TwiSty: a Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling

Ben Verhoeven,† Walter Daelemans† and Barbara Plank*
†CLiPS Research Center, University of Antwerp, Belgium
*University of Groningen, The Netherlands
{ben.verhoeven,walter.daelemans}@uantwerpen.be,b.plank@rug.nl

Abstract

Personality profiling is the task of detecting personality traits of authors based on writing style. Several personality typologies exist, however, the Myers-Briggs Type Indicator (MBTI) is particularly popular in the non-scientific community, and many people use it to analyse their own personality and talk about the results online. Therefore, large amounts of self-assessed data on MBTI are readily available on social-media platforms such as Twitter. We present a novel corpus of tweets annotated with the MBTI personality type and gender of their author for six Western European languages (Dutch, German, French, Italian, Portuguese and Spanish). We outline the corpus creation and annotation, show statistics of the obtained data distributions and present first baselines on Myers-Briggs personality profiling and gender prediction for all six languages.

Keywords: computational stylometry, author profiling, multilingual corpus

1. Introduction

Personality prediction is one of the most difficult author profiling tasks in computational stylometry. It involves detecting personality traits on the basis of writing style. Several typologies of personality traits exist, but the two most well-known are Big Five (Goldberg, 1990) and Myers-Briggs Type Indicator (Briggs Myers and Myers, 2010).

Modeling author attributes such as personality plays an increasingly important role in numerous applications, from business intelligence to personalized translation (Mirkin et al., 2015). A large body of work exists on predicting individual author attributes from linguistic input (Rosenthal and McKeown, 2011; Nguyen et al., 2011; Eisenstein et al., 2011; Volkova et al., 2013; Alowibdi et al., 2013; Ciot et al., 2013; Volkova et al., 2015). Apart from demographic features, such as age or gender, there is also a growing interest in predicting psychological properties such as personality, attested by a growing literature and recent shared tasks on this topic (Celli et al., 2013; Celli et al., 2014; Rangel et al., 2015).

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

However, computational personality recognition is hampered by the availability of limited amounts of labeled data (Nowson and Gill, 2014). Many early existing data sets contain written essays of a certain topic, which are written in highly canonical language. Such controlled settings inhibit the expression of individual traits much more than spontaneous language. As such data is hard to obtain, only limited amounts were available.

With the availability of social media text, recent efforts shifted toward using such data (Schwartz et al., 2013a; Schwartz et al., 2013b; Park et al., 2015; Kosinski et al., 2015). For example, Kosinski et al. (2015) collected a large amount of social media data with Big Five (Kosinski et al., 2015) annotations through a tailored Facebook app. Another approach, suggested by Plank and Hovy (2015), is to use the large amounts of textual data voluntarily produced on social media (i.e., Twitter) together with self-assessed Myers-Briggs Type Indicators (Briggs Myers and Myers, 2010), abbreviated MBTI, to collect large amounts of labeled data. Myers-Briggs classifies users along four dimensions (Introvert–Extravert, iNtuitive–Sensing, Thinking–Feeling, Judging–Perceiving), amounting to 16 different types, e.g., INTJ, ESFP, etc. As such, Myers-Briggs personality types have the distinct advantage of being readily available in large quantities on social media, in particular Twitter, which is highly non-canonical and known for an almost unlimited vocabulary size (Eisenstein, 2013). As in most existing data collections, the labeling is based on the self-testing of the authors based on publicly available tests, and may contain noise if the questions of the test were not answered truthfully or if the test taken was not a good predictor of personality type.

Prior work focused almost exclusively on English (with a few exceptions, see Section 6), a well-represented language on Twitter. English is in fact the most frequent language on Twitter. Figure 1 shows the language distribution found in a sample of 65M tweets (randomly sampled over 2013). We see that 45% of all data is estimated to be English.[1] This ranking of languages is similar to what has been reported earlier (Baldwin et al., 2013) and it remains rather stable if we use a larger sample. The six languages in our new corpus are all among the 15 most frequent languages on Twitter.

