Overview of the Celebrity Profiling Task at PAN 2019

Matti Wiegmann,[1,2] Benno Stein,[1] and Martin Potthast[3]

[1]Bauhaus-Universität Weimar
[2]German Aerospace Center
[3]Leipzig University

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

Abstract Celebrity profiling is author profiling applied to celebrities. The focus on celebrities has several advantages: Celebrities are prolific social media users supplying lots of writing samples, lots of personal details are public knowledge, and they try to build a consistent public persona either themselves or with the help of agents. In addition, a number of demographics apply only to this group of people. In this overview of the first shared task on celebrity profiling at PAN 2019, we survey and evaluate eight submitted models that try to predict the gender, the year of birth, the fame, and the occupation of 48,335 English-speaking celebrities based on text obtained from their Twitter timelines. Anticipating some key results we can report that the models work well for predicting binary gender or for distinguishing the most famous celebrities from the less famous ones. Also the occupations sports, politics, and performer are easily identified. The models work less well for the prediction of rare demographics such as non-binary gender and occupations that are not single-topic (e.g., manager, science, and professional). Predicting the year of birth works best for the years between ca. 1980-2000 (i.e., ages ca. 20-40), but less well for older celebrities, and not at all for younger ones.

1 Introduction

Author profiling aims to correlate writing style with author demographics. It has applications in marketing, forensic linguistics, psycholinguistics, and the social sciences. Especially today's omnipresence of social media and the resulting availability of text data from large portions of the population caused a surge of interest in profiling technology. On social media, celebrities occupy an exalted position. Rallying up to millions of followers, they serve as role models to many and exert a direct influence on public opinion, sometimes for the better, e.g., by lending their voices to the disenfranchised, and sometimes for the worse. Unsurprisingly, the "rich and famous" are subjects to research in the social sciences and economics alike, especially with regard to their presence on social media. The celebrity profiling task at PAN 2019 introduces this population for the first time to the author profiling community.

The task was to predict four demographics of celebrities, given their history of tweets on Twitter:

- *Gender* as male, female, or, for the first time, non-binary.
- *Year of birth* within a novel, variable-bucket evaluation scheme.
- *Fame* as rising, star, or superstar.
- *Occupation* or “claim to fame,” as sports, performer, creator, politics, manager, science, professional, or religious.

The evaluation data for this task was sampled from the Webis Celebrity Profiling Corpus 2019 [49]: 48,335 Twitter timelines of celebrities with on average 2,181 tweets per celebrity. The labels gender, year of birth, and occupation were obtained from Wikidata, the degree of fame was derived from the follower count.

As a quick overview, 92 teams registered for the task, 12 showed some sign of activity, e.g., by requesting a virtual machine, and eight made a successful software submission. Performance was measured using *cRank*, the harmonic mean of the macro-averaged multi-class F_1 for gender, fame, occupation, and a leniently calculated F_1 for year of birth. This measure is stricter than average accuracy, since it prefers consistent results, emphasizing performance on classes reflecting rare demographics. The winning submission achieved an outstanding *cRank* of 0.593. Most submissions prefer feature-based machine learning utilizing word-level features over neural approaches, reporting higher performance of the former in preliminary experiments.

After reviewing related work, we give a more detailed description of the task, the construction of the task’s evaluation data, and the reasoning underlying our performance measures in Section 3. In Section 4, we survey the software submissions, in Section 5, we report the evaluation results and carry out an in-depth analysis with regard to 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. [24], Koppel et al. [15], Schler et al. [41], and Argamon et al. [1], focusing on age, gender, and personality from genres with longer, grammatical documents such as blogs and essays. Table 1 overviews most of the works done in author profiling over the past 20 years, reporting on text genre, author count, word count, and the demographics studied. The most commonly used genre in recent years is Twitter tweets, first used in 2011 to predict gender [4] and age [22]. Later work also used Facebook posts [12], Reddit [13], and Sina Weibo [47]. Recently added demographics include education [6], ethnicity [46], family status [47], income [30], occupation [28], location of origin [12], religion [32], and location of residence [6].

At PAN, author profiling has been studied since 2013, covering different demographics including age and gender [36, 35, 39], personality [33], language variety [37], genres including blogs, reviews, and social media posts [39], and cross-domain prediction [34]. Profiling research related to aspects such as behavioral traits [16], medical conditions [7], and native language identification (NLI) have been excluded from our survey, since these have developed into subfields of their own right.

Table 1. Survey of author profiling studies, sorted by year of publication. A '*' indicates an estimation based on an average of 12.7 words per tweet from the reported number of tweets, a '?' indicates unavailable information.

