Workshop on Computational Personality Recognition: Shared Task

Fabio Celli
CIMeC
corso Bettini 31
38068 Rovereto, Italy
fabio.celli@unitn.it

Fabio Pianesi
FBK
via Sommarive 18
38123 Trento, Italy
pianesi@fbk.eu

David Stillwell
Psychometrics Centre
Free School Lane
Cambridge CB2 3RQ, UK
ds617@cam.ac.uk

Michal Kosinski
Psychometrics Centre
Free School Lane
Cambridge CB2 3RQ, UK
mk583@cam.ac.uk

Abstract

In the *Workshop on Computational Personality Recognition (Shared Task)*, we released two datasets, varying in size and genre, annotated with gold standard personality labels. This allowed participants to evaluate features and learning techniques, and even to compare the performances of their systems for personality recognition on a common benchmark. We had 8 participants to the task. In this paper we discuss the results and compare them to previous literature.

Introduction and Background

Personality Recognition (see Mairesse et Al. 2007) consists of the automatic classification of authors' personality traits, that can be compared against gold standard labels, obtained by means of personality tests. The Big5 test (Costa & MacCrae 1985, Goldberg et al. 2006) is the most popular personality test, and has become a standard over the years. It describes personality along five traits formalized as bipolar scales, namely:
1) **Extraversion** (x) (sociable vs shy)
2) **Neuroticism** (n) (neurotic vs calm)
3) **Agreeableness** (a) (friendly vs uncooperative)
4) **Conscientiousness** (c) (organized vs careless)
5) **Openness** (o) (insightful vs unimaginative).

In recent years the interest of the scientific community in personality recognition has grown very fast. The first pioneering works by Argamon et al 2005, Oberlander & Nowson 2006 and the seminal paper by Mairesse et al. 2007, applied personality recognition to long texts, such as short essays or blog posts. The current challenges are instead related to the extraction of personality from mobile social networks (Staiano et al 2012), from social network sites (see Quercia et al. 2011, Golbeck et al 2011, Bachrach et al. 2012, Kosinski et al. 2013) and from languages different from English (Kermanidis 2012, Bai et al 2012). There are also many other applications that can take advantage of personality recognition, including social network analysis (Celli & Rossi 2012), recommendation systems (Roshchina et Al. 2011), deception detection (Enos et Al. 2006), authorship attribution (Luyckx & Daelemans 2008), sentiment analysis/opinion mining (Golbeck & Hansen 2011), and others.

In the *Workshop on Computational Personality Recognition (Shared Task)*, we invited contributions from researchers or teams working in these areas or other related fields. Despite a growing number of works in personality recognition, it is still difficult to gauge their performance and quality, due to the fact that almost all the scholars working in the field run their experiments on very different datasets, and use very different evaluation procedures (Celli 2013). These problems are exacerbated by the fact that producing gold standard data for personality recognition is difficult and costly.

In 2012 there has been a competition on personality prediction from Twitter streaming data[1] with about 90 teams participating, thus showing the great interest of the industry and the research community about this field. The *Workshop on Computational Personality Recognition (Shared Task)* is different from a simple competition, because we do not want to focus just on systems' performances, but rather we would like to provide a benchmark for discovering which feature sets, resources, and learning techniques are useful in the extraction of personality from text and from social network data. We released two datasets, different in size and domain, annotated with gold standard personality labels. This allowed participants to compare the performance of their personality recognition systems on a common benchmark, or to exploit personality labels for other related tasks, such as social network analysis.

In this paper we summarize the results of the *Workshop on Computational Personality Recognition (Shared Task)*, discussing challenges and possible future directions. The paper is structured as follows: in the next section we provide an overview of previous work on personality recognition. Then in the following sections we present the datasets and the shared task, we report and discuss the results, and finally we draw some conclusions.

Previous Work

In recent years there have been many different attempts to automatically classify personality traits from text or from other cues, like social network usage. Oberlander & Nowson 2006 (Ob06) classified extraversion, stability, agreeableness and conscientiousness of blog authors using n-grams as features and Naïve Bayes (NB) as learning algorithm. They

[1] http://www.kaggle.com/c/twitter-personality-prediction

experimented with different percentiles (using only the authors with the highest and lowest scores, in table 1 we report the results of 50% splitting) and reported that binary classes and automatic feature selection yield the best improvement over the baseline. Mairesse et al. 2007 (Ma07) ran personality recognition in both conversation (using observer judjements) and text (using self assessments via Big5). They exploited two lexical resources as features, LIWC (Pennebaker et al. 2001) and MRC (Coltheart 1981), and predicted both personality scores and classes (we report results over classes in table 1) using Support Vector Machines (SVMs) and M5 trees respectively. They also reported a long list of correlations between Big5 personality traits and two lexical resources they used. Iacobelli et Al. 2011 (Ia11) used as features word n-grams extracted from a large corpus of blogs, testing different extraction settings, such as the presence/absence of stop words or inverse document frequency. They found that bigrams, treated as boolean features and keeping stop words, yield very good results using SVMs as learning algorithm, although the features extracted are few in a very large corpus.

As for the extraction of personality recognition from social network sites, Golbeck et al. 2011a (G11f) predicted personality scores of 279 Facebook users, exploiting both linguistic features (from LIWC) and social features (i.e. friend count, relationship status). Golbeck et al. 2011b (G11t) also predicted the personality of 279 Twitter users, exploiting LIWC, structural features (i.e. hastags, links) and sentiment features, and using a Gaussian Process (GP) as learning algorithm. Quercia et Al. 2011 (Qu11) used network features (followers, following, klout[2] score) to predict the personality scores of 335 Twitter users. They used M5 rules as learning algorithm. Bai et Al. 2012 (Bi12) predicted personality classes of 335 users of RenRen, a popular Chinese social network. They exploited network features such as friend count, self comments and recent statuses counts and experimented either with a median split and 3 percentile grips. They obtained good results using decision trees (C4.5), the best performance was achieved using a median split (results reported in table 1). Bachrach et al. 2012 (Bc12) made an extensive analysis of the network traits (i.e. such as size of friendship network, uploaded photos, events attended, times user has been tagged in photos) that correlate with personality of 180000 Facebook users. They predicted personality scores using multivariate linear regression (mLR), and reported good results on extraversion.

A comparison of the results described here is reported in table 1. Basically there are two different approaches to personality recognition: bottom-up and top-down. The Bottom-up approach (Oberlander & Nowson 2006, iacobelli et al. 2011, Bachrach et al. 2012) starts from the data and seeks for linguistic cues associated to personality traits, while the top-down approach (Argamon et al. 2005, Mairesse et Al. 2007, Golbeck et al. 2011a) makes heavy use of external resources, such as LIWC and MRC, and tests the correlations between those resources and personality traits. The former approach seems to achieve the best improvement from the baselines,

Author	Alg.	Eval.	Traits	Users	Result	B.line
Ob06	NB	acc	xnac	71	.866	.549
Ma07	SVM	acc	xnaco	2.4m	.57	.5
Ia11	SVM	acc	xnaco	3m	.767	.5
G11f	M5	mae	xnaco	279	.115*	.118*
G11t	GP	mae	xnaco	279	.146*	.147*
Qu11	M5	rmse	xnaco	335	.794*	-
Bi12	C4.5	f	xnaco	335	.783	-
Bc12	mLR	rmse	xnaco	180m	.282*	-
Ce13	-	f	xnaco	2.4m	.686	.6

Table 1: Overview of Personality Recognition from Text and Personality Recognition for Social Networks. *=lower scores are best. Results are averaged over the five traits.

but it is more prone to overfitting and should be done on very large corpora, while the latter is more robust but yields smaller improvements. Celli 2013 proposed a combined approach, finding that it is useful for domain adaptation.

Datasets and Shared Task

We provided two gold standard labelled datasets: Essays and MyPersonality.

Essays (Pennebaker & King 1999) is a large dataset of stream-of-consciousness texts (about 2400, one for each author/user), collected between 1997 and 2004 and labelled with personality classes. Texts have been produced by students who took the Big5 test. The labels, that are self-assessments, are derived by z-scores computed by Mairesse et al. 2007 and converted from scores to nominal classes by us with a median split. Since this corpus has been used by different scholars (Mairesse et al. 2007, Argamon et al. 2005, Celli 2013) it has been included in the shared task as a reference to previous work.

myPersonality[3] is a sample of personality scores and Facebook profile data that has been used in recent years for several different researches (i.e. Bachrach et al. 2012, Kosinski et al. 2013). It has been collected by David Stillwell and Michal Kosinski by means of a Facebook application that implements the Big5 test (Costa & McCrae's NEO-PI-R domains and facets), among other psychological tests. The application obtained the consent from its users to record their data and use it for the research purposes. The dataset used for this workshop is a subset (250 users and about 9900 status updates) of the myPersonality sample. We selected only the users for which we had both information about personality and social network structure. The final dataset contains Facebook statuses in raw text, gold standard personality labels (self-assessments obtained using an 100-item long version of the IPIP personality questionnaire) and several social network measures, including: network size, betweenness centrality, density, brokerage and transitivity. We included personality labels both as scores and classes. Classes have been derived from scores with a median split, as for Essays. The status updates in myPersonality have been anonymized manually. For instance each proper name of person has been replaced with a fixed string (`*PROPNAME*`). Famous

[2]http://klout.com/kscore

[3]http://mypersonality.org

names, such as "Chopin" and "Mozart", and locations, such as "New York" and "Mexico", have not been replaced.

Participants of the shared task were required to use at least one of the datasets provided by the organizers for their experiments. They were allowed to add annotation levels using any type of external resource. Participants had one and a half months to develop their own tools and analyze data. Since the main focus of the shared task is not about competition, the datasets are not divided into development, training and test set. Participants were free to split the training and test sets as they wish, although we suggested to use Weka[4] (Witten & Frank 2005) with 66% training and 33% test splitting. As for the evaluation metrics, we suggested to use Precision, Recall and F1-measure to evaluate predictions over classes, and Mean Absolute Error to evaluate predictions over personality scores. As for the baselines, we suggested to refer to Mairesse et al. 2007 for the Essays corpus, while no baseline has been provided for myPersonality corpus, because it was the first time this corpus has been used in this format. We suggested to compute random or majority baselines if needed.

Results and Discussion

We had 8 teams participating to the shared task. Each team tested different solutions to the extraction of personality.

Verhoeven et al. (**Ve**) predicted Facebook personality classes with high F-measure. They used 2000 frequent trigrams as initial features taken from the Essays corpus, then exploited ensemble methods based on meta-learning to generalize features across generes and trained SVMs classifiers. They reported an improvement in the performance using ensemble methods with respect to a single SVM classifier. They also reported poor performances when predicting personality classes on Essays exploiting training data from myPersonality corpus. Table 2 shows the best results of the ensemble classifier on myPersonality (table 4 in Verhoeven et al. 2013). Farnadi et al. 2013 (**Fa**) proposed F-measure weighted by class-size average as evaluation measure. They experimented on myPersonality and Essays using four different sets of features: lexical (LIWC), network measures (social), status update timestamps (time) and other measures like posts per user, capital letters, repeated words (others). They tested three different learning algorithms: NB, SVMs and Nearest Neighbors (kNN) and reported that NB with time features performs poorly, while the other feature/algorithm combinations work well. They also tested cross-domain learning, reporting the best results in the generalization from Essays to myPersonality, using NB and kNN as learning algorithms (the best results, averaged over the five traits, are reported in table 2). They obtained poorer results when generalizing from myPersonality to Essays, although they outperformed the baselines, thus indicating that the size of the training set matters. Tomlinson et al. 2013 (**To**) exploited linguistic nuances to predict conscientiousness on myPersonality. In particular they used depth of the verbs in wordnet troponymy hierarchy, word objectivity taken from sentiwordnet and predicate occurrence with 1st and 3rd person agents and patients. They reported that conscientiousness is normally distributed across users, and an approach that predicts only the outliers (i.e. users in the 1st and 4th quartiles) might allow better results. Markovic et al. 2013 (**Ma**) predicted personality classes on myPersonality exploiting a very large feature space, including social and demographic info, lexical resources, Part Of Speech Tags, word emotional values (AFINN, see Nielsen 2011) and word intensity scale (H4Lvd). They achieved a very high performance using ranking algorithms for feature selection, SVMs and Boosting (B) as learning algorithms. Their results highlight the importance of feature selection in personality recognition. Alam et al. 2013 (**Al**) tested SVMs, Bayesian Logistic Regression (BLR) and Multinomial Naïve Bayes (mNB) to predict personality classes. They used unigrams as features and achieved the best performances with Multinomial Naïve Bayes. Mohammad et al. 2013 (**Mo**) predicted personality classes on Essays using unigrams, word specificity, and different lexical resources for emotion/sentiment analysis and psycholinguistics (including MRC and LIWC). They report that the Hashtag lexicon, a resource that links words to emotions expressed with hashtags extracted from Twitter, yields the most significant improvement over the baseline. Appling et al (**Ap**) manually annotated myPersonality corpus with speech acts (see Austin 1962) and reported their correlations to personality traits. Iacobelli et al 2013 (**Ia**) attempted to use structured classification, thus exploiting the correlations between personality labels to improve classifier's performance. They used unigrams, bigrams and trigrams as features, and Conditional Random Fields (CRF), SVMs, NB and Logistic Regression as learning algorithms. They found that neuroticism is negatively correlated to agreeableness, but for other traits they could not find any significant correlation, hence their results are poorer than expected.

Team	Alg.	Eval.	Traits	Data	Result
Ve	SVM	f	xeaco	es,mp	.72
Fa	SVM,kNN,NB	wf	xeaco	mp+es	.586
To	LR	rmse	c	mp	.63*
Ma	SVM,B	f	xeaco	mp	.904
Al	SVM,BLR,mNB	f	xeaco	mp	.586
Mo	SVM	f	xeaco	es	.57
Ia	NB	acc	xeaco	es	.563

Table 2: Overview of the results. *=lower score is best. We report only the best results, averaged over the five traits.

The Results are summarized in table 2. Participants exploited a wide range of resources, approaches and techniques. They bring out several interesting points: (i) feature selection with ranking algorithms over a large initial feature space is very effective; (ii) bottom-up approaches based solely on words (unigrams, bigram, trigrams) are not very effective, and seem to work best with probabilistic algorithms, like NB; (iii) top-down approaches, based on lexical resources (including the ones for sentiment analysis) and social info, in general seem to help personality recogntion more than bottom-up approaches, based on words or n-grams; (iv) ensemble methods are effective and (v) cross-

[4]http://www.cs.waikato.ac.nz/ml/weka/

domain learning is possible. These results are very important, because it is the first time we can compare bottom-up and top-down approaches on a common benchmark, and -at the same time- many lexical resources that were never exploited before in this kind of task.

Conclusions and Future Work

In this paper we have compared different systems for personality recognition from text on a common benchmark. The results show that personality recognition is a challenging task, due to the fact that there are no strong predictive features, or rather they are very sparse. The system of Markovic et al shows how feature selection, over a very large feature space, can boost the performance of a classifier, outperforming the state-of-the-art.

We are going to make the datasets used for this shared task available as a benchmark for future works[5]. We encourage to test new lexical resources and new approaches that can shed more light on the theoretical aspects of personality and human interactions in general.

References

Argamon, S., Dhawle, S., Koppel, M., and Pennebaker, J. W. 2005. Lexical Predictors of Personality Type. In *Proc. of Joint Annual Meeting of the Interface and the Classification Society of North America*. 1–16.

Austin, J.L. 1962. *How to Do Things With Words*. Harvard University Press, Cambridge MA.

Bai S., Zhu T. and Cheng L. 2012. Big-Five Personality Prediction Based on User Behaviors at Social Network Sites. In *eprint arXiv:1204.4809*.

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D.J., 2012 Personality and Patterns of Facebook Usage. In *Proc. of Web Science 2012*. 36–45.

Celli, F. 2013. *Adaptive Personality recognition from Text*. Saarbrücken, DE: Lambert Academic Publishing.

Celli, F., and Rossi, L. 2012. The role of Emotional Stability in Twitter Conversations. In *Proc. of Workshop on Semantic Analysis in Social Media - EACL*. 10–17.

Coltheart, M. 1981. MRC psycholinguistic database. In *Quarterly Journal of Experimental Psychology*. 33:A. 497–505.

Costa P.T., Jr. and MacCrae, R.R. 1985. In *The NEO Personality Inventory manual*. Psychological Assessment Resources.

Enos, F. Benus, S. Cautin, R.L., Graciarena, M., Hirschberg, J., and Shriberg, E. 2006. Personality Factors in Human Deception Detection: Comparing Human to Machine Performance. In *Proc. of INTERSPEECH - ICSLP*. 813–816.

Golbeck, J., Robles, C., and Turner, K. 2011a. Predicting Personality with Social Media. In *Proc. of the 2011 annual conference extended abstracts on Human factors in computing systems*. 253–262.

[5]data is available for download from the workshop website http://mypersonality.org/wiki/doku.php?id=wcpr13

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. 2011b. Predicting Personality from Twitter. In *Proc. of International Conference on Social Computing*. 149–156.

Golbeck, J. and Hansen, D.L. 2011. Computing political preference among twitter followers. In *Proc. of CHI 2011*, 1105–1108.

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogand R., Ashtone M.C., Cloningerf C.R., and Gough H.G. 2006. The international personality item pool and the future of public-domain personality measures. *Journal Res Pers.* 40(1): 8496.

Kermanidis, K.L. 2012. Mining Authors' Personality Traits from Modern Greek Spontaneous Text. In *Proc. of Workshop on Corpora for Research on Emotion Sentiment & Social Signals, in conjunction with LREC*. 90–93.

Kosinski, M., Stillwell, D.J. and Graepel, T. 2013. Private traits and Attributes are predictable from Digital records of human Behavior. *Proc. of the National Academy of Sciences*. 1–4.

Luyckx, K. and Daelemans, W. 2008. Personae: a corpus for author and personality prediction from text. In *Proc. of LREC-2008, the Sixth International Language Resources and Evaluation Conference*, 2981–2987.

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. In *Journal of Artificial intelligence Research*, 30: 457–500.

Nielsen, F.A. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, In *Proc. of the ESWC2011 Workshop on Making Sense of Microposts*. 93–98.

Oberlander, J., and Nowson, S. 2006. Whose thumb is it anyway? classifying author personality from weblog text. In *Proc. of the 44th Annual Meeting of the Association for Computational Linguistics ACL*. 627–634.

Pennebaker, J.W., Francis, M.E., and Booth, R.J. 2001 *Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum, Mahwah, NJ.

Pennebaker, J.W., and King, L.A. 1999. Linguistic styles: Language use as an individual difference. In *Journal of Personality and Social Psychology*, 77: 1296-1312.

Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. 2011. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *Proc. of SocialCom2011*. 180–185.

Roshchina, A., Cardiff, J., and Rosso P. 2011. A comparative evaluation of personality estimation algorithms for the TWIN recommender system. In *Proc. of the 3rd international Workshop on Search and mining user-generated contents*. 11–18.

Staiano, J., Lepri, B., Aharony, N., Pianesi, F., Sebe, N., and Pentland, A.S. 2012. Friends don't Lie - Inferring Personality Traits from Social Network Structure. In *Proc. of International Conference on Ubiquitous Computing*. 321–330.

Witten, I.H., and Frank, E. 2005. *Data Mining. Practical Machine Learning Tools and Techniques with Java implementations*. Morgan and Kaufman.