Reddit: A Gold Mine for Personality Prediction

Matej Gjurković and **Jan Šnajder**
University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{matej.gjurkovic,jan.snajder}@fer.hr

Abstract

Automated personality prediction from social media is gaining increasing attention in natural language processing and social sciences communities. However, due to high labeling costs and privacy issues, the few publicly available datasets are of limited size and low topic diversity. We address this problem by introducing a large-scale dataset derived from Reddit, a source so far overlooked for personality prediction. The dataset is labeled with Myers-Briggs Type Indicators (MBTI) and comes with a rich set of features for more than 9k users. We carry out a preliminary feature analysis, revealing marked differences between the MBTI dimensions and poles. Furthermore, we use the dataset to train and evaluate benchmark personality prediction models, achieving macro F1-scores between 67% and 82% on the individual dimensions and 82% accuracy for exact or one-off accurate type prediction. These results are encouraging and comparable with the reliability of standardized tests.

1 Introduction

Personality refers to individual and stable differences in characteristic patterns of thinking, feeling, and behaving (Corr and Matthews, 2009). There has been an increasing interest in automated personality prediction from social media from both the natural language processing and social science communities (Nguyen et al., 2016). In contrast to traditional personality tests – whose use so far has mostly been limited to human resource management, counseling, and clinical psychology – automated personality prediction from social media has a far wider applicability, such as in social media marketing (Matz et al., 2017) and dating web-sites and applications (Finkel et al., 2012).

Most work on personality prediction rests on one of the two widely used personality models: Big Five and MBTI. The Big Five (Goldberg, 1990) is a well-established model which classifies personality traits along five dimensions: extraversion, agreeableness, conscientiousness, neuroticism, and openness. In contrast, the Myers-Briggs Type Indicator model (MBTI) (Myers et al., 1990) recognizes 16 personality types spanned by four dimensions: Introversion/Extraversion (how one gains energy), Sensing/iNtuition (how one processes information), Thinking/Feeling (how one makes decisions), and Judging/Perceiving (how one presents herself or himself to the outside world). Despite some controversy regarding test validity and reliability (Barbuto Jr, 1997), the MBTI model has found numerous applications, especially in the industry[1] and for self-discovery. Although the Big Five and MBTI models are built on different theoretical perspectives, studies have shown their dimensions to be correlated (McCrae and Costa, 1989; Furnham, 1996).

The perennial problem of personality prediction from social media is the lack of labeled datasets. This can be traced back to privacy issues (e.g., on Facebook) and prohibitively high labeling costs. The few existing datasets suffer from other shortcomings related to non-anonymity (which makes the users more reluctant to express their true personality), limited expressivity (e.g., on Twitter), low topic diversity, or a heavy bias toward personality-related topics (e.g., on personality forums). Specifically for MBTI, the only available datasets are the ones derived from Twitter (Verhoeven et al., 2016), essays (Luyckx and Daelemans, 2008), and personality forums.[2] Clearly, the lack of adequate benchmark datasets hinders the development of personality prediction models for social media.

In this paper we aim to address this problem by introducing MBTI9k, a new personality prediction dataset labeled with MBTI types. The dataset is

[1] http://www.cpp.com
[2] http://www.kaggle.com/datasnaek/mbti-type

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 87–97
New Orleans, Louisiana, June 6, 2018.

derived from the popular discussion website Reddit, the sixth largest website in the world and also one with the longest time-on-site.[3] What makes Reddit particularly suitable is that its content is publicly available and that many users provide self-reported MBTI personality types. Furthermore, the comments and posts are anonymous and cover a remarkably diverse range of topics, structured into more than a million discussion groups.[4] Altogether, the MBTI9k dataset derived from Reddit addresses all the abovementioned shortcomings of the existing personality prediction datasets.

We use the MBTI9k dataset to carry out two studies. In the first, we extract a number of linguistic and user activity features and perform a preliminary feature analysis across the MBTI dimensions. Our analysis reveals that there are marked differences in the values of these features for the different poles of each MBTI dimension. In the second study, we frame personality prediction as a supervised machine learning task and evaluate a number of benchmark models, obtaining promising results considerably above the baselines.

In sum, the contributions of our paper are threefold: (1) we introduce a new, large-scale dataset labeled with MBTI types, (2) we extract and analyze a rich set of features from this dataset, and (3) we train and evaluate benchmark models for personality prediction. We make the MBTI9k dataset and the extracted features publicly available in the hope that it will help stimulate further research in personality prediction.

The rest of the paper is structured as follows. The next section briefly reviews related work. Section 3 describes the acquisition of the MBTI9k dataset. In Section 4 we describe and analyze the features, while in Section 5 we evaluate the prediction models and discuss the results. Section 6 concludes the paper and outlines future work.

