Author Profiling using Complementary Second Order Attributes and Stylometric Features

Konstantinos Bougiatiotis* Anastasia Krithara

Institute of Information and Telecommunication, N.C.S.R "Demokritos", Greece

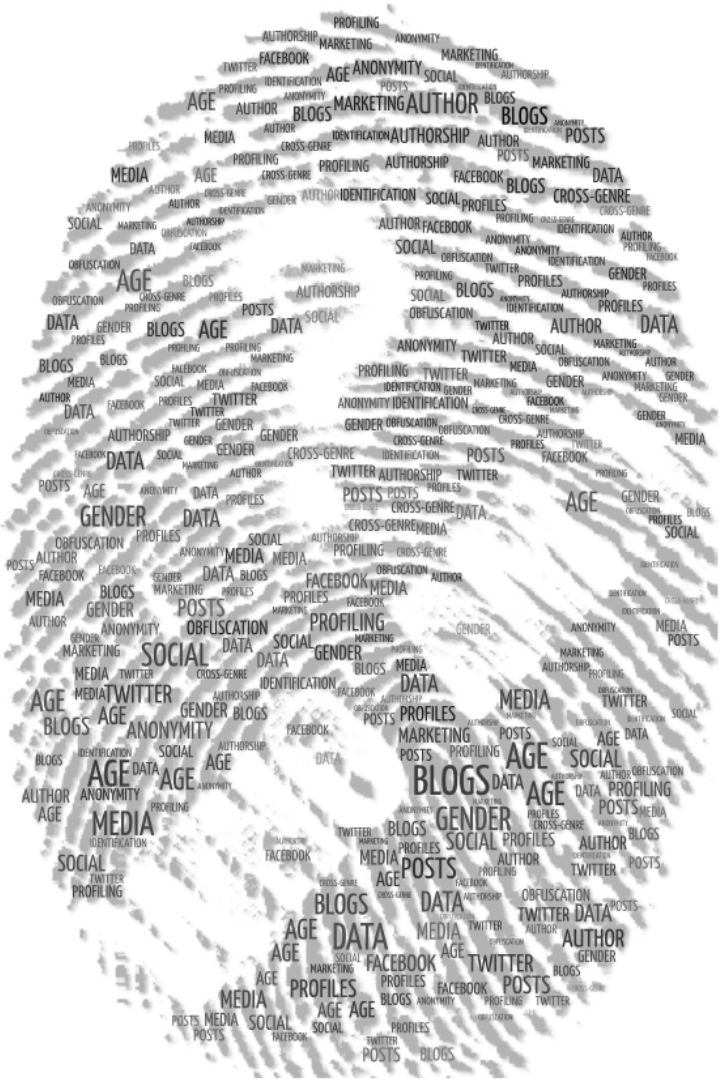
September 3, 2016

Outline

Introduction

Author Profiling

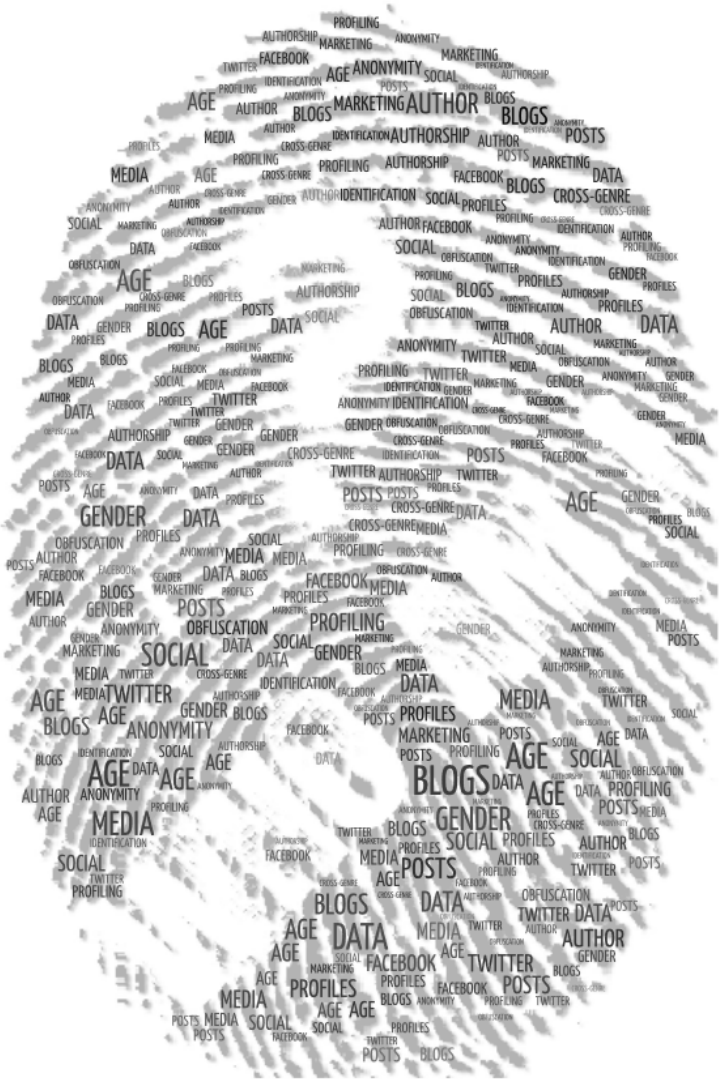
- Find specific **characteristics of authors**, by studying their texts



