Age and Gender Recognition Using Tweets in a Multilingual Setting

Mounica Arroju
mounica@uw.edu
University of Washington, Tacoma

Aftab Hassan
afhah@uw.edu
University of Washington, Tacoma

Golnoosh Farnadi
golnoosh.farnadi@ugent.be
Ghent University and Katholieke Universiteit Leuven

Objective
The author profiling’s task of 2015 is to identify age, gender and personality traits of Twitter users from their tweets in English, Spanish, Italian and Dutch.

Age class: 18-24, 25-34, 35-49, 50-xx.
Gender: Female vs. Male.
Personality scores: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness.

Dataset
The training dataset from Twitter, given by PAN organizers.

Approach
We build two multilingual models, one for identifying age and gender of the users and another for predicting their personality traits.

Age and Gender Prediction
- We extract the word n-gram (uni, bi and trigram) features.
- We apply a linear model with stochastic gradient descent that iteratively optimizes the gradient descent and updates the model with each training example.

Personality Prediction
- We tokenize the words and match them against the Linguistic Inquiry and Word Count (LIWC) dictionary.
- LIWC maps words to categories. By adding the TF-IDF value of the token, we create a feature vector.
- We treat the prediction problem as a multi-target regression problem and use an ensemble of regression chains (ERCC).
- ERCC let us leverage the prediction result for one personality trait to make a prediction for another.

Results
- Results of predicting age, gender and personality traits (Extraversion(Extr), Agreeableness (Agr), Conscientiousness (Con), Emotional Stability (Ems), Openness(Open)) of all four language tweets using our models on the test dataset reported by PAN organizers.

<table>
<thead>
<tr>
<th>Language</th>
<th>Age</th>
<th>Gender</th>
<th>Con</th>
<th>Ext</th>
<th>Agr</th>
<th>Openness</th>
<th>Ems</th>
<th>Global RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.70</td>
<td>0.77</td>
<td>0.184</td>
<td>0.193</td>
<td>0.172</td>
<td>0.184</td>
<td>0.249</td>
<td>0.70</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.69</td>
<td>0.75</td>
<td>0.175</td>
<td>0.190</td>
<td>0.177</td>
<td>0.190</td>
<td>0.215</td>
<td>0.65</td>
</tr>
<tr>
<td>Dutch</td>
<td>0.53</td>
<td>0.55</td>
<td>0.153</td>
<td>0.153</td>
<td>0.162</td>
<td>0.183</td>
<td>0.223</td>
<td>0.68</td>
</tr>
<tr>
<td>Italian</td>
<td>0.58</td>
<td>0.45</td>
<td>0.140</td>
<td>0.150</td>
<td>0.162</td>
<td>0.194</td>
<td>0.194</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Conclusion
We obtained an average 68.5% accuracy for identifying users’ attributes in four different languages.

Our model got better results in inferring age for the test dataset (i.e., 70% and 69% for English and Spanish, respectively) compared to the train dataset (i.e., 69% and 48% for English and Spanish, respectively).

Acknowledgments
This work was funded in part by the SBO-program of the Flemish Agency for Innovation by Science and Technology (IWT-SBO-Nr. 110067)