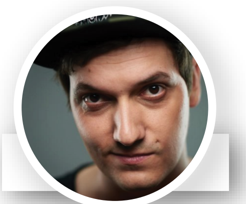
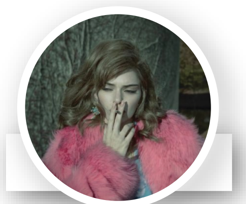
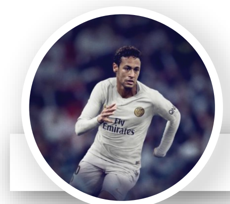
Overview of the Celebrity Profiling Task at PAN 2020



LeFloid
@LeFloid



Kendall
@KendallJenner



Neymar Jr
@nejmarjr

Lil Wayne WEEZY F
@LilTunechi

Matti Wiegmann, Benno Stein, Martin Potthast
Bauhaus-Universität Weimar
webis.de

Celebrity Profiling

Motivation

Celebrity Profiling 2020:

Given the Twitter feeds of the followers of a celebrity, determine the demographics.

Celebrity Profiling

Motivation

Celebrity Profiling 2019:

Given the Twitter feeds ~~of the followers~~ of a celebrity, determine the demographics.

Why Celebrities?

- They write many public, high-quality texts.
- Many personal demographics are public knowledge.

Celebrity Profiling

Motivation

Celebrity Profiling 2019:

Given the Twitter feeds ~~of the followers~~ of a celebrity, determine the demographics.

Why Celebrities?

- They write many public, high-quality texts.
- Many personal demographics are public knowledge.

➜ This is not the case for many users on social media.

Celebrity Profiling

Motivation

Celebrity Profiling 2020:
Given the (?) of a celebrity, determine the demographics.

How can we profile users that do not write a lot?

Celebrity Profiling
Motivation

Celebrity Profiling 2020:
Given the Twitter profile of a celebrity, determine the demographics.

How can we profile users that do not write a lot?

- Author Metadata: Biography, profile picture, ...

Celebrity Profiling
Motivation

Celebrity Profiling 2020:
Given the behavior on Twitter of a celebrity, determine the demographics.

How can we profile users that do not write a lot?

- ~~Author Metadata: Biography, profile picture, ...~~
- Author Behavior: Retweets, Likes, ...

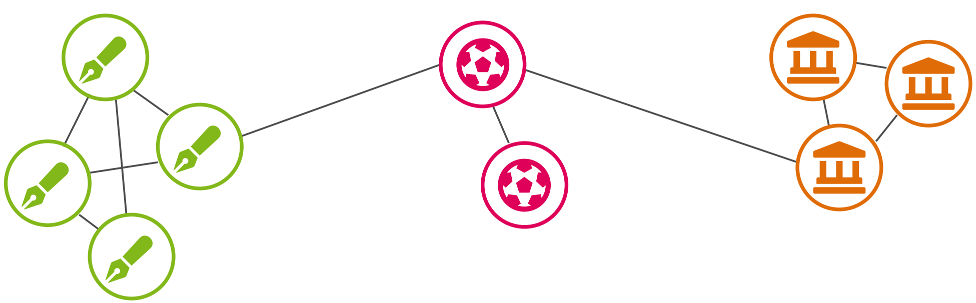
Celebrity Profiling

Motivation

Celebrity Profiling 2020:
Given the Twitter feeds of the followers of a celebrity, determine the demographics.

How can we profile users that do not write a lot?

- ~~Author Metadata: Biography, profile picture, ...~~
- ~~Author Behavior: Retweets, Likes, ...~~
- Social Graph: Homophily.