2 Background and Related Work

Personality and language are closely related – as a matter of fact, the Big Five model emerged from a statistical analysis of the English lexicon (Digman, 1990). Ensuing research in psychology attempted to establish links between personality and language use (Pennebaker and King, 1999), setting the ground for research on automated personality prediction. Most early studies in personality prediction relied on small datasets derived from essays (Argamon et al., 2005; Mairesse et al., 2007), e-mails (Oberlander and Gill, 2006), conversations extracted from electronically activated recorders (Mehl et al., 2001; Mairesse et al., 2007), blogs (Iacobelli et al., 2011), or Twitter (Quercia et al., 2011; Golbeck et al., 2011).

In contrast, MyPersonality (Kosinski et al., 2015) was the first project that made use of a large, user-generated content from social media, with over 7.5 million Facebook user profiles labeled with Big Five types. A subsequent study by Kosinski et al. (2013) on this dataset found the users' digital traces in the form of likes to be a very good predictor of personality. Schwartz et al. (2013) used the MyPersonality database in a first large-scale personality prediction study based on text messages. Over 15.4 million of Facebook statuses collected from 75 thousand volunteers were analyzed using both closed- and open-vocabulary approaches. The study found that the latter yields better results when more data is available, which was later also confirmed on other social media sites, such as Twitter (Arnoux et al., 2017).

The growing interest in personality prediction gave rise to two shared tasks (Celli et al., 2013; Rangel et al., 2015), which relied on benchmark datasets labeled with Big Five types. The overarching conclusion was that the personality prediction is a challenging task because there are no strongly predictive features. However, the results suggested that n-gram based models consistently yield good performance across the different languages.

Presumably due to its controversy, the MBTI model has thus far been less used for personality prediction. This has changed, however, with the work of Plank and Hovy (2015), who made use of the MBTI popularity among general public and collected a dataset of over 1.2 million status updates on Twitter and leveraged users' self-reported personality types (Plank and Hovy, 2015). Soon thereafter, Verhoeven et al. (2016) published a multilingual dataset TwiSty.

Our personality prediction dataset is derived from Reddit. Reddit has previously been used as a source of data for various studies. De Choudhury and De (2014) studied mental health discourse and concluded that Reddit users openly share their experiences and challenges with mental illnesses in their personal and professional lives. Schrading et al. (2015) studied domestic abuse and found that

[3] http://www.alexa.com/topsites
[4] http://redditmetrics.com

abuse-related discussion groups have more tight-knit communities, longer posts and comments, and less discourse than non-abusive groups. Wallace et al. (2014) tackled irony detection and concluded that Reddit provides a lot of context, which can help in dealing with the ambiguous cases. Shen and Rudzicz (2017) achieved good results in anxiety classification using the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2003), n-grams, and topic modeling. To the best of our knowledge, ours it the first work on using Reddit as a source of data for personality prediction.

3 Dataset

3.1 Description

Discussions on Reddit are structured into user-created discussion groups, the so-called *subreddits*, each focusing on one topic. Each subreddit consists of user posts, which may contain text, links, video, or image content. Users can comment on other users' posts, as well as upvote or downvote them. The posts in each subreddit are ranked by the number of comments, so that the most commented posts appear at the top. Apart from being moderated, many subreddits come with their own discussion ground rules, which generally improve the discussion quality. The database of Reddit posts and comments is available on Google Big Query and covers the period from 2005 till the end of 2017, currently totaling more than 3 billion comments and increasing at the rate of 85 million comments per month.

3.2 Flairs

One distinctive feature of Reddit are the special user descriptors called *flairs*. A flair is an icon or text that appears next to a username. It is specific to each subreddit, and in some subreddits users use flairs to introduce themselves. Specifically, in subreddits devoted to MBTI discussions, such as *reddit.com/r/MBTI* and *reddit.com/r/INTP*, users typically use flairs to report their MBTI types. In addition to the MBTI type, many users also provide information about their age, gender, personality types of their partners, marital status, medical diagnoses (e.g., "*Aspie*", to indicate a person with Asperger's syndrome), other personality theories' types (Enneagram, Socionics), and even stereotypes such as "*Dumb Emotional Sensor*" (meant to indicate the sensing-feeling MBTI types).

A problem with flairs is that they are worded in different, often ambiguous ways. In some cases it may be difficult to determine whether the flair refers to a personality type. For example, "*KentJude*" is not an MBTI type even though it contains the ENTJ acronym, a clue being that it is not written in all caps. In other cases, determining the type requires some inference. For instance, from "*INTP-T (MBTI) INTP (KTS) INTj (Socionics)*" one can infer that the user took the 16personalities test,[5] which maps Big Five's neuroticism to Assertive/Turbulent dimension, and that the user's MBTI type is INTP and not INTj, because INTj is a Socionics type, for which the last letter is written in lowercase. A more contrived example is the flair "*to Infjnity and Beyond...*", meant to indicate the INFJ MBTI type.

