Overview of the Celebrity Profiling Task at PAN 2020

Matti Wiegmann,[1] Benno Stein,[1] and Martin Potthast[2]

[1]Bauhaus-Universität Weimar
[2]Leipzig University

pan@webis.de https://pan.webis.de

Abstract Celebrity profiling is author profiling applied to celebrities. As a subpopulation of social media users, celebrities are prolific authors for whom all kinds of personal information is public knowledge, and whose large followership enables new kinds of author profiling tasks: At this year's PAN, we study for the first time author profiling of social media users where an author's age, gender, and occupation has to be predicted by analyzing ten of their followers, rather than the author's original writing. This paper presents this novel approach to profiling, the 2,380-author dataset we created for to study it, and the three models that participants proposed to solve the problem in diverse ways. The participants' follower-based profiling models achieve F_1-scores that far exceed random guessing, even reaching the performance-level of a baseline author profiling model when predicting occupations. Our evaluation reveals that follower-based profiling models have similar strengths and weaknesses as the author-based profiling models for celebrity profiling: They work best if the classes are topically coherent, as for the "sports" occupation, but less so in the opposite case, as for the "creator" occupation. Additionally, while predicting the age of the celebrities is still difficult, the follower-based models show a trend to predict younger users better than the author-based ones on our dataset.

1 Introduction

Author Profiling, the task of predicting the demographics of an author from their texts, is a central task in authorship analysis with many applications in the social sciences, forensics, and marketing. Author profiling technology has been developed on, and applied to many demographics, genres, and related tasks and often achieves good results, but all common approaches require lots of high-quality text for training from the authors in question. Especially on social media, which is the currently dominant genre in the field of author profiling, authors with both many public, high-quality texts and verified personal demographics are few and far between. With current technology, it is not possible to profile users that write only a few textual posts and only interact by reading, liking, and forwarding the messages of other authors. Since these passive authors are very frequent on social media, one can profile them only based on other factors. One

such factor that provides information about passive authors are the messages posted by other authors who are closely connected to them. Social media theory points out that users with similar demographics and interests form online communities and that online communities develop sociolects (language variation [11]), so inspecting the author's friends, followers, and their social graph relations may also hint at an author's demographics. Since celebrities are well-connected, influential, and elevated figures in their communities, they are a suitable subpopulation to study algorithms that profile passive users based on the social graph using the posts of connected authors.

After introducing the task of celebrity profiling [46] for the first time, we organized a corresponding shared task at PAN 2019 [47], asking participants to profile the age, gender, fame, and occupation of a celebrity given their Twitter timeline. For the shared task on celebrity profiling at PAN 2020, we tackle the problem of profiling passive authors, namely by asking participants to predict three demographics of a celebrity—age as a 60-class problem, gender as a 2-class problem, and occupation as a 4-class problem—, given only the tweets of the celebrity's followers. For this task, we constructed a new dataset containing 2,320 celebrities, each annotated with all three demographics, and the Twitter posts of 10 randomly selected, but active followers, each with a sufficient amount of original, English tweets. For consistency, we reused the ranking measure from the previous celebrity profiling task: the harmonic mean of the macro-averaged multi-class F_1 for gender, occupation, and a leniently calculated F_1 for age. Three teams submitted a diverse range of models, all outperforming a baseline model trained on the followers' texts, improving strongly above random guessing, and closing in on another baseline trained on the celebrities' tweets. We thus demonstrated that the task is, in fact, solvable. An in-depth evaluation reveals similar strengths and weaknesses of the models compared to the previous celebrity profiling task: Topically homogeneous occupations (e.g., sports) are easier to predict than heterogeneous ones (e.g., creators), and younger users are easier to predict than older ones.

After reviewing the related work in Section 2, we describe in more detail the task, the construction of the task's datasets, the reasoning underlying our performance measures, and our baselines in Section 3. In Section 4, we survey the software submissions, in Section 5, we report the evaluation results and present our analysis concerning the performance of different approaches and individual demographics of the task.

