Ensemble-based classification for author profiling using various features

M. Czoków[1], M. Meina[1], K. Brodzińska[2], B. Celmer[1], M. Patera[1], J. Pezacki[1] and M. Wilk[1]

[1]Faculty of Mathematics and Computer Science, [2]Faculty of Physics, Astronomy and Informatics
Nicolaus Copernicus University, Toruń, Poland

Introduction

This poster summarize our approach to author profiling task – a part of evaluation lab PAN'13. We have used ensemble-based classification on a large set of features. Here, all the features are roughly described and evaluation for random forest (ensemble-based classifier obtaining the best accuracy) is presented.

Work methodology

Analytic dataset consists of:

- 8 groups of features,
- total number of features is 311 for English and 476 for Spanish.

Two approaches:

1. Random forest — eventually applied.
2. Committee for 8 weak classifiers.
   - 8 subsets of features
   - for each subset four classifiers tested (kNN, Linear SVM, SVM with RBF and Naive Bayes)
   - for each subset of features the best classifier took part in voting

Topic specific features

We applied Latent Semantic Analysis:

- With each document we associate 150 coefficients of different topics.
- In order to obtain this we create tf-idf weighted term-doc matrix M and approximate its singular value decomposition:

$$M \approx U_k \Sigma_k V_k,$$

  where U_k and V_k can be interpreted as term-topic matrix and topic-document matrix.
- In order to avoid overfitting, topic specific features enclosed in the analytic dataset are generated with application of 10-folds cross-validation.

Structural features

- Features that describe structure of conversations, e.g.: the number of conversations, paragraphs, sentences, special characters and words per sentences.
- Statistics for documents with more than one conversation:
  - minimum, maximum and average conversation length,
  - average edit distance between each pair of conversations.
- Statistics concerning hyperlinks and images.

Cluster analysis

- We created clusters on the base of two groups of features:
  - structural,
  - topic specific.
- To the set of features we added distances from centroids.
- **Behaviour profile ⇒ author profile**.

centroid	href_no	sen_no	word_no	href_word_ratio	avg_conv_len	new_line_no	tab_no
English corpora							
C1	0.820	6.372	119.764	0.027	395.533	12.103	7.460
C2	3.354	99.882	2419.265	0.000	11429.932	91.313	7.083
C3	23.879	45.204	921.405	0.009	1306.874	93.641	47.736
C4	3.712	43.678	962.547	0.000	3315.166	29.639	8.439
Spanish corpora							
C1	0.146	3.839	98.389	0.002	385.496	6.427	7.766
C2	3.745	1.203	4.152	0.992	27.819	6.0677	5.186
C3	0.850	46.452	1183.494	0.000	2542.832	19.344	78.775
C4	1.317	250.837	5945.458	0.000	25741.812	19.375	197.689

Dictionary-based features

In each document we counted number of:

- abbreviations,
- emoticons,
- badwords,
- basic emotion words (e.g. *anger, disgust, fear, joy, sadness, surprise*),
- connective words (e.g. *nevertheless, whatever, secondly*)
- words that have little semantical value (e.g. *I, the, own, him*)
- persuasive words (e.g. *you, money, save, new, results, health, easy*).

Text difficulty & readability

Features based on the following readability formulas: Flesch Reading Ease, Flesch-Kincaid Grade Level, Dale-Chall. These statistics are based on the number of words, sentences, syllables and difficult words (there is Dale-Chall list of 3 000 familiar words and thus, words, which are not on that list, are considered as *difficult*).

Results

Classification accuracy for Random Forest.

	gender	age	gender + age
English	0.632 ± 0.0019	0.611 ± 0.0019	0.653 ± 0.0019
Spanish	0.611 ± 0.0071	0.596 ± 0.0089	0.626 ± 0.0091
baseline	0.3333	0.5	0.1650

Experiment was conducted using k-cross validation with ($k = 10$). Minimum samples per leaf = 5, size of a set of feature for each tree was equal to $\sqrt{n\_features}$. Number of trees in the forest around 650.

Submission	Accuracy Total	Accuracy Gender	Accuracy Age	Adult Gender	Adult Age	Adult Both	Predator Gender	Predator Age	Predator Both	Runtime (incl. Spanish)
meina13	0.3894	0.5921	0.6491	6	8	6	72	41	41	383821541
pastor13	0.3813	0.5690	0.6572	1	8	0	72	32	32	2298561
mechti13	0.3677	0.5816	0.5897	2	6	2	52	29	20	1018000000
santosh13	0.3508	0.5652	0.6408	9	9	9	69	32	29	17511633
yong13	0.3488	0.5671	0.6098	6	1	1	28	30	17	577144695
ladra13	0.3420	0.5608	0.6118	9	9	9	72	33	33	1729618
ayala13	0.3292	0.5522	0.5923	3	2	1	53	34	26	23612726
gillam13	0.3268	0.5410	0.6031	1	4	0	72	30	30	615347

