Look! Who's Talking? Projection of Extraversion Across Different Social Contexts

Scott Nowson
Xerox Research Centre Europe
6 chemin de Maupertuis, Meylan 38240, France
scott.nowson@xerox.com

Alastair J. Gill
Department of Digital Humanities
King's College London
26-29 Drury Lane, London WC2B 5RL, UK
alastair.gill@kcl.ac.uk

ABSTRACT

Automatic classification of personality from language depends upon large quantities of relevant training data, which raises two potential problems. First, collecting personality information from the author or speaker can be invasive and expensive, especially in sensitive contexts. Second, issues of context or genre can reduce the usefulness of available training resources for broader personality classification. One approach to dealing with the first issue is to use external judges rather than the text's author. In this paper, we test the extent to which these personality perceptions are useful for training a classifier between different linguistic genres. Following disappointing cross-training results, we explore the projection of personality through specific linguistic factors. We find that while some differences are between the genres overall, some indicate that indeed personality is evidenced differently across situations. It is clear that care is needed leveraging resources from different domains for computational personality recognition.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; J.4 [**Social and Behavioral Sciences**]: Psychology; I.2.6 [**Artificial Intelligence**]: Learning

Keywords

Personality recognition, classification, text

1. INTRODUCTION

Personal Language Analytics is a branch of text mining in which the object of analysis is the author of a document rather than the document itself. Language use in text can reveal a lot about a person. One of the most important individual differences is personality, arguably one of the central tenets of the notion of self. Personality is a valuable source of information for applications such as user modeling or social media engagement.

WCPR'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2659522.2659530.

Yet, as a field of study, computational personality recognition is still considerably under-explored. One of the stumbling blocks for the field is the availability of quality labeled data. While there are some datasets available, it is difficult to construct new resources. Collecting ground truth labels requires active participation of a large volume of individuals which can be expensive and is often considered sensitive.

One alternative approach to data collection is the use of personality perception – wherein the personality labels are judgments made by third parties. One of the core datasets of this workshop is one such corpus: social video blogs ([1], described in more detail later).

Another issue with working in automatic personality classification is understanding how the manifestation of traits varies between data sources. The focus of this work is therefore not strictly on simple classification. Rather, the intention is to ask how the features which convey personality in one genre map to the projection of personality in another.

This paper explores the differences between the social video data with a related dataset - the EAR corpus of recorded conversation [7]. Following a discussion of supporting literature, the paper reports two exploratory comparative studies. First, personality models developed on the vlogs are applied to classify the EAR corpus. Second, a more in-depth analysis of the variance of features within the monologue and dialogue corpora is performed. Results suggest that while there are inherent differences between the datasets themselves, it does appear that personality is projected in a fundamentally different way between corpora.

2. BACKGROUND LITERATURE

From the earliest works of computational personality recognition in 2006 ([9], [6]) through to this workshop, there has been a significant body of work. In the interests of space, here we look at two specific considerations: combining multiple data sources, and impressions of personality.

There have been some attempts to work outside the confines of data availability. Following earlier work on a small, carefully constructed corpus of personal blogs [9], [8] identified a larger source of courser-grained data: results of a personality quiz meme consisting of five binary questions per trait which thousands of people posted to their blogs. Statistically indicative features (word n-grams) were drawn from the clean corpus and used to classify the larger. Results were promising, even for small feature sets (< 40 items).

Another approach is to leverage ensemble learning methods across distinctly different genres, as per [12]. In this work, a classifier trained on essays was used to augment the

results of a classifier trained on Facebook status updates. The final proof-of-concept meta-learner was used to determine the final labels for the Facebook data with considerably improved performance over stand-alone classification.

In general, human perception of personality is more accurate when judges are presented with a greater volume of more varied data. However, studies which have examined relatively small amounts of online data – for example short email or blog texts or Facebook profiles – have found that Extraversion is consistently the most accurately perceived trait. Judges not only show agreement with each other, but also with self assessments also used in the study [3, 5]. In addition, Facebook members were also aware of how Extraversion is expressed in their profile pages [3].