Dataset	Genre	Authors	Words	Demographics
2006				
Schler et al. [41]	Blogs	37,478	7,885	Gender
2007				
Estival et al. [10]	Emails	1,033	3,259	Age, Gender, Education, Native lang., Personality (Big Five), Residence
Estival et al. [11]	Emails	1,033	2,085	Age, Education, Gender, Personality (MBTI)
2011				
Burger et al. [4]	Twitter	183,729	283*	Gender
Nguyen et al. [21]	Blogs	1,997	27,303	Age
Rosenthal and McKeown [40]	Blogs	24,500	(?)	Age
2012				
Bergsma et al. [3]	Publications	4,500	(?)	Gender, Native language
2013				
Ciot et al. [8]	Twitter	8,618	12,700*	Gender
Mikros [19]	Blogs	100	20,323	Gender
Schwartz et al. [42]	Facebook	136,000	4,129	Age, Gender, Personality (NEO-PI-R)
Rangel Pardo et al. [36]	Blogs	346,100	632	Age, Gender
2014				
Rangel Pardo et al. [35]	Multiple	346,100	632	Age, Gender
Verhoeven and Daelemans [44]	Essays	749	976	Age, Birthplace, Gender, Personality (Big Five)
2015				
Kapociute-Dzikiene et al. [14]	Essays	186	286	Age, Gender
Preotiuc-Pietro et al. [28]	Twitter	5,191	26,415*	Occupation (SOC)
Plank and Hovy [26]	Twitter	1,500	12,880	Gender, Personality (MBTI)
Rangel Pardo et al. [33]	Twitter	1,070	1,205	Age, Gender, Personality (Big Five)
Volkova and Bachrach [46]	Twitter	5,000	2,540	Age, Children, Education, Gender, Income, Intelligence, Optimism, Political alignment, Ethnicity, Religion, Relationship, Satisfaction
2016				
Rangel Pardo et al. [39]	Multiple	346,100	632	Age, Gender
Verhoeven et al. [45]	Twitter	18,168	25,400	Gender, Personality (MBTI)
Wang et al. [48]	Sina Weibo	742,323	(?)	Age, Education, Gender, Relationship
2017				
Emmery et al. [9]	Twitter	6,610	31,750*	Gender
Fatima et al. [12]	Facebook	479	2,156	Age, Birthplace, Gender, Education, Extroversion, Nat. lang., Occupation
Litvinova et al. [17]	Essays	500	145	Age, Education, Gender, Personality
Preotiuc-Pietro et al. [29] - D1	Twitter	3,938	15,587*	Age, Gender, Politics
Preotiuc-Pietro et al. [29] - D2	Twitter	13,651	23,717*	Politics
Rangel Pardo et al. [38]	Twitter	19,000	1,195	Dialect, Gender
2018				
Carmona et al. [6]	Twitter	5,000	17,195*	Education, Residence
Gjurkovic and Snajder [13]	Comments	23,503	24,861	Personality (MBTI)
Preotiuc-Pietro and Ungar [30]	Twitter	4,098	16,785*	Age, Education, Gender, Income, Race
Ramos et al. [32]	Facebook	1,019	2,178	Age, Education, Gender, Personality (Big Five), Religion
Rangel Pardo et al. [34]	Twitter	19,000	1,195	Gender
Tighe and Cheng [43]	Twitter	250	31,011*	Personality (Big Five)

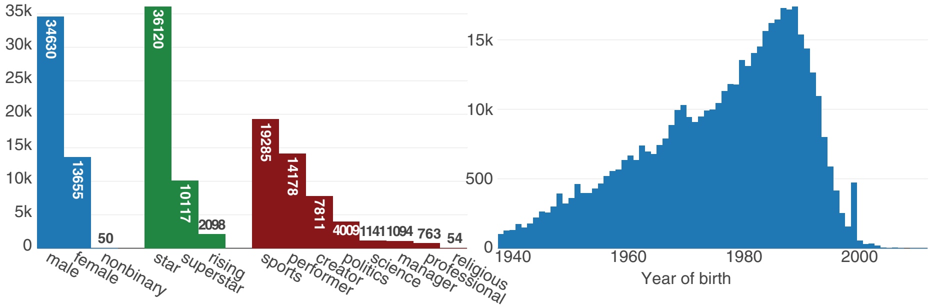


Figure 1. Number of authors in the complete dataset by classes of the demographics: (left) gender in blue, fame in green, occupation in red, and (right) year of birth.

3 Task Description

The task's goal was to evaluate technology to predict the four demographics gender, year of birth, degree of fame, and occupation of a celebrity from their history of tweets on Twitter. Participants were given a large training dataset comprising 33,836 celebrities with up to 3,200 tweets each, and submissions were evaluated on a test dataset comprising 14,499 celebrities using our TIRA evaluation service [27]. Performance was evaluated using a combination of the multi-class F_1-scores of each demographic.

3.1 Evaluation Data

The data used for this task was sampled from the Webis Celebrity Profiling Corpus 2019 [49], which links the Twitter accounts of celebrities with their corresponding Wikidata entries. A celebrity in this corpus is defined as a person who has a verified Twitter account and who is notable as per Wikipedia's notability criteria. Given the list of all verified Twitter accounts, they were heuristically linked to their respective Wikidata entries by matching Twitter's free-form name and the "@"-handle with Wikidata's item name, omitting ambiguous matches, non-person, and memorial accounts. An evaluation of the matching heuristic revealed a very low error rate of 0.6%. In total, the corpus contains 71,706 celebrities as Twitter-Wikidata matches, where Wikidata supplies 239 different demographics, albeit sparsely distributed. For each celebrity, we crawled all available tweets from their timelines,[1] and then filtered out all celebrities who declared a non-English profile language,[2] or were born before 1940, as well as all tweets that did not contain mainly text. Finally, to compile the evaluation data for our task, we sampled all celebrities for the most widely available demographics, namely gender, occupation, and year of birth. Figure 1 shows histograms for each demographic in the sample of the corpus used for this task.[3] Altogether, the evaluation data comprises 48,335 celebrities with an average 2,181 tweets.

[1] Twitter's API limits access to only the latest 3200 tweets. According to the total number of tweets noted in the respective user-profile, this is the complete timeline in 98.05% of cases.

[2] Note that the dataset still contains some bilingual and non-English tweeting celebrities.

[3] The high number of celebrities for year of birth 2000 is an error in Wikidata that we noticed only at the time of writing. We removed them in our subsequent analyses.

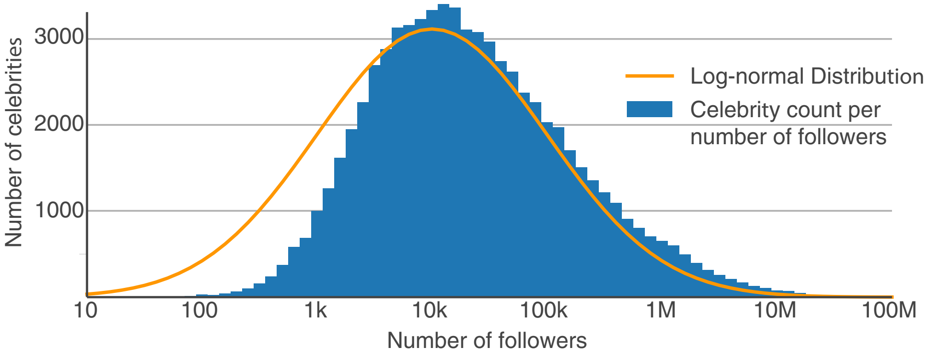


Figure 2. Histogram (blue) of the number of followers each celebrity in the dataset has and a log-normal distribution (yellow) to determine class boundaries of the fame demographic.

