Statistical Learning Methods for Profiling Analysis

Notebook for PAN at CLEF 2015

Lesly Miculicich Werlen

Computer Science Deparment - University of Neuchâtel
Lesly.Miculicich@unine.ch

Abstract Author profiling is the task to infer some information about an author by analyzing her/his writing style. It's application in forensics, business intelligence and psychology makes this topic interesting for researching. In this notebook, we present our baseline approach using SVM and Linear Discriminant Analysis (LDA) classifiers. We analyze features obtained from LIWC dictionaries, these are frequencies of use words by categories, which gives a general view about how the author writes and what he/she is talking about. According the experimental results, those are significant features to differentiate gender, age-group and personality. Although they are relatively few (not more than 100), they allow to discriminate with an acceptable accuracy.

1 Introduction

Studies have demonstrated evidence of differences in the writing style according the gender and age of the authors. These differences are detected with the use of function-words and content-words. The function-words define how the person use the grammar and build sentences. On the other hand, content-words indicate what the person is talking about. For example, Pennebaker [7] found that women tend to use more personal pronouns and words referring to emotions. By the contrary, men tend to use more nouns, prepositions and big words (defined as words with more than 6 characters).

In the case of age, Pennebaker [7] found that younger writers use more personal pronouns in first person and past tense verbs; while older writers tend to use more articles, nouns prepositions and future tense verbs. Another example of this differences is in Schler et al. [9], where these authors found that men's writing is more related to money, job and TV, while women's writing is more related to family, sex and eating. In the case of age, younger people use to write more about sports, friends and emotions; while older people write more about money, job, and family.

More recently studies expanded this analysis to determinate weather the personality influence the writing style too. For example Yarkoni [12] presented a detailed work were he found that extroverted people are more likely to speak about leisure activities, family and other persons than non extroverted. People open to new experiences talk more about friends, time and positive emotions than other people; and some other similar relations for all different personalities.

Based on this evidence, we believe that the word categories are important features to determine the profile of an author of a text, and we will try to measure how much they can tell us about the authors of tweets. Therefore, in this notebook, we present a baseline author profile classifier based on statistical learning over word category features. First, we present the feature pre-processing, extraction and selection. Then, the classification models to identify gender, age group, and personality. And finally, we present the description of the experiments, the results and the conclusions.

2 Features

2.1 LIWC

The studies mentioned previously used the Linguistic Inquired and Word Count (LIWC) [6]. This tool propose a list of word categories, each category is formed by agreement between at least three judges. Then, given a text, it counts the number of words that the text have per each category. The idea is to know how frequent a person use each word category and with that data estimate some information about him or her.

The LIWC categories are grouped in linguistic dimensions (e.g., part-of-speech); content dimensions(e.g., emotions and activities); or spoken dimensions (e.g. fillers and no fluencies markers). LIWC has dictionaries in a variety of languages, for this experiment English, Italian, Spanish and Dutch dictionaries are used. Each dictionary contains the list of predefined categories and the words associated with them, for example for the category *positive emotions* some associated words are *“fun", “nice", “succes",* etc. They were created over reiterative process of human judgment and were tested in several times in different studies to assure their validity.

2.2 Additional Categories

Additionally, we include other two groups of categories: punctuation marks and tweet. We include seven categories in punctuation marks: *question mark, exclamation, period, comma, colon, semi-colon* and *all punctuation*. The last one groups any punctuation mark including the mentioned before. The tweet categories are added for the nature of the corpus data and because they are frequently employed by tweet users: *emoticons, hyper-link, hashtag* and *references to other users*. In Table 1 we can see a summary of the total categories analyzed in this experiment.

Table 1. Number of categories used as features per language’s dictionary

Dictionary	Linguistic	Content	Spoken	Punctuations	Tweet	Total
English	24	40	3	7	4	78
Dutch	14	55	2	7	4	82
Italian	14	74	0	7	4	99
Spanish	14	55	2	7	4	82

Table 2 shows one example of our analysis with LIWC and the additional categories in the tweet: *"ummm yesterday, I was with @Tim in a beautiful concert :) ."*.

