Personality Predictive Lexical Cues and Their Correlations

Xiaoli He
Rutgers University
New Brunswick, NJ, USA
hxl.daybreak@gmail.com

Gerard de Melo
Hasso Plattner Institute / University of Potsdam
Potsdam, Germany
http://gerard.demelo.org

Abstract

In recent years, a number of studies have used linear models for personality prediction based on text. In this paper, we empirically analyze and compare the lexical signals captured in such models. We identify lexical cues for each dimension of the MBTI personality scheme in several different ways, considering different datasets, feature sets, and learning algorithms. We conduct a series of correlation analyses between the resulting MBTI data and explore their connection to other signals, such as for Big-5 traits, emotion, sentiment, age, and gender. The analysis shows intriguing correlation patterns between different personality dimensions and other traits, and also provides evidence for the robustness of the data.

1 Introduction

The notion of personality refers to an individual's characteristic patterns of thinking, feeling, and behaving (Sherman et al., 2013). Studies have shown that personality influences an individual's language usage (Schwartz et al., 2013b; Tucker, 1968; Hirsh and Peterson, 2009). Hence, language may reveal subtle cues about an individual's personality. Since an individual's personality is known to be fairly stable across long periods of time, the relation between personality and language usage is expected to be analyzable given sufficiently large amounts of textual data.

Motivation. Many people now routinely post information pertaining to their daily life, thoughts, emotions, and opinions on different online social media platforms. A number of studies have shown that such social media text may enable automated personality predictions, as reviewed in more detail in Section 2. Yet, automatic personality prediction is still a challenging problem, and there remain several unresolved issues.

First, due to privacy concerns and the high labeling cost, the number of publicly available labeled datasets is limited, and the sample size in such datasets is often rather small (especially when compared with the high dimensionality of n-gram features). Beyond this, some datasets only provide a small number of sentences per sample. These limitations make it difficult to know to what extent results in individual studies generalize across different datasets.

Second, it is non-trivial to compare results across different studies, as they adopt different feature representations and machine learning methods, and consider different personality models.

Two particularly well-known personality models are Myers-Briggs Type Indicators (MBTI) and Big-5 traits, which we introduce in more detail in Section 2. In the field of personality psychology, studies have shown clear correlations between *self-reported* MBTI and Big-5. For instance, MBTI's INTROVERSION–EXTRAVERSION correlates with Big-5 EXTRAVERSION, MBTI's SENSING–INTUITION and JUDGING–PERCEIVING correlate with Big-5's OPENNESS trait, and MBTI's JUDGING–PERCEIVING also correlates with Big-5 CONSCIOUSNESS (Tobacyk et al., 2008).

This raises the question of whether signals from naturally occurring text exhibit similar connections, and how they relate to other psychological and demographic variables.

Goals and Contributions. The goal of this paper is thus to empirically compare personality cues at the lexical level across different datasets, personality models, and methods. In our study we focus primarily on lexical signals based on multiple MBTI datasets from heterogeneous sources, which we compare against lexical cues for Big-5 traits. Additionally, we explore connections to sentiment and emotion lexicons, as well to demographic cues.

Proceedings of Recent Advances in Natural Language Processing, pages 514–523
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_058

2 Background and Related Work

Personality Models. Different models have defined different traits (sub-dimensions) of personality. Two prominent ones are the MBTI and Big-5 schemes. The Myers-Briggs Type Indicator model (MBTI) (Myers et al., 1985) consists of the following four dimensions:

1. INTROVERSION–EXTRAVERSION (I–E): where a person focuses their attention;
2. INTUITION–SENSING (N–S): the way a person tends to take in information;
3. THINKING–FEELING (T–F): how a person makes decisions;
4. JUDGING–PERCEIVING (J–P): how a person deals with the world.

Numerous concerns have been raised regarding the validity and reliability of MBTI. For instance, MBTI assumes binary categorical labels for the aforementioned four dimensions, denoted e.g. as ESTJ, although the majority of people appear to exhibit a combination of different traits along a dimension. Still, MBTI is perhaps the most widely known model, and frequently mentioned in online profiles. In contrast, the Big-5 model (Goldberg, 1990) considers continuous scores along the following five dimensions:

1. EXTRAVERSION (extroversion): describes how outgoing and social a person is;
2. AGREEABLENESS: reflects how warm, friendly, and tactful a person is;
3. OPENNESS: considers how open-minded and authority-challenging a person is;
4. CONSCIENTIOUSNESS: reflects how self-disciplined and organized a person is;
5. NEUROTICISM (emotionism): indicates a person's ability to remain stable and balanced.

Personality Assessment. In psychological assessments, personality is typically measured by means of standardized questionnaires that evaluate particular aspects of personality. This form of personality measurement has generally been found to be fairly stable and consistent. However, a major disadvantage is that experts first need to carefully compile long lists of questions, and individuals then need to explicitly fill out the questionnaire.

This has motivated research into computational analyses of naturally occurring text with the aim of obtaining automated assessments that correlate with the professional ones. In this regard, recent studies have considered several different social media platforms and personality scales. Different models have been developed, from simple Logistic or Linear Regression ones (Arnoux et al., 2017), support vector machines (Biel et al., 2013; Kumar and Gavrilova, 2019), to more complex models such as stability selection (Plank and Hovy, 2015), Gaussian process models (Arnoux et al., 2017), and ensemble methods aggregating multiple classifiers or regressors (Kumar and Gavrilova, 2019).