Wikidata employs a high number of labels for certain demographics. To render the prediction tasks feasible, we simplified the labels as follows:

- *Gender.* From the eight different gender-related Wikidata labels, we kept *male* and *female* and merged the remaining six to *non-binary*.
- *Fame.* To determine the degree of fame, we calculated the distribution of follower counts shown in Figure 2, overlaid a matching log-normal distribution and used its standard deviation to separate the three classes *rising* (less than 1,000 followers), *star* (more than 1,000 and less than 100,000 followers), and *superstar* (more than 100,000 followers).[4]
- *Occupation.* The 1,379 different occupations were grouped into eight classes:
  1. *Sports* for occupations participating in professional sports, primarily athletes.
  2. *Performer* for creative activities primarily involving a performance like acting, entertainment, TV hosts, and musicians.
  3. *Creator* for creative activities with a focus on creating a work or piece of art, for example, writers, journalists, designers, composers, producers, and architects.
  4. *Politics* for politicians and political advocates, lobbyists, and activists.
  5. *Manager* for executives in companies and organizations.
  6. *Science* for people working in science and education.
  7. *Professional* for specialist professions like cooks and plumbers.
  8. *Religious* for professions in the name of a religion.

  We arrived at these groups of celebrity occupations by reconstructing the graph induced by Wikidata's `subclass of` property, connecting all occupations in the corpus. By manually analyzing the graph, the most reasonable sub-structures of closely connected professions were identified.
- *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 *2012* as extracted from Wikidata's `Day of Birth` property.

[4] We attribute the gap under the left half of the log-normal distribution curve to the fact that rising celebrities are less likely to possess a verified Twitter account, thus missing from our corpus.

The different demographics in the dataset are not entirely independent. While the correlation of some class combinations like year of birth and fame, and gender and fame are insignificant, others have notable dependencies: Figure 4 in Appendix B shows that there is a clear imbalance between gender and occupation, and occupation and year of birth. Female celebrities tend to be younger and more likely have a performing or creator occupation, while male celebrities strongly tend to be famous for sports when young, and politics and religion otherwise. Celebrities working in performing occupations like acting or music tend to be more famous than others.

We split the sampled data 70:30 into a training dataset of 33,836 celebrities and a test dataset of 14,499 celebrities (test dataset 1); from the latter we sub-sampled another small-scale test dataset of 956 authors (test dataset 2).

3.2 Performance Measures

In previous years at PAN, the performance of author profiling approaches has been measured as average of the accuracies measured for each demographic in question. This measure is unfit for celebrity profiling, since the demographics are imbalanced and some have many classes. To measure participant performance, we rather average the per-demographic performance using the harmonic mean, promoting a consistent performance across demographics:

$$\mathrm{cRank} = \frac{4}{\frac{1}{\mathrm{F}_{1,\mathrm{fame}}} + \frac{1}{\mathrm{F}_{1,\mathrm{occupation}}} + \frac{1}{\mathrm{F}_{1,\mathrm{gender}}} + \frac{1}{\mathrm{F}_{1,\mathrm{birthyear}}}}.$$

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 $T \in \{$gender, fame, 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, promoting correct predictions of small classes:

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

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

$$m = (-0.1 \cdot \mathrm{truth} + 202.8).$$

This way of measuring the prediction performance for the age demographics addresses a shortcoming of the fixed-age-interval scheme: Defining strict age intervals (i.e. 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 gets more difficult with increasing age as people grow mature and their writing style presumably changes more slowly over the years.

4 Survey of the Submitted Approaches

Eight participants submitted software to this task, six of whom also submitted notebooks describing their approach. Five of these six approaches are based on traditional feature engineering, and three also report negative experiments with deep learning models, whereas only Pelzer [23] employed a neural language model (ULMFiT). The most popular algorithm choices are logistic regression and support vector machines (SVM), the most popular features are exclusively based on content, whereas only Moreno-Sandoval et al. [20] also added grammatical and custom features. To cope with the small classes in the gender and occupation demographics, two participants resorted to oversampling the classes during training, one to downsampling, and one applied class weighting. Three participants grouped the year of birth into eight maximum-sized intervals and predicting them instead. The most popular preprocessing steps are the replacement or removal of hyperlinks, mentions, hashtags, and emojis, while stop words and punctuation are rarely touched. Below, each approach is described in more detail.

Radivchev et al. [31] uses support vector machines to predict fame and occupation and logistic regression to predict year of birth and gender, using tf-idf vectors of the 10,000 most frequent bigrams of 500 randomly selected tweets per celebrity as features. The authors determined class priors to cope with small classes in gender and occupation prediction and grouped the year of births into eight intervals, reversing the window function used for performance measurement. Tweets are preprocessed by removing retweets and all symbols except letters, numbers, @'s, and #'s, replacing hyperlinks with <url> and mentions with <user>, collapsing spaces, and adding a <sep> token at the end of each tweet. The optimal configuration of learning algorithms for each demographic was determined via grid search over several hyperparameter settings for both the SVM and logistic regression. The authors tried multiple alternative approaches, reporting sub-par results for preserving retweets and replacing emojis with <emoji> during preprocessing, using character 3-grams and 4-grams as features, and employing multi-layered perceptrons or a deep pyramid CNN on GloVe embeddings.

Moreno-Sandoval et al. [20] uses logistic regression to predict fame, gender, and year of birth, and a multinomial naive Bayes model to predict occupation, using n-gram features with a minimum frequency of 9 for gender, 6 for year of birth, 3 for occupation, and none for fame, as well as the features average number of emojis, hashtags, mentions, hyperlinks, retweets, words per tweet, word-length, the lexical diversity, the kurtosis and skew of word-length and word-count, respectively, and the number of tweets written in each of the grammatical genders: the first, second, and third person singular and the first and third person plural. Years of birth are combined into eight larger intervals and oversampled. Preprocessing of texts was done for fame, gender, and year of birth in the form of replacing hashtags, mentions, hyperlinks, and emojis with special tokens. The model configurations described above were obtained by testing several combinations of (1) the five algorithms naive Bayes, Gaussian naive Bayes, naive Bayes complement, logistic regression, and random forest, and (2) whether to apply preprocessing, (3) oversampling, and (4) whether to include the features.

Martinc et al. [18] uses logistic regression for all four demographics, with tf-idf vectors of word unigrams, word-bounded character trigrams, and 4-character suffix trigrams of the first 100 tweets per timeline as features. The suffix trigrams were based on the 10%-80% most frequent words and were weighted with 0.8, the character trigrams 4-80% with 0.4-weighting, and the word unigrams 10-80% with 0.8-weighting. No resampling was applied and all years were predicted without regrouping. The text for both trigram features was preprocessed by replacing hashtags, mentions, and hyperlinks with special tokens and the text for the word unigrams by additionally removing all punctuation and stop words. The authors determined the logistic regression algorithm to be optimal after performing a grid search over different hyperparameter combinations of linear SVMs, SVMs with RBF kernel, logistic regression, random forest, and gradient boosting classifiers. Experiments with BERT-based fine-tuning approaches were reported as non-competitive.

Asif et al. [2] utilizes one model for each combination of the four demographics and the 50 languages the authors detected in the dataset, using the most discriminative words as features. To determine the best learning algorithm for each combination, the authors selected the best-performing one after testing support vector machines, logistic regression, decision trees, Gaussian naive Bayes, random forests, and k-nearest neighbor classifiers. The most discriminative word features for each demographic were determined by aggregating word counts for all users of one class, normalizing these counts by the frequency of the class, and summing the pairwise intra-class distance in relative frequencies. This calculation results in a ranking of words for each demographic, indicating which words are more frequently used by members of one class compared to members of all other classes, where the occurrences of the highest-ranking words were used as features. All tweets are preprocessed by removing hyperlinks, punctuation, stop words, numbers, alphanumeric words, escape characters, #'s, and @'s.

Petrik and Chuda [25] use multiple random forest classifiers with 200 decision trees based on the tf-idf vectors of the top 5000 1-, 2-, and 3-grams. To train the models, the authors used the synthetic minority oversampling technique in combination with Tomek links to balance the examples for each class. The timeline text is preprocessed by removing mentions and stop words, collapsing letter repetitions, and replacing hyperlinks and emojis with special tokens. Additionally, the authors report on experiments with RCNNs, which did not deliver promising results and were hence discarded.

Pelzer [23] applies a transfer learning strategy by training an ULMFiT instance on the celebrity timelines. The classifiers constructed from this instance predicted a class for every tweet in a given timeline and used the majority of all per-tweet predictions to infer the celebrity's demographic. The authors further refined their model by regrouping the year of birth into fewer classes and downsampled the examples of all demographics to get a more balanced training dataset. The author reports on slow prediction times of 8 minutes per celebrity; this approach was only evaluated on the second, small-scale test dataset.

Baselines. Since this is the first edition of the task, we did not resort to providing or reimplementing other unproven models as baselines. Instead, we created three sets of random predictions to compare participant predictions against: (1) *baseline-uniform* randomly draws from a uniform distribution of all classes and reflects the data-agnostic lower bound, (2) *baseline-rand* randomly selects a class according to the prior likelihood of appearance in the test dataset, and (3) *baseline-mv* always predicts the majority class of the test dataset.

Table 2. Results of the celebrity profiling task for both test datasets. Bold font indicates the highest, and underlined font the second-highest value.

(a) Primary metric cRank and minor F_1 scores.

Participant	Test dataset 1					Test dataset 2				
	cRank	gender	age	fame	occup	cRank	gender	age	fame	occup
Radivchev	**0.593**	**0.726**	**0.618**	**0.551**	**0.515**	**0.559**	**0.609**	**0.657**	**0.548**	0.461
Moreno-Sandoval	0.505	0.644	0.518	0.388	0.469	0.497	0.561	0.516	0.518	0.418
Martinc	0.462	0.580	0.361	0.517	0.449	0.465	0.594	0.347	0.507	**0.486**
Fernquist	0.424	0.447	0.339	0.493	0.449	0.413	0.465	0.467	0.482	0.300
Petrik	0.377	0.595	0.255	0.480	0.340	0.441	0.555	0.360	0.526	0.385
Pelzer	–	–	–	–	–	0.499	0.547	0.518	0.460	0.481
Asif	–	–	–	–	–	0.402	0.588	0.254	0.504	0.427
Bryan	–	–	–	–	–	0.231	0.335	0.207	0.289	0.165
baseline-rand	0.223	0.344	0.123	0.341	0.125	–	–	–	–	–
baseline-uniform	0.138	0.266	0.117	0.099	0.152	–	–	–	–	–
baseline-mv	0.136	0.278	0.071	0.285	0.121	–	–	–	–	–

(b) Accuracy.

Participant	Test Dataset 1					Test Dataset 2				
	Mean	Gender	Age	Fame	Occup	Mean	Gender	Age	Fame	Occup
Radivchev	**0.744**	**0.926**	**0.511**	**0.784**	**0.757**	**0.743**	**0.930**	**0.517**	0.770	0.757
Moreno-Sandoval	0.614	0.863	0.365	0.543	0.684	0.627	0.861	0.376	0.547	0.722
Martinc	0.710	0.897	0.457	0.756	0.732	0.712	0.915	0.448	0.753	0.733
Fernquist	0.661	0.763	0.469	0.770	0.644	0.666	0.784	0.466	**0.776**	0.640
Petrik	0.616	0.914	0.081	0.767	0.703	0.597	0.852	0.345	0.529	0.661
Pelzer	–	–	–	–	–	0.622	0.862	0.364	0.556	0.704
Asif	–	–	–	–	–	0.696	0.905	0.346	**0.776**	**0.758**
Bryan	–	–	–	–	–	0.515	0.722	0.173	0.763	0.402
baseline-rand	0.419	0.586	0.233	0.577	0.279	–	–	–	–	–
baseline-uniform	0.179	0.336	0.153	0.105	0.122	–	–	–	–	–
baseline-mv	0.542	0.717	0.298	0.751	0.400	–	–	–	–	–

5 Evaluation of the Results

Table 2a shows the performance of the eight participants who submitted a software to the celebrity profiling task, ranked by the cRank score. The winning approach by Radivchev et al. [31] achieves 0.593 on the first and 0.559 on the second test dataset, closely followed by Moreno-Sandoval et al. [20] with 0.505 on the first, and Pelzer [23] with 0.499 on the second test dataset. All submitted approaches beat the baselines, most by a significant margin. The performance measured for our two test datasets is quite similar, comparing participants who submitted runs for both. The scores are less varied on the second test dataset: The leading participant's performance is lower, and Petrik and Chuda's approach improves slightly, overtaking that of Fernquist as fourth in the ranking. These differences can be attributed to the smaller size of the second test

Table 3. F1-scores on the first test dataset for each demographic individually.

