Author profiling

Lovre Mrčela, Marko Ratković, Ante Žužul

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

{lovre.mrcela,marko.ratkovic,ante.zuzul}@fer.hr

June 14, 2016

Contents

Introduction

What is author profiling?

Introduction

What is author profiling?

- author profiling deals with determining author's personality and identity based on written texts

Introduction

What is author profiling?

- author profiling deals with determining author's personality and identity based on written texts

Tasks:

Introduction

What is author profiling?

- author profiling deals with determining author's personality and identity based on written texts

Tasks:

- age–group *(classification)*
- gender *(classification)*
- personality traits *(regression)*
    - extraversion, stability, agreableness, conscientiousness, openness to experience

Introduction

Dataset:

Introduction

Dataset:

- PAN competition
- 4 languages: English, Spanish, Italian and Dutch

Contents

Approach

Approach:

Approach

Approach:

- text preprocessing
- feature extraction
- feature selection, finding the optimal model
- evaluate the model on the test set

Preprocessing tweets

Preprocessing tasks:

Preprocessing tweets

Preprocessing tasks:

- substituting *url* with **URL** tag
- substituting *usernames* (referenced in replies) with **REPLY** tag
- removing stop words from set of user tweets
- converting all tweets to lowercase
- trimming consecutive repetitions of a letter in words (e.g *'coooool'* to *'coool'*)

Set of features

Features used for this problem:

Set of features

Features used for this problem:

- tf–idf vectorized tweets
- number of emoticons, replies, hashtags, ...
- average length and standard deviation of posts and words respectively

Overview of additional features

Table : Overview of additional features values for each age–group, per language.

Language	English			
Age–group	18-24	25-34	35-49	50-XX
Post length	60.714	85.853	86.680	93.753
Post length deviation	29.656	29.401	29.532	32.192
Word length	4.908	5.984	6.312	6.013
Word deviation	3.493	4.726	5.088	4.452
Emoticon count	0.057	0.064	0.046	0.038
Hashtags	0.127	0.658	0.267	0.514
Character repetitions	0.040	0.012	0.017	0.003
Exclamation marks	0.137	0.207	0.195	0.527
User replies	0.492	0.540	0.632	1.293

Overview of additional features

Table : Overview of additional features values for each age–group, per language.

Language	Spanish			
Age–group	18-24	25-34	35-49	50-XX
Post length	75.728	85.246	92.804	101.991
Post length deviation	31.377	31.234	30.719	29.200
Word length	4.935	5.295	5.647	5.409
Word deviation	3.356	3.882	4.230	3.963
Emoticon count	0.135	0.104	0.053	0.030
Hashtags	0.168	0.340	0.259	0.231
Character repetitions	0.065	0.022	0.030	0.022
Exclamation marks	0.183	0.244	0.257	0.276
User replies	0.579	0.715	0.818	0.854

Overview of additional features values for gender

Table : Overview of additional features values for gender, per language.

Language	English		Spanish	
Gender	Female	Male	Female	Male
Post length	76.786	77.222	86.030	86.949
Post length deviation	31.131	32.908	30.594	29.574
Word length	5.529	5.718	5.328	5.281
Word length deviation	4.187	4.385	3.916	3.786
Emoticons count	0.075	0.039	0.082	0.102
Hashtags	0.380	0.394	0.357	0.190
Character repetitions	0.026	0.020	0.041	0.025
Exclamation marks	0.275	0.133	0.252	0.221
User replies	0.641	0.548	0.745	0.698

Overview of results of age–group classification per language

Table : Overview of additional features values for gender, per language.

Language	Italian		Dutch	
Gender	Female	Male	Female	Male
Post length	91.555	87.513	77.442	77.239
Post length deviation	32.908	30.594	29.574	30.829
Word length	5.898	6.153	5.255	5.229
Word length deviation	4.202	4.705	3.489	3.476
Emoticons count	0.214	0.072	0.119	0.072
Hashtags	0.540	0.700	0.424	0.120
Character repetitions	0.008	0.006	0.026	0.016
Exclamation marks	0.228	0.168	0.271	0.121
User replies	0.725	0.556	0.647	0.909

Models

Model used for classification:

Models

Model used for classification:

- Logistic Regression
- Naive Bayes Classifier
- Decision Tree Classifier
- Random Forest Classifier
- SVC (using *rbf*, linear, poly and sigmoid kernels)

Models

Model used for regression:

Models

Model used for regression:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- SVR (using *rbf*, linear, poly and sigmoid kernels)

Contents

Testing

Testing:

Testing

Testing:

- official test set unavailable (due to the ongoing competition)
- subset for training (70%) and testing (30%)
- 10–fold cross–validation

Overview of results

Table : Overview of results of age–group classification per language. Baseline scores are given for comparison

Language	Acc.	Prec.	Rec.	$F1^{(micro)}$	$F1^{(macro)}$
English	0.782	0.717	0.640	0.652	0.782
English (baseline)	0.326	0.081	0.250	0.122	0.326
Spanish	0.933	0.975	0.900	0.924	0.933
Spanish (baseline)	0.600	0.150	0.250	0.187	0.600

Overview of results

Table : Overview of results of gender classification per language. Baseline scores are given for comparison

Language	Accuracy	Precision	Recall	F1
English	0.956	0.888	1.000	0.941
English (baseline)	0.347	0.347	1.000	0.516
Spanish	1.000	1.000	1.000	1.000
Spanish (baseline)	0.400	0.400	1.000	0.571
Italian	1.000	1.000	1.000	1.000
Italian (baseline)	0.416	0.000	0.000	0.000
Dutch	1.000	1.000	1.000	1.000
Dutch (baseline)	0.454	0.454	1.000	0.625

Overview of results

Table : Overview of RMSE of personality traits regression per language. Baseline scores are given for comparison

Language	Extra	Stab	Agree	Consc	Open
English	0.122	0.179	0.153	0.140	0.130
English (baseline)	0.164	0.235	0.182	0.167	0.155
Spanish	0.080	0.143	0.103	0.155	0.131
Spanish (baseline)	0.123	0.220	0.149	0.211	0.183
Italian	0.048	0.123	0.067	0.086	0.099
Italian (baseline)	0.136	0.166	0.116	0.162	0.162
Dutch	0.087	0.112	0.129	0.064	0.041
Dutch (baseline)	0.137	0.197	0.155	0.115	0.116

Contents

Conclusion

- We succeeded to obtain results similar to other published works.
- Possible upgrade:
  - *Latent Semantic Analysis*

The End

Questions??