DATA-DRIVEN DEVELOPMENT OF PERSONALITY PREDICTIVE LEXICA FROM SOCIAL MEDIA

By

XIAOLI HE

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Computer Science

Written under the direction of

Gerard de Melo

And approved by

_____

_____

_____

New Brunswick, New Jersey

May, 2020

ABSTRACT OF THE THESIS

Data-Driven Development of Personality Predictive Lexica from Social Media

by XIAOLI HE

Thesis Director:

Gerard de Melo

Automatic personality prediction is getting more popular because it is convenient and reliable. Lexicon-based analysis has been successful in the fields of sentiment analysis and emotion. Many studies have used linear models for personality prediction, which suggests that we can also use lexical-based analysis for personality prediction. In the current study, we developed weighted word lexicons (words and scores) on each dimension of MBTI personality. The lexicons are built based on eight MBTI datasets, different features (unigram, 1-2 grams, 1-2-3 grams) and weightings (TF, TF-IDF, TF-logIDF), and different supervised learning models. Then we ran correlation analysis between our MBTI lexicons and other existing lexicons, such as Big-5, emotion, sentiment, age, gender. The correlation analysis shows interesting and reasonable correlation between different personality dimensions and other psychological traits, and it also provides evidence for the robustness of our lexicons.

Acknowledgements

I would like to express my great appreciation to Dr. Gerard de Melo for his great mentorship throughout the whole project. I would like to thank Dr. Yongfeng Zhang and Dr. Manish Singh, for being my committee members in this special time. I would also like to show my special appreciations for the authors of the datasets I have used in my thesis. Thank you for sharing the datasets. I really appreciate and admire your efforts for creating and maintaining the datasets. At last, I would like to thank Dr. James Abello for giving me the chance to pursuing my secondary degree in Computer Science, and thank Dena Orkin for all the help with enrollment and graduation.

Dedication

This dissertation is dedicated to my supportive family and my dear boyfriend in China, without whom I cannot accomplish so much. I also want to show my sincere thanks to my dear roommates Yandi, Mengxue and Auntie Hao. Thank you for being my second family in the U.S.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Personality is an individual's characteristic patterns of thinking, feeling, and behaving (Sherman, Nave, & Funder, 2013). Studies have shown that personality influences an individual's language usage (Tucker, 1968; Hirsh & Peterson, 2009; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013). In other words, language contains rich information about an individual's personality. Since an individual's personality is quite stable across a relatively long period of time, the relation between personality and language usage should be consistent and analyzable.

In traditional psychology studies, personality is usually measured by a standard questionnaire that measures different aspects of personality. Different models have defined different traits (sub-dimensions) of personality. There are mainly two types of personality scales: Big-5 and MBTI.

Myers-Briggs Type Indicator model (MBTI) (Myers, McCaulley, & Most, 1985) has the following four dimensions:

1. Introversion-Extraversion (I-E): where a person focuses his/her attention;
2. INtuition-Sensing (N-S): the way a person takes in information;
3. Thinking-Feeling (T-F): how a person makes decisions;
4. Judging-Perceiving (J-P): how a person deals with the world.

In contrast, Big-5 (Goldberg, 1990) has five dimensions:

1. Extraversion (extroversion): indicates how outgoing and social a person is;
2. Agreeableness: indicates how warm, friendly and tactful a person is;

3. Openness: indicates how open-minded and authority-challenging a person is;

4. Conscientiousness: indicates how self-disciplined and organized a person is;

5. Neuroticism (emotionism): indicates a person's ability to remain stable and balanced.

Normally Big-5 provides continuous scores on the five dimensions, while MBTI provides binary labels for the above four dimensions, such as ESTJ. Therefore, Big-5 can be treated as a regression problem and MBTI as a binary classification problem on each of their dimensions.

The accuracy of the traditional ways of personality measurement is quite high and stable. However, the shortcoming is that the professional scales contain long lists of questions, and also it requires an individual to fill out the questionnaire explicitly. Is there a faster alternative?

In the 21st century, online social media, such as Twitter, Facebook, Reddit, has plays an important role in daily life. People tend to post their daily lives, thoughts, emotions, opinions on different social media platforms. These contents may provide rich information about an individual's personality. In other words, social media provides a tremendous opportunity for automatic personality prediction.

