Automatic Extraction of Personality from Text: Challenges and Opportunities

Abstract

In this study we examined the possibility to extract personality traits from a text. We created an extensive dataset by having experts annotate personality traits in a large number of texts from multiple online sources. From these annotated texts we selected a sample and made further annotations ending up with a large low-reliability dataset and a small high-reliability dataset. We then used the two datasets to train and test several machine learning models to extract personality from text, including a language model. Finally, we evaluated our best models in the wild, on datasets from different domains. Our results show that the models based on the small high-reliability dataset performed better (in terms of R2) than models based on large low-reliability dataset. Also, the language model based on the small high-reliability dataset performed better than the random baseline. Finally, and more importantly, the results showed our best model did not perform better than the random baseline when tested in the wild. Taken together, our results show that determining personality traits from a text remains a challenge and that no firm conclusions can be made on model performance before testing in the wild.

Introduction

Since the introduction of the personality concept, psychologists have worked to formulate theories and create models describing human personality and reliable measure to accordingly. The filed has been successful to bring forth a number of robust models with corresponding measures. One of the most widely accepted and used is the Five Factor Model BIBREF0. The model describes human personality by five traits/factors, popularly referred to as the Big Five or OCEAN: Openness to experience, Conscientiousness, Extraversion, Agreeableness, and emotional stability (henceforth Stability). There is now an extensive body of research showing that these factors matter in a large number of domains of

people's life. Specifically, the Big Five factors have been found to predict life outcomes such as health, longevity, work performance, interpersonal relations, migration and social attitudes, just to mention some domains (e.g. BIBREF1, BIBREF2, BIBREF3, BIBREF4).To date, the most common assessment of personality is by self-report questionnaires BIBREF5.

In the past decade however, personality psychologist, together with computer scientist, have worked hard to solve the puzzle of extracting a personality profile (e.g., the Big Five factors) of an individual based on a combination of social media activities BIBREF6. However, in the aftermath of Cambridge Analytica scandal, where the privacy of millions of Facebook users was violated, this line of research has been met with skepticism and suspicion. More recent research focuses on text from a variety of sources, including twitter data (e.g. BIBREF7, BIBREF8). Recent development in text analysis, machine learning, and natural language models, have move the field into an era of optimism, like never before. Importantly, the basic idea in this research is that personality is reflected in the way people write and that written communication includes information about the author's personality characteristics BIBREF9. Nevertheless, while a number of attempts has been made to extract personality from text (see below), the research is standing remarkably far from reality. There are, to our knowledge, very few attempts to test machine learning models "in the wild". The present paper aims to deal with this concern. Specifically, we aim to (A) create a model which is able to extract Big Five personality from a text using machine learning techniques, (B) investigate whether a model trained on a large amount of solo-annotated data performs better than a model trained on a smaller amount of high quality data, and, (C) measure the performance of our models on data from another two domains that differ from the training data.

Related Work

In BIBREF10 the authors trained a combination of logistic and linear regression models on data from 58,466 volunteers, including their demographic profiles, Facebook data and psychometric test results,

such as their Big Five traits. This data, the myPersonality dataset BIBREF11, was available for academic research until 2018, although this access has since been closed down. A demonstration version of the trained system is available to the public in form of the ApplyMagicSauce web application of Cambridge University.

In 2018 media exposed the unrelated (and now defunct) company Cambridge Analytica to considerable public attention for having violated the privacy and data of millions of Facebook users and for having meddled in elections, with some of these operations misusing the aforementioned research results. This scandal demonstrates the commercial and political interest in this type of research, and it also emphasizes that the field has significant ethical aspects.

Several attempts have been made to automatically determining the Big Five personality traits using only text written by the test person. A common simplification in such approaches is to model each trait as binary (high or low) rather than on a more realistic granular spectrum.

The authors of BIBREF12 trained a Bayesian Multinomial Regression model on stylistic and content features of a collection of student-written stream-of-consciousness essays with associated Big Five questionnaire results of each respective student. The researchers focused on the classifier for stability. The original representation of the numerical factor was simplified to a dichotomy between positive and negative, denoting essay authors with values in the upper or lower third respectively, and discarding texts from authors in the more ambiguous middle third. The resulting classifier then achieved an accuracy of 65.7 percent. Similar performance for the other factors was claimed as well, but not published.