Participant	Gender			Fame		
	female	male	nonbinary	star	superstar	rising
Radivchev	**0.881**	**0.951**	**0.307**	**0.874**	0.469	**0.261**
Moreno-Sandoval	0.820	0.909	0.260	0.488	**0.662**	0
Martinc	0.816	0.933	0	0.853	0.425	0.160
Petrik	0.858	0.947	0	0.867	0.329	0.093
Fernquist	0.405	0.853	0	0.868	0.264	0.079
baseline-rand	0.277	0.712	0	0.752	0.070	0.063
baseline-uniform	0.299	0.466	0	0	0.289	0
baseline-mv	0	0.835	0	0.863	0	0

Participant	Occupation							
	performer	creator	sports	manager	politics	science	professional	religious
Radivchev	**0.785**	**0.572**	**0.895**	**0.229**	**0.740**	**0.320**	**0.214**	0.272
Moreno-Sandoval	0.798	0.429	0.860	0.226	0.661	0.318	0.178	**0.363**
Martinc	0.740	0.501	0.872	0.152	0.731	0.271	0.100	0
Petrik	0.645	0.411	0.783	0.006	0.622	0	0	0
Fernquist	0.736	0.445	0.868	0	0.698	0	0	0
baseline-rand	0.299	0.175	0.398	0.007	0.072	0.017	0.008	0
baseline-uniform	0.167	0.139	0.191	0.035	0.084	0.042	0.035	0
baseline-mv	0	0	0.577	0	0	0	0	0

dataset and less to the fact that the second dataset contains exclusively English tweets. To verify this claim, we compiled an English-only version of the first test dataset, and the results shown Table 4, Appendix A, are nearly identical for all participants.

Table 2b shows the accuracies for all submitted approaches, allowing for a comparison of the general, unweighted correctness of class predictions with the cRank measure. Accuracies are generally higher for all participants, a natural consequence of the imbalanced dataset and the existence of small classes. This can be seen by comparing the results of the baseline-mv, which is almost competitive under accuracy but irrelevant under cRank. The differences in the per-demographic performance can be explained further by inspecting the class-wise F_1 shown in Table 3. An important observation is that the top three approaches succeed more frequently in predicting small classes correctly, greatly benefiting cRank without notably impacting accuracy. We assume that the good performance on small classes is due to downsampling and the class weighting applied by the top two approaches, whereas models without these strategies mostly fit toward the majority classes. Overfitting toward the majority class is also the likely explanation for the difference in ranking between accuracy and cRank.

Gender. Predicting the binary sex of an author is a widely studied benchmark task for author profiling approaches. All participants achieved a respectable accuracy in predicting celebrity gender, frequently surpassing 0.9 accuracy, while F_1 scores are near the 0.6-0.7 range. Table 3 and Table 5, Appendix A, show the class-wise F_1 for all demographics, which explain the achieved performance values on gender prediction: The best approaches are best at predicting non-binary gender, while binary gender classification is close to fit for practical use. Interestingly, the averaged confusion matrix for gender in Figure 3 shows that non-binary celebrities are mostly misclassified as female, which can not be explained by imbalanced data and thus justifies further research.

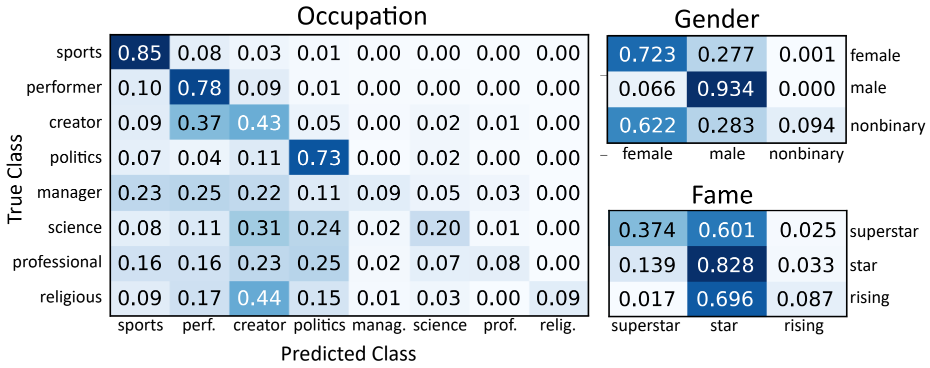
Year of Birth. Our approach to age prediction, departing from fixed-size intervals to a lenient evaluation of year of birth prediction, notably impacts participant performance. Some models reduce the difficulty of the task by reconstructing intervals and using classification algorithms with notably better performance than the alternatively used strategy of predicting each year individually. No submission tries to solve the prediction with regression algorithms. The confusion matrices for the winning model exemplified in Figure 5, Appendix B, illustrate the difficulty of predicting the year of birth, with that approach especially struggling to separate celebrities born before the 1980s. This is a well-known difficulty and has been addressed in our evaluation by being more forgiving on older celebrities.

Fame. The degree of fame is a particularly imbalanced class, reflected in the accuracy where only four participants could beat the *baseline-mv* on the first test dataset and only three on the second. On the contrary, participants are much better at separating classes correctly as shown by their F_1 scores, although there is a trend toward the majority class as can be seen by the confusion matrix in Figure 3. We cannot claim that this task is solved but we have shown that both the most and least famous celebrities can partially be distinguished by their writing.

Occupation. As with the other demographics, occupation was predicted far better than the baselines by all participants and the results were highly influenced by the performance on small classes, although not exclusively. All models work better on occupations with a clear topic, like performer containing actors and musicians, sports, and politics. For occupations that cover multiple topics, like creator, manager, professional, and science, all models are rather weak while still beating the baselines. Ignoring the trend toward majority classes, the confusion matrix for the winning approach in Figure 5, Appendix B, and the averaged one in Figure 3 both show that science is frequently confused with politics and creator, religious with creator, and creator with performer.

