Personality Traits on Twitter
—or—
How to Get 1,500 Personality Tests in a Week

Barbara Plank
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140, DK-2300 Copenhagen S
bplank@cst.dk

Dirk Hovy
Center for Language Technology
University of Copenhagen, Denmark
Njalsgade 140, DK-2300 Copenhagen S
dirk.hovy@hum.ku.dk

Abstract

Psychology research suggests that certain personality traits correlate with linguistic behavior. This correlation can be effectively modeled with statistical natural language processing techniques. Prediction accuracy generally improves with larger data samples, which also allows for more lexical features. Most existing work on personality prediction, however, focuses on small samples and closed-vocabulary investigations. Both factors limit the generality and statistical power of the results. In this paper, we explore the use of social media as a resource for large-scale, open-vocabulary personality detection. We analyze which features are predictive of which personality traits, and present a novel corpus of 1.2M English tweets annotated with Myers-Briggs personality type and gender. Our experiments show that social media data can provide sufficient linguistic evidence to reliably predict two of four personality dimensions.

1 Introduction

Individual author attributes play an important role in customer modeling, as well as in business intelligence. In either task, Natural Language Processing (NLP) is increasingly used to analyze and classify extra-linguistic features based on textual input. Extra-linguistic and linguistic features are assumed to be sufficiently correlated to be predictive of each other, which in practice allows for mutual inference (Pennebaker et al., 2003; Johannsen et al., 2015). A whole body of work in NLP is concerned with attribute prediction from linguistic features (Rosenthal and McKeown, 2011; Nguyen et al., 2011; Eisenstein et al., 2011; Volkova et al., 2013; Alowibdi et al., 2013; Ciot et al., 2013; Volkova et al., 2015). Apart from demographic features, such as age or gender, there is also a growing interest in personality types.

Predicting personality is not only of interest for commercial applications and psychology, but also for health care. Recent work by Preoţiuc-Pietro et al. (2015) investigated the link between personality types, social media behavior, and psychological disorders, such as depression and post-traumatic stress disorder. They found that certain personality traits are predictive of mental illness. Similarly, Mitchell et al. (2015) show that linguistic traits are predictive of schizophrenia.

However, as pointed out by Nowson and Gill (2014), computational personality recognition is limited by the availability of labeled data, which is expensive to annotate and often hard to obtain. Given the wide array of possible personality types, limited data size is a problem, since low-probability types and combinations will not occur in statistically significant numbers. In addition, many existing data sets are comprised of written essays, which usually contain highly canonical language, often of a specific topic. Such controlled settings inhibit the expression of individual traits much more than spontaneous language.

In this work, we take a data-driven approach to personality identification, to avoid both the limitation of small data samples and a limited vocabulary. We use the large amounts of personalized data voluntarily produced on social media (e.g., Twitter) to collect sufficient amounts of data. Twitter is highly non-canonical, and famous for an almost unlimited vocabulary size (Eisenstein, 2013; Fromreide et al., 2014). In order to enable data-driven personality research, we combine this data source with self-assessed Myers-Briggs Type Indicators (Briggs Myers and Myers, 2010), denoted MBTIs. Myers-Briggs uses four binary dimensions to classify users (INTROVERT–EXTROVERT,

Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pages 92–98,
Lisboa, Portugal, 17 September, 2015. ©2015 Association for Computational Linguistics.

Intuitive–Sensing, Thinking–Feeling, Judging–Perceiving), e.g., INTJ, ENTJ, etc., amounting to 16 different types. MBTIs have the distinct advantage of being readily available in large quantities on social media.

We are aware of the ongoing discussion in the psychological literature about the limited expressiveness of MBTI, and a preference for Big Five (Goldberg, 1990; Bayne, 1994; Furnham, 1996; Barbuto Jr, 1997). We are, however, to some extent agnostic to the theoretical differences. MBTI does presumably still capture aspects of the users' personality. In fact, several dimensions are correlated to the Big Five (Furnham, 1996).

Over a time frame of one week, we collect a corpus of 1.2M tweets from 1,500 users that self-identify with an MBTI. We provide an analysis of the type distribution and compare it to existing statistics for the general population. We train predictive models and report performance for the individual dimensions. In addition, we select the most relevant features via stability selection (Meinshausen and Bühlmann, 2010) and find that—apart from linguistic features—gender and count statistics of the user are some of the most predictive features for several dimensions, even when controlling for gender.

