What Your Words Say About You

Identification of Myers-Briggs Type Indicator Using Text Mining Techniques and Predictive Modeling

RONNIE SCHNEIDER ● KYNDRA BEEBEHISER ● NATASHA HARRIS ● GABRIEL MCBRIDE

Portland State UNIVERSITY

TERADATA University Network ANALYTICS CHALLENGE

Problem & Motivation

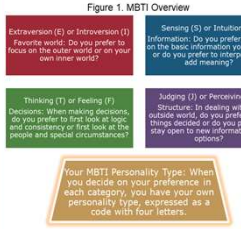
Businesses today are faced with increasing challenge and competition, and increasing opportunity to use data to gain competitive advantage. [1] One potential use of data is to enhance the productivity and effectiveness of employees, teams, and groups within the company.

Businesses today are faced with increasing challenge and competition, and increasing opportunity to use data to gain competitive advantage. [1] One potential use of data is to enhance the productivity and effectiveness of employees, teams, and groups within the company.

> *We cannot safely assume that other people's minds work on the same principles as our own. All too often, others with whom we come in contact do not reason as we reason, or do not value the things we value, or are not interested in what interests us.*
> *- Isabel Briggs Myers*

Research has shown that people are attracted to careers that allow them to make use of their natural personality preferences. The Myers-Briggs Type Indicator (MBTI) instrument is a well-known approach to identifying and understanding preferences. MBTI (see Figure 1) is a personality type system that divides everyone into 16 distinct types across 4 axes: Introversion (I) - Extroversion (E), Intuition (N) - Sensing (S), Thinking (T) - Feeling (F), Judging (J) - Perceiving (P). [2]

Figure 1. MBTI Overview



Understanding the results of the test can help strengthen many areas of a person's life: relationships, education, spirituality, and workplace. Knowing one's MBTI can put a person on the path to the right career and the right company [3].

Once a person is part of an organization, type understanding improves performance and enhances professional development. It is also an extremely helpful tool for those in managerial roles to know their team members' MBTI; the use and understanding of the team's dynamic can help build a stronger company by assembling teams, facilitating communication, and motivating employees. [4]

Approach

Research has shown that people are attracted to careers that allow them to make use of their natural personality preferences. The Myers-Briggs Type Indicator (MBTI) instrument is a well-known approach to identifying and understanding preferences. MBTI (see Figure 1) is a personality type system that divides everyone into 16 distinct types across 4 axes: Introversion (I) - Extroversion (E), Intuition (N) - Sensing (S), Thinking (T) - Feeling (F), Judging (J) - Perceiving (P). [2]

Dataset

The public domain data set was obtained at https://www.kaggle.com/datasnaek/mbti-type. The data were collected from the Meyers Briggs Forum on PersonalityCafe.com. Founded in 2008, PersonalityCafé offers articles, tests, and interactive discussion forums related to psychological personality profiles. The site is described as "a great place to go for learning more about your personality type." [8] The purpose of this dataset is to help see if any patterns can be detected in specific types and their style of writing; suggested uses include using machine learning to evaluate MBTI validity, or production of a machine learning algorithm to attempt to determine a person's MBTI type [9]. We emphasize that the gathering and use of input for this model - posts or other written material - should be done only in accordance with the applicable terms of use.

The dataset contains over 8600 rows of data. On each row is an MBTI forum member's:

- Self-identified MBTI personality type (4 letter code)
- Section of each of 50 posts. Each entry is separated by "|||" (3 pipe characters)

To explore the distribution of types, the data set was divided two ways: into 16 4-letter types, and into the 8 individual elements. Figure 2 shows the distribution of types, as well as a comparison with the United States population. [10] Figure 3 presents the same comparison, as well as the distribution within the set of each element, or subtype. The distribution within the data set is very uneven, with introverts, not surprisingly for a text message forum, far outweighing extroverts. We also observed that only 39 of the 8600+ records were type ESTJ; this minimum influenced the selection of data subsets for analysis.