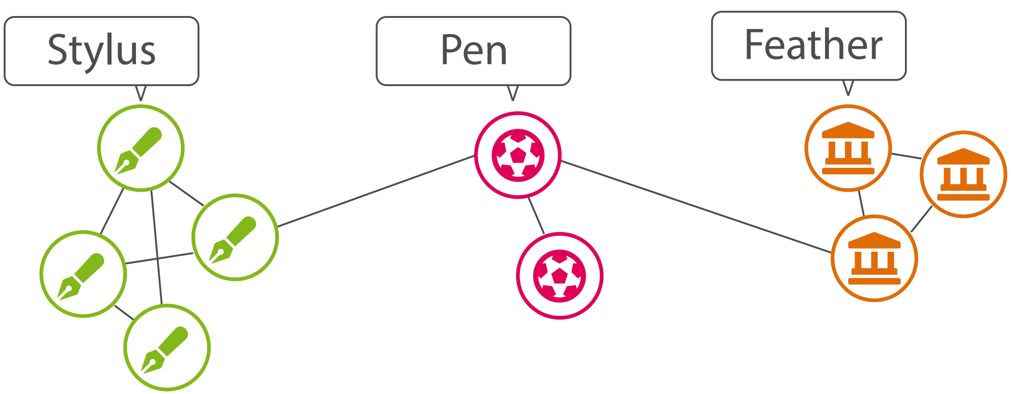
Celebrity Profiling

Motivation

Celebrity Profiling 2020:
Given the Twitter feeds of the followers of a celebrity, determine the demographics.

How can we profile users that do not write a lot?

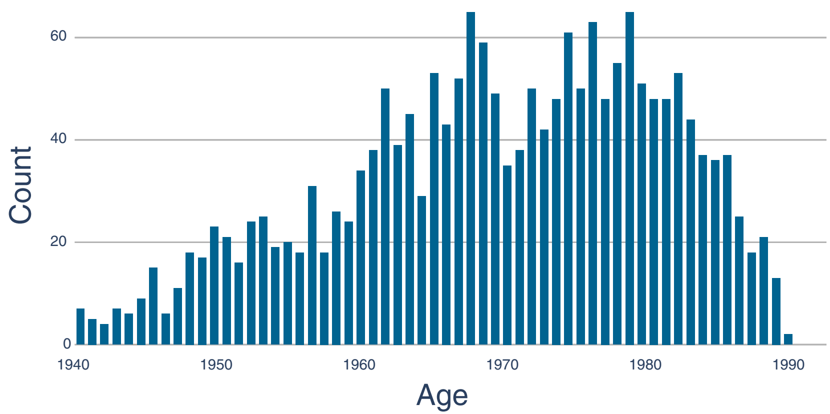
- ~~Author Metadata: Biography, profile picture, ...~~
- ~~Author Behavior: Retweets, Likes, ...~~
- Social Graph: Homophily and language variation.



Celebrity Profiling
Task

Celebrity Profiling 2020:
Given the Twitter feeds of the followers of a celebrity, determine the demographics:

- **Age**,



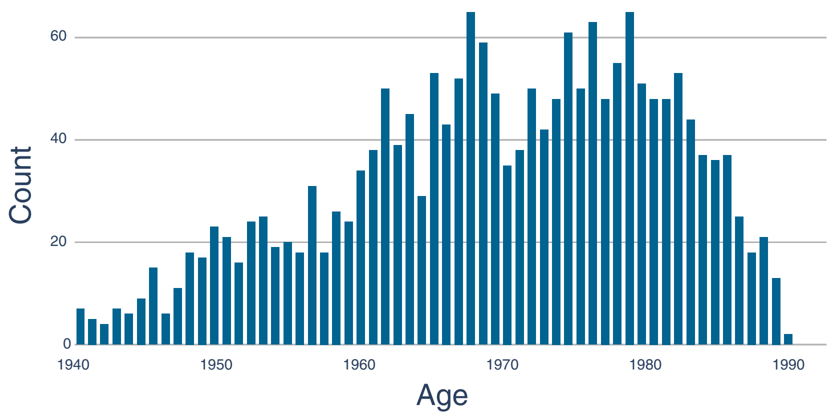
Celebrity Profiling

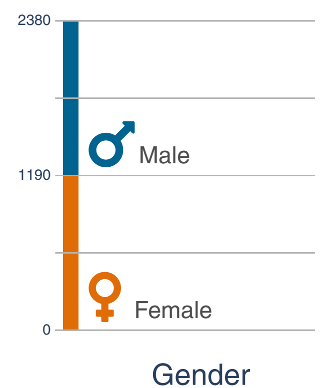
Task

Celebrity Profiling 2020:

Given the Twitter feeds of the followers of a celebrity, determine the demographics:

- **Age**,
- **Gender**,

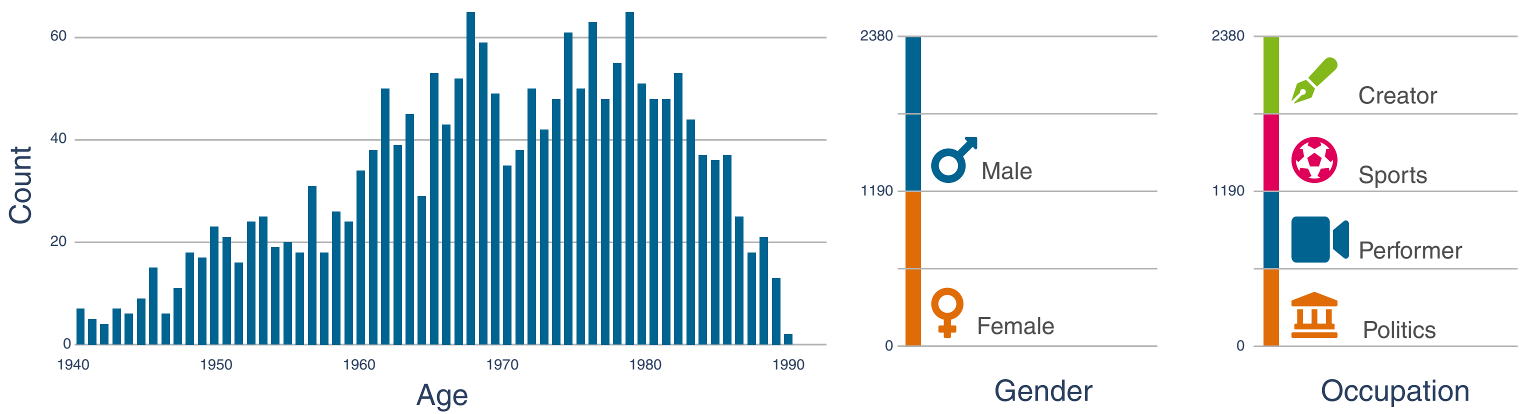




Celebrity Profiling
Task

Celebrity Profiling 2020:
Given the Twitter feeds of the followers of a celebrity, determine the demographics:

- ❑ **Age**,
- ❑ **Gender**, and
- ❑ **Occupation**.

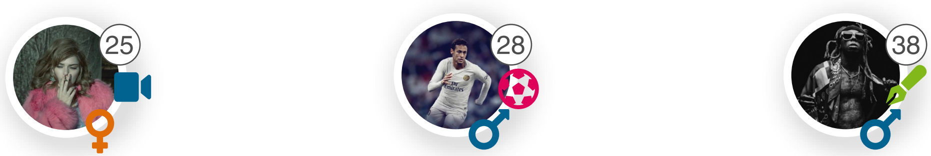


Celebrity Profiling

Data

Dataset creation:

1. Extract celebrities with matching profiles from a Corpus [ACL 2019].

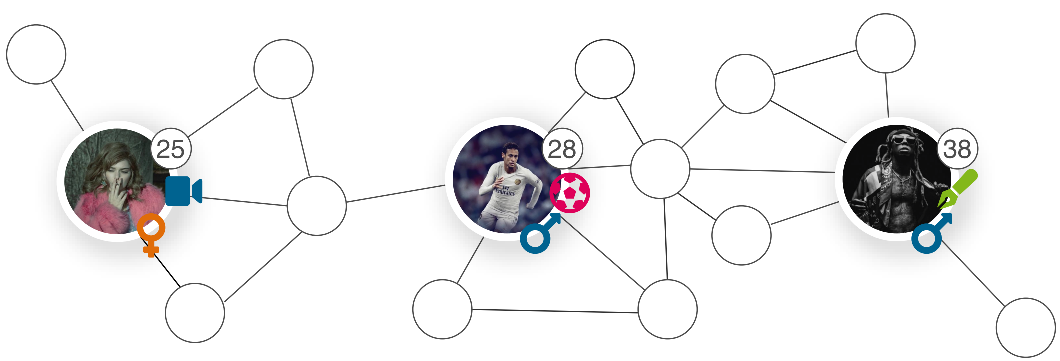


Celebrity Profiling

Data

Dataset creation:

1. Extract celebrities with matching profiles from a Corpus [ACL 2019].
2. Download follower network.

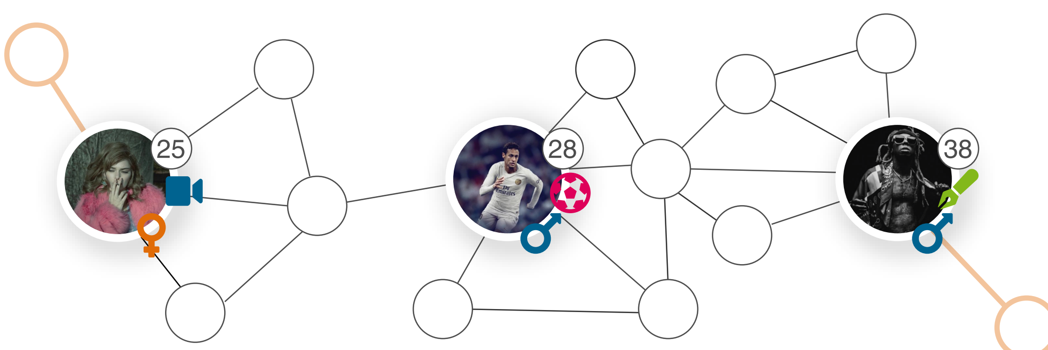


Celebrity Profiling

Data

Dataset creation:

1. Extract celebrities with matching profiles from a Corpus [ACL 2019].
2. Download follower network.
3. Eliminate inactive users.
   - Users with few connections in the network.

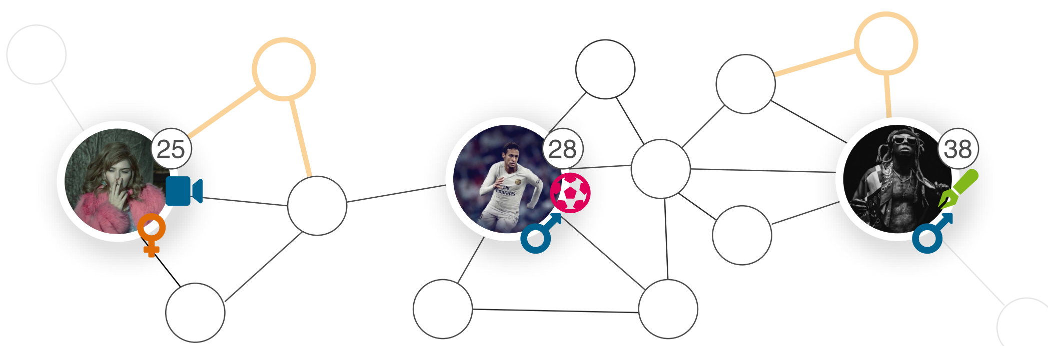


Celebrity Profiling

Data

Dataset creation:

1. Extract celebrities with matching profiles from a Corpus [ACL 2019].
2. Download follower network.
3. Eliminate inactive users, passive users.
   - Users with less than 100 original, English tweets.