A number of regression models were trained and tested for Big Five analysis on texts in BIBREF13. To obtain training data the authors carried out a personality survey on a microblog site, which yielded the texts and the personality data from 444 users. This work is a rare example of the Big Five being

represented an actual spectrum instead of a dichotomy, using an interval $[-1, 1]$. The performance of the systems was therefore measured as the deviation from the expected trait values. The best variant achieved an average Mean Absolute Percentage Error (i.e. MAPE over all five traits) of 14 percent.

In BIBREF14 the authors used neural networks to analyze the Big Five personality traits of Twitter users based on their tweets. The system had no fine-grained scoring, instead classifying each trait only as either yes (high) or no (low). The authors did not provide any details about their training data, and the rudimentary evaluation allows no conclusions regarding the actual performance of the system.

Deep convolutional neural networks were used in BIBREF8 as classifiers on the Pennebaker & King dataset of 2,469 Big Five annotated stream-of-consciousness essays BIBREF9. The authors filtered the essays, discarding all sentences that did not contain any words from a list of emotionally charged words. One classifier was then trained for each trait, with each trait classified only as either yes (high) or no (low). The trait classifiers achieved their respective best accuracies using different configurations. Averaging these best results yielded an overall best accuracy of 58.83 percent.

The authors of BIBREF15 trained and evaluated an assortment of Deep Learning networks on two datasets: a subset of the Big Five-annotated myPersonality dataset with 10,000 posts from 250 Facebook users, and another 150 Facebook users whose posts the authors collected manually and had annotated using the ApplyMagicSauce tool mentioned above. The traits were represented in their simplified binary form. Their best system achieved an average accuracy of 74.17 percent.

In BIBREF7 the accuracy of works on Big Five personality inference as a function of the size of the input text was studied. The authors showed that using Word Embedding with Gaussian Processes provided the best results when building a classifier for predicting the personality from tweets. The data consisted of self-reported personality ratings as well as tweets from a set of 1,323 participants.

In BIBREF16 a set of 694 blogs with corresponding self-reported personality ratings was collected. The Linguistic Inquiry and Word Count (LIWC) 2001 program was used to analyze the blogs. A total of 66 LIWC categories was used for each personality trait. The results revealed robust correlations between the Big Five traits and the frequency with which bloggers used different word categories.

Model Training

We employed machine learning for our text-based analysis of the Big Five personality traits. Applying machine learning presupposes large sets of annotated training data, and our case is no exception. Since we are working with Swedish language, we could not fall back on any existing large datasets like the ones available for more widespread languages such as English. Instead our work presented here encompassed the full process from the initial gathering of data over data annotation and feature extraction to training and testing of the detection models. To get an overview of the process, the workflow is shown in Figure FIGREF4.

Data annotation is time intensive work. Nevertheless, we decided to assemble two datasets, one prioritizing quantity over quality and one vice versa. The two sets are:

$D_\textrm {\textit {LR}}$: a large dataset with lower reliability (most text samples annotated by a single annotator),

$D_\textrm {\textit {HR}}$: a smaller dataset with higher reliability (each text sample annotated by multiple annotators).

By evaluating both directions we hoped to gain insights into the best allocation of annotation resources for future work. Regarding the choice of machine learning methods we also decided to test two approaches:

support vector regression (SVR): a well-understood method for the prediction of continuous values,

pre-trained language model (LM) with transfer learning: an LM is first trained on large amounts of non-annotated text, learning the relations between words of a given language; it is then fine-tuned for classification with annotated samples, utilizing its language representation to learn better classification with less training data. LM methods currently dominate the state-of-the-art in NLP tasks BIBREF17.

Each method was used to train a model on each dataset, resulting in a total of four models: $\textrm {\textit {SVR}}(D_\textrm {\textit {LR}})$ and $\textrm {\textit {LM}}(D_\textrm {\textit {LR}})$ denoting the SVR and the language model trained on the larger dataset, and $\textrm {\textit {SVR}}(D_\textrm {\textit {HR}})$ and $\textrm {\textit {LM}}(D_\textrm {\textit {HR}})$ based on the smaller set with more reliable annotations.

