Ensemble-based classification for author profiling using various features

Notebook for PAN at CLEF 2013

Michał Meina[1], Karolina Brodzińska[2], Bartosz Celmer[1], Maja Czoków[1], Martyna Patera[1], Jakub Pezacki[1], and Mateusz Wilk[1]

[1] Faculty of Mathematics and Computer Science,
Nicolaus Copernicus University, Toruń, Poland
{mich, bcelmer, maja, patera, kpezacki, mtszwlk}@mat.umk.pl
[2] Faculty of Physics, Astronomy and Informatics,
Nicolaus Copernicus University, Toruń, Poland
karola@fizyka.umk.pl

Abstract This paper summarize our approach to author profiling task – a part of evaluation lab PAN'13. We have used ensemble-based classification on large features set. All the features are roughly described and experimental section provides evaluation of different methods and classification approaches.

1 Introduction

Main goal of authorship analysis is to retrieve information carried by the text about specified characteristics of its author. Characteristics may relate to demography, culture, nationality, personality, etc. Such analysis, for example, may be applied to discover the relation between profile of a person and his/her opinion on particular subject. It may also help in recognition of criminal and terrorist activities. In this paper we describe our approach to Author Profiling task, which was a part of the PAN 2013 competition[3]. The goal was to determine gender and age of given chat conversations' authors.

We decided to apply ensemble-based classification methods because of their potential for more effective recognition of complex patterns. Ensemble methods combines several weak classifiers into one classifier, which is more effective than the individual ones. The features were created on the base of structure, stylometry and semantics of the text. The diversity of explored properties of the dataset assures high independence of the features.

2 Data

The dataset was divided into two language groups (English and Spanish). Each group included conversations stored in XML files (grouped by author). There were 236 000 files (564 413 conversations, 180 809 187 words) in English and 75 900 files (126 453 conversations, 21 824 198 words) in Spanish.

[3] http://pan.webis.de/

In the data preprocessing phase we employed standard techniques for text cleaning and tokenization. Regular expressions was used to strip out most of html tags leaving special tokens for hyperlinks and embedded images. Some of feature extraction procedures required word and sentence tokenization. Word tokenization is easy task and can be addressed by regular expressions. For sentence splitting we have used unsupervised algorithm [11] to build a model for abbreviation words, collocations, and words that are at the beginning of sentences.

We noticed that many conversations in dataset seemed to be a spam (probably produced by some chatterbots used for commercial purposes). It is reasonable to assume that age and gender characteristics in spam-like text have different rationale. One can say, for example, that advertisement of certain products is more oriented towards one gender group than the other one. Also mixing together chatter bots and humans conversations can introduce unnecessary noise in phase of learning of a classifier. Because of that we put extra effort to discriminate spam-like over human-like chats.

3 Feature Engineering

With each document we associated a collection of features, which were employed by classification algorithms to identify the age and gender of the document's author. In the final versions of our analytic dataset (that was used for classifier training) there are 311 and 476 features, respectively for the texts written in English and Spanish.

These features can be divided into groups roughly described below.

3.1 Structural Features

Structure of a document is proven to be important feature in various classification problems regarding text mining and authorship profiling. There were not much previous work, however, that can be easily adopted in our approach. Therefore we have engineered very simple features, that not necessarily carry information about authors profile but can be used to group the documents into similarly structured conversations (eg. long and shorts ones). This approach enables us to discover more subtle high-level features in various document groups.

Some examples of such *structural* features are: the number of conversations, paragraphs, sentences and words per sentences, number of special characters, etc. If the person conducted more than one conversation we measure minimum, maximum and average conversation length. The usage (absolute count and ratio) of hyperlinks and images, and whether they were used at the beginning or at the end of the conversation. After more thorough investigation we discovered that this particular feature can be also useful in spam detection. We also noticed that chatterbots seem to perform very similar conversations therefore we measured the Jaccard similarity coefficient of individual conversations and enclosed this average edit distance into an analytic dataset.

3.2 Parts of Speech

Argamon et.al [5] reports that usage of particular part of speech in some cases can be exploited effectively in gender detection of texts' authors. Therefore we have measured

relative frequencies of particular parts of speech in whole conversation of each author. In order to measure these proportions we have used part-of-speech tagger (pos-tagger) available in nltk toolkit. For English texts we used pre-trained tagger (more details in [6]) and for Spanish texts we trained trigram tagger (with respectively bigram and unigram tagger as backoffs) over annotated corpora [12].

