Too Neurotic, Not Too Friendly: Structured Personality Classification on Textual Data

Francisco Iacobelli and **Aron Culotta**
Department of Computer Science
Northeastern Illinois University
Chicago, IL 60625

Abstract

Personality plays a fundamental role in human interaction. With the increasing amount of online user-generated content, automatic detection of a person's personality based on the text she produces is an important step to labeling and analyzing human behavior at a large scale. To date, most approaches to personality classification have modeled each personality trait in isolation (e.g., independent binary classification). In this paper, we instead model the dependencies between different personality traits using conditional random fields. Our study finds a correlation between Agreeableness and Emotional Stability traits that can improve Agreeableness classification. However, we also find that accuracy on other traits can degrade with this approach, due in part to the overall problem difficulty.

Introduction

Personality plays a fundamental role in human interaction. Therefore, with the increasing amount of information that users post online, the ability to automatically detect personality can provide insights into the cognitive processes of a large number of individuals. In particular, one piece of information that internet users generate in large quantities is text: status updates, tweets, blogs, reviews, etc.

Costa and McRae (1992) describe five personality traits in a continuous scale: extroversion, neuroticism (or its inverse, emotional stability), agreeableness, openness to experience and conscientiousness. These traits have become the de-facto personality traits for automatic classification.

There have been many attempts to test classification algorithms for personality in textual data, but in this paper we include two algorithms that have not been reported previously. These correspond to a structured approach to classification –that is, they model the dependencies among output labels as well as the input features. For example, they may learn that high neurotics tend to be less friendly and therefore use that to classify one of those traits more accurately.

In this paper we are not concerned so much with the best feature set to obtain the best accuracy. Instead our interest is to compare the structured approach to sensible non-structured approaches commonly used for text classification.

Background

The automatic classification of personality has been the focus of many studies. Although some studies have examined speech (Mehl, Gosling, and Pennebaker 2006), most of the results reported by these studies relate specifically to written language, and have applied both data-driven words and phrases (e.g., Oberlander and Gill, 2006; Nowson, 2006) and lexical features grouped by psychological categories (e.g., Pennebaker and King, 1999; Gill, Nowson, and Oberlander, 2009).

Studies using groupings of lexical features have adopted the LIWC dictionary (Pennebaker and Francis 1999) or a subset of it to compare baseline classification algorithms to Naïve Bayes and sequential minimal optimization algorithm (SMO) for a support vector classifier (Witten and Frank 2005; Platt 1999). Grouping features that were most related to style, Argamon and colleagues (2005) analyzed the same essay corpus used by Pennebaker and King (1999). Their best accuracy scores were around 57-60% for extroversion and neuroticism using SMO. Mairesse and colleagues (2007) tested several classifiers on that same data set using a different subset of LIWC features and obtained accuracies around 55%–62% using SMO. Their best accuracy scores were on Openness and the worst were on Conscientiousness. In addition they performed a similar analysis of conversational data that resulted in accuracies of around 64%–74% (Mairesse et al. 2007).

Other studies have explored structural and lexical features of a corpus of emails (approximately 9,800 messages) which were used to classify several dimensions, including the five personality traits (Estival et al. 2007). Although accuracies were in the range of 53–57%, this study used the largest corpus for personality classification of which we are aware.

Feature representations have also included structural information given by n-grams. For example, Oberlander and Nowson (2006) tested n-grams in a small corpus of blogs. SMO yielded the best accuracies (around 83–93%) and the best classification was on Conscientiousness. They did not include Openness in their report. However, when these trained classifiers were applied to a much larger corpus, the accuracies dropped to approximately 55%, which may have been a result of over-fitting (Nowson and Oberlander 2007).

Lastly, Iacobelli et. al. (2011) used the large corpus of blogs from (Nowson and Oberlander 2007) and a bigram

based feature representation and obtained accuracies between 70%–84%. The best classification was for Openness and the worst for emotional stability. As it is the case with Oberlander and Nowson, the authors warned of potential over-fitting of the data. It is important to note that the data contained posts from individuals over a long period of time and that seems to influence the accuracy scores.

In addition to classification, all of these studies have found overlap between the features that describe the different personality traits. This overlap is consistent with some correlations between personality traits that have been documented over a long period of time (Richardson 1968; Ode and Robinson 2009). The intuition of this paper is that those correlations may help boost classification. Therefore, we are more concerned with introducing a structured classification approach (Lafferty, McCallum, and Pereira 2001) to assess the helpfulness of these correlation between traits, rather than finding the feature set that best classifies the data.