Contributions We present TwiSty, a novel corpus developed to aid research in author profiling. It contains personality and gender annotations for a total of 18,168 authors spanning six languages (Dutch, German, French, Italian, Portuguese and Spanish). TwiSty is freely

[1] We estimate the language distribution by running `langid` (Lui and Baldwin, 2012).

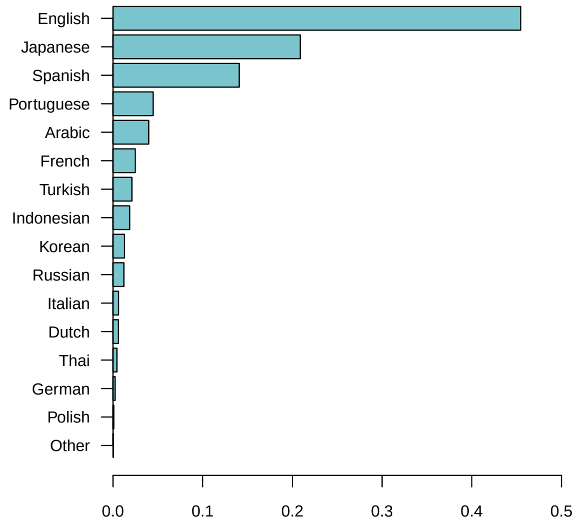


Figure 1: Distribution of languages (% of tweets) estimated from a 65M tweets sample (from 2013).

available at `http://www.clips.uantwerpen.be/datasets`. We also present experiments for personality and gender prediction for all of the languages.

2. Corpus creation

For each language, the corpus creation task is a two-step procedure. First, we mine for user profiles that self-report MBTI. These profiles are then manually checked and further annotated for gender. Once the set of users are identified, their data is downloaded and preprocessed. These two steps are described next.

Identifying users In order to collect the users, we query the Twitter interface for any of the 16 Myers-Briggs type indicators, plus a keyword that is typical for the language (see Table 1). We generally used a translation of the word *personality* and some very frequent pronouns or verb forms. All context words were verified as representative by a native speaker of the language. In this way, we retrieved a list of tweets potentially mentioning a personality type. These are then manually checked to determine that the tweet effectively describes the personality type of the author in the appropriate language. All users for whom such a tweet remained had their gender annotated. Gender decisions were based on the user's name, handle, description and profile picture.

Language	Context Words
German	*ich, bist, Persönlichkeit, dass*
Italian	*che, fatto, sono, personalità*
Dutch	*ik, jij, het, persoonlijkheid*
French	*suis, c'est, personnalité*
Portuguese	*sou, personalidade*
Spanish	*soy, tengo, personalidad*