Technically, each of these four models consists of five subvariants, one for each Big Five personality trait, though for the sake of simplicity we will keep referring to the four main models only. Furthermore, to enhance legibility we will omit the dataset denotation in the model name when it is clear from the context which version is meant (e.g. in result tables).

Model Training ::: Data

As we intended our models to predict the Big Five personality traits on a scale from -3 to 3, rather than binary classification, we required training data that contained samples representing the whole data range for each trait. Given that no such dataset was available for the Swedish language, we set up our own large-scale collection and annotation operation.

The data was retrieved from four different Swedish discussion forums and news sites. These sources

were selected such as to increase the chances of finding texts from authors with a variety of different personalities. Specifically, the four sources are:

Avpixlat: a migration critical news site with an extensive comment section for each editorial article. The debate climate in the comment section commonly expresses disappointment towards the society, immigrants, minority groups and the government.

Familjeliv: a discussion forum with the main focus on family life, relationships, pregnancy, children etc.

Flashback: an online forum with the tagline “freedom of speech - for real”, and in 2018 the fourth most visited social media in SwedenBIBREF18. The discussions on Flashback cover virtually any topic, from computer and relationship problems to sex, drugs and ongoing crimes.

Nordfront: the Swedish news site of the Nordic Resistance Movement (NMR - Nordiska motståndsrörelsen). NMR is a nordic national socialist party. The site features editorial articles, each with a section for reader comments.

Web spiders were used to download the texts from these sources. In total this process yielded over 70 million texts, but due to time constraints only a small fraction could be annotated and thus form our training datasets $D_\textrm {\textit {LR}}$ and $D_\textrm {\textit {HR}}$. Table TABREF19 details the size of the datasets, and how many annotated texts from each source contributed to each dataset. $D_\textrm {\textit {HR}}$ also contains 59 additional student texts created by the annotators themselves, an option offered to them during the annotation process (described in the following section).

Model Training ::: Annotation

The texts were annotated by 18 psychology students, each of whom had studied at least 15 credits of personality psychology. The annotation was carried out using a web-based tool. A student working with this tool would be shown a text randomly picked from one of the sources, as well as instructions to annotate one of the Big Five traits by selecting a number from the discrete integer interval -3 to 3. Initially the students were allowed to choose which of the five traits to annotate, but at times they would be instructed to annotate a specific trait, to ensure a more even distribution of annotations. The tool kept the samples at a sufficiently meaningful yet comfortable size by picking only texts with at least two sentences, and truncating them if they exceeded five sentences or 160 words.

The large dataset $D_\textrm {\textit {LR}}$ was produced in this manner, with 39,370 annotated texts. Due to the random text selection for each annotator, the average sample received 1.02 annotations - i.e. almost every sample was annotated by only one student and for only one Big Five trait. The distribution of annotations for the different factors is shown in Figure FIGREF23. We considered the notable prevalence of -1 and 1 to be symptomatic of a potential problem: random short texts like in our experiment, often without context, are likely not to contain any definitive personality related hints at all, and thus we would have expected results closer to a normal distribution. The students preferring -1 and 1 over the neutral zero might have been influenced by their desire to glean some psychological interpretation even from unsuitable texts.

For $D_\textrm {\textit {HR}}$, the smaller set with higher annotation reliability, we therefore modified the process. Texts were now randomly selected from the subset of $D_\textrm {\textit {LR}}$ containing texts which had been annotated with -3 or 3. We reasoned that these annotations at the ends of the spectrum were indicative of texts where the authors had expressed their personalities more clearly. Thus this subset would be easier to annotate, and each text was potentially more suitable for the annotation of multiple factors.

Eventually this process resulted in 2,774 texts with on average 4.5 annotations each. The distribution for the different factors is shown in Table FIGREF24, where multiple annotations of the same factor for one text were compounded into a single average value.

The intra-annotator reliability of both datasets $D_\textrm {\textit {LR}}$ and $D_\textrm {\textit {HR}}$ is shown in Table TABREF21. The reliability was calculated using the Krippendorff's alpha coefficient. Krippendorff's alpha can handle missing values, which in this case was necessary since many of the texts were annotated by only a few annotators.

Table TABREF22 shows how many texts were annotated for each factor, and Figure FIGREF25 shows how the different sources span over the factor values.

Avpixlat and Nordfront have a larger proportion of annotated texts with factors below zero, while Flashback and especially Familjeliv have a larger proportion in the positive interval. The annotators had no information about the source of the data while they were annotating.

Model Training ::: Feature Extraction

To extract information from the annotated text data and make it manageable for the regression algorithm, we used Term Frequency-Inverse Document Frequency (TF-IDF) to construct features from our labeled data. TF-IDF is a measurement of the importance of continuous series of words or characters (so called n-grams) in a document, where n-grams appearing more often in documents are weighted as less important. TF-IDF is further explained in BIBREF19. In this paper, TF-IDF was used on both word and character level with bi-gram for words and quad-grams for characters.

Model Training ::: Regression Model

Several regression models were tested from the scikit-learn framework BIBREF20, such as RandomForestRegressor, LinearSVR, and KNeighborsRegressor. The Support Vector Machine Regression yielded the lowest MAE and MSE while performing a cross validated grid search for all the models and a range of hyperparameters.

Model Training ::: Language Model

As our language model we used ULMFiT BIBREF21. ULMFiT is an NLP transfer learning algorithm that we picked due to its straightforward implementation in the fast.ai library, and its promising results on small datasets. As the basis of our ULMFiT model we built a Swedish language model on a large corpus of Swedish text retrieved from the Swedish Wikipedia and the aforementioned forums Flashback and Familjeliv. We then used our annotated samples to fine-tune the language model, resulting in a classifier for the Big Five factors.

Model Training ::: Model Performance

The performance of the models was evaluated with cross validation measuring MAE, MSE and $\textrm {R}^2$. We also introduced a dummy regressor. The dummy regressor is trained to always predict the mean value of the training data. In this way it was possible to see whether the trained models predicted better than just always guessing the mean value of the test data. To calculate the R^2 score we use the following measurement:

where y is the actual annotated score, $\bar{y}$ is the sample mean, and e is the residual.

Model Training ::: Model Performance ::: Cross Validation Test

The models were evaluated using 5-fold cross validation. The results for the cross validation is shown in table TABREF33 and TABREF34.

For both datasets $D_\textrm {\textit {LR}}$ and $D_\textrm {\textit {HR}}$, the trained models predict the Big Five traits better than the dummy regressor. This means that the trained models were able to catch signals of personality from the annotated data. Extraversion and agreeableness were easiest to estimate. The smallest differences in MAE between the trained models and the dummy regressor are for extraversion and conscientiousness, for models trained on the lower reliability dataset $D_\textrm {\textit {LR}}$. The explanation for this might be that both of the factors are quite complicated to detect in texts and therefore hard to annotate. For the models based on $D_\textrm {\textit {HR}}$, we can find a large difference between the MAE for both stability and agreeableness. Agreeableness measures for example how kind and sympathetic a person is, which appears much more naturally in text compared to extraversion and conscientiousness. Stability, in particular low stability, can be displayed in writing as expressions of emotions like anger or fear, and these are often easy to identify.

Model Training ::: Model Performance ::: Binary Classification Test

As set out in Section SECREF2, earlier attempts at automatic analysis of the Big Five traits have often avoided modelling the factors on a spectrum, instead opting to simplify the task to a binary classification of high or low. We consider our $[-3, 3]$ interval-based representation to be preferable, as it is sufficiently granular to express realistic nuances while remaining simple enough not to overtax annotators with too many choices. Nevertheless, to gain some understanding of how our approach would compare to the state of the art, we modified our methods to train binary classifiers on the large and small datasets. For the purposes of this training a factor value below zero was regarded as low and values above as high, and the classifiers learnt to distinguish only these two classes. The accuracy during cross validation was calculated and is presented in Table TABREF36. Note that a direct comparison with earlier systems is

problematic due to the differences in datasets. This test merely serves to ensure that our approach is not out of line with the general performance in the field.

We conducted a head-to-head test (paired sample t-test) to compare the trained language model against the corresponding dummy regressor and found that the mean absolute error was significantly lower for the language model $\textrm {\textit {LM}}(D_\textrm {\textit {HR}})$, t(4) = 4.32, p = .02, as well as the $\textrm {\textit {LM}}(D_\textrm {\textit {LR}})$, t(4) = 4.47, p = .02. Thus, the trained language models performed significantly better than a dummy regressor. In light of these differences and the slightly lower mean absolute error $\textrm {\textit {LM}}(D_\textrm {\textit {HR}})$ compared to the $\textrm {\textit {LM}}(D_\textrm {\textit {LR}})$ [t(4) = 2.73, p = .05] and considering that $\textrm {\textit {LM}}(D_\textrm {\textit {HR}})$ is the best model in terms of $\textrm {R}^2$ we take it for testing in the wild.