Table 2. Example of analysis with LIWC and additional categories for the tweet: *"ummm yesterday, I was with @Tim in a beautiful concert :) ."*

Dimension	Category	Words	Counting
Linguistic	Word count	*all words*	10
	Function words	*“I",“ was", “with", “in", “a"*	5
	Pronoun	*“I"*	1
	Article	*“a"*	1
	...	....	...
Content	Affection	*“beautiful"*	1
	Positive Emotion	*“beautiful"*	1
	Time	*“yesterday"*	1
	...	....	...
Spoken	No fluency	*“ummm"*	1
Punctuation	All punctuation	, .	2
	...	...	...
Tweet	Emoticon	*:)*	1
	Reference to other user	*“@Tim"*	1

2.3 Feature Extraction

The tweets were given in XML files. Each XML file was processed to extract only the text information given as scale. First, the text was tokenized using white space (including tab, change of line among others) and punctuation marks as separation characters. We consider the apostrophe as part of the word, for example *“she's"* is consider a single word. This choose is related to the LIWC dictionaries that consider them in this way. Once the tokens are obtained, they were counted in the corresponding categories. As mentioned before, in the dictionary each category has a set of words, so, if the token is part of the set of word of a category then it sums one to that category. One token can appear in many categories (e.g. *“I"* as *pronoun* and *function word*).

The granularity of the models is set by user and not by tweet, so, there is a vector of categories per each user. Once we complete the counting per user, we divide each count by the total number of words by user in other to obtain the relative frequencies. Finally, we keep the frequencies in a matrix format, where the columns are the word categories and the rows are the users. Thus, each row represent the distribution over LIWC categories for the given user.

One additional step is performed before the feature selection, the relative word frequencies x are scaled by calculating the $z - score$ respect to each category according to the following formula:

$$z = \frac{x - \mu}{\sigma}. \tag{1}$$

where, μ and σ are the mean and standard deviation of the frequencies in each category.

The frequency of use of words in not uniform in a language. Some of them are highly used (e.g., function words) and some others have low frequency of use (e.g., topic related words), the relative frequencies are scaled because we need to compare them obtaining their use related to each particular category and not to the general use of language.

2.4 Feature Selection

Even we have a reduce set of features, we need to ignore noisy and irrelevant features before applying a classification scheme. Additionally, it derives in an easier linguistic explanation about the key features to discriminate among the different classes.

Gender and Age Group Fourth feature selection methods were evaluated to determine the more suitable for the data: Manual selection, Information Gain, Odd ratio and Support Vector Machine Recursive Feature Elimination (SVM RFE). The manual selection was based on [7] and [5], where it is explained which are the more general categories to differentiate an author according she/he's age and gender. The Information Gain and Odd Ratio were based on the study of Sebastiani [11] where he compares different methods for feature reduction in text categorization. The SVM RFE proposed by Guyon [3] is a backward feature elimination using SVM, it eliminates one feature at time given a ranking criteria. The three last methods were implemented with Weka [4]. For their evaluation, three different classifiers were tested and the results were compared according the accuracy of classification. The best subset was obtained with SVM RFE.

Personality In the case of personality, the number of classes to discriminate is larger than the previous models and it is more difficult to associate specific categories to each class. Consequently, the methods mentioned before did not have significant improvement in comparison of using the full set of features. So for this case, we applied Forward-Backward Feature Selection, trying to improve the Root Mean Squared Error (RMSE).

3 Classification

3.1 Gender and Age Group

These classes were defined as categorical. We have two classes for gender: “Male" and “Female", and fourth classes for age: “18-24", “25-34", “35-49", and “50-xx". The classification was made with ν-SVM [10], which is a variant of the original SVM but with an easier interpretation for the cost parameter called ν. In the experiments, ν was set to 0.01 and we used radial kernel. The implementation was made in R with the library “e1071".

3.2 Personality

In the case of personality, we define one model per each personality. Our first approach was to define the classes as categorical without taking into account the order of the score of the personality, so we have 11 classes from “-0.5" to “0.5" with one decimal of difference. We choose two classifiers: ν-SVM with ν set to 0.01 and with radial kernel, and Linear Discriminant Analysis (LDA). The implementation was done in R with the libraries “e1071" and “MASS".

