Author Profiling from Facebook Corpora

Fernando Chiu Hsieh, Rafael Felipe Sandroni Dias, Ivandré Paraboni
University of São Paulo, School of Arts, Sciences and Humanities
São Paulo, Brazil
fernando.hsieh@usp.br , rafaelsandronidias@gmail.com, ivandre@usp.br

Abstract

Author profiling - the computational task of prediction author's demographics from text - has been a popular research topic in the NLP field, and also the focus of a number of prominent shared tasks. Author profiling is a problem of growing importance, with applications in forensics, security, sales and many others. In recent years, text available from social networks has become a primary source for computational models of author profiling, but existing studies are still largely focused on age and gender prediction, and are in many cases limited to the use of English text. Other languages, and other author profiling tasks, remain somewhat less popular. As a means to further this issue, in this work we present initial results of a number of author profiling tasks from a Facebook corpus in the Brazilian Portuguese language. As in previous studies, our own work will focus on both standard gender and age prediction tasks but, in addition to these, we will also address two less usual author profiling tasks, namely, predicting an author's degree of religiosity and IT background status. The tasks are modelled by making use of different knowledge sources, and results of alternative approaches are discussed.

Keywords: Author profiling, Document Classification, Corpus-based approaches

1. Introduction

Author profiling - the computational task of prediction author's demographics from text - has been a popular research topic in the NLP field, and also the focus of a number of prominent shared tasks (Pardo et al., 2017). Author profiling is a problem of growing importance, with applications in forensics, security, sales and many others (Rangel et al., 2015).

In recent years, text available from social networks has become a primary source of data for computational author profiling. However, existing studies of this kind are often focused on age and gender prediction, and are in many cases limited to the use of English text and a few others. Other languages, and other author profiling tasks, by contrast, remain somewhat less popular.

In this work we address the issue of computational author profiling from a Facebook corpus in the Brazilian Portuguese language. In doing so, we address standard age and gender prediction tasks, and also two less-known alternatives: predicting an author's degree of religiosity and his/her IT background status. Religiosity prediction may be seen as an extension of personality recognition (Mairesse et al., 2007), and degrees of religiosity are indeed known to correlate with certain personality traits in the Big Five model (de Andrade, 2008). IT background status prediction, on the other hand, may be seen as a means to group authors into (e.g., professional) communities.

In order to predict age, gender, religiosity and IT background status, we propose a number of author profile models based on standard document classification methods (from bag of words to word embeddings) and using different knowledge sources (from purely textual to psycholinguistic features). In doing so, our goal is to determine which methods are best suited for the four tasks at hand based on Facebook text in the Brazilian Portuguese language.

The rest of this paper is structured as follows. A number of existing author profiling models are reviewed in Section 2. Our own approach and corpus are described in Section 3. Evaluation and its results are described in Section 4. Final remarks are presented in Section 6.

2. Background

Studies on gender and age prediction are popular in NLP and related fields, and have in many cases been carried out in the light of the PAN-CLEF Shared task series (Pardo et al., 2017). Outside the scope of shared tasks, however, a direct comparison among existing methods may be complicated by the fact that different studies may rely on different problem definitions (e.g., regression versus classification), different datasets, languages (in most cases, English), and evaluation metrics (e.g., accuracy, F-measure etc.). These difficulties notwithstanding, in what follows we briefly discuss a number of recent studies of this kind.

The work in (Nguyen et al., 2014) presents a comparative study between an automated system and an experiment with human subjects to predict the gender and age of Twitter users, and discusses the limitations of current computational approaches to gender and age prediction from text. The study highlights the difference between social and biological identities, showing that more than 10% of the Twitter users do not employ language that would normally be associated with their biological sex, and that older Twitter users are often perceived to be younger. The study is based on Dutch text translated into English, and compares a linear regression model for age prediction and logistic regression for gender prediction with guesses provided by human subjects. A model based on majority vote reaches an accuracy of 0.84, a result that is arguably similar to existing automatic classification systems on English Twitter data.

