

# Celebrity Profiling

Matti Wiegmann<sup>1,2</sup> Benno Stein<sup>1</sup> Martin Potthast<sup>3</sup>

<sup>1</sup>Bauhaus-Universität Weimar

<sup>2</sup>German Aerospace Center

<sup>3</sup>Leipzig University

<first>.<last>@[uni-weimar|dlr|uni-leipzig].de

## Abstract

Celebrities are among the most prolific users of social media, promoting their personas and rallying followers. This activity is closely tied to genuine writing samples, which makes them worthy research subjects in many respects, not least profiling. With this paper we introduce the Webis Celebrity Corpus 2019. For its construction the Twitter feeds of 71,706 verified accounts have been carefully linked with their respective Wikidata items, crawling both. After cleansing, the resulting profiles contain an average of 29,968 words per profile and up to 239 pieces of personal information. A cross-evaluation that checked the correct association of Twitter account and Wikidata item revealed an error rate of only 0.6%, rendering the profiles highly reliable. Our corpus comprises a wide cross-section of local and global celebrities, forming a unique combination of scale, profile comprehensiveness, and label reliability. We further establish the state of the art's profiling performance by evaluating the winning approaches submitted to the PAN gender prediction tasks in a transfer learning experiment. They are only outperformed by our own deep learning approach, which we also use to exemplify celebrity occupation prediction for the first time.

## 1 Introduction

Author profiling is about predicting personal traits of individual authors based on their writing style. Frequently studied traits are demographics such as gender, age, native language or dialect, and even personality. Applications of author profiling include marketing, social science, risk assessment, and forensics. Given the high expectations that are implied by these and similar applications, the creation of a valid automatic profiler for a given trait, let alone many, depends on the availability of carefully constructed corpora. Corpus construction for

author profiling has always been difficult for lack of large-scale distant supervision sources that provide for genuine pieces of writing from many different authors alongside personal information. In part, the aforementioned selection of demographics that are frequently studied reflects the availability of corresponding ground truth. In this regard, one source of ground truth, available in large quantities, high diversity of traits, and near-perfect label reliability, has been overlooked: celebrities.

The contributions of our research are threefold:<sup>1</sup> First, in Section 2, we survey the state of the art in constructing author profiling corpora for the first time, compiling a taxonomy of construction strategies applied. Second, in Section 3, we report on the construction of the first large-scale corpus of celebrity profiles, describing our acquisition approach based on a reliable matching of Twitter accounts to Wikidata items. Third, in Section 4, we carry out a prediction experiment on the most widely studied trait, gender, comparing the performance of our own deep learning approach with that of the four best-performing ones submitted to the recent PAN author profiling competitions from 2015 to 2018. Moreover, we exemplify the prediction of celebrity occupations.

## 2 Related Work

We analyzed 29 publications on author profiling the authors of which explicitly describe their data acquisition and corpus construction strategies. The strategies have been reviewed, abstracted, and mapped into a taxonomy, which in turn enabled us to identify specific quality criteria.

Table 1 overviews these publications and reports key figures, personal traits, and the underlying acquisition strategy. Note that a large part of this research builds upon the pioneering works done

