Use of Language and Author Profiling: Identification of Gender and Age

Francisco Rangel[1,2], Paolo Rosso[2]

[1] Autoritas Consulting S.A., C/ Lorenzo Solano Tendero 7
28043 Madrid, Spain
francisco.rangel@autoritas.es
http://www.kicorangel.com

[2] Natural Language Engineering Lab, ELiRF,
Universitat Politècnica de València, Camino de Vera S/N
46022 Valencia, Spain
prosso@dsic.upv.es
http://users.dsic.upv.es/~prosso

Abstract. "In the beginning was the Word, and the Word was with God, and the Word was God". Thus, John 1:1[1] begins his contribution to the Holy Bible (one of the most-distributed book in the world with hundreds of millions of copies[2]), the importance of the word lies in the essence of human beings. The discursive style reflects the profile of the author, who decides, often unconsciously, about how to choose and combine words. This provides valuable information about the personality of the author. In this paper we present our approach to identify age and gender of authors based on their use of language. We propose a representation based on stylistic features and obtain encouraging results with a SVM-based approach on the PAN-AP-13[3] dataset.

1 Introduction

Knowing the profile of an author could be of key importance. For instance, from a forensic linguistics perspective being able to know what is the linguistic profile of a suspected text message (language used by a certain type of people) and identify characteristics (language as evidence) just by analyzing the text would certainly help considering suspects. Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, what types of people like or dislike their products.

In previous work we carried out a statistical study of how the language is used in Spanish in different channels of Internet, concretely what grammatical categories

[1] http://www.biblegateway.com/passage/?search=John+1&version=KJV
[2] http://en.wikipedia.org/wiki/List_of_best-selling_books
[3] Dataset for the Author Profiling task of the PAN 2013 http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html

people use in channels such as Wikipedia[4], newsletters, blogs, forums, Twitter[5] and Facebook[6]. In a recent work we investigated how the use of language could provide us enough evidences to identify the six basic emotions of Ekman[7]. We proposed a set of stylistic features and obtained competitive results in the identification of such emotions. We also carried out an exhaustive analysis of how the language varies by gender, topic and emotion, and all the possible combinations.

Based on the results of our previous works, in this paper we focus on the cognitive traits that make us different by gender and age. For that, we propose a set of features to represent texts written by anonymous authors, on the basis of stylistic features. With this set of features we aim to model the differences by age and gender in order to use them in a machine learning approach. We used a SVM method that we trained and tested with the PAN-AP-13 dataset. The obtained results are encouraging, although more in-depth features need to be investigated.

In Section 2 we present the state of the art, describing related work on author profiling. Furthermore we present the theoretical framework based on research on neurology. In Section 3 we describe our approach in detail together with the proposed features. In Section 4 we present the dataset used, the machine learning method and the evaluation measures. In Section 5 we discuss the experimental results. In Section 6 we draw some conclusions and discuss the future work.

2 Related work

Several are the interesting works on author profiling from the perspective of the common theoretical framework which involves several disciplines such as psychology, (computational) linguistics or even neurology.

2.1 Computational linguistics approaches

Several areas such as psychology, linguistics and, more recently, natural language processing are interested on studying how the use of the language varies according to the profile of the author. Pennebaker et al. (2003) connected the use of the language with traits such as gender, age, native language and so on. Argamon et al. (2003) used function words and the part-of-speech to predict gender of the authors of written texts from the British National Corpus, and Holmes and Meyerhoff (2003); Burger and Henderson (2011) have also investigated in obtaining age and gender from formal texts. Authors like Koppel et al. (2003); Schler et al. (2006); Goswami et al. (2009) used combinations of simple lexical and syntactic features to determine the gender and age of authors of anonymous blog posts. Peersman et al. (2011) retrieved a dataset from Netlog[8], with self-annotated age and gender of their authors and

[4] http://dumps.wikimedia.org/eswiki/20121227/eswiki-20121227-pages-meta-current.xml.bz2
[5] https://twitter.com/
[6] https://www.facebook.com/
[7] Joy, surprise, sadness, disgust, anger, fear
[8] http://www.netlog.com/

Goswami et al. (2009) demonstrated that the use of language in blogs correlates with age (for example, with the increase of the use of prepositions and determiners), but could not determine similar correlation with gender. Zhang and Zhang (2010) experimented with short segments of blog posts and Nguyen et al. (2013) studied the use of language and age in Twitter, the most well-known platform of short texts (140 characters long). All of them based their studies on gathering stylistic features like non-dictionary words as slang words, part-of-speech, function words, hyperlinks, the average length of the sentences, and sometimes combined with content features as single words with the highest information gain.