4 Experiments and Results

4.1 Training

The corpus to develop the models is a training set of tweets in English, Spanish, Italian and Dutch given by PAN 2015 Author Profiling task [8]. The validation was made measuring the accuracy for gender and age-group, and RMSE for personality (according the specification of the task). We used the full training set with leave-one-out validation. The results are showed in Tables 3 and 4.

Table 3. Gender and Age Group: Accuracy in %. Leave-one-out validation with training data set. Feature selection with SVM RFE and classification with SVM

	English	Dutch	Italian	Spanish
Gender	86	97	92	94
Age group	77	-	-	69

Table 4. Personality: RMSE rounded to two decimals. Leave-one-out validation with training data set. Feature selection with Back-Forward Propagation and classification with LDA and SVM

	English		Dutch		Italian		Spanish	
Personality	LDA	SVM	LDA	SVM	LDA	SVM	LDA	SVM
Extroverted	0.16	0.16	0.14	0.13	0.21	0.18	0.16	0.19
Stable	0.22	0.21	0.12	0.22	0.18	0.18	0.18	0.20
Agreeable	0.16	0.15	0.23	0.15	0.21	0.16	0.15	0.16
Conscientious	0.15	0.15	0.18	0.12	0.18	0.12	0.14	0.19
Open	0.15	0.15	0.16	0.12	0.21	0.15	0.12	0.17

In the feature selection, the experiment shows that there are some categories to discriminate between gender which are independent of language while other are different for each language, and the same patter for the other models. Tables 5, 6, and 7 contain the common categories that were found in one or more languages.

The training and testing are implemented separately. The outputs are the models and the vectors of means and standard deviation calculated with the training data. These vectors are used to calculate the $z - score$ of the testing data. This step corresponds to Software 1 of TIRA [2].

Table 5. Selected Features for Gender: Common features among one or more languages

Linguistic	Prepositions, word count, you, pronouns
Content	Family, affect, space, swear, feel, emotions, body, home, work, TV, money, future, motion, school, inclusion (and, we, both), exclusion (or, either, but)
Spoken	None
Punctuations	Question mark, exclamation mark, colon
Tweet	Emoticon, reference to other users, hyper-links

Table 6. Selected Features for Age-group: Common features among one or more languages

Linguistic	Prepositions
Content	Anger, body, optimist, insight, discrepancy, inhibition(block, constraint, deny)
Spoken	None
Punctuations	Comma
Tweet	Reference to other users, hyper-links

Table 7. Selected Features for Personality: Common features among one or more languages

Extroverted	Word count, big words, pronouns, I, we, us, others, article, social, family, emoticons, reference to other users
Stable	Pronouns, oneself, we, us, others, article, affection, positive emotions, optimist, anxiety, sadness, anger, emoticons, reference to other users, hyper-links
Agreeable	Pronouns, I , others, prepositions, inhibition, sadness, certain, see, listen, discrepancy, causation, cognitive process, emoticons, reference to other users, hyper-links
Conscientious	Pronouns, I , us, others, time, present, past, work, motion, home, optimist, positive emotions, number, reference to other users, hyper-links
Open	Pronouns, I , us, others, negation, preposition, number, affection, optimist, certain, discrepancy, cause, tentative, see, insight, emoticons, reference to other users, hyper-links

4.2 Testing

The corpus to test was given by PAN 2015 Author Profiling task [8]. The parameters are the input files and the models. This step corresponds to Software 2 on TIRA [2].

Table 8 shows the result for the testing. In almost all cases our solution performs better than the average with less runtime than the majority. The best global results were in Dutch and English, and the worse with Italian. In the case of gender and age-group, the results were good comparing with the state of the art using similar features, Argamon et al. [1] reported 72% accuracy to distinguish gender and 67% for age-group (having 3 groups). Specially in the case of Spanish, where we obtained 92% of accuracy in gender. Nevertheless, the accuracy of classification of age-group was close to average. In the case of personality, the results are also good taking into account the difficulty of the data: bigger number of classes to discriminate many of them with very few or none samples to train, and the few quantity of features used (less than 100). The selected runs for testing personality where using SVM classifier.

The global ranking of our solution for English was 7th over 22, Dutch 5th over 20, Italian 9th over 19, and Spanish 8th over 21.

