Journal of Intelligent Information Systems

Author profiling from Text: new tasks and towards cross-domain modelling

--Manuscript Draft--

Manuscript Number:	
Full Title:	Author profiling from Text: new tasks and towards cross-domain modelling
Article Type:	Manuscript
Keywords:	Natural Language Processing; Author Profiling
Corresponding Author:	Ivandre Paraboni University of Sao Paulo São Paulo, SP BRAZIL
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Sao Paulo
Corresponding Author's Secondary Institution:	
First Author:	Rafael Felipe Dias
First Author Secondary Information:	
Order of Authors:	Rafael Felipe Dias Ivandre Paraboni
Order of Authors Secondary Information:	
Funding Information:	FAPESP (2016/14223-0) Dr. Ivandre Paraboni

jiis manuscript No.
(will be inserted by the editor)

Author profiling from Text: new tasks and towards cross-domain modelling

Rafael F. Sandroni Dias · Ivandré Paraboni

Received: date / Accepted: date

Abstract Author profiling is understood as the computational task of predicting demographic characteristics of a target author based on the text that they have written. Gender and age recognition are arguably the most well-studied author profiling tasks, and particularly so in the case of the English language. Other tasks and languages, by contrast, are less commonly addressed. Based on these observations, this paper presents a series of experiments addressing a number of existing and also a number of relatively novel author profiling tasks, the use of psycholinguistic knowledge in author profiling and the issue of cross-domain modelling. Preliminary results suggest that, whilst all proposed tasks may be accomplished (at least to a certain extent), the use of word-based features still trumps psycholinguistic knowledge, and that cross-domain accuracy loss may in some cases be alleviated by simply adding more training data from an available text source.

Keywords Natural Language Processing · Author Profiling · Cross-domain Profiling

1 Introduction

in Natural Language Processing (NLP), author profiling is understood as 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 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 [23,24,22], with practical applications including marketing research, on-line fraud detection, copyright and plagiarism investigations, among many others.

School of Arts, Sciences and Humanities
University of São Paulo
Av Arlindo Bettio, 1000 - São Paulo, Brazil
E-mail: {rafaelsandroni,ivandre}@usp.br

A piece of text may provide multiple clues about its author's demographics. Among these, the most commonly knowledge source is word usage, based on the assumption that different categories of people will make (perhaps slightly) different word choices. Thus, for instance, we may expect the vocabulary of an individual to vary, at least to some extent, according to the number of years spent in education, their gender, professional occupation, religious or political views, among many other possibilities. Thus, the challenge in lexically-motivated author profiling is to distinguish author categories based on word usage.

Existing studies are often focused on age and gender prediction, and are in many cases limited to the use of English text. In addition to that, existing models are usually based on text features, and seldom benefit from existing resources such as the LIWC psycholinguistic dictionary [33], which is perhaps best-known for its role in personality prediction [13].

Based on these observations, this paper aims to further research in the field of author profiling by discussing a series of experiments involving various tasks and methods.

As in many (or most) studies in the field, we shall focus the relation between word usage and author's demographics. Despite some obvious limitations of this approach (namely, two individuals who belong to different categories may in principle write exactly the same piece text, and are therefore indistinguishable), we assume that having some estimate of an author's profile may be still useful for many applications, and perhaps particularly so in the case of text forensics.

More specifically, our experiments are intended to shed light on three main issues: (a) potentially novel tasks of this kind, (b) the use of psycholinguistic knowledge in author profiling and (c) the issue of cross-domain modelling. Each of these issues is briefly discussed in turn as follows.

First, regarding the issue of possible author profiling tasks, we notice that gender [11,32] and age [2,4] recognition are arguably the most well-studied tasks in the field, and particularly so in the case of the English language. However, the increasing availability of text corpora labelled with other types of information about their authors gives rise to many additional possibilities. These include the recognition of personality traits [13,12], political leaning [19,31,34], moral foundations [10], user type [11], income [8] and others. A number of less common tasks along these lines will be addressed in the current work as well.