Many studies have shown that social media data can give reliable predictions on personality. Those studies have focused on different social media platforms and personality scales. Different models have been developed, from simple Logistic Regression/Linear Regression (Arnoux et al., 2017), support vector machine (Biel, Tsiminaki, Dines, & Gatica-Perez, 2013; Kumar & Gavrilova, 2019), to more complex models such as stability selection (Plank & Hovy, 2015), Gaussian process models (Arnoux et al., 2017) and ensemble methods using different classifiers/regressors (Kumar & Gavrilova, 2019). The studies have also used different features, such as ngrams (unigram, 1-2 grams, 1-2-3 grams) (Yarkoni, 2010; Biel et al., 2013; Plank & Hovy, 2015), word embeddings (Arnoux et al., 2017; Siddique, Bertero, & Fung, 2019). For the studies using ngrams as features, they have applied different weightings, such as TF-IDF (Biel et al., 2013; Siddique et al., 2019), or no weightings (Yarkoni, 2010; Kern et al., 2014; Plank & Hovy, 2015). Other NLP studies have also mentioned using relative word frequency, such as the age and gender study by Sap et al. (2014).

Despite the richness of the above studies, automatic personality prediction is still a challenging problem, and there are still some issues unsolved.

First, due to privacy and high labeling costs, the number of publicly available labeled datasets are limited, and the sample size of each dataset varies, and in general is relatively small (especially when compared with the high dimentionality of ngram features). Also, some datasets only have limited number of sentences in each sample. The limitation of available datasets makes it difficult to generalize results from individual studies.

Second, it is hard to compare results from different studies. On one hand, in the field of natural language processing, some studies use MBTI and some use Big-5, however, few studies have tried to compare between these two different models, especially on the level of individual dimensions. But in the field of personality psychology, studies have shown clear correlations between self-reported MBTI and Big-5. For example, MBTI-IE correlates with Big-5 Extraversion, MBTI-SN and JP correlate with Big-5 Openness, and MBTI JP also correlates with Big-5 Consciousness (Tobacyk, Livingston, & Robbins, 2008). On the other hand, even within the same personality scale, it is hard to compare results from different studies because they used different datasets, features and models.

Third, few of the existing studies have focused on the contributions of individual words on personality prediction. Most studies have focused on improving performance of personality prediction on a given dataset using different methods and features. 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, Francis, & Booth, 2001). But this may have masked the contribution of individual words in the context of open vocabulary scenario.

Lexicon-based analysis has been quite successful in areas such as sentiment analysis and emotion analysis (Ding, Liu, & Yu, 2008; Mohammad, Kiritchenko, & Zhu, 2013; Kiritchenko, Zhu, & Mohammad, 2014; Zhu, Kiritchenko, & Mohammad, 2014). With this approach, a dictionary of words (or bag of words) is generated, with a positive or negative value assigned to each word, indicating the predicted power or correlation strength between the word and the specific domain. In traditional personality psychology, psychologists have developed closed-book vocabularies by self-rating on personality-trait adjectives or verbs (Ashton, Lee, &

Goldberg, 2004; Ashton, Lee, Perugini, et al., 2004). Both aspects imply that the lexicon-based analysis should also be applicable to automatic personality prediction using social media data.

Given the limitations and gaps in the existing studies, the main goal of the current study is to develop a predictive lexicon for each dimension of personality scales using social media data. We gathered four MBTI datasets from Twitter, four derived MBTI datasets from Reddit, one Big-5 dataset from YouTube vblog. Since most of the datasets were using MBTI, we mainly focused on developing MBTI lexicons. To get a better generalization of the lexicons, we applied different methods (stability selection using Logistic regression, Penalized Ridge classification, support vector classification, multi-layer perceptron), features (unigram, 1-2 grams, 1-2-3 grams), and weightings (original, tf-idf, tf-logidf, relative frequency) across different MBTI datasets. Also, we developed Big-5 lexicons using the above approach based on the YouTube dataset. It may not be as powerful as the MBTI lexicons, but it is a good start to compare dimensions of the two different personality scales. Moreover, we applied correlation anlaysis between our MBTI lexicons and the state-of-art Big-5 lexicons, and also in other fields such as age, gender, sentiment and emotion.

Chapter 2

Datasets

2.1 Overview

We collected 8 MBTI datasets and one Big-5 dataset. The datasets are either publicly available online, or requested from the authors by email.

As shown in Figure 2.1, in the eight MBTI datasets, 'kaggle' is from Kaggle Twitter MBTI dataset, it contains 8600 samples. 'twitter_100g', 'twitter_500g', 'twitter_2000g' are from (Plank & Hovy, 2015). Each contains 1500 samples, and they differ in the number of tweets in each sample (100, 500 or 2000). 'reddit' is from (Gjurković & Šnajder, 2018), where it contains 9149 rows of comments from different Reddit authors with more than 1000 words. The original dataset 'reddit' is too large for the calculation of TF-IDF features for 1-2 grams and 1-2-3 grams, so we randomly splitted it into three smaller datasets: 'reddit0', 'reddit1', 'reddit2', and only used those datasets for some analysis (see the next chapter for more details).

Figure 2.1 shows the distribution of each dimension in different MBTI datasets. In each dimension, the first type is coded as 0, and the second type is coded as 1. For example, in 'IE', 'INTROVERT' is represented as 0, and 'EXTRAVERT' is represented as 1. Figure 2.1 shows that overall, each dataset has more INTROVERTs and THINKING types. It has been reported before that INTROVERTs prefer online communications (Goby, 2006; Plank & Hovy, 2015). More interestingly, there are some differences between Reddit users and Twitter users. Reddit has more INTUITIVEs, and Twitter has more JUDGINGs. Reddit has a little more THINKINGs than Twitter.

Figure2.2 shows the only Big-five dataset we gathered from YouTube Vblog (Biel et al., 2013). The text is the manually made transcript of a Vblog, and the Big-5 score is the impression score given by another group of subjects (not self-reported). That is also the reason

why we did not focus on this dataset in our analysis.

2.2 Data preprocessing

We only considered the language information from each dataset (tweets or Reddit comments), and the personality type. We tokenized the text data, and implemented the following preprocessing steps:

1. Change each letter to its lower case;
2. Remove tokens that: are English stop words, contain only numbers, mention one of the personality types,
3. Replace URLs , Hashtags, usernames with '@URL' , '@HASHTAG', '@USER'.

2.3 Feature Extraction

Since eventually we want to create a lexicon of words with weightings, it is natural to start with ngrams. We extracted binary ngram features for each sample, then transformed list of ngrams to a sparse 0-1 vector for each record. In particular, we extracted unigrams, 1-2 grams, 1-2-3 grams for all datasets. We set a minimum threshold for the features, and dropped the features that appear less than 1% in each dataset. For the 1-2 grams and 1-2-3 grams, we excluded tokens which have punctuations in the first or middle, such as ('!', 'I'), ('today', '!', 'I'), ('!', 'today', 'I'). We also excluded tokens that only contain numbers. Originally, we planned to use all three types of ngrams for all models, but 1-2-3 grams led to out of memory issue on my MacBook Pro with Intel Core i5, and based on the result the stability selection study, the performance of 1-2-3 grams was similar compared to unigrams and 1-2 grams. Therefore, we only used 1-2-3 grams for stability selection.

2.4 Ngram weighting

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

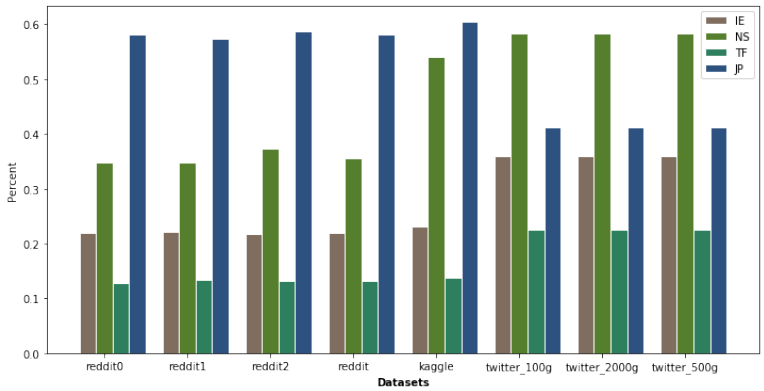

Figure 2.1: MBTI distribution on each dataset, mapping of each dimension: ’I’: 0, ’E’: 1, ’N’: 0, ’S’: 1, ’T’: 0, ’F’: 1, ’J’: 0, ’P’: 1

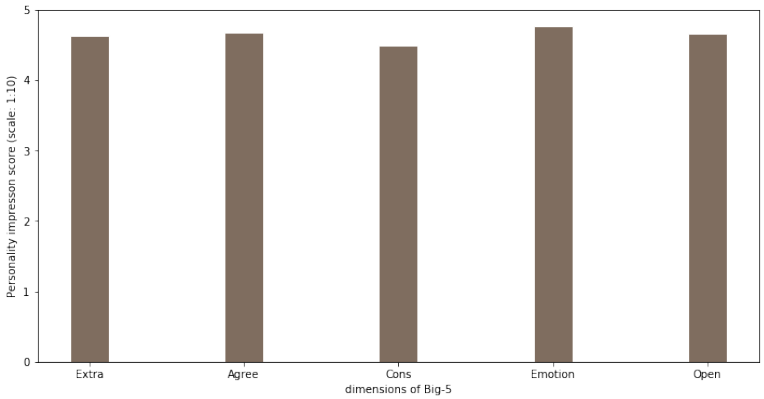

Figure 2.2: Distribution of Big-5 score on each dimention of the YouTube dataset (scale: 1-10)

1. Relative frequency, $\frac{freq\ (word,\ doc)}{freq\ (*,\ doc)}$, is actually the term frequency (TF) in a document. It considers information from the same document.

2. TF-log (IDF), is the standard defination of Term Frequency -Inverse Document Frequency (tf-idf) features. It rules out the situation where some words appear more frequently in general, which is not meaningful for personality prediction. The logarithmically scaled IDF dampens the effect of the ratio.

3. TF-IDF, is slightly different from the second one in that we did not include logarithmical scale for IDF.

Chapter 3

Generating weighted lexicon using different models

Overall, we have three types of features – unigrams, 1-2 grams, 1-2-3 grams, and four possible weightings: no weighting, relative frequency, TF-logIDF, TF-IDF, and 9 datasets. Then, in order to find a universal personality lexicon, we tried almost each combination of feature and weighting on three kinds of linear models and muti-layper perceptron as a comparison (See 3.1 for details).

The idea of using linear models is derived from Sap et al. (2014). They compared the formula of linear multivariate models $y = (\Sigma_{f \in features} w_f * x_f) + w_0$ with the formula of a weighted lexicon: $usage_{lex} = \Sigma_{word \in lex} w_{lex}(word) * \frac{freq\ (word,\ doc)}{freq\ (*,\ doc)}$. They proved that if relative term frequency is used as feature, many multivariate modeling techniques can be seen as learning a weighted lexicon plus an intercept. In other words, the weight of a word in a lexicon can be obtained by the coefficients from linear multivariate regression and classification models.

In our case, if we use weighted ngrams as features, the coefficients from the models can be seen as the weights for the lexicon. In particular, we treated each dimension of personality as a single and independent classification/regression problem. For each combination of feature and weighting, we tried three types of linear models (stability selection with Logistic regression/Lasso, penalized Ridge classifier/regressor, support vector classifier/regressor with linear kernal) and multi-layer perceptron. After getting the coefficients/weights for each method, we normalized the coefficients in different methods into their z-scores, then we were able to compare values between different methods, and get a universal MBTI lexicon for each dimension.

3.1 Different models

3.1.1 Stability selection

The first method is stability selection (Meinshausen & Bühlmann, 2010). The method works by resampling the training data and applying the same model on each resampling. The idea is that the features get selected more often are good features.

We used RandomizedLogisticRegression in sklearn 0.20.dev for MBTI datasets, and RandomizedLasso for the Big-5 dataset. We used all three features with no weightings for this model. We ran 100 resampling procedures, on each resampling, 75% of the samples were randomly chosen. After the step of stability selection, we applied Logistic regression for MBTI (linear regression for Big-5) on the selected features (ngrams), and saved their coeffiecients.

3.1.2 Penalized Ridge Classification/Regression

The second linear model is Ridge Regression (i.e, linear least squares regression with L1 regularization) for Big-5, and Ridge classification for MBTI. The classifier first converts the target values into (-1, 1) and then treats the problem as a regression task. 'L1' penalty produces sparse features, which is more suitable for our case. We splitted each dataset into training set and test set randomly with a ratio of 3:1 (we also used the same ratio for the following two models). Then we used RidgeCV and RidgeClassifierCV in sklearn.linear_model to get the coefficients for each feature and type of weightings.

For this model and the following two models, we used only unigram, 1-2 grams and three weightings. There are two reasons why we gave up on 1-2-3 grams. First, the results of stability selection show no obvious improvement from 1-2 grams to 1-2-3 grams. Second, the number of features increased only a little bit from 1-2 grams to 1-2-3 grams, but the computation costs of weightings get much worse. Given the little gain we get from the trigrams, we only tested on unigram, and 1-2 grams in the following models.

3.1.3 Penalized support vector Classification with linear kernel

Support vector machine is well-suited for high dimensionality. Considering the high dimentionality and sparsity of our feature space, we used support vector classification/regression with a linear kernel and L1 penalty in a 10-fold cross-validation setup. We hope to compare its performance with the Ridge models, to avoid potential overfitting.

3.1.4 Penalized Multi-Layer Perceptron

In the last we used a simple feedforward neural network – multi-layer perceptron, in case the data is not linearly separable.