The work in (Sap et al., 2014) presents predictive lexica for age and gender using regression and classification models from language usage in social networks text with associated demographic labels. The weighted lexica obtained average

prediction accuracy of 0.83 (age) and 0.82 (gender) for the English language.

The work in (Álvarez-Carmona et al., 2015) proposes gender and age recognition models that combine second order features (already employed in their own previous work at PAN 2013 and 2014) with latent semantic analysis (LSA). The resulting model is the overall winner of the PAN-2015 shared task, ranking among the top three systems in the three languages of the task (English, Dutch and Spanish).

The work in (op Vollenbroek et al., 2016) makes use of a SVM model trained on English, Dutch and Spanish Twitter data for gender and age prediction from unknown, non-Twitter text. Given the goal of building a cross-genre model for these tasks, the work avoids the use of language-specific features, focusing instead on n-gram counts, capitalisation, punctuation, word and sentence length, out-of-vocabulary words, vocabulary richness, function words, part-of-speech and emoticons. The model also includes second-order features representing relative values of some of these absolute measures. The resulting model in (op Vollenbroek et al., 2016) was the overall winner of the PAN-2016 cross-genre author profiling task (Rosso et al., 2016), obtaining 0.53 joint accuracy (i.e., age and gender prediction put together) when combining results from English, Spanish and Dutch texts. However, the authors pointed out that their present (cross-genre) results were considerably lower than those obtained in previous (single-genre) author profiling tasks at the PAN-CLEF series.

The work in (Basile et al., 2017) presents a model called N-GrAM for gender prediction from Twitter text in English, Spanish, Arabic and European Portuguese. The model makes use of a linear SVM model with word unigrams and character 3- to 5-grams as features. Interestingly, language- and domain-dependent features such as POS tags and Twitter handles were found to decrease overall accuracy. The proposed model was the best-performing participant in the PAN-CLEF-2017 shared task (Pardo et al., 2017), with 0.83 average accuracy when considering the four languages combined.

Finally, the work in (Guimarães et al., 2017) presents a method for binary age classification from Twitter, and it is one of the few studies to address this task using text in Brazilian Portuguese. The model makes use of Twitter-related features such as punctuation, text length, sharing status, and topic. The study compares a number of classification methods, including MLP, DCNN, Random Forest and SVMs. Among these, best results were obtained by using DCNN, with 0.94 average F-measure.

3. Current work

We designed an experiment to compare a number of document classification models applied to four prediction tasks from a Brazilian Portuguese Facebook corpus. These tasks are described individually in the following sections.

3.1. Author profiling tasks

In the present work we consider both standard gender and age-bracket prediction, and two less usual author profiling tasks, namely, predicting author's degree of religiosity and IT background status. The focus on religiosity and IT background is mainly motivated by our choice of corpus (cf. next section), in which both kinds of information are readily available. Both issues however constitute potentially relevant research questions on their own right.

Age bracket prediction was modelled as a three-class problem (18-20, 23-25, and 28-61 years-old). As in (Rangel et al., 2015), instances in the intermediate age brackets (i.e., 20-23 and 25-28) were disregarded in order to minimise the possible mismatch between a Facebook user age and the actual time of the publication.

Gender and *IT-background status* prediction were modelled as binary classification tasks (male / female and yes / no).

Religiosity is represented by self-reported scores ranging from 1 (not religious at all) to 5 (highly religious), and the corresponding prediction task was modelled as a three-class problem (1-2 degrees, 3 degrees, and 4-5 degrees) so as to obtain nearly-balanced groups[1].

3.2. Computational models

In what follows we will consider five author profiling models. First, given that author profiling may be seen as a document classification task, we will consider models based on both word counts (bag-of-words) and TF-IDF counts. Second, since char n-gram models have been popular in the related task of author prediction (Tschuggnall et al., 2017) we will also investigate their use in our present tasks. Third, given that other author profiling tasks - in particular, personality prediction as in (Mairesse et al., 2007) - may benefit from the use of psycholinguistic lexical information, we will make use of LIWC (Pennebaker et al., 2001) and related features as well. Finally, as in many recent NLP tasks, we will also consider the use of distributed word representations (Mikolov et al., 2013). These models are detailed below.