Our results indicate that certain personality distinctions, namely Introvert–Extrovert (I–E) and Thinking–Feeling (T–F), can be predicted from social media data with high reliability, while others are very hard to model with our features. Our open-vocabulary approach improves considerably as the amount of available data increases.

Contributions In this paper we i) demonstrate how large amounts of social media data can be used for large-scale open-vocabulary personality detection; ii) analyze which features are predictive of which personality dimension; and iii) present a novel corpus of 1.2M English tweets (1,500 authors) annotated for gender and MBTI. The code is available at: `https://bitbucket.org/bplank/wassa2015`

2 Data

Our question is simple: given limited amounts of time (one week, including corpus creation and statistical analysis), how much personality type information can we gather from social media—and is it informative? Using MBTI types and the sheer amounts of user-generated data, we show that social media can be a valuable resource.

I	ISTJ 75	ISFJ 77	**INFJ 257**	INTJ 193
	ISTP 22	ISFP 51	INFP 175	INTP 111
E	**ESTP 15**	ESFP 26	ENFP 148	ENTP 70
	ESTJ 36	ESFJ 36	ENFJ 106	ENTJ 102

Table 1: The 16 MBTI (total users: 1,500) and their raw count. Most frequent/rarest type in bold.

E–I	539 (36%)	961 (64%)
N–S	1162 (77%)	338 (23%)
T–F	624 (42%)	876 (58%)
J–P	882 (59%)	618 (41%)
female–male	939 (63%)	561 (37%)

Table 2: Distribution over dimensions and gender.

Identifying users In order to collect our data, we first search for users that self-identify with one of the 16 MBTIs. We search for mentions of any of the 16 types, plus "Briggs", which we found to be less often misspelled than "Myers". We then manually check all files and remove all tweets that contain more than one type. This typically relates to people describing a switch, referring to another person, or bot posts; this step removes around 30% of the tweets. We additionally label each user as male or female, if discernible. We remove all users whose gender could not be discerned.

In the end, our collection contains 1,500 distinct users with type and gender information. Table 1 shows the distribution over types, Table 2 shows the distribution over each dimension and gender. Figure 1 compares the MBTI type distribution of our Twitter corpus to general population estimates[1] (cf. §3).

We observe that the distribution in our corpus is shifted towards introverts and females (Figure 1 and Table 2). It has been observed before (Goby, 2006) that there is a significant correlation between online–offline choices and the MBTI dimension of Extravert–Introvert. Extroverts are more likely to opt for offline modes of communication, while online communication is presumably easier and more accessible for introverts. Our corpus reflects this observation.

[1] `http://www.capt.org/mbti-assessment/`

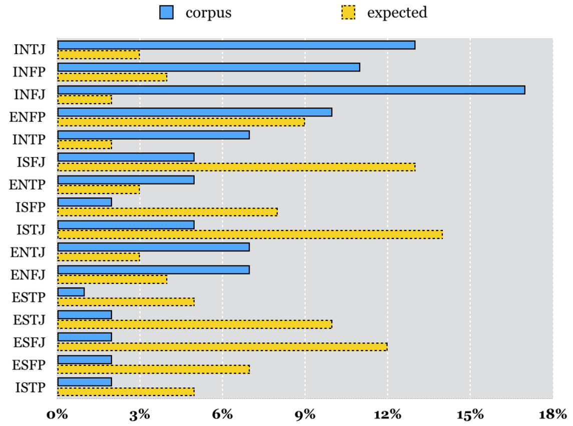

Figure 1: Comparison of MBTI distribution in Twitter corpus and general US population.

Corpus collection For each user, we download their most recent tweets. We require them to have at least 100 tweets, and collect a maximum of 2000 tweets. The final corpus contains 1.2 million tweets (19M tokens, average tweet: 16.1 tokens).