3.3 Exploration of Sequences of Parts of Speech

We employ *n-gram language model* (separate for age and gender problem) to create a vector of features. An n-gram in our case is a contiguous sequence of n pos-tagged words from a given sequence. In the learning phase for each class C we selected all files ascribe to it and we count the number of occurrences of each n-gram and (n-1)-gram in chosen files. Next, on the basis of determined values, for the given class C, the probabilities $\mathbb{P}_C(x_i|x_{i-n+1}, ..., x_{i-1})$ are estimated, for more details see [7].

Obtained probabilities are employed to generate vector of features for a given conversation. To this end for each sentence $x_{-n+2}, ..x_{l+1}$ in the conversation and for each class C is calculated a vector of the posterior probabilities $\mathbb{P}(C|x_{-n+2}, ..x_{l+1})$. In order to calculate this value, the probability of occurrence of the sentence in the files ascribe to the class C is calculated:

$$\mathbb{P}_C(x_{-n+2}, ..x_{l+1}) = \prod_{i=1}^{l+1} \mathbb{P}_C(x_i|x_{i-n+1}, ..., x_{i-1}). \tag{1}$$

In the equation 1 we make a $(n-1)^{th}$ *order Markov assumption*. Next, the parameter $\mathbb{P}_C(x_{-n+2}, ..x_{l+1})$ is multiplied by the prior class probability $\mathbb{P}(C)$. Values obtained in this way for all classes and normalized to sum up to 1, create a vector of the posterior probabilities $\mathbb{P}(C|x_{-n+2}, ..x_{l+1})$. We sum these vectors from all sentences in the conversation. The obtained vector is again normalized to sum up to 1 and is returned as the final feature vector.

For both languages we employed 3 models for $n \in \{4, 5, 6\}$. So, for age we have $2 \cdot 3 = 6$ features and for gender $3 \cdot 3 = 9$. A n-gram language model has a very large number of parameters. For both languages, we applied part-of-speech taggers, which maps words to the set of 46 different symbols, what results in 46^n different possible n-grams. Even with a huge set of training sentences, many of the n-grams do not occur in the training set. It is serious problem, since some of these sequences appear in conversations, which we want to classify. We apply discounting method [7] to smooth date for 4-gram model. This method slightly decreases values of non-zero probabilities $\mathbb{P}_C(x_i|x_{i-3}, ..., x_{i-1})$ and ascribes positive, close to 0 values to all $\mathbb{P}_C(x_i|x_{i-3}, ..., x_{i-1})$, for which $x_{i-3}, ..., x_{i-1}, x_i$ does not occur in the training set. This procedure is not employed for models with $n = \{5, 6\}$ since keeping their all parameters in memory is too expensive. For them we estimate parameters on the basis of parameters of models with lower n. If for n-gram model $\mathbb{P}_C(x_i|x_{i-n+1}, ..., x_{i-1})$ is missing, we estimate it by $\delta \cdot \mathbb{P}_C(x_i|x_{i-n+2}, ..., x_{i-1})$, where $\mathbb{P}_C(x_i|x_{i-n+2}, ..., x_{i-1})$ is taken from (n-1)-gram model. If $\mathbb{P}_C(x_i|x_{i-n+2}, ..., x_{i-1})$ is also missing, we apply $\delta \cdot \lambda \cdot \mathbb{P}_C(x_i|x_{i-n+3}, ..., x_{i-1})$. For English conversations $\delta = 0.2, \lambda = 0.5$ for gender and $\delta = 0.9, \lambda = 0.9$ for age. For Spanish conversations $\delta = 0.2, \lambda = 0.5$ for gender and $\delta = 0.7, \lambda = 0.9$ for age. These values were chosen experimentally.

3.4 Text difficulty & readability

In order to determine text difficulties we applied several readability tests for the documents: Flesch Reading Ease, Flesch-Kincaid Grade Level and Dale-Chall readability formula. They are based on the number of words, sentences, syllables and difficult words (there is Dale-Chall list of 3 000 familiar words [4] and thus, words, which are not on that list, are considered as *difficult*). For details see [9] and [8].

