Michael Suggs and Lauren Simmonds

13 October 2021

# Personality and Online Presence NLP

**The Question**: Are online posts and speech patterns predictive of personality type?

Much of modern socialization occurs online, through social media platforms, forums, group chats, etc., and our reliance on digital communication is increasing.  Despite this increase, there is also a degree of detachment that comes with communicating over the internet. For example, important contextual factors for speaking in-person are missing, and it can be difficult to understand the tone of a message without seeing body language, hearing word inflection, sarcasm, and other underlying language components. These gray areas in online messaging could be potentially resolved if a user's message patterns gave insight into the type of person he or she is. We want to know if personality types are distinguishable in online posts, and if these patterns can then predict the personality types of other posters. The main personality metric we intend to predict is the *Myers-Briggs Type Indicator (MBTI)* due to its popularity in creating 16 different personality "profiles" for people to identify themselves. Despite the natural language processing (NLP) component, the nature of our research question is to classify people under certain personality factors, making the MBTI a proper evaluation metric to for us to use in supervised learning.

## Data

**MBTI**: https://www.kaggle.com/datasnaek/mbti-type?select=mbti_1.csv

The main dataset for exploration was collected from user data on PersonalityCafe, where people can have their MBTI personality type available on their profile. Every row in the dataset corresponds to a unique user out of the 8,675 rows and there are two columns:

- the user’s personality type, and
- up to 50 posts from the user.

**Stream-of-Consciousness Essays**: https://doi.org/10.1037/0022-3514.77.6.1296

Originally compiled in 1999 by Pennebaker and King, this dataset contains a selection of stream-of-consciousness essays submitted by individuals representing three distinct categories: journals from substance abuse inpatients, student writing assessments, and journal abstracts from social psychologists. These samples were compared against codings, self-reports, and behavioral measures from reference subjects to assign boolean presence (or absence) for each of the five factors of the Big Five personality index, aka the OCEAN model. In total, this consists of 2648 essays split into seven variables:

- a unique identifier for each essay, comprised of composition year and an anonymized identifier,
- the body of the essay as a singular string,
- a single column for each factor of the five-factor model (extraversion, neuroticism, openness to experience, agreeableness, and conscientiousness).

**Social Media Scraping**

To expand the amount and variety of available data, we further intend to explore scraping popular social media website(s) for additional self-reported personality metrics we may associate with submissions made by said users. One likely avenue of exploration in this endeavor is the subreddit /r/mbti where users are encouraged to provide their Myers-Briggs personality type as their user “flair” (a decorative tag-like feature that appears next to a given users name). Through the publicly available Reddit API, both the contents of these flairs and submissions made by the user in question may be obtained, allowing for a similar dataset to the aforementioned MBTI dataset to be constructed.

As is common in natural language processing tasks, Twitter may also prove a fruitful area for data collection for either supervised learning or data for classification and prediction. Given the cyclic trending nature of popular online personality tests, it is not uncommon for users to provide their results in the biography section of their profiles. However, even in the absence of this, Twitter also provides hashtags in which users may, in essence, self-categorize their posts; for example, it may be common to post about ones' MBTI results via the #mbti tag, or potentially within tags representing specific MBTI personality types such as #infp.

**SENTANCES1**

Holland codes, or the RIASEC, is another inventory used to determine a person’s working style. Both it and the MBTI are used in career exploration tests, and it follows a classification scheme as well. The Big Five personality index is the current standard in personality testing within psychology but rather than classifying people, it uses a Likert-based scale to score participant responses on a spectrum. The SENTANCES1 datasets are split between these two different personality metrics. Participants who completed either the RIASEC or the Big Five could participate in another post-survey metric. Those who participated were presented with the beginnings of unfinished sentences and were prompted to “fill in the blank” with how they would finish each sentence. These sentence fragments are paired with the RIASEC and Big Five scores of users, so initial language patterns can be used to determine the most likely personality score of a message’s author.

- Columns for age and sex
- Participant responses for each of the six sentence-stem prompts

- Depending on the RIASEC or Big Five test, the personality test results of the user

## Methods and Tasks

### Data Preparation

1. Remove non-word elements such as links, emoji and emoticons, and markdown artifacts.
2. Standardize input text within and across datasets by unifying case, removing stopwords and punctuation, and performing some combination of lemmatization, tokenization, or stemming.
3. Splitting combined post strings on their separators and removing anonymized references to other users.
4. Further anonymize and process the SENTANCE1 data by setting aside information such as age and gender not directly pertinent to our initial analysis; possibly further this by combining the provided sentence fragments with their starting prompts.

### Models of Interest

- Support Vector Machines
- Logistic Regressions
- Bayesian Classifiers
- Pre-trained models (e.g., Google's BERT)

### Potential Explorations

1. Comparing model results with other personality inventory results.

-Big Five and RIASEC data from SENTANCES1, and testing on scraped data from the Twitter or Reddit APIs, and/or the Stream of Consciousness Essays

-This would be good to check if the classification model can generalize to personalities as a whole, rather than only the MBTI

2. Controlling for sex or race differences in language, to reduce biasing in the model

-There have been ethical concerns with classifying human traits. When used in certain circumstances, biased models that misclassify someone can hold extremely negative results. For example, sex- and race-based speech patterns could cause the model to misclassify a personality type in a job hiring scenario, which leads to ignoring someone who was the best candidate.

# References

https://towardsdatascience.com/text-analytics-what-does-your-linkedin-profile-summary-say-about-your-personality-f80df46875d1

Identifying personality types using document classification methods

A Survey of Automatic Personality Detection from Texts

Survey Analysis of ML Methods for Natural Language Processing for MBTI Personality Type Prediction