Personality Detection in the Wild

Textual domain differences may affect the performance of a trained model more than expected. In the literature systems are often only evaluated on texts from their training domain. However, in our experience this is insufficient to assess the fragility of a system towards the data, and thus its limitations with respect to an actual application and generalizability across different domains. It is critical to go beyond an evaluation of trained models on the initial training data domain, and to test the systems “in the wild”, on texts coming from other sources, possibly written with a different purpose. Most of the texts in our training data have a conversational nature, given their origin in online forums, or occasionally in opinionated editorial articles. Ideally a Big Five classifier should be able to measure personality traits in any human-authored text of a reasonable length. In practice though it seems likely that the subtleties involved in personality detection could be severely affected by superficial differences in language and form. To gain some understanding on how our method would perform outside the training domain, we selected our best model $\textrm {\textit {LM}}(D_\textrm {\textit {HR}})$ and evaluated it on texts from two

other domains.

Personality Detection in the Wild ::: Cover Letters Dataset

The cover letters dataset was created during a master thesis project at Uppsala University. The aim of the thesis project was to investigate the relationship between self-reported personality and personality traits extracted from texts. In the course of the thesis, 200 study participants each wrote a cover letter and answered a personality form BIBREF22. 186 of the participants had complete answers and therefore the final dataset contained 186 texts and the associated Big Five personality scores.

We applied $\textrm {\textit {LM}}(D_\textrm {\textit {HR}})$ to the cover letters to produce Big Five trait analyses, and we compared the results to the scores from the personality questionnaire. This comparison, measured in the form of the evaluation metrics MAE, MSE and $\textrm {R}^2$, is shown in Table TABREF39. As it can be seen in the table, model performance is poor and $\textrm {R}^2$ was not above zero for any of the factors.

Personality Detection in the Wild ::: Self-Descriptions Dataset

The self-descriptions dataset is the result of an earlier study conducted at Uppsala University. The participants, 68 psychology students (on average 7.7 semester), were instructed to describe themselves in text, yielding 68 texts with an average of approximately 450 words. The descriptions were made on one (randomly chosen) of nine themes like politics and social issues, film and music, food and drinks, and family and children. Each student also responded to a Big Five personality questionnaires consisting of 120 items. The distribution of the Big Five traits for the dataset is shown in figure FIGREF42.

Given this data, we applied $\textrm {\textit {LM}}(D_\textrm {\textit {HR}})$ to the self-description texts to

compute the Big Five personality trait values. We then compared the results to the existing survey assessment using the evaluation metrics MAE, MSE and $\textrm {R}^2$, as shown in Table TABREF40. As it can be seen in the table, model performance was poor and $\textrm {R}^2$, like the results for the cover letters dataset, was not above zero for any of the Big Five factors.

Conclusions

In this paper, we aimed to create a model that is able to extract Big Five personality traits from a text using machine learning techniques. We also aimed to investigate whether a model trained on a large amount of solo-annotated data performs better than a model trained on a smaller amount of high-quality data. Finally, we aimed to measure model performance in the wild, on data from two domains that differ from the training data. The results of our experiments showed that we were able to create models with reasonable performance (compared to a dummy classifier). These models exhibit a mean absolute error and accuracy in line with state-or-the-art models presented in previous research, with the caveat that comparisons over different datasets are fraught with difficulties. We also found that using a smaller amount of high-quality training data with multi-annotator assessments resulted in models that outperformed models based on a large amount of solo-annotated data. Finally, testing our best model ($\textrm {\textit {LM}}(D_\textrm {\textit {HR}})$) in the wild and found that the model could not, reliably, extract people’s personality from their text. These findings reveal the importance of the quality of the data, but most importantly, the necessity of examining models in the wild. Taken together, our results show that extracting personality traits from a text remains a challenge and that no firm conclusions can be made on model performance before testing in the wild. We hope that the findings will be guiding for future research.