Age, Gender and Personality Recognition Using Tweets in a Multilingual Setting

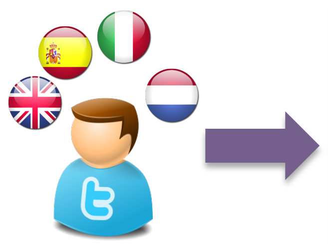
Mounica Arroju
mounica@uw.edu
University of Washington, Tacoma

Aftab Hassan
aftabh@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.

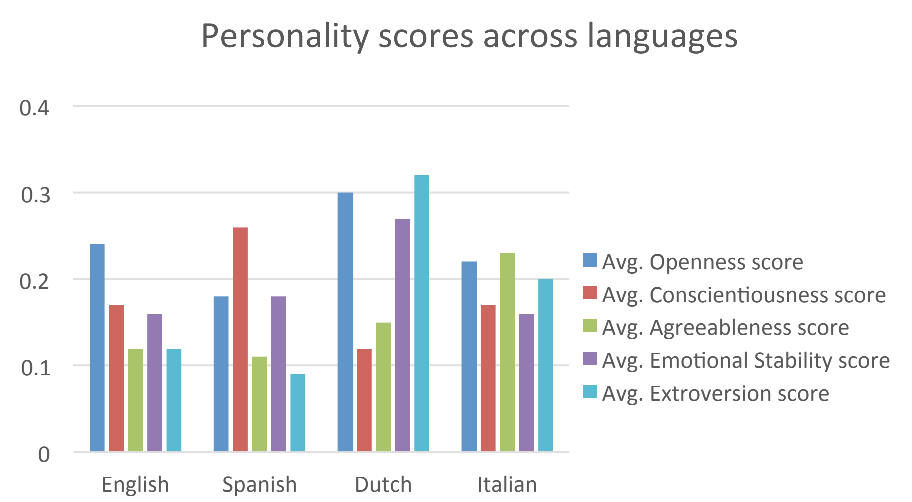


Age class: 18-24, 25-34, 35-49, 50-xx.
Gender: Female vs. Male.
Personality scores: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness.

Dataset

The training dataset from Twitter, given by PAN organizers.

Statistic	English	Spanish	Dutch	Italian
# Female users	41	25	2	12
# Male users	38	36	5	12
Avg. tweets per male user	189	201	195	190
Avg. tweets per female user	194	195	202	202
Avg. length of tweets per user	72	73	74	72
Majority age group	18-24	25-34	Unknown	Unknown



Approach

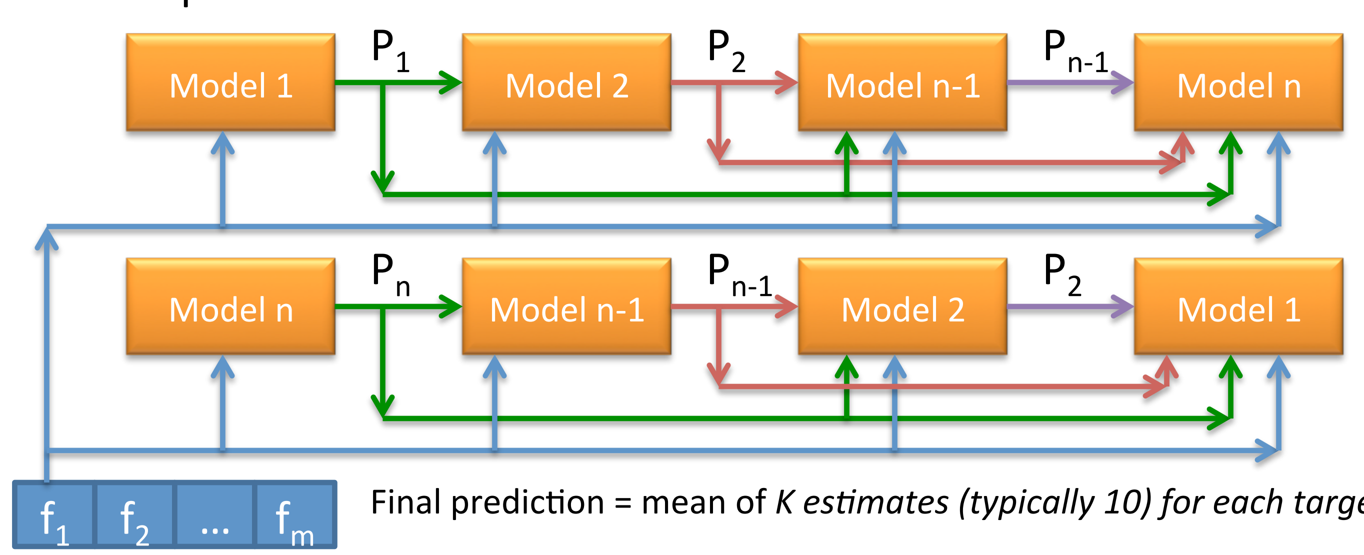
We build two multilingual models, one for identifying age and gender of the users and another for predicting their personality traits.

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.



Final prediction = mean of *K estimates (typically 10) for each target*

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.

Age	Gender	Con	Ext	Agr	Openness	Ems	Global	RMSE
English								
0.70	0.77	0.1481	0.1636	0.1513	0.1584	0.2349	0.70	0.1713
Spanish								
0.69	0.75	0.1785	0.1980	0.1727	0.1469	0.2125	0.65	0.1817
Dutch								
-	0.53	0.1553	0.1573	0.1672	0.1103	0.2235	0.68	0.1627
Italian								
-	0.58	0.1345	0.1480	0.1520	0.1620	0.1941	0.71	0.1581

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)

CENTER FOR DATA SCIENCE
UNIVERSITY of WASHINGTON | TACOMA
Institute of Technology

UNIVERSITEIT GENT

KU LEUVEN

cwds.uw.edu