We will consider three standard text models that do not rely on external knowledge sources: the *TF-IDF* model, which consists of TF-IDF counts for the 3k most frequent terms; the *BoW*, which is a Bag-of-words model based on word counts, also keeping the 3k most frequent words; and the *Char* model, which is a standard 3-5 character n-gram model.

In addition to that, we will also consider a lexical model conveying 68 psycholinguistic features obtained from two external knowledge sources: the LIWC psycholinguistic dictionary (Pennebaker et al., 2001) and additional MRC-like (Coltheart, 1981) psycholinguistic properties. This model is presently labelled as *LIWC+P*.

From the Brazilian Portuguese of the LIWC dictionary (Filho et al., 2013), we computed 64 features representing psycholinguistic word categories (e.g., anger, family, wealth etc.). Each feature represents the number of words found in the corresponding category divided by the total number of words in the document. Moreover, from the psycholinguistic properties database in (dos Santos et al.,

[1] The choice for this scale is similar to (de Andrade, 2008). For a more thorough account of religiosity - including, e.g., the distinction between organisational, non-organisational and subjective religiosity - see (Koenig and Bussing, 2010).

2017a), we computed four features representing concreteness, imageability, subjective frequency and age of acquisition for Brazilian Portuguese text. Each of these features represents the average score of all document words in the corresponding category.

Finally, we will also consider a model based on word embeddings hereby called *Word2Vec*. This consists of the average word vectors obtained from a skip-gram-600 model (Mikolov et al., 2013) using window size=5, min_count=10, built from a 50-million Twitter corpus.

All the above models were built using logistic regression. Language-specific resources in *LIWC+P* and *Word2Vec* models (i.e., psycholinguistic dictionaries and Twitter data) concern the Brazilian Portuguese language only.

3.3. Data

In our experiment we make use of a portion of the *b5* corpus of texts and accompanying author demographics for the Brazilian Portuguese language (Ramos et al., 2018). The corpus has been applied to a number of NLP/NLG tasks ranging from personality recognition (dos Santos et al., 2017b; Silva and Paraboni, 2018) to personality-dependent content selection (Paraboni et al., 2017) and lexical choice (Lan and Paraboni, 2018).

All models were built from the 2.2 million word *b5-post* subcorpus of Facebook status updates. This dataset conveys up to 1,000 status updates from 1019 users of Brazilian Portuguese Facebook. The data is partially labelled with age, gender, degrees of religiosity and IT status information[2].

Gender, age and IT-background information were generally obtained from Facebook and, in some cases, provided by some of the participants of the *b5-post* data collection task upon request. Degrees of religiosity were provided by a small subset of participants upon request.

Table 1 summarises the number of documents (i.e., user's Facebook time lines) to be classified in each of the four prediction tasks.

Task	Documents
Age bracket	516
Gender	1018
Religiosity	440
IT background	814

Table 1: Data size for each classification task

Modelling age prediction as a classification task (as opposed to, e.g., a regression problem) requires deciding how to define age brackets. Our present choice is influenced by the concentration of individuals in their 20's (including a large proportion of undergraduate students) in our data. This distribution is illustrated in Figure 1.

4. Evaluation

After stop words removal using NLTK[3], the five models - *TF-IDF*, *Bow*, *Char*, *LIWC+P* and *Word2Vec* - were trained from the entire dataset for each classification problem. Results from 10-fold cross validation logistic regression are summarised as follows: age bracket in Table 2, gender in Table 3, religiosity in Table 4 and IT background status in Table 5.