<sup>1</sup>Code and corpus: <https://github.com/webis-de/ACL-19>

Dataset	Genre	Lang.	Authors	Words	Personal Traits	Label Acquisition Strategy
Mikros (2013)	Blogs	1	100	20,323	Gender	AIS
Nguyen et al. (2011)	Blogs	1	1,997	27,303	Age	AIS+U
Rosenthal and McKeown (2011)	Blogs	1	24,500	(?)	Age	AIS
Schler et al. (2006)	Blogs	1	37,478	7,885	Gender	AIS
PAN13 (2013)	Blogs	2	346,100	632	Age, Gender	AIS
Wang et al. (2016)	Sina Weibo	1	742,323	(?)	Age, Education, Gender, Relationship	AIS
Burger et al. (2011)	Tweets	12+	183,729	283*	Gender	AIU
MEX-A3T (2018)	Tweets	1	5,000	17,195*	Education, Residence	AIU
Gjurkovic and Snajder (2018)	Comments	1	23,503	24,861	Personality (MBTI)	AIU
Plank and Hovy (2015)	Tweets	1	1,500	12,880	Gender, Personality (MBTI)	AIU
Preotiuc-Pietro et al. (2015)	Tweets	1	5,191	26,415*	Occupation (SOC)	AIU
Ramos et al. (2018)	Facebook	1	1,019	2,178	Age, Education, Gender, Personality (Big Five), Religion	AIU
PAN17 (2017a)	Twitter	4	19,000	1,195	Dialect, Gender	AIU
Twisty (2016)	Twitter	6	18,168	25,400	Gender, Personality (MBTI)	AIU
Preotiuc-Pietro et al. (2017) - D2	Tweets	1	13,651	23,717*	Politics	AIU
TAT en (2007a)	Emails	1	1,033	3,259	Age, Gender, Education, Native lang., Personality (Big Five), Residence	ARS
TAT ar (2007b)	Emails	1	1,033	2,085	Age, Education, Gender, Personality (MBTI)	ARS
Fatima et al. (2017)	Facebook	4	479	2,156	Age, Birthplace, Gender, Education, Extroversion, Nat. lang., Occupation	ARS
Litvinova et al. (2017)	Essays	1	500	145	Age, Education, Gender, Personality	ARS
Preotiuc-Pietro and Ungar (2018)	Tweets	1	4,098	16,785*	Age, Education, Gender, Income, Race	ARS
PAN15 (2015)	Tweets	4	1,070	1,205	Age, Gender, Personality (Big Five)	ARS
Tighe and Cheng (2018)	Tweets	1	250	31,011*	Personality (Big Five)	ARS
Clips CSI (2014)	Essays	1	749	976	Age, Birthplace, Gender, Personality (Big Five)	ARS
Preotiuc-Pietro et al. (2017) - D1	Tweets	1	3,938	15,587*	Age, Gender, Politics	ARS
Schwartz et al. (2013)	Facebook	1	136,000	4,129	Age, Gender, Personality (NEO-PI-R)	ARS
Ciot et al. (2013)	Tweets	4	8,618	12,700*	Gender	ORS
Emmery et al. (2017)	Tweets	1	6,610	31,750*	Gender	ORS
Volkova and Bachrach (2015)	Tweets	1	5,000	2,540	Age, Children, Education, Gender, Income, Intelligence, Optimism, Political alignment, Ethnicity, Religion, Relationship, Satisfaction	ORS
Kapociute-Dzikiene et al. (2015)	Essays	1	186	286	Age, Gender	OIS
Bergsma et al. (2012)	Papers	1	4,500	(?)	Gender, Native language	OIS
<b>Our work</b>	Tweets	37	71,706	29,968	up to 239	OIS

Table 1: Survey of author profiling corpora. A \* indicates an estimation based on an average of 12.7 words per tweet from the reported number of tweets and a ? unavailable information. Row groups reflect acquisition strategy.

by Pennebaker et al. (2003), Koppel et al. (2002), Schler et al. (2006), and Argamon et al. (2009); recent works add novel traits, trait relations, multilingualism, and microblogs. The largest annual shared task on author profiling is part of the PAN competition (Rangel Pardo et al., 2013, 2014, 2015, 2016, 2017b, 2018). Profiling research related to aspects such as behavioral traits (Kumar et al., 2018), medical conditions (Choudhury et al., 2013), or native language identification (NLI) have been excluded from our survey, since these have developed into subfields of their own right.

Three criteria describe the quality of the surveyed resources: the representativeness of the targeted population, the comprehensiveness in terms of author, text, and label size, and the reliability of label attributions. Table 2 shows our taxonomy of label acquisition strategies for reliability and comprehensiveness evaluation: labels provided by the author or by others (A/O), labels provided independently or on request (I/R), and labels re-

	Independent		Requested
	Structured	Unstructured	Structured
<b>Author</b>	(AIS) Profile forms	(AIU) Posts, Comments	(ARS) Questionnaires
<b>Others</b>	(OIS) Wikidata	(OIU) News, Mentions	(ORS) Crowdsourcing


Table 2: Taxonomy of label acquisition strategies with common example applications.

trieved in structured or unstructured form (S/U). The six resulting strategies, disregarding R-U combinations as inapplicable, describe the general strategy and hint possible issues: (1) subjectivity or misunderstandings by experts, volunteer annotators, or crowdsourcing workers versus deception and self-serving bias by author-self-reported labels, (2) self-selection bias and per-author cost in requested labels versus few and stale trait choices in independent reporting, and (3) imprecision, incompleteness, and misunderstandings in unstructured versus restricted choices in structured labeling.