5.1 Discussion

In general, all submitted approaches work better, the more examples there are, and the more clearly classes can be separated by topic. The final ranking was influenced the most by the resampling strategies to avoid fitting to majority classes and the addition of grammatical or stylistic features to avoid the misclassification of occupations without a coherent common topic. From a classification perspective, we see the most potential for improvement in using all available text data to build celebrity representations, instead of just excepts, but still excel at finding small classes, for example, using few-shot models like prototypical or highway networks. From an author profiling perspective, much is still unclear about the expression of fame, non-binary gender, and non-topical occupation groups. The best algorithms can partially separate these demographics and



	sports	perf.	creator	politics	manag.	science	prof.	relig.
sports	0.85	0.08	0.03	0.01	0.00	0.00	0.00	0.00
performer	0.10	0.78	0.09	0.01	0.00	0.00	0.00	0.00
creator	0.09	0.37	0.43	0.05	0.00	0.02	0.01	0.00
politics	0.07	0.04	0.11	0.73	0.00	0.02	0.00	0.00
manager	0.23	0.25	0.22	0.11	0.09	0.05	0.03	0.00
science	0.08	0.11	0.31	0.24	0.02	0.20	0.01	0.00
professional	0.16	0.16	0.23	0.25	0.02	0.07	0.08	0.00
religious	0.09	0.17	0.44	0.15	0.01	0.03	0.00	0.09

female	male	nonbinary	
0.723	0.277	0.001	female
0.066	0.934	0.000	male
0.622	0.283	0.094	nonbinary

superstar	star	rising	
0.374	0.601	0.025	superstar
0.139	0.828	0.033	star
0.017	0.696	0.087	rising

Figure 3. Averaged normalized confusion matrices for gender, fame, and occupation prediction of the top 5 approaches by cRank.

errors are systematical rather than random, but a more fundamental understanding of differences in writing is necessary to make progress.

Although we are satisfied with the results of the celebrity profiling task and the insights gained, we see some opportunities to refine our task setup. For the next iteration, we will consider narrowing the range of years of birth to 1940-2000, omitting occupations religious and professional, and revising the fame boundaries. The existence of several small classes turned out to be the major challenge of this task. We see this as an important aspect of author profiling and especially forensics, since correctly identifying rare demographics is most desirable in practice. A certain degree of class imbalance is hence necessary, albeit the degree of imbalance in all four demographics affected a reliable evaluation and prevented participants from focusing on small classes in particular. To improve the general robustness and ease of use of our dataset, we will remove all non-English tweets and celebrities supplying too little text.

Besides the prediction of small classes, year of birth prediction has been a major factor influencing algorithm performance. The intention behind our approach was to overcome the inherent weakness of interval-based age prediction and to provide an incentive to participants to develop more fine-grained predictions. This was not picked up by participants, since participants simply defined their own interval-based classification based on our scoring formula. For the next iteration, we will consider a distance-based performance scoring for year of birth prediction.

6 Conclusion and Outlook

In the celebrity profiling task at PAN 2019, we invited participants to predict the demographics gender, year of birth, fame, and occupation of 48,335 twitter timelines of celebrities. Eight participants submitted models and six submitted notebooks describing their approach. Participants found traditional machine learning on content-based features to be most reliable, where the best-performing models added some style-based features and resampled the training examples to compensate class imbalance. Although a lot of progress has been made in this task, several open challenges remain: (1) a reliable prediction of rare demographics, like non-binary gender, very young celebrities born after 2000, and "rising" stars, (2) the prediction of occupations without clear topical separation, like professional, manager, science, and creator, and (3) the discrimination of authors born before 1980.

Acknowledgments

We want to thank our participants for their effort and dedication and the CLEF organizers for hosting PAN and the celebrity profiling task.

Bibliography

[1] Argamon, S., Koppel, M., Pennebaker, J., Schler, J.: Automatically Profiling the Author of an Anonymous Text. Commun. ACM 52(2), 119–123 (Feb 2009)

[2] Asif, M., Shahzad, N., Ramzan, Z., Najib, F.: Word Distance Approach for Celebrity profiling—Notebook for PAN at CLEF 2019. In: [5]

[3] Bergsma, S., Post, M., Yarowsky, D.: Stylometric analysis of scientific articles. In: HLT-NAACL. pp. 327–337. The Association for Computational Linguistics (2012)

[4] 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)

[5] Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.): CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, CEUR-WS.org (Sep 2019)

[6] 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)

[7] Choudhury, M.D., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: ICWSM. The AAAI Press (2013)

[8] Ciot, M., Sonderegger, M., Ruths, D.: Gender inference of twitter users in non-english contexts. In: EMNLP. pp. 1136–1145. ACL (2013)

[9] Emmery, C., Chrupala, G., Daelemans, W.: Simple queries as distant labels for predicting gender on twitter. In: NUT@EMNLP. pp. 50–55. Association for Computational Linguistics (2017)

[10] Estival, D., Gaustad, T., Pham, S., Radford, W., Hutchinson, B.: Author profiling for english emails (12 2007)

[11] Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: TAT: an author profiling tool with application to arabic emails. In: ALTA. pp. 21–30. Australasian Language Technology Association (2007)

[12] Fatima, M., Hasan, K., Anwar, S., Nawab, R.M.A.: Multilingual author profiling on facebook. Inf. Process. Manage. 53(4), 886–904 (2017)

[13] Gjurkovic, M., Snajder, J.: Reddit: A gold mine for personality prediction. In: PEOPLES@NAACL-HTL. pp. 87–97. Association for Computational Linguistics (2018)

[14] Kapociute-Dzikiene, J., Utka, A., Sarkute, L.: Authorship attribution and author profiling of lithuanian literary texts. In: BSNLP@RANLP. pp. 96–105. INCOMA Ltd. Shoumen, BULGARIA (2015)

[15] Koppel, M., Argamon, S., Shimoni, A.: Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing 17(4), 401–412 (2002)

[16] Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated corpus of hindi-english code-mixed data. In: LREC. European Language Resources Association (ELRA) (2018)

[17] Litvinova, T., Seredin, P., Litvinova, O., Zagorovskaya, O.: 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. pp. 69–73. Association for Computational Linguistics (2017)