Past studies have also considered different features, such as word unigrams, word n-grams (Plank and Hovy, 2015; Yarkoni, 2010; Biel et al., 2013), and word embeddings (Arnoux et al., 2017; Siddique et al., 2019). For the studies using n-grams as features, some apply TF-IDF weighting schemes (Siddique et al., 2019; Biel et al., 2013), while others use unweighted features (Plank and Hovy, 2015; Yarkoni, 2010; Kern et al., 2014).

While the above studies have mostly sought to improve the performance of personality prediction on a given dataset using a variety of different methods and features, our study focuses on assessing the contribution of individual words and n-grams as signals for personality prediction, and their relationship to other lexical cues. In previous work, a few studies have focused on broader associations between personality and aggregate word categories (Yarkoni, 2010), such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001). However, this may mask the contribution of individual words in the context of an open vocabulary scenario.

Lexical Analyses. Lexicon-driven analyses have proven fruitful in areas such as sentiment analysis (Ding et al., 2008; Mohammad et al., 2013; Kiritchenko et al., 2014; Islam et al., 2020) and emotion analysis (Kulahcioglu and de Melo, 2018; Raji and de Melo, 2020; Raji and de Melo, 2021), especially when there is no labeled data, as well as in social science and digital humanities (Pennebaker et al., 2001). With this approach, a dictionary of words (or bag of words) is generated, with a positive or negative value assigned to each word, reflecting the predictive power or correlation strength between the word and the specific target label or variable. Sap et al. (2014) explored lexical cues for age and gender. In traditional personality research, psychologists have developed closed-book vocabularies by self-rating on personality trait adjectives or verbs (Ashton et al., 2004b,a).

In light of the above, exploring automatically induced lexical cues for personality prediction is a promising endeavor, and the resulting lexical signals can also be compared with lexical cues for other variables.

3 Lexical Cue Induction

In order to determine which words and n-grams are most correlated with specific personality variables, we assume a supervised learning setup with labeled training data that allows us to train a separate linear model for each target variable and identify salient lexical cues along with their weights. We compare several different variants with different feature representations and learning algorithms.

3.1 Feature Representations

Data Preprocessing. Our study considers only the linguistic information for each sample, along with the personality type labels, ignoring multimodal signals and metadata. The text is tokenized and the following preprocessing steps are applied:

1. Lower-casing;
2. Removing English stop words, tokens consisting only of numbers, and tokens mentioning personality types;
3. Replacing URLs, hashtags, usernames with '@URL' , '@HASHTAG', '@USER'.

Feature Extraction. We extract unigram, 1-2 gram (unigram + bigram), and 1-2-3 gram (unigram + bigram + trigram) feature sets for each of the datasets. Due to the combinatorial explosion of n-grams, we apply a minimum frequency threshold, dropping any n-grams appearing less than 1% in each dataset. We further exclude tokens that consist solely of numbers. For the 1-2 grams and 1-2-3 grams, we also exclude tokens with punctuation in the first character or in the middle, such as ('!', 'I'), ('today', '!', 'I'), ('!', 'today', 'I').

N-gram Weighting. Weighting is often used to adjust the importance of individual features. Besides using n-grams directly, we also used three types of weightings for each n-gram:

1. Relative term frequency, $\frac{\mathrm{freq}(w,d)}{\mathrm{freq}(*,d)}$, is defined as the relative term frequency (TF) of a word w within a document d.
2. TF-logIDF is a common definition of Term Frequency-Inverse Document Frequency (TF-IDF) weighting that incorporates a logarithmic scaling of IDF to dampen the effect of the ratio. In general, TF-IDF representations downweight words that appear universally across many documents, as these are less likely to be sufficiently discriminative in personality prediction.
3. TF-IDF differs from the above form in that we omit the logarithmical scaling of IDF.

3.2 Learning Algorithms

Linear models have often been used to induce weighted lexicons. Sap et al. (2014) compared the formula of linear multivariate models $y = (\sum_f w_f x_f) + w_0$ (summing over features f) with the use of a weighted lexicon L with term weights $w_L(t)$ that is applied to a document d with frequencies $f(t,d)$ as $\sum_{t \in L} w_L(t) \frac{f(t,d)}{f(*,d)}$. They prove that if relative term frequency is used as the feature representation, many multivariate modeling techniques can be viewed as learning a weighted lexicon plus an intercept.

Hence, the weight of a word in a lexicon can be obtained based on the coefficients from linear multivariate models. We thus treat each personality dimension as a distinct and independent classification or regression problem. For each combination of feature and weighting, we investigate four types of learning algorithms.

Stability Selection. In stability selection (Meinshausen and Bühlmann, 2010), the training data is repeatedly resampled in a bootstrap operation, and a model is learned for each such iteration, and features selected more frequently are presumed to be more robust indicators. As the base model, we use randomized logistic regression for MBTI datasets, and Randomized Lasso for the Big-5 dataset. We run 100 resampling procedures, such that on each resampling, 75% of the samples are randomly chosen. After the step of stability selection, we apply logistic regression for MBTI (linear regression for Big-5) on the selected features (n-grams), and save their coefficients.

Penalized Ridge Classification/Regression. We further consider Ridge Regression, i.e, linear least squares regression with L1 regularization. For MBTI, we apply Ridge Classification, i.e., the target classification is mapped to $\{-1, 1\}$ so as to cast the problem as a regression task. The L1 penalty encourages sparse features, which is well-suited for our goal of identifying salient lexical cues. We split each dataset into training set and test set randomly with a ratio of 3:1 (also using the same ratio for the following two approaches).