Second, we notice that author profiling models make use of a wide range of input text representations, from simple n-gram counts to more recent word embedding models. Interestingly, however, the use of psycholinguistic knowledge - such as provided by the LIWC dictionary [33] - seems to be largely restricted to the recognition of personality traits [13]. Based on the observation that LIWC includes dozens of word categories such as 'see', 'feel', 'health', 'power', 'reward', 'money' etc. that are potentially useful to other author profiling tasks, the use of LIWC features will also be addressed in the present work.

Finally, we notice that author profiling models will arguably produce optimal results for a given task when using training data from the same domain as the test data. By contrast, when training data in the required domain is not available, we may still resort to data from a different text source, in what is known as *cross-domain author* profiling. For instance, given the goal of profiling the author of an e-mail when a suitable e-mail training corpus is not available, we may attempt to use a model built from, e.g., Facebook data instead. Cross-domain author profiling poses a number of challenges stemming from text genre, size and overall quality [15], and it is also the focus of some of our present experiments.

2 Background

Author profiling - particularly in the case of gender and age recognition - has been the focus of increasingly large number of studies in recent years, many of which developed around the PAN-CLEF competitions [23,24]. As methods are generally similar, the this section will focus in existing work on gender and age recognition only, addressing the issues of single- and cross-domain author profiling separately.

2.1 Single-domain author profiling

Practical single-domain age and gender recognition from text poses a number of well-known difficulties [17] stemming from text genre, quality and size, among others. Although results may remain modest in some settings (particularly in the case of age recognition), computational models of this kind attempt to circumnavigate many of these difficulties by investigating a plethora of machine learning methods and text representations.

Table 1 summarises a number of recent studies in gender and age recognition for the English (en), Spanish (sp), Portuguese (pt), Arabic (ar), Urdu (ur) and Swedish (sw) languages, as well as their main learning features (word, character and part-of-speech n-grams, word embeddings, TF-IDF and LIWC counts) and methods (SVM, Logistic Regression, Convolutional, Recurrent and Long Short-term Memory networks.)

Table 1 Recent author profiling gender/age recognition studies.

Study reference	Task	Language	Features	Method
[2]	age	en,sp,pt,ar	word/char n-grams	SVM
[14]	gender	en,sp,pt,ar	POS n-grams	Logistic Reg.
[30]	gender	en,sp,pt,ar	word embeddings	CNN
[4]	gender,age	en,ar,ur	word/char n-grams	SVM
[26]	gender	en	POS n-grams, TFIDF	Logistic Reg.
[9]	gender	en,sp,fr,ru,sw	LIWC	SVM
[1]	age	pt	word embeddings	CNN
[6]	gender	en	word embeddings	CNN, LSTM
[11]	gender,age	en	words	RNN, LSTM
[32]	gender	en,sp,ar	word embeddings, image	CNN

Given the wide range of task definitions, text genres, datasets and target languages under consideration, a direct comparison between existing studies is not straightforward. The exception is the three first studies [2,14,30] in Table 2, all of which developed in the light of the PAN-CLEF 2017 gender and language variety identification task in Twitter [24]. Among these, the work in [2], which is arguably the simplest model among the three, was the overall winner of the competition, having slightly outperformed a model with added POS information in [14]. The much more sophisticated model in [30], by contrast, was outperformed by 10 out of 22 competitors.

The work in [4] is another example of simple and effective strategy for gender and age recognition in text. As in the case of [2], the study also makes use of of word and char n-grams with SVMs.

POS information plays a central role also in [26]. 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 [9] is one of the few attempts to use psycholinguistic features obtained from the LIWC dictionary [33]. The study evaluates the role of different word categories on gender prediction, and differences in LIWC data availability across the five languages under consideration.