Table 1: Specific context words for each language

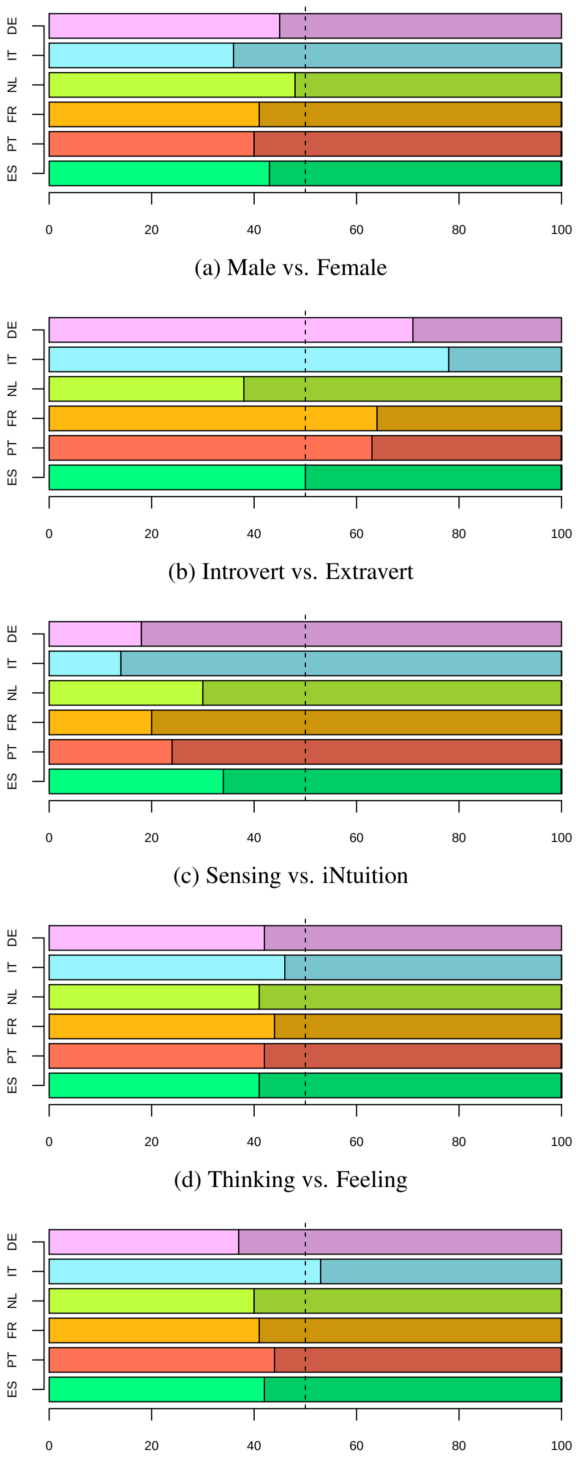
Corpus collection Once the set of users is identified, we fetch their tweets using the Twitter API. For each user, we retrieve all recent tweets. The average number of tweets per user is around 2,000. The statistics for each language corpus are shown in Table 2. We observe almost the same order of languages when comparing the corpus sizes and the estimated frequency of the language on Twitter (see Figure 1). Only Italian and Dutch have switched places.
Because of multilinguality and code switching, especially on Twitter, we perform language identification on each tweet. To have the highest accuracy possible we follow an approach similar to Lui and Baldwin (2014) and use three language identification tools and majority-voting. As language identifiers we used `ldig` (Nakatani, 2012), `langid` (Lui and Baldwin, 2012) and `langdetect` (Nakatani, 2010). After language detection, we found that on average 70-75% of the tweets in each subcorpus were confirmed as being in the language of the subcorpus. Our corpus release contains both the tweet ids with confirmed language, as well as the other additional mined tweets (for potential future research). Further details on the corpus creation as well as more statistics can be found in a technical report (Verhoeven et al., 2016).

	# Authors	# Tweets	Avg.	% in-lang
German	411	952,549	2,318	74.9
Italian	490	932,785	1,904	70.6
Dutch	1,000	2,083,484	2,083	74.0
French	1,405	2,786,589	1,983	71.6
Portuguese	4,090	8,833,132	2,160	71.9
Spanish	10,772	18,547,622	1,722	72.8

Table 2: Tweet counts per language. Avg.: average tweets per author, % in-lang: % of tweets identified as in language.

3. Analysis

When looking at some of the corpus statistics, a number of observations can be made. Figure 2 shows the proportions of each trait (gender and the personality dimensions) for each language. We observe that there is a general trend over all the languages in the corpus towards female, INTROVERT, INTUITIVE, FEELING, PERCEIVING users. Interestingly, most languages have similar distributions for all the traits, which suggests that the Twitter population (that would self-report their personality type) is quite similar over different languages. This holds especially for the INTUITIVE and SENSING traits.
INTROVERT is dominant over EXTRAVERT for four languages (DE, IT, FR and PT). This reflects the prior finding for English that online communication is easier and more accessible for introverts (Goby, 2006; Plank and Hovy, 2015). Also, we find the same advertising/sensationalism bias as shown in the study for English. Infrequent MBTI types in the general population, i.e., INFJ, INFP, INTJ) are among the most frequent types in our Twitter sample. In fact, as also observed in the case for English, people like to tweet about rare events and compare themselves to famous people, as shown in the following Italian and French example:



(a) Male vs. Female

(b) Introvert vs. Extravert

(c) Sensing vs. iNtuition

(d) Thinking vs. Feeling

(e) Judging vs. Perceiving

Figure 2: Proportions of the traits for all languages.