Introduction

Author Profiling

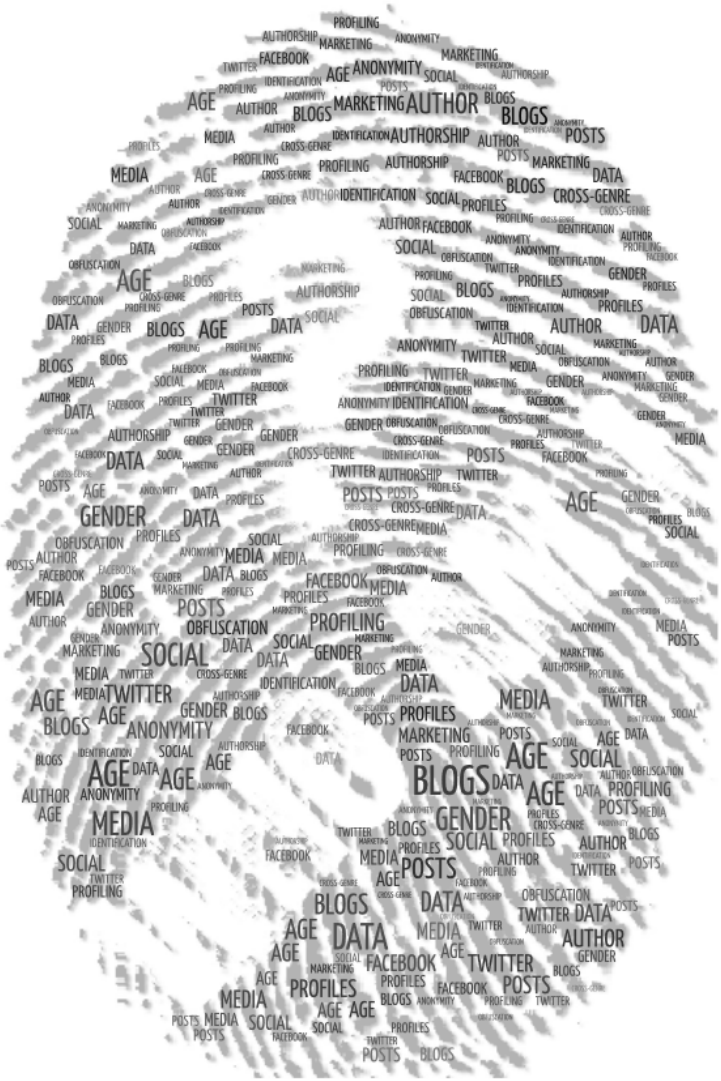
- Find specific **characteristics of authors**, by studying their texts
- **Age**, **gender**, personality traits, emotions



Introduction

Author Profiling

- Find specific **characteristics of authors**, by studying their texts
- **Age**, **gender**, personality traits, emotions
- Marketing, Security, Forensics, ...



