Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1227–1234
Marseille, 11–16 May 2020

Cross-domain Author Gender Classification in Brazilian Portuguese

Rafael Felipe Sandroni Dias, Ivandré Paraboni
School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)
{rafaelsandroni,ivandre}@usp.br

Abstract

Author profiling models predict demographic characteristics of a target author based on the text that they have written. Systems of this kind will often follow a single-domain approach, in which the model is trained from a corpus of labelled texts in a given domain, and it is subsequently validated against a test corpus built from precisely the same domain. Although single-domain settings are arguably ideal, this strategy gives rise to the question of how to proceed when no suitable training corpus (i.e., a corpus that matches the test domain) is available. To shed light on this issue, this paper discusses a cross-domain gender classification task based on four domains (Facebook, crowd sourced opinions, Blogs and E-gov requests) in the Brazilian Portuguese language. A number of simple gender classification models using word- and psycholinguistics-based features alike are introduced, and their results are compared in two kinds of cross-domain settings: first, by making use of a single text source as training data for each task, and subsequently by combining multiple sources. Results confirm previous findings related to the effects of corpus size and domain similarity in English, and pave the way for further studies in the field.

Keywords: author profiling, gender classification, cross-domain profiling

1. Introduction

Author profiling is the computational task of predicting demographic characteristics of a target author based on the text that they have written. For instance, by analysing text from social networks or customer's product reviews, we may infer an author's gender, age, personality traits or other kinds of information. Author profiling tasks have been a popular NLP research topic, and have regularly featured in the PAN-CLEF shared task series (Rangel et al., 2016; Rangel et al., 2017; Rangel et al., 2018). Practical applications include marketing research, on-line fraud detection, copyright and plagiarism investigations, among others.

Author profiling models will often follow a *single-domain* approach, that is, a model is trained from a corpus of labelled texts in a given domain, and it is subsequently validated against a test corpus built from precisely the same domain. Although single-domain settings are arguably ideal, this strategy gives rise to the question of how to proceed when no suitable training corpus (i.e., a corpus that matches the test domain) is available, and in which case we may have to resort to so-called *cross-domain* author profiling.

Cross-domain author profiling has become a popular research topic in recent years (Rangel et al., 2016; Medvedeva et al., 2017), and our own work focuses on the issue of cross-domain gender classification in the Brazilian Portuguese language. The focus on gender is motivated by the observation that gender-labelled corpora are widely available, and that results for gender classification are usually higher than those obtained in other author profiling tasks (dos Santos et al., 2020).

In the present work we train gender classifiers on text in four domains (Facebook, crowd sourced opinions, Blogs and E-gov requests) by making use of word- and psycholinguistics-based features alike. Next, we compare gender classification results in two kinds of cross-domain setting: first, by making use of a single text source as training data for each task, and subsequently by combining multiple text sources. Results confirm previous findings related to the effects of corpus size and domain similarity in English (Medvedeva et al., 2017), and pave the way for further studies in the field.

The rest of this paper is organised as follows. Section 2 reviews existing work in author profiling and single- and cross-domain gender classification. Section 3 describes the four corpora taken as the basis to our experiments. Section 4 introduces our single-domain classifiers and reference results for the task. Section 5 addresses the issue of cross-domain gender classification from a single text source at a time, and Section 6 considers the use of combined (or multi-domain) text sources. Finally, Section 7 presents our final remarks and suggestions of future work.

2. Background

Author profiling - particularly in the case of gender and age recognition - has been the focus of an increasingly large number of studies in recent years, many of which developed around the PAN-CLEF competitions (Rangel et al., 2016; Rangel et al., 2017). This section reviews existing work on author gender classification, addressing the issues of single- and cross-domain author profiling separately.

2.1. Single-domain Author Profiling

Practical single-domain gender classification from text poses a number of well-known difficulties (Nguyen et al., 2014) stemming from text genre, quality and size, among others. Although results may remain modest in some settings, computational models of this kind attempt to circumvent existing difficulties by investigating a plethora of machine learning methods and text representations. Some recent studies of this kind are discussed below.