(1) Ma ho fatto il test delle 16 personalità e sono INTP e quindi logico Siamo solo il 3% della popolazione Lo è anche Einstein e scienziati
But I took the 16 personality test and I am INTP and that is logical We are only 3% of the population. Also Einstein and other scientists are INTP.

(2) Selon le test MBTI, je serais INFJ. Ils représentent environ 1% de la population. Voilà pourquoi je suis si incomprise, tout s'explique !
According to the MBTI test, I am INFJ. They represent around 1% of the population. And that's why I'm so misunderstood, everything is explained!

Although in collecting the Dutch corpus, we also observed a high number of tweets mentioning the rarity of their profile, the Dutch corpus statistics go against the INTROVERT trend. The proportion of EXTRAVERT is a lot higher. It is not yet clear to us, what the reason might be for this EXTRAVERT shift, however, the data shows that by far most Dutch users are EN[FT]P.
Italian seems to be the only language with more JUDGING users than PERCEIVING users, but given that the dataset is rather small and the difference not so big, we do not believe this to be an important difference.
Despite somewhat different biases for the languages, collecting linguistic data in this way has the advantage that it reflects actual language use. In the next section, we assess to what extent personality and gender are predictable from only the linguistic input.

4. Experiments

This section describes our experimental setup, where we train models to predict gender and each of the four Myers-Briggs personality dimensions. We use a concatenation of 200 (language-confirmed) tweets per user. For some authors, we don't have this many tweets so we leave them out. The number of instances for each language is mentioned under the language code in Table 3.

Model We use a LinearSVC as implemented in `sklearn` with standard parameters, as a grid search on the C parameter did not improve results. We also tested LogisticRegresssion, which gave comparable results.

Preprocessing The tweets are preprocessed in two steps: (1) URLs, hashtags and usernames are normalized to URL, hashtag or username placeholders, respectively, (2) tokenization with `happierfuntokenizer`[2].

Features We use binary features for word n-grams (unigrams and bigrams) and character n-grams (trigrams and tetragrams).

Evaluation Evaluation is performed using 10-fold cross-validation. We compare our model to a weighted random baseline (WRB, the sum of the squared class proportions) and majority baseline (MAJ). We report precision, recall and f-score. Precision and recall are the weighted average over the two classes. F-score is the harmonic mean of precision and recall. Note that prior work reported on accuracy only (Plank and Hovy, 2015) .

5. Results

Table 3 shows our results for gender and personality prediction on all six languages. For gender prediction, the model outperforms both random and majority baseline considerably across all languages. The largest improvement is observed on Dutch (NL), where our model reaches an F-score of 82.61 (vs. 51.41 for the majority baseline).
Personality prediction is a more difficult task, yet our study shows promising results. All our personality experiments

[2] `http://wwbp.org/`

Lang	Task	WRB	MAJ	P	R	F
	I–E	60.22	72.61	71.43	73.1	72.27
	S–N	71.03	82.43	67.95	82.43	74.49
DE	T–F	51.16	57.62	58.38	59.69	**59.03**
387	J–P	53.68	63.57	60.27	63.82	61.99
	Gender	50.28	53.75	77.72	77.52	**77.62**
	I–E	65.54	77.88	76.42	79.23	77.78
	S–N	75.60	85.78	73.58	85.78	79.21
IT	T–F	50.31	53.95	51.66	52.60	52.13
443	J–P	50.19	53.05	46.63	47.40	*47.01*
	Gender	54.78	65.46	73.90	72.69	**73.29**
	I–E	53.02	62.28	61.82	64.02	**62.90**
	S–N	57.66	69.57	69.39	71.63	**70.49**
NL	T–F	51.47	58.59	59.26	60.65	**59.95**
920	J–P	52.00	60.00	56.50	59.57	57.99
	Gender	50.04	51.41	82.62	82.61	**82.61**
	I–E	54.77	65.44	65.35	67.68	**66.49**
	S–N	68.00	80.00	77.60	80.24	78.90
FR	T–F	50.65	55.68	57.88	58.56	**58.22**
1,250	J–P	52.13	60.32	55.06	58.64	56.79
	Gender	51.84	59.60	83.77	83.84	**83.80**
	I–E	53.36	62.97	66.06	67.34	**66.69**
	S–N	63.60	76.08	71.02	75.98	73.42
PT	T–F	51.27	57.98	61.23	62.01	**61.62**
3,867	J–P	50.87	56.61	56.10	56.97	56.53
	Gender	52.15	60.36	87.54	87.56	**87.55**
	I–E	50.00	50.49	61.09	61.09	**61.09**
	S–N	55.42	66.47	60.23	62.91	61.54
ES	T–F	51.63	59.04	59.35	60.12	**59.73**
9,445	J–P	51.53	58.75	55.60	56.56	56.08
	Gender	51.00	57.06	87.61	87.63	**87.62**