3.3 Acquisition

Our idea was to use the self-reported MBTI type from the user's flair as that user's personality type label. We make a sensible assumption that, if a user provides his or her MBTI type in the flair, in most cases this will be because she took at least one personality test. The assumption is born out by our analysis of users' comments, which revealed that most users with self-reported MBTI types report on taking multiple personality tests, and many of them even demonstrate a good knowledge of the MBTI theory.

The acquisition of the dataset aimed for high precision at the expense of recall, in the sense that we prefer to have fewer users with reliable MBTI labels rather than more users with uncertain MBTI labels. The acquisition proceeded in five steps:

1. First, we acquired a list of all users who have any mention of an MBTI type in their flair field, and compiled a list of flairs for all users. Many of the so-obtained flairs were false positives, for the reasons outlined above;
2. We next used regex-based pattern matching to (1) identify the flairs that refer to MBTI types, (2) tag ambiguous flairs, and (3) filter out the remaining flairs;
3. We examined the ambiguous flairs and discarded those we could not resolve (e.g., *XNFJ*, indicating extravert/introvert indefinity). We grouped the remaining flairs by users and checked for consistency of MBTI types (users may change their flairs and may have different

[5] http://www.kaggle.com/datasnaek/mbti-type

flairs for different subreddits), removing all users with a non-unique MBTI type;

4. At this point, some MBTI types turned out to be heavily underrepresented (e.g., merely 16 ESFJ and 23 ESTJ users), so we decided to compensate for this by complementing the dataset as follows. For each underrepresented type, we performed a full-text search over MBTI subreddit comments (not the flairs), searching for user's self-declaration of that specific type using a handful of simple but strict patterns ("*I am (an)* ⟨*type*⟩" and variants thereof). We then manually inspected the comments and filtered out the false positives, adding the remaining users to the dataset;

5. Lastly, we acquired all posts and comments of the users shortlisted in steps 3 and 4 above, dating from January 2015 to November 2017.

While the above procedure yields a high-precision labeled dataset, we acknowledge the presence of a selection bias in our dataset. More concretely, our dataset includes only the users who are acquainted with MBTI and who participated in MBTI-related subreddits, who know what a flair is and decided to use it to disclose their MBTI type, and who have written at least one comment. Moreover, additional bias is likely to be introduced by steps 2 and 4 above. The terms "Reddit user" and "Redditor" should be interpreted with these limitations in mind.

The resulting dataset consists of 22,934,193 comments (totaling 583,385,564 words) from 36,676 subreddits posted by 13,631 unique users and 354,996 posts (totaling 921,269 words) from 20,149 subreddits posted by 9,872 unique users. The dataset contains more than eight times more words than used in the aforementioned large-scale research by Schwartz et al. (2013), making it the largest available personality-labeled dataset in terms of the number of words.

3.4 Analysis

Our dataset offers many exciting possibilities for analysis, some of which we hope will be pursued in follow-up work. As a first step, we provide a basic descriptive analysis of the dataset, followed by some more interesting analyses in Section 4 meant to showcase the potential utility of the dataset.

Table 1 shows the distribution of Redditors across MBTI types and across the individual MBTI dimensions. For comparison, the first column shows the distribution estimated for the US population.[6] The data reveal that Redditors are predominantly of introverted, intuitive, thinking, and perceiving types. Incidentally, this distribution bears similarity to the distribution of gifted adolescents (Sak, 2004), and is also aligned with the data that shows that Reddit visitors are more educated than the average Internet user.[7]

Table 2 offers a different perspective on the data: the number of subreddits broken down by the number of distinct MBTI types of the users that participated in these subreddits. Interestingly, the majority (almost 47%) of subreddits attract users of the same type. Conversely, there are only 534 subreddits (1.45%) in which all 16 types participated; while this is a small fraction of the dataset, we believe it might still be sufficient for a comparative analysis between the types.

