Author profiling using stylometric and structural feature groupings

Pan 2015

Andreas Grivas, Anastasia Krithara, George Giannakopoulos

{agrv, akrithara, ggianna}@demokritos.gr

National Center for Scientific Research "Demokritos", Athens, Greece

Introduction

2015 Author Profiling task:

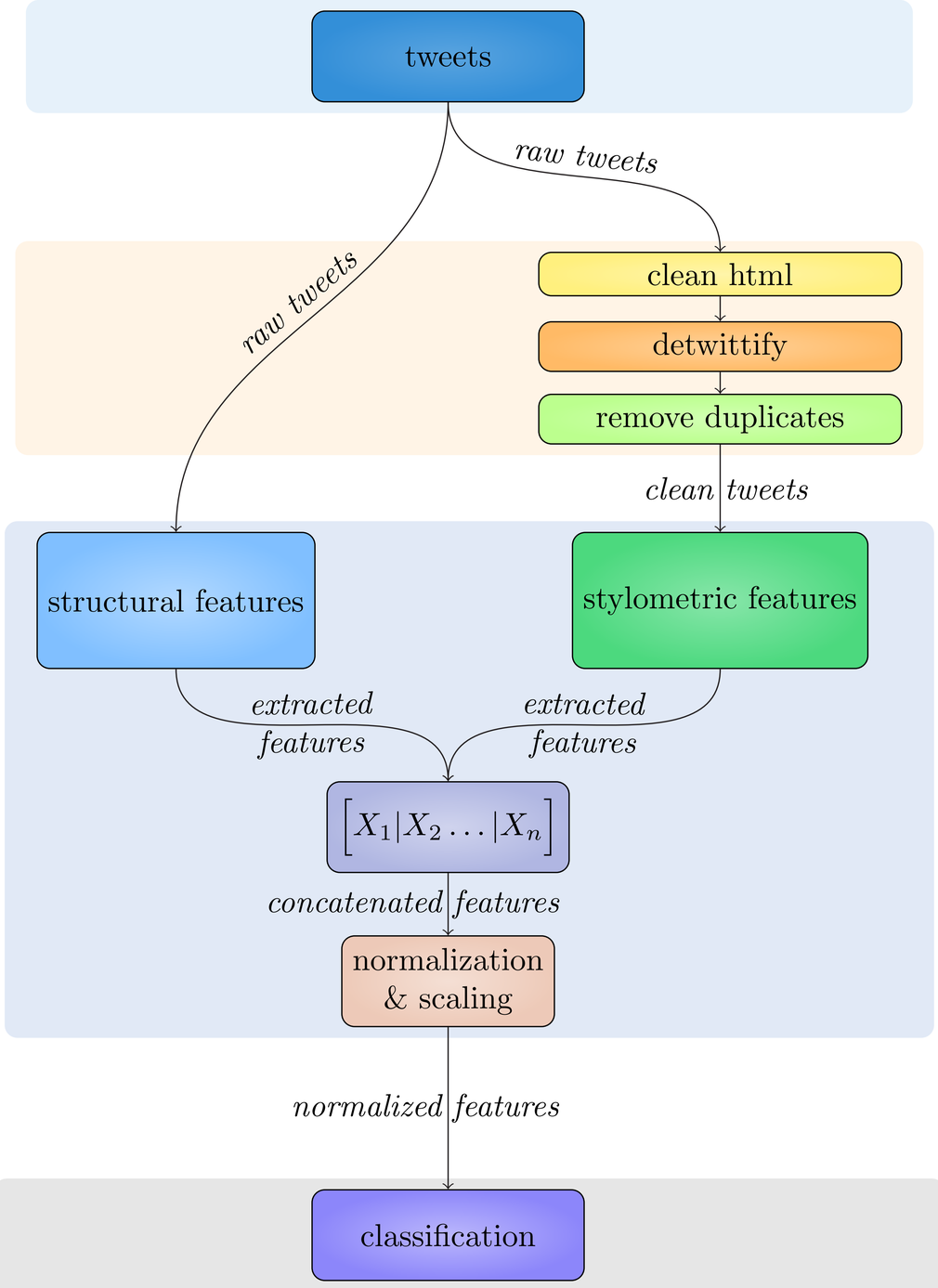
- 3 subtasks (gender, age, personality traits)
- 152 user profiles
- 4 languages (English, Spanish, Italian and Dutch)

Approach

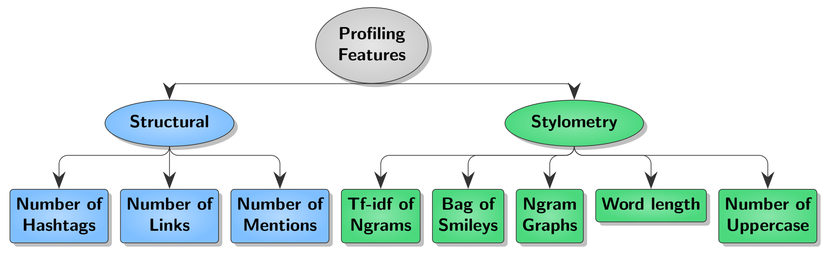
Steps followed for the task:

1. Concatenate tweets of each user
2. Pass text through pre-processing pipeline
3. Extract feature groupings (stylometric, structural)
4. Predictors:
   - Gender & age → classify with Support Vector Machine
   - Personality traits → regression with Support Vector Regression

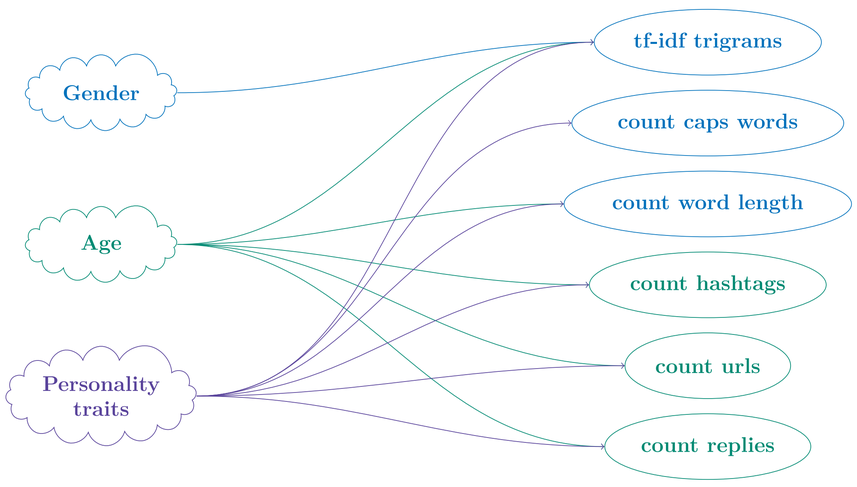
System Pipeline



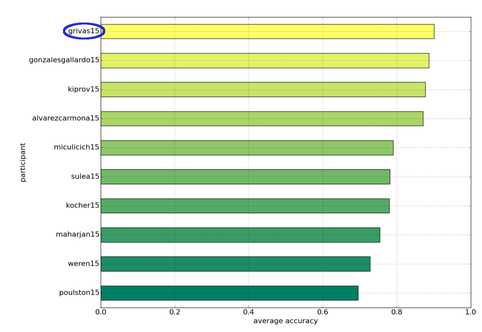
Feature groupings



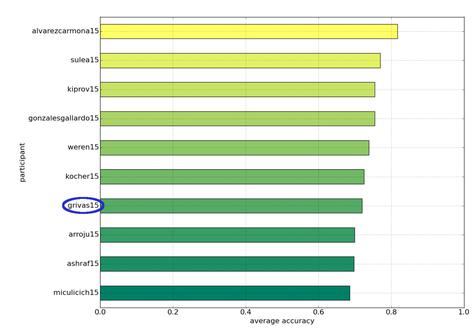
Features used for each subtask



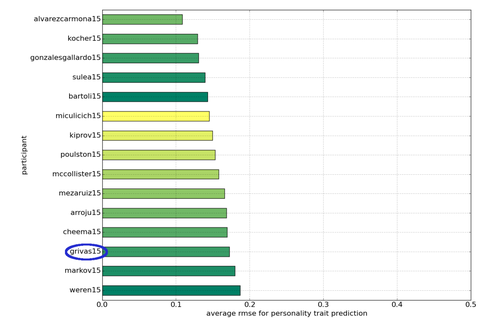
Results



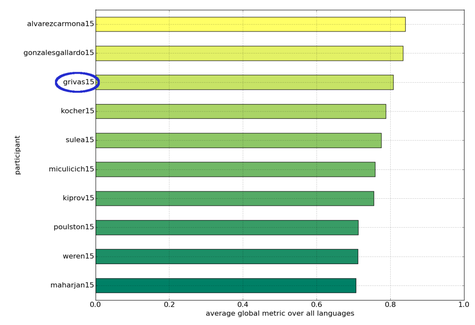
(a) Gender subtask all languages average



(b) Age subtask all languages average (English, Spanish)



(c) Personality subtask all languages average



(d) Global ranking all languages

Conclusions

- ✓ trigrams can capture gender information regardless of language and generalize well for datasets of this size
- ✓ need better feature analysis for the case of age and personality traits

Future Work

- Develop a more sophisticated approach for age and personality trait estimation
- Extensively evaluate results and feature selection
- Attempt to use fewer documents per user - more batches - use average for final decision

Supported by:

Grant Agreements No. FP7-610928

Demokritos Persona Reveal