Grammar-Based and Lexicon-Based Techniques to Extract Personality Traits from Text

Maira B. Carvalho (m.brandaocarvalho@uvt.nl)
Tilburg Center for Cognition and Communication
Tilburg University
Tilburg, The Netherlands

Max M. Louwerse (mlouwerse@uvt.nl)
Tilburg Center for Cognition and Communication
Tilburg University
Tilburg, The Netherlands

Abstract

Language provides an important source of information to predict human personality. However, most studies that have predicted personality traits using computational linguistic methods have focused on lexicon-based information. We investigate to what extent the performance of lexicon-based and grammar-based methods compare when predicting personality traits. We analyzed a corpus of student essays and their personality traits using two lexicon-based approaches, one top-down (Linguistic Inquiry and Word Count (LIWC)), one bottom-up (topic models) and one grammar-driven approach (Biber model), as well as combinations of these models. Results showed that the performance of the models and their combinations demonstrated similar performance, showing that lexicon-based top-down models and bottom-up models do not differ, and neither do lexicon-based models and grammar-based models. Moreover, combination of models did not improve performance. These findings suggest that predicting personality traits from text remains difficult, but that the performance from lexicon-based and grammar-based models are on par.

Keywords: language; personality; traits; machine learning; computational linguistics; lexicon-based; grammar-based

Introduction

In our daily interactions, we guide our behavior towards other people using information that is collected throughout these interactions, but also using knowledge about the world and social groups (Rich, 1979). These judgments are oftentimes made unconsciously.

Models of users’ behavior, thinking and feeling typically rely on the personality traits that can be identified (McCrae & John, 1992). Trait theory is an approach to the study of human personality in which it is believed that humans exhibit habitual patterns of behavior, thought and emotion. It is presumed that there is a relatively small number of dimensions that can be used to describe personality (O’Connor, 2002). Independent analyses have consistently yielded five broad dimensions, called the Big Five (or Five Factor Model): openness to experience, conscientiousness, extraversion, agreeableness and neuroticism (McCrae & John, 1992).

Personality traits are generally identified on the basis of data collected from the users who fill out standardized questionnaires. However, such an approach has certain drawbacks. Firstly, it can be costly for the researcher and time-consuming for the user (Gauch, Speretta, Chandramouli, & Micarelli, 2007). Secondly, people are not reliable sources of information about themselves: there is evidence to suggest that self-descriptions are heavily influenced by the social groups in which a person finds himself (McGuire & Padawer-Singer, 1976), and dissimulation can be a problem in self-reports (Wright, 2014).

Automatic inference of personality offers the advantages of being less intrusive and possibly more environmentally valid. Indeed, a range of studies have investigated the extent to which personality traits can be predicted from user behavior. The lion’s share of these studies use linguistic data as sources of information, both language and speech (Beukeboom, Tanis, & Vermeulen, 2012; Gawda, 2009; Mairesse, Walker, Mehl, & Moore, 2007; Mehl, Robbins, & Holleran, 2012; Oberlander & Gill, 2006; Oberlander & Nowson, 2006). Linguistic data has also shown to indirectly shed light on personality. For instance, linguistic cues have shown to be linked to deception (Louwerse, Lin, Drescher, & Semin, 2010), and to different registers of communication (Louwerse, McCarthy, McNamara, & Graesser, 2004). Furthermore, a person’s emotional state is reflected in language use (Tausczik & Pennebaker, 2009), not only by explicit lexical content but also by implicit semantic associations (Recchia & Louwerse, 2014).

In considering the cognitive science literature that aims to extract behavioral information from linguistic data, two approaches can be distinguished. On the one hand, studies use lexical cues to extract information from text, for example emotional expression (Kahn, Tobin, Massey, & Anderson, 2007), deception (Newman, Pennebaker, Berry, & Richards, 2003), political orientation (Dehghani, Sagae, Sachdeva, & Gratch, 2013), moral foundations (Graham, Haidt, & Nosek, 2009), romantic relationship outcomes (Ireland et al., 2011), among others. On the other hand, extracting behavioral information from explicit lexical information can be problematic. First, in controlled experimental settings it is easy for participants to carefully monitor their semantic content. For instance, in deception studies participants might avoid using specific words. Second, sparsity issues may emerge if algorithms detect specific word use.