[18] Martinc, M., Škrlj, B., Pollak, S.: Who is hot and who is not? Profiling celebs on Twitter—Notebook for PAN at CLEF 2019. In: [5]

[19] Mikros, G.: Authorship Attribution and Gender Identification in Greek Blogs. In: Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO). pp. 21–32 (2013)

[20] Moreno-Sandoval, L., Puertas, E., Plaza-del-Arco, F., Pomares-Quimbaya, A., Alvarado-Valencia, J., Ureña-Lòpez, L.: Celebrity Profiling on Twitter using Sociolinguistic Features—Notebook for PAN at CLEF 2019. In: [5]

[21] Nguyen, D., Smith, N., Rosé, C.: 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. pp. 115–123. ACM (2011)

[22] 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), http://doi.acm.org/10.1145/2065023.2065035

[23] Pelzer, B.: Celebrity Profiling with Transfer Learning—Notebook for PAN at CLEF 2019. In: [5]

[24] Pennebaker, J., Mehl, M., Niederhoffer, K.: Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology 54, 547–577 (2003)

[25] Petrik, J., Chuda, D.: Twitter feeds profiling with TF-IDF—Notebook for PAN at CLEF 2019. In: [5]

[26] Plank, B., Hovy, D.: Personality traits on twitter - or - how to get 1, 500 personality tests in a week. In: WASSA@EMNLP. pp. 92–98. The Association for Computer Linguistics (2015)

[27] 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 - Lessons Learned from 20 Years of CLEF. Springer (2019)

[28] 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)

[29] Preotiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.H.: Beyond binary labels: Political ideology prediction of twitter users. In: ACL (1). pp. 729–740. Association for Computational Linguistics (2017)

[30] 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), https://aclanthology.info/papers/C18-1130/c18-1130

[31] Radivchev, V., Nikolov, A., Lambova, A.: Celebrity Profiling using TF-IDF, Logistic Regression, and SVM—Notebook for PAN at CLEF 2019. In: [5]

[32] 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)

[33] Rangel Pardo, 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)

[34] Rangel Pardo, 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), http://ceur-ws.org/Vol-2125/

[35] Rangel Pardo, 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.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK. CEUR Workshop Proceedings, CEUR-WS.org (Sep 2014)

[36] Rangel Pardo, 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)

[37] Rangel Pardo, 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), http://ceur-ws.org/Vol-1866/

[38] Rangel Pardo, 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.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, vol. 1866. CLEF and CEUR-WS.org (Sep 2017), http://ceur-ws.org/Vol-1866/

[39] Rangel Pardo, 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), http://ceur-ws.org/Vol-1609/16090750.pdf

[40] Rosenthal, S., McKeown, K.R.: Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In: ACL. pp. 763–772. The Association for Computer Linguistics (2011)

[41] 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)

[42] Schwartz, H., Eichstaedt, J., Kern, M., Dziurzynski, L., Ramones, S., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M., Ungar, L.: Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. In: PLoS ONE. p. 8(9): e73791 (2013)

[43] Tighe, E.P., Cheng, C.K.: Modeling personality traits of filipino twitter users. In: PEOPLES@NAACL-HTL. pp. 112–122. Association for Computational Linguistics (2018)

[44] Verhoeven, B., Daelemans, W.: Clips stylometry investigation (CSI) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno,

A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014. pp. 3081–3085. European Language Resources Association (ELRA) (2014)

[45] Verhoeven, B., Daelemans, W., Plank, B.: Twisty: A multilingual twitter stylometry corpus for gender and personality profiling. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016. European Language Resources Association (ELRA) (2016)

[46] 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)

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

[48] Wang, Y., Xiao, Y., Ma, C., Xiao, Z.: Improving users' demographic prediction via the videos they talk about. In: EMNLP. pp. 1359–1368. The Association for Computational Linguistics (2016)

[49] Wiegmann, M., Stein, B., Potthast, M.: Celebrity Profiling. In: Proceedings of ACL 2019 (to appear) (2019)

A Tables

Table 4. Absolute difference in accuracy between the first test dataset and its exclusively English subset.

Participant	Mean	Gender	Age	Fame	Occup
Radivchev	0,07	0,04	0,07	0,08	0,06
Moreno-Sandoval	0,02	0,01	0,01	0,02	0,02
Martinc	0,04	0,04	0,03	0,06	0,03
Fernquist	0,04	0,01	0,04	0,05	0,03
Petrik	0,07	0,08	0,02	0,09	0,11

Table 5. F1-scores on the second test dataset for each demographic individually.

Participant	**Gender**			**Fame**		
	female	male	nonbinary	star	superstar	rising
Radivchev	**0.874**	**0.952**	0	0.858	0.396	0.350
Moreno-Sandoval	0.772	0.902	0	0.641	**0.466**	0.246
Martinc	0.835	0.943	0	0.848	0.383	0.178
Fernquist	0.449	0.866	0	0.869	0.258	0.111
Petrik	0.759	0.894	0	0.620	0.434	0.292
Pelzer	0.732	0.906	0	0.691	0.369	0.144
Asif	0.825	0.937	0	**0.870**	0.189	0.120
Bryan	0.014	0.838	0	0.865	0	0

Participant	**Occupation**							
	performer	creator	sports	manager	politics	science	professional	religious
Radivchev	0.763	**0.527**	**0.900**	0.250	**0.756**	0.150	**0.200**	0
Moreno-Sandoval	0.740	0.417	0.893	0.242	0.715	0.190	0.080	0
Martinc	0.730	0.470	0.869	**0.300**	0.736	0.142	**0.200**	0
Fernquist	0.617	0.362	0.785	0	0.632	0	0	0
Petrik	0.708	0.344	0.854	0.086	0.700	0.142	0.160	0
Pelzer	0.764	0.376	0.874	0.148	0.717	**0.246**	0.105	0
Asif	**0.776**	0.481	0.884	0	0.773	0.095	0	0
Bryan	0.318	0.108	0.550	0	0.218	0	0	0

B Figures

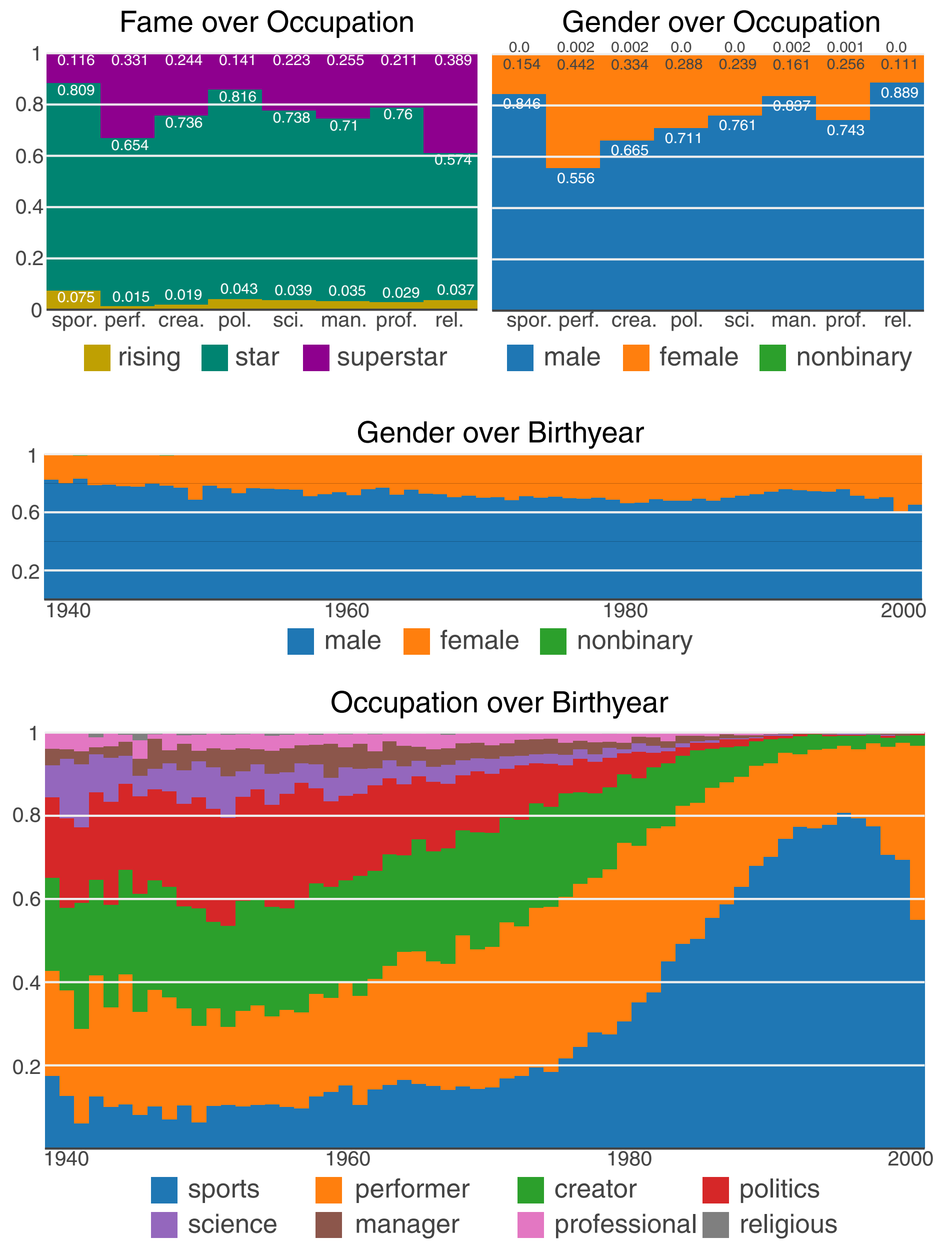


Figure 4. Co-occurrence likelihood of different occupations with (left) different degrees of fame and (right) different genders.

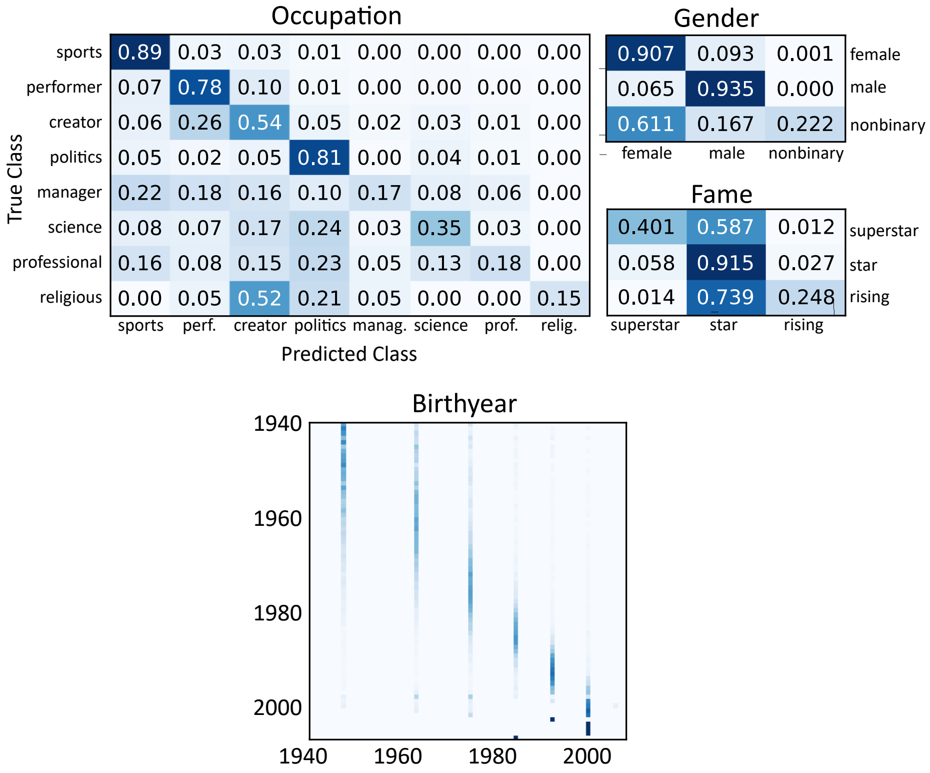


Figure 5. Confusion matrices of the winning approach for all four traits.