### 3 The Webis Celebrity Corpus

This section introduces the Webis Celebrity Corpus 2019, detailing how we identified celebrities at scale, compiled a large corpus of their writing, and linked it with Wikidata to obtain personal profiles. A corpus analysis and validation follows.

#### 3.1 Who is a Celebrity?

To operationalize the term “celebrity”, we say that a person has a celebrity-like status, be it locally or globally, if he or she possesses a *verified* Twitter account, and at the same time, is deemed *notable* enough to be the subject of a Wikipedia article and a Wikidata item. Importantly, Twitter verifies “that an account of public interest is authentic” (Twitter, 2018), awarding a blue checkmark badge: . Notability at Wikipedia pertains to people who are “worthy of notice,” “remarkable,” or “famous or popular” (Wikipedia, 2018). While verified accounts also include organizations, and while most notable people at Wikipedia/Wikidata are not considered celebrities, it is their intersection which provides for a good approximation. To collect celebrity profiles at scale, we join these sources of information.

#### 3.2 Corpus Construction

We crawled all 297,878 verified Twitter accounts,<sup>2</sup> and linked them with Wikidata items. This is a non-trivial task: a Twitter account name and its corresponding Wikidata item need not have an exact string match, and there may be false matches. Table 3a shows the six candidate names we obtained from the unique, static Twitter “@”-names and the free-form display names.

Table 3b shows the linking results. Accounts were marked as *human* or *not human* based on Wikidata’s `instance of` property. In the sequence of name candidates I-VI, a *human* match was kept, even if successive candidates matched non-human items. If items differed between languages for the same candidate, matches were marked *ambiguous*. Matches containing one of the eight death-related Wikidata properties and a date of death before Twitter’s launch in March 2006 were marked *memorial*. All mismatches identified during our subsequent corpus validation were marked as *error*. After excluding matches with private timelines, 71,706 valid account-item matches remained.

<sup>2</sup>Official list: <https://twitter.com/verified>, retrieved May 2018

#### 3.3 Corpus Validation

A large ground truth for evaluating our Twitter-Wikidata matches is provided by Wikidata itself: 89,451 items about humans include a Twitter username; 28,454 of these usernames intersect with the 297,878 verified Twitter accounts we crawled. Comparing these 28,454 true matches with those obtained by our matching heuristic, we distinguish three cases: (1) 20,579 are linked correctly, (2) 124 are linked incorrectly (0.6% error rate), and (3) 7,751 are not linked (27.7% miss rate). Thus, our heuristic achieves a very high precision of 0.994 at a reasonably high recall of 0.723.

Table 3b (bottom row group) breaks down the number of matches by type and name candidate. The most successful name candidate is I, yielding 92% of all matches, but only half the erroneous ones. Name candidates II, III, and VI contribute negligibly, while candidates IV and V provide only for 5% of the matches combined, but 45% of all errors. At an overall error rate of 0.6%, though, candidates IV and V produced 3,416 correct and only 56 incorrect matches, rendering them still viable.

#### 3.4 Corpus Analysis

The corpus we created contains 29,968 words on average per author and 1,523 different Wikidata properties, of which 239 are personal traits relevant for profiling. Table 4 shows a selection of those traits, the most common value and for how many celebrities they are annotated. The remaining properties split into 1,224 external references (i.e., links to other sites) and 60 miscellaneous properties (mostly internal references and multimedia data). Of the 239 traits, 45 are attributed to more than 1,000, and 5 to more than 55,000 users simultaneously. The extracted Wikidata properties are highly specific and frequently feature over 100 different values per property within our corpus, although most are Zipf-distributed and can easily be aggregated or reduced to smaller dimensions, as we will demonstrate with occupation in Section 4. It should be noted that labels, such as ethnicity, religion, and native language, are present mostly for minorities rather than the majority.