Given the wide range of task definitions, text genres, datasets and target languages under consideration, a direct comparison between existing approaches to gender classification is not straightforward. A major exception is the works in Basile et al. (2017), in Martinc et al. (2017) and in Sierra et al. (2017), all of which developed in the light of the PAN-CLEF 2017 gender and language variety identification task in Twitter (Rangel et al., 2017). The work in Basile et al. (2017), the overall winner of the competition, is also arguably the simplest model among the three, making use of a SVM classifier based on character n-grams. This model outperformed the work in Martinc et al. (2017), which presented a similar approach with added part-of-speech (POS) information. The much more sophisticate approach in Sierra et al. (2017), by contrast, made use of Convolutional Neural Networks (CNNs), but it was outperformed by 9 out of 22 systems.

The work in Fatima et al. (2017) is another example of simple and effective strategy for gender (and age) classification from text. As in the case of Basile et al. (2017), the study also makes use of word and character n-grams with SVMs. POS information plays a central role also in Reddy et al. (2017). A model based on TF-IDF-weighted POS n-grams outperforms a number of simple alternatives (e.g., bag of words and others) in the gender classification task in a Trip Advisor hotel recommendations domain.

The work in Isbister et al. (2017) is one of the few attempts to use psycholinguistic features obtained from the LIWC dictionary (Tausczik and Pennebaker, 2010). The study evaluates the role of different word categories on gender prediction, and differences in LIWC data availability across five languages.

The work in Gopinathan and Berg (2017) makes use of two kinds of deep neural network for gender classification in Twitter text: a character-level convolutional bidirectional long short-term memory (LSTM), and a word-level bidirectional LSTM using Global Vectors (GloVe). A stacked architecture combining the character and word models is shown to outperform each of the individual models alone, and also a number of standard (e.g., bag of words and n-grams) baseline systems.

The work in Kim et al. (2017) addresses gender classification and other tasks in Twitter text by modelling each task as a vertex classification problem on graphs based on two types of recursive neural units (RNUs): Naive Recursive Neural Unit (NRNU) and Long Short-Term Memory Unit (LSTMU). These models were found to outperform a number of baseline systems that use lexica, logistic regression, label propagation and others.

Finally, the PAN-CLEF 2018 competition (Rangel et al., 2018) introduced a gender classification task based on a combination of text and image data. Among the participant systems, the work in Takahashi et al. (2018) presented a neural network model called Text Image Fusion Neural Network (TIFNN) to leverage both data sources, and it was the overall winner of the competition.

2.2. Cross-domain Author Profiling

Since 2013, the PAN-CLEF initiative series has addressed the issues of age and gender classification from text and, in Rangel et al. (2016), these tasks were addressed in a cross-domain setting. In this case, models were trained on Twitter data, and subsequently tested on blogs, social media and hotel reviews text written in English, Spanish, and Dutch.

Of particular interest to the present study, the work in Medvedeva et al. (2017) points out that author profiling models are typically domain-specific and based on supervised methods, and therefore show limited portability to other domains. Based on this observation, the study presented a number of experiments assessing whether results obtained by the best-performing cross-domain model at PAN-CLEF 2016 truly carry over domains beyond Twitter in English and Spanish.

Among other findings, the analysis in Medvedeva et al. (2017) suggests that cross-domain author profiling is successful to a certain extent, and that results can be generally explained according to three aspects: size of training data (i.e., using more data improves results, and this may be beneficial even in a cross-domain setting), differences between genres (e.g., tweets are closer to blog publications than to hotel reviews, and that impacts the model outcome), and quality of data (e.g., Twitter texts are more noisy and arguably more difficult to classify.) The authors suggest that the main influencing factor in cross-domain profiling is the difference between training and test genres, and that, when domains are sufficiently close, an increase in the amount of training data can improve results. These issues will also be the focus of some of our own experiments described in the next sections.

3. Corpora

Before presenting our experiments in single- and cross-domain gender classification in the next sections, in what follows we describe the text corpora taken as the basis to these experiments.

3.1. Overview

We will address the gender classification task in four text genres (or domains) - Facebook status updates, crowd-sourced Opinions, Blogs and E-gov requests - as discussed below. These domains were selected based on their differences in style, vocabulary and size, all of which likely to impact the accuracy of the underlying tasks. Table 1 presents the class distribution (Male / Female) and additional descriptive statistics for each domain.

In Table 1, columns ‘Male’ and ‘Female’ present the actual number of learning instances available from each corpus. This results in tasks of different complexity, ranging from small (Opinion) to large (E-gov).

The ‘Documents’ column presents the total number of text documents (or authors) in each domain. Since not all documents are gender-labelled, these totals do not always correspond to the sum of male/female instances. The entire sets of documents were nevertheless considered when creating word embedding models for each domain as discussed in the next section.

