Psychological profiling through textual analysis

Article *in* Literary and Linguistic Computing · August 2013

DOI: 10.1093/llc/fqs070

CITATIONS
40

READS
1,493

3 authors, including:

John Noecker
Duquesne University
4 PUBLICATIONS **73** CITATIONS
SEE PROFILE

Patrick Juola
Duquesne University
95 PUBLICATIONS **2,946** CITATIONS
SEE PROFILE

Psychological profiling through textual analysis

John Noecker Jr, Michael Ryan and Patrick Juola
Duquesne University

Abstract

In this article, we examine the application of computational stylometry to psychological profiling. We adapt several techniques, which have proven useful for author identification to the problem of identifying an individual author's Myers-Briggs personality type indicator from the statistical features of the text. The Myers-Briggs type indicator assigns four binary classifications to define personality type: Extrovert–Introvert, Intuitive–Sensing, Thinking–Feeling, and Judging–Perceiving. For this study, we use the Personae corpus, which consists of 145 Dutch-language texts all pertaining to a specific topic, each labeled with the Myers-Briggs personality profile of the author (Luyckx K and Daelemans W, *Personae: A Corpus for Author and Personality Prediction from Text, In Proceedings of the 6th Language Resources and Evaluation Conference*. Marrakech, Morocco: International Conference on Language Resources and Evaluation. 2008). Our system builds upon earlier work by Luyckx and Daelemans (*Using Syntactic Features to Predict Author Personality from Text. In Proceedings of Digital Humanities 2008*, Oulu, Finland: Digital Humanities, pp. 146–9.) to provide a set of best practices for personality profiling. We propose a more sophisticated modeling technique, combined with more advanced feature selection and state-of-the-art analysis methods from author identification to achieve a significant improvement over previous systems.

Correspondence:
John Noecker Jr, Evaluating Variations in Language Laboratory, 600 Forbes Avenue, Duquesne University, Pittsburgh, PA 15282, USA.
E-mail:
jnoecker@jgaap.com

1 Introduction

An author's writing style has been hypothesized to contain clues to many aspects of the writer's life. Studies have shown that it is possible to identify the author (Binongo, 2003), gender (Koppel *et al.*, 2002), education level (Juola and Baayen, 2005), and even personality indicators (Argamon *et al.*, 2005) from the statistical properties of a text. In this article, we build on earlier work in stylometric personality identification and propose a method for identifying Myers-Briggs personality type indicators for Dutch-language corpora. This method is unique because it is much simpler than previous approaches and provides significantly better performance.

2 Background

As one of the major products of the human mind, language should be expected to provide some sort of indicator of the mind that produced it. In addition to indicating identity, it may also be able to show something about process. The use of linguistic indicators as cues for medical diagnosis has a long history. A simple example is the sentential complexity measures as revised by Brown *et al.* (2005), where the absence of highly complex sentences can be a sign of cognitive trouble. Similarly, other studies have shown that particular linguistic features can be a sign of specific emotional attitudes—when people are depressed, for example, they use more first-person singular pronouns (Rude *et al.*, 2004). Perhaps more hopefully,

Literary and Linguistic Computing, Vol. 28, No. 3, 2013.

doi:10.1093/llc/fqs070 Advance Access published on 8 January 2013

an increased use of causal words such as 'because,' as well as an increased use of cognitive words such as 'realize' predicts recovery from trauma (Pennebaker *et al.*, 2003). These are simple examples of lexical features that can be used to categorize people by their mental states. Other examples of such studies—and examples of intense interest to law enforcement, one expects—include the studies of lying and linguistic style performed by Newman *et al.* (2003) as well as by Hancock (2007).

One obvious 'indicator' is of course, personality: researchers have developed many ways to classify people by their personality. The Myers-Briggs type indicator (MBTI) assigns four binary classifications to define personality (Myers and Myers, 1980):

> Extroversion–Introversion: E's draw energy from social interaction, whereas I's might tend toward solitude. These are two different 'attitudes'.
> iNtuition–Sensing: S's prefer to rely on information they perceive directly, whereas N's are more comfortable with theoretical information. These are two ways of 'perceiving' the world around us.
> Thinking–Feeling: F's tend to make decisions by assigning actions a 'personal, subjective value,' whereas T's prefer to use a logical process. These are two ways of 'judging' actions.
> Judgement–Perception: J's are more comfortable coming to a judgment of their perceptions, whereas P's prefer to simply perceive what is there without quickly drawing a conclusion. These are two different 'lifestyles', and essentially show a preference for either the NS or TF aspects of personality.

Many previous studies (Argamon *et al.*, 2005; Nowson and Oberlander, 2007) have used the so-called 'Big Five' personality factors: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Although exactly which personality profiling system is 'best' is a controversial subject, studies have shown that MBTI profiles correlate well with four of the 'Big Five' personality factors (McCrae and Costa, 1989).

Similarly, many researchers have attempted to infer personality from writing style. Some examples are Argamon *et al.* (2005) and Mairesse *et al.* (2007). Our study builds on earlier work on personality classification by Luyckx and Daelemans (2008b). By using the Tillburg Memory-Based Learner with a variety of parts-of-speech-based feature sets, Luyckx and Daelemans achieved fairly good results on the MBTI classification task, with an average accuracy of approximately 71% on a corpus, described later, of their own development. More detailed results from this experiment are provided later for comparison. Luyckx and Daelemans take their feature set from earlier work by Stamatatos *et al.* (2001), which suggests that this feature set is promising for the author identification task. By using correlation between MBTI and the 'Big Five' personality factors, Luyckx and Daelemans suggest that their overall performance is as good as, or better than, previous major experiments (Argamon *et al.*, 2005; Mairesse *et al.*, 2007; Nowson and Oberlander, 2007). Thus, the Luyckx and Daelemans method of MBTI personality profiling is the current best-performing stylometric personality profiling method as of this writing.

Some recent studies in the author identification task have suggested that even very simple feature sets can provide good performance for authorship attribution. Juola and Ryan (2008) propose that character N-gram models are some of the highest performing models in their system, JGAAP (jgaap.com). The Java Graphical Authorship Attribution Program (JGAAP) is an open source program for authorship attribution that provides a wide variety of feature extraction and analysis techniques within a modular framework. Furthermore, Noecker and Juola (2009) propose the use of simple dot-product-based nearest neighbor methods as the best performing analysis method in the same system. These methods thus appear to be the current 'best practices' methods for author identification. In this article, we assess the suitability of these state-of-the-art authorship identification methods for the task of personality profiling.

3 Materials and Methods

3.1 The personae corpus

For this study, we used the Personae corpus (Luyckx and Daelemans, 2008a). The corpus is approximately

200,000 words of Dutch-language text and consists of 145 individual student essays with controlled 'topic, register, genre, age, and education level.' Each essay is ~1,400 words and contains a factual description of a documentary about Artificial Life as well as the student's opinion of the documentary. Each essay is labeled with metadata about the student's gender, MBTI profile, mother tongue, and region. The distribution of the documents is as shown in Table 1.

3.2 Features and analysis

To perform the experiment, we used the JGAAP system for the feature extraction and R, a free software environment for statistical computing (r-project.org), for the analysis. We used a normalized dot-product nearest neighbor metric for the analysis method, as suggested by Noecker and Juola (2009). Before analysis, the documents were preprocessed to standardize whitespace and character case. Any sequence of whitespace characters in the document was converted to a single space, and all characters were converted to lower case. We tried a variety of feature sets including words and characters, as well as word and character N-gram models. As predicted in Juola and Ryan (2008), character N-grams were the best performers, with increased performance for larger N. We settled on character tetragrams to limit the size of the feature vectors while still providing good performance. We constructed a feature vector key using the character tetragrams, with each document represented as a 24,767 dimensional vector by the relative frequencies with which each tetragram appeared in the document.

Table 1 Corpus distribution

MBTI profile	Number of documents	Corpus (%)
E	80	55.2
I	65	44.8
N	78	53.8
S	67	46.2
T	40	27.6
F	105	72.4
J	117	80.7
P	28	19.3

3.3 Personality models

As with the Luyckx and Daelemans experiment, we chose to train models independently for each binary classification, rather than attempting to model the entire MBTI personality profile as a whole. We have taken two basic approaches to the problem of modeling each binary classifier. The first was to use the nearest neighbor classifier on a subset of documents from each of the binary choices for a personality aspect (e.g. E versus I) and assign a profile based on that of the document 'closest' to the test document in the character tetragram normalized dot-product feature space. The second approach involved the creation of a single, centroid-based model for each aspect of personality type. In this approach, the test document is assigned a category based on which centroid vector (e.g. E versus I) it was 'closest' to in the same space.

