Using Machine Learning to Predict Myers-Briggs Personality Types from Text

Project Proposal

Felix Mohr

Karlsruhe, Germany

---

1. Introduction

In his 1921 book *Psychological Types*, Swiss psychoanalyst Carl Gustav Jung developed a theory according to which a person's personality is determined by four dimensions. The four psychological dimensions identified by Jung are

- *extroversion* vs. *introversion*
- *intuition* vs. *sensing*
- *thinking* vs. *feeling*
- *judging* vs. *perceiving*

According to Jung, there exist complex interrelations between different dominant and subdominant personality traits present in a person. The complex character of Jung's theory makes it difficult for a layperson to apply his typology. Isabel Briggs Myers and Katharine Briggs therefore developed the *Myers-Briggs Type Indicator (MBTI)*, which is based on Jung's theory, yet easier to understand. Although frequently critisized, the MBTI is widely used in various contexts like psychotherapy, self-improvement seminars and human resources.

2. Problem Statement

A person's style of communication often allows us to make accurate assumptions about their personality traits quite easily. If a person's MBTI is a

good indication of their personality, we hence can expect their writing style to correlate with the expected writing style of people classified by the same Myers-Briggs indication.

In this project, we will apply machine learning techniques to examine whether a person's MBTI correlates with their writing style. If interrelations can be detected, we can – in opposition to some critics – conclude that the MBTI in fact is a meaningful indication of a person's personality traits.

It has been shown by various research projects (e. g. [1], [5] and [3]) that machine learning techniques can be applied to make assumptions about personality based on text data. Identifying the MBTI from text written by a person will allow for several applications, e. g. in improving the personalized user experience presented to them on a web application.

3. Datasets and Inputs

We will use data from the *PersonalityCafe* forum[1] made publicly available via kaggle[2]. This dataset contains all posts in the personality cafe forum for 8675 users which are labelled with their respective MBT indications. There exist 16 classes; for each of the four personality dimensions, the MBTI assigns one of two values. The available classes are not balanced; e. g. there exist 1999 extroverted, but 6676 introverted users. To account for this imbalance, we will apply balancing techniques.

Additionally, we will use *Google's* pretrained *word2vec* [4] vectors, which also are publicly available[3]. This can possibly help us in developing stronger classification models.

4. Solution Statement

We will apply state-of-the-art machine learning and text mining techniques to the stated problem. Using the labelled dataset, we can use various classification algorithms to identify patterns that are characteristic for the different MBTI types. Specifically, we will try using Random Forest classifiers. If results are not satisfactory, other classification algorithms like

[1] http://personalitycafe.com/forum

[2] https://www.kaggle.com/datasnaek/mbti-type

[3] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

Support Vector Machines (SVMs) may be used. Before applying these classification algorithms, we will perform balancing techniques like undersampling to account for class imbalances in the data at hand. As we are handling textual data, dimensionality reduction techniques like *Principal Component Analysis (PCA)* may be useful for dimensionality reduction of bag of word vectors.

Our goal is the development of a classifier that is able to identify a person's personality traits from text. This is a challenging problem, as even for a human observer it is not trivial to derive personality traits from written text. Also, our training dataset is not quite large. Taking this into account, we will develop four models – each of these models categorizes a person according to one of the four dimensions described above. We assume that it is possible to derive the four personality dimensions from text independently.

5. Benchmark Model

As a simple baseline model, we will transform the posts written by a respective user to their *bag of words* representation. Four out-of-the-box *Random Forest* classifiers will be trained; each learns to classify a user according to one of the four personality dimensions.

6. Evaluation Metric

It is our goal to reach a high accuracy in classifying users by the posts they have written. The accuracy is defined as

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

where tp is the number of *true positives*, tn are the *true negatives* etc. I. e., the accuracy of a classifier denotes the proportion of objects that are being correctly classified. As we are equally interested in people showing any personality traits, accuracy is the most suitable evaluation metric to the stated problem. As mentioned above, we will use balancing techniques to make sure that the accuracy can be applied to our data.

7. Project Design

As a first step, we will have to thoroughly analyze the data at hand. This will allow us make assumptions about features that are helpful in predicting personality types by writing style. We will examine the applicability of various types of features to our problem. These include the following:

(a) Features identifying patterns in writing style – e. g. the use of punctuation, part of speech features and the number of average words per post

(b) *Bag of Words* features

(c) Features that are based on word2vec vectors

The analysis of applicable features will include statistical examinations (means, variances etc.).

To make use of the word2vec vectors, we will build on the findings made by the authors of [2]. The authors develop the *word mover's distance* as a means to compare different documents for their similarity. To accomplish this, every document is represented by the mean of its word2vec vectors, and documents are considered more similar if their respective means are less distant, compared by the Euclidean distance. We therefore assume that it is possible to represent users in our dataset by the mean of the word2vec representation of the words written by them, and that we can apply machine learning algorithms to classify them based on their mean vectors. To visualize word2vec vectors, PCA has been applied frequently. As a step in data exploration, we will visualize different users based on their word2vec representation by applying PCA to reduce the 300-dimensional vectors to two dimensions. These two dimensions account only for a small amount of the variability in the word2vec representations, but it may still be possible that even in two dimensions, patterns to discriminate users based on their personality traits exist. Even if this is not the case, word2vec features may still be valuable to our problem.

As mentioned in the solution statement, we will develop four classifiers – each of them classifies objects in one of the four personality dimensions. These classifiers might again make use of different classification algorithms, each of which uses the features in one of the categories (a)–(c) to make independent assumptions about personality traits. The four final classifiers will

then be used to combine the preliminary results of the respective classification models. However, the concrete architecture of our classifiers will be based on optimal performance and might hence differ.

Algorithms that might be used in the development of classifiers might include the following:

- PCA can be used to reduce the dimensionality of our input features. In particular, bag of words vectors usually are quite high-dimensional and should be transformed into a lower-dimensional representation.
- Random Forests are a robust classification algorithm that frequently achieves high classification performance. We will use random forests to generate classification models.
- To tune our classifiers, we will perform *grid searches* for parameter optimization.

If some of the mentioned algorithms do not produce satisfactory results, alternative approaches might be exercised.

[1] Jennifer Golbeck. Predicting personality from social media text. *AIS Transactions on Replication Research*, 2(1):2, 2016.

[2] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France, 07–09 Jul 2015. PMLR. URL `http://proceedings.mlr.press/v37/kusnerb15.html`.

[3] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79, 2017.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL `http://arxiv.org/abs/1301.3781`.

[5] Ben Verhoeven and Walter Daelemans. Clips stylometry investigation (csi) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *LREC 2014-NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION*, pages 3081–3085, 2014.