An alternative approach lies in using grammar-based cues. By performing a manual analysis of seven syntax markers, Gawda (2009) identified an increased use of certain features
in emotional narratives written by individuals with antisocial personality disorder. Current computational linguistic tools are able to extract many grammar-based linguistic features automatically and efficiently, and it is reasonable to assume that such features could also carry information about personality. Biber (1988) conducted a study on linguistic variation across speech and writing, and computed the frequency of 67 linguistic features (e.g. frequency of auxiliary verbs, pronouns, main verbs, adjectives); he was able to identify different writing genres using naturally occurring word patterns. Graesser, McNamara, Louwerve, and Cai (2004) proposed the tool Coh-Metrix, which allows the user to analyze texts on discourse, cohesion, and world knowledge. The Suite of Linguistic Analysis Tools (SALAT) calculates scores for aspects such as syntactic complexity (Kyle, 2016) and cohesion (Crossley, Kyle, & McNamara, 2015). These tools open up a range of possibilities for the investigation of the relationships between personality and language use.

In conclusion, two approaches can be identified in extracting personality traits from language use: a lexicon-driven approach and a grammar-driven approach. As pointed above, the majority of cognitive science literature has focused on the lexicon-based approach. The question is to what extent the findings from a grammar-driven approach are comparable with the lexicon-driven approach that currently dominates the literature. We address this question in the current work.

**Extracting personality traits from text**

Most studies on extracting personality from text focused on identifying words, collocations and general linguistic features that occur in texts produced by one group of people versus another group, aiming to uncover which features are informative when trying to differentiate the groups. Early attempts relied on word counting and predefined dictionaries that sort words in categories (Tausczik & Pennebaker, 2009). This approach, albeit basic, has been used in many studies that show links between word usage and certain psychological processes and personalities, e.g. Beukeboom et al. (2012), and Mehl et al. (2012). Other researchers used bottom-up approaches to associate linguistic features with personality types. Oberlander and Gill (2006) collected large corpora of text labeled with the personality traits of the author and performed stratified corpus comparisons. Interesting findings included the fact that people who scored high in extraversion used more inclusive expressions and connectives, while those with low score were more tentative and used adjectives less frequently. The authors also noted that people with high neuroticism scores had preference for multiple punctuation.

Another approach is to treat the problem as a supervised classification task, employing machine learning techniques to identify the personality of the author of a given text. Oberlander and Nowson (2006) investigated a corpus of weblog posts from 71 participants, who completed a personality questionnaire online as part of the study. The authors used Support Vector Machine (SVM) classifiers and feature sets consisting of n-grams extracted from the text and selected according to different levels of restriction. The same approach was later applied to a larger sample of bloggers (Iacobelli, Gill, Nowson, & Oberlander, 2011). Argamon, Dhawle, Koppel, and Pennebaker (2005) used SVMs and four sets of lexical features to differentiate high and low extraversion and neuroticism, using a corpus of around 2400 student essays and personality assessments, collected by Pennebaker and King (1999). Mairesse et al. (2007) also worked on the same corpus, employing a series of classification and regression techniques and features from both the Linguistic Inquiry and Word Count (LIWC) and the Medical Research Council (MRC) Psycholinguistic Database. Their results confirmed previous findings and reveal new correlations between linguistic markers and personality, such as use of swear words and use of pronouns. As for the accuracy of automatic classification, the authors reported accuracies that are, according to their evaluation, significantly above chance; however, it is not clear whether these values are high enough to be useful in real applications (Mairesse et al., 2007).

Following the work by Mairesse et al. (2007), the task of automatic identification of personality from text gained a lot of attention from the research community, mostly due to the Workshop on Computational Personality Recognition (Celli, Pianesi, Stillwell, & Kosinski, 2013). As part of a shared task, the organizers made available two datasets of text labeled with the personality traits of the authors – including the Essay Corpus by Pennebaker and King (1999). As a result, many researchers tackled the problem with different learning algorithms (e.g. Naive Bayes, SVM, kNN, ensemble methods, logistical regression) and using different features such as n-grams, LIWC, MRC, lexical nuances, part-of-speech tags, emotional values from the AFINN database, word intensity scale, sentiment analysis and word associations to emotions (Celli et al., 2013).

Although the results from these attempts are encouraging, it had been noted that top-down approaches based on lexical resources seem to perform better than bottom-up approaches based only on words or n-grams (Celli et al., 2013). Nevertheless, there are benefits in employing approaches that do not rely on pre-defined vocabularies, for example allowing exploration of topics not previously considered, easier application in different genres and languages, and saving the effort of creating the word lists (Schwartz et al., 2013).

Schwartz et al. (2013) used a large dataset of Facebook posts (over 15.4 million Facebook messages collected from 75 thousand volunteers) to perform an open-vocabulary analysis of correlations between personality types and vocabulary use. The goal of the work was to discover unexpected relationships that would not necessarily be evident from using pre-defined word categories. Although the focus of the work was mainly to explore and gain insights on the data, the authors also used the approach to predict personality from text, with results that are comparable to previous literature. Liu, Wang, and Jiang (2016) also attempted to predict personality
from text while avoiding to rely on predefined vocabularies. The authors proposed a model that expands latent Dirichlet allocation (LDA) to include the assumption that topic distribution depends not only on the characteristics of the corpus itself, but also on the five personality traits of writers. However, the topics identified by their model seem to be affected disproportionately by individuals with less common personality combinations, and for this reason the model must be trained with a massive, representative corpus for which the personality of the writers is known. Since obtaining such corpora is difficult, the applicability of their approach seems to be limited.