2.2 Author profiling tasks

The task of obtaining author profiles has an emerging interest in the scientific community, as can be seen in the number of related tasks around the topic. The task on *Author Profiling at PAN 2013*[9] encouraged researchers to identify age and gender of the authors of a large amount of anonymous texts (Rangel et al., 2013). Participants had to infer from blog posts what age and gender the authors are, in a real scenario with a large-size corpus and high amount of spam data, for example, automatically generated by robots.

Similarly, the shared task on Native Language Identification at *BEA-8 Workshop*[10] promotes researchers to identify native language of an author based on a sample of their writing. Finally, the task on *Personality Recognition at ICWSM 2013*[11] intends to be a common research framework where to investigate the Big Five traits[12].

In a similar vein, the interest in this type of research is evident in the Kaggle[13] platform, where companies and research departments can share their needs and independent researchers can join the challenge of solving them. We can find challenges as *Psychopathy Prediction Based on Twitter Usage*[14], *Personality Prediction Based on Twitter Stream*[15] or *Gender Prediction from HandWriting*[16]. This shows the rise of interest on this kind of problems.

2.3 Neurology: A theoretical framework

Neurology is the science which focuses its interest on treating disorders on neural system, such as aphasia. Aphasia is the loss of ability to produce or understand language due to lesions in brain areas related to these functions. At the end of XIX century and as result of studies about that disease, the German neurologist and

---

[9] http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-profiling.html
[10] https://sites.google.com/site/nlisharedtask2013/
[11] http://mypersonality.org/wiki/doku.php?id=wcpr13
[12] Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
[13] http://www.kaggle.com/
[14] http://www.kaggle.com/c/twitter-psychopathy-prediction
[15] http://www.kaggle.com/c/twitter-personality-prediction
[16] http://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting

psychiatrist Karl Wernicke and the French physician, anatomist and anthropologist Paul Pierre Broca defined two brain areas involved in the comprehension and production of the language, respectively Wernicke's Area and Broca's Area. (Falk, D., 2004)

Wernicke's area is in the cerebral cortex in the posterior half of the superior temporal gyrus and in the adjacent part of the middle temporal circunvolution, and its main role is the auditive decoding in the linguistic function, related with the language comprehension and with the control of the content of the message.

Broca's area is in the third inferior frontal gyrus, in the frontal lobe of the left hemisphere of the brain, for the vast majority of people, the hemisphere which rules the language. It controls many of the social skills of people, processes the grammar, is involved in the production of the speech, in the processing of the language and in its comprehension. It controls the ability to express and conciliate emotions, the skill for reading facial emotion in other people, the emotions and the skill for establishing social relationships. This area seems to be responsible for processing style words.

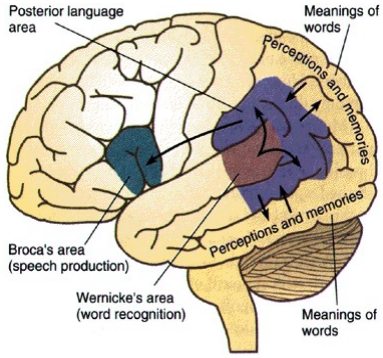

Fig. 1. Broca and Wernicke's areas of the brain

There are two basic questions to answer when we write a speech, WHAT to say and HOW to say it. The first question relates to the object that we want to communicate and defines the content of the speech. It is not a personal issue but a matter referred to what is being communicated. The second question responds to the way the author is going to communicate the content, for example, the style of the discourse itself. Therefore, it is a matter inherent to the communicator, to the way the communicator builds his discourse, and it is determined essentially by his profile. From the viewpoint of the receiver, the questions turn on WHAT is said and WHO is saying what, and both questions are related to two kind of words, those oriented to communicate contents, to answer the question WHAT, and those oriented to connect the discourse, to give its style and form, to answer the question HOW/WHO.

