Personal-ITY: A Novel YouTube-based Corpus for Personality Prediction in Italian

Elisa Bassignana
Dipartimento di Informatica
University of Turin
elisa.bassignana@edu.unito.it

Malvina Nissim
CLCG
University of Groningen
m.nissim@rug.nl

Viviana Patti
Dipartimento di Informatica
University of Turin
viviana.patti@unito.it

Abstract

We present a novel corpus for personality prediction in Italian, containing a larger number of authors and a different genre compared to previously available resources. The corpus is built exploiting Distant Supervision, assigning *Myers-Briggs Type Indicator* (*MBTI*) labels to YouTube comments, and can lend itself to a variety of experiments. We report on preliminary experiments on Personal-ITY, which can serve as a baseline for future work, showing that some types are easier to predict than others, and discussing the perks of cross-dataset prediction.

1 Introduction

When faced with the same situation, different humans behave differently. This is, of course, due to different backgrounds, education paths, and life experiences, but according to psychologists there is another important aspect: personality (Snyder, 1983; Parks and Guay, 2009).

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 (Vinciarelli and Mohammadi, 2014).

Such characteristics are formalised in *Trait Models*, and there are currently two of these models that are widely adopted: *Big Five* (John and Srivastava, 1999) and *Myers-Briggs Type Indicator* (*MBTI*) (Myers and Myers, 1995). The first examines five dimensions (OPENNESS TO EXPERIENCE, CONSCIENTIOUSNESS, EXTROVERSION, AGREEABLENESS and NEUROTICISM) and for each of them assigns a score in a range. The second one, instead, considers 16 fixed personality types, coming from the combination of the opposite poles of 4 main dimensions (EXTRAVERT-INTROVERT, INTUITIVE-SENSING, FEELING-THINKING, PERCEIVING-JUDGING). Examples of full personality types are therefore four letter labels such as ENTJ or ISFP.

The tests 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 on social media, has prompted the task of *Personality Detection*, i.e., the automatic prediction of personality from written texts (Youyou et al., 2015; Argamon et al., 2009; Litvinova et al., 2016; Whelan and Davies, 2006).

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 and psychological and social assessment.

As a contribution to personality detection in Italian, we present Personal-ITY, a new corpus of YouTube comments annotated with MBTI personality traits, and some preliminary experiments to highlight its characteristics and test its potential. The corpus is made available to the community[1].

2 Related Work

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), two datasets annotated with the *Big Five* traits have been released in 2013

[1] https://github.com/elisabassignana/Personal-ITY

Corpus	Model	# user	Avg.
PAN2015	Big Five	38	1258
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.

(Essays (Pennebaker and King, 2000) and myPersonality[2]) and two in 2014 (YouTube Personality Dataset (Biel and Gatica-Perez, 2013) and Mobile Phones interactions (Staiano et al., 2012)).

For the 2015 PAN Author Profiling Shared Task (Pardo et al., 2015), personality was added to gender and age in the profiling task, with tweets in English, Spanish, Italian and Dutch. These are also annotated according to the *Big Five* model.

Still in the Big Five landscape, Schwartz et al. (2013) collected a 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.

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 TWISTY (Verhoeven et al., 2016). The latter is a corpus of data collected from Twitter annotated with MBTI personality labels and gender for six languages (Dutch, German, French, Italian, Portuguese and Spanish) and a total of 18,168 authors. We are interested in the Italian portion of TWISTY.

Table 1 contains an overview of the available Italian corpora labelled with personality traits. We include our own, which is described in Section 3.

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 2 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).

Trait	WRB	MAJ	f-score
EI	65.54	77.88	77.78
NS	75.60	85.78	79.21
FT	50.31	53.95	52.13
PJ	50.19	53.05	47.01
Avg	60.41	67.67	64.06

Table 2: TWISTY scores from the original paper. Note that all results are reported as *micro*-average F-score.

3 Personal-ITY

First, we explain two major choices that we made in creating Personal-ITY, namely the source of the data and the trait model. Second, we describe in detail the procedure we followed to construct the corpus. Lastly, we provide a description of the resulting dataset.

Data 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.

Trait Model Our model of choice is the MBTI. The first benefit of this decision is that this model 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 3.1). 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.