Penalized Support Vector Classification with Linear Kernel. Support vector machines are well-suited for high-dimensional vector representations. Considering the high dimensionality and sparsity of our feature space, we consider support vector classification/regression with a linear kernel and L1 penalty in a 10-fold cross-validation setup.

Penalized Multi-Layer Perceptrons. Lastly, to better account for non linearly separable data, we consider a feed-forward neural network with a 100-dimensional hidden layer and RelU activation function, trained using Adam optimization with an initial learning rate of 0.001.

4 Lexical Cue Analysis

In the following, we empirically assess lexical cues induced using the aforementioned techniques.

4.1 Datasets

Our analysis is based on 8 MBTI datasets and one Big-5 dataset, all consisting of naturally occurring English language text annotated with personality traits. For the former, we illustrate the respective data distributions in Figure 1. In particular, *kaggle* refers to the Kaggle Personality Cafe MBTI dataset, which provides 8,600 samples collected from the discussion forums of the Personality Cafe website. The Twitter datasets *twitter_100g*, *twitter_500g*, *twitter_2000g* are obtained from Plank and Hovy (2015). Each such dataset contains 1,500 samples, but they differ in the number of tweets per sample (100, 500, or 2,000). The *reddit* dataset is taken from Gjurković and Šnajder (2018), and provides 9,149 rows of comments from different Reddit authors with more than 1,000 words each. Due to the computational burden of the feature computation for bi- and tri-grams, we additionally also consider splits into smaller subsets (*reddit0*, *reddit1*, *reddit2*), which are mainly used for analysis (see the next section for further details).

Figure 1 shows the distribution within each dimension in the different MBTI datasets. For each dimension, the first type is coded as 0, and the second type is coded as 1. For example, for I–E, INTROVERT is represented as 0, and EXTRAVERT is represented as 1. Figure 1 shows that, overall, each dataset has more INTROVERT and THINKING individuals. It has been reported that INTROVERT individuals prefer online communication (Plank and Hovy, 2015; Goby, 2006), though this overrepresentation may also have other causes. Interestingly, there are some differences between users in the different datasets. The Reddit data has more INTUITIVE users, while the Twitter data has more JUDGING users. The Reddit data includes slightly more users with the THINKING trait than the Twitter dataset.

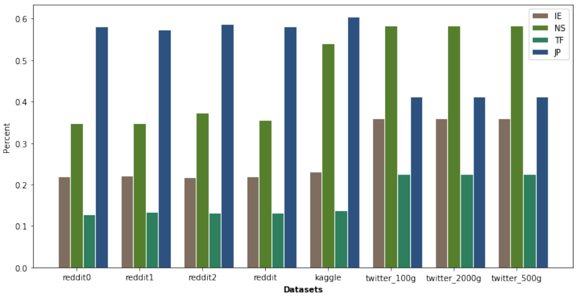

Figure 1: MBTI distribution on each dataset, where the dimensions I–E, N–S, T–F, and J–P are each mapped from 0 to 1

We only considered a single Big-5 dataset, based on YouTube video blogs (Biel et al., 2013). The texts are the manually created transcripts, and the Big-5 score is not self-reported but rather the impression score assigned by a separate group of subjects, unlike most other datasets in our study.

4.2 Prediction Quality of Different Models

The prediction accuracies obtained for each combination of feature, weighting scheme, and model are given in Table 1, for each dimension of MBTI. The three numbers in a given cell represent the results from the three different weighting schemes: relative frequency, TF-logIDF, TF-IDF. Note that, owing to scalability considerations for trigrams, the 1-2-3 gram feature set was only considered for stability selection.

The results suggest that 1) the accuracy is fairly similar across different weighting schemes; 2) the accuracies consistently increase from unigrams to 1-2 grams, but only modestly with 1-2-3 gram features. We additionally plot the results using 1-2 grams weighted by TF-logIDF for each method in Figure 2. Each sub-figure shows the results from one model. Within each sub-figure, different bars indicate the results for different personality dimensions. Figure 2 conveys the following two messages: First, it shows that for the three linear models with penalty, dimension N–S obtains the highest accuracy, I–E the second highest, while J–P

