Problem Statement

Hypothesis 1 – Is it possible to classify, with a certain degree of accuracy (how much?), a person's MBTI personality type through how they expresses themselves in written communication?

1 Context

The Myers Briggs Type Indicator (or MBTI for short) is a personality type system that divides everyone into 16 distinct personality types across 4 axis: Introversion (I) – Extroversion (E) Intuition (N) – Sensing (S) Thinking (T) – Feeling (F) Judging (J) – Perceiving (P). Personality type tests have their own limits but the are useful to guide individuals in self-awareness and in how they relate to the world around them. A clearer understanding of who we are can help us in our daily life.

Currently, types are determined after a person takes a introspective self-report (a questionnaire). While these tests have been found to be a good measure of a personality type, we want to determine if it is also possible to classify individuals through their written communication.

2 Criteria for success

We are able to develop a classification model with enough accuracy to classify individuals in each personality type.

3 Scope of solution space

The scope of the solution will be the +380.000 posts gathered from www.personalitycafe.com forum for 8675 individuals.

4 Constraints within solution space

The main limitation we have is inherent to the theory behind the analysis and our available dataset. The axes of MBTI profile tests are ranges not dichotomous variables. Being an E or an I does not mean you are 100% one or the other, everybody falls in between in some degree or other, we are talking about type dominance. However, we do not have this information available to us. Therefore, we are putting in the same basket someone who is a very strong E and someone who is an E but falls very close to being an I.

Another constraint is that the dataset seems to be skewed towards a certain personality type.

5 Stakeholders to provide key insight

There are no stakeholders in this project since it is a personal capstone project. There are, however, experts in the field who could weight in on this topic.

6 Key data sources

Our dataset was obtained from Kaggle. Posted by user Mitchell J. 3 years ago. This dataset contains 8675 rows of data, on each row is a person's:

- Type (This persons 4 letter MBTI code/type)
- A section of each of the last 50 things they have posted (Each entry separated by "|||" (3 pipe characters))