3 Statistical Analysis and Comparison

Using Twitter data naturally introduces a selection bias. We only have access to users who use Twitter and self-report their MBTI, while in previous studies participants were recruited to fill out a questionnaire and write texts specifically for the experiment.[2]

In order to quantify the differences to the general population, we compare the obtained MBTI distribution to general population estimates. Figure 1 shows that our Twitter distribution differs significantly from the general population (Spearman, $p < 0.05$) and exhibits different biases. There are many more introverts, and the data is shifted towards females (63%). While self-expression is easier for introverts online (Goby, 2006), our corpus also shows advertising/sensationalism bias. People like to tweet about rare events, e.g.,

> "Took a Myers-Briggs Personality Test. Received INFJ. Turns out only 1-2% of the population are that type #Interesting".

Interestingly, infrequent MBTIs in the general population (the first three bars in Figure 1, i.e., INFJ, INFP, INTJ) are amongst the most frequent types in our Twitter sample. Upon manual inspection of the data, we found that of the users reporting infrequent types, more than 60% belong to the three most frequent types in our corpus.

Despite the different biases, collecting linguistic data in this way has the advantage that it reflects actual language use, allows large-scale analysis and is less affected by interviewer biases.

4 Experiments

Model In order to predict each of the four dimensions from data, we train a logistic regression classifier.[3] As features, we use binary word n-grams ($n \in \{1, 2, 3\}$), gender, and several discretized count-based meta-features, i.e., counts of tweets, followers, statuses (total of tweets and retweets), favorites (number of favorited tweets) and listed counts (number of lists on which the Twitter user appears). Preliminary experiments showed that removing stopwords (and thus, removing personal pronouns) harms performance. The data is pre-processed, i.e., tokenized,[4] hashtags, URLs and usernames are replaced with unique tokens. We also remove any tweets containing a mention of one of the 16 MBTIs.

Feature selection In addition to type prediction, we perform feature selection to obtain insights into the classes. We use *stability selection* (Meinshausen and Bühlmann, 2010) to select the most discriminative features. We do *not* use the results of this selection in the predictive models.

We want to find the features that carry a high weight, irrespective of the conditions, in the entire data set. The conditions in this case are the *data composition* and *regularization*. In order to simulate different data compositions, we sample 100 times from the data. We use a sample size of 75% with replacement. For each sample, we fit a logistic regression model with a randomly set L_1 regularization constant, which encourages sparse feature weights. We average the weight vectors of all 100 induced models and select the features with the highest positive weight, representing the probability of being selected in each sample.

[2] Most of these questionnaires are administered in Psychology introduction classes, which introduces its own bias, though. See Henrich et al. (2010).

[3] Using the `sklearn` toolkit.

[4] Tokenizer from: `http://wwbp.org/`

5 Results

Table 3 shows the prediction accuracy for a majority-class baseline and our models on the full data set (10-fold cross-validation). While the model clearly improves on the I–E and F–T distinctions, we see no improvements over the baseline for S–N, and even a slight drop for P–J. This indicates that for the latter two dimensions, we either do not have the right features, or there is not linguistic evidence for them, given that they are more related to perception. The results from Luyckx and Daelemans (2008) on Dutch essays also suggest that P–J is difficult to learn.

Given the heavy gender-skew of our data, we run additional experiments in which we control for gender. The gender-controlled dataset contains 1070 authors. The results in Table 4 show the same tendency as in the previous setup.

	I–E	S–N	T–F	P–J
Majority	64.1	77.5	58.4	58.8
System	**72.5**	77.4	**61.2**	55.4

Table 3: Accuracy for four discrimination tasks with 2000 tweets/user.

	I–E	S–N	T–F	P–J
Majority	64.9	79.6	51.8	59.4
System	**72.1**	79.5	**54.0**	58.2

Table 4: Prediction performance for four discrimination tasks with 2000 tweets/user controlled for gender.

Figure 2 shows the effect of increased data size on prediction accuracy for the two best dimensions. Already from as little as 100 tweets, our model outperforms the baseline and is comparable to other studies. More data leads to better prediction accuracy. For I–E, there seems to be more headroom, while the accuracy of T–F plateaus after 500 tweets in the original dataset and slightly decreases in the gender-controlled setup. The trend on I–E also holds when controlling for gender as a confounding factor, while for T–F the highest performance is obtained with 500 tweets/user. In general, though, the results emphasize the benefits of large-scale analysis, especially for distinguishing the I–E dimension.

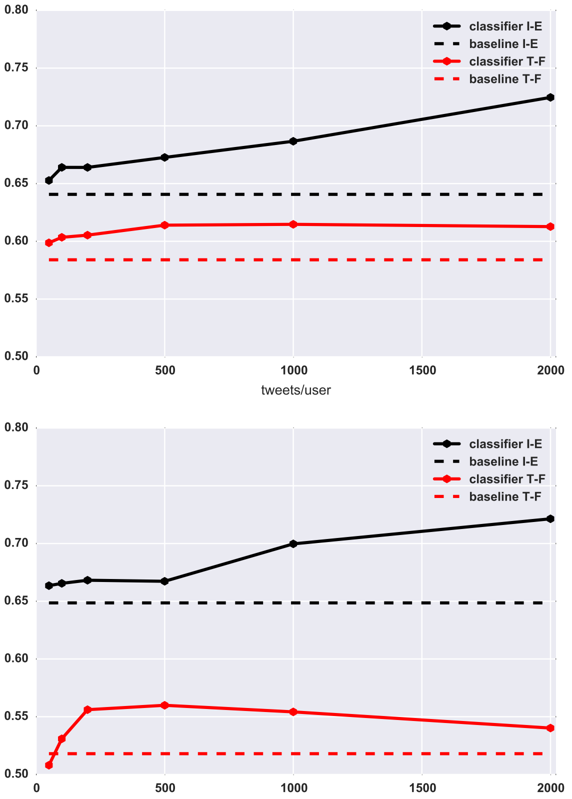

Figure 2: Learning curves and majority baselines for I–E and T–F on whole data set (top) and gender-balanced (bottom). x-axis = #tweets/user, y-axis = classification accuracy.

5.1 Predictive features

Table 5 shows the top 10 features for I–E and F–T found by stability selection. Our results show that linguistic features are by far the most predictive features for personality.

However, meta-features of the user account can also provide strong cues. More followers seem to indicate extroverts: a follower count of 100-500 users is a moderately strong indicator for extroverts (0.37). Interestingly, a status count of 1000–5000 tweets is a strong feature for introvert prediction (0.77), while less than 500 statuses correlate with extroverts (0.43). Similarly, if a user is member of 5-50 lists, it is indicative of introverts (0.64), while being in less than 5 lists is predictive of extroverts (0.55). These results support the finding that introverts prefer online media for communication (Goby, 2006).

Gender is another non-linguistic cue. In the gender-controlled experiment for the I–E dimension, gender is no longer a predictive feature, in contrast to the original dataset. For the F–T dis-

INTROVERT	EXTROVERT
someone 0.91	**pull** 0.96
probably 0.89	**mom** 0.81
favorite 0.83	**travel** 0.78
stars 0.81	**don't get** 0.78
b 0.81	**when you're** 0.77
writing 0.78	posted 0.77
, the 0.77	#HASHTAG is 0.76
status count< 5000 0.77	comes to 0.72
lol 0.74	**tonight !** 0.71
but i 0.74	join 0.69
THINKING	FEELING
must be 0.95	out to 0.88
drink 0.95	difficult 0.87
red 0.91	the most 0.85
from the 0.89	couldn't 0.85
all the 0.88	me and 0.8
business 0.85	in @USER 0.8
to get a 0.81	**wonderful** 0.79
hope 0.81	what it 0.79
june 0.78	trying to 0.79
their 0.77	! so 0.78

Table 5: Stability selection: most predictive features and their probabilities in the original dataset. Features in bold are predictive in both gender-balanced and original dataset (top 10 in both).

tinction, however, gender is actually fairly well-correlated with the respective classes for both types of experiments, albeit somewhat weaker for the gender-controlled setup (for T, GENDER=MEN is 0.57 in the original vs. 0.27 in the controlled experiment; for F, GENDER=FEMALE is 0.78 vs. 0.54). This indicates that gender is still an effective feature in predicting the F–T dimension when controlling for its distributional effect, while it is less important for distinguishing I–E.

6 Related work

Personality information can be valuable for a number of applications. Mitchell et al. (2015) studied self-identified schizophrenia patients on Twitter and found that linguistic signals may aid in identifying and getting help to people suffering from it.

Luyckx and Daelemans (2008) present a corpus for computational stylometry, including authorship attribution and MBTIs for Dutch. The corpus consists of 145 student (BA level) essays. They controlled for topic by asking participants to write about a documentary on artificial life. In a follow-up study, they extended the corpus to include reviews and both Big Five and MBTI information (Verhoeven and Daelemans, 2014). Instead, we focus on English and social media, a more spontaneous sample of language use.