3.5 Dictionary-based Features

We wanted to examine the intensity of words and expressions of particular types. In each document we counted number of abbreviations, emoticons and badwords. We useed NodeBox [1] to count the number of basic emotion words (*anger, disgust, fear, joy, sadness, surprise*), connective words (*nevertheless, whatever, secondly, etc. and words like I, the, own*, him which have little semantical value) and persuasive words (*you, money, save, new, results, health, easy,* etc.).

3.6 Errors

Numbers of errors and language mistakes is determined by using LanguageTool [2] in accordance with the list of 27 standardized ISO 27 error' types that can be found in [3].

3.7 Topic Specific

By topic specific features we understand coefficients corresponding to the representation of the document as a linear combination of 150 (for each language) "statistical topics" estimated using Latent Semantic Analysis (LSA) technique [**?**]. In short, crucial point of LSA is k-rank approximation of singular value decomposition of term-doc matrix:

$$M \approx U_k \Sigma_k V_k, \tag{2}$$

where M is tf-idf weighted term-doc matrix, U_k and V_k can be interpreted as term-topic matrix and topic-document matrix both in low rank approximation. We have computed this decomposition for English and Spanish corpus separately using rank $k = 150$ (chosen experimentally). Next, previously unseen document can be represented in latent (topic specific) space by "fold-in" operation:

$$d' = \Sigma_k^{-1} U_k^T d, \tag{3}$$

where d is vector (bag-of-words) representation of a document and $d' = w_i$ is k-length vector in which $|w_i|$ indicates how this document contribute to i-th topic. We enclosed those values in analytic dataset avoiding classifier overfitting using 10-folds to obtain those topic specific features. Each 10% of documents was treated as unseen and folded into latent representation estimated with 90% rest of corpus.

3.8 Structural and Topic Specific Centroids

We employed a cluster analysis on two subsets of features: structural and topic specific (defined in previous section). This is used to differentiate behavioral profiles of authors. The basic behavior profile can be perceived as a preliminary authorship analysis; most significantly we can distinguish human from chatter bots (similarly to Gianvecchio et. al. [10] but using different features). Centroid of a cluster either structural $\{C_1, ..., C_4\}$ or topical $S_1, ..., S_m$ ($m = 30$ for English and $m = 17$ for Spanish) is used as typical behavior or conversation topic therefore euclidean distance to each cluster is enclosed into analytic dataset.

We obtained clusters using K-Means algorithm with Silhouette score as criterion for estimation the number of clusters. Below table depicts structural centroids.

Table 1. Classification accuracy

centroid	href_count	sentence_count	word_count	href_word_ratio	avg_conv_len	new_line_count	tab_count
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

3.9 Natural Language Model

N-grams on words were used in order to preserve language model. We estimated top-n n-grams for each gender-age group, next we merged the results and computed Mutual Information (equation below) to measure how much information every ngrams carry about each group.

$$I(C,T) = \sum_{c\in 0,1} \sum_{t\in 0,1} \mathbb{P}(C=c, T=t) log_2 \left(\frac{\mathbb{P}\left(C=c, T=t\right)}{\mathbb{P}\left(C=c\right)\mathbb{P}\left(T=t\right)} \right), \qquad (4)$$

where $\mathbb{P}(C = 0)$ represents the probability that randomly selected author is a member of particular age-gender group and $\mathbb{P}(C = 1)$ represents probability that it isn't. Similarly, $\mathbb{P}(T = 1)$ represents the probability that a randomly selected chatter contains a given n-gram, and $\mathbb{P}(T = 0)$ represents the probability that it doesn't.

4 Experimental Results

We tested how individual classifiers (Naive Bayes, Random Tree, SVM) as well as ensemble methods (Random Forest, Classifiers Committees) work with our features set. The random forest method gives the best results and those are presented in Table 2.

Table 2. Classification accuracy

	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	0.1650	0.5	0.3333

Experiment was conducted using k-cross validation with ($k = 10$). The final Random Forest classifier was trained on 12-core machine using about 30GB of RAM. Training took 45 minutes. Parameters of the classifier were estimated by trial-and-error: minimum samples per leaf = 5, size of feature set for each tree was equal to $\sqrt{n\_features}$. The classification accuracy convergence for the number of tree in the forest larger than 1000, but due to memory constraints on test machine we lowered this parameter to 666, which surprisingly fits the memory almost exactly. Also because memory constraints we did not used a n-gram model from subsection 3.9.

