahmadi, S., Richardson, C., and Gargan, I. (1999). Sensational interests and general personality traits. *The Journal of Forensic Psychiatry*, 10(3):567–582.
- Egan, V. and Campbell, V. (2009). Sensational interests, sustaining fantasies and personality predict physical aggression. *Personality and Individual Differences*, 47(5):464–469.
- Eysenck, H. J. (1944). Types of personality: a factorial study of seven hundred neurotics. *Journal of mental Science*, 90(381):851–861.
- Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.-F., and De Cock, M. (2016). Computational personality recognition in social media. *User modeling and user-adapted interaction*, 26(2-3):109–142.
- Farnadi, G., Tang, J., De Cock, M., and Moens, M.-F. (2018). User profiling through deep multimodal fusion. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 171–179.
- Fast, L. A. and Funder, D. C. (2008). Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology*, 94(2):334–346.
- Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2).
- Flynn, J. R. (1987). Massive iq gains in 14 nations: What iq tests really measure. *Psychological bulletin*, 101(2):171.
- Freud, S. and Strachey, J. (1930). Civilization and its discontents. *The Standard Edition*, 21:64–145.

- Friggeri, A., Lambiotte, R., Kosinski, M., and Fleury, E. (2012). Psychological aspects of social communities. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 195–202. IEEE.
- Furnas, G. et al. (1988). Using latent semantic analysis to improve information retrieval. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 281–285. ACM Press.
- Garrett, M. (2018). Trump campaign phased out use of cambridge analytica data before election. CBS News. <https://tinyurl.com/y4xtnof5>. Last checked April 2018.
- Giacomin, M. and Rule, N. O. (2019). Eyebrows cue grandiose narcissism. *Journal of Personality*, 87(2):373–385.
- Gignac, G. E. and Szodorai, E. T. (2016). Effect size guidelines for individual differences. *Personality and Individual Differences*, 102:74–78.
- Golbeck, J., Robles, C., and Turner, K. (2011). Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, pages 253–262.
- Golbeck, J. A. (2016). Predicting personality from social media text. *AIS Transactions on Replication Research*, 2(1):2.
- Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American psychologist*, 48(1):26.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96.
- Gottlieb, B. W., Gottlieb, J., Berkell, D., and Levy, L. (1986). Sociometric status and solitary play of ld boys and girls. *Journal of Learning Disabilities*, 19(10):619–622.
- Gow, A. J., Whiteman, M. C., Pattie, A., and Deary, I. J. (2005). Goldberg's 'ipip' big-five factor markers: Internal consistency and concurrent validation in scotland. *Personality and Individual Differences*, 39(2):317–329.
- Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2.
- Grijalva, E., Newman, D. A., Tay, L., and Donnelan, M. B. (2015). Gender differences in narcissism: a meta-analytic review. *Psychological Bulletin*, 141(2):261–310.

- Griskevicius, V., Tybur, J. M., Sundie, J. M., Cialdini, R. B., Miller, G. F., and Kenrick, D. T. (2017). Blatant benevolence and conspicuous consumption: when romantic motives elicit strategic costly signals. *Journal of Personality and Social Psychology*, 93(1):85–102.
- Gunkel, P. (2013). 638 primary personality traits. *Ideonomy: The Science of Ideas*. <http://ideonomy.mit.edu/essays/traits.html>.
- Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., and Lero Vie, M. (2013). How universal is the big five? testing the five-factor model of personality variation among forager–farmers in the bolivian amazon. *Journal of personality and social psychology*, 104(2):354.
- Hagger-Johnson, G., Egan, V., and Stillwell, D. (2011). Are social networking profiles reliable indicators of sensational interests? *Journal of Research in Personality*, 45(1):71–76.
- Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York Inc., New York, NY, USA, 2 edition.
- Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, 58(1):78–79.
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.
- Hoerl, A. and Kennard, R. W. (2000). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86.
- Hofstee, W. K., Kiers, H. A., De Raad, B., Goldberg, L. R., and Ostendorf, F. (1997). A comparison of big-five structures of personality traits in dutch, english, and german. *European Journal of Personality*, 11(1):15–31.
- Holtzman, N. S. (2011). Facing a psychopath: detecting the dark triad from emotionally-neutral faces, using prototypes from the Personality Faceaurus. *Journal of Research in Personality*, 45(6):648–654.
- Holtzman, N. S. and Strube, M. J. (2010). Narcissism and attractiveness. *Journal of Research in Personality*, 44(1):133–136.
- Holtzman, N. S. and Strube, M. J. (2011). *The intertwined evolution of narcissism and short-term mating: An emerging hypothesis.*, pages 210–220. Wiley, Hoboken, NJ.

