Author Profiling using Complementary Second Order Attributes and Stylometric Features

Konstantinos Bougiatiotis and Anastasia Krithara
{bogas.ko, akrithara}@iit.demokritos.gr
National Center for Scientific Research "Demokritos", Athens, Greece

Introduction

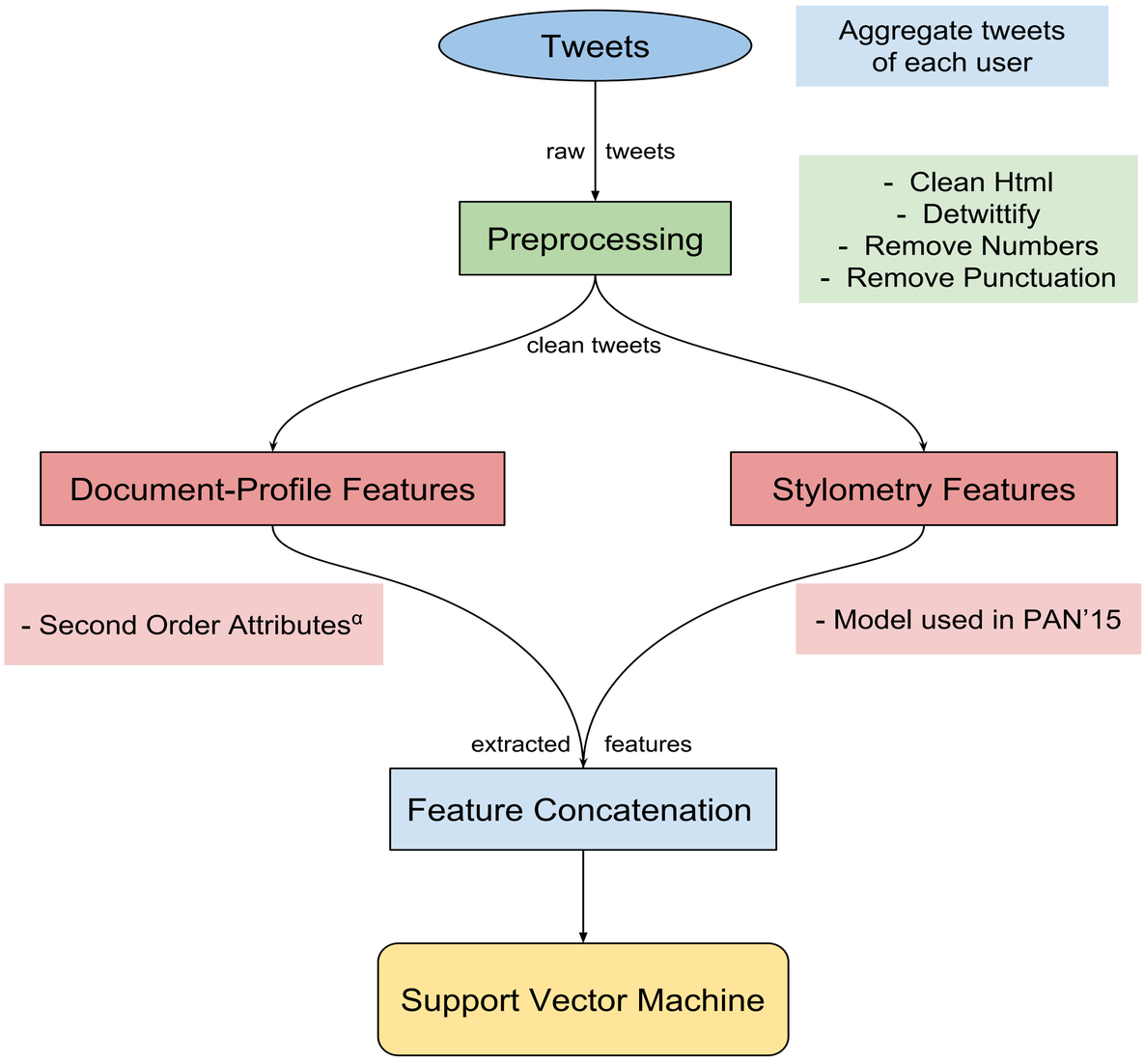
Overview

- *Author Profiling*: Predict **author traits** based solely on text
- *Novelty*: PAN'16 features **cross-genre** evaluation(train on Twitter texts and test on other genres)

