Matching Theory and Data with Personal-ITY: What a Corpus of Italian YouTube Comments Reveals About Personality

Elisa Bassignana[♢♡] **Malvina Nissim**[♣] **Viviana Patti**[♡]
[♡]Dipartimento di Informatica, Università degli Studi di Torino, Italy
[♢]Department of Computer Science, IT University of Copenhagen, Denmark
[♣] CLCG – Faculty of Arts, University of Groningen, The Netherlands
[♡]viviana.patti@unito.it, [♢]elba@itu.dk
[♣]m.nissim@rug.nl

Abstract

As a contribution to personality detection in languages other than English, we rely on distant supervision to create Personal-ITY, a novel corpus of YouTube comments in Italian, where authors are labelled with personality traits. The traits are derived from one of the mainstream personality theories in psychology research, named *MBTI*. Using personality prediction experiments, we (i) study the task of personality prediction in itself on our corpus as well as on TWISTY, a Twitter dataset also annotated with MBTI labels; (ii) carry out an extensive, in-depth analysis of the features used by the classifier, and view them specifically under the light of the original theory that we used to create the corpus in the first place. We observe that no single model is best at personality detection, and that while some traits are easier than others to detect, and also to match back to theory, for other, less frequent traits the picture is much more blurred.

1 Introduction

Human Personality is a psychological construct aimed at explaining the wide variety of human behaviours in terms of a few, stable and measurable individual characteristics (Snyder, 1983; Parks and Guay, 2009; Vinciarelli and Mohammadi, 2014). Research in psychology has formalised these characteristics into what are known as *Trait Models*. Two are the major models widely adopted also outside of purely psychological research (see Section 2.1): *Big Five* (John and Srivastava, 1999) and *Myers-Briggs Type Indicator* (*MBTI*) (Myers and Myers, 1995).

The psychological tests commonly used to detect prevalence of traits include human judgements regarding semantic similarity and relations between adjectives that people use to describe themselves and others. This is because language is believed to be a prime carrier of personality traits (Schwartz et al., 2013). This aspect, together with the progressive increase of available user-generated data from social media, has prompted the task of *Personality Detection*, i.e., the automatic prediction of personality from written texts (Whelan and Davies, 2006; Argamon et al., 2009; Celli et al., 2013; Youyou et al., 2015; Litvinova et al., 2016; Verhoeven et al., 2016). Personality detection can be useful in predicting life outcomes such as substance use, political attitudes and physical health. Other fields of application are marketing, politics, psychological and social assessment and, in the computational domain, dialogue systems (Ma et al., 2020) and chatbots (Qian et al., 2018).

As a contribution to personality detection in languages other than English, we have developed Personal-ITY, a novel corpus of YouTube comments in Italian, which are annotated with MBTI personality traits. The corpus creation methodology, described in detail in (Bassignana et al., 2020), makes use of a Distant Supervision approach that can also serve as a blueprint to develop datasets for other languages. In this work, we use Personal-ITY to cast light not only on the feasibility of personality detection *per se* on this corpus, but also on the relationship between our data and the theory it is based on. More specifically, we want to investigate if our distantly obtained labels are meaningful with respect to the psychological theory they come from, and whether language does indeed reflect the traits that should be associated with such labels. To do this, we run a series of in- and cross-dataset experiments; on top of

performance analysis we conduct an in depth study on the relevant features used by the classifiers, and how they might relate to the source psychological theory.

Personal-ITY is available at `https://github.com/elisabassignana/Personal-ITY`.

2 Background

Personality profiling is addressed both from a psychological viewpoint (traits model for the classification of personality) and from a computational perspective in the field of Natural Language Processing. We provide relevant background from both sides, as they are intertwined in our work.

2.1 Psychological models

There are two main personality trait models widely accepted and used by the research community, also outside of psychology: *Big Five* and *Myers-Briggs Type Indicator* (*MBTI*).

Big Five (John and Srivastava, 1999), also know as Five-Factor Model (FFM) or the OCEAN model, was developed from the 1980s onwards. This theory outlines five global dimensions of personality and describes people by assigning a score in a range for each of them. The five traits considered are: OPENNESS TO EXPERIENCE, CONSCIENTIOUSNESS, EXTROVERSION, AGREEABLENESS and NEUROTICISM. Thus a person's personality would be defined through five corresponding scores indicating the positive or negative degree to which each dimension is expressed. Interestingly, the Big Five tests usually assign scores on the basis of semantic associations between personality traits and words considering the texts of the users' answers, rather than relying on neuropsychological experiments. The use of the Big Five model in computational approaches to personality began more than one decade ago and it is now widely accepted in academia.