- Holtzman, N. S. and Strube, M. J. (2013). Dark personalities tend to create a physically attractive veneer. *Social Psychological and Personality Science*, 4(4):461–467.
- Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Kufner, A. C., Deters, F. G., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., et al. (2019). Linguistic markers of grandiose narcissism: A liwc analysis of 15 samples. *Journal of Language and Social Psychology*, 38(5-6):773–786.
- Holtzman, N. S., Vazire, S., and Mehl, M. (2010). Sounds like a narcissist: Behavioral manifestations of narcissism in everyday life. *Journal of Research in Personality*, 44(4):478–484.
- Huddy, L. (2003). Group identity and political cohesion. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*. <https://doi.org/10.1002/9781118900772.etrds0155>.
- IBM Watson (2019). The science behind the service. <https://tinyurl.com/yytpxuon>. Last checked Dec 2020.
- Jackson, D. N. (1970). A sequential system for personality scale development. In *Current topics in clinical and community psychology*, volume 2, pages 61–96. Elsevier.
- Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.
- Jensen, A. R. (2001). Vocabulary and general intelligence. *Behavioral and Brain Sciences*, 24(6):1109.
- John, O. P. and Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.
- Johnson, C. E. (2019). Does sharing information with friends and family cause men to adhere more strongly to masculine norms? Master’s thesis, Georgia Southern University. <https://digitalcommons.georgiasouthern.edu/etd/1914/>.
- Jonason, P., Baughman, H. M., Carter, G. L., and Parker, P. (2015). Dorian gray without his portrait: Psychological, social, and physical health costs associated with the Dark Triad. *Personality and Individual Differences*, 78:5–13.
- Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2016). Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*.
- Karp, R. M., Shenker, S., and Papadimitriou, C. H. (2003). A simple algorithm for finding frequent elements in streams and bags. *ACM Transactions on Database Systems (TODS)*, 28(1):51–55.

- Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. *arXiv preprint arXiv:2006.16236*.
- Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., Kosinski, M., Ramones, S. M., and Seligman, M. E. P. (2014). The online social self: An open vocabulary approach to personality. *Assessment*, 21(2):158–169.
- Khandani, A. E., Kim, A. J., and Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787.
- Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*.
- Kleanthous, S., Herodotou, C., Samaras, G., and Germanakos, P. (2016). Detecting personality traces in users’ social activity. In *International conference on social computing and social media*, pages 287–297. Springer.
- Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
- Kleinberg, J. M. (2003). An impossibility theorem for clustering. In *Advances in neural information processing systems*, pages 463–470.
- Kohut, H. (1977). *The Restoration of the Self*. New York: International Universities Press.
- Kosinski, M. (2014a). The development and piloting of an online iq test. http://mypersonality.org/wiki/lib/exe/fetch.php?media=report_-_in_full.pdf. Last checked April 2018.
- Kosinski, M. (2014b). *Measurement and prediction of individual and group differences in the digital environment*. PhD thesis, University of Cambridge.
- Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543.
- Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.
- Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079.
- Laleh, A. and Shahram, R. (2017). Analyzing facebook activities for personality recognition. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 960–964. IEEE.