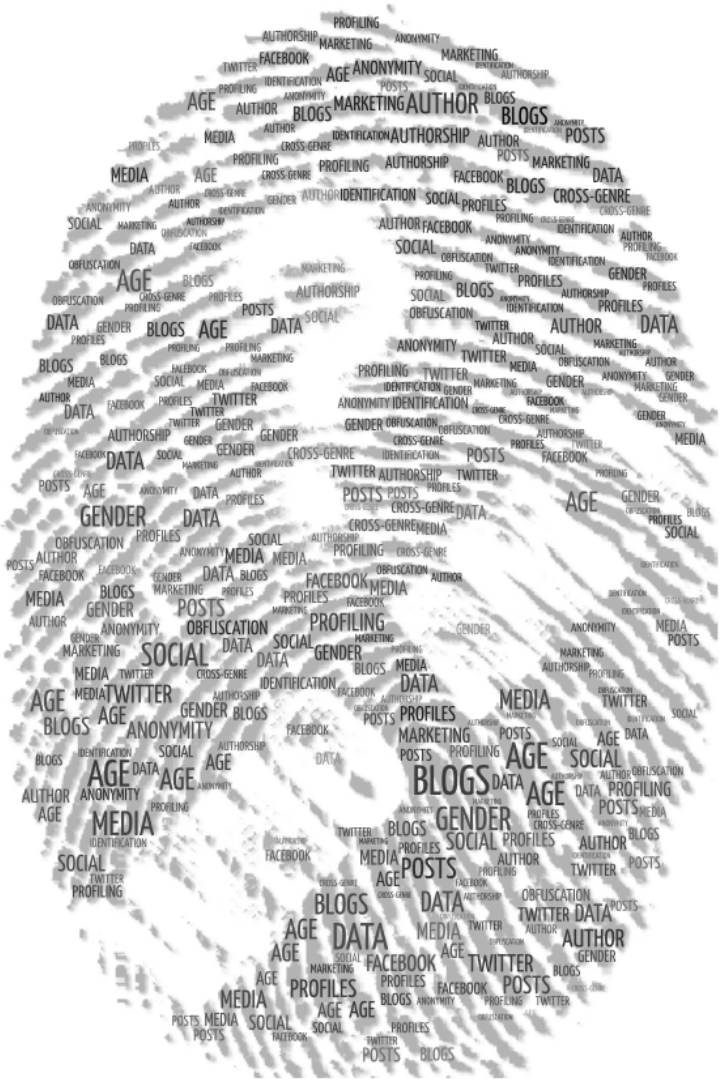
Introduction

Author Profiling

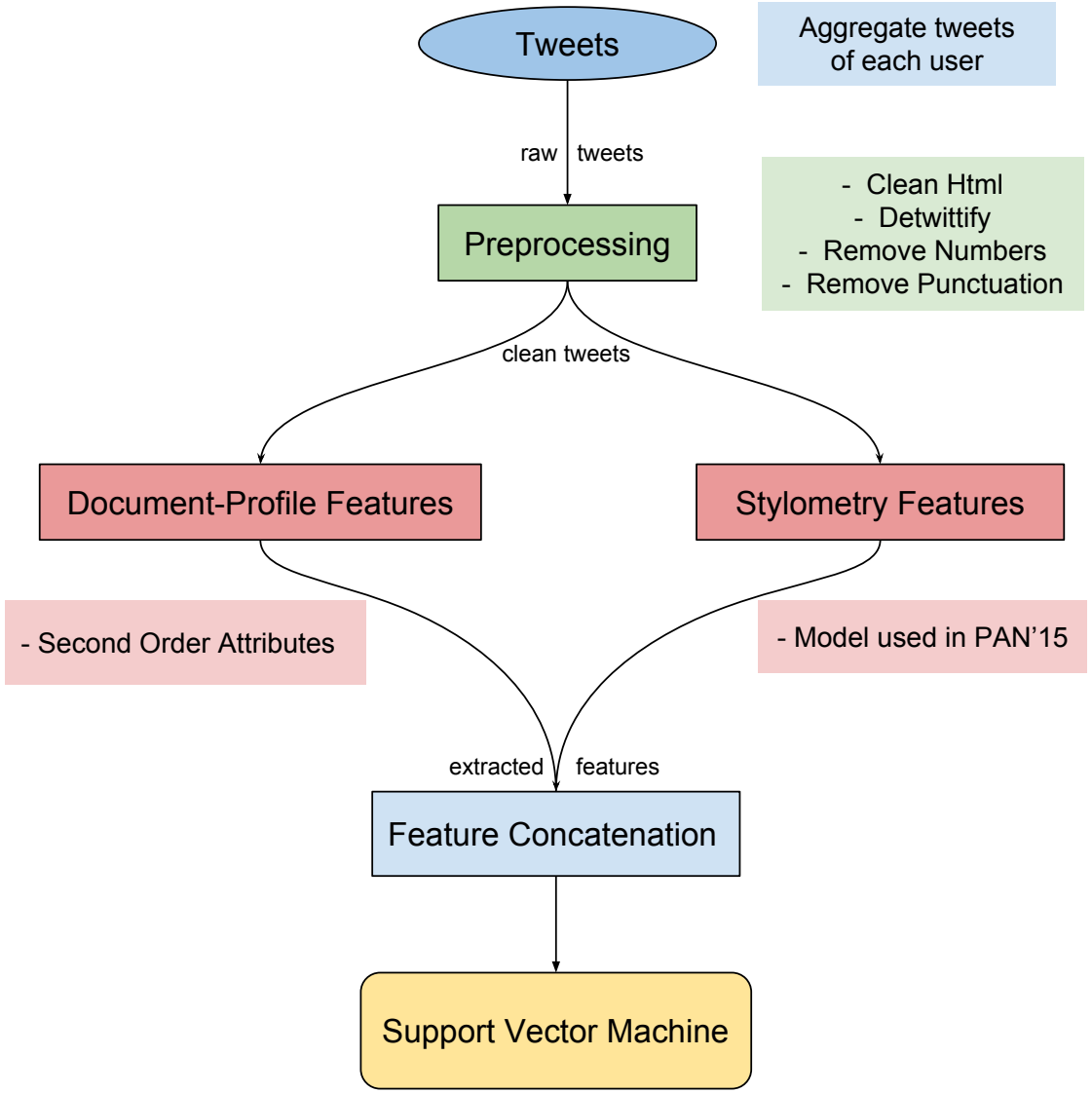
- Find specific **characteristics of authors**, by studying their texts
- **Age**, **gender**, personality traits, emotions
- Marketing, Security, Forensics, ...