Even when using social media, most prior work on personality detection can be considered small-scale. The 2014 Workshop on Computational Personality Recognition hosted a shared task of personality detection on 442 YouTube video logs (Celli et al., 2014). Celli et al. (2013) also examined Facebook messages of 250 users for personality. In contrast, our study uses 1.2M tweets from 1,500 different users.

The only prior large-scale open-vocabulary work on social media studies Facebook messages (Schwartz et al., 2013a; Schwartz et al., 2013b; Park et al., 2015). To date, their study represents the largest study of language and personality. Through a Facebook app, they collected personality types and messages from 75,000 Facebook users. They found striking variations in language use with personality, gender and age. Our approach is simpler, requires no tailored app, and can be used to collect large amounts of data quickly.

7 Conclusions

We use the self-reported Myers-Briggs type of Twitter users to collect a large corpus of tweets and train predictive models for each dimension.

Our results show that we can model the I–E (INTROVERT–EXTROVERT) and F–T (FEELING–THINKING) distinction fairly well. Learning the other two dimensions turns out to be hard. We find that linguistic features account for most of the predictive power of our models, but that meta-information, such as gender, number of followers, statuses, or list membership, add valuable information.

The distribution of Myers-Briggs personality types observed in our Twitter corpus differs from the general population, however, the data reflects real language use and sample sizes with sufficient statistical power. Our results suggest that while theoretically less well-founded than traditional approaches, large-scale, open-vocabulary analysis of user attributes can help improve classification accuracy and create insights into personality profiles.

Acknowledgements

We would like to thank the reviewers for their constructive feedback, as well as Lyle Ungar and Ben Verhoeven for suggestions.

References

Jalal S Alowibdi, Ugo A Buy, and Philip Yu. 2013. Empirical evaluation of profile characteristics for gender classification on twitter. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 365–369. IEEE.

John E Barbuto Jr. 1997. A critique of the myers-briggs type indicator and its operationalization of carl jung's psychological types. *Psychological Reports*, 80(2):611–625.

Rowan Bayne. 1994. The" big five" versus the myers-briggs. *PSYCHOLOGIST-LEICESTER-*, 7:14–14.

Isabel Briggs Myers and Peter Myers. 2010. *Gifts differing: Understanding personality type*. Nicholas Brealey Publishing.

Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. 2013. Workshop on computational personality recognition (shared task). In *Proceedings of the Workshop on Computational Personality Recognition*.

Fabio Celli, Bruno Lepri, Joan-Isaac Biel, Daniel Gatica-Perez, Giuseppe Riccardi, and Fabio Pianesi. 2014. The workshop on computational personality recognition 2014. In *Proceedings of the ACM International Conference on Multimedia*, pages 1245–1246. ACM.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Jacob Eisenstein, Noah Smith, and Eric Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of ACL*.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL*.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating NER for Twitter #drift. In *LREC*.

Adrian Furnham. 1996. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. *Personality and Individual Differences*, 21(2):303 – 307.

Valerie Priscilla Goby. 2006. Personality and online/offline choices: Mbti profiles and favored communication modes in a singapore study. *CyberPsychology & Behavior*, 9(1):5–13.

Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL*.

Kim Luyckx and Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. In *LREC*.

Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Dong Nguyen, Noah A Smith, and Carolyn P Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Scott Nowson and Alastair J Gill. 2014. Look! who's talking?: Projection of extraversion across different social contexts. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pages 23–26. ACM.

Greg Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, David J Stillwell, Michal Kosinski, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, Hansen Andrew Schwartz, and Lyle H Ungar. 2015. The role of personality, age and gender in tweeting about mental illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, NAACL.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics.

Hansen Andrew Schwartz, Johannes C Eichstaedt, Lukasz Dziurzynski, Margaret L Kern, Eduardo Blanco, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013a. Toward personality insights from language exploration in social media. In *AAAI Spring Symposium: Analyzing Microtext*.

Hansen Andrew 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. 2013b. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9).

Ben Verhoeven and Walter Daelemans. 2014. CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *LREC*.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media (demo). In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, Austin, TX, January.