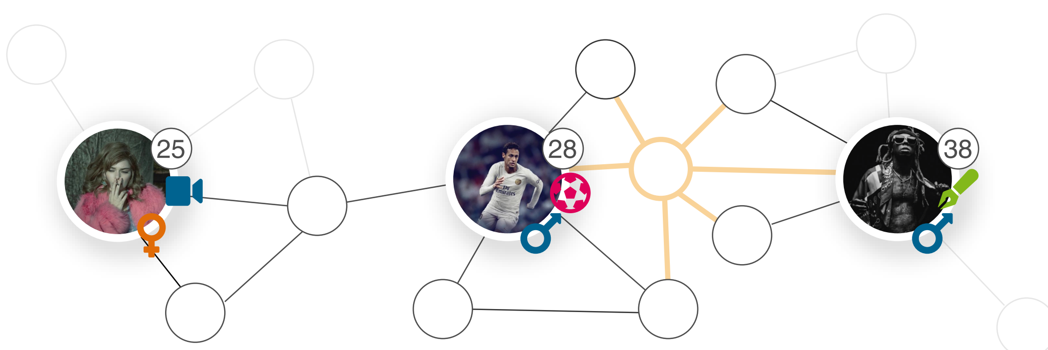


Celebrity Profiling

Data

Dataset creation:

1. Extract celebrities with matching profiles from a Corpus [ACL 2019].
2. Download follower network.
3. Eliminate inactive users, passive users, and other hub users.
   - Users with many followers or atypical behavior.

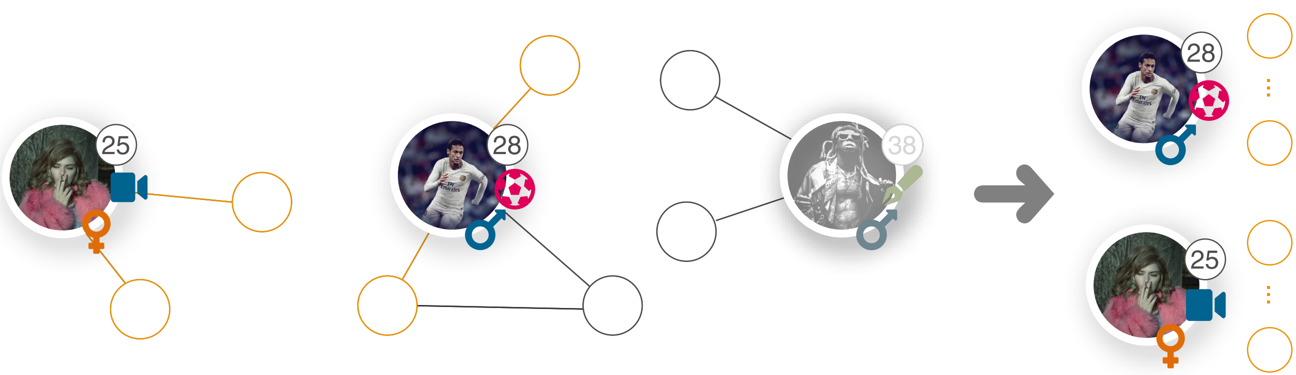


Celebrity Profiling

Data

Dataset creation:

1. Extract celebrities with matching profiles from a Corpus [ACL 2019].
2. Download follower network.
3. Eliminate inactive users, passive users, and other hub users.
4. Sample 10 followers per celebrity in a balanced dataset.
   - **Training dataset**: 1,980 celebrities.
   - **Test dataset**: 400 celebrities.



Celebrity Profiling

Evaluation

Performance is measured as the harmonic mean of the classwise averaged F_1.