Ethics Statement

Personality profiling must be carefully evaluated from an ethical point of view. In particular, often, personality detection involves ethical dilem-

[2] http://mypersonality.org

mas regarding appropriate utilization and interpretations of the prediction outcomes (Weiner and Greene, 2017). 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 were publicly available on the YouTube platform at the time of the collection. As we will explain in detail in this Section, the information collected are comments published under public videos on the YouTube platform by authors themselves. For a major 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 the real identifiers in the platform, allowing to trace the identity of the authors, are not released. Note also that the corpus was created for academic purposes and is not intended to be used for commercial deployment or applications.

3.1 Corpus Creation

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).

Users left comments to some 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?"]). We exploited such comments to create Personal-ITY with the following procedure.

First, we searched for as many Italian YouTube videos about MBTI as possible, ending up with a selection of ten with a conspicuous number of comments as the ones above[3].

Second, we retrieved all the comments to these videos using an AJAX request, and built a list of authors and their associated MBTI label. A label was associated to a user if they included an MBTI combination in one of their comments. Table 3 shows some examples of such associations. The association process is an approximation typical of DS approaches. To assess its validity, we manually checked 300 random comments to see whether the mention of an MBTI label was indeed referred to the author's own personality. We found that in 19 cases (6.3%) our method led to a wrong or unsure classification of the user's personality (e.g. *O tutti gli INTJ del mondo stanno commentando questo video oppure le statistiche sono sbagliate :-)*). We can assume that our dataset might therefore contain about 6-7% of noisy labels.

Comment	User - MBTI label
Io sono ENFJ!!!	User1 - ENFJ
Ho sempre saputo di essere connessa con Lady Gaga! ISFP!	User2 - ISFP

Table 3: Examples of automatic associations *user - MBTI personality type*.

Using the acquired list of authors, we meant to obtain as many comments as possible written by them. The YouTube API, however, does not allow to retrieve all comments by one user on the platform. In order to get around this problem we relied on video similarities, and tried to expand as much as possible our video collection. Therefore, as a third step, we retrieved the list of channels that feature our initial 10 videos, and then all of the videos within those channels.

Fourth, through a second AJAX request, we downloaded all comments appearing below all videos retrieved through the previous step.

Lastly, we filtered all comments retaining those written by authors included in our original list. This does not obviously cover all comments by a relevant user, but it provided us with additional data per author.

3.2 Final Corpus Statistics

For the final dataset, we decided to keep only the authors with a sufficient amount of data. More specifically, we retained only users with at least five comments, each at least five token long.

Personal-ITY includes $96,815$ comments by 1048 users, each annotated with an MBTI label. The average number of comments per user is 92

[3]Links to the 10 YouTube videos:
https://www.youtube.com/watch?v=VCo9RlDRpz0
https://www.youtube.com/watch?v=N4kC8iqUNyk
https://www.youtube.com/watch?v=Z8S8PgW8t2U
https://www.youtube.com/watch?v=wHZOG8k7nSw
https://www.youtube.com/watch?v=lO2z3_DINqs
https://www.youtube.com/watch?v=NaKPl_y1JXg
https://www.youtube.com/watch?v=814o4VBXlGY
https://www.youtube.com/watch?v=GK5J6PLj218
https://www.youtube.com/watch?v=9P95dkVLmps
https://www.youtube.com/watch?v=g0ZIFNgUmoE

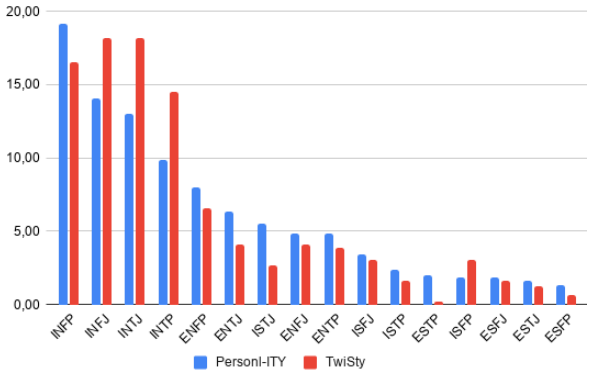


Figure 1: Distribution of the 16 personality types in the YouTube corpus and in the Italian section of TWISTY.