- Lau, R. R., Sigelman, L., and Rovner, I. B. (2007). The effects of negative political campaigns: a meta-analytic reassessment. *Journal of Politics*, 69(4):1176–1209.
- Lawley, D. (1953). A modified method of estimation in factor analysis and some large sample results. In *Uppsala symposium on psychological factor analysis*, volume 17, pages 35–42. Taylor & Francis.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Luo, J., Sorour, S. E., Goda, K., and Mine, T. (2015). Predicting student grade based on free-style comments using Word2Vec and ANN by considering prediction results obtained in consecutive lessons. Paper presented at the eight International Conference on Educational Data Mining (EDM). <https://files.eric.ed.gov/fulltext/ED560772.pdf>.
- Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 18(1):50–60.
- Markovikj, D., Gievska, S., Kosinski, M., and Stillwell, D. J. (2013). Mining facebook data for predictive personality modeling. In *Seventh International AAAI Conference on Weblogs and Social Media*. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/download/6179/6311>.
- Maslow, A. H. (1950). Self-actualizing people: a study of psychological health. *Personality, Symposium*, 1:11–34.
- McConaughy, S. H. and Ritter, D. R. (1986). Social competence and behavioral problems of learning disabled boys aged 6-11. *Journal of Learning Disabilities*, 19(1):39–45.
- McCrae, R. R. and Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.
- Megvii (2015). Face++ Photo Analysis. <https://tinyurl.com/y27fatdo>. Last checked April 2018.
- Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.

- Möttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., et al. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. *European Journal of Personality*, 34(6):1175–1201.
- Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D., and Rentfrow, J. (2018). Musical preferences predict personality: evidence from active listening and facebook likes. *Psychological Science*, 29(7):1145–1158.
- Nguyen, D., Trieschnigg, D., Dođruöz, A. S., Gravel, R., Theune, M., Meder, T., and De Jong, F. (2014). Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961.
- Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., and Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.
- Patton, J. H., Stanford, M. S., and Barratt, E. S. (1995). Factor structure of the barratt impulsiveness scale. *Journal of clinical psychology*, 51(6):768–774.
- Paulhus, D. L. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? *Journal of Personality and Social Psychology*, 74(5):1197–1208.
- Paulhus, D. L. (2001). Normal narcissism: Two minimalist accounts. *Psychological Inquiry*, 12(4):228–230.
- Paulhus, D. L. and John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, 66(6):1025–1060.
- Paulhus, D. L. and Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, 36(6):556–563.
- Peciña, M., Azhar, H., Love, T. M., Lu, T., Fredrickson, B. L., Stohler, C. S., and Zubieta, J.-K. (2013). Personality trait predictors of placebo analgesia and neurobiological correlates. *Neuropsychopharmacology*, 38(4):639.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

- Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin. <https://repositories.lib.utexas.edu/handle/2152/31333>.
- Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahwah: Lawrence Erlbaum Associates*, 71.
- Pennebaker, J. W. and King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.
- Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54:547–577.
- Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Pew Research Center (2014). Religious landscape study. <https://www.pewforum.org/religious-landscape-study/religious-family/agnostic/>. Last checked April 2018.
- Pincus, A. L., Pimentel, C. A., Cain, N. M., Wright, A. G. C., and Levy, K. N. (2009). Initial construction and validation of the pathological narcissism inventory. *Psychological Assessment*, 21(3):365–379.
- Preotiuc-Pietro, D., Carpenter, J., Giorgi, S., and Ungar, L. (2016). Studying the Dark Triad of personality through Twitter behavior. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16*, pages 761–770, New York, NY, USA. ACM.
- Preotiuc-Pietro, D., Liu, Y., Hopkins, D., and Ungar, L. (2017). Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 729–740.
- Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., and Crowcroft, J. (2012). The personality of popular facebook users. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 955–964.
- Quilty, L. C., Sellbom, M., Tackett, J. L., and Bagby, R. M. (2009). Personality trait predictors of bipolar disorder symptoms. *Psychiatry Research*, 169(2):159–163.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