Another interesting and important aspect of the dataset is the language used for posts and comments. We ran the langid[8] language identification tool on all comments and posts of each of the user. The results suggest that the majority of users write more than 97% of their comments in English. This is in line with the web traffic data, according to which 76.4% of Reddit visitors come from native English-speaking countries.[7]

We make two versions of the dataset available: (1) a dataset of all comments and posts, each annotated with the MBTI type of the author, and (2) a subset of this dataset, referred to as MBTI9k dataset, which contains the comments of all users who contributed with more than 1000 words. Moreover, to remove the topic bias, we expunged from the MBTI9k dataset all comments from 122 subreddits that revolve around MBTI-related topics (making up 7.1% of all comments) and replaced all explicit mentions of MBTI types (and related terminology, such as cognitive functions (Mascarenas, 2016)) with placeholders. Besides comments, for each user we provide the MBTI type and a set of precomputed features (cf. Section 4). We make both datasets publicly available,[9] and use MBTI9k for the subsequent analyses.

[6] https://www.capt.org/products/examples/20025HO.pdf
[7] https://www.alexa.com/siteinfo/reddit.com
[8] https://github.com/saffsd/langid.py
[9] http://takelab.fer.hr/data/mbti

Type	% USA	% comm	% post	% MBTI9k
INTP	3.3	22.3	26.8	25.3
INTJ	2.1	17.2	20.6	20.0
INFJ	1.5	11.2	12.9	11.1
INFP	4.4	11.0	13.3	11.6
ENFP	8.1	6.1	7.4	6.6
ENTP	3.2	6.1	7.4	6.7
ENTJ	1.8	5.3	2.8	3.9
ISTP	5.4	5.2	3.7	4.8
ISTJ	11.6	3.4	1.3	2.4
ENFJ	2.5	3.3	1.1	2.3
ISFJ	13.8	2.4	0.7	1.3
ISFP	8.8	2.3	0.7	1.6
ESTP	4.3	1.2	0.5	0.9
ESFP	8.5	1.1	0.3	0.7
ESTJ	8.7	1.0	0.3	0.5
ESFJ	12.3	0.8	0.2	0.4
Dimension				
Introverted	50.7	75.1	80.0	78.1
Extroverted	49.3	24.9	20.0	21.9
Sensing	73.3	17.4	7.7	12.6
Intuitive	26.7	82.6	93.3	87.4
Thinking	40.2	61.7	63.4	64.4
Feeling	59.8	38.3	36.6	35.6
Judging	54.1	44.6	39.9	41.8
Perceiving	45.9	55.4	61.1	58.2

Table 1: Distributions of MBTI types and dimensions in US general public and on Reddit

# types	# subred.	%	# types	# subred.	%
1	17222	46.96	9	729	1.99
2	5632	15.36	10	640	1.75
3	3105	8.47	11	567	1.55
4	2034	5.55	12	512	1.4
5	1540	4.2	13	443	1.21
6	1217	3.32	14	377	1.03
7	964	2.63	15	362	0.99
8	798	2.18	16	534	1.46

Table 2: Distribution of subreddits by the number of distinct MBTI types of participating users

4 Feature Extraction and Analysis

4.1 Feature Extraction

For each of the 9,111 Reddit users from the MBTI9k dataset we extracted a set of features. These can be divided into two main groups: linguistic features (extracted from user's comments) and user activity features. Next we describe these features in more detail, followed by a preliminary feature analysis.

Linguistic features. The linguistic features include both content- and style-based features. The simplest of them are tf- and tf-idf-weighed character n-grams (lengths 2–3) and word n-grams (lengths 1–3), stemmed with Porter's stemmer. The total number of n-gram features is 11,140. For each user we also compute the type-token ratio, the ratio of comments in English, and the ratio of British English vs. American English words.

We used LIWC (Pennebaker et al., 2015), a widely used NLP tool in personality prediction, to extract 93 features. These range from part-of-speech (e.g., pronouns, articles) to topical preferences (e.g., bodily functions, family) and different psychological categories (e.g., emotions, cognitive processes). Complementary to LIWC, we used a number of psycholinguistic words lists, including perceived happiness, affective norms (e.g., valence, arousal, and dominance), imageability, and sensory experience, described in Preoţiuc-Pietro et al. (2017), as well as two lists of word meaningfulness ratings from the MRC Psycholinguistic Database (Coltheart, 1981). For each user, we calculated the average ratings for every word from these dictionaries, which gave us 26 features, denoted PSYCH.

User activity features. User activity features were extracted from comment and post metadata. The *global* features include the number of comments (all comments and comments on MBTI-related subreddits) and the number of subreddits commented in. The *posts* features include the overall post score (difference between the number of up and down votes), number of posts on "over 18" subreddits, the number of "self posts" (posts linking to other Reddit posts), and the number of gilded posts (posts awarded with money by other users).

Another group of features are topical affinity features. We computed comment counts for the user across subreddits and encoded these as a a single vector, together with the entropy of the corresponding distribution. In addition, we derive topic distributions from user's comments (1) using LDA models with 50 and 100 topics (2) by manually grouping top-200 subreddits into 35 semantic categories, and encode these as 50-, 100-, and 35-dimensional vectors, respectively.

We speculate that the temporal aspect of one's activities might be relevant for personality type prediction. We therefore include the time intervals between comment timestamps (the mean, median, and maximum delay), as well as daily, weekly, and monthly distributions of comments, encoded as vectors of corresponding lengths.

Feature group	E/I	S/N	T/F	J/P
char_tf	29.03	45.16	35.48	51.61
word_tf	35.48	25.81	12.9	32.26
liwc	19.35	0.0	25.81	9.68
lda100	6.45	0.0	9.68	3.23
psy	3.23	0.0	12.9	0.0
word	3.23	9.68	0.0	0.0
char	0.0	12.9	0.0	0.0
posts	0.0	6.45	0.0	3.23

Table 3: Percentage of each feature group in top-30 relevant features for each dimension

Feature group	E/I	S/N	T/F	J/P
global	33.33	33.33	100.0	66.67
psy	25.0	41.67	70.83	41.67
liwc	40.86	29.03	62.37	39.78
day_of_week	0.0	0.0	28.57	100.0
word_an_tf	28.22	32.07	38.17	27.3
char_an_tf	19.28	27.06	36.26	21.47
word_an	7.4	19.58	27.28	24.72
char_an	4.45	14.4	30.3	8.82
meaning	0.0	0.0	50.0	0.0
lda100	9.0	12.0	15.0	9.0
posts	5.0	20.0	5.0	10.0
char	0.12	0.88	28.99	0.24
month	0.0	25.0	0.0	0.0
word	0.16	1.23	21.67	1.12
time_diffs	0.0	16.67	0.0	0.0
subcat	0.0	2.86	8.57	0.0
lda50	0.0	6.0	4.0	0.0
hour	0.0	0.0	0.0	4.17
sub	0.04	0.48	0.14	0.0

Table 4: Percentage of highly relevant features ($p<0.001$) in total number of features per feature group and dimension

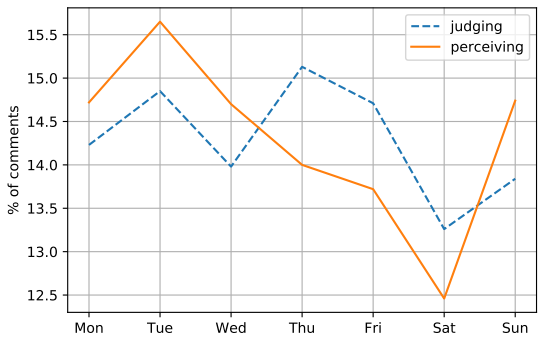
4.2 Feature Analysis

Feature relevance. We estimate the relevance of each feature for each MBTI dimension using a t-test: feature relevance is inversely proportional to the p-value under the null hypothesis of no difference in feature values for the two classes. Table 3 shows the proportion of features from each feature group in the set of top-30 most relevant features for each MBTI dimension. For instance, tf-weighted character n-grams (char_tf) account for about 29% of top-30 most relevant features in the extravert-introvert (E/I) dimension. The main observation is that different features are relevant for different dimensions. Generally, tf-idf-weighted character n-grams are the most relevant features for all dimensions except for E/I, for which tf-idf-weighted word n-grams are most relevant. However, while LIWC, PSYCH, and LDA100 account for 48% of top-30 most relevant features for the T/F dimension, they have no relevance for the S/N dimension. Post features seem to be relevant only for S/N and J/P dimensions.

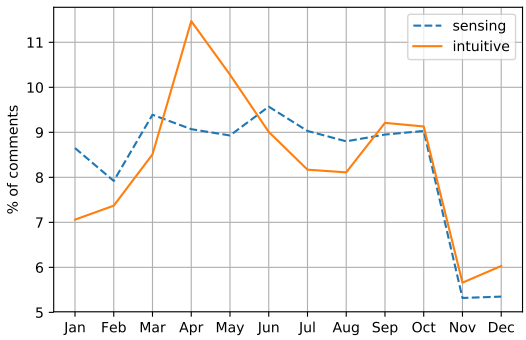
Table 4 offers a complementary view on feature relevance: it shows the proportion of highly relevant features (p-value < 0.001) from each of the feature groups for each dimension. The global, PSYCH, and LIWC features are used in substantial (>50%) proportions for one or more dimensions. The relevance of PSYCH and LIWC features is not surprising, given that these were tailored to model psycholinguistic processes. They seem most indicative for the T/F dimension and, unlike post features, the least relevant for the S/N dimension.