The ‘Vocabulary’ column presents the number of unique words in each corpus. Once again, this suggests tasks of different complexity, ranging from very limited (Opinion, which convey opinions about only eight topics, as discussed below), to broad vocabularies (Blog).

Finally, the ‘Words’ and Words / docs’ columns present word counts and their average document sizes. Corpus word counts range from small (Opinion) to large (E-gov),

	Learning instances		Text statistics			
Domain	Male	Female	Documents	Vocabulary	Words	Words / docs
Facebook	441	578	1,019	63,165	2,434,215	2,389
Opinion	285	148	433	11,004	187,118	432
Blog	1,038	1,564	5,801	207,947	9,119,406	5,801
E-gov	28,805	15,893	49,449	77,396	3,760,126	76

Table 1: Corpora descriptive statistics

and average document sizes range from small (E-gov) to large (Blog).

Further details regarding each individual domain are discussed in the next sections. In what follows we briefly describe the kinds of text available from each domain.

Facebook texts are provided by the *b5-post* corpus (Ramos et al., 2018; dos Santos et al., 2017), a collection of over 194k status updates written by 1019 users of Brazilian Facebook that has been previously taken as the basis in a number of single-domain author profiling and personality classification tasks (Hsieh et al., 2018; Silva and Paraboni, 2018a; Silva and Paraboni, 2018b). Facebook status updates naturally cover a wide range of topics, including significant proportions of information about the authors themselves (e.g., what they are doing, what they are eating etc.) and, as in the case of social network languages in general, are often informal and noisy.

Opinion texts were obtained from an ongoing data collection task (dos Santos and Paraboni, 2019), conveying over 3400 short texts written by 433 on-line Brazilian microvolunteers. Opinion texts are mostly impersonal and highly focused on the topic under discussion, and are more formal than Facebook text. Texts consist of short moral stances produced in response to questions about eight contemporary topics including drug legalisation, abortion policies, death penalty, and others.

Blog texts are taken from the BlogSetBR corpus (dos Santos et al., 2018) of Brazilian personal blogs (2.4 million words) written by over 4000 authors. Blog texts cover a wide range of topics, from highly personal issues to, e.g., international politics, and may include third-party material or even experts in foreign languages.

Finally, E-gov texts were obtained from a collection of online requests made to the e-sic citizen information service provided by the Brazilian government[1]. E-gov requests are highly impersonal, addressing issues related to companies, taxes, authority and public policies, among many others. E-gov requests range from highly formal (e.g., official letters written by a council or other government department) to informal (e.g., short requests made by individuals regarding their rights, social benefits etc.)

4. Single-domain Gender Classification

Our fist experiment concerns standard single-domain gender classification. In doing so, our goal is to assess the degree of difficulty posed by each of our target domains, and to compare the use of psycholinguistics- and word-based features in each task. Reference results from this initial experiment will be taken into account when addressing the issue of cross-domain gender classification in the subsequent sections.

4.1. Data

The experiment makes use of the four corpora described in the previous section, namely, Facebook, Opinion, Blog and E-gov, and addresses the four gender profiling tasks supported by these datasets. To this end, each corpus was randomly split into training (80%) and test (20%) subsets in a stratified fashion, and the test portion of each dataset was reserved for the purpose of evaluation in Section 4.3.

4.2. Models

Existing author profiling models make use of a wide range of learning methods, from SVMs to (more recently) deep neural networks. Interestingly, however, there is evidence to suggest that some of the simplest approaches may be actually difficult to surpass (Basile et al., 2018). Motivated by this observation, and also by the small size of some of our current datasets, the present experiment will focus on the use of logistic regression and multi-layer perceptron (MLP) methods only.

The experiment will compare the results obtained by four gender classification models: a model based on psycholinguistic knowledge as provided by the LIWC dictionary (Pennebaker et al., 2001), and two word-based models: one using TF-IDF counts, and one using weighted skipgram word embeddings. A majority class baseline is also added for illustration purposes. These models are summarised as follows.

- *LR-LIWC*: psycholinguistics-motivated model using multinomial logistic regression.
- *LR-Tfidf*: k-best TF-IDF counts with ANOVA f-value univariate feature selection, using multinomial logistic regression.
- *MLP-skipgram*: TF-IDF average skipgram word embedding model, using multi-layer perceptron classifiers.
- *Baseline* : a simple majority class baseline system.