These online results mirror those found for personality perception (and expression) in daily life: not only is Extraversion accurately perceived from listening to a sampled recording of daily activities, but that extravert behaviour is closely related to 'implicit folk theories', such as talking more and socialising [7].

3. DATA

This paper employs two datasets:

1. a corpus of manually transcribed, personal video blogs (henceforth 'Vlog'), which are monologous in nature
2. a corpus of manually transcribed, conversational dialogue (EAR)

Table 1 summarises some of the key statistics of the two corpora. Both are relatively balanced for gender and are not constrained by topic. This should minimise any potential interactive effects of these factors.

Table 1: Summary statistics of the two corpora

Property	Vlog	EAR
Genre	monologue	dialogue
Timing	continuous	sampled
Participants	404	96
Rating mechanism	10 item TIPI	44 item BFI FFM
Judges per subject	5	5-7
Total word count	236762	96277
Average WC	586	1003
Reliability (ICC[1,k])	.76	.84

Note for EAR corpus ICC is average reported across all traits

3.1 Monologue Video - Vlog

The primary data for this workshop is the Vlog corpus of [1]. Personality perception labels were collected via crowdsourcing on Amazon's Mechanical Turk platform. The judges were self-selected from the varied population of users, or 'Turkers.' The inventory used was the TIPI [4], whose ten items are designed and validated for personality in constrained environments. The judges, five per video, completed the items after viewing only the first minute - leading to 'first impression' style assessment of personality.

3.2 Dialogue Audio - EAR

The comparison data for this study comes from the study of [7], which used the EAR recording system to sample the spoken language of 96 participants over a two day period. In addition to the participants themselves completing a personality questionnaire (BFI-FFM), 5-7 judges listened to the whole recording and then rated the participant's personality using the same measure. Unlike in the Vlog corpus, the same judges did all ratings here. Mehl and colleagues [7] note relatively high agreement within judges (we use their average score in the current analysis) and also between the judges and the participant's self-ratings.

3.3 Trait selection

Though both corpora are labeled with all traits of the five factor model, for this initial study the focus is on Extraversion (E). There are a number of reasons for this:

- Extraversion is the most accurately perceived and best understood trait. As such, it is also the most commonly studied trait.
- Judge reliability is strongest in the Vlog corpus for Extraversion (the trait returns the highest Intraclass Correlation Coefficient, reported as ICC in table 1).
- Extraversion is considered the social trait, and since both corpora are social in nature (EAR is conversational, while vlogging is a form of social media) we expect significant evidence of trait projection.

4. METHOD

4.1 Coding Classes of Personality

For the purposes of this workshop, personality traits are considered here as discrete variables: each trait consists of a high- and low-scoring group, with those between considered mid-scoring. This is in part to account for the smaller EAR dataset, but it is also an attempt to abstract away from any differences in the differing personality models.

These three classes were determined by statistically derived boundaries on the judge rating scores. The high group consists of those who scored *greater* than the mean score *plus* half the standard deviation, while the low group is conversely scores *less* than the mean *minus* half the standard deviation. The mid group is those who scored between the previous values.[1] In both corpora this creates an effectively balanced corpus across the three classes. The balance is reasonably well maintained across gender.

4.2 Classification

For this exploratory study, the textual features to be used are those provided by the Linguistic Inquiry and Word Count tool (LIWC 2007, [11]). This feature set is commonly used in personality studies as they lend themselves well to psychological interpretation.

Different from [8] who developed a model based on the smaller data set to classify the larger, this work applies models in the more natural direction of building on the larger while testing on the smaller. However, similarly to [12] the focus of this work is not in optimizing the approach for maximum results. As previously mentioned, the intention is an initial exploration as to the similarities and differences of our two chosen corpora.

In order to examine the utility of using the Vlog dataset to classify the EAR data two experiments were performed:

[1]extreme classes: $mean + / - (0.5 * StDev)$

1. 10-fold cross-validation on the EAR corpus as a baseline

2. test the same 10-folds of EAR data on a model trained on all of the Vlog data

Classification models are built using Naïve Bayes[2] for both binary (high-low) and ternary (high-medium-low) tasks.