The MBTI (Myers and Myers, 1995) theory is based on the conceptual theory of the Psychological Types proposed by the psychiatrist Carl Jung, who had speculated the main dimensions able to describe how people experience the world. Assessment is based on a self-reported psychological questionnaire that helps researchers to classify people into one personality type out of sixteen. The sixteen labels are the product of binary labels over four different dimensions, as follows: (EXTRAVERT-INTROVERT, INTUITIVE-SENSING, FEELING-THINKING, PERCEIVING-JUDGING). Each person is assumed to have one dominant quality from each category, thus producing sixteen unique types. Examples of full personality types are therefore four letter labels such as ENTJ or ISFP.

The initial intention of the test was to help women who were entering the industrial workforce for the first time during the Second World War to identify the "best, most comfortable and effective" job for them based on their personality type. In later years, MBTI continued to be used in order to predict validity of employees' job performance and to help students in their choice of career or course of study.

Although several studies suggest that the MBTI test lacks convincing validity data for these types of applications as it can measure preferences and not ability, it continues to be popular because it is very easy to administer it and it is not difficult to understand.

2.2 Personality Detection

Most approaches to automatic personality detection are supervised models trained on silver or gold labelled data. In this section, we revise existing datasets and standard methods of personality detection.

Corpora There exist a few datasets annotated for personality traits. For the shared tasks organised within the *Workshop on Computational Personality Recognition* (Celli et al., 2013), four English datasets annotated with the *Big Five* traits have been released. For the 2013 edition, the data contained "Essays" (Pennebaker and King, 2000), which is a large dataset of stream-of-consciousness texts collected between 1997 and 2004, and "myPersonality"[1], a corpus collected from FaceBook including information on user social network structures. For the 2014 edition, the data consisted of the "YouTube Personality Dataset" (Biel and Gatica-Perez, 2013), which contains a collection of behavioural features, speech transcriptions, and personality impression scores for a set of 404 YouTube vloggers, and "Mobile Phones", which is a

[1] `http://mypersonality.org`

Corpus	Model	# user	Avg.
PAN2015	Big Five	38	1,258
TWISTY	MBTI	490	21,343
Personal-ITY	MBTI	1048	10,585

Table 1: Summary of Italian corpora with personality labels. Avg.: average tokens per user.

collection of call logs and proximity data of 53 subjects living in a student residency of a major US university, collected through a special software incorporated in their phones (Staiano et al., 2012).

Schwartz et al. (2013) collected a Big Five annotated dataset of FaceBook comments (700 millions words) written by 136.000 users who shared their status updates. Interesting correlations were observed between word usage and personality traits.

For the 2015 PAN Author Profiling Shared Task (Pardo et al., 2015), personality was added to gender and age in their standard profiling task, with tweets in English, Spanish, Italian and Dutch annotated according to the *Big Five* model assigning a score in a range [-0.5; +0.5] for each trait.

If looking at data labelled with the MBTI traits, we find a corpus of 1.2M English tweets annotated with personality and gender (Plank and Hovy, 2015), and the multilingual dataset TWISTY (Verhoeven et al., 2016). The latter is a corpus of data collected from Twitter using a Distant Supervision approach. It is annotated with MBTI personality labels and gender for six languages (Dutch, German, French, Italian, Portuguese and Spanish), and includes a total of 18,168 authors.

As we concentrate on Italian, we report in Table 1 an overview of the available Italian corpora labelled with personality traits. We include information on our own Personal-ITY corpus, which is described in Section 3. For TWISTY, we only report information for the Italian portion.

Computational work exists also on comparing (labels from) the two models (Celli and Lepri, 2018). Furnham et al. (2003) defined some correlations between various dimensions across the two trait models: Big Five Extraversion is correlated with MBTI Extraversion-Introversion, Openness to Experience is correlated with Sensing-Intuition, Agreeableness with Thinking-Feeling and Conscientiousness with Judging-Perceiving. In order to increase the amount of data we could work with, and to obtain a general, usable model, we tried to convert the PAN 2015 Italian data to MBTI annotations. We considered the mid value in the Big Five range as threshold between the opposite poles of MBTI dimensions. This experiments didn't led to any informative results, probably due the the small dimension of the corpus (see Table 1): almost all the few users present have been annotated with the same MBTI label.

Detection Approaches Regarding detection approaches, Mairesse et al. (2007) tested the usefulness of different sets of textual features making use of mostly SVMs. At the PAN 2015 challenge (see above) a variety of algorithms were tested (such as Random Forests, decision trees, logistic regression for classification, and also various regression models), but overall most successful participants used SVMs. Regarding features, participants approached the task with combinations of style-based and content-based features, as well as their combination in *n*-gram models (Pardo et al., 2015).