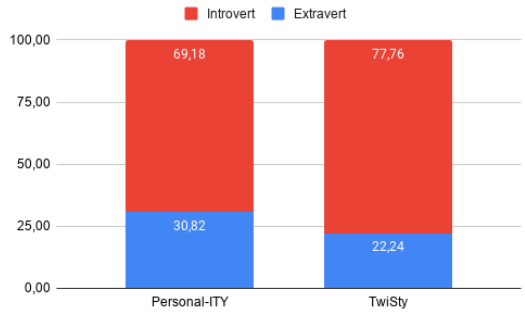
and each message has on average 115 tokens.

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. The 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. In Figure 1, we also see that 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 advantage that it reflects actual language use and allows large-scale analysis (Plank and Hovy, 2015).

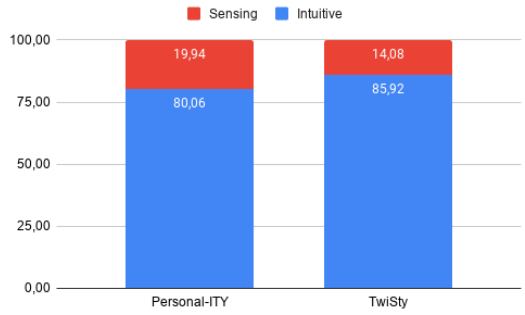
Figure 2 shows more in detail, trait by trait, the distribution of the opposite poles through the users in Personal-ITY and in TWISTY. As we might have expected, in line with what is observed in Figure 1, the two datasets present very similar trends. Such similarities between Personal-ITY and TWISTY are these similarities are a further confirmation of the reliability of the data we collected.

4 Preliminary Experiments

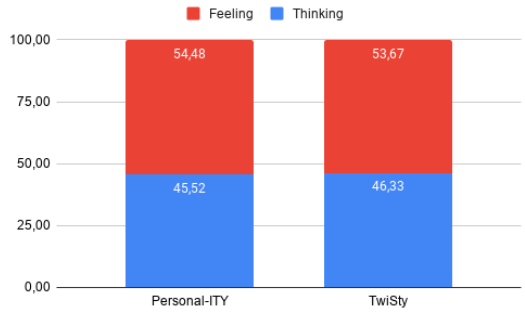
We ran a series of preliminary experiments on Personali-ITY which can also serve as a baseline for future work on this dataset. We pre-processed texts by replacing hashtags, urls, usernames and



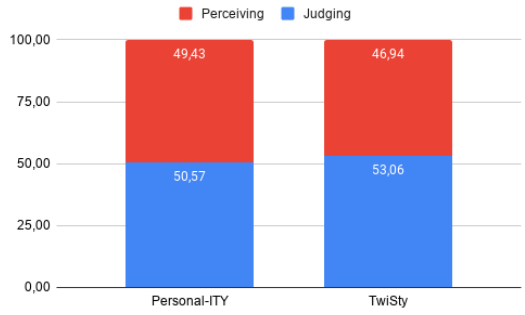
(a) *Extravert - Introvert*



(b) *Sensing - Intuitive*



(c) *Thinking - Feeling*



(d) *Judging - Perceiving*

Figure 2: Comparison of the distributions of the four MBTI traits between Personal-ITY and the Italian part of TWISTY.

emojis with four corresponding placeholders. We adopted the `sklearn` (Pedregosa et al., 2011) implementation of a linear SVM (LinearSVM), with standard parameters. We tested three types of features. 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. At a more stylistically level, we considered the use of emojis, hashtags, pronouns, punctuation and capitalisation. Lastly, we also experimented with embeddings-based representations, by using, on the one hand, YouTube-specific (Nieuwenhuis and Nissim, 2019) pre-trained models, on the other hand, more generic embeddings, such as the Italian version of GloVe (Pennington et al., 2014), which is trained on the Italian Wikipedia[4]. We looked for all the available embeddings of the words written by each author, and used the average as feature. If a word appeared more than once in the string of comments, we considered it multiple times in the final average.

We used 10-fold cross-validation, and assessed the models using macro f-score. Note that the original TWISTY paper uses micro f-score. Thus, for the sake of comparison, we include also micro-F in Table 5 for the MAJ baseline and our lexical n-gram model. Table 4 shows the results of our experiments with different feature types.[5] Overall, lexical features (n-grams) perform best. Combining different feature types did not lead to any improvement. Classification was performed with four separate binary classifiers (one per dimension), and with one single classifier predicting four classes, i.e, the whole MBTI labels at once. In the latter case, we observe that the results are quite high considering the increased difficulty of the task. Table 5 reports the scores of our models on TWISTY. As for Personal-ITY, best results were achieved using lexical features (tf-idf *n*-grams); stylistic features and embeddings are just above the baseline. Our model outperforms the one in (Verhoeven et al., 2016) for all traits (micro-F).

To test compatibility of resources and to assess model portability, we also ran cross-domain experiments on Personal-ITY and TWISTY. In the first setting, we tested the effect of merging the two datasets on the performance of models for personality detection, maintaining the 10-fold cross-validation setting and by using the model performing better on average for YouTube and Twitter data (a character *n*-grams model). Table 6 contains the result of such experiments[6]. Scores are almost always lower compared to the in-domain experiments (excepts for NS as regards Twitter scores reported in Table 5: 46.15 $\rightarrow$ 48.31), but quite increased compared to the majority baseline.

Trait	MAJ	Lex	Sty	Emb	FL
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 4: Results of the experiments on Personal-ITY. FL: prediction of the full MBTI label at once, with a character *n*-gram model.

	micro F		macro F			
Trait	MAJ	Lex	MAJ	Lex	Sty	Emb
EI	77.75	**79.18**	43.69	55.23	43.69	43.69
NS	85.92	**85.92**	46.15	46.15	46.15	46.15
FT	53.67	**55.31**	34.79	52.98	35.34	34.70
PJ	53.06	**54.08**	34.56	53.01	35.20	34.90
Avg	67.6	**68.62**	39.80	**51.84**	40.09	39.86

Table 5: Results of our experiments on TWISTY.

Trait	MAJ	Lex
EI	41.64	50.57
NS	44.93	48.31
FT	35.04	51.31
PJ	30.66	48.24
Avg	38.07	49.61

Table 6: Merging Personal-ITY with TWISTY.

In the second setting, instead, we divided both corpora in fixed training and test sets with a proportion of 80/20 and ran the models using lexical features, in order to run a cross-domain experiment. For direct comparison, we run the model in-domain again using this split. Results are shown

[4] `https://hlt.isti.cnr.it/wordembeddings`

[5] In Tables 4–5, we report the highest scores based on averages of the four traits. Considering the dimensions individually, better results can be obtained by using specific models.

[6] Prediction of the full label at once.

Train	Personal-ITY			TwiSty		
Test	IN	CROSS		IN	CROSS	
	Pers	MAJ	Twi	Twi	MAJ	Pers
EI	58.94	44.94	49.33	55.66	44.59	44.59
NS	52.88	47.87	47.31	47.87	45.31	45.31
FT	49.20	37.58	47.09	65.26	39.13	51.04
PJ	54.43	32.41	32.50	56.87	36.56	38.54
Avg	53.86	40.70	44.06	56.42	41.40	44.87

Table 7: Results of the cross-domain experiments. MAJ = baseline on the cross-domain testset.

in Table 7. Cross-domain scores are obtained with the best in-domain model.[7] They drop substantially compared to in-domain, but are always above the baseline.

5 Conclusions

The experiments show that there is no single best model for personality prediction, as the feature contribution depends on the dimension considered, and on the dataset. Lexical features perform best, but they tend to be strictly related to the context in which the model is trained and so to overfit.

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.

Personal-ITY is made available to further investigate the above and other issues related to personality detection in Italian. The corpus can lend itself to a psychological analysis of the linguistic cues for the MBTI personality traits. On this line, it is interesting to investigate the presence of evidences linking linguistic features with psychological theories about the four considered dimensions (Extravert-Introvert, iNtuitive-Sensing, Feeling-Thinking, Perceiving-Judging). First results in this direction are presented in (Bassignana et al., 2020).

Acknowledgments

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

[7] Better results can be obtained with other specific models.

References

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 YouTube Comments Corpus for Personality Profiling in Italian Social Media. In Viviana Patti, Malvina Nissim, and Barbara Plank, editors, *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, (PEOPLES@COLING 2020)*. Association for Computational Linguistics.

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.

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.

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.

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.

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.

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.

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.

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.