# Evaluating Content-Independent Features for Personality Recognition

Ben Verhoeven
CLiPS, University of Antwerp
Antwerp, Belgium
ben.verhoeven@uantwerpen.be

Juan Soler Company
Universitat Pompeu Fabra
Barcelona, Spain
juan.soler@upf.edu

Walter Daelemans
CLiPS, University of Antwerp
Antwerp, Belgium
walter.daelemans@uantwerpen.be

## ABSTRACT

This paper describes our submission for the WCPR14 shared task on computational personality recognition. We have investigated whether the features proposed by Soler and Wanner [10] for gender prediction might also be useful in personality recognition. We have compared these features with simple approaches using token unigrams, character trigrams and liwc features. Although the newly investigated features seem to work quite well on certain personality traits, they do not outperform the simple approaches.

## Keywords

personality recognition from text; author profiling; computational stylometry

## 1. INTRODUCTION

The goal of personality recognition is assigning a personality profile to the author of a text. This personality profile is usually conceptualised as consisting of five different traits (Big Five) that are present to a certain degree in an individual. [3]):

- Extraversion (EXT)
- Emotional stability (EMS)
- Agreeableness (AGR)
- Conscientiousness (CON)
- Openness to experience (OPN) .

Currently this task is most effectively handled using supervised machine learning methods; see e.g. [6, 5, 7] and references therein. These systems will make a binary decision whether the author's personality can be described by a trait or not, disregarding the degree in which each trait is present. Possible applications of this task include social network analysis and user modelling in conversational systems.

Recently, Soler and Wanner [10] introduced a small set of mainly content-independent features that accounted for state-of-the-art performance in gender prediction. We will investigate whether this set of features (Soler2014 features) can also achieve state-of-the-art performance in personality recognition.

## 2. DATA AND APPROACH

The WCPR14 organizers provided the participants with a training (348 texts) and test set (56 texts) of transcribed video blogs [1]. In our approach, we used the training set to perform tenfold cross-validation experiments with different types of features in order to tune the parameters of the machine learning algorithm and to establish which features work best for the detection of each trait. All machine learning experiments were performed using Scikit-learn's support vector machine algorithm [8].

The following often used feature types were used as default approaches:

- Token unigrams (frequency threshold 5).
- Character trigrams (frequency threshold 10).
- LIWC features [9].
- The previous three feature types combined.

All feature types above were used in the form of relative frequencies and their values are thus between 0 and 1. We compare these baseline feature types with a feature set (Soler2014) proposed by [10] that was shown to work well for gender prediction. As in previous research, we treated prediction of each trait as a separate binary classification problem. So on the basis of cross-validation, we decide on the optimal feature set for each personality trait, and these feature sets of the training data are then used to construct the final system and predict the class of the test set items.

### 2.1 Soler2014 Features

Our hypothesis is that given the fact that the Soler2014 features, introduced for gender identification, seem to capture the style of the authors rather than the content of the texts, they can possibly be applied successfully to similar tasks such as personality recognition too. If the style of the authors can be analyzed to distinguish between genders, maybe it can be used to differentiate between personalities as well.

*WCPR'14,* November 7, 2014, Orlando, FL, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3129-6/14/11 ...$15.00.
http://dx.doi.org/10.1145/2659522.2659527.

The features that were used can be classified in five different groups. In each of these groups, the texts are analyzed at a different level, starting from the simplest one (analyzing the characters) and advancing up to sentence structure. The set of features that was used in this work is a superset of the features described in [10]. The five types of features in question are: Character-based, Word-based, Sentence-based, Dictionary-based and Syntactic features.

**Character-based features** capture the frequency of punctuation marks, upper case characters, some other characters such as hyphens, quotes or parenthesis and the total number of characters per text among others. The use of some of these features is to a major extent motivated by individual stylistic preferences, therefore, it could be related to a certain type of personality of the authors.

**Word-based features** analyze the words and their structure. Some of the features that are in this group are the total number of words per text, the vocabulary richness, the mean value of characters per text, the mean number of proper nouns per text, the percentage of words that are stop words or the usage of acronyms. Some of these features like vocabulary richness have proven very useful for Gender Identification and could be a factor in this task as well.

**Sentence-based features** only measure the number of sentences in a text and the number of words per sentence. This is a superficial analysis of the sentences, a more in-depth approach is done in the Syntactic features group.

**Dictionary-based features** measure the frequency of specific words in the analyzed texts. The first two dictionaries that are used are polarity dictionaries containing words that are either emotionally positive or negative. These dictionaries were used for the first time in [4] and contain approximately 6800 words classified by their polarity. The features that are extracted using these dictionaries are the percentage of words of the texts that are polarity words. The usage of discourse markers, interjections, curse words and abbreviations is also measured.

**Syntactic features** use the dependency parser described in [2] to measure the structure of the sentences of the text by analyzing the dependencies between words. In this group of features the usage of each one of the dependencies is measured as well as the length of these dependencies (defined as the distance between the head and the dependant in words). Since the dependencies between words form a dependency tree, the shape of these dependency trees can be used to characterize the writing of the authors.

