Celebrity Profiling

What this is about

Celebrity profiling is author profiling applied to celebrities.

- They are **prolific** social media users, supplying **lots of writing samples**.
- Lots of **personal details** are public knowledge.
- They build a **consistent public persona**, either themselves or with the help of agents.
- A number of **demographics** apply only to this population.

Celebrities are a great population to study!

Contributions

1. We built a large **corpus of celebrity profiles** by matching Twitter usernames with Wikidata items.
2. We **compared** profiling celebrities with the SOTA on the general population and held a competition on profiling. With this, we showed where celebrity profiling works and where it does not.
3. We obtained some insights into **celebrities on Twitter** by analyzing our corpus.

DATA AND CODE

https://github.com/webis-de/ACL-19

https://pan.webis.de

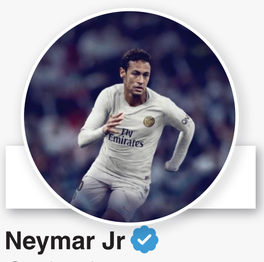
Corpus Construction

1. Find authors

We started with a list of all 297,878 verified Twitter accounts.

Kendall
@KendallJenner

LeFloid
@LeFloid



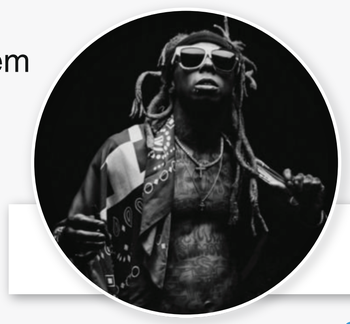
Neymar Jr
@nejmarjr

2. Link to Wikidata

We then created several candidate names, resolved them via Wikipedia in different languages and acquired the corresponding Wikidata items.

Name candidates		
Clean display name	✗	Lil Wayne Weezy F
Split @-reference	✗	Lil Tunechi
Remove middle-name	✗	Lil F
Remove last names	✓	Lil Wayne

↓ Get Wikidata item



Lil Wayne WEEZY F
@LilTunechi

Lil Wayne (Q15615)

American rapper, singer, record executive and businessman
Dwayne Michael Carter, Jr. | Weezy | Weezy F Baby | jalm

3. Verify matches

We removed matches of non-human and memorial accounts, detected errors, and ambiguous accounts.

Coca-Cola (Q2813)

✗ instance of: **patent medicine**
➜ not a human

Jimi Hendrix (Q5928)

✓ instance of: **human**
✓ Twitter username: **JimiHendrix**
✗ date of death: **18 September 1970**
✗ place of death: **Kensington**
✗ manner of death: **accident**
✗ cause of death: **barbiturate overdose**
➜ memorial

Jimi Hendrix
@JimiHendrix

4. Evaluate matching

For evaluation we reversed the procedure: acquired verified tweeters from Wikidata and counted misses and errors.

89,451	Wikidata items with a Twitter ID	
28,454	of those are verified	
20,579	we got right	recall: **0.723**
7,751	we missed	
124	we got wrong	precision: **0.994**

Final corpus

Celebrities	71,706
Average № words	29,968
Languages	37 77% English

Demographics	**Top Attribute**
90.1% Sex	71.7% Male
87.9% Occupation	15.3% Actor
84.4% Date of birth	–
39.2% Educated at	2.1% Harvard
16.9% Languages spoken	54.9% English
9.4% Political party	16.4% Republican
0.5% Race	66.5% African Am.
0.4% Religion	23.5% Islam

Corpus overview with selected demographics.

Experiment Results

Can we profile celebrities like other authors?

We profiled the **benchmark demographic gender** on four **general population** datasets, the respective SOTA models, and our own data and model. **Results are comparable**, no matter which data we trained on.

We held a competition at **PAN** to predict four demographics of celebrities. The performances of the eight submitted algorithms show:

What works?

- ✓ **Binary gender**, as usual
- ✓ Distinguishing the **most from the least famous** celebrities
- ✓ Predicting the occupations **sports, politics, and performers**
- ✓ **Age** in the range of ~20–40 years

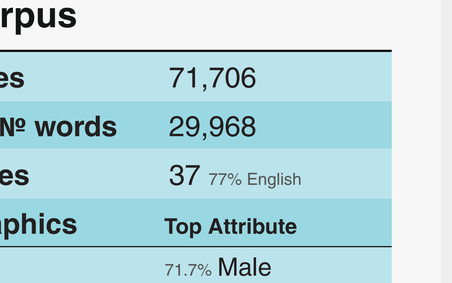
What does not work?

- ✗ **Rare demographics**: non-binary gender and religious occupations
- ✗ **Cross-topic occupations**: i.e. managers and scientists
- ✗ **Ages** outside the range of ~20–40 years

Corpus Analysis

Some insights into the population of celebrities on Twitter:

- Young **men** are most often famous for doing sports.
- Most young **women** are actors or musicians.
- Politics and management are an old man's business.
- The occurance ratio of occupations is more balanced for celebrities on Twitter older than 40.



Occurance ratios of (bottom) celebrity occupations over different years of birth and (top) the gender balance between different occupation groups.

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

(1) Bauhaus-Universität Weimar first.last@uni-weimar.de
(2) German Aerospace Center (DLR) first.last@dlr.de
(3) Leipzig University first.last@uni-leipzig.de