In this paper, we investigate how we can predict user personality from written text by using features that do not rely on closed-vocabularies, and compare the results to the state of the art. In the next section, we describe our approach.

**Procedure**

**Dataset**

We use the Essay Corpus (Pennebaker & King, 1999), which consists of 2468 essays, written by introductory psychology students of the University of Texas as part of their course assignments. The students also completed the Five Factor Inventory personality questionnaire (John, Donahue, & Kentle, 1991), so that all essays could be marked with five personality scores for each Big-5 trait (Openness to Experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism). In addition to the scores, the corpus contains binary values for each trait (high/low), which were obtained using a median split over the scores. The class distribution of the binary values is shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>OPN</th>
<th>CON</th>
<th>EXT</th>
<th>AGR</th>
<th>NEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>1196</td>
<td>1214</td>
<td>1191</td>
<td>1158</td>
<td>1235</td>
</tr>
<tr>
<td></td>
<td>(48.46%)</td>
<td>(49.19%)</td>
<td>(48.26%)</td>
<td>(46.92%)</td>
<td>(50.04%)</td>
</tr>
<tr>
<td>High</td>
<td>1272</td>
<td>1254</td>
<td>1277</td>
<td>1310</td>
<td>1233</td>
</tr>
<tr>
<td></td>
<td>(51.54%)</td>
<td>(50.81%)</td>
<td>(51.74%)</td>
<td>(53.08%)</td>
<td>(49.96%)</td>
</tr>
</tbody>
</table>

Table 1: Class distribution in dataset.

**Features**

We employed four groups of feature sets, which were chosen to investigate to what extent the performance of lexicon-based and grammar-based computational linguistic methods are comparable.

For the lexicon-based features, we used two main approaches: top-down (LIWC and MRC) and bottom-up (topic modeling with latent Dirichlet allocation (LDA)). For the grammar-based features, we selected the original Biber features (Biber, 1988).

**Lexicon-based, top-down**

1) A total of 80 LIWC features were extracted using the LIWC2007 software, which outputs relative frequencies of words found in each pre-defined category, and a few structural features such as word count and words per sentence.

2) The 14 MRC features refer to word length, number of syllables or phonemes, and values for frequency of use, imageability, concreteness, meaning, age of acquisition, among others. The MRC features were calculated by averaging the scores of the essay words found in the database (as opposed to averaging over total word count).

**Lexicon-based, bottom-up**

For topic modeling, we pre-processed the corpus by lemmatizing the words using NLTK’s WordNet lemmatizer, and removing non-English words (i.e. words that were not found in Wordnet). Given the relatively small size of the corpus, we did not filter out words based on frequency. Then, we trained three LDA models with different number of topics (30, 65 and 100). Each document in the dataset was converted to a vector that represents the proportion in which each topic appears in the document. To train the model, we used the library Gensim, with 10 passes and default hyperparameters.

To illustrate the topics found in the corpus, these are the ten most relevant words for each of the three most frequently appearing topics extracted by the 30-topic LDA model: “think, go, get, really, like, write, minute, time, wonder, need”; “go, get, really, time, friend, home, want, much, like, miss”; and “life, people, thing, know, think, time, one, make, feel, way”.

**Grammar-based, top-down**

For this study, we use the 67 features selected by Biber (1988) to reflect the linguistic structure of the text. These features primarily operate at the word level, such as parts-of-speech, and fall into categories such as tense and aspect markers, adverbials, pronouns, questions, nominal forms, passives, subordination features, prepositional phrases, coordinations and negations, and so on. These features were extracted from the text using software developed in-house.

**Combinations**

In addition to considering these models separately, we investigated the model combinations in order to determine their complementary value.

**Classifiers**

We trained five Support Vector Machine (SVM) classifiers with linear kernel, one for each personality trait. SVMs were chosen due to previous reports of them performing better on this task than other algorithms (Mairesse et al., 2007), and linear kernels were employed to retain interpretability of the model. For the implementation, we used the machine learning library Scikit-learn, which in turn uses an implementation based on Libsvm. The classifiers were trained without parameter tuning (i.e. penalty parameter C=1.0).