Experiments on TWISTY were performed by the corpus creators themselves using a LinearSVM with word (1-2) and character (3-4) *n*-grams. Their results (reported in Table 4a for the Italian portion of the dataset) are obtained through 10-fold cross-validation; the model is compared to a weighted random baseline (WRB) and a majority baseline (MAJ). Let us notice that the model proposed in (Verhoeven et al., 2016) for the Italian language is the only one not reaching any baseline (for all the other languages the model proposed reach at least the weighted random baseline). This also prompted us to work on the Italian language, where there is still ample room for improvement on the development of resources and models for the personality detection task. More in general, our choice has to be seen as an intention of improving the state of the art for languages other than English (Joshi et al., 2020).

Recent computational approaches on personality prediction from texts investigated the use of deep learning (Majumder et al., 2017) and regression models (Akrami et al., 2019).

3 Data

To run experiments on personality detection, we have created a dedicated corpus with MBTI labels, exploiting distant supervision: Personal-ITY (Bassignana et al., 2020). Here, we summarise the choices that we made regarding the source of the data and the theoretical trait model, the procedure followed to construct the corpus, and provide a description of the resulting dataset. In addition, we also partly use the existing TWISTY (see Section 2.2 and Table 1).

Because we deal with personal data, and because we do believe profiling is a sensitive task in general, we also provide an Ethics Statement.

Ethics Statement

Personality profiling must be carefully evaluated from an ethical point of view. In particular personality detection can involve ethical issues regarding the appropriate use and interpretation of the prediction outcomes (Weiner and Greene, 2017). Also, concerns have been raised regarding the inappropriate use of these tests with respect to invasion of privacy, cultural bias and confidentiality (Mehta et al., 2019).

The data included in the Personal-ITY dataset was publicly available on the YouTube platform at the time of the collection. As we explain in this Section (but see (Bassignana et al., 2020) for details), the information collected consists in comments published under public videos on the YouTube platform by the authors themselves. For an increased protection of user identities, in the released corpus only the YouTube usernames of the authors are mentioned, which are not unique identifiers. The YouTube IDs of the corresponding channels, which are instead unique identifiers on the platform and would allow to trace back the identity of the authors, are not released. The corpus was created for academic research purposes, and is not intended for commercial deployment or applications.

3.1 Source and Theoretical Model

YouTube is the source of data for our corpus. The decision is grounded on the fact that compared to the more commonly collected tweets, YouTube comments can be longer, so that users are freer to express themselves without limitations. Additionally, there is a substantial amount of available data on the YouTube platform, which is easy to access thanks to the free YouTube APIs.

Our theoretical trait model of choice is MBTI. Although it has been extensively criticized for a number of limitations (McCrae and Costa, 1989; Boyle, 1995 03; Pittenger, 2005), and the Big Five model seems to be more widely accepted in psychology, our choice has been driven by two main reasons. The first benefit of this decision is that MBTI is easy to use in association with a Distant Supervision approach (just checking if a message contains one of the 16 personality types; see Section 3.2). Another benefit is related to the existence of TWISTY. Since both TWISTY and Personal-ITY implement the MBTI model, analyses and experiments over personality detection can be carried out also in a cross-domain setting.

3.2 Corpus Creation and Description

The fact that users often self-disclose information about themselves on social media makes it possible to adopt *Distant Supervision* (DS) for the acquisition of training data. DS is a semi-supervised method that has been abundantly and successfully used in affective computing and profiling to assign silver labels to data on the basis of indicative proxies (Go et al., 2009; Pool and Nissim, 2016; Emmery et al., 2017).

We observed that some YouTube Italian users were used to leave comments to videos on the MBTI theory, in which they were stating their own personality type (e.g. *Sono ENTJ...chi altro?* [en: "I'm ENTJ...anyone else?"]; *INTP, primo test che effettivamente ha ragione* [en: "INTP, the first test that is actually right"]). We exploited such comments to create Personal-ITY.

The methodology, explained in detail in (Bassignana et al., 2020), consisted in creating automatically a list of YouTube users annotated with MBTI personality labels starting from the comments cited above. In the second macro-step, we adopted a Distant Supervision approach in order to retrieve as much as possible texts written by the authors whose personality was known. The result of this procedure led to a final corpus with a conspicuous number of users and comments, where only authors with at least five comments, each at least five token long, are included.