5. RESULTS

5.1 Classifier Accuracy

Table 2 contains the accuracy results of the classification experiments. Along with a majority class baseline, the results from both the 3-class (Vlog n=404, EAR n=96) and 2-class (Vlog n=270, EAR n=62) models are reported.

Table 2: Classification results on the EAR corpus

Task	Ternary (HML)	Binary (HL)
Baseline	35%	55%
Crossfold validation	47%	84%
Cross Training	41%	62%

As previously stated, optimization of results is not the focus of this work. However, it is worth noting that this pattern of results is observable with a number of different classification algorithms.

5.2 Analysis

In order to better understand the differences found between the two datasets, we performed post-hoc Spearman correlations between the three levels of Extraversion and the percentage usage of the LIWC variables. In addition, the relative frequency ratio of use of features between the corpora was examined.

Given the multiple correlations and risk of false positives (Type I error), in Table 3 we only discuss LIWC features which achieve a relationship with Extraversion of $p \leq 0.01$ in either the Vlog or EAR corpus.

We note in interpreting this data that the Interpersonal and Contextual features tend to be used less frequently across both corpora (often less than 1 per cent mean usage) and so due to their scarcity may be less reliable predictors of personality than those more frequently used features, such as Psychological or General features.

6. DISCUSSION

The cross-validation results on the EAR corpus show the best performance, particularly in the binary task. This is understandable given the limited size of the corpus. It is worth noting, however, that across the classification folds results were considerably varied.

The performance of the cross-training experiments is worse, though this is expected to a degree. However, while they are above baseline, the degree of performance is greater than might be expected, especially in the binary case. Given the limited size of the feature set used for classification, this suggests potentially significant differences with the data when represented at this level.

[2]as implemented in scikit-learn [10]

Table 3: Spearman correlations between LIWC variables and Extraversion levels and corpus relative usage indicators

LIWC category	E Vlog	E EAR	Vlog/EAR Ratio
General			
WC	0.07	0.62***	<
Dic	-0.20***	-0.10	
funct	-0.17***	-0.10	
assent	0.16***	-0.10	<<
Interpersonal			
i	-0.02	0.03	
we†	0.04	0.19	<
you	0.14**	-0.06	
shehe‡	0.10*	0.27**	<<
they†	-0.06	-0.07	<
ipron	-0.15**	-0.09	
family†	0.08	0.38***	>
friend†	0.01	0.24*	>
Psychological			
affect	0.11*	0.28**	
posemo	0.09	0.15	
negemo	0.03	0.30**	
cogmech	-0.21***	0.08	
Contextual			
sexual†	0.15**	0.25*	
swear†	0.09	0.28**	<<
relig†	0.13**	0.32**	>
death†	-0.02	0.27**	
space	0.13**	-0.07	

† indicates LIWC categories showing a mean frequency of less than 1 per cent across both corpora; ‡ indicates frequency of less than 1 per cent in the Vlog corpus
'<<' indicates a relative frequency ratio of ≤1:2 in the direction indicated; '<' indicates a ratio of ≤4:5

In terms of correlations found for LIWC features, there are a number of similarities between the corpora. For example, in both datasets people perceived to be extravert are more likely to use words relating to personal Contextual categories such as 'religion' and 'sexual' ($p \leq 0.05$ for EAR data) than those perceived as introverted. Also extraverts are more likely to use general 'affect' language ($p \leq 0.05$ for Vlog).

For Interpersonal language, in the EAR data (mirrored in the Vlogs) there is a greater use of third person pronouns by extraverts. In addition, the EAR shows a greater number of references to family (also to friends at a $p \leq 0.05$ level). These appear to show evidence of the perception of extravert sociability, and how to some extent this can vary across genres. A further example in the Vlog data is that perceived extraverts make more direct reference to others (second person pronouns, e.g., 'you'), which matches known extravert behaviour of addressing an audience previously found for written blogs ([2]). This is a clear example of variance in personality projection across genres and mediums. While spoken vlogs are similar to written blogs (by-and-large both monologues), they are different to spoken dialogues.