Pan'16

- Languages: English, Spanish and Dutch(gender only)
- Focus on **cross-genre** evaluation



General Workflow



Tweets

Concatenate the tweets of each user ➔ **Profile Based Approach**

Tweets

Concatenate the tweets of each user → **Profile Based Approach**

- Raw Tweet: Noisy data, HTML tags, links, etc

Sample Tweet

```
Thanks for the follow back
<a href="/WolfgangDigital"
class="twitter-atreply
pretty-link js-nav"
data-mentioned-user-id="391
869708"
><s>@</s><b>WolfgangDigital
</b></a> I&#39;ll be
keeping an eye out for any
vacancies you advertise in
the near future.
```

Tweets

Concatenate the tweets of each user → **Profile Based Approach**

- Raw Tweet: Noisy data, HTML tags, links, etc
- Cleaning HTML

Sample Tweet

```
Thanks for the follow back
@WolfgangDigital I&#39;ll
be keeping an eye out for
any vacancies you
advertise in the near
future.
```

Tweets

Concatenate the tweets of each user → **Profile Based Approach**

- Raw Tweet: Noisy data, HTML tags, links, etc
- Cleaning HTML
- Detwittify (remove hashtags, replies etc)

Sample Tweet

```
Thanks for the follow back
I&#39;ll be keeping an eye
out for any vacancies you
advertise in the near
future.
```

Tweets

Concatenate the tweets of each user → **Profile Based Approach**

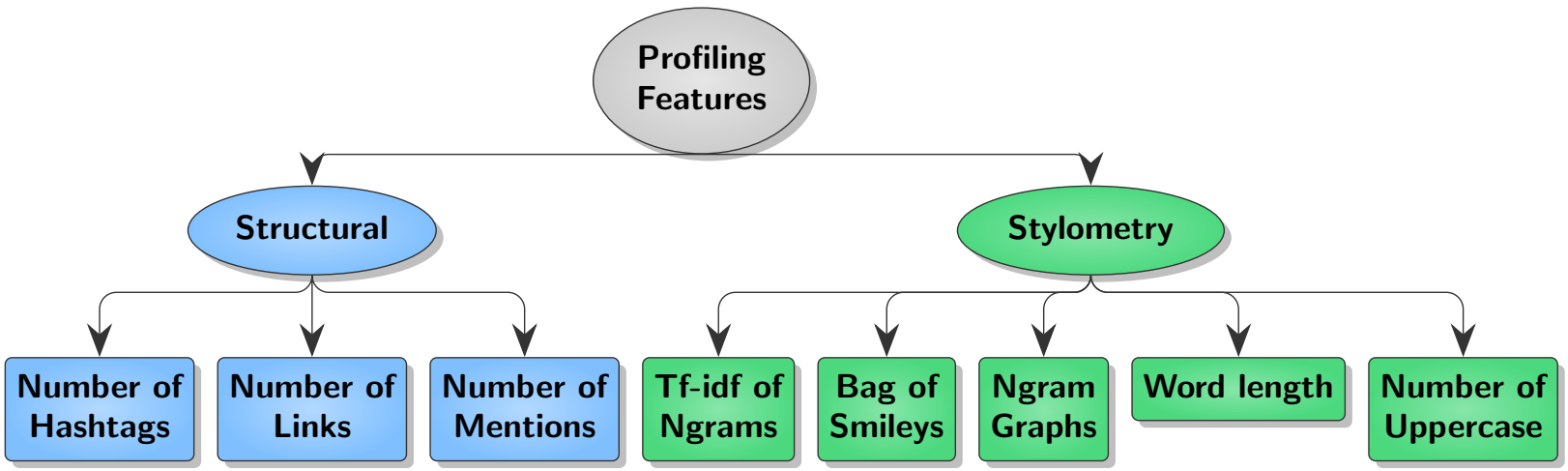
- Raw Tweet: Noisy data, HTML tags, links, etc
- Cleaning HTML
- Detwittify (remove hashtags, replies etc)
- Remove all non-letter characters (numbers, ...)

Sample Tweet

```
Thanks for the follow back
I ll be keeping an eye out
for any vacancies you
advertise in the near
future
```

Stylometric and Structural Features - PAN'15

Experimented with many features:



Finally settled on term-frequencies **3-grams**(age) and **unigrams**(gender)

Second Order Attributes-SOA

- Idea originally from PAN'13 winning Team (INAOE, Mexico)[1]
- **2-step method**, similar approach to Naive Bayes

[1] 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

Second Order Attributes-SOA

- Idea originally from PAN'13 winning Team (INAOE, Mexico)[1]
- **2-step method**, similar approach to Naive Bayes

Intuition

1. Associate the different terms in our collection with target profiles (age or gender classes) $\rightarrow$ Calculate **words-classes vectors** based on word frequency

[1] 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

Second Order Attributes-SOA

- Idea originally from PAN'13 winning Team (INAOE, Mexico)[1]
- **2-step method**, similar approach to Naive Bayes

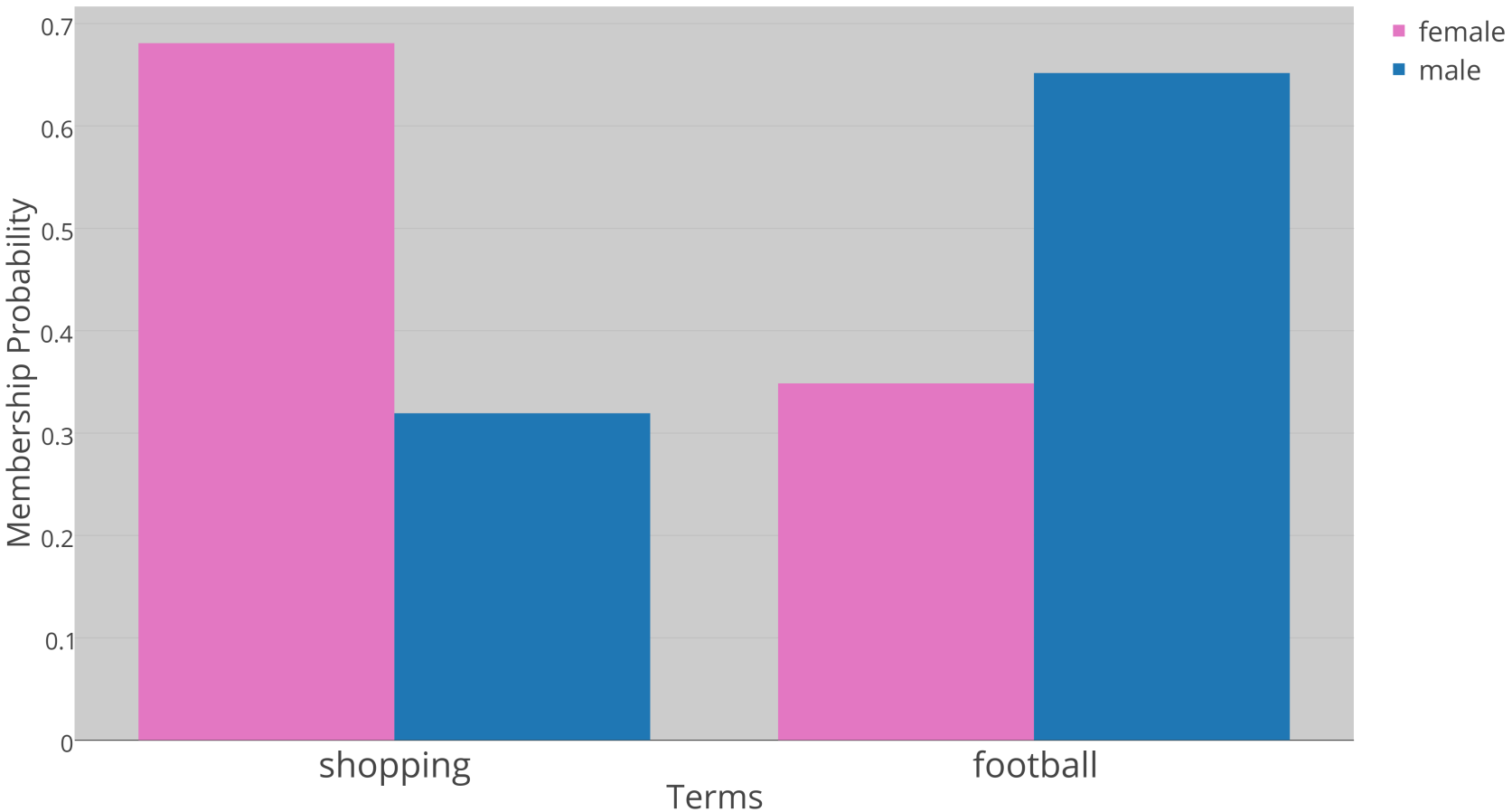
Intuition

1. Associate the different terms in our collection with target profiles (age or gender classes) $\rightarrow$ Calculate **words-classes vectors** based on word frequency
2. Project the documents in the profile space according to the weighted aggregation of their terms $\rightarrow$ Calculate **document-classes vectors**

[1] 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

Example of Age Specific Terms

Example of Gender Specific Terms



Example illustration of generated SOA

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

Vocabulary V

Target Profiles P

	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

Novelties introduced:

- Use **complementary classes** documents for each word-class relation

Intuition

Counter **skewed class** distribution of data → Use **complementary classes** for each term-profile relation → More **even amount** of data for each class → **Robust estimates and lesser bias**

Weighted SOAComplementary

Novelties introduced:

- Use **complementary classes** documents for each word-class relation
- Add **weighting** term to boost the influence of terms in documents of rare profiles

Intuition

Exploit knowledge of **prior distribution** of documents into classes $\rightarrow$ The rarer a profile, the higher the influence of the terms included in it $\rightarrow$ Weighting term **inversely proportional** to the probability of the profile $\rightarrow$ Cope with the **sparsity** of specific profiles

Classification

Experimented with many different classifiers:(*sklearn* implementations)

- Naive Bayes
- Decision Trees
- Random Forests
- **SVM**

Classification

Experimented with many different classifiers:(*sklearn* implementations)

- Naive Bayes
- Decision Trees
- Random Forests
- **SVM**

- Age: **RBF** kernel
- Gender: **Linear** kernel

Classification

Experimented with many different classifiers:(*sklearn* implementations)

- Naive Bayes
- Decision Trees
- Random Forests
- **SVM**

- Age: **RBF** kernel
- Gender: **Linear** kernel
- Hyper-parameters selected through grid search