Class	N	TF-IDF	BoW	Char	LIWC+P	Word2Vec
18-20	182	**0.60**	0.57	0.56	0.51	0.58
23-25	189	**0.56**	0.48	0.48	0.40	0.48
28-61	145	**0.61**	0.55	0.54	0.45	0.59

Table 2: Age bracket F-measure results

Class	N	TF-IDF	BoW	Char	LIWC+P	Word2Vec
female	578	**0.90**	0.86	0.84	0.80	0.88
male	440	**0.86**	0.81	0.79	0.72	0.85

Table 3: Gender F-measure results

Class	N	TF-IDF	BoW	Char	LIWC+P	Word2Vec
1-2	217	**0.67**	0.60	0.55	0.61	0.59
3	96	0.17	0.19	0.18	0.22	**0.26**
4-5	127	**0.52**	0.40	0.38	0.41	0.43

Table 4: Religiosity F-measure results

5. Discussion

As expected for a small corpus of this kind, overall results vary across dataset size and class definition. Best results are observed in the case of the two larger binary classes, namely gender classification and IT background status. Age bracket classification was moderately accurate, and degrees of religiosity classification had the lowest accuracy of all, with particularly low results for the sparse intermediate (degree 3) class.

Regarding the computational methods under consideration, we notice that the *TF-IDF* model generally outperforms the alternatives, with *Word2Vec* as a second best (particularly in the case of gender identification). By contrast, the lexical *LIWC+P* method generally produced the lowest results of all, which seems to suggest that psycholinguistic knowledge may be more suitable for personality and sentiment recognition than for the present author profiling tasks.

6. Final remarks

We have presented initial results of an author profiling task from Facebook text for the Brazilian Portuguese language. Our work has focused on both standard gender and age prediction, and less usual tasks of religiosity and IT background status prediction. We have compared a number of logistic regression models, and overall best results were observed when using TF-IDF counts. As future work, we intend to build higher-order n-gram models of Portuguese (Pereira and Paraboni, 2007; de Novais and Paraboni, 2012) and larger word embedding representations to further the use of deep learning methods applied to the current tasks.

[2]The latter is a result of having a significant proportion of Computer Science students among the target participants.

[3]http://www.nltk.org/

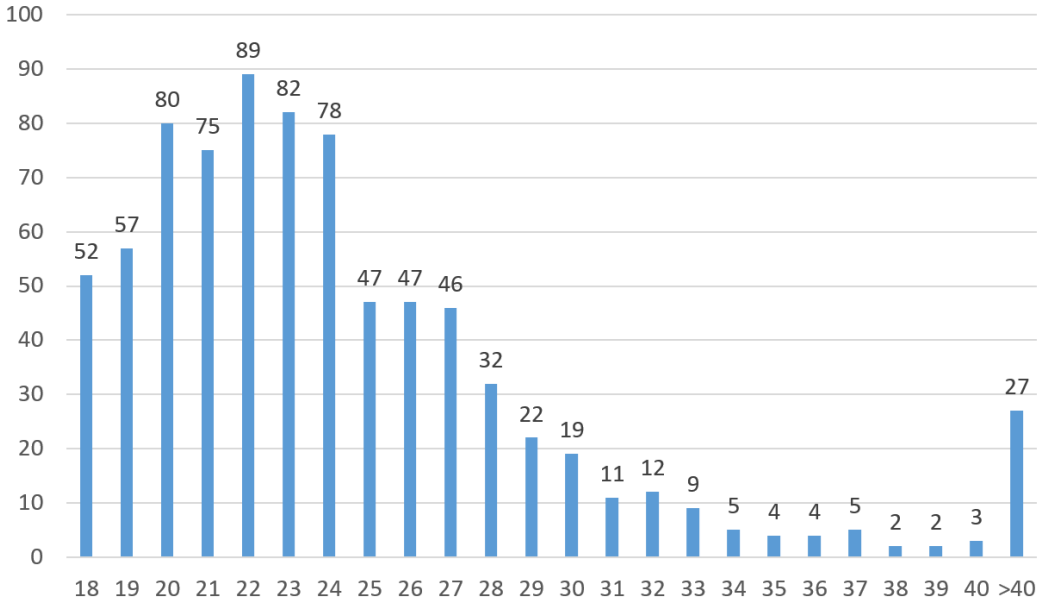

Figure 1: Age distribution in the b5-post corpus.

Class	N	TF-IDF	BoW	Char	LIWC+P	Word2Vec
no	491	**0.80**	0.76	0.73	0.75	0.73
yes	323	**0.64**	0.60	0.59	0.55	0.59

Table 5: IT background F-measure results

7. Acknowledgements

This work has been supported by the Brazilian Ministry of Education PET programme, and by grant # 2016/14223-0, São Paulo Research Foundation (FAPESP).

