Multilingual Personality Profiling on Twitter

Ben Verhoeven
CLiPS Research Center, University of Antwerp, Belgium
ben.verhoeven@uantwerpen.be

Barbara Plank
University of Groningen, The Netherlands
b.plank@rug.nl

Walter Daelemans
CLiPS Research Center, University of Antwerp, Belgium
walter.daelemans@uantwerpen.be

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 MBTI (Briggs Myers and Myers 2010). Modeling author attributes, such as personality, plays an important role in numerous applications from business intelligence to personalized translation (Mirkin et al. 2015).

In this study, we build on recent work that explores the use of social media as a resource for large-scale personality prediction (Plank and Hovy 2015). The idea of using social media for large-scale personality prediction was not new (cf. the myPersonality project (Kosinski et al. 2015)), but Plank and Hovy (2015) mine whatever is publicly available on Twitter which requires far less effort than setting up a campaign on Facebook to have people do a personality test. However, all these previous efforts were solely for English.

TwiSty is a new corpus developed to aid research in author profiling (Verhoeven, Daelemans and Plank 2016b). The corpus contains Twitter profiles with associated gender and self-reported MBTI personality profile for six Western European languages: Italian, Dutch, German, French, Portuguese and Spanish. These languages are all in the top 20 of the most frequent languages on Twitter (based on the language distribution estimated through language identification on a Twitter sample of 65 million tokens).

Corpus creation

TwiSty is based on the idea that many people post information about their personality online. We look for people that report the outcome of their MBTI personality test (Briggs Myers and Myers 2010) on Twitter. We follow Plank and Hovy (2015) in their choice to use the MBTI typology of personality, because this test is a lot more popular in the general public and therefore more present online than Big Five. We

gather[1] the data by searching Twitter for combinations of the MBTI types and some frequent language-specific words, e.g. *ENFJ tengo* (Spanish) or *ISTP jij* (Dutch). The MBTI types are specific combinations of four letters that refer to the person's personality on four axes: Extraversion-Introversion, iNtuition-Sensing, Thinking-Feeling, Judging-Perceiving (Briggs Myers and Myers 2010). The downloaded tweets are then manually checked to make sure the personality type in the tweet describes the author. We further manually enrich the corpus with gender information. This yields a list of profiles with associated self-assessed personality profile and gender information for each language.

For each of these profiles, we download all the most recent tweets. The average number of tweets per user is around 2,000. Because not everyone speaks the same language all the time, especially not on Twitter, we perform language identification on each tweet. We used a majority-voting approach of three implementations of such algorithms: `ldig` (Nakatani 2012), `langid` (Lui and Baldwin 2012) and `langdetect` (Nakatani 2010). We found that 70-75% of tweets in each subcorpus were confirmed as being in that language. Our corpus release[2] contains both the tweet ids with confirmed language, as well as the other ones we mined (for potential future research).

Table 1 contains the number of authors, the total number of tweets for each subcorpus (i.e. for each language), the average number of tweets per author, as well as the percentage of tweets for which the language was confirmed. For English, we refer to Plank and Hovy (2015) who gathered MBTI profiles of 1,500 authors with 2 million tweets.

	# Authors	# Tweets	Avg./Author	% Confirmed
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	1,417	2,792,472	1,982	71.4
Portuguese	4,090	8,833,132	2,160	71.9
Spanish	10,772	18,547,622	1,722	72.8

Table 1: Tweet counts per language.

This novel corpus allows many interesting lines of research. Personality prediction is one of them and it is in dire need of multilingual attention. In fact, most prior studies are monolingual and are mostly restricted to English, with a few exceptions, e.g. Dutch (Verhoeven and Daelemans 2014; Luyckx and Daelemans 2008), Modern Greek (Komianos et al. 2012) and Italian (Celli 2012). We are only aware of one earlier research that studied personality prediction on different languages (cfr. Rangel et al. 2015) but it suffers immensely from data scarcity.

Experiments

We have performed first experiments on this data by training models to predict gender and each of the four Myers-Briggs personality dimensions. We use a concatenation of 200 (language-confirmed) tweets per user. We use the state-of-the-art methodology here with `sklearn`'s LinearSVC as classification algorithm in a 10-fold cross-validation setup with character and word n-grams as features. The results of these experiments can be found in Table 2.

[1] Details on the corpus creation can be found in our technical report (Verhoeven, Daelemans and Plank 2016a).
[2] Freely available at: http://www.clips.uantwerpen.be/datasets/twisty-corpus

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 2: 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 (MAJ). The number of instances is indicated below the language code.

For gender prediction, the model outperforms both random and majority baseline considerably across all languages. Personality prediction is a more difficult task, yet our study shows promising results. All our personality experiments (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.

References

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

Celli, F. (2012) 'Unsupervised personality recognition from social networking sites', *ICDS 2012: The Sixth International Conference on Digital Society*, pp. 59–62.

Goldberg, L. R. (1990) 'An Alternative "Description of Personality": the Big-Five factor structure.', *Journal of personality and social psychology* 59.6, p. 1216.

Komianos, V. et al. (2012) 'Predicting Personality Traits from Spontaneous Modern Greek Text: Overcoming the Barriers', *Artificial Intelligence Applications and Innovations*, pp. 530–539.

Kosinski, M. et al. (2015) 'Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines', *American Psychologist* 70.6, pp. 543–556.

Lui, M. and Baldwin, T. (2012) 'langid.py: An off-the-shelf language identification tool', *Proceedings of the ACL 2012 system demonstrations*, Jeju, Korea: ACL, pp. 25–30.

Luyckx, K. and Daelemans, W. (2008) 'Personae: a Corpus for Author and Personality Prediction from Text.', *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Mirkin, S. et al. (2015) 'Motivating Personality-aware Machine Translation', *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.

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

Plank, B. and Hovy, D. (2015) 'Personality Traits on Twitter–or–How to Get 1,500 Personality Tests in a Week', *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Lisbon, Portugal: ACL.

Rangel, F. et al. (2015) 'Overview of the 3rd Author Profiling Task at PAN 2015', *CLEF 2015 Working Notes*, Toulouse, France: CEUR.

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', *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Verhoeven, B., Daelemans, W. and Plank, B. (2016a) 'Creating TwiSty: Corpus Development and Statistics', *CLiPS Technical Report Series (CTRS)* 6.

— (2016b) 'TwiSty: a multilingual Twitter Stylometry corpus for gender and personality profiling', *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia: ELRA.