2 Related Work

The study of author profiling techniques has a rich history, with the pioneering works done by Pennebaker et al. [28], Koppel et al. [19], Schler et al. [42], and Argamon et al. [3], focusing on age, gender, and personality from genres with longer, grammatical documents such as blogs and essays. The most commonly used genre in recent years is Twitter tweets, first used in 2011 to predict gender [7] and age [27]. Later work also used Facebook posts [13], Reddit [15], and Sina Weibo [45]. Recently added demographics include education [10], ethnicity [44], family status [45], income [31], occupation [30], location of origin [13], religion [33], and location of residence [10].

At PAN, author profiling has been studied since 2013, covering different demographics including age and gender [39, 38, 41], personality[34], language variety [40], genres like blog posts, reviews, and social media messages [41], predicting across gen-

res [36], and profiling author characteristics outside the domain of demographics, such as the authors inclination to spread fake news [35] of detecting if an author writes like a bot [37]. The population of celebrities, introduced to author profiling by Wiegmann et al. [46], has been studied at PAN since 2019 with the first shared task on celebrity profiling [47] with the goal on predicting age, gender, occupation, and fame of 48,335 celebrities given the respective Twitter timeline.

Methodologically, author profiling has been comparatively stable over the last decade: most approaches utilize supervised machine learning based on the authors' texts and varying stylometric and psycholinguistic features to encode non-lexical information. The additional features proved to be important to the degree that even advanced neural network architectures are only competitive if these features are explicitly encoded [14]. The biggest methodological improvements, experimentally shown for selected demographics, are the usage of message-level attention, recently proposed by Lynn et al. [21] and of network homophily by encoding information from the social graph. The pioneering work by Kosinski et al. [20] shows that the common likes of Facebook users suffice to predict demographics like gender, sexual orientation, ethnicity, and substance use behavior with up to 0.9 accuracy. Recent advances in graph encoding algorithms [16] motivated the use of node embeddings as supplemental features when predicting age and gender on Facebook [12], occupation and income [1], racism and sexism [22], and suicide ideation [23] on Twitter. Similar approaches have also been explored in related fields to, for example, profile the bias and factuality of news agencies [5]. An even more advanced approach to predict the occupation of authors was suggested by Pan et al. [24], who jointly encoded the adjacency matrix of the follower graph with the biographies of all authors in the network using graph convolutional neural networks. Additionally, the metadata of related authors in the social graph is central in other user analysis tasks, like geolocation prediction [4].

Besides text-based author profiling through the homophily of social networks, several studies explore language variation and convergence on social media. Essentially, language variation and convergence explains how groups of people adopt lexical changes and are, together with the psycholinguistic preferences of social groups studied by Pennebaker et al. [28], the reason author profiling is possible. The works that explore language variation have shown, for example, that online language does not convergence to a common "netspeak" but often follows the geographic and demographic [11] similarities of online communities. Besides real-world factors, a significant impact on lexical variation is attributed to social factors. For example, Pavalanathan and Eisenstein [25] show that lexical variation decreases with the size of the intended audience, which means that social media texts have less lexical variation if they are addressed to a larger audience. Similarly, Tamburrini et al. [43] have shown that an author's words are based on the social identity of the conversion-partner. The specific impact of the network structure on the language variations was studied by [17] who found that language variation is adopted more quickly if individuals are more closely connected. Based on the related work, it is reasonable to assume that the same linguistic processes of lexical variation and convergence used by Pennebaker et al. [28] to profile individuals based on the individuals' texts also apply to social groups, and it is also possible to profile individuals to a degree based on the social groups' texts.

3 Task Description: Follower-based Author Profiling

We introduce the task of follower-based author profiling: It's goal is to predict the three demographics of a social media user from the writing of their followers. We operationalize the task using celebrities and their followership on Twitter, asking for the prediction of their age, gender, and occupation. Our training dataset contains the timelines of ten randomly chosen followers per celebrity with at least 100 original English tweets for each of the 2,000 celebrities, balanced by gender and occupation. Likewise, the test dataset contains another 200 celebrities. The performance of the submissions was judged by the harmonic mean of the multi-class F_1 scores of each demographic, and evaluated using the TIRA evaluation platform [29]. All data and code are publicly available.[1]

3.1 Evaluation Data

The dataset for our shared task is has been sampled from the Webis Celebrity Profiling Corpus 2019 [6]. This corpus contains the Twitter IDs of 71,706 celebrities and extensive demographic information collected about them on Wikidata. We started by extracting all celebrities from the corpus where the target demographics age, gender, and occupation are known simultaneously, omitting all celebrities with demographics outside of the targeted demographic spectrum:

- *Gender.* From the eight different gender-related Wikidata labels, only *male* and *female* were kept, since all others are rare and too diverse for a meaningful grouping.
- *Occupation.* From the 1,379 different occupation-related Wikidata labels, only those belonging to the following, manually determined super-classes were kept (all others are too rare):
  - *Sports* for occupations participating in professional sports, primarily athletes.
  - *Performer* for creative activities primarily involving a performance like acting, entertainment, TV hosts, and musicians.
  - *Creator* for creative activities with a focus on creating a work or piece of art, for example, writers, journalists, designers, composers, producers, and architects.
  - *Politics* for politicians and political advocates, lobbyists, and activists.
- *Age.* Unlike the profiling literature on age prediction, we did not define a static set of age groups, but used the year of birth between *1940* and *1999* as extracted from Wikidata's `Day of Birth` property. Figure 1 shows the distribution of the years of birth in the training and test datasets.

For this selection of celebrity profiles, we downloaded the Twitter IDs of up to 100,000 followers, starting with the most recent. To limit excessive downloading of follower profiles, we first acquired the user descriptions for all followers and discarded all but the most active users with more than ten followers, more than ten followees, and at least fifteen messages. Afterward, all users with more than 100,000 followers

[1] See https://pan.webis.de/data.html and https://github.com/pan-webis-de/pan-code.

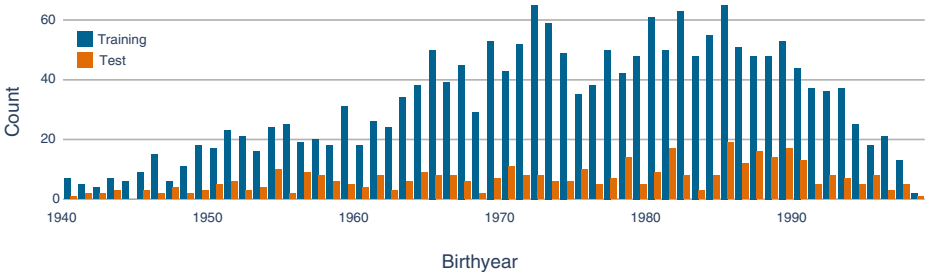

Figure 1. Histogram showing the age distribution over both datasets, training and test.

or 1,000 followees were discarded. Finally, the timelines of all remaining followers were downloaded, omitting all retweets, replies, and non-English tweets. To compile the dataset, we randomly selected ten followers per celebrity which had at least 100 tweets left. This initial compilation of the evaluation dataset contained 10,585 celebrity profiles with ten followers per celebrity and with at least 100 original, English Tweets per follower. From this initial compilation, we selected the largest possible sample of profiles balanced by occupation and by gender, yielding 2,320 celebrities for training and test, and leaving 8,265 celebrities for an unbalanced, supplemental dataset. We split the 2,320 celebrity dataset roughly 80:20 into a 1,920-author training dataset and a 400-author test dataset test, and handed out the training and supplemental datasets to the participants, keeping the test data hidden for the cloud-based evaluation on TIRA.

3.2 Performance Measures

For consistency and comparability with the 2019 edition of celebrity profiling, performance is measured again as the harmonic meanf of the per-demographic performance:

$$\text{cRank} = \frac{3}{\frac{1}{\text{F}_{1,\text{age}}} + \frac{1}{\text{F}_{1,\text{gender}}} + \frac{1}{\text{F}_{1,\text{occupation}}}}.$$

Let T denote the set of classes labels of a given demographic (e.g., gender), where $t \in T$ is a given class label (e.g., female). The prediction performance for $T \in \{\text{gender, occupation}\}$ is measured using the macro-averaged multi-class F_1-score. This measure averages the harmonic mean of precision and recall over all classes of a demographic, weighting each class equally, and thus promoting correct predictions of small classes:

$$\text{F}_{1,T} = \frac{2}{|T|} \cdot \sum_{t_i \in T} \frac{\text{precision}(t_i) \cdot \text{recall}(t_i)}{\text{precision}(t_i) + \text{recall}(t_i)}.$$

We also apply this measure to evaluate the prediction performance for the demographic T = age, but change the computation of true positives: a predicted year is counted as correct if it is within an ε-environment of the true year, where ε increases linearly from 2 to 9 years with the true age of the celebrity in question:

$$\varepsilon = (-0.1 \cdot \text{truth} + 202.8).$$

This way of measuring the prediction performance for the age demographic addresses a shortcoming of the traditional "fixed-age interval scheme:" Defining strict age intervals (e.g., 10-20 years, 20-30, etc.) overly penalizes small prediction errors made at the interval boundaries, such as predicting an age of 21 instead of 20. Furthermore, we decided against combining precise predictions with an error function like mean squared error, since we presume that age prediction is more difficult for older users since the writing style presumably changes more slowly with increasing maturity.

3.3 Baselines

Since our task is rather novel, few competitive baselines are available, so that we resort to two basic approaches instead: The baseline `n-gram` which uses the follower timelines, and the baseline `oracle`, which is identical to `n-gram` but uses the celebrities' timelines instead of the follower timelines. Both baselines solve the task with a multinominal logistic regression [26], where the inputs are the TF-IDF vectors of the respective tweets. The texts are preprocessed by lowercasing, replacing hashtags, usernames, emoticons, emojis, time expressions, and numbers with respective special tokens, removing all remaining newlines and non-ASCII characters, and collapsing spaces. The TF-IDF vectors are constructed from the word 1-grams and 2-grams of all concatenated tweets of the celebrities or followers, respectively, with a per-celebrity frequency of at least 3. We added special separator tokens to encode the end of a tweet and the end of a follower timeline. Due to the lenient calculation of $F_{1,\mathrm{age}}$, the age prediction was simplified to the five years: 1947, 1963, 1975, 1985, and 1994.

4 Survey of the Submitted Approaches

Three participants submitted their software to our shared task. Altogether, the submissions were methodologically diverse, covering creative feature engineering, thorough feature selection, and contemporary deep learning methods. As opposed to last year, neither approach is generally superior to the other ones, with each showing individual strengths and weaknesses in some demographics. The overall ranking of the approaches is shown in Table 1; in what follows, each approach is reviewed in more detail.

The approach of Price and Hodge [32] utilizes a logistic regression classifier for each individual demographic. The model does not directly use representations of the text, but entirely relies on hand-crafted features as input; specifically: the average tweet length per celebrity, the average of all word vectors of the followers' tweets, and the to-token-ratios of the POS-tags, stop words, named entity types, number of links, hashtags, mentions, and emojis. To optimize their model, the authors used 20% of the training dataset for validation in order to pre-evaluate three competing algorithms for each demographic: logistic regression, random forest, and support vector machines. The optimal setting of hyperparameters was determined via five-fold cross-validation on the remaining 80% of the training dataset for each evaluated algorithm, where the optimal parameters were determined using the macro-F_1 score. The final model selection on the left-out validation dataset using the official evaluation measures showed that the logistic regression model was best-suited for all demographics.

Team	cRank	Age	Gender	Occupation
Baseline `oracle`	**0.631**	**0.500**	**0.753**	0.700
Price and Hodge [32]	0.577	0.432	0.681	**0.707**
Koloski et al. [18]	0.521	0.407	0.616	0.597
Alroobaea et al. [2]	0.477	0.315	0.696	0.598
Baseline `n-gram`	0.469	0.362	0.584	0.521
Random	0.333	0.333	0.500	0.250

Table 1. Results of the celebrity profiling task at PAN 2020. Bold scores mark the best overall performance, underlined scores the best performance achieved by a participant.

The approach of Koloski et al. [18] utilizes a logistic regression classifier to predict the age in eight classes, another logistic regression classifier to predict the occupation, and an SVM to predict the gender of the celebrities. The model primarily uses lexical representations as features, but limits the input text to 20 tweets per follower and thus 200 tweets in total per celebrity. Specifically, the features are computed by (1) preprocessing the text into three versions: the original tweets, the tweets without punctuation, and the tweets without punctuation and stop words; (2) computing the top 20,000 most frequent character 1-grams and 2-grams and word 1-grams, 2-grams, and 3-grams; and (3) extracting 512 dimensions with a singular value decomposition to be used as features. To optimize their model, the authors first split the training dataset 90:10 into a training and validation set, and used the training split in a five-fold cross-validation to find the optimal n-gram limit, feature dimensionality, and age prediction strategy. Specifically, six alternative feature counts between 2,500 and 50,000 were tested, seven alternative feature dimensions between 128 and 2048, and three different strategies to solve the age prediction task: as a regression task, as a classification task with 60 classes, and as a classification task with eight classes. After optimizing parameters, the authors selected their model based on their performance on the validation dataset, comparing XGBoost, logistic regression, and linear SVMs for each demographic.

The approach of Alroobaea et al. [2] utilizes an LSTM neural network for classification; however, no further details are revealed about its architecture. The model uses exclusively the followers' texts as a TF-IDF matrix as input. The text itself is preprocessed by removing links, HTML-style tags, stop words, non-alphanumeric tokens, and typical punctuation marks, replacing mentions with @, and stemming all remaining tokens with NLTK's Snowball stemmer. The authors did not report on any experiments to optimize their model.

5 Results and Discussion

Table 1 shows the results of the participants with successful submissions as well as the performance of the three aforementioned baselines. All participants managed to surpass the random expectation and improve on the `n-gram` baseline, the winning approach by 0.11 F_1 in the combined metric cRank. The best performance of the submitted solutions already closes in on the `oracle` baseline, which shows that the followers'

Table 2. Class-wise performance for each individual demographic. Listed are the F_1 scores for age, gender, and occupation. For ease of interpretation, the age is evaluated over five classes and the table lists the centroid year of birth of each class together with the mean absolute error (MAE). Bold scores mark the best overall performance, underlined scores the best performance achieved by a participant.

Team	**Age**						**Gender**		**Occupation**			
	1994	1985	1963	1975	1947	MAE	Female	Male	Creat.	Perf.	Polit.	Sports
Baseline `oracle`	0.215	**0.632**	0.476	**0.396**	**0.129**	**7.37**	0.708	**0.762**	0.419	0.645	**0.864**	0.772
Price and Hodge [32]	0.274	0.463	0.420	0.319	0.036	9.49	0.661	0.697	**0.457**	**0.731**	0.776	**0.830**
Koloski et al. [18]	**0.402**	0.389	0.480	0.165	0.000	11.14	0.354	0.689	0.292	0.629	0.693	0.632
Alroobaea et al. [2]	0.000	0.111	**0.497**	0.361	0.000	10.89	**0.712**	0.676	0.454	0.519	0.678	0.721
Baseline `n-gram`	0.362	0.445	0.415	0.226	0.000	10.12	0.434	0.678	0.248	0.578	0.645	0.488

texts contain noticeable hints about the demographics of the followee. Table 2 shows the F_1 scores for each individual class. The results show, that, although the submitted approaches are quite diverse, their weaknesses are structural and allow some cautious conclusions about the underlying profiling problem. First, it is easier to predict the age of the youngest celebrities from follower tweets than from their own, but age prediction gets increasingly difficult with increasing age. Second, predicting the celebrities' gender from their follower tweets works better for male celebrities. Third, predicting the occupation based on follower tweets competes with the oracle baseline.

The best-performing submission for predicting the age of the celebrities from their follower tweets from Price and Hodge achieved an F_1 score of 0.432, which is with a distance of 0.07 directly in-between the baselines `n-gram` and `oracle`. Judging by the multi-class F_1 scores shown in Table 1, the age prediction task is the most difficult demographic to predict this year. For ease of analysis, we evaluate the age prediction subtask as a five-class problem over the ranges of birth years with the centroids 1994, 1985, 1963, 1975, and 1947. The results of the multi-class F_1 scores shown in Table 1, the class-wise F_1 scores shown in Table 2, and the misclassifications depicted in the confusion matrices in Figure 2 (top) allow for three observations: First, most submitted models simply perform better on the majority classes. Since no participant employed resampling to balance the training data, this effect may be due to the unbalanced training data. The confusion matrices illustrate this effect, where all models skew towards the center range of birth years, except for the one of Koloski et al., who optimized the age-prediction strategy to achieve the opposite effect: their model skews towards never predicting the center age group. Second, both the `n-gram` baseline and the model of Koloski et al. significantly outperform the `oracle` baseline on predicting the youngest celebrities, born between 1990 and 1999. This observation is not explained by the class imbalance or sampling: Although both, Koloski et al. and the `n-gram` baseline, resample the age classes from 60 classes down to five or eight, respectively, they still significantly outperform the `oracle` baseline, which also reduces the number of age groups to predict. The results do not fully explain this behavior, but it may hint at useful information contained in follower tweets towards better detecting the youngest celebrities. However, the increased performance when predicting young celebrities does not improve the performance in general, since the `oracle` baseline, followed by the model

Figure 2. Confusion matrices for the demographics age, gender, and occupation (top to bottom).

of Price and Hodge, still achieve better multi-class F_1 scores and mean absolute errors. Third, all models poorly predict the oldest celebrities born between 1940 and 1955, although, as shown in Figure 1, this class has as many subjects as the 1990–1999 year range while covering a broader age spectrum.

The best-performing submission for predicting the gender of the celebrities from their follower tweets from Alroobaea et al. achieved an F_1 score of 0.696, which is with a distance of 0.057 closer to the `oracle` than with 0.112 to the `n-gram` baseline. Predicting the binary gender has been included as a baseline task since it is very commonly done when predicting demographics, and typically achieves accuracies above the mark of 0.9. Based on the observed results, gender prediction is more difficult for the sampled celebrities. The F_1 scores and the confusion matrices, as shown in Figure 2 (middle), allow for one observation: The models tend towards predicting a celebrity as male rather than as female. This kind of skew is typically explained by imbalanced data or dataset sampling. However, both explanations are unlikely, since our dataset is balanced and has 200 celebrities per class, which is usually sufficient to avoid biased data. The best-performing model in this demographic tends to predict female over male, and the `oracle` baseline, using the celebrities' timelines, does so, too.

The best-performing submission for predicting the occupation of the celebrities from their follower tweets from Price and Hodge achieved an F_1 score of 0.707, which is marginally better than the `oracle` baseline by 0.007, on average. Predicting the occupation is the easiest part of our shared task. We assume that occupation prediction relies heavily on topic markers in the text, and that these topics are the common

ground for discussion between the followers of a celebrity. In this respect, it is surprising that the submission supposedly encoding the least lexical but most stylometric features achieved the best performance. The results of the F_1 scores and the confusion matrices shown in Figure 2 (bottom) allow for one further observation: Although the class-wise results are mixed between the different submissions, politicians, performers, and athletes (sports) are consistently predicted well, while creators are consistently misclassified as either performer or politicians. These results are mostly consistent with the results of the 2019 task, albeit, this year, politicians were less frequently misclassified than athletes.

6 Conclusion and Outlook

This paper overviews the shared task on celebrity profiling at PAN 2020. The goal of the task was to determine three demographics of celebrities on Twitter based on their followers writing, rather than their own: the age as a 60-class problem with lenient evaluation, the gender as a two-class problem, and the occupation as a four-class problem. The submitted models rely on a variety of proven methods: feature-based machine learning with stylometric or n-gram features, and LSTMs on TF-IDF matrices. The individual demographics' results point towards similar difficulties as were found the corresponding shared task of 2019: the topically more diverse occupations "creator" and "performer" are harder to profile, as are older authors over younger ones. Our results impressively demonstrate that it is possible to profile authors based on their followers' texts almost as well as on their own. However, there is still much potential to explore different approaches and gain further insights. Technologically, utilizing the messages of followers to improve author profiling models is a promising future direction.

Acknowledgments

We thank our participants for their effort and dedication, and the CLEF organizers for hosting PAN and the shared task on celebrity profiling.

Bibliography

[1] Aletras, N., Chamberlain, B.P.: Predicting twitter user socioeconomic attributes with network and language information. In: Proceedings of the 29th on Hypertext and Social Media, pp. 20–24 (2018)

[2] Alroobaea, R., Almulihi, A.H., Alharithi, F.S., Mechti, S., Krichen, M., Belguith, L.H.: A Deep learning Model to predict gender, age and occupation of the celebrities. In: [8]

[3] Argamon, S., Koppel, M., Pennebaker, J., Schler, J.: Automatically Profiling the Author of an Anonymous Text. Commun. ACM **52**(2), 119–123 (Feb 2009), ISSN 0001-0782, https://doi.org/10.1145/1461928.1461959

[4] Bakerman, J., Pazdernik, K., Wilson, A.G., Fairchild, G., Bahran, R.: Twitter Geolocation: A Hybrid Approach. ACM Trans. Knowl. Discov. Data **12**(3) (2018), https://doi.org/10.1145/3178112

[5] Baly, R., Karadzhov, G., An, J., Kwak, H., Dinkov, Y., Ali, A., Glass, J., Nakov, P.: What was written vs. who read it: News media profiling using text analysis and social media context. arXiv preprint arXiv:2005.04518 (2020)
[6] Bevendorff, J., Potthast, M., Hagen, M., Stein, B.: Heuristic Authorship Obfuscation. In: Korhonen, A., Màrquez, L., Traum, D. (eds.) 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), pp. 1098–1108, Association for Computational Linguistics (Jul 2019), URL https://www.aclweb.org/anthology/P19-1104
[7] Burger, J., Henderson, J., Kim, G., Zarrella, G.: Discriminating Gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309, ACM (2011)
[8] Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.): CLEF 2020 Labs and Workshops, Notebook Papers, CEUR Workshop Proceedings, CEUR-WS.org (Sep 2020)
[9] Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.): CLEF 2019 Labs and Workshops, Notebook Papers, CEUR Workshop Proceedings, CEUR-WS.org (Sep 2019), URL http://ceur-ws.org/Vol-2380/
[10] Carmona, M.Á.Á., Guzmán-Falcón, E., Montes-y-Gómez, M., Escalante, H.J., Pineda, L.V., Reyes-Meza, V., Sulayes, A.R.: Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: IberEval@SEPLN, CEUR Workshop Proceedings, vol. 2150, pp. 74–96, CEUR-WS.org (2018)
[11] Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: Diffusion of lexical change in social media. PLOS ONE **9**(11), 1–13 (11 2014), https://doi.org/10.1371/journal.pone.0113114, URL https://doi.org/10.1371/journal.pone.0113114
[12] Farnadi, G., Tang, J., De Cock, M., Moens, M.F.: User profiling through deep multimodal fusion. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 171–179 (2018)
[13] Fatima, M., Hasan, K., Anwar, S., Nawab, R.M.A.: Multilingual author profiling on facebook. Inf. Process. Manage. **53**(4), 886–904 (2017)
[14] Giachanou, A., Rosso, P., Crestani, F.: Leveraging emotional signals for credibility detection. In: Proceedings of the 42nd International ACM SIGIR, pp. 877–880, ACM (2019), https://doi.org/10.1145/3331184.3331285
[15] Gjurkovic, M., Snajder, J.: Reddit: A gold mine for personality prediction. In: PEOPLES@NAACL-HTL, pp. 87–97, Association for Computational Linguistics (2018)
[16] Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 855–864 (2016)
[17] Kershaw, D., Rowe, M., Stacey, P.: Towards Modelling Language Innovation Acceptance in Online Social Networks. In: Proceedings of the Ninth ACM WSDM, pp. 553–562, ACM (2016), https://doi.org/10.1145/2835776.2835784
[18] Koloski, B., Pollak, S., Škrlj, B.: Know your Neighbors: Efficient Author Profiling via Follower Tweets. In: [8]
[19] Koppel, M., Argamon, S., Shimoni, A.: Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing **17**(4), 401–412 (2002)
[20] Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the national academy of sciences **110**(15), 5802–5805 (2013)
[21] Lynn, V., Balasubramanian, N., Schwartz, H.A.: Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (Jul 2020), https://doi.org/10.18653/v1/2020.acl-main.472, URL https://www.aclweb.org/anthology/2020.acl-main.472

[22] Mishra, P., Del Tredici, M., Yannakoudakis, H., Shutova, E.: Author profiling for abuse detection. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1088–1098 (2018)
[23] Mishra, R., Sinha, P.P., Sawhney, R., Mahata, D., Mathur, P., Shah, R.R.: Snap-batnet: Cascading author profiling and social network graphs for suicide ideation detection on social media. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 147–156 (2019)
[24] Pan, J., Bhardwaj, R., Lu, W., Chieu, H.L., Pan, X., Puay, N.Y.: Twitter homophily: Network based prediction of user's occupation. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th ACL, pp. 2633–2638, Association for Computational Linguistics (2019), https://doi.org/10.18653/v1/p19-1252
[25] Pavalanathan, U., Eisenstein, J.: Audience-modulated variation in online social media. American Speech **90** (05 2015), https://doi.org/10.1215/00031283-3130324
[26] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
[27] Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting Age and Gender in Online Social Networks. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents, pp. 37–44, SMUC '11, ACM, New York, NY, USA (2011), ISBN 978-1-4503-0949-3, https://doi.org/10.1145/2065023.2065035, URL http://doi.acm.org/10.1145/2065023.2065035
[28] Pennebaker, J., Mehl, M., Niederhoffer, K.: Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology **54**, 547–577 (2003), ISSN 0066-4308, https://doi.org/10.1146/annurev.psych.54.101601.145041
[29] Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer (Sep 2019), ISBN 978-3-030-22948-1, https://doi.org/10.1007/978-3-030-22948-1_5
[30] Preotiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through twitter content. In: ACL (1), pp. 1754–1764, The Association for Computer Linguistics (2015)
[31] Preotiuc-Pietro, D., Ungar, L.H.: User-level race and ethnicity predictors from twitter text. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, pp. 1534–1545, Association for Computational Linguistics (2018), URL https://aclanthology.info/papers/C18-1130/c18-1130
[32] Price, S., Hodge, A.: Celebrity Profiling using Twitter Follower Feeds: Notebook for PAN at CLEF 2020. In: [8]
[33] Ramos, R., Neto, G., Silva, B.B.C., Monteiro, D.S., Paraboni, I., Dias, R.: Building a corpus for personality-dependent natural language understanding and generation. In: LREC, European Language Resources Association (ELRA) (2018)
[34] Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France, CEUR Workshop Proceedings, CEUR-WS.org (Sep 2015), ISSN 1613-0073, URL http://ceur-ws.org/Vol-1391
[35] Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: [8]

[36] Rangel, F., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Cross-domain Authorship Attribution and Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France, CEUR Workshop Proceedings, CEUR-WS.org (Sep 2018), ISSN 1613-0073, URL http://ceur-ws.org/Vol-2125/

[37] Rangel, F., Rosso, P.: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In: [9], URL http://ceur-ws.org/Vol-2380/

[38] Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd Author Profiling Task at PAN 2014. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) Working Notes Papers of the CLEF 2014 Evaluation Labs, CEUR Workshop Proceedings, CEUR-WS.org (Sep 2014), ISSN 1613-0073, URL http://ceur-ws.org/Vol-1180/

[39] Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, CEUR-WS.org (Sep 2013), ISBN 978-88-904810-3-1, ISSN 2038-4963, URL http://ceur-ws.org/Vol-1179

[40] Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland, CEUR Workshop Proceedings, CEUR-WS.org (Sep 2017), ISSN 1613-0073, URL http://ceur-ws.org/Vol-1866/

[41] Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal, CEUR Workshop Proceedings, CEUR-WS.org (Sep 2016), ISSN 1613-0073, URL http://ceur-ws.org/Vol-1609/16090750.pdf

[42] Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 199–205, AAAI (2006)

[43] Tamburrini, N., Cinnirella, M., Jansen, V., Bryden, J.: Twitter users change word usage according to conversation-partner social identity. Social Networks **40**, 84?89 (01 2015), https://doi.org/10.1016/j.socnet.2014.07.004

[44] Volkova, S., Bachrach, Y.: 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 (2015)

[45] Wang, X., Bendersky, M., Metzler, D., Najork, M.: Learning to Rank with Selection Bias in Personal Search. In: SIGIR, pp. 115–124, ACM (2016)

[46] Wiegmann, M., Stein, B., Potthast, M.: Celebrity Profiling. In: Korhonen, A., Màrquez, L., Traum, D. (eds.) 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), pp. 2611–2618, Association for Computational Linguistics (Jul 2019), URL https://www.aclweb.org/anthology/P19-1249

[47] Wiegmann, M., Stein, B., Potthast, M.: Overview of the Celebrity Profiling Task at PAN 2019. In: [9], URL http://ceur-ws.org/Vol-2380/