Among individual classifiers we tested also popular classifiers such as kNN (k = 5), linear SVM, SVM with RBF and Naive Bayes. We also conducted experiments for two simple ensembles composed of those classifiers: majority and weighted committee. Due to the time constraint we did not performed k-fold cross-validation. All classifiers were trained on 99% of feature vectors and tested on the remaining 1%. The date set was divided into training and test set by means of stratified cross-validation. The experiments were conducted on 48-core machine with 200 GB of RAM available. The obtained results are presented in Table 3.

Table 3. Classification accuracy. Spanish language.

	gender	age	gender + age
kNN	0.534	0.535	0.263
Naive Bayes	0.553	0.520	0.016
Linear SVM	0.6123	0.595	0.357
SVM with RBF	0.529	0.573	0.279
Majority committee	0.584	0.552	0.264
Weighted committee	0.573	0.552	0.242
baseline	0.1650	0.5	0.3333

We also performed similar tests on the subsets of features. The description of each of 9 subset can be found in Sect. 3 - "Feature engineering". We tested all of the four classifiers: kNN, Linear SVM, SVM with RBF and Naive Bayes. Further, for each subset of features we chose the best classifier and built committee from all 9 thus obtained classifiers. The results that we achieved are presented in Table 4.

The subsets of features: (I) Structural features, (II) Parts of speech, (III) Exploration of sequences of parts of speech, (IV)) Test difficulty, (V) Dictionary-based features, (VI) Errors, (VII) Topical features, (VIII) Topical centroids, (IX) Structural centroids.

Table 4. Classification accuracy. Subset of features. Spanish language.

	gender	age	gender + age
I.	0.541 (Naive Bayes)	0.564 (Linear SVM)	0.306 (Naive Bayes)
II.	0.556 (Linear SVM)	0.562 (Linear SVM)	0.305 (Linear SVM)
III.	0.581 (Linear SVM)	0.597 (Linear SVM)	0.346 (Liear SVM)
IV.	0.514 (Linear SVM)	0.561 (Linear SVM)	0.289 (Linear SVM)
V.	0.536 (Linear SVM)	0.561 (Linear SVM)	0.3014 (Linear SVM)
VI.	0.568 (kNN)	0.582 (Naive Bayes)	0.315 (Naive Bayes)
VII.	0.529 (Naive Bayes)	0.562 (Linear SVM)	0.363 (kNN)
VIII.	0.541 (Naive Bayes)	0.565 (Linear SVM)	0.284 (SVM)
IX.	0.541 (Linear SVM)	0.561 (Linear SVM)	0.280 (Linear SVM)
Majority committee	0.604	0.561	0.308
Weighted committee	0.664	0.405	0.069
baseline	0.1650	0.5	0.3333

Due to the long time of execution the tests for English language were not performed.

References

1. http://nodebox.net/code/index.php/Linguistics
2. http://www.languagetool.org/
3. http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#lqissue-typevalues
4. Word list: Dale-chall list of simple words, http://www.usingenglish.com/resources/wordcheck/list-dale-chall+list+of+simple+words.html
5. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. TEXT 23, 321–346 (2003)
6. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly, Beijing (2009), http://www.nltk.org/book
7. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th annual meeting on Association for Computational Linguistics. pp. 310–318. ACL '96 (1996)
8. Dale, E., Chall, J.S.: A formula for predicting readability. Educational Research Bulletin 27(1), pp. 11–20+28 (1948)
9. Flesch, R.: A new readability yardstick. Journal of Applied Psychology 32(3), p221 – 233 (June 1948)
10. Gianvecchio, S., Xie, M., Wu, Z., Wang, H.: Humans and bots in internet chat: measurement, analysis, and automated classification. IEEE/ACM Trans. Netw. 19(5), 1557–1571 (Oct 2011)
11. Kiss, T., Strunk, J.: Unsupervised multilingual sentence boundary detection. Comput. Linguist. 32(4), 485–525 (Dec 2006)
12. Martí, M.A., Taulé, M., Márquez, L., Bertran, M.: Cess-ece: A multilingual and multilevel annotated corpus (2007), http://www.lsi.upc.edu/~mbertran/cess-ece

Table 5. The table lists 50 features with the highest Information Gain ratio

	English		Spanish	
	Feature	Inf. gain	Feature	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