3.3.1 *Document model*

The first model is similar to that used in the authorship attribution experiments mentioned above (Juola and Ryan, 2008; Noecker and Juola, 2009). To avoid train-on-test bias, a random subset of documents was chosen for each experiment. Each of the four binary classifications was tested independently. A single test document was used, and at least one training document with the same MBTI binary class as the training document. A random number of training documents were also included to vary the size of the candidate and distractor sets. The test document was never included in the random training sample and was labeled according to a nearest neighbor cosine distance metric on the documents' feature vectors. One hundred thousand such experiments were run and an accuracy rating was calculated for each of the eight models, which we then compared with the models of Luyckx and Daelemans (2008b).

3.3.2 *Centroid model*

In this model, a centroid was calculated for each of the eight binary classifications. Each test document was compared with the centroid from each personality category and assigned to the category closest to it in the feature space. The centroid model was calculated by using the average relative frequency of

each event per document. Using this centroid model, we predict each of the four categories independently for all of the documents in the test corpus. To avoid bias or over-fitting, we use leave-one-out cross-validation. We then calculate the average precision, recall, and *F* score of each model. Our precision is calculated as the number of documents we correctly identified as belonging to a certain category over the total number of documents we identified as belonging to that category. The recall score is calculated as the number of documents we correctly identified as belonging to a category over the number of documents that actually belong in that category. The *F* score is provided as a measure of the test's accuracy and is the harmonic mean of precision and recall. These metrics were chosen as they mirror the data provided by Luyckx and Daelemans (2008b) and thus provide an easy method of comparison.

4 Results

4.1 Document model

Table 2 shows our accuracy on each of the eight MBTI classifications, as well as the accuracy from Luyckx and Daelemans. Our current results are, on average, an 11% increase over those of the previous experiment. Table 3 shows the accuracy broken down by category. Here, the current results are about a 9% increase from the results attained by Luyckx and Daelemans (2008b). Baselines for the by-category model are given by Luyckx and Daelemans (2008b). Random baselines were found by randomly assigning personality, whereas majority baselines are formed by always choosing the most common profiles.

Table 2 Document model accuracy by model

Model	Current result (%)	Luyckx and Daelemans (2008b) (%)
E	79.48	60.00
I	72.48	64.14
N	77.55	56.62
S	75.35	55.17
T	65.78	65.52
F	86.10	73.79
J	91.15	82.07
P	56.91	57.93
Average	75.60	64.41

Table 3 Document model accuracy by category

Category	Current result (%)	Luyckx and Daelemans (2008a,b) (%)	Majority Baseline (%)
Attitudes (E versus I)	76.35	65.52	55.20
Perceiving (N versus S)	76.54	62.07	53.80
Judging (T versus F)	80.47	73.79	72.40
Lifestyle (J versus P)	84.45	82.07	80.70
Average	79.45	70.86	65.53

4.2 Centroid model

Table 4 shows the precision, recall, *F* scores, and accuracy for each of the centroid models as well as the results previously published by Luyckx and Daelemans. Observe that the current results offer significantly higher precision and *F* score, with an average raw-percentage improvement of 23% and 10%, respectively, whereas the recall by Luyckx and Daelemans (2008b) was on an average ∼4% increase over the centroid method. Random baselines come from Luyckx and Daelemans (2008b).

Table 4 Centroid model accuracy

Model	Precision	Recall	*F* score	Accuracy
E	93	63	75	77
I	67	94	78	77
N	75	95	84	80
S	91	63	74	80
T	57	98	72	79
F	99	72	84	79
J	100	74	85	79
P	48	100	65	79
Average	79	82	77	79

5 Discussion and Conclusion

These results represent a noticeable improvement over previous attempts at the personality classification task by Luyckx and Daelemans (2008b). Our best performing methods result in a 41.6% relative

increase in precision and a 15.2% relative increase in *F* score, while suffering only a 4.5% decrease in recall over Luyckx and Daelemans (2008b). All of our results are very much better than chance, supporting our hypothesis that writing style is useful for predicting Myers-Briggs personality type indicator. Overall, we feel that these results represent an advancement in the state-of-the-art of stylometric personality profiling.