The work in [1] addresses the issue of age bracket recognition using deep learning, and it is among the few to focus on the Brazilian Portuguese language (also to be addressed in our own experiments, cf. next section.) Unlike the age-related task proposed in the PAN-CLEF series, however, age classification is presently modelled as a binary classification task (young / adult). Results suggest that CNN-based models outperform a number of standard baseline alternatives.

The use of deep neural networks is also the focus of the work in [6], which presents a character-level Convolutional Bidirectional Long Short-Term Memory (LSTM) and a word-level Bidirectional LSTM using Global Vectors (GloVe) for gender recognition from Twitter text. A stacked architecture combining the character and word models outperforms the individual models and also a number of bag of words and n-grams baseline systems.

The work in [11] addresses gender, binary age bracket and user type (individual, organisation and other) classification on Twitter. These author profiling tasks are modelled as a vertex classification task 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 base on lexica, logistic regression, label propagation and others.

Finally, the PAN-CLEF 2018 competition [22] introduced a gender classification task based on a combination of text and image data. Among the participant systems, the work in [32] 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 recognition from text and, in [23], 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 [15] 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

Table 2 Corpora descriptive statistics

Domain	Vocabulary	Words	Documents	Words / docs
Facebook	63,165	2,434,215	1019	2389
Opinion	11,004	187,118	433	432
Blog	207,947	9,119,406	1572	5801
E-gov	77,396	3,760,126	49,449	76

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 [15] 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.

3 Author profiling tasks and datasets

In this section we describe the datasets (or text corpora) taken as the basis to our work, and the kinds of author profiling tasks that each dataset supports. This will be followed by the experiments proper, in the subsequent sections.

3.1 Overview

We address the author profiling task in four text genres (or domains): Facebook status updates, Opinions, Blogs and E-gov requests, and supporting 19 tasks in total 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 2 presents descriptive statistics for each domain.

The ‘Vocabulary’ column in Table 2 presents the number of unique words in each corpus. This shows that one domain is very limited (Opinion, conveying opinions about eight topics only, as discussed below), two domains show moderate complexity (Facebook and E-gov), and one domain has a broad vocabulary (Blog).

The ‘Documents’ columns presents the number of documents (or authors) in each domain. This shows that one domain is small (Opinion), two domains are moderately-sized (Facebook and Blog) and one is large (E-gov).

Finally, the ‘Words / docs’ column presents the average document size in each domain. This shows that document sizes range from small (E-gov) to large (Blog). Further details regarding each individual domain are discussed in the next sections.

3.2 Facebook domain

Facebook texts are provided by the *b5-post* corpus [20], a collection of over 194k status updates (2.2 million words in total) written by 1019 users of Brazilian Facebook. Facebook status update 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.

The *b5-post* corpus is partially labelled with gender, age, degrees of religiosity, ranging from 1 ('no religious at all') to 5 ('highly religious'), and IT background status information (denoting whether the author has any background knowledge in the field or not.) Age and religiosity information are grouped into discrete classes of approximately similar number of instances. More specifically, age information is modelled as three class brackets: from 18- to 20-years old, from 23 to 25, and 28 to 61. Following [21] and others, intermediate instances are discarded in order to minimize class overlap given that a user's status updates may span across several years, whereas their age information refers to the fixed point in time in which the texts were collected. Similarly, degrees of religiosity are modelled as three classes r12, r3 and r45. The profiling tasks supported in this domain and their class distributions are summarised in Table 3.

Table 3 Author profiling tasks and class distribution in the Facebook domain.

Task	Class	Instances
Gender	male	441
Gender	female	578
Age	a18-20	183
Age	a23-25	189
Age	a28-61	146
Religiosity	r12	217
Religiosity	r3	96
Religiosity	r45	128
IT background	yes	325
IT background	no	491

3.3 Opinion domain

Opinion texts were obtained from an ongoing data collection task, conveying over 3400 short texts (187k words in total) written by 433 on-line Brazilian micro-volunteers. Opinion texts are mostly impersonal and highly focused on the topic under discussion, and are more formal than Facebook text.

The Opinion domain consists of short moral stances produced in response to questions about eight contemporary topics including drug legalisation, abortion policies, death penalty, and others. Texts are labelled with age, gender, degrees of religiosity, IT background, education level (from 0=none to 4=post-graduate) and political views (from 1=left to 5=right). Age, religiosity, education and political views data were grouped into discrete classes. Due to differences in class distribution, however, we notice that age brackets in one dataset will not match age brackets available from the others. The profiling tasks supported in this domain and their class distributions are summarised in Table 4.

Table 4 Author profiling tasks and class distribution in the Opinion domain.

Task	Class	Instances
Gender	male	285
Gender	female	148
Age	a0-23	165
Age	a24-30	153
Age	a31-99	115
Religiosity	r1	107
Religiosity	r23	186
Religiosity	r45	140
IT background	yes	293
IT background	no	140
Education	e12	167
Education	e3	128
Education	e4	138
Politics	p12	157
Politics	p3	160
Politics	p45	116

3.4 Blog domain

Blog texts are taken from the BlogSetBR corpus [27] 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 international politics, and may include third-party material or even exerts in foreign languages.

Blog texts are partially labelled with gender, age (three classes) and education level (four classes) about their authors. The profiling tasks supported in this domain and their class distributions are summarised in Table 5.

Table 5 Author profiling tasks and class distribution in the Blog domain.

Task	Class	Instances
Gender	male	1038
Gender	female	1564
Age	a10-25	668
Age	a26-40	1020
Age	a40+	914
Education	e1	303
Education	e2	463
Education	e3	465
Education	e4	341

3.5 E-gov domain

Finally, e-gov texts were obtained from a collection of on-line 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.) Texts are partially labelled with gender, age, education (three classes), and profession information (academy, public or private sector, or self-employed). From the geographic information available from the corpus, texts were also labelled with their corresponding Brazilian geographic region (five classes, e.g., north, south etc.), and binary information representing whether the request originated from a state capital city or not. The profiling tasks supported in this domain and their class distributions are summarised in Table 6.

Table 6 Author profiling tasks and class distribution in the E-gov domain.

Task	Class	Instances
Gender	male	28805
Gender	female	15893
Age	a17-30	17219
Age	a31-42	17559
Age	a43+	14671
Education	e1	6268
Education	e2	16926
Education	e3	20682
Profession	academy	12158
Profession	private	5238
Profession	public	14932
Profession	self	3532
Region	c	8702
Region	ne	8324
Region	n	2829
Region	se	19227
Region	s	5166
Capital	yes	23698
Capital	no	19743

4 Experiment 1: Single-domain author profiling

Our fist experiment investigates a number of common and less popular author profiling tasks alike using Brazilian Portuguese as our target language. In doing so, our goal is to assess the degree of difficulty posed by each task in different domains, and to compare the use of psycholinguistics- and word-based features for each problem.

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

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 19 author profiling tasks supported by these datasets. The corpora were randomly split into training (80%) and test (20%) datasets in a stratified fashion. The test portion of each dataset was reserved for evaluation described 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 [3]. 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 compares results obtained by four author profiling models: a model based on psycholinguistic knowledge as provided by the LIWC psycholinguistic dictionary [18], 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*: LIWC-based 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 multilayer 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 LIWC dictionary [5]. *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 7. We notice that the two larger corpora - Blog and E-gov - require much larger feature sets than Facebook and Opinion.

Finally, *MLP-skipgram* makes use of multilayer perceptron classifiers using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (lbfgs) solver, and additional parameter tuning as follows. *MLP-skipgram* takes as an input a TF-IDF-weighted averaging skipgram word embedding model [16]. Both self- and pre-trained word embeddings configurations were considered[2]. We also 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 7) were retained.

Finally, we followed [25] and others and considered reducing the embedding dimensionality itself. As in the case of the previous models based on TF-IDF vectors, we will once again make use of univariate feature selection with an ANOVA

[2] Pre-trained embeddings were taken from [7].

Table 7 Optimal k values for the *LR-Tfidf* models.

Domain	Task	k
Facebook	Gender	3500
Facebook	Age	5500
Facebook	Religiosity	1000
Facebook	IT background	5500
Opinion	Gender	1000
Opinion	Age	1500
Opinion	Religiosity	1000
Opinion	IT background	3000
Opinion	Education	2000
Opinion	Politics	1500
Blog	Gender	27500
Blog	Age	25500
Blog	Education	28000
E-gov	Gender	9000
E-gov	Age	11000
E-gov	Education	14500
E-gov	Profession	29500
E-gov	Region	5000
E-gov	Capital	6000

f-value function. The choice for this particular method was however motivated by computational efficiency. For a possibly more sophisticated approach, see for instance the use of principal component analysis (PCA) for embedding dimensionality reduction in [25].

The alternative strategies for computing word embeddings and related network parameters are summarised in Table 8.

Table 8 Parameters under consideration for the *MLP-skipgram* models.

Parameter	Values
w: word embedding size	{50, 100, 300, 600}
s: word embedding source	{self, pre}
x: k-best feature set size	from 30 to w, at 10% intervals
filter: k-best word filtering	{yes, no}
it: iterations	from 100 to 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

Optimal parameter values were obtained by performing grid search over the entire training dataset. These are summarised in Table 9. 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, taken from [7].

Table 9 Optimal parameter values for the *MLP-skipgram* models.

Domain	Task	w	s	x	filter	it	l	n	f	alpha
Facebook	Gender	100	self	100	yes	400	3	75	tanh	1e-04
Facebook	Age	100	self	60	no	200	1	10	logi	1e-06
Facebook	Religiosity	100	self	100	yes	200	1	40	relu	1e-05
Facebook	IT background	50	self	50	no	400	1	5	logi	1e-06
Opinion	Gender	100	self	80	yes	250	3	25	tanh	1e-07
Opinion	Age	100	self	40	no	200	1	30	tanh	1e-04
Opinion	Religiosity	50	self	45	no	200	1	20	relu	1e-05
Opinion	IT background	100	pre	30	yes	150	1	10	relu	1e-06
Opinion	Education	50	pre	30	yes	200	1	20	relu	1e-07
Opinion	Politics	50	self	50	no	400	1	25	relu	1e-08
Blog	Gender	600	pre	600	no	200	1	300	relu	1e-05
Blog	Age	300	pre	300	no	200	3	120	relu	1e-07
Blog	Education	600	self	600	yes	250	3	400	relu	1e-04
E-gov	Gender	1000	pre	1000	yes	200	1	500	relu	1e-05
E-gov	Age	600	pre	600	no	200	1	400	relu	1e-05
E-gov	Education	1000	pre	1000	no	200	1	500	relu	1e-05
E-gov	Profession	1000	pre	1000	no	200	1	500	relu	1e-05
E-gov	Region	1000	pre	1000	yes	250	2	300	relu	1e-05
E-gov	Capital	1000	pre	1000	no	200	1	500	relu	1e-05

4.3 Results

For each of the 19 author profiling tasks under consideration, Table 10 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.

Table 10 Weighted F1 scores. Best results for each task are highlighted.