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

Celebrity Profiling

Evaluation

Performance is measured as the harmonic mean of the classwise averaged F_1.

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

Variable-bucketed age evaluation:

- Predict author age directly.
- Count near-misses as correct, depending on the age of the author.
- Apply multi-class evaluation.

Celebrity Profiling

Results

Baseline:

- Algorithm: Logistic regression.
- Features: Bags of word 1 and 2-grams, TD-IDF weighted.
- Age was predicted in 5 classes: 1947, 1963, 1975, 1985, and 1994.

Celebrity Profiling

Results

Baseline:

- Algorithm: Logistic regression.
- Features: Bags of word 1 and 2-grams, TD-IDF weighted.
- Age was predicted in 5 classes: 1947, 1963, 1975, 1985, and 1994.

Trained and tested on all followers' tweets as a lower bound.

Participant	Test dataset			
	cRank	Age	Gender	Occupation
baseline-follower	0.47			

Celebrity Profiling

Results

Baseline:

- Algorithm: Logistic regression.
- Features: Bags of word 1 and 2-grams, TD-IDF weighted.
- Age was predicted in 5 classes: 1947, 1963, 1975, 1985, and 1994.

Trained and tested on all followers' tweets as a lower bound.

Trained and tested on the celebrities' tweets as a goalpost.

Participant	Test dataset			
	cRank	Age	Gender	Occupation
baseline-oracle	0.63			
baseline-follower	0.47			

Celebrity Profiling

Results

As proof of concept: Profiling users from their followers' texts works.

- Baseline was beaten by a healty margin.

Participant	Test dataset			
	cRank	Age	Gender	Occupation
baseline-oracle	0.63			
Hodge and Price	0.58			
Koloski et al.	0.52			
Alroobaea et al.	0.47			
baseline-follower	0.47			

Celebrity Profiling

Results

As proof of concept: Profiling users from their followers' texts works.

- Baseline was beaten by a healty margin.
- Submissions predict young users (20-30) better by .2 F_1.

Participant	Test dataset			
	cRank	Age	Gender	Occupation
baseline-oracle	0.63	0.50		
Hodge and Price	0.58	0.43		
Koloski et al.	0.52	0.41		
Alroobaea et al.	0.47	0.32		
baseline-follower	0.47	0.36		

Celebrity Profiling
Results

As proof of concept: Profiling users from their followers' texts works.

- Baseline was beaten by a healty margin.
- Submissions predict young users (20-30) better by .2 F_1.
- Submissions skew towards the "Male" class.

Participant	**Test dataset**			
	cRank	Age	Gender	Occupation
baseline-oracle	0.63	0.50	0.75	
Hodge and Price	0.58	0.43	0.68	
Koloski et al.	0.52	0.41	0.62	
Alroobaea et al.	0.47	0.32	0.70	
baseline-follower	0.47	0.36	0.58	

Celebrity Profiling

Results

As proof of concept: Profiling users from their followers' texts works.

- Baseline was beaten by a healty margin.
- Submissions predict young users (20-30) better by .2 F_1.
- Submissions skew towards the "Male" class.
- Submissions beat the oracle on occupation, although "Creators" is a problematic class (.46 F_1).

Participant	**Test dataset**			
	cRank	Age	Gender	Occupation
baseline-oracle	0.63	0.50	0.75	0.70
Hodge and Price	0.58	0.43	0.68	0.71
Koloski et al.	0.52	0.41	0.62	0.60
Alroobaea et al.	0.47	0.32	0.70	0.60
baseline-follower	0.47	0.36	0.58	0.52

Celebrity Profiling

Outlook

We still have many open questions:

- Does the communities' text reflect the demographics of a celebrity?

Celebrity Profiling

Outlook

We still have many open questions:

- Does the communities' text reflect the demographics of a celebrity?
- Do celebrities influence the writing of their fans?
- What are the rules of style formation?

See you at CLEF 2021!