Visualization

Figure 2. Comparison of MBTI Type Distribution

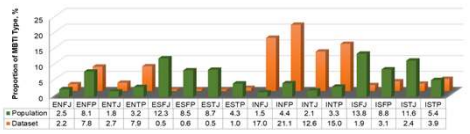
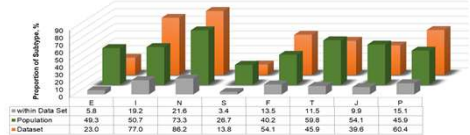
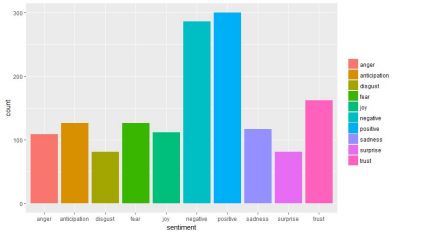


Figure 3. Summary and Comparison of MBTI Type Distribution



A convenient representation of the text content of the data set is the word cloud. Clouds for two representative types are presented in Figure 4. These clouds are generated using data that has been cleaned and processed. Some terms common to nearly all the types were removed to emphasize the differences.

Figure 4. Sample Word Clouds



Sentiment Breakdown for ENFJ Posts

Tools and Analytics

Analysis Tool

The analysis tool used for this project is the open source programming language, R. The data exploration, visualization, processing, analysis, and result formatting (Notebook capability) were all done in R. The project consists of two major parts: the preparation of the text data, and the analysis. An overview of the entire process is presented in Figure 5.

Figure 5

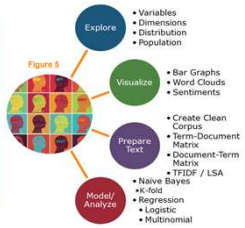
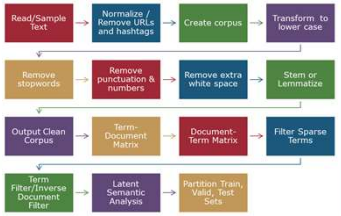


Figure 6



Text Processing

The goal of text processing is to winnow out the meaningful words from the vast bulk of text. The steps for 'cleaning' the data – changing it from raw text to an ordered data set - are represented in Figure 6. The approach for this project was a 'bag of words', meaning that the presence/absence of words is the focus, not the underlying meaning.

Analysis

The problem at hand is one of supervised learning, specifically classification: identifying the class of a new record based upon its attributes, using a model trained with a given data set. Methods of prediction considered included Neural Network, Tree Analysis, Support Vector Machines, Naïve Bayes classification, and Logistic Regression. After researching these methods and their suitability for text classification, we selected the latter two as being well-suited to classification, and straightforward to implement [13] [14]. Naïve Bayes is a probabilistic model that calculates the likelihood of a record belonging to a given class based upon its features. [15] Logistic regression is a regression model that generates the logit of the outcome variable (class yes/no decision), rather than a value. [11]

During development a set of tuning parameters was developed and implemented as input into individual processing and analysis functions within the code.

Work to improve accuracy continues with the addition of a custom lemmatization dictionary (e.g. replace is/are/was with "be") and k-fold cross validation during model training.

Results

Both Naïve Bayes and Regression models were trained to predict MBTI type based upon the data set of types and posts. Table 1 presents a summary of iterations of the Naïve Bayes algorithm performed to seek the best prediction accuracy for the full MBTI type, and each subtype. We found that prediction using sub-elements of type yielded much better accuracy in both models. This result was not surprising considering the many uncertainties associated with the data.

- Did the persons who made those original posts correctly declare their own type?
- The type classification instrument produces scores for each attribute. A person may be classified "I" with a score very close to the middle between E and I.
- Limited information for some of the types, and inconsistent amount of data even between subtypes.

The results for each subtype are listed on the left. Our results compare favorably with a benchmark using the same data set [16].

Subtype NB Accuracy	
E/I	0.66
N/S	0.79
T/F	0.72
J/P	0.58

Benchmark Accuracy Comparison			
Classifier	Benchmark	Naïve Bayes	Regression
E/I	0.75	0.66	0.74
Overall	0.26	0.28	0.47

Table 1. Naïve Bayes Model Development

	Case	Parameter set	Method	samples	trainperc	sparse_factor	dexfreq	tfidfdim	Naïve Bayes Accuracy
16 MBTI Types	A1.	a.	simplefreq	39	0.6	0.5	50	60	8%
	A1.	b.	simplefreq	All	0.6	0.98	50	60	1%
	A1.	c.	simplefreq	All	0.9	0.98	50	60	0%
	A1.	d.	simplefreq	All	0.6	0.98	50	120	1%
	A1.	e.	simplefreq	All	0.6	0.98	50	240	1%
	A2.	a.	tfidf_lsa	39	0.6	0.5		60	6%
	A2.	b.	tfidf_lsa	All	0.6	0.98		60	10%
	A2.	c.	tfidf_lsa	All	0.9	0.98		60	6%
	A2.	d.	tfidf_lsa	All	0.6	0.98		120	14%
	A2.	e.	tfidf_lsa	All	0.6	0.98		240	19%
	A3.	a.	tfidf	39	0.6	0.5			6%
	A3.	b.	tfidf	All	0.6	0.98			28%
	A3.	c.	tfidf	All	0.9	0.98			27%
	A3.	d.	tfidf	All	0.6	0.98			28%
	A3.	e.	tfidf	All	0.6	0.98			28%
E/I	B1.	a.	simplefreq	39	0.6	0.5	50	60	46%
	B1.	b.	simplefreq	All	0.6	0.98	50	60	68%
	B1.	c.	simplefreq	All	0.9	0.98	50	60	65%
	B1.	d.	simplefreq	All	0.6	0.98	50	120	67%
	B1.	e.	simplefreq	All	0.6	0.98	50	240	67%
	B2.	a.	tfidf_lsa	39	0.6	0.5		60	51%
	B2.	b.	tfidf_lsa	All	0.6	0.98		60	42%
	B2.	c.	tfidf_lsa	All	0.9	0.98		60	67%
	B2.	d.	tfidf_lsa	All	0.6	0.98		120	51%
	B2.	e.	tfidf_lsa	All	0.6	0.98		240	56%
	B3.	a.	tfidf	39	0.6	0.5			51%
	B3.	b.	tfidf	All	0.6	0.98			67%
	B3.	c.	tfidf	All	0.9	0.98			63%
	B3.	d.	tfidf	All	0.6	0.98			66%
	B3.	e.	tfidf	All	0.6	0.98			66%
E/I	C1.		tfidf	All	0.6	0.98			66%
N/S	C2.		tfidf	All	0.6	0.98			79%
T/F	C3.		tfidf	All	0.6	0.98			72%
J/P	C4.		tfidf	All	0.6	0.98			58%

Contributions and Uniqueness

In addition to creating functions to process text data and train models, the resulting models were implemented into a predictive function. Given a text sample, the model generates a prediction: "This person is predicted using NB model to be an ISFJ". These models can be exported and implemented into a readily available platform, such as an Excel workbook.

A potential application of the predictive model would be to create teams within the business environment by selecting individuals that are highly compatible with each other, or have diverse and complementary skills, depending on the need. For example, a USC Research Report discusses how certain personality types are preferred when selecting a systems architecting team [17]. For that function, team members with systematic and strategic analysis problem solving skills, or specifically the "NT" (intuition and thinking classifiers) type, were identified as preferred.

Another application of this project is to introduce text mining using R. The R Notebook includes the commented text mining outline, user-defined functions, visualization techniques, and learning notes.

While the current model accuracy is not ideal, and MBTI validity is a subject of some debate, we believe that the exercise of considering and discussing personality preferences and strengths has intrinsic value.