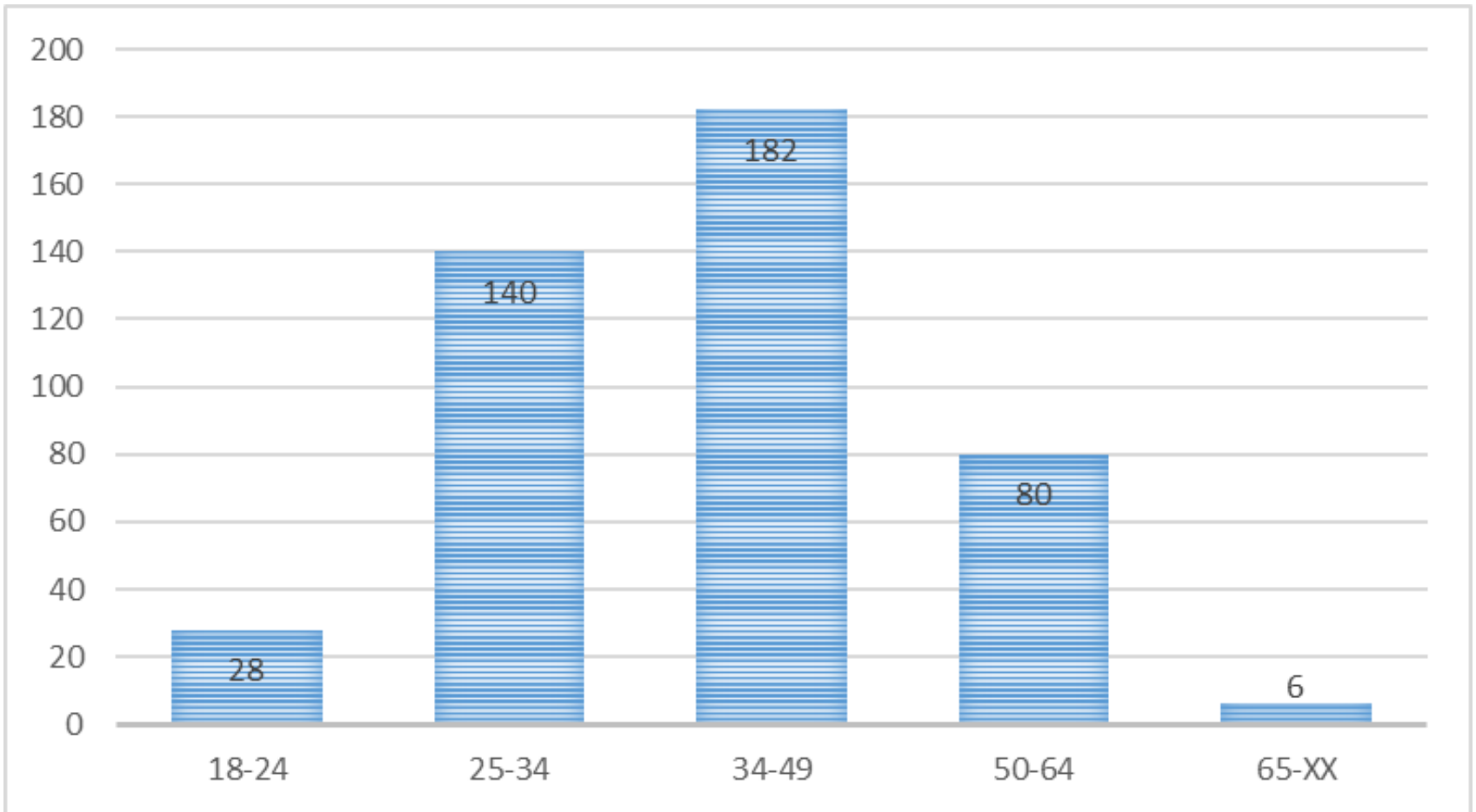
Dataset

Much more data than PAN'15:

- **1070 Users**: 436 English | 250 Spanish | 384 Dutch
- **562812 Texts**: 277792 English | 208620 Spanish | 76800 Dutch

  - Age: **Imbalanced** dataset over age classes
  - Gender: **Uniform** distribution of male/female samples

English Dataset Age Distribution



% Accuracy of 4-fold CV tests

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**

% Accuracy of 4-fold CV tests

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**

% Accuracy of 4-fold CV tests

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**

Average Joint Accuracy

Team	Global	English	Spanish	Dutch
Busger et al.	**0.5258**	0.3846	**0.4286**	0.4960
Modaresi et al.	0.5247	0.3846	**0.4286**	0.5040
...	...	...	...	...
Bougiatiotis & Krithara	0.4519	**0.3974**	0.2500	0.4160
...	...	...	...	...
Deneva	0.4014	0.2051	0.2679	**0.6180**
...	...	...	...	...

Average Joint Accuracy

Team	Global	English	Spanish	Dutch
Busger et al.	**0.5258**	0.3846	**0.4286**	0.4960
Modaresi et al.	0.5247	0.3846	**0.4286**	0.5040
...	...	...	...	...
Bougiatiotis & Krithara	0.4519	**0.3974**	0.2500	0.4160
...	...	...	...	...
Deneva	0.4014	0.2051	0.2679	**0.6180**
...	...	...	...	...

- Average Accuracy: **45.19%**
- Position: 6^{th} (22 teams overall)
- **1^{st}** Position on **global ranking** for the English language

Conclusions

- ✓ Descriptive and stylometric features model age and especially gender well enough.
- ✓ **Fusion** schemes seem to boost the performance
- ✓ **Age** subtask considerably **more difficult** across all models and languages
- ✓ Difference in performance between the test datasets highlight the **added difficulty of the cross-genre task**

Ongoing-Future Work

- Model age and gender in a **unified profile space** → Tackle the assumption of independence between tasks
- Examine more sophisticated **fusion schemes** and deploy **ensemble learning** techniques to exploit the difference in the representation spaces of each method
- Emphasis on cross-genre specialization, **important features per genre**, varying document length, per language-models, ...