Temporal features. While day-of-week distribution turned out to be a good predictor for T/F and J/P dimensions, posting time differences are relevant only for S/N dimension. Day-of-week proportion of 100% for J/P basically means that all points in the distribution are indicative for that particular dimension. In contrast, the monthly distribution proportion of 25% suggests that only four months in a year are relevant for the S/N dimension. More insight is given by Fig. 1a, which shows the distribution of comments across days of week for the J/P and S/N dimensions. Perceiving types tend to comment more on Tuesdays and Sundays, while judging types comment more on other days. The intuitive types are more active during April and May, while sensing types prefer to comment during January and July.

Word usage. The use of specific words or word classes is known to correlate well with personality traits. Extraversion is characterized by the use of social- and family-related words (Schwartz et al., 2013) and the use of exclamation marks. This is consistent with the most relevant word features for the E/I dimension in our dataset: *Friend*, *Social*, *comm_mbti*, *only*, *i'm an extrovert*, *fri*, *at least*, *drivers*, *Affiliation*, *Exclam*, *origin*, *!!* (word classes from LIWC and PSYCH are shown capitalized). The most relevant words for the S/N dimension are also somewhat expected: *Is_self_mean*, *Is_self_median*, *–*, *i*, *'*, *is a*, *my_*, *it*, *"a*, *Avg_img*, *my*, *_he*, *cliché*, *Sixltr*, *exist*. By definition, sensing types are more concrete while intuitives are more abstract, which seems to be reflected in the imageability feature (e.g., *Avg_img*). Intuitives tend to use more rare (*e.g., cliché*), more complex, and longer words (as signaled, e.g., by the *Sixltr* fea-



(a) J/P distribution across days of week



(b) S/N distribution across months

Figure 1: Temporal distribution of comments

ture: words with more than six characters). Sensing types also seem to share posts with content they found outside Reddit more than intuitives (e.g., *Is_self* features). The feelers tend to use more words about love, feelings, and emotions. They also use more social and affectionate words as well as pronouns and exclamations, as evidenced by the most relevant words for the T/F dimension: *love*, *Feel*, *Posemo*, *valence*, *Emotion*, *happy*, *i*, *polarity*, *!*, *i love*, *Ppron*, *SOCIAL*, *Exclaim*, *Affect*, *Pronoun*, *_so*, *e!*. i The most relevant words for J/P also seem to reflect the common stereotypes, such as that judgers are more plan, work, and family oriented: *Work*, *husband*, *Home*, *help*, *for*, *plan*, *sit*, *hit*, *joke*, *fo*. We leave a more detailed analysis for future work.

5 Personality Prediction

In line with standard practice, we frame the MBTI personality prediction task as four independent binary classification problems, one for each MBTI dimension. In addition, we consider the 16-way multiclass task of predicting the MBTI type, which we accomplish simply by combining together the predictions for the four individual dimensions.

5.1 Experimental Setup

We experiment with three different classifiers: a support vector machine (SVM), ℓ_2-regularized logistic regression (LR), and a three-layer multilayer perceptron (MLP). We use nested stratified cross-validation with five folds in the outer loop and 10 (for LR) or 5 (for SVM) folds in the inner loop; the inner loop is used for model selection with macro F1-score as the evaluation criterion. To investigate the merit of the different features, we (1) train all models with features selected using the t-test and (2) the LR model with each of the feature group separately. Feature selection and standard scaling are applied on training set only, separately for each of the cross-validation folds, and the number of features is also being optimized. Class weighting is used to account for class imbalance. A majority class classifier (MCC) is used as baseline. We use the implementation from Scikit-learn (Pedregosa et al., 2011) for all models.

5.2 Results

Per-dimension prediction. Table 5 shows prediction results for each dimension in terms of the macro F1-score, averaged across the five folds. Although we are using relatively simple models, we achieve surprisingly good results which are well above the baseline. Models using a combination of all features (LR_all and MLP_all) achieve the best results across all dimensions.

Looking into the individual dimensions, the best model for the E/I dimension is MLP_all, but its score is only slightly above the LR word n-gram model. Character n-grams and, to some extent, LIWC and PYSCH were also predictive for the E/I dimension. Models based on topical and user-activity based features did not achieve results above the baseline. Results are similar for the S/N dimension, where MLP_all again outperforms other models, while word-ngram features seem to perform rather well. The overall lowest results are for the T/F dimension, which is consistent with the findings of Capraro and Capraro (2002). Here, n-gram based features perform only slightly better than dictionary-based (LIWC, PSYCH) and topic-based (LDA) features, but overall the differences in model scores are lower. Lastly, for the J/P dimension, the best-performing model is LR_all, well above all models that use a single feature group.