Table 3: Results for gender and personality classification for six languages. The result in italics is the only one not reaching any baseline, all the others reach at least the weighted random baseline (WRB). Results in bold also outperform the majority baseline. The number of instances is indicated below the language code.

(except for one, the P-J dimension for Italian) outperform the weighted random baseline, which should be regarded as the main point of comparison. For four languages (Dutch, French, Portuguese, Spanish) our model even outperforms the higher majority baseline consistently for two dimensions, namely INTROVERT–EXTRAVERT and THINKING–FEELING. The other two dimensions are more difficult to predict and our model does not reach majority baseline (with only one exception, S–N for Dutch). This has been observed earlier on English by Plank and Hovy (2015). They found the exact same dimensions where no improvement was achieved, and a similar trend was described by Luyckx and Daelemans (2008) for the last dimension (J–P). This suggests that INTROVERT–EXTRAVERT as well as THINKING–FEELING are predictable from linguistic input alone, while this is much less the case for the other two dimensions. However, for languages for which we have fewer than 500 authors, namely Italian and German, the model usually does not outperform the majority baseline.

6. Discussion and related work

Because different dataset types and sizes, collection methods, evaluation metrics, and preprocessing methods make direct comparisons impossible, we conclude from our gender identification results that they are comparable to or better than the best published results on gender identification from Twitter for the different languages in our corpus. See Burger et al. (2011) for another comparative multilingual study on gender identification from twitter data, but using an approach that is difficult to compare to ours (learning all languages with one classifier).

Predicting Myers-Briggs type indicators from linguistic input has been studied in the seminal paper of Luyckx and Daelemans (2008). They created a corpus for Dutch, consisting of 145 student essays about a documentary on artificial life. Recently, the CSI (CLiPS Stylometry Investigation) corpus was introduced, which includes Dutch reviews as well as essays and annotations for both Big Five and MBTI annotations (Verhoeven and Daelemans, 2014). In contrast, we here focus on social media data, in particular Twitter, and self-assessed (and self-reported) MBTI personality types. In many prior studies, participants were asked to participate in a personality test and produce essay(s).

Collecting personality data from social media has been done before (Schwartz et al., 2013a; Schwartz et al., 2013b; Park et al., 2015). For instance, the myPersonality dataset (Kosinski et al., 2015) contains personality types and messages from 75,000 users collected through a Facebook app. Earlier work using social media data is mostly smaller scale, e.g., the YouTube video blog corpus (Biel and Gatica-Perez, 2013) used in the 2014 shared task (Celli et al., 2014) contains 404 users, or a Facebook dataset of 250 users (Celli et al., 2013), or a Twitter datasets of 102 users (Farnadi et al., 2016). Our approach is simpler, requires no tailored app, and can be used to collect large amounts of data quickly.

Finally, most prior research efforts on personality profiling focus on the Big Five (Schwartz et al., 2013b; Farnadi et al., 2016), which is not completely comparable to the MBTI taxonomy. This is unfortunate for reasons of comparison. However, there are strong correlations between the traits used in both systems especially in the INTROVERT–EXTRAVERT and JUDGING–PERCEIVING dimensions.