Both *LR-LIWC* and *LR-Tfidf* make use of multinomial logistic regression with liblinear solver, L2 penalty and balanced class weights. *LR-LIWC* takes as an input the 64 psycholinguistic features provided by the Brazilian Portuguese

[1] https://esic.cgu.gov.br/

LIWC dictionary (Balage Filho et al., 2013). *LR-Tfidf* consists of a standard TF-IDF unigram feature vector, subsequently reduced with k-best univariate feature selection using ANOVA f-value as a score function. Optimal k values were obtained by performing grid search over the training dataset in the 1000..30000 range at 500 intervals. These are summarised in Table 2. We notice that the two larger corpora - Blog and E-gov - require much larger feature sets than Facebook and Opinion.

Domain	k value
Facebook	3,500
Opinion	1,000
Blog	27,500
E-gov	9,000

Table 2: Optimal k values for the *LR-Tfidf* models.

MLP-skipgram makes use of multilayer perceptron classifiers using the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (lbfgs) solver, and additional parameter tuning as follows. Documents are represented as the weighted average of its skipgram word embeddings (Mikolov et al., 2013) multiplied by the TF-IDF scores of their words. Both self- and pre-trained word embeddings configurations were considered[2], and we considered using both full input texts (as in the original corpus) and filtered versions in which only words corresponding to the k best terms (cf. Table 2) were retained.

Finally, we followed Raunak (2017) and others and considered reducing the embedding dimensionality itself, using once again univariate feature selection with an ANOVA f-value function. The choice for this particular method is however primarily motivated by computational efficiency. For a possibly more sophisticated approach, we report to Raunak (2017), in which embeddings dimensionality reduction is achieved by making use of principal component analysis (PCA).

The alternative strategies for computing word embeddings and related network parameters are summarised in Table 3.

Parameter	Values
w: word embedding size	{50, 100, 300, 600, 1000}
s: word embedding source	{self, pre}
x: k-best feature set size	30..w, at 10% intervals
filter: k-best word filtering	{yes, no}
it: iterations	100..500, at 50 intervals
l: hidden layers	{1, 2, 3}
n: neurons per layer	from 5 to x, at 5% intervals
f: activation function	{ReLu, Tanh, Logistic}
alpha: MLP alpha value	1e-03..08

Table 3: Parameters under consideration for the *MLP-skipgram* models.

Optimal parameter values were obtained by performing grid search on training data. These are summarised in Table 4. Due to the computational costs involved in performing grid search over the two larger corpora (Blog and E-gov), however, only larger (300 and 600) embedding models and ReLu activation function were considered. Moreover, in the case of E-gov, the alpha parameter was kept constant (1e-05.) The word embedding model of size 1000 was only available in pre-trained format.

4.3. Results

Table 5 shows weighted F1 scores obtained by the four models - the majority class baseline, *LR-LIWC*, *LR-Tfidf*, and *MLP-skipgram* - applied to the test data in each domain.

As expected, the majority class baseline never outperforms the alternatives. Perhaps more surprisingly, however, the strategy based on psycholinguistic features *LR-LIWC* does fare much better than the baseline either. On the other hand, the use of TF-IDF counts in *LogRef-Tfidf* generally represents a substantial gain over the previous two models, and the combination of word embeddings and neural models in *MLP-skipgram* increases results even further, although not always outperforming the simpler *LR-Tfidf* approach.

5. Cross-domain Gender Classification from Individual Data Sources

Single-domain author profiling will arguably produce optimal results for certain tasks such as gender classification. However, when training data of the required type is not available, it may be necessary to resort to an alternative text source as a substitute. This strategy - known as cross-domain author profiling - gives rise to the question of how cross-domain compares to single-domain profiling or, to be more precise, how much loss (e.g., in F1 scores) should be expected.

Assuming single-domain gender classification results (cf. previous section) to be a gold standard, our second experiment aims to identify which domains (i.e., domains other than the test domain itself), if taken as training data, would produce results that are closest to this gold standard. In other words, we would like to identify which training domain would obtain the smallest loss in F1 scores.

5.1. Data

The present experiment makes use of the same four corpora in the previous experiment, namely, Facebook, Opinion, Blog and E-gov.

