Overview of the Celebrity Profiling Task at PAN 2020

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,

![Age and Gender Distribution](chart.png)
Celebrity Profiling

Task

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

- **Age**,  
- **Gender**, and  
- **Occupation**.
Dataset creation:

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

1. Extract celebrities with matching profiles from a Corpus [ACL 2019].
2. Download follower network.
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.
Performance is measured as the harmonic mean of the classwise averaged $F_1$.

\[
c\text{Rank} = \frac{3}{\frac{1}{F_{1,\text{gender}}} + \frac{1}{F_{1,\text{occupation}}} + \frac{1}{F_{1,\text{age}}}}
\]
Celebrity Profiling
Evaluation

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

$$c\text{Rank} = \frac{3}{\frac{1}{F_{1,\text{gender}}} + \frac{1}{F_{1,\text{occupation}}} + \frac{1}{F_{1,\text{age}}}}$$

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.
Celebrity Profiling

Results

Baseline:

- Algorithm: Logistic regression.
- Features: Bags of word 1 and 2-grams, TD-IDF weighted.

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

<table>
<thead>
<tr>
<th>Participant</th>
<th>Test dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cRank</td>
</tr>
<tr>
<td>baseline-follower</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Celebrity Profiling

Results

Baseline:

- Algorithm: Logistic regression.
- Features: Bags of word 1 and 2-grams, TD-IDF weighted.

Trained and tested on all followers’ tweets as a lower bound.
Trained and tested on the celebrities’ tweets as a goalpost.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Test dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cRank</td>
</tr>
<tr>
<td>baseline-oracle</td>
<td>0.63</td>
</tr>
<tr>
<td>baseline-follower</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Celebrity Profiling

Results

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

- Baseline was beaten by a healthy margin.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Test dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cRank</td>
</tr>
<tr>
<td>baseline-oracle</td>
<td>0.63</td>
</tr>
<tr>
<td>Hodge and Price</td>
<td>0.58</td>
</tr>
<tr>
<td>Koloski et al.</td>
<td>0.52</td>
</tr>
<tr>
<td>Alroobaea et al.</td>
<td>0.47</td>
</tr>
<tr>
<td>baseline-follower</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Sep 25 • WIEGMANN
Celebrity Profiling

Results

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

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

<table>
<thead>
<tr>
<th>Participant</th>
<th>Test dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cRank</td>
</tr>
<tr>
<td>baseline-oracle</td>
<td>0.63</td>
</tr>
<tr>
<td>Hodge and Price</td>
<td>0.58</td>
</tr>
<tr>
<td>Koloski et al.</td>
<td>0.52</td>
</tr>
<tr>
<td>Alroobaea et al.</td>
<td>0.47</td>
</tr>
<tr>
<td>baseline-follower</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Celebrity Profiling

Results

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

- Baseline was beaten by a healthy margin.
- Submissions predict young users (20-30) better by $0.2 \text{ F}_1$.
- Submissions skew towards the “Male” class.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Test dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cRank</td>
</tr>
<tr>
<td>baseline-oracle</td>
<td>0.63</td>
</tr>
<tr>
<td>Hodge and Price</td>
<td>0.58</td>
</tr>
<tr>
<td>Koloski et al.</td>
<td>0.52</td>
</tr>
<tr>
<td>Alroobaea et al.</td>
<td>0.47</td>
</tr>
<tr>
<td>baseline-follower</td>
<td>0.47</td>
</tr>
</tbody>
</table>
Celebrity Profiling

Results

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

- Baseline was beaten by a healthy margin.
- Submissions predict young users (20-30) better by .2 F₁.
- Submissions skew towards the “Male” class.
- Submissions beat the oracle on occupation, although “Creators” is a problematic class (.46 F₁).

<table>
<thead>
<tr>
<th>Participant</th>
<th>Test dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cRank</td>
</tr>
<tr>
<td>baseline-oracle</td>
<td>0.63</td>
</tr>
<tr>
<td>Hodge and Price</td>
<td>0.58</td>
</tr>
<tr>
<td>Koloski et al.</td>
<td>0.52</td>
</tr>
<tr>
<td>Alroobaea et al.</td>
<td>0.47</td>
</tr>
<tr>
<td>baseline-follower</td>
<td>0.47</td>
</tr>
</tbody>
</table>
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!