Senior Thesis Proposal: “Percieving Personality Through Sentiment Analysis”

Sidharth Dhawan

Professor Christiane Fellbaum

Motivation:

The goal of my thesis is to learn something about the personality (for example, happiness levels, or different axes of personality as dictated by the Meyers-Briggs personality test) from the way a person writes letters. There are many applications that require an understanding of the personality; for example, many companies try to discover the personality of applications through interviews. When someone is being sent to jail or put on trial, they may require a psychiatric evaluation. For applications like this, my algorithm may be able to save costs and resources.

Data sets that I will use in the thesis are:

(1) Presidential papers: this contains a list of the president's speeches, statements, public addresses, etc. This data has been compiled into large text volumes that should be quite easy to search. This data is most complete for the presidents in the twentieth and twenty-first centuries. Before the twentieth century, public addresses are not included in the dataset, although diaries documenting the president's daily activities and records of their addresses to congress are. Overall, in discovering personality, it will probably be most useful to use data from the twentieth-century presidents (of which there are about 15). Furthermore, it will be important to perform the learning using the same total document length for every president. In other words, the cumulative word count of documents used for each president should be the same.

(2) Presidential correspondences: these are available online via the library of congress. Data is available for a select number of presidents (and other famous people) from before the twentieth century. The data consists of all (or at least many) of the written letters exchanged between these famous people and close friends or family– they are usually private documents, as opposed to the above papers, which tend to be public in nature. For this reason, the correspondences dataset may be far more useful when trying to understand a person's personality. For each famous person, the full set of letters can be downloaded in text format via the library of Congress website, making the data set very easy to search.

Background and Related Work:

There has been a great deal of related work in the fields of sentiment analysis, as well as in the fields of psychological profiling. However, my algorithm is very different from any currently existing approach in its methodology and goals. The goal is to see if we can use the existing tools of sentiment analysis to learn things about psychology.

Approach and Plan:

As stated before, the goal of the thesis will be to learn something about the personality of each president or famous person through the bulk of their written correspondences. Personality traits will attempt to be learned through techniques in sentiment analysis and Natural Language Processing.

One option is to diagnose levels of happiness or optimism. This may be learned by a combination of factors. Firstly, a person's happiness probably depends on a combination of the misfortunes they have encountered in their everyday lives, combined with a sense of how much they tend to ruminate on positive or negative emotions (i.e. their optimism). Even if stated objectively, it is possible to identify an event as negative or positive. For example, an increase in a positive entity or a decrease in a negative entity is usually associated with a positive emotion, and vice versa for a negative emotion. In this case, it may be useful to use an impartial biography of the person to come up with events that are strongly negative or strongly positive (such as the death of a family member or getting an important bill passed), and try to see how these events are discussed in letters. It may also be useful to come up with a sentiment dictionary (which identifies levels of positivity and negativity within words of a given detaset) to help identify which words each president uses in a positive or negative context. A large and frequently used body of negative words may suggest a pessimistic outlook. Rumination on events that are negative or strongly negative may suggest lower levels of happiness or optimism, whereas a larger number of objective positive statements may imply optimism. Furthermore, description of the same events with positive or negative adjectives (i.e. “there is a deplorable lack of states left in the union” versus “there is a plentiful number of states left in the union”) can also be revealing. Optimism is a much more objective trait to learn, and to verify with a set of independent readers.

Additionally, the presidents could be rated on one or more other axes of the meyers-briggs personality scale. Fortunately, the presidents have already been pre-rated by psychologists along the MTBI scale. This can be a good final metric to judge the validity of the NLP-based metric designed.
Possibly one of the easiest metrics to measure via NLP analysis of written letters is the Thinking-Feeling personality axis. A thinking person will usually base decisions on what their principles, while a feeling person will go with the gut instinct and do what they feel is right. This may be very easy to discover through sentiment analysis, which deals with people's subjective and / or “emotional” statements, and with what constitutes an “emotional” versus an “objective” statement. One good starting point for this might be, as stated, finding “emotional” words within a text, versus opinions that are expressed via objective or evaluative statements. For example:

- evaluative opinions: “The German defense is very strong.”
- emotional opinions: “I feel sad for Argentina. You know nothing of defense.”

Furthermore, sometimes sentiments are implied in objective statements, while other times, they are stated outright.

- implied sentiments: “The battery life on my phone is less than three hours”
- opinionated sentiments: “The battery life on my phone is subpar.”

As mentioned previously, there are sentiment dictionaries and other, similar techniques that can be used to identify which objective words and statements can be considered negative within the any given context, which would help identify whether a sentence which seems objective might contain hidden sentiment. Using such techniques, one could also potentially infer other dimensions of the MBTI personality type.

Evaluation:

Evaluating the final results of the algorithm would require human opinion to judge correctly. Fortunately, it is relatively easy to find MTBI analyses of many famous people by performing a simple search. For some factors that are relatively easy or objectively judged, such as a person's level of optimism, the Amazon Mechanical Turk could be a useful tool. This is a software that allows any dataset to be analyzed via crowdsourcing – in other words, independent people could be paid to read letters and judge the level of optimism of the writer. To complete a final evaluation of the model, I will train it with some fraction of the data, and ask it to make personality predictions for the remaining data. Its prediction will be compared against these human evaluations.