5.2. Models

For each of the four domains under consideration, a gender classification model was built using the previous *LR-Tfidf* approach for simplicity (cf. Section 4.2.) The choice for this particular model is motivated by the observation that results (cf. the previous Table 5) are sufficiently close to those obtained by the best-performing multi-layer perceptron models, but at a much lower computational cost.

In this setting, cross-domain predictions made by each model (e.g., the use of Facebook model to predict gender in the Opinion, Blog and E-gov domains etc.) are to be compared with the previous single-domain gold standard results obtained by performing 10-fold cross-validation on each individual dataset.

[2] Pre-trained embeddings taken from Hartmann et al. (2017).

Domain	w	s	x	filter	it	l	n	f	alpha
Facebook	100	self	100	yes	400	3	75	tanh	1e-04
Opinion	100	self	80	yes	250	3	25	tanh	1e-07
Blog	600	pre	600	no	200	1	300	relu	1e-05
E-gov	1000	pre	1000	yes	200	1	500	relu	1e-05

Table 4: Optimal parameter values for the *MLP-skipgram* models.

Domain	Baseline	LR-LIWC	LR-Tfidf	MLP-skipgram
Facebook	0.41	0.50	**0.80**	0.73
Opinion	0.52	0.63	0.70	**0.74**
Blog	0.45	0.66	0.75	**0.78**
E-gov	0.51	0.60	**0.79**	**0.79**

Table 5: Weighted F1 scores. Best results for each task are highlighted.

	Test			
Training	Blogs	E-gov	Facebook	Opinion
Blog	-	**0.22**	0.08	**0.29**
E-gov	**0.09**	-	**0.01**	0.35
Facebook	0.14	0.24	-	0.38
Opinion	0.31	0.25	0.52	-

Table 6: Cross-domain F1 loss relative to single-domain results. Rows represent a training domain and columns represent a test domain. Lower (and therefore better) results in each test domain are highlighted.

5.3. Results

We compared single- and cross-domain author profiling results by measuring F1 weighted loss, hereby understood as the weighted F1 score obtained in single-domain task minus the weighted F1 score obtained in the corresponding cross-domain task. Loss scores that are closer to zero are therefore better, indicating that the impact of using a different training domain is small.

Results for all possible domain combinations are shown in Table 6, with rows representing each source training domain, and columns representing target test domains.

From these results we notice that the perceived loss in using cross-domain approach is generally substantial. The only major exception is the case of gender classification in the Facebook domain using the model trained on E-gov data, which obtained minimal (0.01) loss and, to a lesser extent, the same task using the model trained on Blog data.

The higher losses observed in the other cases may be explained by the size of the training data. In particular, we notice that loss is somewhat smaller when using the larger Blog and E-gov models (on the two top rows of the table) as training data, although is not always the case. In the case of the (simpler) Opinion test domain, for instance, there seems to be little difference between using the large E-gov corpus as training data and the much smaller Facebook corpus. These issues will be further addressed in a complementary experiment described in the next section.

Training	Male	Female
All except Facebook	30,654	17,079
All except Opinion	30,810	17,509
All except Blog	29,531	16,619
All except E-gov	2,290	1,764

Table 7: Male/Female class distribution for cross-domain gender classification using multi-domain data sources.

6. Cross-domain Gender Classification from Multiple Sources

Results from the previous experiment support the well-established notion that more data usually helps classification tasks in general. Based on this observation, we envisaged a third experiment in which gender prediction in a given test domain is attempted by using training data provided by *all other available sources combined*, that is, by using as training data all available text except for the test domain itself. In doing so, we would like to investigate whether cross-domain F1 loss may be reduced by simply making use of more training data regardless of which domain the data come from. Thus, for instance, we will predict Facebook author's gender by using a model built from Opinion, Blog and E-gov texts combined, and so forth.

Other than concatenating training data from multiple text corpora, the present setting is similar to the previous experiment. We will once again test all four text genres available, and we will measure weighted F1 loss by comparing their results to those obtained from the single-domain gold standard as discussed in Section 5.2.

6.1. Data

For each of the four test domains - Facebook, Opinion, Blog and E-gov - we created multi-domain training data sets by combining all the three remaining sources (i.e., by concatenating all text sources except for the test domain itself.) The resulting class distribution is summarised in Table 7.

By concatenating data sources in this way, we notice that the first three datasets now have approximately the same size. The exception is the case in which the E-gov dataset is removed, since this corpus is the largest of all in number of instances. This issue will be further discussed in the next sections.

Test domain	Multi-domain	Best single-domain
Blog	**0.05**	0.09
E-gov	**0.16**	0.22
Facebook	**0.01**	**0.01**
Opinion	**0.26**	0.29

Table 8: Cross-domain F1 loss obtained from multiple- and single-domain models (from the previous experiment 2.) Lower (and therefore better) results for each test domain are highlighted.

6.2. Models

From each of the four multi-domain datasets described in the previous section, a gender classification model was built, once again by using the *LR-Tfidf* strategy (cf. Section 4.2.)

6.3. Results

Table 8 summarises weighted F1 loss results obtained by each of the four multi-domain models (left) accompanied by the results obtained by the best single-domain models addressed in the previous section, which are presently reproduced for ease of comparison.

A potentially interesting outcome of this experiment is the case of gender classification in the Facebook domain, in which results of both single- and multi-domain data sources remain essentially the same as in the original single-domain task, that is, with near zero loss. Facebook gender classification seems to be fairly easily accomplished based on a variety of text sources, an effect that may be at least partially explained by the observation that authors in this domain tend to write more about themselves than in the other domains under consideration.

Another result worth mentioning is the case of gender classification in blogs. Although the previous model (built from a cross-domain source) still has a considerably high (0.09) F1 loss, the use of multiple sources combined reduced the loss to 0.05, suggesting that some of the present difficulties may be circumvented by simply using more training data, from perhaps *any* available source.

Finally, regarding the more problematic E-gov and Opinion domains, we notice that the use of more training data in the multi-domain setting does reduce F1 loss, but the current levels are likely to be still unacceptable for practical applications. In the case of the E-gov domain, it is possible that by simply using a (much) larger training dataset, F1 loss may get closer to single figures, as this was by far the largest corpus of all. In the case of the Opinion domain, however, it is not immediately clear why the model performs so poorly, and more research seems to be required.

7. Final Remarks

In this paper we have presented three experiments addressing the issue of author gender classification in various text genres in the Brazilian Portuguese language, and discussed the issue of cross-domain gender classification from single and combined data sources.

The first experiment examined a number of single-domain gender classification models built from different corpora. Generally speaking, we notice that pure text-based representations (as provided by TF-IDF counts or TF-IDF weighted word embeddings) outperform the use of psycholinguistic features. This is in principle a positive outcome, particularly for applications focused on languages for which a suitable LIWC dictionary may not be available.

The second experiment focused on situations in which training data from the intended test domain is not available, and in which case we may resort to cross-domain author profiling. Although substantial losses were observed, we notice that this is not always the case and, in in some scenarios, cross-domain loss may become acceptably small if more training data is provided.

The observation that more data may alleviate the losses in cross-domain gender classification led to a third experiment in which heterogeneous training dataset were built by combining multiple text sources. Results once again show significant losses in comparison with the single-domain setting, but to a lesser extent than in the previous experiment. This suggests that cross-domain gender classification may be in principle feasible in some cases, provided that a sufficiently large amount of data is available, and it is consistent with the findings in Medvedeva et al. (2017) regarding cross-domain gender classification in English.

Regardless of training data size, however, we notice that cross-domain strategies may be more suitable to some domains than others. Once again, this is consistent with previous studies that have addressed the quality and the degree of difference between training and test domains as in, e.g., Medvedeva et al. (2017), but more research is still required to determine which of these (or other) factors may affect cross-domain learning in the present setting, and to which extent the present cross-domain strategies may be generalised to other (perhaps less directly comparable) author profiling tasks.

An important limitation of the current work is that all experiments were focused on gender classification only. As future work, we intend to expand the current experiments by addressing other author profiling tasks such as age and personality classification.

Also as future work, we intend to take a closer look into the effects of dataset size on task performance, and build cross-domain models by making use of larger amounts of training data. Another possible investigation along these lines is the use of profiling strategies based on lexical knowledge (Sap et al., 2014; Schwartz et al., 2013). Methods of this kind, which are clearly attractive for reasons of computational efficiency and potential for generalisation, still require further investigation in cross-domain settings.

Finally, we also intend to enrich an existing authorship attribution system (Custódio and Paraboni, 2018) with the output of the current gender classifiers and similar models. In doing so, we expect to improve overall accuracy in author identification with the aid of automated author profiling methods.

Acknowledgements

This work received support by FAPESP grant # 2016/14223-0 and from the University of São Paulo grant # 668/2018.