Domain	Task	Baseline	LR-LIWC	LR-Tfidf	MLP-skipgram
Facebook	Gender	0.41	0.50	**0.80**	0.73
Facebook	Age	0.20	0.35	0.48	**0.54**
Facebook	Religiosity	0.33	0.33	0.47	**0.54**
Facebook	IT background	0.45	0.46	0.66	**0.72**
Opinion	Gender	0.52	0.63	0.70	**0.74**
Opinion	Age	0.21	0.33	0.51	**0.60**
Opinion	Religiosity	0.25	0.25	0.46	**0.63**
Opinion	IT background	0.55	0.58	**0.76**	0.72
Opinion	Education	0.21	0.26	0.45	**0.59**
Opinion	Politics	0.20	0.36	0.47	**0.55**
Blog	Gender	0.45	0.66	0.75	**0.78**
Blog	Age	0.22	0.43	0.52	**0.54**
Blog	Education	0.13	0.29	0.40	**0.46**
E-gov	Gender	0.51	0.60	**0.79**	**0.79**
E-gov	Age	0.19	0.42	0.59	**0.60**
E-gov	Education	0.30	0.46	0.62	**0.65**
E-gov	Profession	0.24	0.42	0.61	**0.64**
E-gov	Region	0.26	0.37	0.65	**0.67**
E-gov	Capital	0.39	0.57	0.72	**0.73**

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.

Since corpus sizes are not equal, comparisons between genres are generally unhelpful. We notice however that results from Opinion texts tend to be superior to those obtained from Facebook for the same tasks (gender, age and religiosity.) despite the fact that the corpus is smaller. This may be explained by the observation that the Opinion domain contains texts that are generally more well-formed, and less noisy than Facebook status updates.

5 Experiment 2: Cross-domain gender recognition from individual data sources

Although single-domain author profiling will arguably produce optimal results for a given task, when training data of the required domain is not available, we may 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 profiling 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 (namely, by effecting the smallest loss in F1 scores.) To this end, we will take the case of the gender information task as a working example. The choice for this task is motivated by the observation that gender information is ubiquitous in a wide range of labelled corpora, and it is readily comparable across domains with little ambiguity (as opposed to, e.g., age information, which has different distributions across domains, as discussed in Section 4.2.) Moreover, gender recognition achieved the best results among the single-domain profiling tasks in the previous experiment, suggesting a higher standard to be achieved by cross-domain alternatives.

5.1 Data

The experiment makes use of the same four corpora in the previous experiment, namely, Facebook, Opinion, Blog and E-gov, all of which providing (male/female) author gender labels.

5.2 Models

For each of the four domains under consideration - Facebook, Opinion, Blog and E-gov - a gender recognition model was built using the previous *LR-Tfidf* approach for simplicity (cf. Section 4.2). 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 single-domain gold standard results obtained by performing 10-fold cross-validation on each individual dataset.

Table 11 Cross-domain F1 loss compared to the single-domain strategy. Rows represent a training domain and columns represent a test domain. Lower (and better) results 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	-

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. Results for all possible domain combinations are shown in Table 11, 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 recognition 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 Blogs data, with a 0.08 loss.

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 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. For instance, the E-gov test domain has a consistently high loss in all scenarios, and in the case of the simpler Opinion test domain there seems to be little difference between using the large E-gov dataset and the much smaller Facebook dataset. These issues will be further addressed in a complementary experiment described in the next section.

6 Experiment 3: Cross-domain gender recognition from multiple data sources

Results from the previous experiment support the simple machine learning principle that more data usually helps classification tasks in general. Based on this observation, we envisage a third experiment in which gender prediction in a given test domain is attempted by using training data provided by all other available sources, that is, by using 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 training data come from) obtained from multiple sources combined. Thus, for instance, we will predict Facebook author's gender by using a model built from data taken from Opinion, Blog and E-gov texts combined, and so forth.

Other than concatenating training data from multiple text sources, the present experiment setting is similar to the previous Experiment 2. We will once again test all four text genres available, and we will measure weighted F1 loss by comparing

Table 12 Male/Female class distribution for cross-domain gender recognition using multi-domain data sources.

Domain	Male	Female
All except Facebook	30654	17079
All except Opinion	30810	17509
All except Blog	29531	16619
All except E-gov	2290	1764

Table 13 Cross-domain F1 loss using multiple- and single-domain models (from the previous experiment 2.) Lower (better) results are highlighted.

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

their results to those obtained from the single-domain gold standard as discussed in the previous sections.

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 12.

By concatenating data sources in this way, we notice that the first three datasets are now similarly sized. The exception is the 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.

6.2 Models

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

6.3 Results

Table 13 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 recognition 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 recognition 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 recognition in blogs. Although the previous (in cross-domain source) model 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 largely circumnavigated by simply using more training data from perhaps any available domain.

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 small-scale experiments addressing a number of author profiling tasks in various text genres, and discussed the issue of cross-domain gender recognition from single and combined data sources.

The first experiment examined a number of existing and potentially novel author profiling tasks in the Brazilian Portuguese language. Generally speaking, 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 the case of gender recognition in situations in which single-domain data 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 particular, cross-domain loss may become acceptably small if more training data is available.

The observation that more data may alleviate the losses in cross-domain gender recognition 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 single-domain profiling, but to a lesser extent than in the previous experiment. This suggests that, at least for the gender recognition task, cross-domain learning may be in principle feasible provided that a sufficiently large amount of data is available, an outcome that seems consistent with previous findings in the field.

Regardless of training data size, however, we notice that certain profiling tasks - e.g., Facebook gender recognition - seem to be more successfully accomplished than others and, accordingly, cross-domain strategies may be more suitable to some domain/task combinations. 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., [15]. More research is however required to determine which other factors may affect cross-domain learning, and to which extent the present cross-domain strategies may be generalised to other (perhaps less directly comparable) author profiling tasks.

As future work, we intend to take a closer look at 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 lexica [28,29]. Methods of this kind, which are clearly attractive for reasons of computational efficiency and potential for generalisation, still require further testing in cross-domain settings.

Acknowledgements This work has been supported by FAPESP grant 2016/14223-0.

References

1. aes, R.G.G., Rosa, R.L., de Gaetano, D., Rodríguez, D.Z., Bressan, G.: Age groups classification in social network using deep learning. IEEE Access **5**, 10805–10816 (2017). DOI 10.1109/ACCESS.2017.2706674
2. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-GrAM: New groningen author-profiling model. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. Dublin (2017)
3. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: Simply the best: Minimalist system trumps complex models in author profiling. In: LNCS vol. 11018, pp. 143–156. Springer, Cham (2018)
4. Fatima, M., Hasan, K., Anwar, S., Nawab, R.M.A.: Multilingual author profiling on facebook. Information Processing & Management **53**(4), 886–904 (2017). DOI https://doi.org/10.1016/j.ipm.2017.03.005
5. Filho, P.P.B., Aluísio, S.M., Pardo, T.: An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: 9th Brazilian Symposium in Information and Human Language Technology - STIL, pp. 215–219. Fortaleza, Brazil (2013)
6. Gopinathan, M., Berg, P.C.: A deep learning ensemble approach to gender identification of tweet authors (2017)
7. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluísio, S.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: 11th Brazilian Symposium in Information and Human Language Technology - STIL, pp. 122–131. Uberlândia, Brazil (2017)
8. Hasanuzzaman, M., Kamila, S., Kaur, M., Saha, S., Ekbal, A.: Temporal orientation of tweets for predicting income of users. In: 55th Annual Meeting of the Association for Computational Linguistics, pp. 659–665. Association for Computational Linguistics, Vancouver (2017)
9. Isbister, T., Kaati, L., Cohen, K.: Gender classification with data independent features in multiple languages. In: European Intelligence and Security Informatics Conference (EISIC-2017), pp. 54–60. IEEE Computer Society, Athens, Greece (2017)
10. Johnson, K., Goldwasser, D.: Classification of moral foundations in microblog political discourse. In: 56th Annual Meeting of the Association for Computational Linguistics, pp. 1–11. Association for Computational Linguistics, Melbourne (2018)
11. Kim, S.M., Xu, Q., Qu, L., Wan, S., Paris, C.: Demographic inference on Twitter using recursive neural networks. In: Proceedings of ACL-2017, pp. 471–477. Vancouver, Canada (2017)
12. Liu, F., Perez, J., Nowson, S.: A language-independent and compositional model for personality trait recognition from short texts. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 754–764. Association for Computational Linguistics, Valencia, Spain (2017)
13. Mairesse, F., Walker, M., Mehl, M., Moore, R.: Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research (JAIR) **30**, 457–500 (2007)
14. Martinc, M., Skrjanec, I., Zupan, K., Pollak, S.: PAN 2017: Author profiling - gender and language variety prediction. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. Dublin (2017)

15. Medvedeva, M., Haagsma, H., Nissim, M.: An analysis of cross-genre and in-genre performance for author profiling in social media. In: LNCS vol. 10456. Springer, Cham (2017)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013)
17. Nguyen, D.P., Trieschnigg, R.B., Dogruoz, A.S., Gravel, R., Theune, M., Meder, T., de Jong, F.M.: Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In: Proceedings of COLING-2014, pp. 1950–1961. Association for Computational Linguistics (2014)
18. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Inquiry and Word Count: LIWC. Lawrence Erlbaum, Mahwah, NJ (2001)
19. Preotiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond binary labels: Political ideology prediction of twitter users. In: 55th Annual Meeting of the Association for Computational Linguistics, pp. 729–740. Association for Computational Linguistics, Vancouver (2017)
20. Ramos, R.M.S., Neto, G.B.S., Silva, B.B.C., Monteiro, D.S., Paraboni, I., Dias, R.F.S.: Building a corpus for personality-dependent natural language understanding and generation. In: 11th International Conference on Language Resources and Evaluation (LREC-2018), pp. 1138–1145. ELRA, Miyazaki, Japan (2018)
21. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: CLEF 2015 Evaluation Labs and Workshop. CEUR-WS.org, Toulouse, France (2015)
22. Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In: L. Cappellato, N. Ferro, J.Y. Nie, L. Soulier (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs, CEUR Workshop Proceedings. CLEF and CEUR-WS.org (2018)
23. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In: CLEF 2016 Evaluation Labs and Workshop, Notebook papers, pp. 750–784. CEUR-WS.org, Évora, Portugal (2016)
24. Rangel, F.M., Rosso, P., Potthast, M., Stein, B.: 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 (2017)
25. Raunak, V.: Simple and effective dimensionality reduction for word embeddings. In: NIPS-2017 Limited Labeled Data workshop (2017)
26. Reddy, T.R., Vardhan, B.V., Reddy, P.V.: N-Gram approach for gender prediction. In: Advance Computing Conference (IACC), pp. 860–865 (2017)
27. dos Santos, H.D.P., Woloszyn, V., Vieira, R.: BlogSet-BR: A Brazilian Portuguese Blog Corpus. In: 11th International Conference on Language Resources and Evaluation (LREC-2018). ELRA, Miyazaki, Japan (2018)
28. Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., Schwartz, H.: Developing age and gender predictive lexica over social media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1146–1151. Association for Computational Linguistics, Doha, Qatar (2014)
29. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M., Ungar, L.H.: Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One **8**(9), e73791 (2013). DOI 10.1371/journal.pone.0073791
30. Sierra, S., y Gómez, M.M., Solorio, T., González, F.A.: Convolutional neural networks for author profiling. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. Dublin (2017)
31. Sylwester, K., Purver, M.: Twitter language use reflects psychological differences between democrats and republicans. PLoS ONE **10**(9), e0137422 (2015)
32. Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., Ohkuma, T.: 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 (2018)
33. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology **29**(1), 24–54 (2010). DOI 10.1177/0261927X09351676

34. Vijayaraghavan, P., Vosoughi, S., Roy, D.: Twitter demographic classification using deep multi-modal multi-task learning. In: 55th Annual Meeting of the Association for Computational Linguistics, pp. 478–483. Association for Computational Linguistics, Vancouver (2017)