7. Conclusions and Outlook

We presented TWISTY, a novel corpus for personality and gender profiling for Twitter. It covers 6 languages (Dutch, French, German, Italian, Spanish and Portuguese) and contains both gender annotation and Myers-Briggs personality type indicators. The corpus was constructed following Plank and Hovy (2015) for English.

An exploratory empirical investigation on our new corpus shows that gender prediction works well on it, and that personality trait identification from text is a more difficult problem. Our results confirm prior findings in that certain personality distinctions, namely INTROVERT–EXTRAVERT (I–E) and THINKING–FEELING (T–F), can be predicted from social media data with success The other two MBTI dimensions are harder to predict from the linguistic signal alone.

8. Acknowledgements

The first author is supported by a PhD scholarship from the FWO Research Foundation - Flanders. Part of this research was carried out in the framework of the AMiCA (IWT SBO-project 120007) project, funded by the government agency for Innovation by Science and Technology (IWT).

9. Bibliographical References

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

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How noisy social media text, how diffrnt social media sources? In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan.

Biel, J.-I. and Gatica-Perez, D. (2013). The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55.

Briggs Myers, I. and Myers, P. (2010). *Gifts differing: Understanding personality type*. Nicholas Brealey Publishing.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1301–1309, Edinburgh, Scotland, UK. ACL.

Celli, F., Pianesi, F., Stillwell, D., and Kosinski, M. (2013). Workshop on computational personality recognition (shared task). In *Proceedings of the Workshop on Computational Personality Recognition*, Boston, MA, USA.

Celli, F., Lepri, B., Biel, J.-I., Gatica-Perez, D., Riccardi, G., and Pianesi, F. (2014). The workshop on computational personality recognition 2014. In *Proceedings of the ACM International Conference on Multimedia*, pages 1245–1246, Orlando, FL, USA. ACM.

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

Eisenstein, J., Smith, N., and Xing, E. (2011). Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.-F., and Cock, M. D. (2016). Computational personality recognition in social media. *User Modeling and User-Adapted Interaction*.

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

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

Kosinski, M., Matz, S., Gosling, S., Popov, V., and Stillwell, D. (2015). Facebook as a social science research tool: Opportunities, challenges, ethical considerations and practical guidelines. *American Psychologist*, 70(6):543–556.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30, Jeju, Korea. ACL.

Lui, M. and Baldwin, T. (2014). Accurate language identification of Twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden. ACL.

Luyckx, K. and Daelemans, W. (2008). Personae: a corpus for author and personality prediction from text. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Mirkin, S., Nowson, S., Brun, C., and Perez, J. (2015). Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, Lisbon, Portugal. ACL.

Nakatani, S. (2010). Language detection library for java. https://github.com/shuyo/language-detection.

Nakatani, S. (2012). Short text language detection with infinity-gram. https://shuyo.wordpress.com/2012/05/17/short-text-language-detection-with-infinity-gram/.

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

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

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., and Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6).

Plank, B. and Hovy, D. (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)*, Lisbon, Portugal. ACL.

Preoţiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A., and Ungar, L. H. (2015). The role of personality, age and gender in tweeting about mental illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psy-*

chology: From Linguistic Signal to Clinical Reality. ACL.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., and Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015. In *CLEF 2015 Working Notes*, Toulouse, France. CEUR.

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

Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Kosinski, M., Stillwell, D., Seligman, M. E., and Ungar, L. H. (2013a). Toward personality insights from language exploration in social media. In *AAAI Spring Symposium: Analyzing Microtext*. AAAI.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013b). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9).

Verhoeven, B. and Daelemans, W. (2014). CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Verhoeven, B., Daelemans, W., and Plank, B. (2016). Creating TwiSty: Corpus development and statistics. *CLiPS Technical Report Series (CTRS)*, 6.

Volkova, S., Wilson, T., and Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing (EMNLP)*. ACL.

Volkova, S., Bachrach, Y., Armstrong, M., and Sharma, V. (2015). Inferring latent user properties from texts published in social media (demo). In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI.