Is Big Five better than MBTI?
A personality computing challenge using Twitter data

Fabio Celli
FBK - MobS and Profilio Company
Trento, Italy
celli@fbk.eu

Bruno Lepri
FBK - MobS
Trento, Italy
lepri@fbk.eu

Abstract

English. Personality Computing from text has become popular in Natural Language Processing (NLP). For assessing gold-standard personality types, Big5 and MBTI are two popular models but still there is no comparison of the two in personality computing. With this paper, we provide for the first time a comparison of the two models from a computational perspective. To do that we exploit two multilingual datasets collected from Twitter in English, Italian, Spanish and Dutch.

Italiano. *Il riconoscimento automatico di personalità è diventato popolare nelle comunità di linguistica computazionale. I test Big Five e MBTI sono due modelli differenti per valutare la personalità, ma ancora non c'è un vero confronto dei due in ambito di riconoscimento automatico di personalità. In questo articolo per la prima volta forniamo una comparazione dei due modelli dal punto di vista computazionale. Per fare questo abbiamo raccolto dati Twitter in Inglese, Italiano, Spagnolo e Olandese in due corpora paralleli annotati con i due test.*

1 Introduction

The last decade has been characterized by the rise of personality computing in Natural Language Processing (NLP) (Vinciarelli and Mohammadi, 2014): for example, several works have dealt with the automatic prediction of personality traits of authors from different pieces of text they wrote in emails, blogs or social media (Mairesse et al., 2007; Iacobelli et al., 2011; Schwartz et al., 2013) (Rangel Pardo et al., 2015). Personality computing is also broadening its application to many fields in academia as well as in industry, including security (Golbeck et al., 2011), human resources (Turban et al., 2017), advertising (Celli et al., 2017) and deception detection (Fornaciari et al., 2013). Historically, there are two popular but very different psychological tests to asses personality: (i) the Big Five (Costa and McCrae, 1985; Costa and McCrae, 2008), which is widely accepted in academia, and (ii) the Myers Briggs Type Indicator (MBTI) (Myers and Myers, 2010), which is very popular and widely used in industry. The Big Five model defines personality along 5 bipolar scales: Extraversion (sociable vs. shy); Emotional Stability (secure vs. neurotic); Agreeableness (friendly vs. ugly); Conscientiousness (organized vs. careless); Openness to Experience (insightful vs. unimaginative). In contrast, the MBTI defines 4 binary classes that combines into 16 personality types: Extraversion/Introversion, Sensing/Intuition, Perception/Judging, Feeling/Thinking. Correlation analyses of the personality measures showed that Big Five Extraversion was correlated with MBTI Extraversion-Introversion, Openness to Experience was correlated with Sensing-Intuition, Agreeableness with Thinking-Feeling and Conscientiousness with Judging-Perceiving (Furnham et al., 2003). A reason for the recently gained popularity of MBTI is the fact that it is easier to collect gold-standard labelled data about MBTI than about Big Five, as an MBTI type is a 4-letter coding (e.g., INTJ) that could be retrieved with simple queries. In a field like personality computing, where data is costly and difficult to collect, this is an enormous advantage.

In this paper we address the question whether it is easier to predict Big Five or MBTI classes with a machine learning approach. To do so, we collect two Twitter datasets in English, Italian, Dutch and Spanish, one annotated with the Big Five personality types and one with MBTI. We believe that this work will be useful for the scientific community of personality computing to better understand the heuristic power of the two models when applied to machine learning tasks.

The paper is structured as follows: in the next section we provide an overview of related works in the field of personality computing in NLP, in Section 3 we describe the datasets we used, in Section 4 we report the results of our experiments and in Section 5 we draw some conclusions.

2 Related Work

Brief overview of personality computing The research in personality computing from text begun more than a decade ago with few pioneering works recognizing personality traits (Big Five traits) from blogs (Oberlander and Nowson, 2006) and self presentations (Mairesse et al., 2007). Other related fields have developed in the same years, like personality computing from multimodal and social signals, such as recorded meetings (Pianesi et al., 2008). In that period the research on MBTI was limited to find correlates between personality types and behavioral expectations, such as job preference (Cohen et al., 2013). Thus, MBTI was marginally used for personality computing until 2015 (Luyckx and Daelemans, 2008); while many works demonstrated the validity of Big Five for the automatic prediction of personality from different sources, including Twitter (Quercia et al., 2011) (Pratama and Sarno, 2015) (Qiu et al., 2012). The most common features used by researchers to perform such tasks were extracted from text, such as sentiment (Basile and Nissim, 2013), Part of Speech (PoS) tags, psycholinguistic tags (LIWC) (Tausczik and Pennebaker, 2010) and from metadata, such as number of followers, density of subject's network, hashtags, Likes and profile pictures. The rise of personality computing by means of the Big Five model brought fruitful collaborations between the communities of computer science and personality psychology (Back et al., 2010), and very interesting findings came out: for example that several personal characteristics extracted from social media profiles such as education, religion, marital status and the number of political preferences have really high correlations with personality types (Kosinski et al., 2013), or that popular users in social media are both extroverts and emotionally stable as well as high in Openness, while influential ones tend to be high in Conscientiousness (Quercia et al., 2012).

Overview of datasets The scarcity of data annotated with gold standard personality labels, difficult and costly to collect, was a major problem and the few large datasets available (MyPersonality, about 75K users, and Essays, about 2K users) soon became standard benchmarks (Celli et al., 2013). These available datasets covered mainly English language, while all the other datasets were much smaller, around 200 or 300 instances. In this scenario a dataset of 1500 instances collected by means of a simple Twitter search came out, and it was in English and annotated with MBTI labels (Plank and Hovy, 2015). This demonstrated that MBTI labels are very common and easy to retrieve from Twitter, unlike Big Five labels. Soon thereafter, TwiSty came out (Verhoeven et al., 2016), a multilanguage dataset of 17K instances annotated with MBTI and including Italian, Dutch, Portuguese, French and Spanish.

State of the art The MBTI model formalizes personality types as classes, while Big Five as scores. Despite this, works in computer science and computational linguistics split between those who use scores (Golbeck et al., 2011) and those who turn Big Five scores into binary classes in order to have a better control on class distribution and easier-to-interpret prediction tasks (Mairesse et al., 2007) (Segalin et al., 2017). In particular, Mairesse et al. obtained an average of 57% accuracy in the prediction of Big Five classes using the LIWC psycholinguistic features, also reporting that Openness to Experience was the easiest trait to model. Verhoeven et al. (Verhoeven et al., 2013) obtained a 72% of F-measure in the prediction of Big Five using trigrams and ensemble methods in a small Facebook dataset trained on a larger essays dataset. In a following study, Verhoeven et al. (Verhoeven et al., 2016) obtained an average of 63.8% of F-measure in the prediction of MBTI on Twitter in multiple languages using word and characters n-grams. Again, Farnadi et al. (Farnadi et al., 2013) obtained an average accuracy of 58.6% to predict Big Five classes on the same dataset using mostly metadata. Finally, Plank and Hovy (Plank and Hovy, 2015) used words and Twitter metadata to predict Extraversion/Introversion and Feeling/Thinking with 72% and 61% of accuracy, respectively. They reported that the best performing features are the linguistic ones.

The different settings and datasets used by previous works in the field makes it impossible to compare the results. Here, we aim to fill this gap.

3 Datasets

We collected from Twitter two multilingual datasets, of 900 users each, one annotated with MBTI and one with Big Five. First we collected the Big Five set by means of queries with Twitter advanced search[1], retrieving the results of different Big Five tests, ranging from the short 10-items test to the 44-items test. The language of the tweets were English, Italian, Spanish and Dutch, so we replicated the language distribution in the MBTI set using a portion of TwiSty (Verhoeven et al., 2016) and Plank's corpus (Plank and Hovy, 2015). The details about language distributions are reported in Figure 1.

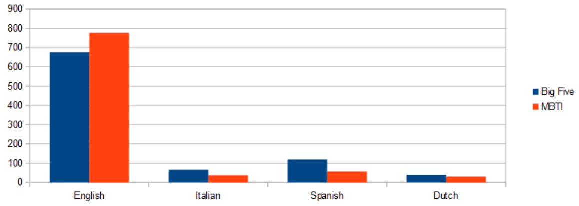

Figure 1: Distribution of the languages in the two datasets. The x-axis represents the number of users.

As expected there are many more tweets containing the results of the MBTI with respect to the Big Five. We use a concatenation of all tweets of a user, and a limit to 40 tweets per user in order to balance those who have too many tweets those that have few. In the end we used two comparable datasets with 900 users each, 265K words in the Big Five one and 290K words in the MBTI one. The classes are balanced in the Big Five set, as we obtained them with a median split from the original scores, on the contrary in the MBTI set there is a strong imbalance in the distribution of Sensing/Intuition and Feeling/Thinking, reported also in Plank's corpus. In the experiments, described in the next section, we balance the classes of both datasets and test different combinations of the features to evaluate the performance of machine leaning algorithms in the prediction of classes derived from the two different personality models.

4 Experiments, Results, Discussion and Limitations

Experimental settings We compared the performance of algorithms for the prediction of Big Five and MBTI classes in 9 binary classification tasks. To do so, we used the following features:

- Character n-grams (1000 features): we extracted from tweets 1000 characters bi-grams and tri-grams with a minimum frequency of 3. We did not remove stopwords and punctuation;

- LIWC match ratio (68 features): we computed the ratio of matches of the words in the LIWC dictionaries in all the four languages. LIWC provides mapping from words to 68 psycholinguistic categories, including words about others, self, space, time, society, family, friendship, sex, and functional words, among others;

- Metadata (10 features): this feature set includes the followers/following ratio, favorite/tweets ratio, listed/tweets ratio, link color, text color, border color, background color, hashtag/words ratio, retweet ratio, whether the profile picture is the default one or not. As feature selection procedure we used a subset selection algorithm (Hall and Smith, 1998) that reduces the degree of redundancy. We balanced the classes assigning weights to the instances in the data so that each class has the same total weight. For the classification we compared SVMs and a meta-classifier that automatically finds the best performing algorithm for the task (Thornton et al., 2013). As evaluation setting we used a 10-fold cross validation, as metric we reported accuracy and averages. For the maximum comparability we also reported the average on the Big Five four traits correlated with MBTI (avg4): extraversion, openness, agreableness and conscientiousness.

Results and discussion Results reported in Table 1 show that, on average, SVMs have higher performance in the prediction of MBTI classes with respect to Big Five, but there is much variability in the prediction of Big Five traits. In particular, we obtained very good performances for Emotional Stability and Agreeableness using a SVMs with polynomial kernel and Random Sub Spaces respectively, but poor with simple SVMs, indicating that the space is not linearly separable. On the contrary, the predictions of the MBTI seems to be more stable, in contrast to the results of Plank and Hovy. We suggest that this different

[1] https://twitter.com/search-advanced

trait	baseline	svm	auto	best feature
extr.	49.6	61.8	66.4 lr	others
stab.	49.8	59.6	74.8 svmk	I
agree.	49.6	61.1	73.3 rss	death
consc.	49.8	60.3	61.6 sdg	death
open.	49.6	53.1	59.4 nb	ngrams
avg4	49.7	**59.0**	**65.1**	-
avg	49.7	59.1	67.0	-
E-I	49.5	63.9	64.7 sdg	hashratio
S-N	49.2	66.3	68.6 bag	negate
F-T	49.8	63.0	63.0 svm	self
P-J	49.5	61.7	63.5 nb	self
avg	49.5	**63.7**	**64.9**	-

Table 1: Results of the experiments with all the languages and 900 instances per each set. Big Five is in the upper part of the Table and MBTI is below. We report accuracies for Support Vector Machines (svm) and AutoWeka (auto), a meta-classifier that automatically finds the best algorithm and settings for the task. The auto meta-classifier used Logistic Regression (lr), Support Vector Machines with polynomial kernel (svmk), Random Sub Spaces (rss), Stochastic Gradient Descent Regression (sdg), Naive Bayes (nb) and Bagging (bag). We also report average accuracy of Big Five traits correlated to MBTI (avg4): Extraversion, Openness to Experience, Agreeableness and Conscientiousness. The best features for the predictions are: words about others (others), first person singular pronoun (I), words about death (death), ngrams (ngrams), words about self (self), negation words (negate), hashtag ratio (hashratio).

result is due to three factors: class balancing, the use of LIWC and the subset feature selection. It is interesting to note that the *reference to others* is the best feature for the prediction of Big Five Extraversion and *first person pronouns* for the prediction of Emotional Stability/Neuroticism. We explain the predictive power of words about death for Agreeableness and Conscientiousness with the fact that this feature is correlated to the negative poles of these traits. The presence of different languages might affect negatively the performance so we ran an experiment using only English (650 users for each set).

Results, reported in Table 2, show that the effect of language variety is minimum, given that English is the most represented language in the datasets. It is interesting to note the changes in the best features: *hashtag ratio* is in English the best feature for Extraversion Big Five, while in the previous experiment it was the best feature for Extraversion MBTI. Here the best feature for Extraversion MBTI is *anger*, that is a clue for the negative class of this trait: Introversion. It is also interesting to note that words about feelings become in English the best feature for Agreeableness, although the performance decreases a little bit with respect to the experiment with all languages.

trait	baseline	svm	best feature
extr.	49.6	66.1	hashratio
stab.	49.6	62.9	I
agree.	49.6	59.7	feel
consc.	49.4	60.2	ngrams
open.	49.5	60.3	ngrams
avg4	49.6	**61.5**	-
avg	49.6	61.8	-
E-I	49.7	61.3	anger
S-N	48.4	68.5	we
F-T	49.3	68.6	self
P-J	49.6	60.2	I
avg	49.5	**64.6**	-

Table 2: Results of the experiments with English only and 650 instances per each set. Big Five is in the upper part of the Table and MBTI is below. We report accuracy for the majority baseline and Support Vector Machines (svm). The best features for the predictions are: hashtag ratio (hashratio), first person singular pronoun (I), words about feelings (feel), ngrams (ngrams), words about self (self), negation words (negate), words about anger (anger), first person plural pronoun (we), words about self (self).