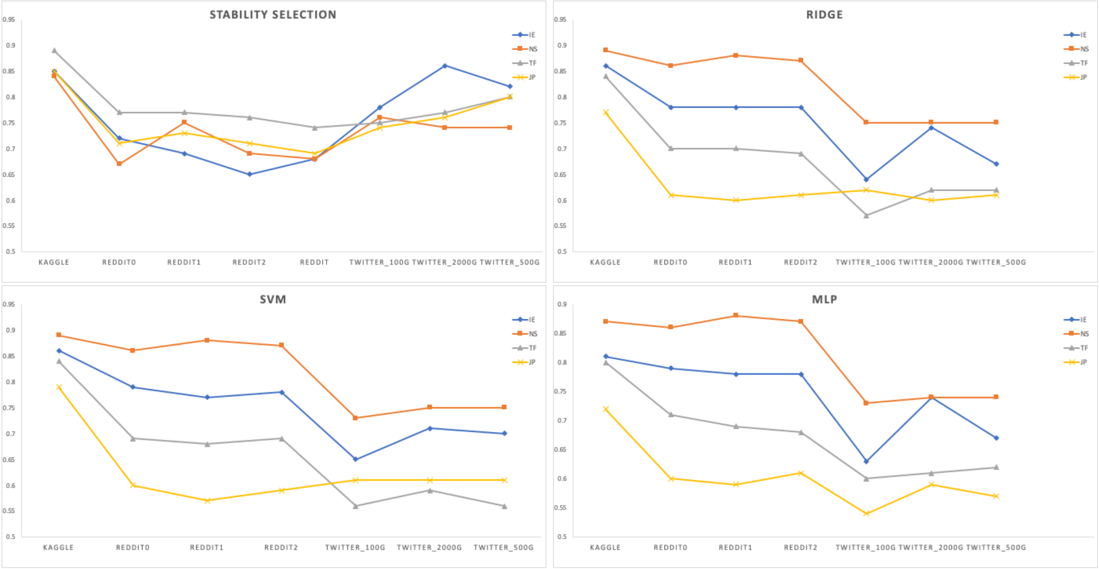

Figure 2: Accuracy of different models using 1-2 grams and TF-logIDF on MBTI dimensions

and T–F exhibit lower accuracies, which are close to the baseline. This is consistent with the previous literature in that word usage usually has reliable predictions along INTROVERT-–EXTRAVERT and SENSATION–INTUITION scales (Plank and Hovy, 2015; Kumar and Gavrilova, 2019), while showing worse performance on JUDGING–PERCEIVING and THINKING–FEELING.

Second, we observe that the performance of ridge regression, SVMs, and MLP are fairly consistent on different datasets. They perform better with the Reddit datasets, in comparison with the last three Twitter datasets. This may be due to the fact that the Reddit datasets have more samples (3,000 for reddit0/1/2 and 9,000 for reddit), while the Twitter datasets only have 1,500 samples. Additionally, the Reddit datasets have more words for each record. The high level of performance on the Kaggle Personality Cafe dataset (with its 8,600 samples) also accords with this hypothesis.

Overall, through our experiments on different datasets, we find that applying linear models on n-gram features consistently obtain fairly reliable predictions on at least two dimensions of MBTI, namely I–E and N–S. We have also run correlation analyses between each of two lexicons for the same dimension, and the results shows good correlation across different datasets and models. With the consistent performance across different models, we can confidently proceed to procure more robust lexicons across different datasets and methods.

4.3 Selecting Top-Ranked Features for MBTI

For each MBTI dimension, we have ∼249 n-gram coefficient sets based on different datasets, features, weightings, and models. We select a small set of top-ranked lexical cues for each such dimension:

1. First, within each such lexicon, we normalize the coefficients using z-scores (enabling us to better compare them across different models).
2. Then, we sort the n-grams with the absolute values of their z-scores, and choose the top 75% n-grams – such that we obtain a subset X_i for each original set of features.
3. For each n-gram in X_i, we calculated the term frequency across all feature sets, as well as the average z-scores, and chose the n-grams that appear in at least 60% among all sets.
4. Eventually, only the n-grams retained after the last step as well as their average z-score serve as the final set of weighted lexical cues for the dimension under consideration.

With the above procedure and the two filtering steps, we select small sets of top-ranked 79, 27, 124, 85 n-grams for I–E, N–S, T–F, J–P. Note that N–S has much fewer words, so we adjusted the thresholds in steps 2 and 3 (grid search in the two dimensional space with a step of 0.01), eventually using (0,8, 0.58) to obtain 85 n-grams for N–S.

Table 2 shows the top individual words for each dimension. Interestingly, it reflects certain stereotypical characteristics of each personality type. For example, EXTRAVERT individuals have more positive words such as *lol*, *haha*, *surprise*, while IN-

			Datasets							
			kaggle_mbti	reddit0_mbti9k	reddit1_mbti9k	reddit2_mbti9k	reddit_mbti9k	twitter_mbti_100g	twitter_mbti_2000g	twitter_mbti_500g
I–E	Stab	1-gram	0.82	0.67	0.64	0.63	0.67	0.73	0.84	0.75
		1-2 grams	0.84	0.71	0.69	0.65	0.67	0.77	0.85	0.8
		1-2-3 grams	0.85	0.72	0.69	0.65	0.68	0.78	0.86	0.82
	Ridge	1-gram	[0.85, 0.85, 0.83]	[0.78, 0.78, 0.78]	[0.77, 0.78, 0.77]	[0.78, 0.78, 0.78]	[0.77, 0.77, 0.77]	[0.65, 0.64, 0.65]	[0.72, 0.72, 0.73]	[0.68, 0.67, 0.66]
		1-2 grams	[0.86, 0.86, 0.84]	[0.78, 0.78, 0.78]	[0.78, 0.78, 0.78]	[0.78, 0.78, 0.78]	[NA, NA, NA]	[0.64, 0.64, 0.64]	[0.74, 0.74, 0.74]	[0.67, 0.67, 0.67]
	SVM	1-gram	[0.84, 0.84, 0.84]	[0.78, 0.79, 0.78]	[0.75, 0.76, 0.77]	[0.77, 0.78, 0.77]	[0.77, 0.77, 0.76]	[0.66, 0.63, 0.61]	[0.75, 0.74, 0.63]	[0.69, 0.68, 0.69]
		1-2 grams	[0.86, 0.86, 0.85]	[0.78, 0.79, 0.78]	[0.77, 0.77, 0.77]	[0.78, 0.78, 0.78]	[NA, NA, NA]	[0.65, 0.65, 0.63]	[0.71, 0.71, 0.67]	[0.67, 0.67, 0.70]
	MLP	1-gram	[0.81, 0.81, 0.80]	[0.75, 0.78, 0.77]	[0.73, 0.77, 0.77]	[0.75, 0.78, 0.77]	[0.74, 0.73, 0.74]	[0.62, 0.62, 0.62]	[0.70, 0.69, 0.70]	[0.63, 0.65, 0.64]
		1-2 grams	[0.81, 0.80, 0.80]	[0.79, 0.78, 0.78]	[0.78, 0.78, 0.78]	[0.78, 0.78, 0.78]	[NA, NA, NA]	[0.62, 0.62, 0.63]	[0.74, 0.74, 0.72]	[0.66, 0.67, 0.65]
T-F	Stab	1-gram	0.87	0.75	0.74	0.74	0.74	0.69	0.76	0.72
		1-2 grams	0.88	0.77	0.76	0.75	0.73	0.74	0.78	0.82
		1-2-3 grams	0.89	0.77	0.77	0.76	0.74	0.75	0.77	0.80
	Ridge	1-gram	[0.83, 0.83, 0.81]	[0.68, 0.70, 0.68]	[0.67, 0.69, 0.69]	[0.69, 0.69, 0.68]	[0.71, 0.71, 0.70]	[0.57, 0.57, 0.57]	[0.64, 0.64, 0.62]	[0.6, 0.62, 0.59]
		1-2 grams	[0.84, 0.84, 0.83]	[0.70, 0.70, 0.70]	[0.67, 0.70, 0.67]	[0.68, 0.69, 0.68]	NA	[0.57, 0.57, 0.57]	[0.62, 0.62, 0.62]	[0.62, 0.62, 0.60]
	SVM	1-gram	[0.82, 0.83, 0.83]	[0.67, 0.68, 0.66]	[0.66, 0.68, 0.69]	[0.68, 0.69, 0.69]	[0.69, 0.69, 0.69]	[0.53, 0.55, 0.57]	[0.59, 0.59, 0.60]	[0.58, 0.58, 0.52]
		1-2 grams	[0.84, 0.84, 0.84]	[0.69, 0.68, 0.69]	[0.66, 0.68, 0.66]	[0.67, 0.69, 0.67]	NA	[0.56, 0.56, 0.56]	[0.59, 0.59, 0.56]	[0.56, 0.56, 0.51]
	MLP	1-gram	[0.79, 0.79, 0.78]	[0.65, 0.67, 0.66]	[0.65, 0.67, 0.66]	[0.65, 0.66, 0.67]	[0.68, 0.67, 0.68]	[0.56, 0.58, 0.57]	[0.60, 0.60, 0.60]	[0.60, 0.60, 0.59]
		1-2 grams	[0.80, 0.80, 0.78]	[0.71, 0.70, 0.71]	[0.68, 0.69, 0.67]	[0.67, 0.67, 0.68]	NA	[0.60, 0.59, 0.58]	[0.60, 0.61, 0.60]	[0.62, 0.62, 0.61]
N–S	Stab	1-gram	0.82	0.75	0.74	0.62	0.68	0.60	0.68	0.68
		1-2 grams	0.84	0.65	0.66	0.70	0.69	0.68	0.77	0.72
		1-2-3 grams	0.84	0.67	0.75	0.69	0.68	0.76	0.74	0.74
	Ridge	1-gram	[0.86, 0.86, 0.87]	[0.86, 0.86, 0.86]	[0.88, 0.88, 0.88]	[0.87, 0.87, 0.87]	[0.88, 0.88, 0.88]	[0.75, 0.75, 0.75]	[0.75, 0.75, 0.75]	[0.75, 0.75, 0.75]
		1-2 grams	[0.89, 0.89, 0.87]	[0.86, 0.86, 0.86]	[0.88, 0.88, 0.88]	[0.87, 0.87, 0.87]	[NA, NA, NA]	[0.75, 0.75, 0.75]	[0.75, 0.75, 0.75]	[0.75, 0.75, 0.75]
	SVM	1-gram	[0.89, 0.89, 0.88]	[0.86, 0.86, 0.86]	[0.88, 0.88, 0.88]	[0.87, 0.87, 0.87]	[0.88, 0.87, 0.88]	[0.73, 0.73, 0.74]	[0.75, 0.75, 0.74]	[0.74, 0.74, 0.74]
		1-2 grams	[0.89, 0.89, 0.89]	[0.86, 0.86, 0.86]	[0.88, 0.88, 0.88]	[0.87, 0.87, 0.87]	[NA, NA, NA]	[0.73, 0.73, 0.72]	[0.75, 0.75, 0.74]	[0.75, 0.75, 0.74]
	MLP	1-gram	[0.86, 0.86, 0.85]	[0.84, 0.86, 0.86]	[0.87, 0.88, 0.88]	[0.86, 0.87, 0.87]	[0.86, 0.87, 0.86]	[0.73, 0.75, 0.72]	[0.74, 0.73, 0.73]	[0.74, 0.74, 0.75]
		1-2 grams	[0.87, 0.87, 0.85]	[0.86, 0.86, 0.86]	[0.88, 0.88, 0.88]	[0.87, 0.87, 0.87]	[NA, NA, NA]	[0.73, 0.73, 0.72]	[0.74, 0.74, 0.74]	[0.74, 0.74, 0.73]
J–P	Stab	1-gram	0.81	0.66	0.68	0.68	0.69	0.70	0.77	0.74
		1-2 grams	0.84	0.71	0.72	0.71	0.69	0.72	0.79	0.78
		1-2-3 grams	0.85	0.71	0.73	0.71	0.69	0.74	0.76	0.80
	Ridge	1-gram	[0.76, 0.76, 0.72]	[0.61, 0.61, 0.61]	[0.58, 0.58, 0.57]	[0.60, 0.60, 0.61]	[0.63, 0.63, 0.62]	[0.60, 0.60, 0.62]	[0.61, 0.61, 0.61]	[0.60, 0.60, 0.60]
		1-2 grams	[0.77, 0.77, 0.74]	[0.61, 0.61, 0.61]	[0.59, 0.60, 0.59]	[0.58, 0.61, 0.58]	NA	[0.62, 0.62, 0.61]	[0.60, 0.60, 0.60]	[0.61, 0.61, 0.61]
	SVM	1-gram	[0.77, 0.77, 0.77]	[0.59, 0.60, 0.58]	[0.57, 0.56, 0.55]	[0.59, 0.55, 0.58]	[0.59, 0.58, 0.58]	[0.58, 0.58, 0.59]	[0.63, 0.61, 0.55]	[0.59, 0.60, 0.57]
		1-2 grams	[0.79, 0.79, 0.78]	[0.60, 0.59, 0.60]	[0.57, 0.57, 0.57]	[0.59, 0.58, 0.59]	NA	[0.61, 0.61, 0.59]	[0.61, 0.61, 0.57]	[0.59, 0.59, 0.61]
	MLP	1-gram	[0.69, 0.70, 0.69]	[0.55, 0.60, 0.59]	[0.56, 0.53, 0.54]	[0.61, 0.60, 0.60]	[0.57, 0.57, 0.58]	[0.50, 0.52, 0.56]	[0.56, 0.57, 0.57]	[0.55, 0.54, 0.53]
		1-2 grams	[0.72, 0.72, 0.72]	[0.60, 0.59, 0.59]	[0.59, 0.57, 0.58]	[0.59, 0.61, 0.59]	NA	[0.54, 0.53, 0.53]	[0.59, 0.59, 0.57]	[0.57, 0.57, 0.57]

Table 1: Model accuracies on classification distinguishing I–E, T–F, N–S, and J–P, across different datasets using different features and weightings

TROVERTS use words expressing uncertainty such as *awkward*, *probably*, *introvert*. SENSING individuals focus on physical reality, while INTUITIVE individuals are driven by thoughts. Accordingly, the top words for S are concrete, such as *soccer*, *jeans*, *cards*, while for N they are more abstract, such as *writing*, *science*, *proof*. For F–T, the FEELING type has more adjectives describing feelings, e.g., *wonderful*, *incredible*, *adorable*, *beautiful*, while the THINKING type has words such as *suppose*, *tastes*, *fix*. For J–P, the words also reflect common stereotypes: *career*, *passion*, *management*, *husband* shows JUDGING individuals are more plan, work, and family oriented. Finally, the PERCEIVING type appears to use more words expressing feelings, such as *sigh*, *jealous*, *wtf*.

5 Correlation Analyses

Given the aggregated weighted lexical cues induced for each MBTI dimension, we seek to assess correlations with extrinsic lexical data covering a series of different phenomena.

5.1 Comparing Personality Models

The first interesting and straight-forward comparison is between MBTI and Big-5.

First, we applied the same experiments on the YouTube dataset by Biel et al. (2013), so as to induce similar Big-5 signals, denoted as YouTube-B2013. Then, we ran Pearson correlation analyses comparing the two. We also compared our MBTI data with a well-established YouTube lexicon from Schwartz et al. (2013b), denoted as YouTube-S2013. The correlation results are given in Table 3. The analysis shows significant correlations between I–E and four dimensions of Big-5. J–P shows strong correlations with Agreeableness, Consiousness, Extraversion, Openness. T–F has a strong correlation with Agreeableness. Compared to Tobacyk et al. (2008), we have found more correlations between the two scales. In the personality literature in psychology, strong correlations have been found between Big-5 and MBTI. Most of the correlations found here can find support in psychology. The values given in brackets denote results that accord with significant correlations found in the psychology literature (Furnham, 1996).

The correlation between MBTI lexicons and our induced Big-5 lexicon (YouTube-B2013) is found to be much weaker. This is because this Big-5 lexicon is only based on one dataset (Biel et al., 2013), and that dataset has only around 400 samples.

I	E	S	N	F	T	P	J
gym	surprise	soccer	writing	wonderful	usa	shit	passion
probably	lol	husband	mode	men	tastes	fuck	crazy
introvert	ppl	jeans	science	feeling	bullshit	training	months
awkward	wine	para	moon	incredible	suppose	sigh	series
friends	hey	cards	shit	anxiety	money	rain	career
stars	bar	wife	proof	feel	pay	summer	yes
party	months	workout	write	adorable	science	ahead	management
tonight	meeting	apple	beer	heart	cost	jealous	pull
dragon	dat	episodes	folks	beautiful	map	wtf	husband
looks	haha	lazy	thx	haha	fix	movie	degrees

Table 2: Top words (unigrams) for each dimension in the MBTI lexicons

		I–E	J–P	N–S	T–F
Extraversion	YouTube-S2013	[0.71∗∗]	[−0.58∗∗]	0.65	[−0.06]
	YouTube-B2013	−0.14	−0.24	0.17	0.29∗
Agreeableness	YouTube-S2013	0.52∗	−0.77∗∗	0.19	[0.84∗∗]
	YouTube-B2013	0.03	−0.38	0.24	0.23
Openness	YouTube-S2013	[0.71∗∗]	[−0.71∗∗]	[0.24]	0.27
	YouTube-B2013	0.08	−0.47∗∗	0.28	−0.07
Conscientiousness	YouTube-S2013	−0.19	[−0.59∗∗]	0.14	[0.13]
	YouTube-B2013	−0.09	−0.19	−0.41∗	−0.01
Emotionism	YouTube-S2013	[0.75∗∗]	−0.07	−0.15	0.23
	YouTube-B2013	0.11	0.09	0.18	−0.20

$\ast: p < .05 \quad \ast\ast: p < .01$

Table 3: Correlation between our MBTI lexicons and two YouTube Big-5 lexicons (Furnham, 1996)

	I–E	J–P	N–S	T–F
Anger	−0.17	0.26	−0.23	−0.26
Fear	−0.26	0.41∗	−0.18	−0.08
Joy	0.20	−0.15	−0.07	0.28∗∗
Sadness	−0.12	0.41∗	0.10	−0.15
Arousal	0.22∗∗	0.00	−0.10	0.01
Dominance	0.26∗∗	−0.15∗∗	−0.27∗∗	0.08
Valence	0.15∗∗	0.02	0.01	0.28∗∗
Sentiment (Ding et al., 2008)	0.25∗∗	−0.13∗	−0.25∗∗	0.25∗∗
Sentiment (NRC)	0.16∗∗	−0.21∗∗	0.02	0.27∗∗
Sentiment (Twitter)	0.16∗	−0.24∗∗	−0.10	0.43∗∗
Sentiment (VADER)	0.36∗∗	−0.41∗∗	−0.17	0.42∗∗

$\ast: p < .05 \quad \ast\ast: p < .01$

Table 4: Correlation between MBTI lexicons and emotion and sentiment

	I–E	J–P	N–S	T–F
Age	−0.03	−0.12∗∗	0.03	−0.05
Gender	−0.09∗	−0.10∗	0.13∗	0.23∗∗
Gender (1-2-3-grams)	0.09	−0.68∗∗	0.72∗∗	0.73∗∗
Gender (age 13–18)	−0.31	0.52∗∗	0.24	0.05
Gender (age 19–22)	0.36	0.33	−0.56	0.09
Gender (age 23–29)	0.46	−0.43∗	−0.12	−0.21
Gender (age 30+)	0.06	−0.11	0.43	0.17

$\ast: p < .05 \quad \ast\ast: p < .01$

Table 5: Correlation between MBTI signals with demographic signals

5.2 Correlation with Emotion and Sentiment

Personality influences an individual's emotions, opinions, and behaviours. This motivates us to study the relationship between MBTI cues and other psychological lexicons, such as sentiment and emotion ones. Little work has been conducted on the correlation between personality and emotion, but the definitions of the MBTI dimensions suggest a possible connection.

We retrieved several emotion and sentiment lexicons and computed their correlation with our data in Table 4. The first four lexicons (Mohammad, 2017) focus on four basic affective categories from the Plutchik model (Plutchik, 1980): anger, fear, joy, and sadness. Only J–P has a positive correlation with fear and sadness, while T–F has a positive correlation with joy. This suggests that the FEELING type tends to use more joy-related words, while PERCEIVING individuals tend to use more fear and sadness related words. The next three lexicons (Mohammad, 2018) are based on the PAD model (Russell and Mehrabian, 1977), which conceptualizes emotion along three dimensional axes – arousal, dominance, and valence. We observe that I–E has significant positive correlations with all three dimensions, reflecting that EXTRAVERTs focus more on outside stimuli, and tend to have more emotional reactions. J–P and N–S are negatively correlated with dominance, meaning that the JUDGING and INTUITION types exhibit higher dominance – i.e., more stability. These two types perhaps also tend to analyze and give solutions, while SENSING and PERCEIVING individuals exhibit stronger feelings, which leads to lower dominance. T–F shows positive correlation with valence – suggesting that the FEELING type may have more emotional reactions.

As personality affects an individual's way of writing and talking, we further hypothesize that people with the same personality may tend to use expressions with similar sentiment. Thus, we also compare our MBTI data with three sentiment lexicons: Ding et al. (2008), NRC (Mohammad et al., 2013), Twitter (Kiritchenko et al., 2014), and VADER (Hutto and Gilbert, 2014). Both I–E and T–F show positive correlations with sentiment, while J–P shows a negative correlation. This suggests that EXTRAVERT, FEELING, and JUDGING types may tend to have more positive sentiment. Lin et al. (2017) developed a Big-5 personality-based sentiment classifier and argue that it performs better than an ordinary sentiment classifier, providing further corroboration for a potential correlation between personality and sentiment analysis.

5.3 Correlation with Demographic Signals

Twitter and Reddit have large user bases, including different gender and age groups. We study potential correlations between these two demographic features and the MBTI dimensions, relying on the age and gender lexicons by Sap et al. (2014), as well as the age-specific gender lexicons by Schwartz et al. (2013a). The correlations are reported in Table 5. Only J–P has a negative correlation with the age lexicons. It appears plausible that older individuals might rely more on judgement than perception.

The two general gender lexicons *gender* and *gender (1-2-3-grams)* show that I–E has a slightly negative correlation with gender, J–P has a strong correlation, while N–S and T–F have moderate positive correlations with gender. Note that for the gender lexicons, male is here treated as negative, and female as positive, and the two lexicons fail to account for other gender identities. The correlation analysis is consistent with the stereotypes that female users tend to use more words about feelings (F) and are more sensible (S) in general. However, it is interesting to see that female users are found to be more JUDGING. When we control for age group, most correlations between gender and personality disappear, and only J–P showed strong positive correlation with gender in the age group 13 to 18, and negative correlation in age group 23 to 29.

6 Conclusion

We have inferred personality predictive lexical signals, i.e., words and n-grams along with their weights, for each MBTI dimension. The data is induced based on several diverse MBTI datasets, using a variety of feature sets, weighting schemes, and learning algorithms. Our focus here is on identifying correlations with other kinds of cues, including Big-5 data, as well as emotion, sentiment, and gender-predictive lexicons. We show that naturally occurring text harbors subtle cues exhibiting correlations that largely accord with findings from psychology on self-reported personality correlations. This provides further evidence for the validity of drawing on such naturally occurring data for automated lexical cue induction.

Ethical Statement

It is important to keep in mind that all results presented here are highly dependent on the characteristics of the respective datasets and on the lexicon induction methodology. As shown in Section 4.1, different datasets provide data from different sources, leading to biases both in the kinds of textual content they provide and in the label distributions. Additionally, using automated predictors for lexicon induction tends to lead to signals reflective of particularly stereotypical cues, and linear models are unable to account for the particular context of a particular word mention. Thus, the particular word-level correlations observed in this study do not entail that such correlations also hold among people exhibiting a particular trait. Last but not least, mere correlations such as those considered in this paper do not license conclusions about particular individuals or groups of individuals, and any studies attempting to predict the personality of individuals or groups of individuals would need to consider a large number of very serious ethical and privacy concerns.

References

Pierre-Hadrien Arnoux, Anbang Xu, Neil Boyette, Jalal Mahmud, Rama Akkiraju, and Vibha Sinha. 2017. 25 tweets to know you: A new model to predict personality with social media. In *Eleventh International AAAI Conference on Web and Social Media*.

Michael C Ashton, Kibeom Lee, and Lewis R Goldberg. 2004a. A hierarchical analysis of 1,710 english personality-descriptive adjectives. *Journal of Personality and Social Psychology*, 87(5):707.

Michael C Ashton, Kibeom Lee, Marco Perugini, Piotr Szarota, Reinout E De Vries, Lisa Di Blas, Kathleen Boies, and Boele De Raad. 2004b. A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *Journal of personality and social psychology*, 86(2):356.

Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. 2013. Hi YouTube! Personality impressions and verbal content in social video. In *Proceedings of the 15th ACM International conference on Multimodal Interaction*, pages 119–126.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international onference on Web Search and Data Mining*, pages 231–240.

Adrian Furnham. 1996. The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Personality and Individual Differences*, 21(2):303–307.

Matej Gjurković and Jan Šnajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 87–97, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Valerie Priscilla Goby. 2006. Personality and online/offline choices: MBTI profiles and favored communication modes in a Singapore study. *Cyberpsychology & behavior*, 9(1):5–13.

Lewis R Goldberg. 1990. An alternative "description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

Jacob B Hirsh and Jordan B Peterson. 2009. Personality and language use in self-narratives. *Journal of research in personality*, 43(3):524–527.

C.J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. ICWSM-14*.

SM Mazharul Islam, Xin Dong, and Gerard de Melo. 2020. Domain-specific sentiment lexicons induced from labeled documents. In *Proceedings of COLING 2020*.

Margaret L Kern, Johannes C Eichstaedt, H Andrew Schwartz, Lukasz Dziurzynski, Lyle H Ungar, David J Stillwell, Michal Kosinski, Stephanie M Ramones, and Martin EP Seligman. 2014. The online social self: An open vocabulary approach to personality. *Assessment*, 21(2):158–169.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Tugba Kulahcioglu and Gerard de Melo. 2018. FontLex: A typographical lexicon based on affective associations. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, pages 62–69, Paris, France. European Language Resources Association (ELRA).

KN Pavan Kumar and Marina L Gavrilova. 2019. Personality traits classification on twitter. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE.

Junjie Lin, Wenji Mao, and Daniel D Zeng. 2017. Personality-based refinement for sentiment classification in microblog. *Knowledge-Based Systems*, 132:204–214.

Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Saif M Mohammad. 2017. Word affect intensities. *arXiv preprint arXiv:1704.08798*.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Isabel Briggs Myers, Mary H McCaulley, and Robert Most. 1985. *MBTI Manual, a guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33. Elsevier.

Shahab Raji and Gerard de Melo. 2020. What sparks joy: The AffectVec emotion database. In *Proceedings of The Web Conference 2020*, pages 2991–2997, New York, NY, USA. ACM.

Shahab Raji and Gerard de Melo. 2021. Guilt by association: Emotion intensities in lexical representations. *ArXiv*, 2104.08679.

James A. Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin E P Seligman, and Lyle H Ungar. 2013a. Characterizing geographic variation in well-being using tweets. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013b. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Ryne A Sherman, Christopher S Nave, and David C Funder. 2013. Situational construal is related to personality and gender. *Journal of Research in Personality*, 47(1):1–14.

Farhad Bin Siddique, Dario Bertero, and Pascale Fung. 2019. GlobalTrait: Personality alignment of multilingual word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7015–7022.

Jerome J Tobacyk, Mary M Livingston, and James E Robbins. 2008. Relationships between Myers-Briggs Type Indicator measure of psychological type and NEO measure of big five personality factors in polish university students: A preliminary cross-cultural comparison. *Psychological reports*, 103(2):588–590.

G. Richard Tucker. 1968. Judging personality from language usage: a filipino example. *Philippine Sociological Review*, 16(1/2):30–39.

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