Table 8. Testing results: “GLOBAL" is the total performance of the solution, “Gender" and “Age" are measured by accuracy in %,“BOTH" is the accuracy when gender and age were both well classified. The personality traits were measure with RMSE rounded to two decimal points, “RMSE" is the average of all personalities traits.

English

Performance	GLOBAL	BOTH	Gender	Age	RMSE	Extrovert.	Stable	Agreeable	Conscient.	Open	Runtime
Best	79	73	86	84	0.14	0.13	0.20	0.13	0.11	0.12	02:38:33
Our solution	**71**	**57**	**79**	**69**	**0.15**	**0.13**	**0.22**	**0.13**	**0.13**	**0.12**	**00:00:12**
Mean	67	51	71	69	0.18	0.16	0.24	0.16	0.16	0.16	03:48:25
Worse	52	22	50	41	0.24	0.23	0.32	0.22	0.22	0.26	05:23:51

Dutch

Performance	GLOBAL	BOTH	Gender	Age	RMSE	Extrovert.	Stable	Agreeable	Conscient.	Open	Runtime
Best	94	-	97	-	0.06	0.08	0.06	0.00	0.10	0.04	00:00:01
Our solution	**85**	**-**	**81**	**-**	**0.12**	**0.12**	**0.13**	**0.10**	**0.14**	**0.10**	**00:00:10**
Mean	78	-	70	-	0.14	0.15	0.17	0.14	0.14	0.11	00:05:17
Worse	67	-	47	-	0.25	0.21	0.28	0.28	0.24	0.24	01:07:09

Italian

Performance	GLOBAL	BOTH	Gender	Age	RMSE	Extrovert.	Stable	Agreeable	Conscient.	Open	Runtime
Best	87	-	86	-	0.10	0.07	0.16	0.05	0.11	0.10	00:00:01
Our solution	**74**	**-**	**64**	**-**	**0.15**	**0.11**	**0.17**	**0.12**	**0.17**	**0.19**	**00:00:12**
Mean	74	-	64	-	0.16	0.12	0.21	0.14	0.15	0.18	00:02:46
Worse	60	-	42	-	0.21	0.19	0.26	0.22	0.25	0.25	00:17:18

Spanish

Performance	GLOBAL	BOTH	Gender	Age	RMSE	Extrovert.	Stable	Agreeable	Conscient.	Open	Runtime
Best	82	77	97	80	0.12	0.13	0.16	0.10	0.10	0.11	00:00:02
Our solution	**73**	**63**	**92**	**68**	**0.16**	**0.19**	**0.20**	**0.13**	**0.14**	**0.17**	**00:00:13**
Mean	67	52	79	62	0.18	0.18	0.22	0.16	0.17	0.16	00:10:36
Worse	50	22	56	36	0.27	0.30	0.29	0.26	0.27	0.27	01:00:24

5 Conclusions

The present approach using LIWC Categories has demonstrated being a good solution regarding the limitation of having a few quantity of features compared with other solutions. According to the testing results, it had better performance than the average state of the art. Moreover, it is simple and efficient. But the most important point is that we can linguistically justify the classification decision because we can know which are the key features for the decision process. Deeper analysis is needed to extract the explanation of correct and incorrect assignment of classes; and to compare the differences in the results using SVM or LDA classifier. We think that it can be improved in future with a finer analysis of features and selection methods, and with a more appropriate definition of the classes and modeling for age-group and personality traits.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM 52(2), 119–123 (2009)
2. Gollub, T., Stein, B., Burrows, S.: Ousting ivory tower research: towards a web framework for providing experiments as a service. In: Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1125–1126. ACM (2012)
3. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning 46(1-3), 389–422 (2002)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
5. Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker, J.W.: Gender differences in language use: An analysis of 14,000 text samples. Discourse Processes 45(3), 211–236 (2008)
6. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates 71 (2001)
7. Pennebaker, J.: The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury Publishing (2011)
8. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: Cappellato L., Ferro N., Gareth J. and San Juan E. (Eds). (Eds.) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR-WS.org (2015)
9. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. vol. 6, pp. 199–205 (2006)
10. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural Computation 12(5), 1207–1245 (2000)
11. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) 34(1), 1–47 (2002)
12. Yarkoni, T.: Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. Journal of Research in Personality 44(3), 363–373 (2010)