The width and depth of these trees can be seen as a metric on how complex the sentences are, and can be useful to characterize the style of the authors. In this group of features, the number of different dependencies per sentence is also measured.

The total number of gender identification features that were used is 98. This is a fairly small number of features that are mostly content-independent (only the dictionary based features depend directly on the content of the texts). Table 1 displays the number of features of each group.

Because the values of all these features have different ranges, which makes learning harder for some machine learning algorithms (e.g. SVM), we have standardised these by mean subtraction and standard deviation division.

| Feature Category | | #Features |
|---|---|---|
| Character-based | C | 16 |
| Word-based | W | 7 |
| Sentence-based | S | 2 |
| Dictionary-based | D | 6 |
| Syntactic | Y | 67 |

**Table 1: Number of features per group**

# 3. RESULTS

In order to establish which feature sets and which parameters to use, we performed tenfold cross-validation on the training set of the provided dataset. Table 2 shows the results of these experiments and the baselines they can be compared with (majority baseline and weighted random baseline (WRB)). The $C$ and $gamma$ parameters that we used with the SVC algorithm for each system can be found in Table 3. The feature sets have been coded in the tables as described below. Their letter codes are combined for experiments where different feature sets were used together (e.g. UCL is a system trained on token unigrams, character trigrams and liwc features).

**U** - token unigrams

**C** - character trigrams

**L** - LIWC features

**S** - Soler2014 features

The results of the tenfold cross-validation show us that we can beat both baselines for EXT, EMS and CON and that we can beat the weighted random baseline for OPN. AGR, however, has a more skewed distribution and appears harder to learn. Although the Soler2014 features can beat at least one baseline for three of the traits, they perform not as good as the simpler feature sets we used. The best performing system for each trait is indicated in bold.

For our final submission, we first used all the systems described in Table 2, trained them on the train data and predicted the output classes of the test data. The results of these experiments can be found in Table 4. Despite there being some really good results in this table (the best results are indicated by italics), it would be methodologically incorrect to cherry-pick these for our final submission. Our choice should depend on the training data only, no knowledge of the test data can be used for this decision. The results matching the best-performing systems on the tenfold are indicated in bold.

The final results, as calculated with the scoring script provided by the workshop organizers, can be found in Table 5.

| Class | P(Avg) | R(Avg) | F1(Avg) |
|---|---|---|---|
| Extra | 0.49 | 0.49 | 0.48 |
| Neuro | 0.61 | 0.61 | 0.61 |
| Agree | 0.69 | 0.68 | 0.68 |
| Cons | 0.45 | 0.46 | 0.45 |
| Open | 0.50 | 0.50 | 0.49 |
| Avg | 0.55 | 0.55 | 0.54 |

**Table 5: Official results of submission**

|  | Baselines |  | Systems |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | WRB | Majority | U | C | L | UCL | S | US | LS | ULS | CS | CLS | UCS | UCLS |
| EXT | 51.3 | 58.0 | **59.6** | 59 | 58.4 | 59.4 | 53.2 | 57.7 | 53.2 | 57.7 | 57.7 | 57.7 | 57.7 | 57.7 |
| EMS | 50.5 | 54.9 | 61.1 | 59.2 | **63.4** | 58.9 | 60.6 | 60.2 | 60.6 | 60.2 | 60.2 | 60.2 | 60.2 | 60.2 |
| AGR | 66.5 | 78.7 | 53.9 | 56.3 | 52.3 | **57.7** | 44.9 | 56.7 | 44.9 | 56.7 | 56.7 | 56.7 | 56.7 | 56.7 |
| CON | 50.6 | 55.6 | 53.7 | 56.6 | **60.1** | 56.6 | 59.6 | 55.7 | 59.6 | 55.7 | 55.7 | 55.7 | 55.7 | 55.7 |
| OPN | 54.3 | 64.7 | 48.5 | 48.8 | 54.4 | **56.0** | 46.7 | 47.9 | 46.7 | 47.9 | 47.9 | 47.9 | 47.9 | 47.9 |

**Table 2: F-scores of all tenfold cross-validation experiments**

|  | Systems |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | U | C | L | UCL | S | US | LS | ULS | CS | CLS | UCS | UCLS |
| $C$ | 2048 | 2048 | 2048 | 2048 | 8 | 32 | 8 | 32 | 2048 | 2048 | 2048 | 2048 |
| $gamma$ | 0.5 | 0.5 | 0.5 | 0.5 | $2^{-9}$ | $2^{-7}$ | $2^{-9}$ | $2^{-7}$ | $2^{-7}$ | $2^{-7}$ | $2^{-7}$ | $2^{-7}$ |

**Table 3: Parameters of the systems in Tables 2 and 4, chosen on basis of the train data**

|  | Baselines |  | Systems |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trait | WRB | Majority | U | C | L | UCL | S | US | LS | ULS | CS | CLS | UCS | UCLS |
| EXT | 51.3 | 58.0 | **47.6** | 52.2 | *57.6* | 50.5 | 43.9 | 47.6 | 43.9 | 47.6 | 47.6 | 47.6 | 47.6 | 47.6 |
| EMS | 50.5 | 54.9 | 64.1 | 59.4 | **60.5** | 61.5 | *67.7* | 60.5 | 67.7 | 60.5 | 60.5 | 60.5 | 60.5 | 60.5 |
| AGR | 66.5 | 78.7 | 61.4 | 50.2 | 53.6 | ***68.3*** | 49.0 | 53.5 | 49.0 | 53.5 | 53.5 | 53.5 | 53.5 | 53.5 |
| CON | 50.6 | 55.6 | 51.3 | 47.6 | **45.2** | 48.1 | *57.3* | 50.5 | *57.3* | 50.5 | 50.5 | 50.5 | 50.5 | 50.5 |
| OPN | 54.3 | 64.7 | 49.3 | *52.4* | 46.5 | **49.4** | 40.8 | 46.9 | 40.8 | 46.9 | 46.9 | 46.9 | 46.9 | 46.9 |

**Table 4: F-scores of all train-test experiments**

We then investigated the performance of the Soler2014 features per category (see Table 1). We first performed experiments using all possible combinations of feature categories for each trait. The same algorithm and parameters were used as for the experiment above on all the Soler2014 features. The best result for each trait together with the best result from the earlier experiments are shown in Table 6.

| Trait | best from category | best from above |
|---|---|---|
| EXT | 55.8 | 59.6 |
| EMS | 63.8 | 63.4 |
| AGR | 44.9 | 57.4 |
| CON | 61.5 | 59.6 |
| OPN | 47.3 | 53.2 |

**Table 6: Best results of category experiments and all experiments above**

We can observe that no combination of feature categories of the Soler2014 features works for the AGR and OPN traits. For the other traits, more promising results were found. We find that certain category combinations perform better - using these parameter settings - than combining all categories as seen above.

In order to estimate the importance of the feature categories for each of the three well-performing traits, we have averaged over the F-scores of all the experiments using that category and plotted these averages for each category and each trait. The best results from the earlier experiments for each trait are included as a reference. This plot can be found in Figure 1 and it shows the importance of the dictionary-based features (D) for the CON and EMS traits and the syntax-based features (Y) for the CON and EXT traits. Other feature categories appear to have much less of an influence on the results.

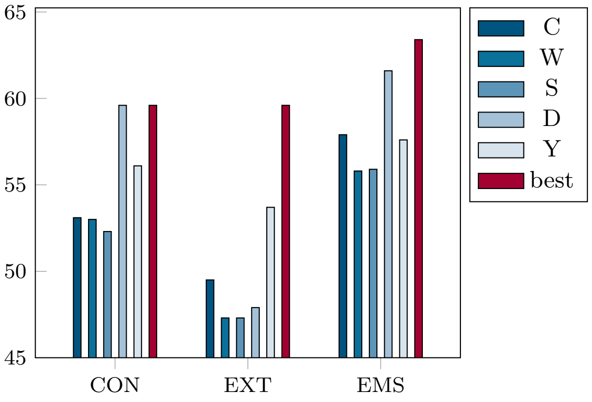

**Figure 1: Per category plot of trait scores**

## 4. DISCUSSION AND CONCLUSION

We have investigated whether content-independent features such as those proposed by Soler and Wanner [10] for gender prediction can also achieve good results for personality recognition. Based on our experiments we can say that these features do indeed look interesting, however further research is essential because we have our doubts about the validity of the results given the size of the dataset.

It seems that the provided dataset consisting of video blog transcripts is too small for the task of personality recognition. Previous research has shown that this task is very hard even with large amounts of data. We have noticed that many eager machine learning algorithms cannot learn a model from this data with their standard parameters. This was also the reason for working with different tuned parameters.

However, tuning the parameters probably means that all our systems are overfit. Evidence for overfitting can be found in the high variability between the tenfold and train-test results, despite the fact that both train and test data come from the same corpus and are similar texts. The same system only performs best for one trait (AGR), which is the one that does not reach baseline.

## 5. ACKNOWLEDGMENTS

This research was funded by the Flemish government agency for Innovation by Science and Technology (IWT) through SBO (Strategic Basic Research) project AMiCA.

## 6. REFERENCES

[1] J.-I. Biel and D. Gatica-Perez. The YouTube Lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. In *IEEE Transactions on Multimedia*, volume 15, pages 41–55, 2013.

[2] B. Bohnet. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[3] L. R. Goldberg. An alternative "Description of Personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, 1990.

[4] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[5] K. Luyckx and W. Daelemans. Personae: a corpus for author and personality prediction from text. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008. European Language Resources Association.

[6] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.

[7] J. Noecker, M. Ryan, and P. Juola. Psychological profiling through textual analysis. *Literary and Linguistic Computing*, 2013.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] J. W. Pennebaker, M. Francis, and R. Booth. *Linguistic Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum, Mahwah, NJ, USA, 2001.

[10] J. Soler Company and L. Wanner. How to use less features and reach better performance in author gender identification. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014.