We collected an average 2,181 tweets per celebrity and 156,411,899 tweets in total ( $\approx$  3 billion words), covering 98.05% of all their tweets.<sup>3</sup> Of all collected tweets, 29.3% are retweets and 20.9%

<sup>3</sup>Though Twitter allows for retrieving only the 3,200 most recent tweets per account, its total number of tweets is given.

(a)		(b)					(c)			
Name candidate generation rule		Celebrity	Error	Memorial	Not hum.	Ambig.	Dataset	Authors		
		all						Training	Test	
I	only alphanumeric characters of the display name									
II	reference name split at capitalization	I	91.8%	50.0%	70.4%	77.6%	82.6%	PAN15 (2015)	152	142
III	reference name split at display name	II	2.8%	3.2%	2.6%	6.2%	1.8%	PAN16 (2016)	428	78
IV	first and last part from I, split at spaces	III	>.1%	0.0%	0.0%	>.1%	0.0%	PAN17 (2017b)	3,600	2,400
V	all but the last part from I	IV	1.8%	23.3%	5.6%	3.8%	5.3%	PAN18 (2018)	2,000	1,900
VI	all but the last two parts from I	V	2.9%	21.8%	9.2%	10.6%	9.6%	Celebrities	31,861	13,614
		VI	0.3%	1.6%	12.3%	1.9%	0.8%			

Table 3: (a) Rules to generate name candidates for Wikidata matching from Twitter reference and display names. (b) Evaluation of matching success as per generation rule. (c) Sizes of the datasets used for evaluation.

Label	Occurrences	Most frequent value	Most frequent value
Sex	65,035 90.1%	Male	71.7%
Occupation	63,017 87.9%	Actor	15.3%
Date of birth	60,493 84.4%	-	-
Educated at	28,134 39.2%	Harvard	2.1%
Sport	18,688 26.1%	Football	30.8%
Languages spoken	12,094 16.9%	English	54.9%
Political party	6,703 9.4%	Republican	16.4%
Genre	6,699 9.3%	Pop Music	21.6%
Race	3,531 0.5%	African Am.	66.5%
Religion	2,960 0.4%	Islam	23.5%

Table 4: Selection of relevant personal traits studied in the related work, how often they have been assigned in our corpus and the most frequent value for each label.

replies. Of the 49.7% remaining tweets, an average of 989 (13,938 words) per celebrity are longer than 20 characters and do not contain links, yielding a conservative estimate of tweets amenable for style analysis. Although celebrities tweeted in 50 different languages, 77% of all timelines consisted of tweets exclusively written in English, followed by 7% in Spanish and 4% in French, while 2,104 celebrities tweeted at least bilingual.

### 3.5 Corpus Reliability and Limitations

Regarding the representativeness of our sample from the population of celebrities, we may cautiously claim to have obtained a wide cross-section of people of elevated status. However, celebrities are excluded who do not use Twitter, whose account is not verified (which is exceedingly unlikely, the more famous they are), or who have no Wikipedia article about themselves. There are no reliable estimates of the true number of celebrities worldwide, but it is safe to assume that our corpus has a bias towards Western culture, and particularly English-speaking celebrities.

Regarding profile comprehensiveness, our corpus provides for comparably long samples of writing per author and a rich set of traits, albeit many traits are available only for a subset of profiles. Most celebrities provide genuine writing samples of themselves at Twitter, but some employ public

relations staff to manage their account. Though a problem for generic author profiling, this does not impede *celebrity profiling*. Celebrities craft public personas as their own unique brands. If a celebrity decides to employ staff to do so, approving their impersonations, these personas are no less genuine and normative than personally crafted personas.

The information about the traits of celebrities obtained from Wikidata can be considered highly reliable. Dedicated volunteers collect all kinds of personal information about celebrities, which are often referenced and under constant review by other Wikipedia and Wikidata editors. As per our taxonomy of label acquisition strategies in Table 2, we employ an OIS strategy: we obtain labels from third-party expert annotators (O), who are independent (I), supplying data in structured form (S).