One particularly interesting aspect of these results is the simplicity of the methods used. Other studies in personality classification have used more linguistically complex feature sets such as use of part-of-speech based features by Luyckx and Daelemans (2008b). Stamatatos *et al*. (2001) have proposed that these kind of syntactic features are a better indicator of author style than token-level features like character tetragrams and have shown these more complex features to be useful for authorship identification. Our results support earlier work by Juola and Ryan (2008), showing some level of transference of the usefulness of simple feature sets like character tetragrams from authorship attribution to personality identification. Likewise, our use of the normalized dot-product nearest neighbor classifier means that our classification process is both simpler and much faster than previous methods. We believe that this is a significant benefit as these methods lend themselves well to future large-scale experiments in personality identification.

The experiment is not, however, without some shortcomings that lend themselves well to future work in personality profiling. The Personae corpus consisted entirely of Dutch-language text and was homogenous with respect to genre and author age and education level. Future work will focus on cross-genre techniques for personality profiling and on expanding these techniques for use with corpora in non-Dutch languages, particularly English. Juola (2008) found that many authorship attribution techniques that work well in English work equally well in other languages, particularly those that are most closely related to English. Thus, there is some basis to the belief that our technique will transfer well to personality profiling with English-language corpora. Still, we hope to expand upon our technique in these areas and hope to perform further analysis to explain exactly why the character tetragram feature set worked so well for this task.

Funding

This work was supported by the National Science Foundation (Grant No. OCI-1032683).

References

Argamon, S., Dhawle, S., Koppel, M., and Pennebaker, J. (2005). *Lexical Predictors of Personality Type, Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America.*

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, **16**(2): 9–17.

Brown, C., Covington, M. A., Semple, J., and Brown, J. (2005). *Reduced idea density in speech as an indicator of schizophrenia and ketamine intoxication. In International Congress on Schizophrenia Research*. Savannah, GA: International Congress on Schizophrenia Research.

Hancock, J. (2007). *Digital Deception: When, Where and How People Lie Online. Oxford Handbook of Internet Psychology*. Oxford, UK: Oxford University Press, pp. 287–301.

Juola, P. and Ryan, M. (2008). *Authorship Attribution, Similarity, and Noncommutative Divergence Measures. Selected Papers from the Chicago DHCS Colloquium*. Chicago, IL: Chicago Colloquium on Digital Humanities and Computer Science.

Juola, P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3): 233–334.

Juola, P. and Baayen, R. H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguist Computing*, **20**(Suppl): 59–67.

Koppel, M., Argamon, S., and Shimoni, A. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, **17**(4): 401–12.

Luyckx, K. and Daelemans, W. (2008a). *Personae: A Corpus for Author and Personality Prediction from Text. Proceedings of the 6th Language Resources and Evaluation Conference.* Marrakech, Morocco: International Conference on Language Resources and Evaluation.

Luyckx, K. and Daelemans, W. (2008b). *Using Syntactic Features to Predict Author Personality from Text. Proceedings of Digital Humanities 2008*. Oulu, Finland: Digital Humanities, pp. 146–9.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, **30**: 457–500.

McCrae, R. R. and Costa, P. T. Jr (1989). Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. *Journal of Personality*, **57.**

Myers, I. B. and Myers, P. (1980). *Gifts Differing: Understanding Personality Type*. Palo Alto, CA: Consulting Psychologists Press.

Newman, M., Pennebaker, J., Berry, D., and Richards, J. (2003). Lying words: predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, **29**: 665–75.

Noecker, J. Jr and Juola, P. (2009). *Cosine Distance Nearest-Neighbor Classification for Authorship Attribution, Proceedings from Digital Humanities 2009*. College Park, MD: Digital Humanities.

Nowson, S. and Oberlander, J. (2007). *Identifying More Bloggers: Towards Large Scale Personality Classi?cation of Personal Weblogs. International Conference on Weblogs and Social Media*. Boulder, CO: International Conference on Weblogs and Social Media.

Pennebaker, J., Mehl, M., and Niederhoffer, M. (2003). Psychological aspects of natural language use: our words, ourselves. *Annual Review of Psychology*, **54**: 547–77.

Rude, S., Gortner, E., and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, **18**: 1121–33.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, **35**(2): 193–214.