Thank you!

Reveal

SEVENTH FRAMEWORK PROGRAMME

Appendix

Backup Slides

PAN'16 Author Profiling Challenge

Tasks:

- Predict **Age** and **Gender**
- Languages: English, Spanish and Dutch(gender only)

PAN'16 Author Profiling Challenge

Tasks:

- Predict **Age** and **Gender**
- Languages: English, Spanish and Dutch(gender only)

Novelties:

- Focus on **cross-genre** evaluation
- **Bigger** dataset (Users: 1070, Tweets: 562812)
- Added '65-xx' age class

SOA-Calculations

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

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

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

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

WSOAC- Calculations

1. Use complementary classes for the word-profile vectors

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

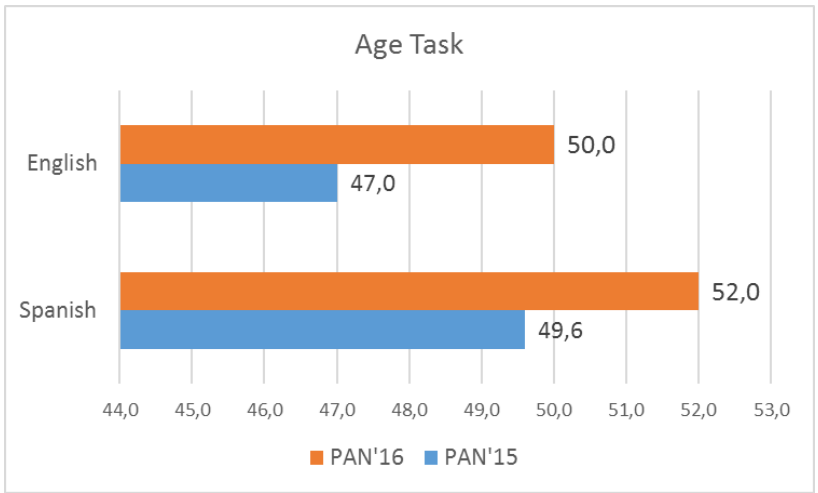
2. Use weights per class for the word-profile vectors

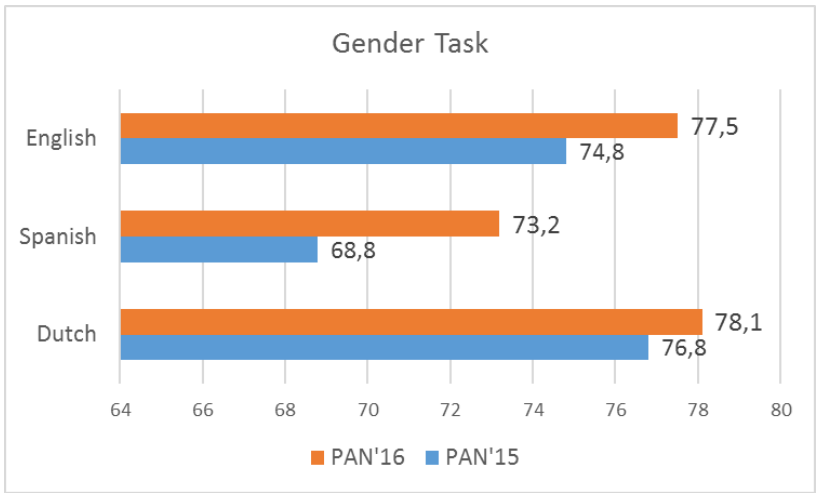
$$t_{i,j} = \sum_{\boldsymbol{k:d_k \notin P_j}} log(1 + \frac{tf_{i,k}}{len(d_k)} * \boldsymbol{w_k})$$

3. "Normalize" document-profile vectors by subtraction of the minimum score(corresponds to the most probable class)

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

Comparison of PAN'15-16 models





Voc, Dict Length? What about Tokenization?

% Accuracy of 4-fold CV tests

Models	English		Spanish		Dutch
	Age	Gender	Age	Gender	Gender
N-grams(PAN'15)	47.0	74.8	49.6	68.8	76.8
LSI	41.8	70.2	50.4	65.2	74.0
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 Data % Accuracy

Dataset	Language	Subtask	Accuracy
Social Media	Dutch	Gender	44.00
	English	Age	30.46
		Gender	53.45
	Spanish	Age	34.38
		Gender	57.81
Blogs	Dutch	Gender	41.60
	English	Age	**55.13**
		Gender	**69.23**
	Spanish	Age	32.14
		Gender	67.86