3.2 Results of different models

The accuracies of each combination of feature, weighting and model are shown in 3.1, 3.2 , 3.3,3.4,for each dimension of MBTI. The three numbers in the same cell represent results from the three different weightings: relative frequency, tf-logidf, tf-idf.

The table shows that 1) the accuracy for different weightings are quite similar; 2) accuracies consistently increase from unigram to 1-2 grams, but not so much with 1-2-3 grams. Therefore, to make the results more intuitive, we plot the results using 1-2 ngrams weighted by tf-logidf for each method in Figure 3.1. Each subfigure shows the results from one model. In each subfigure, different lines show the results for different personality dimensions.

Figure 3.1 conveys the following two messages: first, it shows that for the three linear models with penalty, dimension NS has the highest accuracy, IE has the second highest accuracy, while JP and TF have lower accuracies which are close to the baseline. This is consistent with previous literature that word usage usually has reliable predictions on INTROVERT–EXTROVERT and SENSATION-INTUITION (Plank & Hovy, 2015; Kumar & Gavrilova, 2019), and worse performance on JUDGING-PERCEIVING and THINKING-FEELING.

Second, it shows that the performance of Ridge, SVM and MLP are consistently similar on different datasets. They perform better with the Reddit datasets, compared with the last three Twitter datasets. It may due to the fact that Reddit datasets have more samples (3000 for reddit0, 1,2 and 9000 for reddit), and the Twitter datasets only have 1500 samples. Also, Reddit

Figure 3.1: Accuracy of different models using 1-2 grams and tf-logidf on MBTI dimensions

datasets have more words for each record. The high performance of dataset kaggle also proves this (it has 8600 samples).

Overall, through our experiments on different datasets, it has shows that applying linear models on ngram features have consistenly reliable predictions on at least two dimensions of MBTI, which are IE and NS. We have also run correlation analysis between each two files of the same dimensions, and the results shows good correlations across different datasets and models (see supplementary tables). With the consistent performance across different models, we are confident to get the lexicons across different datasets and methods.

Table 3.1: Model accuracies on dimension of IE across different datasets using different features and weightings

			Datasets							
			kaggle_mbti	reddit0_mbti9k	reddit1_mbti9k	reddit2_mbti9k	reddit_mbti9k	twitter_mbti_100g	twitter_mbti_2000g	twitter_mbti_500g
Models	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, 0.7]
	MLP	1-gram	[0.81, 0.81, 0.8]	[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]	[0.7, 0.69, 0.7]	[0.63, 0.65, 0.64]
		1-2 grams	[0.81, 0.8, 0.8]	[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]

[hp]

Table 3.2: Model accuracies on dimension of TF across different datasets using different features and weightings

			Datasets							
			kaggle_mbti	reddit0_mbti9k	reddit1_mbti9k	reddit2_mbti9k	reddit_mbti9k	twitter_mbti_100g	twitter_mbti_2000g	twitter_mbti_500g
Models	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	0.8
	Ridge	1-gram	[0.83, 0.83, 0.81]	[0.68, 0.7, 0.68]	[0.67, 0.69, 0.69]	[0.69, 0.69, 0.68]	[0.71, 0.71, 0.7]	[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]	[0.7, 0.7, 0.7]	[0.67, 0.7, 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, 0.6]
	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, 0.6]	[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]	[0.6, 0.6, 0.6]	[0.6, 0.6, 0.59]
		1-2 grams	[0.8, 0.8, 0.78]	[0.71, 0.7, 0.71]	[0.68, 0.69, 0.67]	[0.67, 0.67, 0.68]	NA	[0.6, 0.59, 0.58]	[0.6, 0.61, 0.6]	[0.62, 0.62, 0.61]

Table 3.3: Model accuracies on dimension of NS across different datasets using different features and weightings

			Datasets							
			kaggle_mbti	reddit0_mbti9k	reddit1_mbti9k	reddit2_mbti9k	reddit_mbti9k	twitter_mbti_100g	twitter_mbti_2000g	twitter_mbti_500g
Models	Stab	1-gram	0.82	0.75	0.74	0.62	0.68	0.6	0.68	0.68
		1-2 grams	0.84	0.65	0.66	0.7	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]

Table 3.4: Model accuracies on dimension of JP across different datasets using different features and weightings

			Datasets							
			kaggle_mbti	reddit0_mbti9k	reddit1_mbti9k	reddit2_mbti9k	reddit_mbti9k	twitter_mbti_100g	twitter_mbti_2000g	twitter_mbti_500g
Models	Stab	1-gram	0.81	0.66	0.68	0.68	0.69	0.7	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	0.8
	Ridge	1-gram	[0.76, 0.76, 0.72]	[0.61, 0.61, 0.61]	[0.58, 0.58, 0.57]	[0.6, 0.6, 0.61]	[0.63, 0.63, 0.62]	[0.6, 0.6, 0.62]	[0.61, 0.61, 0.61]	[0.6, 0.6, 0.6]
		1-2 grams	[0.77, 0.77, 0.74]	[0.61, 0.61, 0.61]	[0.59, 0.6, 0.59]	[0.58, 0.61, 0.58]	NA	[0.62, 0.62, 0.61]	[0.6, 0.6, 0.6]	[0.61, 0.61, 0.61]
	SVM	1-gram	[0.77, 0.77, 0.77]	[0.59, 0.6, 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, 0.6, 0.57]
		1-2 grams	[0.79, 0.79, 0.78]	[0.6, 0.59, 0.6]	[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, 0.7, 0.69]	[0.55, 0.6, 0.59]	[0.56, 0.53, 0.54]	[0.61, 0.6, 0.6]	[0.57, 0.57, 0.58]	[0.5, 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]	[0.6, 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]

3.3 Generating the weighted lexicon for MBTI from different methods

For each dimension of MBTI, we have around 249 ngram-coefficient files which used different datasets, features, weightings and models.

To get the lexicon for each dimension of MBTI:

1. First, for each file, we converted the coefficients into z-score (so it is more reasonable to compare across different models).

2. Then, we sorted the ngrams with the absolute values of their z-scores, and chose the top 75% ngrams – so we have a set X_i for each file.

3. For each ngram in X_i, we calculated its term frequency in all files, as well as their average z-scores, and chose the ngrams that appear at least 60% in the total 249 files.

4. Eventually, the ngrams 'survived' in the last step, as well as their average z-score is the lexicon for that dimension.

With the above procedure and the two thresholds in step 2 and 3, we get 79, 27, 124, 85 ngrams for IE, NS, TF, JP. Note that NS has much fewer words, and we played with the two thresholds (grid search in the two dimensional space with a step of 0.01), and eventually used (0,8, 0.58) and get 85 ngrams for NS. Table 3.5 shows the top words for each dimension. Interestingly, it shows the stereo-characteastic of each personality type. For example, EXTRAVERTs has more positive words such as *lol, haha, surprise*, while INTROVERTs has words expressing uncertainty like *awkward, probably, introvert*. SENSING types are more concrete while INTUITIVEs are more abstract. Therefore, the top words for S are abstract like *writing, science, proof*, while for N it is more concrete, such as *husband, wife, apple*. For F/T, FEELING type has more adjectives describing feelings, like *wonderful, incredible, adorable, beautiful*, while THINKING type has words like *suppose, tastes, fix*. For J/P, the words also reflect the common stereotypes: *career, passion, management, husband* shows JUDGINGs are more plan, work and family oriented. And PERCEIVINGs have used more words expressing feelings such as *sigh, jealous, wtf*.

Table 3.5: Top words for each dimension in the MBTI lexicons

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

Chapter 4

Correlation analysis

With our MBTI lexicons on each dimension, it would be interesting to see its correlations with lexicons from other areas. It also provides evidence for the robusness of our lexicons.

4.1 Correlation analysis between MBTI and Big-5

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

First, we applied the same experiments on the YouTube dataset (Biel et al., 2013), and generated our own Big-5 lexicons. Then we ran a Pearson correlation analysis on the two lexicons. We also compared our MBTI lexicons with a well established YouTube lexicons from Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al. (2013). The correlation results are shown in Table 4.1. The analysis shows significant correlations between IE and four dimensions of big-5: Agreeableness, Emotionism (Neuroticism), Extraversion and Openness. JP has strong correlations with Agreeableness, Consiousness, Extraversion, Openness. TF has a strong correlation with Agreeableness. Compared to Tobacyk et al. (2008), we have found more correlations between the two scales. In personality psychology literature, there are strong correlations between Big-5 and MBTI. Most of the correlations we found here can find support from the psychology society. The values in parentheses are the significant correlations found in the psychology literature (Furnham, 1996).

Besides, it can been seen from the table that the correlation between MBTI lexicons and our own Big-5 lexicon is much weaker. It makes sense since it is only based on one dataset, and the dataset has only around 400 samples.

