UniNE at CLEF 2015: Author Profiling

Mirco Kocher

University of Neuchâtel, Switzerland

Swiss National Science Foundation

Task

Predict an author's demographics from her Twitter tweets

- Output: gender, age group, Big Five personality traits
  - Gender – nominal values male and female
  - Age group – ordinal measurements 18-24, 25-34, 35-49, and >50
  - Personalities – interval scale from -0.5 to +0.5
    - Extraversion, neuroticism, agreeableness, conscientiousness, and openness
- Four different datasets:
  - Dutch and Italian – determine gender & personality
  - English and Spanish – predict gender, age & personality
- *TIRA* platform for evaluation
  - Encapsulated system with restricted data access
  - Fair comparison of the time needed to produce an answer
- Total 22 different participants

Strategy

Term	@unknown	…	@alexa	…	@james	…	@john	…
the	0.09		0.09		0.10		0.06	
of	0.06		0.06		0.06		0.04	
to	0.03		0.00		0.01		0.05	
#tags	0.02		0.04		0.02		0.00	
…								

200

Number of samples in training set

- L_1-norm to select three nearest neighbors:
  - @unknown – @alexa: 0.05 -> 2nd nearest neighbor
  - @unknown – @james: 0.03 -> 1st nearest neighbor
  - @unknown – @john: 0.09 -> 3rd nearest neighbor
  - … …
- Determine gender and age with majority voting
  - Select nearest if three different age groups are returned

Rank	Gender	Age group
1	Male	25-34
2	Female	35-49
3	Male	18-24

  - Predict: Male 25-34
- Each personality trait is arithmetic mean of candidates traits

Rank	Extraversion	Neuroticism	…	Openness
1	0.3	0.5		0.2
2	0.2	0.1		0.2
3	0.1	0.4		0.1

  - Predict: 0.2 0.3 0.2

Results

Language	Final	Joint	RMSE	Runtime	Rank
Dutch	0.8469	0.8125	0.1186	00:00:01	6
English	0.7037	0.5563	0.1489	00:00:04	8
Italian	0.8260	0.7778	0.1259	00:00:01	4
Spanish	0.7735	0.6705	0.1235	00:00:02	4

- Joint = accuracy of age and gender
- RMSE = root mean squared error of personality traits
- Final = (joint + 1 – RMSE) / 2
- No age prediction in Dutch and Italian required
- English and Spanish datasets are significantly bigger
- Compared with all 22 participants:
  - Fastest runtime (8 seconds), median runtime for classification of other systems is over ten minutes (excluding their training time)

Evaluation

- Explanation of our proposed assignment (*e.g.* in English):
  - Usually, the relative frequency differences with very frequent words such as *when*, *is*, *in*, *that*, *to*, or *it* can explain the decision
- Overall gender & age detection
  - Median accuracy of 70% for age group prediction in English
  - Lower age prediction accuracy in Spanish (60%)
  - Almost 85% accuracy for gender recognition in Spanish
  - Lower gender accuracy in Dutch, English, & Italian (72%, 73%, & 61%)
- Neuroticism factor:
  - Tendency to experience negative emotions
  - Most difficult trait in all four languages for all participants
  - Other four traits have mean RMSE around 0.15
  - This has mean RMSE around 0.2
  - Maybe more complicated to determine from written text
  - Possibly less reliable answers on self-assessment tests

Conclusion

- Macro-averaging ranks us 4th out of 22 participants
- Simple supervised approach can solve the profiling problem
- No language dependent adjustments necessary
- Parameter training not needed
- Most frequent terms tend to select most discriminative features
- Reduced set of comprehensible features could clearly explain the decision taken

Acknowledgments

This research was supported, in part, by the NSF under Grant #200021_149665/1.

Mirco Kocher – University of Neuchâtel – Rue Emile-Argand 11 – 2000 Neuchâtel, Switzerland – Mirco.Kocher@unine.ch

http://style.unine.ch/