3 Automatic identification of gender and age based on stylistic features

We focused our interest on the cognitive approach based on the neurology studies of Broca and Wernicke and tried to represent the way the users express themselves, the way they use the language, that is, the style authors write. Based on the study of Pennebaker (2011) for English, where stylistic features identify some profile traits, we carried out some statistical research in Spanish analyzing a large number of documents[17] from Wikipedia, newsletters, forums, blogs, Twitter and Facebook and obtaining the frequency of use of the different grammatical categories.

Table 1. Distribution of grammatical categories per channel

POS	WIKI	NEWS	BLOGS	FORUMS	TW	FB
ADJ	13.57	12.50	13.67	9.27	6.62	12.06
ADV	2.78	3.46	3.87	**4.74**	**6.30**	3.49
CONJ	1.52	2.10	1.80	4.18	**7.00**	2.64
Q	3.34	4.47	4.15	5.34	5.53	4.29
DET	2.88	3.48	2.78	4.18	6.40	4.02
INTJ	0.35	0.04	0.06	0.42	0.38	0.07
MD	0.01	0.03	0.02	0.00	0.00	0.00
PREP	4.00	5.49	5.07	**8.94**	**13.81**	6.15
PRON	0.65	0.92	1.12	**2.22**	**3.32**	1.39
NOM	**50.33**	47.05	46.59	42.63	34.08	47.07
VERB	20.55	20.47	**20.88**	18.08	16.56	18.83

Table 2. Frequency of person and number in pronouns and verbs

POS	PER	NUM	WIKI	NEWS	BLOG	FOR	TW	FB
PRON	1	SIN	13.61	14.58	18.85	**54.47**	**65.81**	22.3
		PLU	0.00	0.00	0.00	0.00	0.00	0.00
	2	SIN	4.58	1.18	2.23	1.54	3.53	3.95
		PLU	1.92	1.75	5.31	4.61	5.62	3.49
	3	SIN	**55.06**	**50.75**	39.26	24.08	12.70	34.68
		PLU	13.42	18.22	16.93	8.91	3.35	17.14
	OTHER		11.41	13.52	17.42	6.39	8.99	18.44
VERB	1	SIN	19.95	17.41	17.50	28.94	24.00	16.61
		PLU	2.10	2.42	4.19	2.68	4.68	4.89
	2	SIN	6.02	1.55	3.58	3.55	6.77	2.95
		PLU	0.46	0.42	0.69	0.98	1.65	0.76
	3	SIN	31.40	34.00	29.92	28.80	31.21	31.21
		PLU	40.07	44.20	45.11	35.05	31.69	43.59

[17] Number of documents per channel: Wikipedia: 3,987,179 Newsletters: 5,191,694 Blogs: 1,083,709 Forums: 673,664 Twitter: 23,873,371 Facebook: 576,723

Table 1 shows the similitude between Wikipedia, newsletters and blogs in the use of adjectives, nouns and verbs to describe objects, people, places and situations. Forums is highlighted for the high use of prepositions, adverbs and pronouns due to the need of authors for directly describing their problems and searching for a solution. In Twitter, people use pronouns and verbs in first person (Table 2) with the highest frequency. This confirms such channel as ego-centered, where authors try to communicate personal thoughts, together with a low use of verbs and high use of adverbs and prepositions, following Twitter's main motto: "what are you thinking about?", "what are you doing?" or "what is happening?".

Following, we employed these findings on the use of grammatical categories in order to identify emotional profile in texts. Texts were classified into the six basic emotions (joy, surprise, anger, disgust, fear, sadness). We analyzed the distribution of the use of grammatical categories by gender in a dataset of 1,200 Facebook comments. Results are shown in Table 3.

Table 3. Distribution of grammatical categories by gender

POS	ALL	MALE	FEMALE
ADJ	6.49	6.53	6.45
ADV	3.93	3.94	3.91
CONJ	9.51	9.55	9.46
Q	5.46	5.76	5.12
DET	7.25	6.81	**7.74**
INTJ	0.23	0.18	**0.30**
MD	0.00	0.00	0.00
PREP	6.06	**6.25**	5.85
PRON	2.45	2.24	**2.67**
NOM	31.89	32.21	31.53
VERB	15.38	15.44	15.32

We can appreciate some important variations in the use of the grammatical categories by gender, for instance, as found for English (Pennebaker, 2011), we also verified for Spanish that men use more prepositions than women (+6.84%), perhaps because they try to hierarchically categorize things into their environment, and women use more pronouns (+19.20%), determinants (+13.66%) and interjections (+66.67%) than men perhaps because they are more interested in social relationships. Such conclusions appear to be parallel with content and style, with Wernicke and Broca's areas and we thought that using such stylistic features could help us to determine gender with some accuracy. We also have the intuition that such features could help us to identify age. Thus, we proposed the following features:

- Frequencies: Ratio between number of unique words and total number of words, words starting with capital letter, words completely in capital letters, length of the words, number of capital letters and number of words with flooded characters (e.g. Heeeelloooo);
- Punctuation marks: Frequency of use of dots, commas, colon, semicolon, exclamations, question marks and quotes;

- Part-of-speech: Frequency of use of each grammatical category, number and person of verbs and pronouns, mode of verb, proper nouns (NER) and non-dictionary words (words not found in dictionary);
- Emoticons[18]: Ratio between the number of emoticons and the total number of words, number of the different types of emoticons representing emotions: joy, sadness, disgust, angry, surprised, derision and dumb;
- Spanish Emotion Lexicon (SEL) (Sidorov et. al, 2012) : We obtained the lemma for each word and then its *Probability Factor of Affective Use* value from the SEL dictionary. If the lemma does not have an entry in the dictionary, we look for its synonyms. We add all the values for each emotion, building one feature per emotion.

We do not use any content/context dependent features in order to obtain more independence from the topics.

4 Methodology

4.1 Training and test datasets

The PAN-AP-13 dataset consists of a large number of anonymous authors labeled with gender and age. For the age group, on the basis of previous work (Koppel et al., 2003) the following classes are considered: 10s (13-17), 20s (23-27) and 30s (33-47). The data is balanced by gender but not by age. Each author can contain from one to tens of posts. The distribution of the number of authors per dataset is shown in Table 4.

Table 4. Distribution of number of authors by age

AGE	NUM. OF AUTHORS	
	TRAIN	TEST
10s	2,500	240
20s	42,600	3,840
30s	30,800	2,720

4.2 Machine learning approach and performance measures

We used the Support Vector Machine method implemented in Weka[19]. We experimented with different parameters and finally we used a Gaussian kernel with g=0.01 and c=2,000.

In order to be able to compare our results with the ones obtained by the teams participating in the PAN 2013 task on Author Profiling, we used Accuracy as

[18] http://es.wikipedia.org/wiki/Anexo:Lista_de_Emoticonos

[19] http://www.cs.waikato.ac.nz/ml/weka/

"closeness of agreement between a measured quantity value and a true quantity value of a measurand". Concretely, we perform the ratio between the number of authors correctly predicted by the total number of authors.

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + true\ negatives + false\ negatives}$$

5 Discussion of the experimental results

In Table 5 our proposal is ranked together with the final results of PAN-AP task[20], separately for age and gender.

Table 5. PAN ranking for Author Profiling by Gender and by Age (Spanish)

POS	TEAM	GENDER		POS	TEAM	AGE
1	Santosh	0.6473		1	Pastor	0.6558
2	Pastor	0.6299		2	Santosh	0.6430
3	Haro	0.6165		**3**	**(*Rangel*)**	**0.6350**
4	Ladra	0.6138		4	Haro	0.6219
5	Flekova	0.6103		5	Flekova	0.5966
6	Jankowska	0.5846		6	Ladra	0.5727
7	**(*Rangel*)**	**0.5713**		7	Yong	0.5705
8	Kern	0,5706		8	Ramirez	0.5651
9	Jimenez	0.5627		9	Aditya	0.5643
10	Ayala	0.5526		10	Jimenez	0.5429
11	Cagnina	0.5516		11	Gillam	0.5377
12	Yong	0.5468		12	Kern	0.5375
13	Mechti	0.5455		13	Moreau	0.5049
14	Weren	0.5362		14	Meina	0.4930
15	Meina	0.5287		15	Weren	0.4615
16	Ramirez	0.5116		16	Jankowska	0.4276
17	*Baseline*	*0.5000*		*17*	Cagnina	0.4148
18	Aditya	0.5000		18	Hidalgo	0.4000
19	Hidalgo	0.5000		19	Farias	0.3554
20	Farias	0.4982		*20*	*Baseline*	*0.3333*
21	Moreau	0.4967		21	Ayala	0.2915
22	Gillam	0.4784		22	Mechti	0.0512

We achieved the 7th position in gender prediction and 3rd position in age prediction. Based on such results, we conclude that the proposed stylistic features perform better for age than for gender identification. Perhaps this is due to the fact that the writing style depends more on the age of the author than on the gender, confirming what

[20] http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/pan13-ap-final-results.pdf

stated in (Goswami et al., 2009) about the correlation between age and use of language. But in any case, the task of identifying age seems to be easier than identifying gender, task which seems to be very difficult because the values obtained are not so high compared to the baseline (50%).