As personality traits are in fact manifested on a continuous scale along each dimension, it makes

Model	Dimensions				Type
	E/I	S/N	T/F	J/P	
LR all	81.6	77.0	**67.2**	**74.8**	40.8
MLP all	**82.8**	**79.2**	64.4	74.0	**41.7**
SVM all	79.6	75.6	64.8	72.6	37.0
LR w_ng	81.0	73.6	66.4	71.8	38.0
LR chr_ng	62.2	64.0	66.4	65.8	26.5
LR liwc	55.0	49.8	65.0	57.4	14.2
LR psych	52.0	48.2	64.0	57.0	12.5
LR lda100	50.0	48.2	62.4	56.2	13.9
LR posts	49.4	53.2	48.0	51.8	9.5
LR subtf	49.6	49.6	50.4	50.2	13.2
MCC	50.04	50.04	50.0	50.02	25.2

Table 5: Macro F1-scores for per-dimension prediction and accuracy of type-level prediction for models with all features, LR models with a single feature group, and the MCC baseline

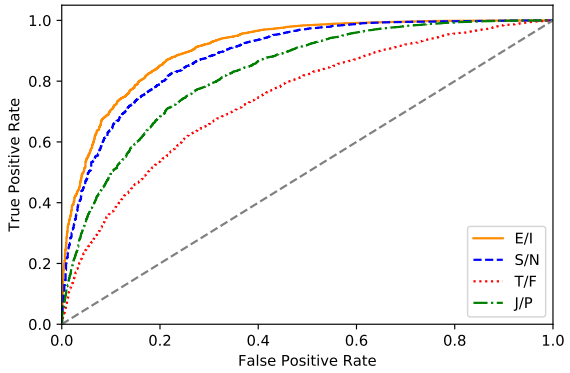


Figure 2: ROC curves for the LR_all model

sense to evaluate type prediction as a confidence-rated classification task using ROC curves. Results are shown on Figure 2. The ROC curve shows the true positive rate (recall) as a function of the false positive rate (fall-out), both of which increase as the classification threshold increases. For instance, the ROC curve for the T/F dimensions tells us that we can detect about 70% of T cases with a fall-out of about 40%.

Type-level prediction. For MBTI type prediction, we concatenated the outputs of the binary models for each individual dimension. Prediction accuracy is shown in the last column of Table 5. The best result is achieved by the MLP_all model, with an accuracy of 42%, while the baseline performs at only 25%. Further insight can be gleaned from Table 6, which shows the breakdown of incorrect predictions for the LR_all model by the number of mismatched dimensions. In 82% of

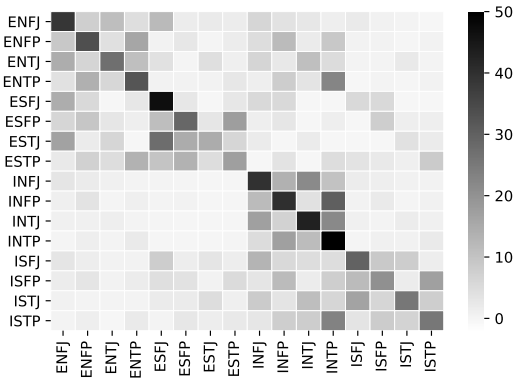


Figure 3: Heatmap of the type prediction confusion matrix

	# mismatches				
	0	1	2	3	4
Count	3757	3715	1384	240	15
%	40.83	40.77	15.61	2.63	0.16

Table 6: The number and percentage of mismatched dimensions between predicted and actual types

cases, the model predicts either the correct type or errs on one dimension, while in more than 97% of cases the model predicts two or more dimensions correctly. The likely mismatches are shown on Fig. 3, showing a heatmap of the type prediction confusion matrix for the LR_all model. The confusion matrix shows that types which are similar in the MBTI theory tend to get grouped together. For example, introverted intuitives tend to be similar and even for people it is often difficult to distinguish between INTP and INTJ. At the same time, INTJ is more similar to INFJ, while INTP is more similar to INFP. The confusion matrix shows that the model was able to capture these nuances.

6 Conclusion

We described MBTI9k, a new, large-scale dataset for personality detection acquired from Reddit. The dataset addresses the shortcomings of the existing datasets, primarily those of user non-anonymity and low topic diversity, and comes with MBTI types and precomputed sets of features for more than 9000 Reddit users.

We carried out two studies on the MBTI9k. In the first, we extracted and analyzed a number of linguistic and user-activity features, demonstrating that there are marked differences in feature values between the different MBTI poles and dimensions. We then used these features to train several benchmark models for personality predic-

tion. The models scored considerably higher than the baseline, ranging from 67% macro F1-score for the T/F dimension to 82% for the S/N dimension. Type-level prediction reaches accuracy of 41% for exact match and 82% for exact or one-off match, which is comparable to the reliability of standardized tests (Lawrence and Martin, 2001). We also found that models using only word n-gram features also perform remarkably well, presumably due to the large size of the dataset.