Submission	Accuracy Total	Accuracy Gender	Accuracy Age	Runtime (incl. English)
santosh13	0.4208	0.6473	0.6430	17511633
pastor13	0.4158	0.6299	0.6558	2298561
haro13	0.3897	0.6165	0.6219	9559554
flekova13	0.3683	0.6103	0.5966	18476373
ladra13	0.3523	0.6138	0.5727	1729618
jimenez13	0.3145	0.5627	0.5429	3940310
kern13	0.3134	0.5706	0.5375	18285830
yong13	0.3120	0.5468	0.5705	577144695
ramirez13	0.2934	0.5116	0.5651	64350734
aditya13	0.2824	0.5000	0.5643	3734665
jankowska13	0.2592	0.5846	0.4276	16761536
meina13	0.2549	0.5287	0.4930	383821541
gillam13	0.2543	0.4784	0.5377	615347
moreau13	0.2539	0.4967	0.5049	448406705
weren13	0.2463	0.5362	0.4615	11684955
cagnina13	0.2339	0.5516	0.4148	855252000

Sequences of parts of speech

- Preprocessing — each sentence tagged into sequence of parts of speech.
- For each document we calculated an average probabilities, that a tagged sequence from this document belongs to the respective classes (separately for gender and age).
- In order to do this we created n-gram models (we calculated conditional probabilities that for a given class a given tag occurs in a sequence, when it is preceded by a given sequence of $n-1$ length).

Errors

Numbers of errors and language mistakes in accordance with the list of 27 standardized ISO 27 errors' types.

Parts of speech

- Preprocessing — each sentance tagged into sequance of parts of speech.
- Frequencies of particular parts of speech in all conversations of each author.
- Much more parts of speech (features) for Spanish (this number is predefined by tagger).

	English Feature	English Inf. gain	Spanish Feature	Spanish Inf. gain
1	min_conv_len	0.0653	gram_n4_30s	0.0416
2	total_connective_words/total_sents	0.0653	gram_n5_30s	0.0363
3	avg_conv_len_words	0.0647	gram_n4_20s	0.0337
4	avg_conv_len	0.0644	gram_n5_20s	0.0246
5	total_abbreviations/total_sents	0.0642	gram_n4_male	0.0228
6	C1	0.0635	gram_n4_female	0.0228
7	gram_n6_20s	0.0631	total_uncategorized_errors/total_sents	0.0209
8	max_conv_len	0.0625	gram_n4_age	0.0207
9	C0	0.0624	gram_n5_age	0.0201
10	gram_n5_20s	0.0622	total_errors/total_sents	0.0197
11	gram_n6_age	0.0612	total_typographical_errors/total_sents	0.0177
12	total_badwords/total_sents	0.0604	new_line_count/sentence_count	0.0172
13	C3	0.0559	gram_n4_gender	0.0169
14	gram_n4_20s	0.0539	gram_n5_female	0.0163
15	gram_n6_30s	0.0524	gram_n5_male	0.0163
16	gram_n5_30s	0.0523	gram_n4_10s	0.0134
17	gram_n5_age	0.0518	gram_n5_gender	0.0127
18	total_abbreviations	0.0514	Fc_n	0.0107
19	word_count	0.0508	sps00_n	0.0107
20	gram_n4_30s	0.0503	gram_n5_10s	0.0100
21	total_badwords	0.0478	href_count	0.0095
22	total_persuasive_words/total_sents	0.0458	sentence_count	0.0090
23	sentence_count	0.0430	total_connective_words/total_words	0.0087
24	new_line_count/word_count	0.0404	Fp_n	0.0086
25	href_count	0.0397	UNK_n	0.0077
26	new_line_count/sentence_count	0.0385	href_word_ratio	0.0073
27	gram_n4_age	0.0380	new_line_count	0.0071
28	gram_n6_female	0.0369	word_count	0.0067
29	gram_n6_male	0.0369	rn_n	0.0066
30	gram_n4_male	0.0345	Fat_n	0.0061
31	gram_n4_female	0.0345	C2	0.0061
32	gram_n5_female	0.0344	Fs_n	0.0060
33	gram_n5_male	0.0344	avg_conv_len_words	0.0059
34	C2	0.0308	total_difficult_words/total_words	0.0057
35	total_difficult_words/total_words	0.0284	max_conv_len	0.0056
36	total_syllables/total_words	0.0283	C1	0.0055
37	gram_n4_10s	0.0268	new_line_count/word_count	0.0055
38	gram_n5_10s	0.0265	total_abbreviations	0.0054
39	gram_n6_gender	0.0252	C3	0.0053
40	gram_n6_10s	0.0250	ncmp000_n	0.0053
41	flasch_reading_easy	0.0241	avg_conv_len	0.0053
42	gram_n5_gender	0.0230	vmip1s0_n	0.0053
43	gram_n4_gender	0.0227	topic-85	0.0052
44	dale_chall_readability_formula	0.0216	topic-55	0.0050
45	total_badwords/total_words	0.0210	topic-17	0.0050
46	flesch_kincaid_grade_level	0.0209	topic-116	0.0049
47	total_emoticons/total_words	0.0206	topic-8	0.0049
48	total_emoticons	0.0202	topic-147	0.0048
49	total_abbreviations/total_words	0.0187	topic-36	0.0048
50	total_emoticons/total_sents	0.0180	pp1cs000_n	0.0048