Methodology

Corpus

The corpora for our experiment consisted of (a) the essays corpus produced by Pennebaker and King (1999) in which 2,469 individuals wrote a stream of consciousness essay for 20 minutes; and (b) the myPersonality corpus which consists of 9,918 Facebook status updates and social network metrics associated to the authors of posts.

Both corpora contain binary judgments as to whether the author of the text is a high or a low scorer of each of the Big Five personality traits. For the essays, those judgments were derived from the z-scores computed for each trait on a personality test (John, Donahue, and Kentle 1991) taken by each participant. In the case of the myPersonality corpus, binary classes were derived from the average scores on a personality test. In this set, well known proper names, such as "Chopin" and "Mozart", and locations, such as "New York" and "Mexico," were kept, while lesser known entities were replaced with a common token.

Data sets

To prepare our data, we treated each corpus as follows: all words were converted to lower case, punctuation symbols were treated as individual types and emoticons (e.g. *:-)* , *:P*) were treated as a single type.

Because multiple posts in the myPersonality corpus could belong to a single author, we merged all the posts that corresponded to one author and treated them as a single document. The reason for doing this is that we did not want to confound the personality classification with authorship information. This resulted in 251 documents.

Lastly, unigrams, bigrams and trigrams were combined as features for each document. Of these, we did not consider features occurring fewer than 5 times and we also eliminated the 20 most frequent.

Classifiers

Because previous research suggests that some traits of personality may be correlated with others, we decided to try structured approaches to personality classification.

Structured classification is an approach to classification in which the dependencies between output variables (e.g. personality traits) are modeled. For example, an decrease in emotional stability may be correlated with a decrease in agreeableness.

Conditional random fields (CRFs) (Lafferty, McCallum, and Pereira 2001) are a widely-used model for structured classification, and have been successfully applied to other multi-label text classification problems (Ghamrawi and McCallum 2005).

CRFs can be understood as the structured analog of logistic regression. For text analysis, given a sequence of words $\{word_1, word_2, ..., word_n\}$ and a set of labels $\{label_1, label_2, ..., label_k\}$, CRFs define the probability of seeing a specific combination of labels (output variables) given that a sequence of words (input variables) has been observed. More generally, for any set of input variables $\mathbf{x}$ and output variables $\mathbf{y}$, CRFs define the following conditional probability:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_k f_k(\mathbf{x}, \mathbf{y}, \Lambda)$$

where $f_k : X \times Y \mapsto \mathbb{R}^+$ is a *factor* with model parameters Λ; and $Z(\mathbf{x})$ is a normalization constant. The set of factors determines which dependencies are modeled. Each factor is an exponentiated dot product of features and parameters: $f_k = \exp\left(\sum_j \lambda_k^j \phi_k^j(\mathbf{x}, \mathbf{y})\right)$ where $\phi : X \times Y \mapsto \mathbb{R}$ computes features over $\mathbf{x}$ and $\mathbf{y}$, each of which is weighted by a corresponding $\lambda_k^j \in \Lambda$.

Here, $\mathbf{x}$ is a word vector, and $\mathbf{y}$ is the vector of binary labels (i.e., the five personality traits). The parameters are optimized to maximize the joint accuracy of these five traits. Contrast this approach to a logistic regression model, which ignores the dependencies between labels, where a multi-label problem with m classes would be modeled by m independent logistic regression models:

$$p(y_1|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} f_{lr}(\mathbf{x}, y_1, \Lambda)$$
$$\dots$$
$$p(y_m|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} f_{lr}(\mathbf{x}, y_m, \Lambda)$$

Following (Ghamrawi and McCallum 2005), we consider two types of factors:

1. **Pairwise Label Factors (CRF-P)**: Includes factors for the interaction between each pair of labels: $f_p(y_i, y_j, \Lambda)$ $\forall i, j$. This introduces $\binom{5}{2} = 10$ factors. These are combined with the logistic regression factors $f_{lr}(y_i, \mathbf{x})$ $\forall i$.

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{i,j} f_p(y_i, y_j, \Lambda) \prod_i f_{lr}(\mathbf{x}, y_i, \Lambda)$$

Given that each y_i is a binary variable, this introduces an additional $(10 \times 2^2) = 40$ parameters over standard logistic regression.