6 Conclusions and future work

We focused our interest on the cognitive approach based on the neurology studies of Broca and Wernicke and tried to represent the way the users express themselves, the way they use the language, that is, the style in which authors write. We carried out some experiments and some important variations in the use of the grammatical categories by gender were appreciated. For example, men use more prepositions than women because they try to hierarchically categorize things into their environment, and women use more pronouns, determinants and interjections than men because they are more interested in social relationships.

We conclude that stylistic features help to identify age and gender of anonymous authors although the task seems to be very difficult mainly for gender detection. We obtained competitive results in comparison with the ones obtained by the PAN-AP task participants. This encourages us to follow the research in this direction in order to understand better how people use language to express themselves and how this could help us to identify the profile of an author.

We must bear in mind with the differences between languages, for example between English and Spanish. For instance, in Spanish the use of pronouns is generally elliptical and it is a choice of the author to use them perhaps to emphasize something, as well as the use of prepositions or determinants in English is more regulated than in Spanish. Due to such specificities, we plan to investigate our proposal to different languages as English.

The features we use for modeling the discursive style are preliminary and simples. As future work we are interested in analyzing the discourse in order to investigate further how people use different words of the different grammatical categories, how they place them in the sentence, and how such stylistic decisions provide us information about the author profile. We also plan to research on the relationship between the demographics such as the gender and age with the emotional profile of the authors and their personality traits, trying to link such tasks in order to build a common framework to allow us to better understand how people use language from a cognitive linguistics viewpoint.

Acknowledgements

The work of the first author was partially funded by Autoritas Consulting SA and by Ministerio de Economía de España under grant ECOPORTUNITY IPT-2012-1220-430000. The work of the second author was carried out in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the FP 7 Marie Curie, the DIANA APPLICATIONS – Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

Argamon, S., Koppel, M., Fine, J., Shimoni, A. (2003). Gender, genre, and writing style in formal written texts. Text, 23(3): 321–346.

Burger, J. D.; Henderson, J.; Kim, G.; and Zarrella, G. (2011). Discriminating gender on Twitter. In EMNLP'11: 1301-1309.

Falk, D. (2004). Prelinguistic evolution in early hominins: Whence motherese? Behavioral and Brain Sciences 27, 491-541.

Goswami, S.; Sarkar, S.; and Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. In ICWSM'09.

Holmes, J., and Meyerhoff, M. (2003). The handbook of language and gender. Oxford: Blackwell.

Koppel, M., Argamon, S., Shimoni, A. R. (2003). Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing 17: 401-412.

Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T. (2013). "How Old Do You Think I Am?": A Study of Language and Age in Twitter. The 7th International AAAI Conference on Weblogs and Social Media. ICWSM'13.

Peersman, C., Daelemans, W., Van Vaerenbergh, L. (2011). Predicting Age and Gender in Online Social Networks. SMUC'11.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology. (54): 547–577.

Pennebaker, J.W. (2011) The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury Press.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G. (2013) Overview of the Author Profiling Task at PAN 2013. In: Forner P., Navigli R., Tufis D.(Eds.), Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013, Valencia, Spain, September 23-26.

Schler, J., Koppel, M., Argamon, S., Pennebaker, J. (2006) Effects of Age and Gender on Blogging. American Association for Artificial Intelligence.

Sidorov, G., Miranda Jiménez, S., Viveros Jiménez, F., Gelbukh, A., Castro Sánchez, N., Velásquez, F., Díaz Rangel, I., Suárez Guerra, S., Treviño, A., Gordon, J. (2012) Empirical Study of Opinion Mining in Spanish Tweets. LNAI 7629-7630.

Zhang, C., and Zhang, P. (2010). Predicting gender from blog posts. Technical Report. University of Massachusetts Amherst, USA.