**Results**

Reservations have been expressed by the scientific community on the application of null-hypothesis statistical testing for comparison of machine learning algorithms for many rea-
Table 2: Average accuracy, precision and recall for each classifier, with 95% confidence interval.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Baseline</th>
<th>Lexicon, top-down</th>
<th>Lexicon, bottom-up</th>
<th>Grammar</th>
<th>Combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(A) Majority</td>
<td>(B) LIWC+MRC</td>
<td>(C) LIWC</td>
<td>(D) TM (30)</td>
<td>(E) TM (65)</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>.517 .562 ± .007 .543 ± .006</td>
<td>.546 ± .006 .555 ± .006 .551 ± .006</td>
<td>.546 ± .007</td>
<td>.533 ± .007 .537 ± .007</td>
<td></td>
</tr>
<tr>
<td>Neuroticism</td>
<td>.531 .545 ± .006 .552 ± .007</td>
<td>.542 ± .005 .542 ± .006</td>
<td>.562 ± .006</td>
<td>.552 ± .006 .553 ± .007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.508 .607 ± .009 .611 ± .010</td>
<td>.653 ± .011 .610 ± .010 .615 ± .011</td>
<td>.599 ± .010</td>
<td>.588 ± .010 .598 ± .009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.517 .560 ± .006 .554 ± .005</td>
<td>.544 ± .004 .551 ± .004 .550 ± .004</td>
<td>.553 ± .006</td>
<td>.562 ± .006 .563 ± .006</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.517 .624 ± .010 .624 ± .010</td>
<td>.764 ± .009 .756 ± .010 .731 ± .011</td>
<td>.635 ± .010</td>
<td>.614 ± .010 .644 ± .010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.531 .561 ± .006 .567 ± .005</td>
<td>.554 ± .003 .555 ± .004 .546 ± .004</td>
<td>.562 ± .005</td>
<td>.570 ± .005 .570 ± .006</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.531 .651 ± .011 .664 ± .010</td>
<td>.843 ± .007 .758 ± .012 .821 ± .009</td>
<td>.678 ± .010</td>
<td>.636 ± .010 .649 ± .010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.500 .565 ± .007 .571 ± .006</td>
<td>.532 ± .007 .527 ± .007 .520 ± .006</td>
<td>.545 ± .006</td>
<td>.552 ± .006 .545 ± .007</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.500 .563 ± .007 .571 ± .007</td>
<td>.529 ± .006 .527 ± .008 .519 ± .006</td>
<td>.545 ± .005</td>
<td>.552 ± .006 .544 ± .006</td>
<td></td>
</tr>
</tbody>
</table>

1. Precision and recall scores consider the “high” class as positive label.
General Discussion

The differences in performance between the different algorithms are very small. Using different feature sets yields similar results, and combining different features does not improve the performance in any meaningful way.

One possible reason could be a floor effect, in which the questionnaire used to assess personality traits in this corpus would not be able to distinguish reliably between subjects at the lower end of the scale. This is unlikely, since the test used in this corpus is a standardized questionnaire that has been validated and used in numerous studies (John et al., 1991).

It is also possible that the use of self-assessments of personality makes this task particularly difficult due to the potential unreliability of self-reports, as discussed in the introduction. Future investigation could incorporate personality assessments made by human observers, to evaluate to which extent self-assessment and observed scores differ, and whether the algorithms could match the performance of human judges. Furthermore, replicating the study with other corpora could also indicate whether different text types could be more suitable for detecting certain personality traits.

In this study, we used a relatively limited set of non-top-down features, namely the features proposed by Biber (1988) and topic modeling. Future work could investigate whether applying other grammar-based and bottom-up lexicon-based features (e.g. cohesion, syntactic complexity, n-grams, skip grams, Word2Vec, semantic similarities) would result in better performances. In addition, we could try to improve the models by using non-linear kernels, performing parameter tuning, and employing ensemble machine learning methods for combining different sets of features.

However, the difficulty of identifying personality traits from text could signal a more fundamental issue. Mischel and Shoda (1995) have argued that individual differences in social behaviors are actually variable across different situations (situationism), and not completely stable as it is proposed by trait theory. As such, if personality scales and textual analyses tap into different social situations, tasks that use questionnaire scores as gold standard will not be able to achieve acceptable performance. Further research is needed to investigate this hypothesis, and whether other stable patterns of behavior could be used as gold standard for automatic personality inferences.

The current study has used the most common personality traits classification, the Big Five, and the most commonly used corpus to identify personality traits, the Essay Corpus, in order to compare the difference between top-down and bottom-up lexicon-based and grammar-based computational linguistic techniques. Our findings show that no differences were obtained between lexicon-based and grammar-based or between top-down and bottom-up approaches, nor complementary advantages for combinations of models, despite the fact that all methods were on par with the performance previously reported.

Acknowledgments

We would like to thank James W. Pennebaker for providing us access to the original personality scores of the Essay Corpus.

References