2. **Pairwise Label-Features Factors (CRF-PF)**: Includes factors for the interaction between each pair of labels, conditioned on the observed variables $\mathbf{x}$: $f_{pc}(y_i, y_j, \mathbf{x})$ $\forall i, j$. This model thus learns context-specific interactions among $\mathbf{y}$:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{i,j} f_{pc}(\mathbf{x}, y_i, y_j, \Lambda)$$

Because each f_{pc} is now a function of $\mathbf{x}$, the number of parameters will be much greater. If there are n parameters in the logistic regression model, then there are $40n$ parameters in this model.

In the context of our experiment, the CRF-P model would learn the correlations between all pairwise combinations of personality traits to boost classification, while CRF-PF would learn many correlations of those combinations depending on the words observed.

Because the tree-width of the corresponding graphical models is sufficiently small for personality classification, exact CRF learning and inference can be performed using the junction tree algorithm (Koller and Friedman 2009). We use the GRMM (Sutton 2006) CRF toolkit in our experiments with default parameter settings.

In addition to the two CRF models (CRF-P and CRF-PF) we classified the data using Logistic Regression (Log-Reg), Naive Bayes (NB) and SMO, the sequential minimal optimization algorithm for a support vector classifier that is part of the Weka (Witten and Frank 2005; Platt 1999) toolkit. Finally, to provide a baseline we used ZeroR, a majority classification approach.[1]

Results

Table 1 shows the results obtained with these classifiers on the essays data. Table 2 shows the results obtained on the myPersonality data. All results are computed using 10-fold cross-validation. We compared the best result for each trait with all others using paired t-tests. Some of the best results were better than others at the $p < 0.05$ level. In those cases, the best result has a superscript indicating the algorithm(s) they outperformed significantly: s = SMO; n = NB; p = CRF-P; f = CRF-PF; l = LogR and z = ZeroR. However, no best result outperformed all of the other algorithms for a given trait at the $p < 0.05$ level.

Discussion

Our results show that SMO did not result in the best accuracies for the Essays data set. This is surprising considering that previous research has found it to be the best performing algorithm for this task. We believe that the choice of features as well as tokenization can have an important influence on classification given that so far most of the research has yielded poor accuracies for this data set. It may be that our feature set boosts the performance of a Naïve Bayes (and probabilistic classifiers). Adjustments in feature sets can improve Naïve Bayes algorithms yielding accuracies comparable to Support Vector Machines(Rennie et al. 2003).

[1]CRF-P, CRF-PF and Log-Reg were trained using GRMM (`http://mallet.cs.umass.edu/grmm/`). NB,SMO and ZeroR were trained using Weka (`http://www.cs.waikato.ac.nz/ml/weka/`).

	SMO	NB	CRF-P	CRF-PF	LogR	ZeroR
EXT	57.60	**61.20**	58.00	57.60	60.00	61.60
NEU	44.80	48.40	51.20	50.00	**52.40**	60.40
AGR	**52.80**	48.40	48.40	50.40	51.20	53.60
CON	52.40	**56.00**	51.60	49.60	52.80	52.00
OPN	54.80	54.40	64.00	61.20	**66.00**^{s}	70.40

Table 1: Accuracies by personality and classifier on the myPersonality data. Superscript denotes significant difference with corresponding algorithms ($p < 0.05$)

	SMO	NB	CRF-P	CRF-PF	LogR	ZeroR
EXT	51.74	56.24	56.08	55.95	**56.61**^{sz}	51.70
NEU	53.93	55.45	55.48	54.67	**57.17**^{sfpz}	48.50
AGR	51.62	53.28	53.05	**54.83**^{sp}	53.54	53.08
CON	53.24	**55.11**^{z}	54.42	54.78	55.06	50.81
OPN	54.32	**61.81**^{sz}	61.14	59.94	61.07	51.52

Table 2: Accuracies by personality and classifier on the essay data. Superscript denotes significant difference with corresponding algorithms ($p < 0.05$)

It is worth noting that our majority classifier resulted in a different baseline than that of Mairesse et. al. (Mairesse et al. 2007) even when using stratified folds –that is, folds that aim at balancing the values for a given class. This discrepancies may stem from minor differences in the prepapration of this version of the corpus[2]. When comparing our approach to their baselines[3], all of our best accuracies perform significantly better than them at the $p < 0.05$ level.

Finally, the majority classification scores for the myPersonality data outperformed the other classifiers in almost all traits. This can be attributed to the small sample size and skewedness of the data after merging the posts that belonged to the same individuals.