- Raskin, R. and Shaw, R. (1988). Narcissism and the use of personal pronouns. *Journal of Personality*, 56(2):393–404.
- Raskin, R. and Terry, H. (1988). A principal-components analysis of the Narcissistic Personality-Inventory and further evidence of its construct-validity. *Journal of Personality and Social Psychology*, 54(5):890–902.
- Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B., and Funder, D. C. (2014). The situational eight DI-AMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personal and Social Psychology*, 107(4):677–718.
- Rodriguez, A. J., Holleran, S. E., and Mehl, M. R. (2010). Reading between the lines: The lay assessment of subclinical depression from written self-descriptions. *Journal of Personality*, 78(2):575–598.
- Rogers, C. R. (1961). *On Becoming a Person*. Boston: Houghton Mifflin.
- Rosenberg, M., Confessore, N., and Cadwalladr, C. (2018). How trump consultants exploited the data of millions. *New York Times*. <https://tinyurl.com/ybj4ulek>. Last checked April 2018.
- Salovey, P. and Mayer, J. D. (1990). Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211.
- Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.
- Saroglou, V. (2010). Religiousness as a cultural adaptation of basic traits: A five-factor model perspective. *Personality and social psychology review*, 14(1):108–125.
- Schmitt, D., Alcalay, L., Allik, J., Alves, I. C. B., Anderson, C., Angelini, A. L., Asendorpf, J., Austers, I., Balaguer, I., Baptista, A., et al. (2017). Narcissism and the strategic pursuit of short-term mating: Universal links across 11 world regions of the International Sexuality Description Project-2. *Psychological Topics*, 26(1):89–137.
- Schneider, M. C. and Bos, A. L. (2014). Measuring stereotypes of female politicians. *Political Psychology*, 35(2):245–266.
- Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., and Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. *Personality and individual differences*, 25(2):167–177.
- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013a). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

- Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. (2013b). Personality, gender, and age in the language of social media: The open-vocabulary approach. *Plos One*, 8(9).
- Sedikides, C., Gregg, A. P., Cisek, S., and Hart, C. M. (2007). The I that buys: Narcissists as consumers. *Journal of Consumer Psychology*, 17(4):254–257.
- Seven Steps to Learn English (2020). Personality adjectives: 100+ words to describe someone in english. <https://tinyurl.com/yxwvc72g>. Last checked April 2020.
- Shamir, R. and Sharan, R. (2002). 11 algorithmic approaches to clustering gene expression data. *Current Topics in Computational Molecular Biology*, page 269.
- Shiramizu, V. K. M., Dozma, L., DeBruine, L. M., and Jones, B. C. (2019). Are Dark Triad cues really visible in faces? *Personality and Individual Differences*, 139:214–216.
- Sindermann, C., Elhai, J. D., and Montag, C. (2020). Predicting tendencies towards the disordered use of facebook’s social media platforms: on the role of personality, impulsivity, and social anxiety. *Psychiatry Research*, 285:112793.
- Skinner, B. (1971). *Beyond freedom and dignity*. Knopf/Random House, New York, NY, US.
- Skorska, M. N., Geniole, S. N., Vrysen, B. M., McCormick, C. M., and Bogaert, A. F. (2015). Facial structure predicts sexual orientation in both men and women. *Archives of Sexual Behavior*, 44(5):1377–1394.
- Smeland, O. B., Wang, Y., Lo, M.-T., Li, W., Frei, O., Witoelar, A., Tesli, M., Hinds, D. A., Tung, J. Y., Djurovic, S., et al. (2017). Identification of genetic loci shared between schizophrenia and the big five personality traits. *Scientific reports*, 7(1):1–9.
- Snell, W. E., Miller, R. S., and Belk, S. S. (1988). Development of the emotional self-disclosure scale. *Sex Roles*, 18(1-2):59–73.
- Sniekers, S., Stringer, S., Watanabe, K., Jansen, P. R., Coleman, J. R., Krapohl, E., Taskesen, E., Hammerschlag, A. R., Okbay, A., Zabaneh, D., et al. (2017). Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature genetics*, 49(7):1107.
- Soto, C. J. and John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Personality and Individual Differences*, 113(1):117–143.
- Spark Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21. <https://doi.org/10.1108/eb026526>.

- Spearman, C. (1904). "General Intelligence" objectively determined and measured. *American Journal of Psychology*, 15(5):201–293.
- Stecklow, S. (2018). Why facebook is losing the war on hate speech in myanmar. Reuters. <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>. Last checked Dec 2020.
- Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. *Developmental review*, 28(1):78–106.
- Stetler, D. A., Davis, C., Leavitt, K., Schriger, I., Benson, K., Bhakta, S., Wang, L. C., Oben, C., Watters, M., Haghnegahdar, T., et al. (2014). Association of low-activity maoa allelic variants with violent crime in incarcerated offenders. *Journal of psychiatric research*, 58:69–75.
- Stillwell, D. J. and Kosinski, M. (2012). mypersonality project: Example of successful utilization of online social networks for large-scale social research. In *Tenth International Conference on Mobile Systems, Applications, and Services (MobiSys)2012*.
- Sumner, C., Byers, A., Boochever, R., and Park, G. J. (2012). Predicting dark triad personality traits from Twitter usage and a linguistic analysis of tweets. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 386–393.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.
- Tett, R. P., Jackson, D. N., and Rothstein, M. (1991). Personality measures as predictors of job performance: a meta-analytic review. *Personnel psychology*, 44(4):703–742.
- Thilakaratne, M., Weerasinghe, R., and Perera, S. (2016). Knowledge-driven approach to predict personality traits by leveraging social media data. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 288–295. IEEE.
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of applied psychology*, 4(1):25–29.
- Thurstone, L. L. (1934). The vectors of mind. *Psychological review*, 41(1):1.
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.
- Valentova, J. V., Kleisner, K., Havlíček, J., and Neustupa, J. (2014). Shape differences between the faces of homosexual and heterosexual men. *Archives of Sexual Behavior*, 43(2):353–361.

- Vazire, S., Naumann, L. P., Rentfrow, P. J., and Gosling, S. D. (2008). Portrait of a narcissist: Manifestations of narcissism in physical appearance. *Journal of Research in Personality*, 42(6):1439–1447.
- Vehtari, A., Gelman, A., and Gabry, J. (2015). Efficient implementation of leave-one-out cross-validation and waic for evaluating fitted bayesian models. *arXiv preprint arXiv:1507.04544*.
- Verdejo-García, A., Lawrence, A. J., and Clark, L. (2008). Impulsivity as a vulnerability marker for substance-use disorders: review of findings from high-risk research, problem gamblers and genetic association studies. *Neuroscience & Biobehavioral Reviews*, 32(4):777–810.
- Vize, C. E., Lynam, D. R., Collison, K. L., and Miller, J. D. (2018). Differences among Dark Triad components: A meta-analytic investigation. *Personality disorders: Theory Research and Treatment*, 9(2):101–111.
- Vrandečić, D. (2020). Architecture for a multilingual wikipedia.
- Wald, R., Khoshgoftaar, T., and Sumner, C. (2012). Machine prediction of personality from facebook profiles. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pages 109–115. IEEE.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in neural information processing systems*, pages 3266–3280.
- Wang, N., Kosinski, M., Stillwell, D., and Rust, J. (2014). Can well-being be measured using facebook status updates? validation of facebook’s gross national happiness index. *Social Indicators Research*, 115(1):483–491.
- Wang, Y. and Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246.
- Watson, J. B. (1919). *Psychology: From the standpoint of a behaviorist*. JB Lippincott.
- Webster, M. (2014). Personality types - english vocabulary word list: Learner’s dictionary.
- Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.
- Worsley, A. (2020). 800 character traits: The ultimate list. The Art of Living. <https://tinyurl.com/yyqjysyo>.

- Xu, A., Liu, Z., Guo, Y., Sinha, V., and Akkiraju, R. (2017). A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3506–3510.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.
- Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.
- Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee.
- Zhang, L., Zhao, L., Zhang, X., Kong, W., Sheng, Z., and Lu, C.-T. (2018). Situation-based interpretable learning for personality prediction in social media. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1554–1562. IEEE.
- Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52.
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.
- Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.

Curriculum Vitae