8. Bibliographical References

Balage Filho, P. P., 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.

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.

Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., and Nissim, M. (2018). Simply the best: Minimalist system trumps complex models in author profiling. In *LNCS vol. 11018*, pages 143–156, Cham. Springer.

Custódio, J. E. and Paraboni, I. (2018). EACH-USP ensemble cross-domain authorship attribution. In *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2018) vol.2125*, Avignon, France.

dos Santos, W. R. and Paraboni, I. (2019). Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. In *Recents Advances in Natural Language Processing (RANLP-2019)*, pages 1069–1075, Varna, Bulgaria.

dos Santos, V. G., Paraboni, I., and Silva, B. B. C. (2017). 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.

dos Santos, H. D. P., Woloszyn, V., and Vieira, R. (2018). BlogSet-BR: A Brazilian Portuguese Blog Corpus. In *11th International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. ELRA.

dos Santos, W. R., Ramos, R. M. S., and Paraboni, I. (2020). Computational personality recognition from facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia*, 25(4):268–287.

Fatima, M., Hasan, K., Anwar, S., and Nawab, R. M. A. (2017). Multilingual author profiling on facebook. *Information Processing & Management*, 53(4):886–904.

Gopinathan, M. and Berg, P.-C. (2017). A deep learning ensemble approach to gender identification of tweet authors.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *11th Brazilian Symposium in Information and Human Language Technology - STIL*, pages 122–131, Uberlândia, Brazil.

Hsieh, F. C., Dias, R. F. S., and Paraboni, I. (2018). Author profiling from facebook corpora. In *11th International Conference on Language Resources and Evaluation (LREC-2018)*, pages 2566–2570, Miyazaki, Japan. ELRA.

Isbister, T., Kaati, L., and Cohen, K. (2017). Gender classification with data independent features in multiple languages. In *European Intelligence and Security Informatics Conference (EISIC-2017)*, pages 54–60, Athens, Greece. IEEE Computer Society.

Kim, S. M., Xu, Q., Qu, L., Wan, S., and Paris, C. (2017). Demographic inference on Twitter using recursive neural networks. In *Proceedings of ACL-2017*, pages 471–477, Vancouver, Canada.

Martinc, M., Skrjanec, I., Zupan, K., and Pollak, S. (2017). PAN 2017: Author profiling - gender and language variety prediction. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin.

Medvedeva, M., Haagsma, H., and Nissim, M. (2017). An analysis of cross-genre and in-genre performance for author profiling in social media. In *LNCS vol. 10456*, Cham. Springer.

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.

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

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)*, pages 1138–1145, Miyazaki, Japan. ELRA.

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., and Stein, B. (2016). Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In *CLEF 2016 Evaluation Labs and Workshop, Notebook papers*, pages 750–784, Évora, Portugal. CEUR-WS.org.

Rangel, F. M., 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.

Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., and Stein, B. (2018). Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In Linda Cappellato, et al., editors, *Working Notes Papers of the CLEF 2018 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org.

Raunak, V. (2017). Simple and effective dimensionality reduction for word embeddings. In *NIPS-2017 Limited Labeled Data workshop*.

Reddy, T. R., Vardhan, B. V., and Reddy, P. V. (2017). N-Gram approach for gender prediction. In *Advance Computing Conference (IACC)*, pages 860–865.

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.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M., and Ungar, L. H. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One*, 8(9):e73791.

Sierra, S., y Gómez, M. M., Solorio, T., and González, F. A. (2017). Convolutional neural networks for author profiling. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum*, Dublin.

Silva, B. B. C. and Paraboni, I. (2018a). Learning personality traits from Facebook text. *IEEE Latin America Transactions*, 16(4):1256–1262.

Silva, B. B. C. and Paraboni, I. (2018b). Personality recognition from Facebook text. In *13th International Conference on the Computational Processing of Portuguese (PROPOR-2018) LNCS vol. 11122*, pages 107–114, Canela. Springer-Verlag.

Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., and Ohkuma, T. (2018). Text and image synergy with feature cross technique for gender identification. In *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2018) vol.2125*, Avignon, France.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

9. Language Resources References

- b5 corpus (Ramos et al., 2018).
- BlogSetBR corpus (dos Santos et al., 2018)
- BRmoral corpus 4.33 (dos Santos and Paraboni, 2019)
- e-sic database (`https://esic.cgu.gov.br/`)
- NILC embeddings (Hartmann et al., 2017)