## 4 Evaluation

To investigate the usefulness of our corpus for author profiling, we carry out a first large-scale profiling experiment by predicting celebrity occupation and gender and evaluating four state of the art approaches that won the PAN 2015-2018 author profiling competitions. Instead of retraining their prediction models, we use the models for gender inference as they have been trained on the PAN training datasets provided to participants of the respective years. Additionally, we train our own baseline gender model on celebrity profiles. Gender is a suitable benchmark trait that is frequently studied in the related work and a recurring trait prediction task at PAN. We observe a successful model transfer, thus mutually corroborating that ours and the PAN corpora capture the same underlying concept of gender.

### 4.1 Preprocessing and Baselines

For our experiments, we extracted a subset of 45,475 English-speaking profiles from our corpus with the traits gender and occupation and split it 70/30 into training and test sets. Table 3c shows

Model	PAN15	PAN16	PAN17	PAN18	Celeb
alvarezcamona15 (2015)	<b>0.859</b>	–	–	–	0.723
nissim16 (2016)	–	0.641	–	–	0.740
nissim17 (2017)	–	–	<b>0.823</b>	–	0.855
danehsvar18 (2018)	–	–	–	<b>0.822</b>	0.817
CNN (Celeb)	0.747	0.590	0.747	0.756	<b>0.861</b>
CNN (Celeb + PAN15)	0.793	–	–	–	–
CNN (Celeb + PAN16)	–	<b>0.690</b>	–	–	–
CNN (Celeb + PAN17)	–	–	0.768	–	–
CNN (Celeb + PAN18)	–	–	–	0.759	–

Table 5: Accuracy of (top) the state of the art gender prediction approaches on their respective datasets and transfer performance to celebrities, and (bottom) our baseline deep learning approach, with and without re-training on the PAN datasets.

this dataset in comparison to the PAN datasets. Our subset has 1,379 different occupations annotated, which we manually assigned to eight groups: sports, performer, creator, politics, manager, science, professional, and religious. We preprocessed the text by lowercasing, replacing mentions with `<user>`, hashtags with `<hashtag>`, hyperlinks with `<url>`, number-groups with `<numbers>`, the most frequent emoticons with `<smiley>`, and we removed all punctuation sequences beyond basic English punctuation marks.

As baseline models for gender and for occupation prediction, we adapted the convolutional neural network (CNN) for text classification introduced by Kim (2014). Our variant of this model builds on the 100-dimensional GloVe (Pennington et al., 2014) Twitter embeddings, uses four parallel 1D-convolution layers with 128 filters each for 1-, 2-, 3-, and 4-grams, a 64-node dense layer for concatenation after the convolutions, and a final classification layer. The models for occupation and gender only differ in the last classification layer and loss function used to facilitate binary (gender) and categorical truth (occupation). We limited the vocabulary to the most common 100,000 words and padded the word-sequence for each author to 5000 words, which is roughly the average per author word count between ours and the PAN datasets. In our tests on the celebrity profiles, this hyperparameter setting achieves more consistent results than fewer or shorter n-gram filters, smaller dense layers, shorter or longer sequence length, or a larger vocabulary. Note that our corpus has labels for more than the two sexes male and female, however, the PAN data did not, so that we excluded profiles with other genders from our experiments, leaving their investigation for future work.

## 4.2 Evaluation Results

Table 5 shows all models’ transfer performance between populations on gender. In general, all models generalize well to the respectively unseen datasets but perform best on the data they have been specifically trained for. The largest difference can be observed on the sub-1,000 author dataset PAN15, where the model of Álvarez-Carmona et al. (2015) suffers a significant performance loss, and PAN16, where the model of Busger op Vollenbroek et al. (2016) performs notably better on the celebrity data. This was a surprise to us that may be explained by the longer samples of writing per profile in our corpus. This hypothesis is also supported by the large increase in accuracy of the baseline model after retraining for two epochs with the PAN15 and PAN16 training datasets, respectively. The occupation model achieved a 0.7111 accuracy.

Altogether, the results of our experiments show that profiling models trained on a random choice of people generalize to celebrities, and vice versa. Our corpus can hence be used for generic author profiling, while providing significantly richer profiles in terms of writing samples and as of yet unexplored personal traits. The scale of our corpus allows for the training of deep learning models, which, at least on our corpus, outperform the state of the art. We expect that further fine-tuning of the model architecture will yield significant improvements.

## 5 Conclusion

This paper introduces the Webis Celebrity Corpus 2019, the first corpus of its kind comprising a total of 71,706 celebrity profiles, 239 profiling-relevant labels, and 3 billion words. Its quality is due to Twitter’s verification process, Wikidata’s accuracy, and our low-error linking strategy between the two sites. Its generalizability qualities for gender prediction have been demonstrated using state-of-the-art approaches.

Our corpus formed the basis for the first celebrity profiling competition, organized as part of the PAN evaluation lab (Wiegmann et al., 2019). The traits studied were the degree of fame, occupation, age, and gender, introducing fame and occupations as novel, celebrity-specific profiling traits, and revisiting the well-known traits age and gender.

In future work, we plan on improving the corpus by incorporating verified accounts from other social networks, and, by inferring new labels for as of yet unlabeled celebrities through link prediction.

## References

- Miguel A. Álvarez-Carmona, A. Pastor López-Monroy, Manuel Montes y Gómez, Luis Villaseñor-Pineda, and Hugo Jair Escalante. 2015. INAOE's participation at PAN'15: Author Profiling Task—Notebook for PAN at CLEF 2015. In (Cappellato et al., 2015).
- 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.
- Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors. 2016. *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*, CEUR Workshop Proceedings. CEUR-WS.org.
- Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. *N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017*. In (Cappellato et al., 2017).
- Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric Analysis of Scientific Articles. In *HLT-NAACL*. The Association for Computational Linguistics.
- John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACM.
- Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim. 2016. *GronUP: Groningen User Profiling—Notebook for PAN at CLEF 2016*. In (Balog et al., 2016).
- Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors. 2017. *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland*, CEUR Workshop Proceedings. CEUR-WS.org.
- Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors. 2015. *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*, CEUR Workshop Proceedings. CEUR-WS.org.
- Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors. 2018. *CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Avignon, France*, CEUR Workshop Proceedings. CEUR-WS.org.
- Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at Ibereval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *IberEval@SEPLN*, volume 2150 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *ICWSM*. The AAAI Press.
- Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender Inference of Twitter Users in Non-English Contexts. In *EMNLP*. ACL.
- Saman Daneshvar and Diana Inkpen. 2018. *Gender Identification in Twitter using N-grams and LSA—Notebook for PAN at CLEF 2018*. In (Cappellato et al., 2018).
- Chris Emmery, Grzegorz Chrupala, and Walter Daelemans. 2017. Simple Queries as Distant Labels for Predicting Gender on Twitter. In *NUT@EMNLP*. Association for Computational Linguistics.
- Dominique Estival, Tanja Gaustad, Son Pham, Will Radford, and Ben Hutchinson. 2007a. Author profiling for English Emails.
- Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007b. TAT: An Author Profiling Tool with Application to Arabic Emails. In *ALTA*. Australasian Language Technology Association.
- Mehwish Fatima, Komal Hasan, Saba Anwar, and Rao Muhammad Adeel Nawab. 2017. Multilingual Author Profiling on Facebook. *Inf. Process. Manage.*, 53(4).
- Matej Gjurkovic and Jan Snajder. 2018. Reddit: A Gold Mine for Personality Prediction. In *PEOPLES@NAACL-HTL*. Association for Computational Linguistics.
- Jurgita Kapociute-Dzikiene, Andrius Utkas, and Ligita Sarkute. 2015. Authorship Attribution and Author Profiling of Lithuanian Literary Texts. In *BSNLP@RANLP*. INCOMA Ltd. Shoumen, BULGARIA.
- Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*. ACL.
- Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4).
- Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-Annotated Corpus of Hindi-English Code-Mixed Data. In *LREC*. European Language Resources Association (ELRA).