Comparing CRF Models From Table 1 and 2 we can compare the two CRF models with the other two probabilistic approaches (Naive Bayes and Logistic Regression). We observe that the CRF improves performance in only one case on both data sets: Agreeableness (AGR). Otherwise, the CRF performs consistently worse than Logistic Regression. We speculate two possible reasons for this: First, overall classification accuracy is quite poor for logistic regression (51-66%). The motivation for structured classification is that the predicted distribution for one class may inform the label for another. However, given the rate of misclassification, these dependencies may be as likely to confuse as to inform classification. Second, the dependencies between labels may not be strong enough to overcome the first problem. To investigate this, we computed the correlation between all pairs of labels in the Essays and myPersonality data. The average absolute value of all pairwise correlations is only 0.12.

[2]Personal communication with the preparer of the corpus

[3]EXT=50.04; NEU=50.08; AGR=50.36; CON=50.57; OPN=50.32

However, we find the strongest correlation is between NEU and AGR (Pearson's $r = -0.20$; i.e., neurotic people are less likely to be friendly). It is natural, then, that the one improvement with CRFs involved the AGR class. This correlation is in line with previous research on personality by John et. al. (1991). In addition, research in clinical psychology (Ode and Robinson 2009) has found that agreeable people have better control over negative emotions –a characteristic of high scorers of NEU (Costa and McCrae 1992)

Conclusions and Future Work

We compared a structured classification approach to personality to explore whether dependencies between the output labels for each trait helped improve classification. We found that of the four probabilistic classification methods used, CRF with pairwise label-features factors performed the best on agreeableness –the hardest trait to classify for the other classifiers. In addition, the strongest correlation among pairs of traits involved agreeableness and neuroticism. However, for other traits, the CRF does not appear to improve accuracy. Although the data has a high rate of misclassification, our research suggest that a correlations between agreeableness and neuroticism exists and that taking this into account in a classification model may help boost accuracy for agreeableness. Future work will include semi-supervised learning of the dependency between output labels to improve classification.

References

Argamon, S.; Dhawle, S.; Koppel, M.; and Pennebaker, J. W. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*.

Costa, P. T., and McCrae, R. R. 1992. Neo PI-R Professional Manual. *Odessa, FL: Psychological Assessment Resources*.

Estival, D.; Gaustad, T.; Pham, S. B.; Radford, W.; and Hutchinson, B. 2007. Author profiling for english emails. In *10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, 262–272.

Ghamrawi, N., and McCallum, A. 2005. Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, 195–200. New York, NY, USA: ACM.

Gill, A. J.; Nowson, S.; and Oberlander, J. 2009. What are they blogging about? personality, topic and motivation in blogs. In *ICWSM 2009*.

Iacobelli, F.; Gill, A.; Nowson, S.; and Oberlander, J. 2011. Large scale personality classification of bloggers. In D'Mello, S.; Graesser, A.; Schuller, B.; and Martin, J.-C., eds., *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 568–577.

John, O. P.; Donahue, E. M.; and Kentle, R. L. 1991. The "big five" inventory: Versions 4a and 5b. tech. rep., berkeley: University of california, institute of personality andSocial research. Technical report, Tech. rep., Berkeley: University of California, Institute of Personality and Social Research.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, 282–289.

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.

Mehl, M. R.; Gosling, S. D.; and Pennebaker, J. W. 2006. 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.

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

Nowson, S. 2006. *The Language of Weblogs: A study of genre and individual differences.* Ph.D. Dissertation, University of Edinburgh.

Oberlander, J., and Gill, A. J. 2006. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes* 42(3):239–270.

Oberlander, J., and Nowson, S. 2006. 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*.

Ode, S., and Robinson, M. D. 2009. Can agreeableness turn gray skies blue? a role for agreeableness in moderating Neuroticism-Linked dysphoria. *Journal of Social and Clinical Psychology* 28(4):436–462.

Pennebaker, J. W., and Francis, M. E. 1999. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum, 1 edition.

Pennebaker, J. W., and King, L. A. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77(6):1296–1312.

Platt, J. C. 1999. Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B.; Burges, C. J. C.; and Smola, A. J., eds., *Advances in kernel methods: support vector learning*. Cambridge, MA, USA: MIT Press. 185–208.

Rennie, J.; Shih, L.; Teevan, J.; and Karger, D. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *ICML*.

Richardson, J. F. 1968. Correlations between the extraversion and neuroticism scales of the e.p.i. *Australian Journal of Psychology* 20(1):15–18.

Sutton, C. 2006. GRMM: GRaphical Models in Mallet. http://mallet.cs.umass.edu/grmm.

Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.