We envision several directions for future work. First, the dataset could be improved in a number of ways. It could be enlarged with older posts dating back to year 2005, or by increasing the number of users by searching for MBTI declarations in comment texts rather than only the flairs. The same technique could be used to amended the dataset with self-reported demographic data, including age, gender, and location.

On the modeling side, taking into account the success of word-based features and the size of the dataset, using deep learning models for personality might be a reasonable next step. The T/F dimension might, however, require more sophisticated features, judging by the modest performance of the benchmark models on that particular dimension.

In perspective, we believe that Reddit has a lot to offer as a source of data for personality prediction and – more generally – author profiling. A large number of users and comments, highly diverse subcommunities, and the numerous interactions between users are a true gold mine for researchers from both natural language processing and social science communities.

References

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society*. pages 1–16.

Pierre-Hadrien Arnoux, Anbang Xu, Neil Boyette, Jalal Mahmud, Rama Akkiraju, and Vibha Sinha. 2017. 25 tweets to know you: A new model to predict personality with social media. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. pages 472–475.

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.

Robert M. Capraro and Mary Margaret Capraro. 2002. Myers-Briggs type indicator score reliability across: Studies a meta-analytic reliability generalization study. *Educational and Psychological Measurement* 62(4):590–602.

Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. 2013. Workshop on computational personality recognition: Shared task. In *Proceedings of the AAAI Workshop on Computational Personality Recognition*. pages 2–5.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33(4):497–505.

Philip J. Corr and Gerald Matthews. 2009. *The Cambridge handbook of personality psychology*. Cambridge University Press, Cambridge.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*. pages 71–80.

John M. Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology* 41(1):417–440.

Eli J. Finkel, Paul W. Eastwick, Benjamin R. Karney, Harry T. Reis, and Susan Sprecher. 2012. Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest* 13(1):3–66.

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.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from Twitter. In *Proceedings of 2011 IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT) and IEEE International Conference on Social Computing (SocialCom)*. pages 149–156.

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

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

Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70(6):543.

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. https://doi.org/10.1073/pnas.1218772110.

Gordon Lawrence and Charles R. Martin. 2001. *Building people, building programs: A practitioner's guide for introducing the MBTI to individuals and organizations*. Center for Applications of Psychological Type.

Kim Luyckx and Walter Daelemans. 2008. Personae: A corpus for author and personality prediction from text. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. pages 2981–2987.

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* 30:457–500.

David Mascarenas. 2016. A Jungian based framework for artificial personality synthesis. In *Proceedings of the Fourth Workshop on Emotions and Personality in Personalized Systems (EMPIRE)*. pages 48–54.

S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences* 114(48):12714–12719. https://doi.org/10.1073/pnas.1710966114.

Robert R. McCrae and Paul T. Costa. 1989. Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality* 57(1):17–40.

Matthias R. Mehl, James W. Pennebaker, D. Michael Crow, James Dabbs, and John H. Price. 2001. The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers* 33(4):517–523.

Isabel Briggs Myers, Mary H. McCaulley, and Allen L. Hammer. 1990. *Introduction to Type: A description of the theory and applications of the Myers-Briggs type indicator*. Consulting Psychologists Press.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* 42(3):537–593.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology* 77(6):1296.

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.

Barbara Plank and Dirk Hovy. 2015. Personality traits on Twitter –or– how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015)*. pages 92–98. http://www.aclweb.org/anthology/W15-2913.

Daniel Preoţiuc-Pietro, Jordan Carpenter, and Lyle Ungar. 2017. Personality driven differences in paraphrase preference. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, pages 17–26. https://doi.org/http://dx.doi.org/10.18653/v1/W17-2903.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our Twitter profiles, our selves: Predicting personality with Twitter. In *Proceedings of 2011 IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT) and IEEE International Conference on Social Computing (SocialCom)*. pages 180–185.

Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at PAN 2015. In *CLEF 2015 labs and workshops, notebook papers, CEUR Workshop Proceedings*. volume 1391.

Ugur Sak. 2004. A synthesis of research on psychological types of gifted adolescents. *Journal of Secondary Gifted Education* 15(2):70–79.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 2577–2583. https://doi.org/10.18653/v1/D15-1309.

Andrew H. Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791. https://doi.org/10.1371/journal.pone.0073791.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology – From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, pages 58–65. http://aclweb.org/anthology/W17-3107.

Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. pages 1632–1637.

Byron C. Wallace, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*. pages 512–516. http://www.aclweb.org/anthology/P14-2084.