Table 4.1: Correlation between our MBTI lexicons and two YouTube lexicons

		IE	JP	NS	TF
Agr	liter	0.52*	-0.77**	0.19	[0.84**]
	ours	0.03	-0.38*	0.24	0.23
Cons	liter	-0.19	[-0.59**]	0.14	[0.13]
	ours	-0.09	-0.19	-0.41*	-0.01
Emot	liter	[0.75**]	-0.07	-0.15	0.23
	ours	0.11	0.09	0.18	-0.2
Extr	liter	[0.71**]	[-0.58**]	0.65	[-0.06]
	ours	-0.14	-0.24	0.17	0.29*
Open	liter	[]0.71**]	[-0.71**]	[0.24]	0.27
	ours	0.08	-0.47**	0.28	-0.07

*: $p < 0.05$, **: $p < 0.01$

4.2 Correlation analysis between MBTI and lexicons from other topics

Personality influences an individual's emotions, opinions and behaviours. Therefore, it would be interesting to compare our MBTI lexicons with other psychological lexicons, such as sentiment and emotion.

4.2.1 Correlation analysis between MBTI and Sentiment and Emotion lexicons

We found eight emotion lexicons from (Mohammad et al., 2013; Kiritchenko et al., 2014; Zhu et al., 2014). The datasets are shown in Table 4.2. Not so much work has been found on the correlation between personality and emotion, but still we can see their correlation from the defination of each dimension.

Emot_NRC has the general positive/negative emotion scores. All four dimensions are significantly correlated with emotion pos/neg scores. And E, J, N, T have positive correlations with emotion. The next three lexicons focus on different dimensions of emotion – arousal, dominance and valence. It shows that IE has significant positive correlations with all three dimensions, which makes sense in that EXTRAVERTs focus more on the outside stimuli, and will have more emotional reaction. JP and NS are negatively correlated with dominance, meaning that JUDGING type and INTUITION types have higher dominance – more stable. It also makes sense that these two types tend to analyze and give solutions, but SENSING and PERCEIVING types will have stronger feelings, which leads to lower dominance. TF has postive correlations with valence – FEELING type has stronger emotions.

The next four lexicons focus on specific types of emotion: anger, fear, joy and sadness. Only JP has positive correlations with fear and sadness, TF has positive correlation with joy. It suggests that FEELING type tends to use more joy-related words, while PERCEIVINGs tend to use more fear- and sadness- related words.

Personality influences individuals' way of writing and talking, which suggests that people with the same personality tends to use similar sentiment expressions. Therefore, we also compare MBTI lexicons with three sentiment lexicons: *senti_NRC* , *senti_Twitter_Eval2015* ,*senti_vader* (Ding et al., 2008; Mohammad et al., 2013; Kiritchenko et al., 2014; Zhu et al., 2014). IE, TF shows postive correlations with sentiment, while JP shows negative correlation. It means that EXTRAVERTs, FEELINGs and JUDGINGs will have more positive sentiment. Lin, Mao, and Zeng (2017) has developed a Big-5 personality-based sentiment classfier and they showed that it performed better than ordinary sentiment classifer. This also provides evidence for the possible correlation between personality and sentiment analysis.

Table 4.2: Correlation between MBTI lexicons and emotion and sentiment

	IE	JP	NS	TF
emot_NRC	0.25**	-0.13*	-0.25**	0.25**
emot_NRC-VAD_arousal	0.22**	0.0	-0.1	0.01
emot_NRC-VAD_dominance	0.26**	-0.15**	-0.27**	0.08
emot_NRC-VAD_valence	0.15**	0.02	0.01	0.28**
emot_NRC_anger	-0.17	0.26	-0.23	-0.26
emot_NRC_fear	-0.26	0.41*	-0.18	-0.08
emot_NRC_joy	0.2	-0.15	-0.07	0.28**
emot_NRC_sadness	-0.12	0.41*	0.1	-0.15
senti_NRC	0.16**	-0.21**	0.02	0.27**
senti_Twitter_Eval2015	0.16*	-0.24**	-0.1	0.43**
senti_vader	0.36**	-0.41**	-0.17	0.42**

4.2.2 Correlation analysis between MBTI and Age and gender lexicons

Gender and age are important demographic information. Twitter and Reddit have huge groups of users, including different genders and age groups. We wonder if there are some correlations between these two demographic features and MBTI dimensions. We used the age and gender lexicons by (Sap et al., 2014), as well as a few gender lexicons grouped by age (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al., 2013). The correlations are shown in Table 4.3. Only JP has a negative correlation with age lexica. It is in the reasonable direction that older

individuals rely more on judgement than perception.

The two gender lexicons, *gender_emnlp14* and *gender_123ngram* show that IE has a slightly negative correlation with gender, JP has a strong correlation, while NS and TF has moderate positive correlations with gender. Note that for gender lexicons, male is negative, and female is positive. The correlation analysis is consistent with the stereotypes that females tend to use more words about feelings (F) and are more sensible (S) in general. However, it is interesting to see that females tends to be more JUDGING. When we controlled for age group, most correlations between gender and personality disappearred, and only JP showed strong positive correlation with gender in age group 13 to 18, and negative correlation in age group 23 to 29. This may suggest the trait of JP for a person may have changed over time, but more analysis needs to be done before we make any conclusion.

Table 4.3: Correlation between MBTI lexicons with age and gender lexicons

	IE	JP	NS	TF
age_emnlp14	-0.03	-0.12**	0.03	-0.05
gender_emnlp14	-0.09*	-0.1*	0.13*	0.23**
gender_123ngram	0.09	-0.68**	0.72**	0.73**
gender_13_18	-0.31	0.52**	0.24	0.05
gender_19_22	0.36	0.33	-0.56	0.09
gender_23_29	0.46	-0.43*	-0.12	-0.21
gender_30_up	0.06	-0.11	0.43	0.17

Chapter 5

General Discussion

The main contribution of this study is that we developed predicted lexicons with ngrams and relative weights on each dimension of MBTI personality. The lexicons are built on eight MBTI datasets on Twitter and Reddit, using different features (unigram, 1-2 grams, 1-2-3 grams) and weightings (TF, TF-IDF, TF-logIDF), and different supervised learning models. Therefore, the lexicons should be robust enough and can be used across different scenarios. The correlation between MBTI lexicons and other existing lexicons, such as the Big-5 lexicons, emotion and sentiment lexicons, also provides evidence for the robusness of our personality lexicon. However, more evaluation needs to be done in the future, such as applying our lexical reprensentation with other datasets, or using our lexical representations as feature for different learning tasks.

The MBTI lexicons can be useful in the following aspects. On one hand, it can be useful for automatic personality prediction via social media data. Our lexical representation can be treated as a type of feature for different predicted models. It may be helpful in the fields of personalized recommendation and health tracking. On the other hand, when we compare our MBTI lexicons with lexicons from different areas, we found limited number of studies correlating MBTI personality with other fields such as sentiment and emotion. Therefore, it will be interesting to explore more systematically about the correlation between MBTI and other psychological terms. Besides, some studies in psychology have developed personality-descriptive vocabulary by self-rating on each word (Ashton, Lee, & Goldberg, 2004). It will be interesting to compare the difference between our lexicons with theirs, since ours are generated from online social media.

Bibliography

Arnoux, P.-H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., & Sinha, V. (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*.

Ashton, M. C., Lee, K., & 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., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., … De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of personality and social psychology*, *86*(2), 356.

Biel, J.-I., Tsiminaki, V., Dines, J., & Gatica-Perez, D. (2013). Hi youtube! personality impressions and verbal content in social video. In *Proceedings of the 15th acm on international conference on multimodal interaction* (pp. 119–126).

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 231–240).

Furnham, A. (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.

Gjurković, M., & Šnajder, J. (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* (pp. 87–97). doi:10.18653/v1/W18-1112

Goby, V. P. (2006). Personality and online/offline choices: Mbti profiles and favored communication modes in a singapore study. *Cyberpsychology & behavior*, *9*(1), 5–13.

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

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

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., … Seligman, M. E. (2014). The online social self: An open vocabulary approach to personality. *Assessment*, *21*(2), 158–169.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, *50*, 723–762.

Kumar, K. P., & Gavrilova, M. L. (2019). Personality traits classification on twitter. In *2019 16th ieee international conference on advanced video and signal based surveillance (avss)* (pp. 1–8). IEEE.

Lin, J., Mao, W., & Zeng, D. D. (2017). Personality-based refinement for sentiment classification in microblog. *Knowledge-Based Systems*, *132*, 204–214.

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

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Myers, I. B., McCaulley, M. H., & Most, R. (1985). *Manual, a guide to the development and use of the myers-briggs type indicator*. consulting psychologists press.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Plank, B., & Hovy, D. (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* (pp. 92–98).

Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., ... Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1146–1151).

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., ... Ungar, L. H. (2013). Characterizing geographic variation in well-being using tweets. In *Proceedings of the 7th international aaai conference on weblogs and social media*. ICWSM.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

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

Siddique, F. B., Bertero, D., & Fung, P. (2019). Globaltrait: Personality alignment of multilingual word embeddings. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 7015–7022).

Tobacyk, J. J., Livingston, M. M., & Robbins, J. E. (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.

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

Yarkoni, T. (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.

Zhu, X., Kiritchenko, S., & Mohammad, S. (2014). Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th international workshop on semantic evaluation (semeval 2014)* (pp. 443–447).