- Tatiana Litvinova, Pavel Seredin, Olga Litvinova, and Olga Zagorovskaya. 2017. Differences in Type-Token Ratio and Part-of-Speech Frequencies in Male and Female Russian Written Texts. In *Proceedings of the Workshop on Stylistic Variation*. Association for Computational Linguistics.
- George Mikros. 2013. Authorship Attribution and Gender Identification in Greek Blogs. In *Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO)*.
- Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. 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*. ACM.
- James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological Aspects of Natural Language Use: Our Words, our Selves. *Annual Review of Psychology*, 54:547–577.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Barbara Plank and Dirk Hovy. 2015. Personality Traits on Twitter - or - How to get 1,500 Personality Tests in a Week. In *WASSA@EMNLP*. The Association for Computer Linguistics.
- Daniel Preotiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An Analysis of the User Occupational Class through Twitter Content. In *ACL (1)*. The Association for Computer Linguistics.
- Daniel Preotiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle H. Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *ACL (1)*. Association for Computational Linguistics.
- Daniel Preotiuc-Pietro and Lyle H. Ungar. 2018. User-level Race and Ethnicity Predictors from Twitter Text. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Association for Computational Linguistics.
- Ricelli Ramos, Georges Neto, Barbara Barbosa Claudino Silva, Danielle Sampaio Monteiro, Ivandré Paraboni, and Rafael Dias. 2018. Building a Corpus for Personality-Dependent Natural Language Understanding and Generation. In *LREC*. European Language Resources Association (ELRA).
- Francisco Manuel 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 (Cappellato et al., 2015).
- Francisco Manuel Rangel Pardo, Manuel Montes-y-Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th Author Profiling Task at PAN 2018: Cross-domain Authorship Attribution and Style Change Detection. In (Cappellato et al., 2018).
- Francisco Manuel Rangel Pardo, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd Author Profiling Task at PAN 2014. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*, CEUR Workshop Proceedings. CEUR-WS.org.
- Francisco Manuel Rangel Pardo, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the Author Profiling Task at PAN 2013. In *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*. CEUR-WS.org.
- Francisco Manuel Rangel Pardo, Paolo Rosso, Martin Potthast, and Benno Stein. 2017a. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In *Working Notes Papers of the CLEF 2017 Evaluation Labs*, volume 1866 of *CEUR Workshop Proceedings*. CLEF and CEUR-WS.org.
- Francisco Manuel Rangel Pardo, Paolo Rosso, Martin Potthast, and Benno Stein. 2017b. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In (Cappellato et al., 2017).
- Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In (Balog et al., 2016).
- Sara Rosenthal and Kathleen R. McKeown. 2011. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations. In *ACL*. The Association for Computer Linguistics.
- Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. AAAI.
- 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, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. In *PLoS ONE*, page 8(9): e73791.

- Edward P. Tighe and Charibeth K. Cheng. 2018. Modeling Personality Traits of Filipino Twitter Users. In *PEOPLES@NAACL-HTL*. Association for Computational Linguistics.
- Twitter. 2018. FAQ: About verified accounts. <https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts>, accessed 15.11.2018.
- Ben Verhoeven and Walter Daelemans. 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 Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*. European Language Resources Association (ELRA).
- Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.* European Language Resources Association (ELRA).
- Svitlana Volkova and Yoram Bachrach. 2015. On Predicting Sociodemographic Traits and Emotions from Communications in Social Networks and their Implications to Online Self-Disclosure. *Cyberpsy., Behavior, and Soc. Networking*, 18(12):726–736.
- Yuan Wang, Yang Xiao, Chao Ma, and Zhen Xiao. 2016. Improving Users' Demographic Prediction via the Videos they Talk about. In *EMNLP*. The Association for Computational Linguistics.
- Matti Wiegmann, Benno Stein, and Martin Potthast. 2019. Overview of the Celebrity Profiling Task at PAN 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*, CEUR Workshop Proceedings. CEUR-WS.org.
- Wikipedia. 2018. Notability Guidelines for People. [https://en.wikipedia.org/wiki/Wikipedia:Notability\\_\(people\)](https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)), accessed 15.11.2018.