Data

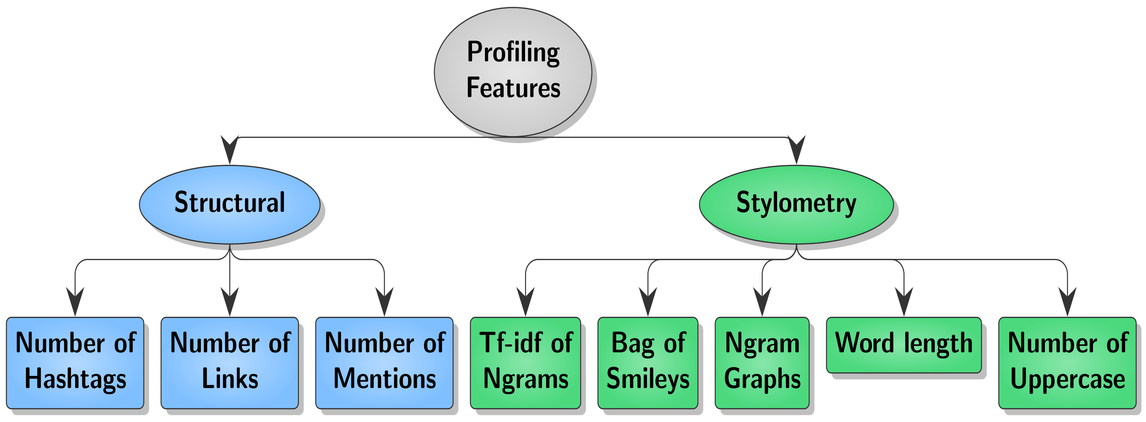
Users: 1070 | Tweets: 562812
Tasks: **Age** and **Gender**
Languages: English, Spanish and Dutch(gender only)

System Workflow



[a] López-Monroy et al.: INAOE's participation at PAN'13: Author Profiling task-Notebook for PAN at CLEF 2013. In: CLEF 2013 Evaluation Labs and Workshop

Stylometric-Structural Features(PAN15')



Finally, selected top **3000** frequent **3-grams** of chars(*age*) and **unigrams** of chars(*gender*)

Supported by:
SEVENTH FRAMEWORK PROGRAMME
Reveal
Grant Agreements No. FP7-610928

Second Order Attributes(SOA)

1. Calculate word-profile vectors → Find **descriptive terms per class**, exploiting the per-class frequency of the words

$$t_{i,j} = \sum_{k:d_k \notin P_j} log(1 + \frac{tf_{i,k}}{len(d_k)} * w_k)$$

2. Map **documents in profile space**, using the word-profile vectors, from step 1, of the containing terms for each document

$$d_{k,j} = \sum_{i:t_i \in d_k} \frac{tf_{i,k}}{len(d_k)} \times \vec{t_i}$$

Sample representation of the SOA method

Vocabulary V

Documents D	"and"	"football"	"shopping"	....
doc	0.09	0.60	0	....
doc2	0.35	0.21	0.28	....
...	....	....	....	....
docD	0.14	0	0.8	....

X

Target Profiles P

Vocabulary V	Female	Male
....	....	....
"and"	0.45	0.55
"football"	0.35	0.65
"shopping"	0.68	0.32
....	....	....

Target Profiles P

Documents D	Female	Male
doc	0.4	0.6
doc2	0.48	0.52
...		
docD	0.72	0.28

Weighted SOAComplementary

- Use the **complementary classes** for each word-class relation → More even amount of data for each class → **Robust estimates** and lesser bias
- Weights **inversely proportional to class frequency** → Terms related with rare profiles, aggregate more weight →*Prior knowledge* will **help sparsely populated classes**

Train Results(4-fold CV)

Models	English		Spanish		Dutch
	Age	Gender	Age	Gender	Gender
N-grams(PAN'15)	47.0	74.8	49.6	68.8	76.8
SOA	47.5	76.2	**54.0**	72.8	76.0
SOAC	49.1	76.8	50.4	71.6	76.8
W-SOAC	49.1	76.8	50.4	72.8	76.8
N-grams + W-SOAC	**50.0**	**77.5**	52.0	**73.2**	**78.1**

Test Results

Dataset	**Language**	**Subtask**	**Accuracy%**
Test-1	Dutch	Gender	44.00
	English	Age	30.46
		Gender	53.45
	Spanish	Age	29.69
		Gender	60.94
Test-2	Dutch	Gender	41.60
	English	Age	**55.13**
		Gender	**69.23**
	Spanish	Age	32.14
		Gender	67.86

Conclusions

✓ Stylometry and Discriminative features both capture gender information well enough. Also boosted performance through **fusion**
✓ Age considerably more difficult than gender to predict, across all languages and regardless of the methodology
✓ Different performance in the two test datasets, highlight the added **difficulty of the cross-genre task**