Crucially, however, there are a number of differences which go beyond minor variance within sub-categories, as in the case of Interpersonal language. One key difference can be attributed to data collection methodology. Volume of language use (e.g. speech rate, word count) frequently evidences the

trait of Extraversion. Extraverts talk significantly more than introverts in the unconstrained environment recorded in the EAR data. In fact as [7] stated, this relationship is even stronger in perception scores than with ground truth values for Extraversion. However, there is no such effect in the Vlog corpus. This is most likely because impressions were based on only a thin-slice (the first minute) of the data, while transcriptions are provided for the entire video.

There are also differences which appear to be due to the general nature of the language used within each context. Swearing, for example, correlates strongly with Extraversion in the EAR corpus, but there is no such effect in the Vlogs. By examining the relative usage, it is clear that swearing is considerably underused in the Vlog corpus. This would suggest, perhaps, that in general swearing in this domain is less acceptable than every day conversation. So, while it is theoretically possible that this may still be a useful feature for some outliers, overall the strength of the signal is muted.

Contrast this with feature relationships which do not present in both corpora. Like swearing, the use of negative emotion language (in particular, the 'anger' category, not shown here) is a strong indicator of Extraversion in the EAR data. Unlike swearing, however, there is no such usage difference for the expression of negative emotion, suggesting that there is no effect for this variable with Extraversion in Vlogs.

Similar effect differences can be seen in the differing relationships between Extraversion and a number of other categories. Extravert vloggers, for example, show a reduced use of 'cognitive mechanism' language, perhaps indicating a greater certainty and reduced hedging (e.g., 'I think'), or coverage of less intellectual subjects that is not useful evidence of Extraversion in dialogue.

7. CONCLUSION AND FUTURE WORK

This work is a preliminary study on the classification of a single variable comparing data from different sources. Despite the seeming similarities between the data sets – at least compared to others available – there are significant differences. For computational personality recognition, while it is disappointing that overall levels of features vary between genres, this is merely a loss of signal strength. This implies a lower confidence when applying models across datasets.

Of greater concern are the instances where the feature signal relates differently to the trait of interest: from a straightforward effect absence/presence, to effect reversal. These more serious differences thus serve to warn for the potential of introducing errors in classification.

In future work, we would be keen to explore in more detail the personality perceptions resulting from additional dialogue and monologue corpora. In addition, a deeper analysis of the effect on classification of all traits of a broader range of linguistic variables is required.

Continuing the work of [7], we would also seek to establish parallel ground truth values for personality traits on collected data. This would allow us to establish relationships between those aspects of language which influence perception and those which have direct correlation with true personality. Exploring these differences will further our understanding of how aspects of personality – both projected and perceived – vary across social contexts.

8. REFERENCES

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

[2] A. J. Gill, S. Nowson, and J. Oberlander. What Are They Blogging About? Personality, Topic and Motivation in Blogs. In *ICWSM 2009*, 2009.

[3] S. D. Gosling, S. Gaddis, and S. Vazire. Personality impressions based on Facebook profiles. In *Proceedings of ICWSM 2007: the International Conference on Weblogs and Social Media*, 2007.

[4] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[5] J. Li and M. Chignell. Birds of a feather: How personality influences blog writing and reading. *Int. J. Human-Computer Studies*, 68:589–602, 2010.

[6] F. Mairesse and M. A. Walker. Words Mark the Nerds: Computational Models of Personality Recognition through Language. In *28th Annual Conference of the Cognitive Science Society*, Vancouver, July 2006.

[7] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862–877, May 2006.

[8] S. Nowson and J. Oberlander. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *In Proceedings of the International Conference on Weblogs and Social*, 2007.

[9] J. Oberlander and S. Nowson. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of COLING/ACL-06: 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*, July 2006.

[10] 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.

[11] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The development and psychometric properties of LIWC2007. This article is published by LIWC Inc, Austin, Texas 78703 USA in conjunction with the LIWC2007 software program., 2007.

[12] B. Verhoeven, W. Daelemans, and T. De Smedt. Ensemble Methods for Personality Recognition. In *Proceedings of WCPR13, Workshop on Computational Personality Recognition at ICWSM13 (7th International Conference on Weblogs and Social Media)*, 2013.