Limitations In order to compare the two personality models, we forced the Big Five outcome, originally scores, into classes. This is one of the reasons why it is more difficult to predict Big Five classes than MBTI, but it is interesting to note that the performance of some Big Five traits can be boosted using non-linear models. Another limitation is related to the fact that we collected different users in the two datasets, with the risk to have some individuals in one dataset or the other that are easier to classify. In any case, it is impossible to collect data of the same users annotated with both MBTI and Big Five with Twitter queries, this is something that could be done only with a costly data collection effort, that we hope future work will do.

5 Conclusion

In this paper we provide for the first time a comparison of Big Five and MBTI from a personality computing perspective. To do so we use two multilingual Twitter datasets, one annotated with Big Five classes and one with MBTI classes. For the first time, we provide an evidence that algorithms trained on MBTI could have better performances than trained on the Big Five, although the Big Five is much more informative and has great variability in performance depending also on the algorithm used for the prediction. We let available the files used for the experiments[2], in order to grant the replicability or improvement of the results.

[2] http://personality.altervista.org/fabio.htm

Acknowledgments

The work of Fabio Celli and Bruno Lepri was partly funded by EIT Digital by City Enabler for Digital Urban Services (CEDUS) and by EIT Distributed Ledger Invoice (DLI).

References

Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. *WASSA 2013*, page 100.

Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. 2013. Workshop on computational personality recognition: Shared task. In *WCPR in conjuction to ICWSM 2013*.

Fabio Celli, Pietro Zani Massani, and Bruno Lepri. 2017. Profilio: Psychometric profiling to boost social media advertising. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 546–550. ACM.

Yuval Cohen, Hana Ornoy, and Baruch Keren. 2013. Mbti personality types of project managers and their success: A field survey. *Project Management Journal*, 44(3):78–87.

Paul T Costa and Robert R McCrae. 1985. *The NEO personality inventory: Manual, form S and form R*. Psychological Assessment Resources.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *In G.J. Boyle, G Matthews and D. Saklofske (Eds.). The SAGE handbook of personality theory and assessment*, 2:179–198.

Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. 2013. Recognising personality traits using facebook status updates. In *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*. AAAI.

Tommaso Fornaciari, Fabio Celli, and Massimo Poesio. 2013. The effect of personality type on deceptive communication style. In *Intelligence and Security Informatics Conference (EISIC), 2013 European*, pages 1–6. IEEE.

Adrian Furnham, Joanna Moutafi, and John Crump. 2003. The relationship between the revised neo-personality inventory and the myers-briggs type indicator. *Social Behavior and Personality: an international journal*, 31(6):577–584.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 149–156. IEEE.

Mark A Hall and Lloyd A Smith. 1998. *Practical feature subset selection for machine learning*. Springer.

Francisco Iacobelli, Alastair J Gill, Scott Nowson, and Jon Oberlander. 2011. Large scale personality classification of bloggers. In *Affective Computing and Intelligent Interaction*, pages 568–577. Springer.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

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

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

Isabel Briggs Myers and Peter B Myers. 2010. *Gifts differing: Understanding personality type*. Davies-Black Publishing.

Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634. Association for Computational Linguistics.

Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. 2008. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60. ACM.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter - or - how to get 1,500 personality tests in a week. *6TH Workshop on computational approaches to subjectivity, sentiment and social media analysis WASSA 2015*, page 92.

Bayu Yudha Pratama and Riyanarto Sarno. 2015. Personality classification based on twitter text using naive bayes, knn and svm. In *Data and Software Engineering (ICoDSE), 2015 International Conference on*, pages 170–174. IEEE.

Lin Qiu, Han Lin, Jonathan Ramsay, and Fang Yang. 2012. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710–718.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 180–185. IEEE.

Daniele Quercia, Renaud Lambiotte, David Stillwell, Michal Kosinski, and Jon Crowcroft. 2012. The personality of popular facebook users. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 955–964. ACM.

Francisco Manuel Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at pan 2015. In *Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391*, pages 1–8.

Andrew H Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):773–791.

Cristina Segalin, Fabio Celli, Luca Polonio, Michal Kosinski, David Stillwell, Nicu Sebe, Marco Cristani, and Bruno Lepri. 2017. What your facebook profile picture reveals about your personality. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 460–468. ACM.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2013. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855. ACM.

Daniel B Turban, Timothy R Moake, Sharon Yu-Hsien Wu, and Yu Ha Cheung. 2017. Linking extroversion and proactive personality to career success: The role of mentoring received and knowledge. *Journal of Career Development*, 44(1):20–33.

Ben Verhoeven, Walter Daelemans, and Tom De Smedt. 2013. Ensemble methods for personality recognition. In *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, pages 35–38.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al.*, pages 1–6.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):1–1.