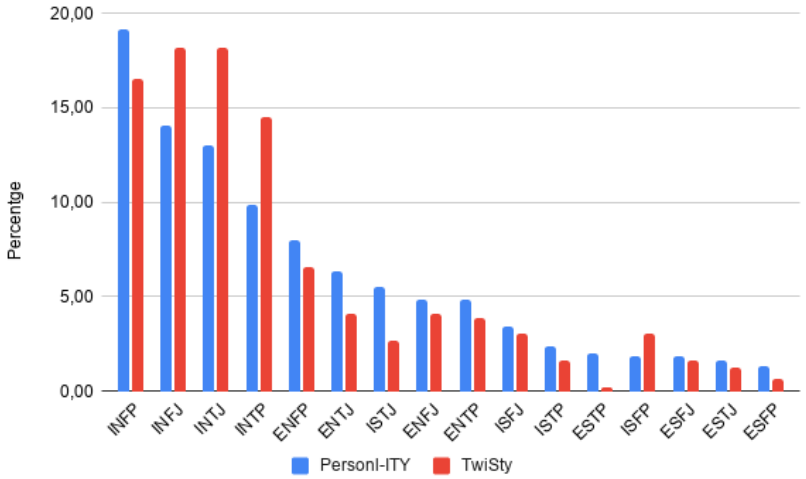

Figure 1: Distribution of the 16 labels in the YouTube corpus and in the Italian part of TWISTY.

Personal-ITY contains 1048 users, each annotated with an MBTI label. The average number of comments per user is 92 and each message is on average 115 tokens. Table 2 shows explicitly the comparison between our new corpus and TWISTY.

Corpus	# Users	Avg. comments/user	Avg. tokens/comment	Avg. tokens/user
Personal-ITY	1048	92	115	10,585
TWISTY	490	1,903	11	21,343

Table 2: Statistical comparison between Personal-ITY and TWISTY.

The amount of the 16 personality types in the corpus is not uniform. Figure 1 shows such distribution and also compares it with the one in TWISTY. It can be observed that the two corpora present quite similar percentages of personality types. Some little differences (e.g., *INFJ*, *INTJ*, *INTP*) can be explained considering the different data sources: Personal-ITY is collected from YouTube, while TWISTY is collected from Twitter. The general unbalanced distribution can be due to personality types not being uniformly distributed in the population, and to the fact that different personality types can make different choices about their online presence. Goby (2006) for example, observed 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 for introverts. Figure 1 confirms this theory as the four most frequent types are introverts in both datasets. The conclusion is that, despite the different biases, collecting linguistic data in this way has the advantages that it reflects actual language use and allows large-scale analysis (Plank and Hovy, 2015).

4 Experiments

We ran a series of initial experiments on Personal-ITY, which we use to peek into the relationship between labels and theory, and which can also serve as a baseline for future work on this dataset.

The choices related to the experimental framework have been driven by the interest on investigating signals and linguistic cues for personality traits, analyzed through the lens of psychological studies on personality. This perspective has therefore led us not use deep learning techniques that would have made this analysis task more complex. Rather, we opt for state-of-the-art approaches commonly used in author profiling, developing interpretable models that can aid the analysis process. These will make it possible to perform a linguistic and psychological analysis, though perhaps at the expense of performance.

Specifically, we used the `sklearn` (Pedregosa et al., 2011) implementation of a linear SVM

(LinearSVM), with standard parameters, and tested three types of features: lexical-, stylistic-, and embeddings-based. We used four placeholders for hashtags, urls, usernames and emojis.

At the lexical level, we experimented with word (1-2) and character (3-4) *n*-grams, both as raw counts as well as tf-idf weighted. Character *n*-grams were tested also with a word-boundary option. Considering stylistic features, we investigated the use of emojis, hashtags, pronouns, punctuation and capitalisation. Lastly, we also experimented with embeddings-based representations, using more generic (Pennington et al., 2014) and YouTube-specific (Nieuwenhuis and Nissim, 2019) pre-trained models. We created one representation per user by averaging the vectors for all words written by that user.

We used 10-fold cross-validation, and assessed the models using macro f-score. We deem this way of averaging over f-scores per class appropriate, since the dataset is quite unbalanced, but we want good performance for each class. For comparison, we calculated a majority baseline (MAJ).

Table 3 shows the results of our experiments with the different feature types described above. Regarding *n*-grams and embeddings representations, we report results for the best configurations, namely character *n*-grams for the lexical features, and GloVe embeddings. Overall, lexical features perform best. Combining different feature types did not lead to any improvement. Classification was performed both with four separate binary classifiers (one per dimension), as well as with one single classifier predicting a total of four classes, i.e, the whole MBTI labels at once. Interestingly, in the latter case, we observe that the results are quite high considering the increased difficulty of the task.

Table 4 reports the scores of our models on TWISTY. Because the original TWISTY paper uses micro f-score, for the sake of comparison, in Table 4a we include the results of our experiments using micro-f for the MAJ baseline and our lexical *n*-gram model. For all traits, our models achieve better results (micro-f) than those reported in the original TWISTY paper (Verhoeven et al., 2016). In Table 4b, instead, there are the macro-f of the experiments we performed on TWISTY. As for Personal-ITY, best results were achieved using lexical features (tf-idf word *n*-grams); results of models with stylistic features and embeddings were just above the baseline.

Trait	Binary classification				Full label at once
	MAJ	*n*-grams	Sty	Emb	*n*-grams
EI	40.55	51.85	40.46	40.55	51.65
NS	44.34	51.92	44.34	44.34	49.04
FT	35.01	50.67	36.27	35.01	50.86
PJ	29.49	50.53	51.04	47.06	51.03
Avg	37.35	**51.24**	43.03	41.74	50.65

Table 3: Results of the experiments on Personal-ITY. Predictions of the full MBTI label at once were performed using the model performing best in the binary classification.

	(Verhoeven et al., 2016)			Our experiments	
Trait	WRB	MAJ	Lex	MAJ	*n*-grams
EI	65.54	77.88	77.78	77.75	**79.18**
NS	75.60	85.78	79.21	85.92	**85.92**
FT	50.31	53.95	52.13	53.67	**55.31**
PJ	50.19	53.05	47.01	53.06	**54.08**
Avg	60.41	67.67	64.06	67.6	**68.62**

(a) Comparison between our experiments on TWISTY and the ones in (Verhoeven et al., 2016). Scores reported are micro f-scores. WRB = weighted random baseline. MAJ = majority baseline.

Trait	MAJ	*n*-grams	Sty	Emb
EI	43.69	55.23	43.69	43.69
NS	46.15	46.15	46.15	46.15
FT	34.79	52.98	35.34	34.70
PJ	34.56	53.01	35.20	34.90
Avg	39.80	**51.84**	40.09	39.86

(b) Results of our experiments on TWISTY by using macro f-scores.

Table 4: Results of our experiments on TWISTY.

To test compatibility of resources and to assess model portability, we also ran cross-domain experiments on Personal-ITY and TwiSty. We divided both corpora in fixed training and test sets with a proportion of 80/20 so that the test set stays the same for in-domain and cross-domain settings. Indeed, we run the in-domain models again using this split. The models use lexical features as they are the one performing better overall the others. Results are shown in Table 5. We ran the experiments both using a binary classification of each trait separately and with the prediction of the full MBTI label at once. This latter leads to results even better on the considered sets. Cross-domain scores are obtained with the best in-domain model[2,3]. They drop substantially compared to in-domain, but are always above the baseline.

We specify that in every experiments done (Tables 3–4–5) we chose to look for the best model able to predict the whole MBTI personality and so we report the highest scores based on averages of the four traits. Considering the four dimensions individually better results can be obtained by using specific models. Table 6 shows an overview of the best models considering each trait independently for each source of data we tested. Results are quite scattered: there is no a single best model for personality predictions, as feature contribution depends on the dimension considered, and on the dataset. This observation confirms the inherent difficulty of the Personality Detection task from written texts.

	Binary classification						Full MBTI label at once					
Train	Personal-ITY			TwiSty			Personal-ITY			TwiSty		
Test	IN	CROSS		IN	CROSS		IN	CROSS		IN	CROSS	
	Pers	MAJ	Twi	Twi	MAJ	Pers	Pers	MAJ	Twi	Twi	MAJ	Pers
EI	58.94	44.94	49.33	55.66	44.59	44.59	54.67	44.94	44.94	59.77	44.59	44.59
NS	52.88	47.87	47.31	47.87	45.31	45.31	48.65	47.87	47.59	47.87	45.31	45.31
FT	49.20	37.58	47.09	65.26	39.13	51.04	54.18	37.58	46.38	61.98	26.32	46.71
PJ	54.43	32.41	32.50	56.87	36.56	38.54	58.07	32.41	38.08	50.27	36.56	46.80
Avg	53.86	40.70	44.06	56.42	41.40	44.87	53.89	40.70	**44.25**	54.97	38.20	**45.85**

Table 5: Results of the cross-domain experiments. MAJ = baseline on the cross-domain testset.

Source	Trait	Tf-idf			Count		
		word	char	char_wb	word	char	char_wb
Personal-ITY	EI						X
	NS						X
	FT			X			
	PJ		X				
TwiSty	EI	X					
	NS					X	
	FT			X			
	PJ	X					
Cross-domain: Tr: Personal-ITY Te: TwiSty	EI				X		
	NS			X			
	FT						X
	PJ						X
Cross-domain: Tr: TwiSty Te: Personal-ITY	EI					X	
	NS					X	
	FT	X					
	PJ						X

Table 6: Best model for each trait considering each source individually.

The in-domain experiments show that performance over the two datasets is very similar overall, though

[2]Binary classification: Train on Personal-ITY: character *n*-grams. Train on TwiSty: tf-idf character *n*-grams.

[3]Full MBTI label at once: Train on Personal-ITY: character *n*-grams. Train on TwiSty: tf-idf word *n*-grams.

with some differences regarding the best and worst predicted traits. However, we observe that cross-domain performance drops by approximately 10 points, independently of the direction of training and testing taken. This underlines differences in the dataset which might not make them fully compatible. For our in-depth linguistic analysis, we choose to concentrate on Personal-ITY, mainly due to the availability of longer author's comments, which can give rise to more interesting insights when studying word-based feature contribution in connection with the source MBTI personality theory.

5 Feature Analysis and Linguistic Cues for Personality Traits

In this section we discuss possible correlations between linguistic cues derived from the experiments described in Section 4 on the Personal-ITY corpus and psychological traits descriptions deriving from the field of Psychology. Table 7 shows the most important features on which our classifier bases its decisions (i.e., the most relevant ones based on the weight they have on the model prediction). We report word *n*-gram features as they are the most interpretable for a psychological analysis.

	Extravert	Introvert	Sensing	Intuition	Thinking	Feeling	Judging	Perceiving
Count word (1,2)	punti	sbagliato	adoro	fuori	**ma io**	**io sono**	**alle**	**bene**
	poi	**che era**	**gli**	**ehm**	nel	giorno	**un video**	12
	ora	fa	**del**	sbagliato	**vero**	beh	**nella**	**emoji ho**
	ne	via	oddio	**vorrei**	anni	marco	ancora	che ti
	ancora	**anzi**	credo	morto	**vedo**	nero	**tuoi**	non ci
	emoji non	hashtag se	qualcuno	che tu	17	sotto	**minecraft**	ho fatto
	volevo	**lui**	solo	alcune	tanto	più	so se	**beh**
	midna	sia	io mi	serie	con le	avevo	**te**	**nether**
	dal	**era**	idea	tutto il	chi	**dentro**	**metti**	tanto
	fosse	**penso**	va	secondo	**capire**	**trovo**	**neanche**	test
Tf-idf word (1,2)	in	nn	matteo	die	**perchè**	più	**user**	perchè
	xd	eren	ahah	die die	emoji	emoji emoji	nn	**emoji**
	che	bella playerinside	**del**	raiden	**vedo**	marco	**minecraft**	test
	ancora	genio genio	adoro	cane	lullaby	nn	**da**	**anche**
	ci	playerinside xd	**di**	cuticole	dei	avevo	**tuoi**	genio genio
	davvero	**marco**	erenblaze	anche	xd	test	hashtag	non
	molto	die	libro	tano	un	genio genio	**alle**	u3000
	di	tifo	marco	**ehm**	**questa**	raiden	di eren	**bene**
	perchè	bella	**in**	copia	eren	in	leo	u3000 u3000
	dei	pixelmon	persone	00 00	grazie	u3000	puoi	un

Table 7: Most predictive word *n*-grams on Personal-ITY.

Below we are going to analyze each MBTI trait independently by interpreting observed features with existing theoretical definitions of the personality types, which we take from (Geyer, 2014).

- EXTRAVERT vs. INTROVERT trait:
  - EXTRAVERT: extraverts tend to use abstract words, to be vague and to use adjectives. They talk a lot (in respect to their opposite) about family, friends, groups and social activities.
  - INTROVERT: introverts, on the other side, tend to use concrete words, to be more precise and so to use nouns, pronouns, articles, numbers and distinctions (*but*, *except*...).

  Referring to Table 7, we find *emoji* in the extraversion pole (recall that we normalized all the emojis in the corpus[4]), while it is not present in the opposite pole. Looking at the definition, we linked this to a more extroverted behavior, people talking about friends, groups and social activities. Similar to that label there is, in the same column, *xd*, that we intended as an emoticon. Moreover, closely related to the definition (being abstract and vague) there is the word *fosse* (subjunctive form of the verb 'to be'). In the introversion column, instead, there are more concrete and precise words as *che era* ('which was'), *era* ('it was') and *penso* ('I think'). With the same intention to be concrete and precise, those people use (proper) nouns and pronouns as *lui* ('he') and *marco*. Lastly, coherently with the introvert definition, we find *anzi* ('rather', 'instead'), word belonging to the distinction set.

[4] https://pypi.org/project/emojis/

- SENSING vs. INTUITION trait:
  - SENSING: these people are realistic, they usually talk about practical activities and about what is already happened. They describe facts specifying a lot of details, tangible information and rely a lot on senses.
  - INTUITION: people with this personality follow intuition, fantasy, imagination and ideas. They usually talk about what is going to happen, future possibilities and relate their discourses to abstract and general principles.

  In the sensing list of words we highlighted *gli* ('the'), *del* ('of the'), *di* ('of') and *in* ('in') as those tokens are used to specify details. In the opposite column, in line with the definition, we just found the words *ehm* and *vorrei* ('I would like to').

- THINKING vs. FEELING trait:
  - THINKING: these people follow logic, objectivity, rationality, causality and consistency.
  - FEELING: their opposite, instead, are more inclined to follow the heart and principles; they look for cooperation, harmony and are more sensitive.

  In line with the definition of thinking we highlighted the *n*-grams *ma io* ('but I'), *vero* ('true'), *vedo* ('I see'), *capire* ('to understand'), *perchè* ('because') and *questa* ('this'). For their opposite we found *io sono* ('I am'), *dentro* ('inside') and *trovo* ('I think', 'I find').

- JUDGING vs. PERCEIVING trait:
  - JUDGING: this trait indicates determined people, who are used to plan everything and that are comfortable with rules and guide lines.
  - PERCEIVING: the opposite are people who like improvisation and tend to keep open options. They are more relaxed and look for liberty.

  In the judging column there are words such as *alle* ('at'), *nella* ('in'), *tuoi* ('yours'), *te* ('you'), *metti* ('put'), *neanche* ('neither'), *da* ('from', 'by') and *user* (which is a normalized label deriving from pre-processing; it replaces references to specific users). These can all be used to make precise plans, and fit with the description of determined people, comfortable with rules and guidelines. The more interesting word *n*-grams in this set, is *minecraft*, which is the name of a video game where planning skills are fundamental. In the opposite pole, perceiving, some distinctive words are: *bene* ('good'), *beh*, the label *emoji* and *anche* ('also').

A final consideration about the analysis above is that the correlations we found are sometime weak and not so explicit, especially for the S-N trait and we observed that this is coherent with the not so high results obtained from experiments in Section 4: it is likely that the absence of strong evidences linking linguistic cues with psychological ones, makes the decision of the classifier hard. Notice that a similar observation concerning the difficulty to predict the S-N trait has been reported in (Plank and Hovy, 2015) on an English corpus, suggesting that such trait could in general be more related to perception, with a weak linguistic signal. A second observation is that Table 7 contains many unexpected tokens which apparently have no explanation. Some examples are: *midna* for extravert, *eren*, *die playerinside* and *pixelmon* for introvert, *erenblaze* for sensing, *die* again also for intuition, *eren*, *17* and *lullaby* for thinking, *raiden* for feeling, *eren* again for judging and *u3000* for perceiving. The explanation we give to the presence of such 'specific' tokens is to be found in the source and in the way we collected the corpus. Channels on YouTube gather a community of users with similar interests and common saying. It is therefore likely that users commenting videos from a given channel develop some sort of shared, own slang. In this perspective, a further analysis of the dataset aimed at detecting topics and word distribution would be useful to identify the presence of a narrow set of specific-domain texts.

For a second qualitative analysis we used *Wordify*[5], a tool developed by Bocconi University whose intent is to identify words that discriminate categories in textual data. Table 8 reports the results of such

[5] https://wordify.unibocconi.it/index

tool on Personal-ITY, and we can observe various correspondences with Table 7 (terms appearing in both Tables) such as, regarding the first trait, *xd* and *davvero* for EXTRAVERT and *eren*, *playerinside xd* and *nn* for INTROVERT. This 'double check' makes bold words in Table 8 reliable features in MBTI personality prediction, even if for most of them we didn't find a direct psychological explanation coherent with the trait definitions.

Trait	Term	Score	Label	Trait	Term	Score	Label
First trait	sì	0,384	E	Second trait	emoji	0,312	N
	xd	0,352	E				
	davvero	0,328	E				
	eren	0,474	I				
	emoji	0,466	I				
	bello **playerinside**	0,400	I				
	playerinside xd	0,324	I				
	nn	0,312	I				
Third trait	**test**	0,414	F	Fourth trait	**hashtag**	0,344	J
	qi	0,364	F		**nn**	0,336	J
	commento	0,338	F		**eren**	0,306	J
	ahah	0,420	T		**test**	0,414	P
	perché	0,336	T		ahah	0,33	P
	ahahah	0,306	T		xd	0,330	P
					sì	0,318	P

Table 8: Results of the *Wordify* tool on Personal-ITY.

6 Conclusions

We presented Personali-ITY, a novel YouTube-based Corpus for Personality Prediction in Italian. An exploratory empirical investigation on our new corpus confirms that identifying MBTI personality trait from social media texts is challenging. Lexical features perform best, but they tend to be strictly related to the context in which the model is trained and so to overfit. Concerning our experiments on TWISTY, our model outperforms the original TWISTY results (Verhoeven et al., 2016) for Italian and provides a new baseline on this corpus. Moreover we performed cross-domain experiments between YouTube and Twitter data, achieving scores above the baseline. Performance drops from the in- to the cross-genre setting show limits in portability, and leave a lot of space for improvements.

In general, better results could, probably, be obtained using more complex models by using neural networks (e.g. LSTM) (Mehta et al., 2019). The choice to use a simpler SVM model was driven by its greater understandability and transparency. These qualities allowed the features analysis of Section 5.

The inherent difficulty of the task itself is confirmed and deserves further investigations, as assigning a definite personality is an extremely subjective and complex task even for humans. The distant supervision approach remains promising, also applied to YouTube data. Indeed, assigning an 'absolute' personality to an individual is difficult. Even if, to some extent, distant supervision is inaccurate and can lead to the creation of corpus containing bias and noise, since there is no control on user statements, we cannot avoid considering that according to the literature also professional psychological tests can lead to inaccurate results. Moreover, DS, despite its inaccuracy, allows the availability of a large amount of data, necessary for addressing the task by using machine learning approaches.

Acknowledgments

The work of Elisa Bassignana was partially carried out at the University of Groningen within the framework of the Erasmus+ program 2019/20.

References

Nazar Akrami, Johan Fernquist, Tim Isbister, Lisa Kaati, and Björn Pelzer. 2019. Automatic extraction of personality from text: Challenges and opportunities. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3156–3164. IEEE.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123, February.

Elisa Bassignana, Malvina Nissim, and Viviana Patti. 2020. Personal-ity: A novel youtube-based corpus for personality prediction in italian. In Felice Dell'Orletta, Johanna Monti, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020), Bologna, Italy, March 1-3, 2021*, CEUR Workshop Proceedings. CEUR-WS.org.

Joan-Isaac Biel and Daniel Gatica-Perez. 2013. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia, IEEE Transactions on*, 15(1):41–55.

Gregory J Boyle. 1995-03. Myers briggs type indicator (mbti): some psychometric limitations. *Australian Psychologist*, 30(1):71–74.

Fabio Celli and Bruno Lepri. 2018. Is big five better than mbti? a personality computing challenge using twitter data. In *CLiC-it*.

Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. 2013. Workshop on computational personality recognition: Shared task. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Chris Emmery, Grzegorz Chrupała, and Walter Daelemans. 2017. Simple queries as distant labels for predicting gender on Twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 50–55, Copenhagen, Denmark, September. Association for Computational Linguistics.

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 - SOC BEHAV PERSONAL*, 31:577–584, 01.

Peter Geyer. 2014. C.G.Jung's psychological types, the MBTI, and ideas of social adjustment. In *Proceedings of AusAPT Meeting Adelaide*, 03.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Valerie Goby. 2006. Personality and Online/Offline Choices: MBTI Profiles and Favored Communication Modes in a Singapore Study. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 9:5–13, 03.

Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of personality: Theory and research*, page 102–138. Guilford Press.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July. Association for Computational Linguistics.

Tatiana Litvinova, P. Seredin, Olga Litvinova, and Olga Zagorovskaya. 2016. Profiling a set of personality traits of text author: What our words reveal about us. *Research in Language*, 14, 12.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50 – 70.

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, sep.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32:74–79, 03.

Robert McCrae and Paul Costa. 1989. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57:17–40, 03.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2019. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27.

I.B. Myers and P.B. Myers. 1995. *Gifts Differing: Understanding Personality Type*. Mobius.

Moniek Nieuwenhuis and Malvina Nissim. 2019. The Contribution of Embeddings to Sentiment Analysis on YouTube. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Francisco M. 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 *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Laura Parks and Russell P Guay. 2009. Personality, values, and motivation. *Personality and individual differences*, 47(7):675–684.

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 Pennebaker and Laura King. 2000. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77:1296–312, 01.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

David Pittenger. 2005. Cautionary comments regarding the myers-briggs type indicator. *Consulting Psychology Journal: Practice and Research*, 57:210–221, 06.

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*, pages 92–98, Lisboa, Portugal, September. Association for Computational Linguistics.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization, 7.

H. Andrew 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.

Mark Snyder. 1983. The influence of individuals on situations: Implications for understanding the links between personality and social behavior. *Journal of personality*, 51(3):497–516.

Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. 2012. Friends don't lie - inferring personality traits from social network structure. In *UbiComp'12 - Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 321–330, 09.

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'16)*, pages 1632–1637, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Irving B. Weiner and Roger L. Greene, 2017. *Ethical Considerations In Personality Assessment*, chapter 4, pages 59–74. Wiley.

Susan Whelan and Gary Davies. 2006. Profiling consumers of own brands and national brands using human personality. *Journal of Retailing and Consumer Services*, 13(6):393–402.

Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.