8. Bibliographical References

Álvarez-Carmona, M., López-Monroy, A., y Gómez, M. M., Villaseñor-Pineda, L., and Escalante, H. (2015). INAOE's participation at PAN'15: Author Profiling task. In *CLEF 2015*.

Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., and Nissim, M. (2017). N-GrAM: New groningen author-profiling model. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 33(4):497–505.

de Andrade, J. M. (2008). *Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil*. Ph.D. thesis, Universidade de Brasília.

de Novais, E. M. and Paraboni, I. (2012). Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19(2):135–146.

dos Santos, L. B., Duran, M. S., Hartmann, N. S., Junior, A. C., Paetzold, G. H., and Aluísio, S. M. (2017a). A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In *TSD-2017*.

dos Santos, V. G., Paraboni, I., and Silva, B. B. C. (2017b). Big five personality recognition from multiple text genres. In *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*, pages 29–37, Prague, Czech Republic. Springer-Verlag.

Filho, P. P. B., Aluísio, S. M., and Pardo, T. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *9th Brazilian Symposium in Information and Human Language Technology - STIL*, pages 215–219, Fortaleza, Brazil.

Guimarães, R. G., Rosa, R. L., de Gaetano, D., Rodríguez, D. Z., and Bressan, G. (2017). Age groups classification in social network using deep learning. *IEEE Access*, 5:10805–10816.

Koenig, H. G. and Bussing, A. (2010). The duke university religion index (DUREL): a five-item measure for use in epidemiological studies. *Religions*, 1(1):78–85.

Lan, A. G. J. and Paraboni, I. (2018). Definite description lexical choice: taking speaker's personality into account. In *11th International Conference on Language Resources and Evaluation (LREC-2018) (to appear)*, Miyasaki, Japan. ELRA.

Mairesse, F., Walker, M., Mehl, M., and Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Nguyen, D.-P., Trieschnigg, R. B., Dogruoz, A. S., Gravel, R., Theune, M., Meder, T., and de Jong, F. M. (2014). Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING-2014*, pages 1950–1961. Association

for Computational Linguistics.

op Vollenbroek, M. B., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., and Nissim, M. (2016). GronUP: Groningen user profiling: Notebook for PAN at CLEF 2016. In *CEUR Workshop Proceedings*, Netherlands.

Paraboni, I., Monteiro, D. S., and Lan, A. G. J. (2017). Personality-dependent referring expression generation. In *Text, Speech and Dialogue (TSD-2017) Lecture Notes in Artificial Intelligence vol. 10415*, pages 20–28, Prague, Czech Republic. Springer-Verlag.

Pardo, F. M. R., Rosso, P., Potthast, M., and Stein, B. (2017). Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.

Pereira, D. B. and Paraboni, I. (2007). A language modelling tool for statistical NLP. In *5th Workshop on Information and Human Language Technology (TIL-2007). Anais do XXVII Congresso da SBC*, pages 1679–1688, Rio de Janeiro. Sociedade Brasileira de Computação.

Ramos, R. M. S., Neto, G. B. S., Silva, B. B. C., Monteiro, D. S., Paraboni, I., and Dias, R. F. S. (2018). Building a corpus for personality-dependent natural language understanding and generation. In *11th International Conference on Language Resources and Evaluation (LREC-2018) (to appear)*, Miyasaki, Japan. ELRA.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., and Daelemans, W. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. In *CLEF 2015 Evaluation Labs and Workshop*.

Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., and Stein, B. (2016). Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. In *7th International Conference of the CLEF Initiative (CLEF 16)*, Berlin Heidelberg New York. Springer.

Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., and Schwartz, H. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar. Association for Computational Linguistics.

Silva, B. B. C. and Paraboni, I. (2018). Learning personality traits from Facebook text. *IEEE Latin America (to appear)*.

Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., and Potthast, M. (2017). Overview of the 5th author profiling task at PAN 2017: Style breach detection and author clustering. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin.