Natural Language Engineering (2023), **29**, pp. 110–137
doi:10.1017/S1351324921000383

ARTICLE

Authorship attribution using author profiling classifiers

Caio Deutsch and Ivandré Paraboni*

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

(Received 7 September 2020; revised 5 November 2021; accepted 8 November 2021; first published online 19 January 2022)

Abstract

Authorship attribution – the computational task of identifying the author of a given text document within a set of possible candidates – has been attracting interest in Natural Language Processing research for many years. At the same time, significant advances have also been observed in the related field of author profiling, that is, the computational task of learning author demographics from text such as gender, age and others. The close relation between the two topics – both of which focused on gaining knowledge about the individual who wrote a piece of text – suggests that research in these fields may benefit from each other. To illustrate this, this work addresses the issue of author identification with the aid of author profiling methods, adding demographics predictions to an authorship attribution architecture that may be particularly suitable to extensions of this kind, namely, a stack of classifiers devoted to different aspects of the input text (words, characters and text distortion patterns.) The enriched model is evaluated across a range of text domains, languages and author profiling estimators, and its results are shown to compare favourably to those obtained by a standard authorship attribution method that does not have access to author demographics predictions.

Keywords: Authorship attribution; Author profiling; Text classification

1. Introduction

Authorship attribution – the task of identifying the author of a given document based on a set of possible candidates (Potthast *et al.* 2017) – is present in a wide range of text forensics and related applications. These include, for instance, corruption investigation models (Chen *et al.* 2011; Juola and Stamatatos 2013), on-line abuse (Vartapetiance and Gillam 2012), fake news (Peng, Choo, and Ashman 2016) and false impersonation detection (Koppel and Seidman 2018), among many others.

Studies in the field will usually draw a distinction between *closed-set* and *open-set* settings (Kestemont *et al.* 2019). In closed-set authorship attribution, the target author of the input document is assumed to be a member of the set of possible candidates and, as a result, the task consists of selecting the correct candidate among the alternatives provided. In open-set authorship attribution, by contrast, the target author may or may not be found within the candidate set, and therefore the task involves determining whether the author is unknown as well. In what follows, we shall focus on closed-set authorship attribution only.

Authorship attribution has been a popular research topic in Natural Language Processing and the subject of several shared tasks in the PAN-CLEF authorship attribution series (Potthast *et al.* 2017; Kestemont *et al.* 2018, 2019). Closed-set authorship attribution, in particular, is usually modelled as a supervised learning task, making use of text corpora labelled with author identifiers representing the classes (or authors) to be identified. Popular methods include the use of support vector machine classifiers (Schwartz *et al.* 2013; Stamatatos 2017), recurrent neural networks (Bagnall 2016; Jafariakinabad and Hua 2019), convolution neural networks (Sari and Stevenson 2016; Shrestha *et al.* 2017; Misra *et al.* 2019) and stacks of ensemble classifiers (Custódio and Paraboni 2019), as we shall discuss later.

Despite advances in recent years, authorship attribution continues to attract interest as a research problem (Kestemont *et al.* 2019). At the same time, significant advances have also been observed in the related field of author profiling, that is, the computational task of learning author demographics from text (Silva and Paraboni 2018; Rangel and Rosso 2019). As in the case of closed-set authorship attribution, author profiling is often modelled as a supervised problem (i.e., relying on text corpora labelled with demographics information.) Systems of this kind have been applied to a wide range of tasks, most noticeably in gender and age classification (Kim *et al.* 2017; Takahashi *et al.* 2018; Rangel *et al.* 2020), but also in the recognition of personality traits (dos Santos, Ramos, and Paraboni 2019), bot detection (Pizarro 2019) and many others.

1.1 Authorship attribution using author profiling

Given the close relation between authorship attribution and author profiling – in the sense that both tasks are focused on gaining knowledge about the individual who wrote a piece of text – in the present work we shall argue that demographics predictions obtained from author profiling methods may help reduce the search space (i.e., the number of author candidates under consideration) in the authorship attribution task and, as a result, improve overall accuracy.

To illustrate this, let us assume, for instance, that we are able to infer the gender (male/female) of the individual who wrote a given piece of text by using a standard author profiling method as in, for example, Basile *et al.* (2017) or many others. In this case, gender predictions may effectively split the set of candidates under consideration into two groups (i.e., men and women), letting an authorship attribution model to focus on the subset of candidate authors of interest. Moreover, as we shall argue in the present study, the same principle may apply not only to standard gender (or age) author profiling but also to many other (perhaps less usual) tasks, including the use of classifiers for education level, political orientation, degrees of religiosity or indeed for potentially any kind of demographics information that may be reliably inferred from labelled corpora using supervised machine learning.

Using author profiling classifiers as an aid to the authorship attribution task might seem intuitive, and it has been indeed addressed in the context of aggressive language detection (Casavantes, López, and González 2019; Garrido-Espinosa, Rosales-Pérez, and López-Monroy 2020) and other tasks. This, however, gives rise to the question of how the two tasks may be combined, and whether using possibly suboptimal profiling estimators (as it may often be the case) in this way may actually harm results. To shed light on these issues, the present work considers the authorship attribution model described in Custódio and Paraboni (2019), which consists of a stack of classifiers focused on different aspects of the input text (words, characters and text distortion patterns.) An ensemble architecture along these lines – which obtained the overall best results in closed-set authorship attribution at the PAN-CLEF shared task (Kestemont *et al.* 2018) – may not only provide a suitable basis for an extension using multiple author profiling classifiers but, as we shall argue, may actually benefit even from suboptimal profiling estimators.

1.2 Goals and contributions

Based on these observations, this work describes a number of experiments using an authorship attribution model enriched with author profiling classifiers. In doing so, our goal is to verify whether the present method may improve results in a stack architecture as proposed in Custódio

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

and Paraboni (2019) by considering a range of domains and languages and a number of standard and less-known author profiling tasks alike.

The main contributions of the present study are summarised below.

- A novel approach to closed-set authorship attribution that enriches an existing top-performing ensemble model with author profiling predictions.
- Proposed approach compares favourably to previous work in the field for a number of domains, languages, candidate set sizes and tasks.
- Author profiling models that go beyond standard gender and age classification, including classifiers for education level, political orientation, degrees of religiosity and others.

The remainder of this article is structured as follows. Section 2 provides an overview of recent computational approaches to author profiling and authorship attribution methods alike. Section 3 describes a pilot experiment intended to illustrate how having access to author demographics information may improve results of the authorship attribution task in the intended stack architecture. Section 4 presents our extended approach to authorship attribution and the author profiling models under consideration. Section 5 describes the evaluation procedure, training and test data sets for our experiments. Section 6 presents results from both individual author profiling classifiers and the extended authorship attribution model. Finally, Section 7 presents final remarks and discusses future extensions.

2. Background

In this section, we briefly review existing work in the author profiling (Section 2.1) and authorship attribution (Section 2.2) fields. For further details, we report also to the results of the recent shared tasks devoted to each task in Rangel and Rosso (2019) and Kestemont *et al.* (2019), respectively.

2.1 Author profiling

Computational author profiling consists of inferring author demographics from text. Gender and age recognition are by far the most popular tasks of this kind found in the literature and are often addressed by using supervised machine learning methods. Author profiling has been the centre of a number of shared tasks in the PAN-CLEF series (Rangel and Rosso 2019), most notably focused on age and gender prediction in the Twitter domain, although other tasks (e.g., recognising personality traits, language variation, bots, etc.), languages (e.g., Arabic, Dutch etc.) and modalities (e.g., learning from both images and texts) have been addressed as well.

As a brief introduction to recent approaches to author gender and age profiling, Table 1 summarises a number of selected studies in the field, including some of the top-performing systems at PAN-CLEF in 2017 (Basile *et al.* 2017), 2018 (Takahashi *et al.* 2018) and 2019 (Pizarro 2019).

We notice that most approaches are based on Twitter data, make use of word- and character n-gram models, and often based on SVM or logistic regression classifiers. Further details are discussed as follows.

The early work in Nguyen *et al.* (2014) introduces a number of useful insights in gender and age prediction alike. The study compares machine and human performance in gender and age prediction from Twitter texts and discusses a number of the limitations of popular computational approaches to these tasks. The study points out differences between social and biological identities, and shows that, for over 10% of Twitter users, there is a mismatch between their biological sex and the kind of language they use on social media, and that older users tend to be perceived to be younger than what they actually are. The study makes use of Dutch tweets translated to English and compares standard computational models (linear regression for age prediction, and

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 1. Selected recent approaches to gender and age author profiling according to task (A=age and G=gender), domain, language (En=English, Sp=Spanish, Pt=Portuguese, Ar=Arabic, Fr=French, Ru=Russian, Sw=Swedish), text features (w=word, c=character, or p=part-of-speech n-grams, LIWC (Language Inquiry and Word Count) counts, w2v=Word2vec word embeddings and method (SVM=support vector machines, LR=logistic regression, NB=Naive Bayes, RF=Random Forest, CNN=Convolutional Neural Networks, MLP=multilayer perceptron, LSTM=Long Short-term Memory networks.)

Study	Tasks	Domains	Languages	Features	Methods
Nguyen *et al.* (2014)	G,A	Twitter	En	w	LR
Basile *et al.* (2017)	G	Twitter	En,Sp,Pt,Ar	w,c	SVM
Reddy, Vardhan, and Reddy (2017)	G	Reviews	En	p,w	LR
Isbister, Kaati, and Cohen (2017)	G	Blogs	En,Sp,Fr,Ru,Sw	LIWC	SVM
Kim *et al.* (2017)	G,A	Twitter	En	w	LSTM
Takahashi *et al.* (2018)	G	Twitter+images	En,Sp,Ar	w2v	CNN
Pizarro (2019)	G	Twitter	En,Sp	w,c	SVM,LR,NB
Rangel *et al.* (2020)	G,A	Twitter	En,Sp,Ar,Pt	w	SVM,LR,MLP,RF

logistic regression for gender prediction) with human evaluation. A majority-vote model obtains an accuracy of 0.84, which is similar to existing author profiling classifiers for English Twitter data.

The work in Basile *et al.* (2017) may be seen as a standard approach to author gender profiling, and it was the overall best-performing participant in the PAN-CLEF-2017 author profiling shared task (Rangel *et al.* 2017). The system obtained 0.83 average accuracy in author gender classification by making use of a linear SVM model with word unigrams and 3.5 character n-gram counts as learning features. Other language- and domain-related features such as part-of-speech (POS) tags and Twitter handles were found to actually harm overall accuracy.

In the work in Reddy *et al.* (2017), by contrast, the use of POS information plays a more prominent role in a gender classification task. The study introduces a TF-IDF (term frequency–inverse document frequency) weighted POS n-gram model that outperforms a number of standard baseline alternatives (e.g., bag of words, etc.) in the hotel reviews domain.

Unlike most data-driven approaches to author profiling, in Isbister *et al.* (2017), author gender classification is addressed with the aid of psycholinguistic features computed from the Language Inquiry and Word Count (LIWC) dictionary (Pennebaker, Francis, and Booth 2001). Results from SVM classifiers highlight the role of different LIWC categories in the task and differences across languages.

The work in Kim *et al.* (2017) addresses the issues of gender, age and user type Twitter profiling in the English language by classifying graph vertices with the aid of recursive neural networks (RNNs.) To this end, network, text and label information are combined into tree structures and fed into individual RNNs. The approach is found to outperform a number of robust baseline systems (lexica, logistic regression, label propagation, text-associated DeepWalk and Tri-Party Deep Network Representations) in the three tasks under consideration.

The work in Takahashi *et al.* (2018) was the overall best-performing system in the PAN-CLEF 2018 author profiling shared task (Rangel *et al.* 2018), addressing the issue of gender classification based on multimodal input, that is, conveying both text and image data. To this end, a neural approach called 'Text Image Fusion Neural Network' (TIFNN) is introduced in order to leverage both data sources and produce gender predictions accordingly.

Finally, we notice that many of the early approaches to gender and age profiling have been recently outperformed by the work in Rangel *et al.* (2020), which enhances previous methods with

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 2. Selected recent approaches to closed-set authorship attribution according to domain, language (En=English, Sp=Spanish, Fr=French, It=Italian, Pl=Polish), text features (w=word, c=character, or p=part-of-speech n-grams; Glove word embeddings, or PCFG=probabilistic context-free grammars) and methods (SVM=support vector machines, LR=logistic regression, NB=Naive Bayes, RF=Random Forest, Markov Chains, fastText, LDA=Latent Dirichlet allocation, Stacked Ensembles, LSTM=Long short-term memory networks, Convolutional Neural Networks (CNN) and Word Similarity)

Study	Domains	Languages	Features	Methods
Hinh, Shin, and Taylor (2016)	Essay	En	c,w	SVM
Stamatatos (2017)	News	En	c,w	SVM
Markov, Stamatatos, and Sidorov (2017)	News	En	c,w	SVM, NB
Shrestha *et al*. (2017)	Twitter	En	c,w	CNN
Rocha *et al*. (2017)	Twitter	En	c,w,p	SVM, RF
Sundararajan and Woodard (2018)	Reviews, news	En	PCFG	Markov Chains
Stevenson, Vlachos, and Sari (2017)	Legal, reviews, news	En	c,w	fastText
Patchala and Bhatnagar (2018)	News, Blogs	En	c,w,p	SVM, NB
Reddy *et al*. (2018)	Reviews	En	c	LR, NB
Jafariakinabad and Hua (2019)	News, Blogs	En	p, Glove	LSTM
Custódio and Paraboni (2019)	fanfics	En,Sp,Fr,Ii,Pl	c,w,p	Ensemble
Sharon Belvisi, Muhammad, and Alonso-Fernandez (2020)	Twitter	En	c,w	Similarity

the use of a novel text representation – called LDSE (Low-Dimensionality Statistical Embedding) – that takes into account the word distributions in each author profiling class. In the present work, however, since author profiling is viewed simply as a tool to improve our main task (i.e., authorship attribution), we shall focus on a simple approach to author profiling along the lines in Basile *et al*. (2017), Reddy *et al*. (2017) and others by making use of word-based models and logistic regression as discussed in Section 4.

2.2 *Authorship attribution*

Closed-set authorship attribution (hereby called authorship attribution, for short) concerns the computational task of selecting the author of a given document from a well-defined set of candidates (Stamatatos 2017). As in the case of author profiling, authorship attribution often resorts to supervised machine learning methods, and it has been the focus of several shared tasks in the PAN-CLEF series (Kestemont *et al*. 2019), in addition to the related tasks of author clustering (Potthast *et al*. 2017), open-set authorship attribution (Kestemont *et al*. 2019) and others.

Table 2 summarises a number of recent studies in closed-set authorship attribution. The list, which is by no means complete, is solely intended to illustrate a variety of recent approaches to the task.

Generally speaking, existing approaches to authorship attribution are largely based on word- and character n-gram models, with some methods (Stamatatos 2017; Markov *et al*. 2017) resorting to text distortion (Granados *et al*. 2011) to omit certain parts of the input text whilst focusing on others. SVM classifiers are among the most popular strategies, and input size – which may be a particular concern in the Twitter domain – has been found to be correlated with overall accuracy (Rocha *et al*. 2017). Most recent studies are devoted to the English language, with the exception of

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

those related to the PAN-CLEF authorship attribution task in Kestemont *et al.* (2018). Individual details are discussed as follows.

The study in Hinh *et al.* (2016) makes use of frame semantics from FrameNet (Baker, Fillmore, and Lowe 1998) to build a bag of frames representation intended to capture an author's writing style. The model comprises features representing frame semantics statistics such as frame element (FE) counts, average number of FEs per frame and others, and it is compared against a baseline model comprising text-related features such as vocabulary size, word and character counts. Results from a SVM classifier show that the frame semantics authorship attribution model is consistently superior to the baseline in a corpus of adversarial stylometric data.

The work in Stamatatos (2017) makes use of a text distortion method inspired from Granados *et al.* (2011), in which rare words are replaced by sequences of a special symbol '$*$', and more frequent words are kept unchanged. In doing so, authorship attribution SVM classifiers are able to focus on the text fragments that are deemed more relevant to the task. Results suggest, among other findings, that the method does improve overall accuracy and that function words are less suitable to text distortion.

Text distortion is also performed at the pre-processing stage of input texts for authorship attribution in Markov *et al.* (2017). In this case, numbers, named entities and highly frequent words are replaced by special symbols. Results from SVM and multinomial Naive Bayes classification suggest that the method compares favourably to a standard bag-of-words approach without pre-processing.

The work in Shrestha *et al.* (2017) investigates a number of CNN architectures for authorship attribution in social media texts, taking as an input character unigram and bigram embeddings, and skip-gram word embedding representations. Results are compared against those obtained by a range of baseline systems, including the use of logistic regression with variable length character n-grams, and Long Short-Term Memory networks (LSTMs) with character bigrams.

The work in Rocha *et al.* (2017) focuses on the issues of closed- and open-set authorship attribution (of which only closed-set scenarios are presently dealt with) with limited input data using short texts from a corpus of 10 million tweets posted by 10,000 users (authors). The work makes use of SVM, Random Forest, distance-based and text compression methods built from word, character and POS n-grams. A number of experiments were carried out by varying both the number of candidate authors under consideration and the number of input texts per author. Among other findings, the study suggests that the text compression method outperforms the alternatives for small input sizes and that overall accuracy decreases linearly as the number of candidate authors increases.

The study at Sundararajan and Woodard (2018) investigates the role of syntax and word choice in authorship attribution, which may be particularly relevant to cross-genre scenarios in which content-based information does not play a significant role. Syntax is investigated with the aid of context-free probabilistic grammars (PCFG) and Markov chain models, and the issue of word choice is addressed by masking out certain words or topics (which may be seen as an instance of text distortion) corresponding to different POS categories. Results suggest that cross-genre scenarios may benefit from syntactic knowledge, whereas both single- and cross-domain scenarios may benefit from lexical knowledge. Moreover, purely syntactic models were found to be insufficient by themselves, and may require combination with more content-oriented (e.g., character-based) models. In particular, common nouns, verbs, adjectives and adverbs were found to help author identification, whereas proper nouns do not.

The study in Stevenson *et al.* (2017) addresses the use of continuous word and character n-gram representations for authorship attribution in four domains using fastText (Joulin *et al.* 2017). In doing so, the model focuses on short word and character sequences, but it does not keep track of longer dependencies. Feature representations and classifiers are built jointly by adapting the fastText shallow architecture, and results suggest that the use of continuous character n-gram

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

representations outperform a number of baseline systems (and the use of continuous word n-grams) in two domains (news and reviews.) On the other hand, the use of topic modelling was still superior in the case of legal texts authorship attribution.

The work in Patchala and Bhatnagar (2018) introduces an authorship attribution model based on topic-independent syntactic templates built from each candidate author of interest, and which are intended to represent an individual's writing style. Results obtained from a number of standard classifiers (e.g., SVM, Naive Bayes and others) suggest that the combination of parsed tree structures and additional syntactic features outperforms the use of individual features alone.

The study in Reddy *et al.* (2018) introduces an instance-based authorship attribution method that relies on author-specific document weights to represent input texts, rather than document features or terms. Document weights are obtained by first computing terms weights, which are subsequently normalised by author. A number of experiments – with and without document weighting – were carried out using standard classifiers (logistic regression, Naive Bayes and Random Forest) based on a bag of words model. Results from a small (10-authors) reviews corpus suggest that document weighting generally increases task accuracy.

The work in Jafariakinabad and Hua (2019) presents a neural model that encodes document information from lexical, syntactic and structural levels for authorship attribution. In this approach, syntactic and lexical sentence representations are jointly encoded, and subsequently an attention-based hierarchical network encodes the syntactic and semantic structures of input texts themselves while rewarding those that help capturing the writing style of their authors. The model is evaluated against a number of SVM and CNN baseline systems, including the approaches in Shrestha *et al.* (2017) and Stevenson *et al.* (2017). Results show the strength of each individual level of document information and suggest that the proposed model outperforms the baseline alternatives and its individual components alike.

Unlike existing machine learning approaches to authorship attribution, the work in Sharon Belvisi *et al.* (2020) takes a forensic approach to the task by comparing the use of standard n-gram and stylometric features (e.g., character, word and punctuation counts etc.) through text similarity. More specifically, the evaluation of different features is carried out by measuring the similarity between representations of different authors using Cosine, Euclidean and Manhattan distances. Results based on a small (40-users) Twitter corpus suggest that the use of idiosyncratic features (e.g., misspellings, abbreviations, emoji counts, etc.) outperforms the use of n-gram counts by a small margin.

Finally, the stack ensemble approach in Custódio and Paraboni (2019) will be taken as the starting point to the present work, and for that reason is discussed in more detail in the next section.

2.3 EACH-USP ensemble approach to authorship attribution

The EACH-USP approach to closed-set authorship attribution described in Custódio and Paraboni (2019) is based on the assumption that identifying the author of a given document may require relying on multiple knowledge sources. To this end, the approach makes use of standard word- and character-based n-gram models, and an additional character-based model subject to text distortion (Granados *et al.* 2011). The output probabilities of the three models – hereby called word, char and distorted char – are combined in a stack architecture (Wolpert 1992) and subject to a second-level logistic regression classifier to determine the author of an input document. This architecture is illustrated in Figure 1.

Text distortion has been introduced in Granados *et al.* (2011) and has been previously considered in authorship attribution (Stamatatos 2017), deception detection (Sánchez-Junquera *et al.* 2020) and other tasks, and it is largely intended to mask out words that are not relevant to the task. In Custódio and Paraboni (2019), by contrast, text distortion is performed at the character level. More specifically, the model replaces every character in the input text – except punctuation

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 3. An example of text distortion in the EACH-USP approach (in Portuguese)

Original text	Distorted text
`Isso é apenas um pequeno, e não muito completo, exemplo de distorção textual (em Português)...`	`**** é ****** ** *******, * *ã* ***** ********, ******* ** ******çã* ******* (** *******ê*)...`

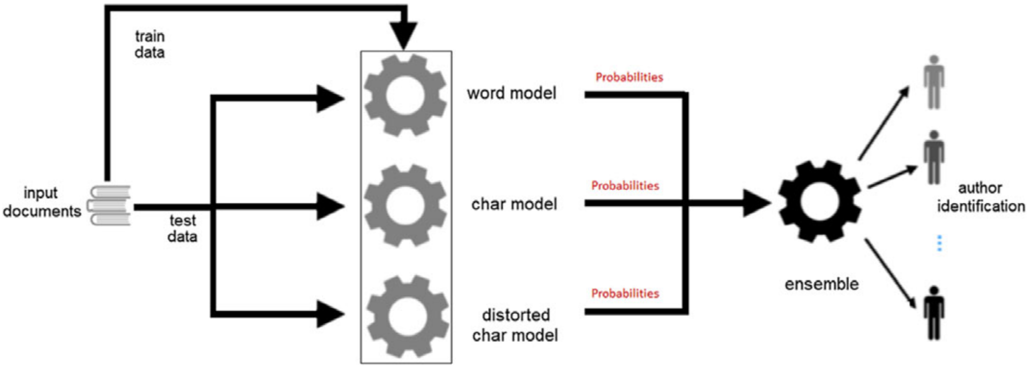

Figure 1. EACH-USP ensemble authorship attribution, adapted from Custódio and Paraboni (2019). Input documents (left) are split into train and test portions, and pre-trained word, char and distorted char models (centre) are built from training data. Test documents are submitted to the individual classifiers independently, and their predictions are taken as the input to the final, second-level classifier (right).

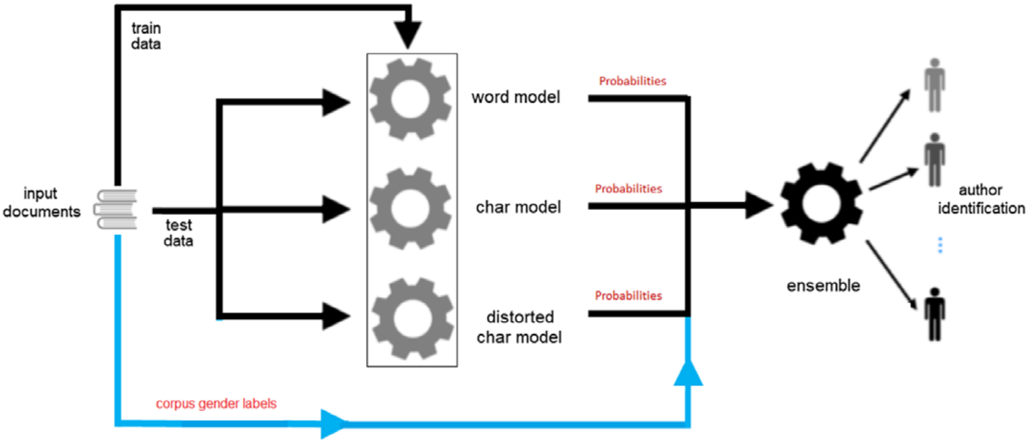

Figure 2. Ensemble authorship attribution with added author gender information. Predictions made by the individual word, char and distorted char models (centre) from the EACH-USP approach (cf. Figure 1), alongside gender labels, are taken as the input to the final, second-level classifier (right).

and diacritics – for a '$*$' symbol so that the model is able to focus on these particular patterns. An example of text distortion of this kind – rendered in Portuguese to show diacritics usage – is illustrated in Table 3.

The attention to punctuation and diacritics patterns has been found to be particularly useful for more general, cross-domain authorship attribution tasks in multiple languages, as in Kestemont *et al.* (2018). As for the combination of the three individual classifier components, this works as follows. First, the set of d input documents is vectorised by making use of a word-, char- or distortion-based feature extraction function $V(d)$ as required by each model (or channel), and the

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

resulting feature set X is normalised by a function $N(X)$. Next, X is subject to PCA dimensionality reduction, and a multinomial classifier generates the probability $P(Y=k)$ for each class k.

First-level classifiers are optimised by making use of a second-level model $\sigma(\sum_c \sum_i (w_{ci} * c_i) + k)$, where c_i is the probability of a candidate author i being the actual author of the given document according to the c classifier, w_{ci} is the weight of c_i, k is a constant and σ is the *sigmoid* function. This produces a new s vector of i probabilities of each candidate author (or class) being the author of the document.

This stack ensemble approach was evaluated at the PAN-CLEF authorship attribution shared task in Kestemont *et al.* (2018), which considered cross-domain authorship attribution scenarios based on fan fiction texts written in five languages, and required identifying the author of a text written in a particular genre (e.g., Harry Potter) based on texts written in a different genre (e.g., Star Wars.) Given the top-performing results reported in Kestemont *et al.* (2018), and the observation that a stack architecture of this kind may be easily extended with any number of additional (e.g., author profiling) classifiers, this approach will be taken as the basis to the present work as well.

3. Pilot study: does author profiling information help authorship attribution?

Before introducing our current profiling-based approach to authorship attribution, we will first examine the extent to which having access to author demographics information may actually help author identification. To this end, we envisaged a simple pilot study in which author gender information available from a labelled corpus is fed directly into the authorship attribution ensemble in Custódio and Paraboni (2019). This strategy, which amounts to using ground truth information instead of predictions made by an author gender classifier, is intended to illustrate whether using gender information may help authorship attribution at all and, if so, what upper and lower limits of accuracy an actual gender classifier would be expected to achieve in order to effectively help authorship attribution. After discussing these issues, the use of actual classifiers will be the focus of our main approach in Section 4.

3.1 Overview

The present analysis makes use of the b5-corpus of Facebook texts described in Ramos *et al.* (2018), which will be further discussed in Section 4.1 as part of our main author profiling experiments. In this experiment, binary gender labels (male/female) available from the corpus are added as a fourth information source to the authorship attribution ensemble in Custódio and Paraboni (2019), that is, in addition to the word, character and text distortion channels described in the previous Section 2.3. The resulting ensemble is illustrated in Figure 1 and essentially differs from the original architecture only by presenting a fourth (blue) channel at the bottom, which is intended to represent the gender label information taken from the input texts.

The use of binary gender information in this way is similar to the *ap.label* approach to be discussed in Section 4.2. In its present form, 0/1 gender labels are combined with the probabilities obtained by the three ensemble components and taken as the input to the second-level authorship attribution stack classifier.

Using gender labels available from the input text will provide us with an upper limit for the accuracy that the ensemble authorship attribution model may be able to achieve when using an optimal author profiling classifier. In practice, however, author profiling classifiers will most likely obtain much lower results. To shed light on this issue, we ran a number of simulations in which different levels of noise were added to the model, so that the actual gender information was corrupted by a certain margin. By comparing multiple authorship attribution scenarios based on gender estimates of varying degrees of robustness, we would like to establish the lower limit of accuracy that a gender classifier would be expected to achieve in this particular setting.

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 4. Authorship attribution mean results with and without gender information in different degrees of accuracy

Metrics	Baseline	100%	90%	80%	70%	60%	50%
F1	0.32	0.42	0.38	0.35	0.33	0.32	0.32
Precision	0.33	0.42	0.39	0.36	0.34	0.33	0.32
Recall	0.32	0.42	0.38	0.35	0.33	0.32	0.32
Accuracy	0.33	0.42	0.38	0.36	0.34	0.33	0.32

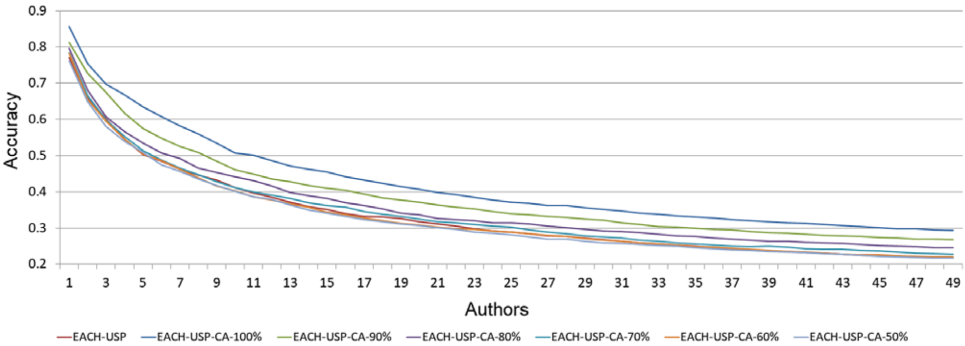

Figure 3. Model accuracy along increasing number of candidate authors.

3.2 Procedure

The 50 authors with the largest amount of text available from the corpus were selected for this analysis, and their texts were split into document units (or posts) at line breaks. The study consisted of comparing authorship attribution results obtained by the standard EACH-USP approach in Custódio and Paraboni (2019), which is presently taken as a baseline system, and its extended version that includes gender information with a certain level of added noise. Assuming that gender labels available from the corpus are 100% correct, we tested a number of scenarios in which gender information was corrupted so as to obtain 90%, 80%, 70%, 60% and 50% accuracy, hence simulating author profiling classifiers of different levels of robustness.

Testing was carried out as follows. First, two authors are randomly selected and taken as the input to both models (with and without gender information.) Next, additional authors are randomly selected one at a time, and the procedure is repeated until reaching 50 authors. For the largest (i.e., 20-author) setting, this corresponds to 5600 train and 2400 test documents. At each turn, we compute accuracy, precision, recall and F1 measures. In order to minimise possible effects of random selection, the experiment is repeated 20 times, and we report its overall mean results.

3.3 Results

Table 4 summarises mean results for the EACH-USP baseline method and the alternatives that have access to additional gender information with different degrees of accuracy, ranging from 100% (hence simulating an optimal gender classifier) to 50%.

From these results, a number of observations are warranted. First, we notice that using an optimal gender classifier (as in the 100% column) would indeed help authorship attribution by a considerable margin, that is, overall accuracy would be increased by 9 points (from 0.33

to 0.42) in this particular scenario. This represents the upper (and in practical terms possibly unachievable) limit for a method based on gender author profiling classifiers. Second, we notice that using a suboptimal gender classifier would still be helpful (i.e., outperforming the use of the baseline ensemble alone) if the classifier accuracy is above 70%, which is therefore the lower limit for a gender classifier in this scenario.

Finally, we also notice that, for both approaches, accuracy decreases uniformly as the number of candidate authors (or classes to be learned) is increased. This effect, which is to be expected in a multi class machine learning setting of this kind, is illustrated in Figure 3, and it is consistent with the findings in Rocha *et al.* (2017), in which experiments in authorship attribution involving multiple candidate set sizes have been discussed at length.

Put together, these results have motivated us to implement a range of author profiling models (which are by definition suboptimal) and use their predictions in an ensemble approach to authorship attribution along the lines of the present pilot study.

4. Authorship attribution using author profiling classifiers

As discussed in the previous section, the use of ground truth author demographics obtained from corpus labels to aid the authorship attribution task suggests that author profiling methods may obtain comparable results in an automatic fashion, that is, without resorting to corpus annotation directly. To put this idea to the test, in this section, we introduce a stack ensemble method that extends the approach in Custódio and Paraboni (2019) with a number of independently-built author profiling classifiers as an aid to authorship attribution in different domains and languages.

Unlike the experiment in the previous section, however, the present approach will not rely on the actual demographics about the authors to be identified, using instead predictions made by multiple models built from a disjoint data set (i.e., which does not include any author under identification.) In other words, author profiling and authorship attribution models are independently built from different data and, despite the use of supervised author profiling methods, the present authorship attribution approach does not require the input documents to be labelled with author demographics, taking as an input only a standard set of documents labelled with unique identifiers as in the existing work in the field.

The reminder of this section will focus on the use of one author profiling classifier at a time, leaving the discussion on how to combine multiple classifiers in a single task to be dealt with in Section 6.4.

4.1 Data

Using author profiling classifiers to aid authorship attribution requires text documents labelled with both author demographics information (in order to train the author profiling classifiers) and unique author identifiers (to train the authorship attribution model proper.) This unfortunately rules out many of the existing corpora available for the purpose of authorship attribution, including those made available by the PAN-CLEF shared tasks (Kestemont *et al.* 2019) since those corpora are generally labelled only with author identifiers, but not with author demographics. Corpora developed for author profiling tasks, on the other hand, will obviously provide demographics information, but author identifiers are often unavailable.

Based on these observations, we selected a number of publicly available corpora in different domains and languages, and whose text documents are suitably labelled for both author profiling and authorship attribution tasks as required by our combined approach. More specifically, our models will be built from text in four domains: blog texts from the Blog Authorship corpus (Schler *et al.* 2006), Facebook posts from the b5-post corpus (Ramos *et al.* 2018), short essay texts about topics of a moral nature (e.g., abortion legalisation, death penalty, etc.) from the BRmoral

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 5. Corpus descriptive statistics representing the number of authors, the text unity taken as an input document, the number of text unities in each corpus, their average number of words and author profiling tasks supported by the existing labels (G=gender, A=age bracket, I=IT background, E=level of education, P=political orientation, R=degree of religiosity.)

Domain	Language	Authors	Text unity	*Unities*	Words/unity	Tasks
Blog	English	19, 320	post	918, 298	145.6	G, A
Facebook	Portuguese	1019	post	128, 310	15.2	G, A, I
Essay	Portuguese	510	sentence	10, 236	21.3	G, A, I, E, P, R
Twitter-Du	Dutch	97	tweet	149, 276	13.1	G
Twitter-Fr	French	786	tweet	1, 044, 604	13.4	G
Twitter-Ge	German	228	tweet	389, 578	12.1	G
Twitter-It	Italian	93	tweet	122, 133	13.0	G
Twitter-Pt	Portuguese	59	tweet	86, 154	12.1	G
Twitter-Sp	Spanish	615	tweet	745, 184	12.8	G

corpus (dos Santos and Paraboni 2019; Pavan *et al.* 2020) and Twitter data from the TwiSty corpus (Verhoeven, Daelemans, and Plank 2016). We notice that, in addition to providing a certain level of variety to our experiments (and hence reducing possible effects of topical bias and others, as discussed in Sari, Stevenson, and Vlachos 2018), some of these choices were motivated by our particular interest in Portuguese NLP, or were aimed at investigating the use author profiling tasks beyond standard gender and age classification.

All data sets are labelled with unique author identifiers. Blogs, Facebook and essay texts are labelled so as to support multiple author profiling tasks in one single language each, whereas Twitter texts are labelled only with author gender information available in six languages (and which may therefore be regarded as six independent corpora.) Descriptive statistics for each corpus are summarised in Table 5 and further discussed below.

Possible author profiling tasks are determined by the labels available from each corpus. All corpora are labelled with author binary gender (G) information, and therefore support binary (male/female) gender classification. With the exception of the Twitter domain, all corpora are also labelled with age (A) information, which has been presently modelled as a 3-class problem based on the distribution of each corpus. IT background is a binary label available from the b5-post (Facebook) and BRmoral (Essay) corpora only, both of which containing a significant proportion of text produced by students in Computer Science and related fields, and which indicates whether each author in the corpus has this kind of background or not. Level of Education (E), political orientation (P) and degree of religiosity (R) are crowd-sourced, self-reported labels available in the essay domain represented by the BRmoral corpus. Each of these labels supports a ternary classification problems (from basic to superior education, from left to right political orientation and from no religious at all to highly religious.) For details regarding the BRmoral corpus and its annotation scheme, we refer to Pavan *et al.* (2020).

4.2 *Author profiling and authorship attribution models*

As in the present work, author profiling classifiers have been developed only as a support to the main task of authorship attribution, in what follows we take a standard approach to the task by making use of TF-IDF unigram counts and multinomial logistic regression with univariate

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

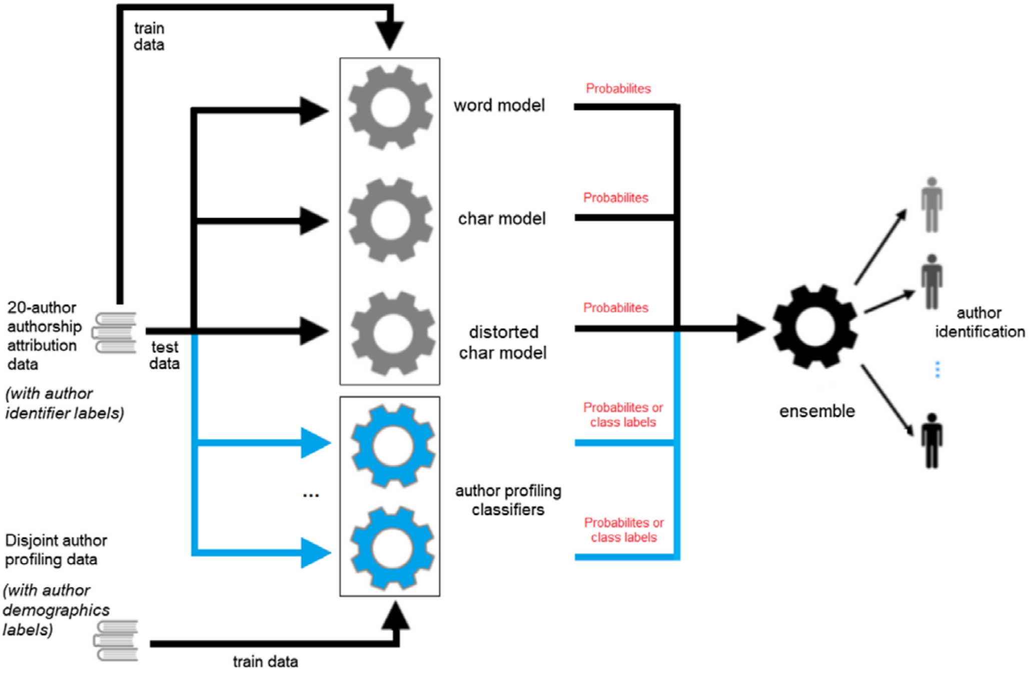
train data
word model
char model
distorted char model
Probabilites
Probabilites
Probabilites
20-author authorship attribution data
(with author identifier labels)
test data
Probabilites or class labels
Probabilites or class labels
author profiling classifiers
...
ensemble
author identification
Disjoint author profiling data
(with author demographics labels)
train data

Figure 4. Ensemble authorship attribution with added author profiling classifiers. For the main authorship attribution task, 20 input documents labelled with author ids only (left) are split into train and test data, and word, char and distorted char models (top middle) are built from this training data. From the reminder of the data and accompanying demographics labels (disjoint data, in the bottom left), the auxiliary author profiling classifiers are built. Test documents are submitted to each individual classifier independently, and their predictions are taken as the input to the final, second-level classifier (right).

feature selection along the lines of Hsieh, Dias, and Paraboni (2018) and others. In all models, logistic regression uses L2 regularisation and newton-cg solver with a 0.0001 tolerance as a stopping criteria. For reasons discussed below, depending on the authorship attribution strategy under consideration, author profiling predictions obtained by performing logistic regression may be taken either as class probabilities or as actual class labels.

Regarding the authorship attribution task proper, the present work essentially extends the EACH-USP stack authorship attribution approach in Custódio and Paraboni (2019) by adding author profiling classifiers to the existing ensemble of word, character and distorted character models as discussed in Section 2. In other words, the actual architecture is similar to the previous Figure 2, except that author demographics information will be presently inferred from text automatically with the aid of author profiling classifiers, rather than taken from ground truth corpus labels. This is illustrated in Figure 4 using two author profiling modules as an example (in light colour, at the bottom) and further discussed below.

Two strategies for adding author profiling predictions to the ensemble, hereby called *ap.prob* and *ap.label*, are presently considered. These strategies differ from each other only in the way their output predictions are represented. In *ap.prob*, we use author profiling predictions represented as probabilities not unlike the output of any of the existing components of the stack ensemble model. In *ap.label*, by contrast, we use class labels predictions (e.g., for gender, age etc.) In doing so, we would like to investigate the extent to which the present authorship attribution tasks may benefit from having access to more fine-grained probabilities or more coarse-grained class label predictions.

The use of author profiling probabilities from a given input document in *ap.prob* is illustrated as follows. Let us consider, for instance, a gender classifier that predicts that the author of the

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 6. Second-level classifier input represented as probabilities in ap.prob

Word	Char	Dist	male	female
0.78	0.82	0.24	0.43	0.57

Table 7. Second-level classifier input represented as class labels in ap.label

Word	Char	Dist	e1	e2	e3
0.78	0.82	0.24	0	1	0

document has a 0.43 probability of being male, and hence a 0.57 probability of being female. In a binary classification task of this kind, both probabilities are taken as an input to the second-level classifier (that is, in addition to the existing probabilities predicted by the original word, char and distorted char classifiers.) Similarly, for ternary author profiling classes (e.g., education level etc.), the three probabilities are considered.

An example of how the input to the second-level classifier is represented in *ap.prob* is illustrated in Table 6, in which probabilities provided by the Word, Char (character) and text distortion (Dist) modules of the EACH-USP ensemble for an individual candidate author are appended to his/her gender probabilities. This creates a set of five probabilities associated with each author, which are to be submitted to the second-level authorship attribution classifier.

Regarding the use of class label predictions in *ap.label*, author profiling probabilities are replaced by class labels directly or, more specifically, by assigning the value 1 to the class of highest probability and 0 to all the others. Thus, for instance, the 0.57 probability of being female in the previous example would be replaced by a 1 value, and the 0.43 probability of being male would be replaced by 0. An example of this representation using a ternary class (Education e1.e3) for an individual author is illustrated in Table 7, in which the class of highest probability is assumed to be e2.

5. Evaluation

This section describes the evaluation of our present approach. First, we discuss how the corpora described in the previous sections were organised into non-overlapping training and test sets for each author profiling and authorship attribution tasks. Next, we describe the evaluation procedure proper, and details of how each of the two tasks were optimised and tested. In the case of the authorship attribution task, statistical significance is to be assessed using the McNemar's test (McNemar 1947).

5.1 Train and test sets

Central to the current approach is the separation between data for our main task – authorship attribution – and for the auxiliary author profiling classifiers. Authorship attribution data consist of a set of train and test documents produced by 20 selected authors as discussed below, and it is labelled only with author identifiers. Author profiling data, by contrast, comprise all documents produced by other authors (i.e., those outside the 20-author group), and it is labelled with author demographics only (e.g., gender, age, etc.)

The organisation of each corpus into training and test sets takes into account the differences in granularity of the author profiling and authorship attribution tasks. The present author profiling

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 8. Demographics distribution in the 20-author test set for Gender (female/male), IT background (no/yes), Age, Education, Religiosity and Politics levels (low/medium/high)

Domain	Gender	IT backg.	Age	Education	Religiosity	Politics
Blog	9/11	–	2/9/9	–	–	–
Facebook	17/3	18/2	3/7 10	–	–	–
Essay	5/15	5/15	9/3/8	9/4/7	9/3/8	9/3/8
Twitter-Du	8/12	–	–	–	–	–
Twitter-Fr	7/13	–	–	–	–	–
Twitter-Ge	10/10	–	–	–	–	–
Twitter-It	4/16	–	–	–	–	–
Twitter-Pt	7/13	–	–	–	–	–
Twitter-Sp	9/11	–	–	–	–	–

models take as an input the set of all texts produced by an individual, which are concatenated as a single document labelled with their corresponding demographics. The authorship attribution models, by contrast, require multiple text samples from each candidate author (or else author identification would become trivial) and, as a result, take as an input individual text unities (i.e., Facebook and blog posts, sentences or tweets) as described in Section 4.1.

As the main focus of the present work is the authorship attribution task, we selected from each corpus the 20 authors with the largest volume of text available. These sets of 20 authors are taken to be the test data of our main authorship attribution approach in each domain and will be considered in a number of experiments in multiple test scenarios conveying from 2 to 20 candidates each. The choice for the authors with the largest possible amount of text data is intended to minimise situations in which the baseline approach in Custódio and Paraboni (2019) may fail due to lack of data, which would have obscured the role of the author profiling classifiers as an aid to the authorship attribution task. The issue of input size in authorship attribution is addressed in detail in Rocha *et al.* (2017). Test set author profiling class distributions are illustrated in Table 8. We notice, however, that these class labels are not taken into account by the present approach and are presented only as a means to illustrate how author profiling estimates may help authorship attribution.

Test sets selected from each domain are naturally more balanced towards some classes than others and, as the comparison among author profiling classifiers (e.g., gender, age, etc.) will require a fixed test set for each domain, class imbalance may impact the results of the present authorship attribution approach. For instance, we notice using a gender classifier is arguably less helpful if, for example, most test authors turn out to be of the same gender. Keeping balanced test sets for all author profiling classes is, however, impractical for a number of reasons. First, we notice that this would require a large number of distinct author profiles to cover all possible class values (e.g., the Essay domain would require a test set consisting of at least 324 distinct authors selected out of a corpus containing only 510 individuals). Moreover, many profiles are considerably rare, or simply do not occur at all in the data (e.g., there are relatively few individuals who belong to the more extreme classes of education, politics and religiosity; most IT people tend to be male, etc.)

In order to minimise these difficulties, in the evaluation of our authorship attribution approach we will keep the natural author profiling class imbalance as is, and we will resort instead to multiple random tests as discussed in Section 5.3. The issue of class imbalance will also be revisited in the light of our results as discussed in Section 6.3.

Finally, leaving the 20 test authors aside, the remaining portion of each corpus is concatenated (i.e., disregarding author identifiers) and taken as training data for the auxiliary author profiling

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 9. Train and test instances for the author profiling and authorship attribution tasks in the 20-author evaluation setting

	Author profiling		Authorship attribution	
Domain	*Train*	Test	*Train*	*Test*
Blog	19, 300	20	22, 856	9796
Facebook	713	20	5600	2400
Essay	490	20	706	303
Twitter-Du	77	20	35, 424	15, 182
Twitter-Fr	766	20	38, 992	16, 711
Twitter-Ge	208	20	38, 573	16, 531
Twitter-It	73	20	31, 441	13, 475
Twitter-Pt	39	20	30, 212	12, 948
Twitter-Sp	595	20	35, 844	15, 362

classifiers in each domain. Thus, the training data for author profiling does not include any text produced by the members of the 20-author group and, conversely, authorship attribution does not rely on the actual demographics information associated with the authors under identification, making instead its own predictions based on a disjoint data set.

The number of train and test documents (i.e., text units, cf. previous Table 5) for each task are summarised in Table 9, based on the largest possible (i.e., 20-author) evaluation setting.

5.2 Author profiling evaluation

Prior to the evaluation of the present authorship attribution approach, we built and evaluated its individual components, that is, the author profiling classifiers that could be built from the existing labels in each corpus. To this end, we performed univariate feature selection over development data using the ANOVA function and the F1 metrics to obtain the k-best text features (i.e., words) in each domain and language. Optimal values were searched within the 3000–20,000 features range at 1000 intervals and are summarised in Table 10.

Evaluation of the author profiling models was carried out by performing 10-fold cross validation over the training data and by considering a simple majority class baseline for illustration purposes. For all author profiling classifiers, we measure mean precision, recall, F1 and accuracy scores.

5.3 Authorship attribution evaluation

Multiple authorship attribution evaluation experiments were carried out by considering random sets of candidate authors drawn from the 20-author test set. With the exception of the Blog domain, tests were carried out by varying the number of candidate authors from 2 to 20. In the case of blogs, only tests involving 5, 10, 15 and 20 candidates were considered due to computational costs. As a means to obtain a balanced (authorship attribution) classification setting, the number of input texts taken from each candidate author is kept constant within each task by considering the smallest set size of the group.

In order to reduce the possible effects of random selection (e.g., in case of author profiling class imbalance when most authors turn out to belong to the same gender, etc.), evaluation was repeated

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 10. Author profiling univariate feature selection k-best values

Domain	*Gender*	IT backg.	*Age*	Education	Religiosity	Politics
Blog	17, 000	–	19, 000	–	–	–
Facebook	13, 000	6000	16, 000	–	–	–
Essay	5000	3000	9000	10,000	9000	7000
Twitter-Du	6000	–	––	–	–	–
Twitter-Fr	19, 000	–	––	–	–	–
Twitter-Ge	8000	–	––	–	–	–
Twitter-It	3000	–	––	–	–	–
Twitter-Pt	6000	–	––	–	–	–
Twitter-Sp	8000	–	––	–	–	–

20 times by varying the candidates randomly, and also by randomly selecting different train and test documents. More specifically, we performed 20 runs $*$ (3 corpora $*$ 19 non-blog candidate set sizes) $+$ 20 runs $*$ (1 corpus $*$ 4 blog candidate set sizes) experiments, making 1220 randomised authorship attribution evaluation tasks in total. Given the large number of evaluation scenarios, in what follows we will report overall mean results only.

The two versions of our current approach – *ap.prob* and *ap.label* – are to be compared against the standard EACH-USP ensemble baseline system in Custódio and Paraboni (2019) whilst measuring mean accuracy scores for all models. In doing so, our goal is to verify whether using author profiling classifiers improves results over the original approach that does not have access to author demographics predictions.

6. Results

Results of the experiments described in the previous section are presented in two parts. Section 6.1 reports results for the individual author profiling classifiers, and Section 6.2 presents results for the authorship attribution task proper.

6.1 Author profiling results

Table 11 presents author profiling results for the four domains under consideration (Facebook, Essay, Blogs and Twitter, respectively) as obtained from test data using our current classifiers and a majority class baseline system. Best macro F1 scores for each class are highlighted.

Although not the main focus of the present work, this admittedly simple analysis should suffice to illustrate that the present author profiling classifiers obtain results considerably above a majority class selector and that this approach may arguably help improve results in the actual authorship attribution task in the same way that using ground truth gender information improved results in the pilot study described in Section 3.

6.2 Authorship attribution results

This section presents mean accuracy results obtained by our present authorship attribution approaches *ap.prob* and *ap.label* and by the baseline system. For each individual domain and task, best results are highlighted.

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 11. Macro (A)ccuracy, (P)recision, (R)ecall and (F)-measure author profiling results

			Majority class				Author profiling			
Domain	Task	Classes	A	P	R	F	A	P	R	F
Blog	Age	3	0.10	0.03	0.33	0.06	0.75	0.73	0.81	**0.76**
	Gender	2	0.45	0.23	0.50	0.31	0.65	0.67	0.66	**0.65**
Facebook	Age	3	0.35	0.12	0.33	0.17	0.45	0.57	0.44	**0.42**
	Gender	2	0.85	0.43	0.50	0.47	0.80	0.60	0.61	**0.60**
	IT background	2	0.90	0.45	0.50	0.46	0.90	0.72	0.72	**0.72**
Essay	Age	3	0.15	0.05	0.33	0.09	0.55	0.52	0.42	**0.37**
	Gender	2	0.75	0.38	0.50	0.43	0.90	0.87	0.87	**0.87**
	IT background	2	0.75	0.38	0.50	0.43	0.85	0.81	0.77	**0.78**
	Politics	3	0.40	0.13	0.33	0.19	0.45	0.31	0.38	**0.32**
	Religiosity	3	0.30	0.10	0.33	0.15	0.55	0.63	0.56	**0.54**
	Education	3	0.45	0.15	0.33	0.21	0.60	0.44	0.48	**0.42**
Twitter	Gender (Du)	2	0.60	0.30	0.50	0.38	0.55	0.46	0.48	**0.43**
	Gender (Fr)	2	0.65	0.33	0.50	0.39	0.80	0.80	0.75	**0.76**
	Gender (Ge)	2	0.50	0.25	0.50	0.33	0.80	0.81	0.80	**0.80**
	Gender (It)	2	0.80	0.40	0.50	0.44	0.75	0.58	0.56	**0.57**
	Gender (Pt)	2	0.65	0.33	0.50	0.39	0.70	0.67	0.63	**0.64**
	Gender (Sp)	2	0.55	0.28	0.50	0.35	0.85	0.85	0.86	**0.85**

6.2.1 Blog domain

Table 12 presents mean accuracy scores for authorship attribution in the Blog domain as obtained by the EACH-USP baseline and by the *ap.prob* and *ap.label* models using age and gender classifiers. Best results for each candidate set (conveying 5, 10, 15 or 20 authors each) are highlighted. All differences between the baseline and the proposed models are significant ($p < 0.0001$).

Results for the Blog domain suggest that, on average, using author profiling classifiers is superior to the standard authorship attribution method by a narrow but significant margin. In particular, the use of (ternary) age labels as predicted by *ap.label* outperforms the baseline and, to a lesser extent, it is also superior to the use of age probabilities as predicted by *ap.prob*. The use of gender classifiers is still useful if compared to the baseline, but the advantage is small, and mean results obtained by both *ap.label* and *ap.prob* are similar.

6.2.2 Facebook domain

Table 13 presents mean accuracy scores for authorship attribution in the Facebook domain as obtained by the EACH-USP baseline and by the *ap.prob* and *ap.label* models using age, gender and IT background classifiers. Best results for each candidate set (conveying from 2 to 20 authors each) are highlighted. All differences between the baseline and the proposed models are significant ($p < 0.0001$).

The use of author profiling classifiers is consistently superior to the standard authorship attribution model alone and, in particular, using label predictions provided by the (ternary) age

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 12. Authorship attribution mean accuracy results for the Blog domain

		Age		Gender	
Authors	Baseline	ap.prob	ap.label	ap.prob	ap.label
5	0.77	0.81	**0.84**	0.81	0.82
10	0.64	0.62	**0.65**	0.63	0.63
15	0.54	0.55	**0.57**	0.55	0.56
20	0.49	0.52	**0.53**	0.52	0.52
Mean	0.61	0.63	**0.65**	0.63	0.63

Table 13. Authorship attribution mean accuracy results for the Facebook domain

		Age		Gender		IT background	
Authors	Baseline	ap.prob	ap.label	ap.prob	ap.label	ap.prob	ap.label
2	0.80	0.83	0.83	0.82	**0.84**	0.82	0.83
4	0.61	0.68	**0.76**	0.68	0.71	0.67	0.68
6	0.52	0.60	**0.70**	0.60	0.62	0.59	0.60
8	0.48	0.56	**0.65**	0.54	0.56	0.52	0.52
10	0.44	0.54	**0.61**	0.53	0.54	0.52	0.51
12	0.42	0.53	**0.57**	0.50	0.50	0.48	0.48
14	0.39	0.51	**0.55**	0.48	0.47	0.46	0.46
16	0.38	0.49	**0.54**	0.46	0.46	0.44	0.43
18	0.37	0.48	**0.51**	0.45	0.44	0.43	0.42
20	0.36	0.47	**0.50**	0.43	0.42	0.42	0.41
Mean	0.48	0.57	**0.62**	0.56	0.56	0.54	0.53

classifier is best of all. On the other hand, predictions made by the binary classifiers (gender and IT background) are less helpful, and the difference between *ap.prob* and *ap.label* is generally small.

6.2.3 Essay domain

Table 14 presents mean accuracy scores for authorship attribution in the Essay domain as obtained by the EACH-USP baseline and by the *ap.prob* and *ap.label* models (abbreviated to prob and label for ease of visualisation) using age, gender, IT background, political orientation, degree of religiosity and education level classifiers. Best results for each candidate set (conveying from 2 to 20 authors each) are highlighted. All differences between the baseline and the proposed models are significant ($p < 0.0001$) except for the comparison between the baseline and *ap.prob* using the religiosity classifier.

Results for the Essay domain suggest that all author profiling classifiers help author identification by a sizeable margin. The single most successful strategy in this domain is the use of degrees of religiosity as predicted by *ap.label*. All ternary classifiers (religiosity, age, politics and

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 14. Authorship attribution mean accuracy results for the Essay domain

		Age		Gender		IT backgr.		Politics		Religiosity		Education	
Authors	Baseline	prob	label	prob	label	prob	label	prob	label	prob	label	prob	label
2	0.78	0.82	0.85	0.75	0.76	0.75	0.77	0.76	0.88	0.78	**0.89**	0.85	0.83
4	0.56	0.59	0.62	0.59	0.67	0.60	0.65	0.56	0.71	0.61	**0.72**	0.60	0.68
6	0.48	0.47	0.53	0.48	0.55	0.51	0.56	0.46	0.58	0.51	**0.64**	0.45	0.58
8	0.38	0.37	0.43	0.39	0.49	0.37	0.47	0.41	0.46	0.42	**0.53**	0.38	0.47
10	0.38	0.40	**0.48**	0.39	0.45	0.36	0.47	0.36	0.43	0.37	**0.48**	0.39	0.43
12	0.33	0.34	0.41	0.36	0.41	0.35	0.39	0.33	0.39	0.37	**0.43**	0.34	0.41
14	0.30	0.34	0.38	0.33	0.37	0.31	0.38	0.34	0.39	0.32	**0.45**	0.31	0.38
16	0.26	0.29	0.38	0.30	0.35	0.28	0.35	0.28	0.33	0.30	**0.40**	0.28	0.37
18	0.26	0.28	0.33	0.29	0.32	0.27	0.35	0.29	0.34	0.29	**0.40**	0.26	0.32
20	0.26	0.28	0.32	0.28	0.31	0.29	0.31	0.27	0.34	0.28	**0.36**	0.28	0.32
Mean	0.40	0.42	0.47	0.42	0.47	0.41	0.47	0.41	0.49	0.42	**0.53**	0.41	0.48

education) outperform the binary classifiers for gender and IT background by a small margin, although binary classifiers are still significantly helpful if compared to the baseline. Moreover, we notice that the use of class labels as predicted by *ap.label* consistently outperforms both the baseline system and the use of probabilities in *ap.prob* in all scenarios.

6.2.4 Twitter domain

Authorship attribution results for the six Twitter data sets are divided into two tables for ease of visualisation, conveying three languages each. Table 15 summarises results for the Dutch, French and German corpora, and Table 16 concerns Italian, Portuguese and Spanish. In all cases, we report mean accuracy scores as obtained by the EACH-USP baseline and by the *ap.prob* and *ap.label* models using gender classifiers (recall that gender is the only kind of author demographics available from this domain, cf. Section 5.1.) Best results for each language and candidate set (conveying from 2 to 20 authors each) are highlighted. All differences between the baseline and the proposed models are significant ($p < 0.0001$).

Results for the Twitter domain suggest that, once again, using author profiling (gender) classifiers help authorship attribution. With the exception of the Dutch corpus (for which both gender labels and probabilities produced similar mean results), the use of gender probabilities as predicted by *ap.prob* is slightly superior to using gender labels as predicted by *ap.label*.

6.3 Discussion

Although results in the previous section vary considerably across domains, tasks and languages, the use of author profiling predictions was found to consistently improve mean accuracy in the authorship attribution task in all corpora and in all settings under consideration, comprising 1220 randomised evaluation scenarios in total. In other words, using author demographics predictions always leads to a certain improvement over the standard method. This, in our view, supports the main research hypothesis of the present study.

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 15. Authorship attribution mean accuracy results for the Twitter domain in Dutch, French and German corpora with gender profiling information

	Dutch			French			German		
Authors	Baseline	ap.prob	ap.label	Baseline	ap.prob	ap.label	Baseline	ap.prob	ap.label
2	0.89	0.90	**0.91**	0.91	**0.95**	**0.95**	0.88	**0.94**	0.93
4	0.78	0.81	**0.83**	0.81	**0.89**	0.86	0.78	**0.88**	0.84
6	0.71	**0.75**	**0.75**	0.75	**0.86**	0.80	0.71	**0.82**	0.79
8	0.67	**0.72**	**0.72**	0.71	**0.83**	0.77	0.68	**0.80**	0.76
10	0.64	**0.68**	**0.68**	0.68	**0.80**	0.73	0.63	**0.77**	0.72
12	0.61	**0.66**	0.65	0.65	**0.78**	0.71	0.61	**0.75**	0.70
14	0.59	**0.64**	**0.64**	0.63	**0.76**	0.69	0.59	**0.73**	0.68
16	0.57	**0.62**	**0.62**	0.61	**0.74**	0.66	0.58	**0.72**	0.67
18	0.55	**0.61**	**0.61**	0.59	**0.73**	0.65	0.56	**0.70**	0.65
20	0.53	**0.60**	0.59	0.58	**0.72**	0.64	0.54	**0.68**	0.63
Mean	0.65	**0.70**	**0.70**	0.69	**0.83**	0.78	0.66	**0.78**	0.74

Table 16. Authorship attribution mean accuracy results for the Twitter domain in Italian, Portuguese and Spanish corpora with gender profiling information

	Italian			Portuguese			Spanish		
Authors	Baseline	ap.prob	ap.label	Baseline	ap.prob	ap.label	Baseline	ap.prob	ap.label
2	0.87	**0.88**	**0.88**	0.92	**0.95**	**0.95**	0.91	0.96	**0.97**
4	0.76	**0.79**	0.78	0.82	**0.88**	0.87	0.82	**0.91**	**0.91**
6	0.69	**0.73**	0.72	0.77	**0.83**	**0.83**	0.76	**0.88**	0.85
8	0.65	**0.70**	0.68	0.72	**0.80**	0.78	0.71	**0.85**	0.81
10	0.61	**0.67**	0.64	0.68	**0.76**	0.75	0.67	**0.83**	0.77
12	0.59	**0.64**	0.62	0.65	**0.74**	0.72	0.65	**0.81**	0.75
14	0.56	**0.61**	0.59	0.63	**0.72**	0.70	0.63	**0.79**	0.72
16	0.53	**0.59**	0.56	0.60	**0.70**	0.67	0.61	**0.77**	0.70
18	0.51	**0.57**	0.54	0.58	**0.68**	0.65	0.59	**0.76**	0.68
20	0.50	**0.56**	0.53	0.56	**0.66**	0.63	0.57	**0.75**	0.67
Mean	0.63	**0.67**	0.65	0.69	**0.77**	0.76	0.69	**0.81**	0.75

Differences across experiments may have been influenced by multiple and possibly intertwined factors. First, there is the issue of author profiling accuracy. Some tasks (or some domains) may be simply more challenging than others, and this may explain, for instance, why the gain perceived by our method in the blog domain is consistently smaller than in the other text genres.

Second, as pointed out in Section 5.1, we notice that using author profiling predictions may be more helpful when the test set is balanced according to the predicted class. Related to this,

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 17. Authorship attribution mean accuracy results for the Blog domain using age and gender classifiers simultaneously

Classifiers	Baseline	ap.prob	ap.label
Individual classifiers:			
Age	0.61	0.63	0.65
Gender	0.61	0.63	0.63
Combinations:			
C1: Age + Gender	0.61	0.65	**0.75**

class distribution may also explain why binary classifiers were generally less helpful than ternary classifiers: from an author identification perspective, having a set of candidate authors split into three classes may be simply more effective than having the same set split into two classes.

Differences between *ap.prob* and *ap.label* are generally small, but using author profiling label predictions in *ap.label* is still superior to using probabilities in most scenarios. The main exception is the Twitter domain, in which there is a certain advantage for *ap.prob*.

6.4 Using multiple author profiling classifiers simultaneously

Given that the use of individual author profiling classifiers always increases overall accuracy in the authorship attribution task, we may ask whether using multiple classifiers simultaneously may do even better. To shed light on this issue, we carried out a series of complementary experiments in which every possible ensemble combination of classifiers is attempted in every domain except for Twitter, which supports one single (i.e., gender) author profiling class.

6.4.1 Blog domain

Results for the Blog domain are summarised in Table 17, in which individual classifier results (top rows) are reproduced from the previous sections for ease of comparison with the combined alternatives (bottom). The overall best alternative (C1) is highlighted.

From these results, we notice that using age and gender classifier simultaneously improves overall results, and particularly so in the case of the *ap.label* strategy.

6.4.2 Facebook domain

Results for the Facebook domain are summarised in Table 18, once again showing the comparison between individual classifier results (top rows) and those obtained by their combinations (bottom). The overall best alternative (C4) is highlighted.

Results from Table 18 suggest that using all three classifiers simultaneously (C4) is considerably superior to the use of any individual classifier alone (on the top rows), or their other possible combinations (C1.C3). However, this outcome should be interpreted carefully given that stacking an arbitrary large number of classifiers may easily lead to overfitting (Custódio and Paraboni 2021). We notice, for instance, that the small difference between C4 and C3 may suggest that adding the IT background classifier (which has the lowest accuracy among the three individual options) is not necessarily helpful in the present case.

6.4.3 Essay domain

Finally, results for the Essay domain are summarised in Table 19, once again showing both individual classifier results (top) and the combined alternatives (bottom). Given the large number of

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 18. Authorship attribution mean accuracy results for the Facebook domain using combinations of age, gender and IT background classifiers

Classifiers	Baseline	ap.prob	ap.label
Individual classifiers:			
Age	0.48	0.57	0.62
Gender	0.48	0.56	0.56
IT background	0.48	0.54	0.53
Combinations:			
C1: Gender + IT background	0.48	0.59	0.59
C2: Age + IT background	0.48	0.63	0.68
C3: Gender + Age	0.48	0.64	0.70
C4: Gender + Age + IT background	0.48	0.67	**0.72**

Table 19. Authorship attribution mean accuracy results for the Essay domain using combinations of age, gender, IT background, politics, religiosity and education classifiers

Classifiers	Baseline	ap.prob	ap.label
Individual classifiers:			
Age	0.40	0.42	0.47
Gender	0.40	0.42	0.47
IT background	0.40	0.41	0.47
Politics	0.40	0.41	0.49
Religiosity	0.40	0.42	0.53
Education	0.40	0.41	0.48
Combinations:			
C1: Gender + Religiosity	0.40	**0.95**	0.79
C2: Gender + Religiosity + Age	0.40	0.44	0.71
C3: Gender + Religiosity + IT background	0.40	0.43	0.68
C4: Gender + Religiosity + Politics	0.40	0.43	0.68
C5: Gender + Religiosity + Education	0.40	0.43	0.71

possible combinations, the present analysis is limited to the best-performing classification pair (which turns out to be Gender + Religiosity), to which we attempted to add a third classifier only. In other words, combinations of four classifiers or more are presently not addressed. The overall best alternative (C1) is highlighted.

For the Essay domain, the best results were obtained by using two classifiers only (i.e., the C1 combination), which was found to consistently outperform all single- and three-classifier alternatives alike, and often by a considerable margin. This advantage, which is particularly striking in the case of the *ap.prob* strategy, significantly deteriorates with the addition of a third classifier.

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Table 20. Authorship attribution mean accuracy results for the Essay domain based on random gender distribution versus and single- and balanced-gender candidate selection

Gender distribution >	Random		Single		Balanced	
Authors	ap.prob	ap.label	ap.prob	ap.label	ap.prob	ap.label
2	0.75	0.76	0.72	0.73	0.77	0.84
4	0.59	0.67	0.53	0.54	0.57	0.68
6	0.48	0.55	0.44	0.43	0.49	0.57
8	0.39	0.49	0.40	0.38	0.40	0.51
10	0.39	0.45	0.34	0.33	0.38	0.46
mean	0.52	0.58	0.49	0.48	0.52	0.61

This outcome, in our view, once again suggests that individual author profiling classifiers need to be added judiciously to the ensemble architecture for optimal results.

6.5 The role of author demographics distribution

The observation that author demographics information may be naturally unbalanced – as it is indeed the case in the present data sets – gives rise to the question of how the use of author profile classifiers may contribute to the overall authorship attribution task in scenarios of different demographics distribution among the candidate authors. For instance, given a candidate set in which all authors belong to the same gender group, we would expect gender classification to be of little or no help and presumably greater in a more gender-balanced set.

To illustrate the role of author demographics distribution in our current authorship attribution approach, we carried out a complimentary analysis in situations with and without class balance. However, due to the data sparsity and inherent class imbalance in our data, the discussion that follows is limited to the case of gender classification in the Essay domain.

Table 20 presents authorship attribution results for unbalanced (i.e., single-gender) and balanced (i.e., with the same number of male and female individuals) candidate sets. These results are shown alongside the previously observed results from random distributions, which are presently reproduced from Section 6.2.3 for ease of visualisation.

Single-gender results (middle columns in Table 20) from both *ap.prob* and *ap.label* models are on average inferior to those obtained in random gender distribution (left columns). By contrast, balanced-gender results (right columns) are equal or slightly superior to those observed in random gender distribution. This pattern may also be visualised in Figure 5.

7. Final remarks

The present study has addressed the authorship attribution task of digital texts by extending an existing stack ensemble model with author profiling classifiers. This approach has been evaluated in a range of text domains, author profiling tasks and languages and was found to be consistently superior to a standard authorship attribution method that does not have access to author demographics predictions even though author demographics information is naturally imbalanced, and classifiers of this kind are generally suboptimal.

Using author profiling classifiers to aid authorship attribution was found to be particularly useful in the case of ternary classes, which have the effect of splitting the set of candidate authors

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

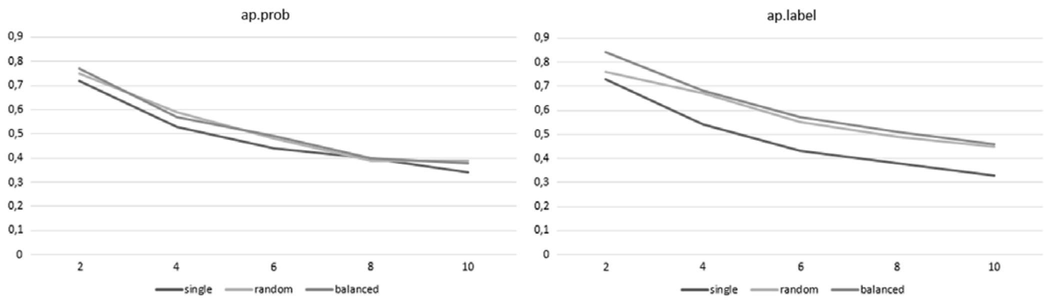

Figure 5. Authorship attribution accuracy based on different (single, random and balanced) gender class distributions for 2–10 candidate authors.

into smaller subsets, and hence facilitate author identification. Moreover, the method appears to suit the stack authorship attribution strategy even when the accuracy of the author profiling task is relatively low, and it was found to boost overall results even further when multiple classifier are considered simultaneously. More research is, however, required to establish which classifiers may be combined in this way and how to guarantee optimal results with no risk of overfitting.

The experiments that were carried out give rise to the question of how to explain the present gains over the standard authorship attribution model. In the present two-level architecture (i.e., consisting of author profiling and authorship attribution levels), however, an analysis of this kind would require investigating the possible interactions between domains, languages, profiling tasks, corpus sizes, number of instances, class imbalance, number of candidate authors and others. For that reason, in the present work, we chose to minimise some of these issues by performing multiple random tests, and we leave a more detailed analysis along these lines to future work, which should seek to pinpoint the exact circumstances under which using author profiling predictions may or may not be useful, and also compare the present approach with other authorship attribution methods.

Finally, yet another important limitation of the present work is that evaluation was focused on candidate sets conveying up to 20 authors in each domain. This limitation, which is inherited from the authorship attribution model taken as the basis for our current work (e.g., from PAN-CLEF shared tasks and others), should also be addressed in future work. However, given the strong correlation between input size and authorship attribution accuracy (Rocha *et al.* 2017), this will most likely require additional corpora with larger amounts of text samples per author.

Acknowledgement. The second author has received support from FAPESP grant 2016/14223-0.

Conflicts of interest. The authors declare none.

References

Bagnall D. (2016). Authorship clustering using multi-headed recurrent neural networks. In **Cappellato L., Ferro N., Macdonald C. and Balog K.** (eds), *CEUR Workshop Proceedings*, vol. 1609, Evora, Portugal. CEUR-WS.org, pp. 791–804.

Baker C.F., Fillmore C.J. and Lowe J.B. (1998). The Berkeley FrameNet project. In *COLING-1998*, Montréal, Quebec, Canada. Association for Computational Linguistics, pp. 86–90.

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.

Casavantes M., López R. and González L.C. (2019). UACh at MEX-A3T 2019: preliminary results on detecting aggressive tweets by adding author information via an unsupervised strategy. In *IberLEF@ SEPLN*, Bilbao, Spain. CEUR-WS.org, pp. 537–543.

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Chen X., Hao P., Chandramouli R. and Subbalakshmi K.P. (2011). Authorship similarity detection from email messages. In *Machine Learning and Data Mining in Pattern Recognition - 7th International Conference, MLDM*, New York, NY, USA. Berlin, Heidelberg: Springer, pp. 375–386.

Custódio J.E. and Paraboni I. (2019). An ensemble approach to cross-domain authorship attribution. In *International Conference of the Cross-Language Evaluation Forum for European Languages CLEF 2019*, Lecture Notes in Computer Science, vol. 11696, Lugano, Switzerland. Springer, pp. 201–212.

Custódio J.E. and Paraboni I. (2021). Stacked authorship attribution of digital texts. *Expert Systems with Applications* **176**, 114866.

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)*, Varna, Bulgaria. INCOMA Ltd., pp. 1069–1075.

dos Santos W.R., Ramos R.M.S. and Paraboni I. (2019). Computational personality recognition from facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia* 25(4), 268–287.

Garrido-Espinosa M.G., Rosales-Pérez A. and López-Monroy A.P. (2020). GRU with author profiling information to detect aggressiveness. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF)*, Malaga, Spain.

Granados A., Cebrián M., Camacho D. and de Borja Rodrguez F. (2011). Reducing the loss of information through annealing text distortion. *IEEE Transactions on Knowledge and Data Engineering* **23**(7), 1090–1102.

Hinh R., Shin S. and Taylor J. (2016). Using frame semantics in authorship attribution. In *IEEE International Conference on Systems, Man, and Cybernetics, SMC-2016*, Budapest, Hungary, pp. 4093–4098.

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)*, Miyazaki, Japan. ELRA, pp. 2566–2570.

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)*, Athens, Greece. IEEE Computer Society, pp. 54–60.

Jafariakinabad F. and Hua K.A. (2019). Style-aware neural model with application in authorship attribution. In *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 325–328.

Joulin A., Grave E., Bojanowski P. and Mikolov T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain. Association for Computational Linguistics, pp. 427–431.

Juola P. and Stamatatos E. (2013). Overview of the author identification task at PAN 2013. In *Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23–26, 2013.*

Kestemont M., Stamatatos E., Manjavacas E., Daelemans W., Potthast M. and Stein B. (2019). Overview of the cross-domain authorship attribution task at PAN 2019. In Cappellato L., Ferro N., Losada D. and Müller H. (eds), *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Kestemont M., Tschugnall M., Stamatatos E., Daelemans W., Specht G., Stein B. and Potthast M. (2018). Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In Cappellato L., Ferro N., Nie J.-Y. and Soulier L. (eds), *Working Notes Papers of the CLEF 2018 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org.

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*, Vancouver, Canada, pp. 471–477.

Koppel M. and Seidman S. (2018). Detecting pseudepigraphic texts using novel similarity measures. *Digital Scholarship in the Humanities* **33**(1), 72–81.

Markov I., Stamatatos E. and Sidorov G. (2017). Improving cross-topic authorship attribution: the role of pre-processing. In *18th International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, Hungary, pp. 289–302.

McNemar Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**(2), 153–157.

Misra K., Devarapalli H., Ringenberg T.R. and Rayz J.T. (2019). Authorship analysis of online predatory conversations using character level convolution neural networks. In *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 623–628.

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*. Association for Computational Linguistics, pp. 1950–1961.

Patchala J. and Bhatnagar R. (2018). Authorship attribution by consensus among multiple features. In *27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pp. 2766–2777.

Pavan M.C., dos Santos V.G., Lan A.G.J., ao Trevisan Martins J., dos Santos W.R., Deutsch C., da Costa P.B., Hsieh F.C. and Paraboni I. (2020). Morality classification in natural language text. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2020.3034050

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Peng J., Choo K.-K.R. and Ashman H. (2016). Astroturfing detection in social media: using binary n-gram analysis for authorship attribution. In 2016 IEEE Trustcom/BigDataSE/ISPA, pp. 121–128.

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

Pizarro J. (2019). Using N-grams to detect Bots on Twitter. In Cappellato L., Ferro N., Losada D. and Müller H. (eds), *CLEF 2019 Labs and Workshops, Notebook Papers*, Lugano, Switzerland. CEUR-WS.org.

Potthast M., Rangel F., Tschuggnall M., Stamatatos E., Rosso P. and Stein B. (2017). Overview of PAN 17: author identification, author profiling, and author obfuscation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2017*. Lecture Notes in Computer Science, vol. 10456. Springer, pp. 275–290.

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

Rangel F. and Rosso P. (2019). Overview of the 7th author profiling task at PAN 2019: bots and gender profiling. In Cappellato L., Ferro N., Losada D. and Müller H. (eds), *CLEF 2019 Labs and Workshops, Notebook Papers*, Lugano, Switzerland. CEUR-WS.org.

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 Cappellato L., Ferro N., Nie, J.-Y. and Soulier L. (eds), *Working Notes Papers of the CLEF 2018 Evaluation Labs*, CEUR Workshop Proceedings, Avignon, France. CLEF and CEUR-WS.org.

Rangel F., 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. CEUR-WS.org.

Rangel F., Rosso P., Zaghouani W. and Charfi A. (2020). Fine-grained analysis of language varieties and demographics. *Natural Language Engineering* **26**(6), 641–661.

Reddy P.B., Reddy T.R., Chand M.G. and Venkannababu A. (2018). A new approach for authorship attribution. In *Advances in Intelligent Systems and Computing*, vol. 701, pp. 1–9.

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

Rocha A., Scheirer W.J., Forstall C.W., Cavalcante T., Theophilo A., Shen B., Carvalho A.R.B. and Stamatatos E. (2017). Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security* **12**(1), 5–33.

Sánchez-Junquera J., nor Pineda L.V., y Gómez M.M., Rosso P. and Stamatatos E. (2020). Masking domain-specific information for cross-domain deception detection. *Pattern Recognition Letters* 135, 122–130.

Sari Y. and Stevenson M. (2016). Exploring word embeddings and character N-grams for author clustering notebook for PAN at CLEF 2016. In *CEUR Workshop Proceedings*, Evora, Portugal. CEUR-WS.org.

Sari Y., Stevenson M. and Vlachos A. (2018). Topic or style? exploring the most useful features for authorship attribution. In *27th International Conference on Computational Linguistics COLING-2018*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 343–353.

Schler J., Koppel M., Argamon S. and Pennebaker J. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, Menlo Park, California, USA. AAAI Press, pp. 199–205.

Schwartz R., Tsur O., Rappoport A. and Koppel M. (2013). Authorship attribution of micro-messages. In *Empirical Methods in Natural Language Processing*, Seattle, Washington, USA. Association for Computational Linguistics, pp. 1880–1891.

Sharon Belvisi N.M., Muhammad N. and Alonso-Fernandez F. (2020). Forensic authorship analysis of microblogging texts using n-grams and stylometric features. In *8th International Workshop on Biometrics and Forensics (IWBF)*, Porto, Portugal. IEEE, pp. 1–6.

Shrestha P., Sierra S., Gonzalez F., Rosso P., Montes-Y-Gomez M. and Solorio T. (2017). Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, Valencia, Spain. Association for Computational Linguistics (ACL), pp. 669–674.

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

Stamatatos E. (2017). Authorship attribution using text distortion. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL-2017)*, Valencia, Spain. Association for Computational Linguistics.

Stevenson M., Vlachos A. and Sari Y. (2017). Continuous n-gram representations for authorship attribution. In *15th Conference of the European Chapter of the Association for Computational Linguistics EACL-2017*, Valencia, Spain, pp. 267–273.

Sundararajan K. and Woodard D.L. (2018). What constitutes style in authorship attribution? In *27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 2814–2822.

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. CEUR-WS.org.

https://doi.org/10.1017/S1351324921000383 Published online by Cambridge University Press

Vartapetiance A. and Gillam L. (2012). Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*, Rome, Italy. CEUR-WS.org.

Verhoeven B., **Daelemans W. and Plank B.** (2016). TwiSty: a multilingual Twitter Stylometry corpus for gender and personality profiling. In *10th International Conference on Language Resources and Evaluation (LREC-2016)*, Portoroz, Slovenia. ELRA, pp. 1632–1637.

Wolpert D.H. (1992). Stacked generalization. *Neural Networks* **5**(2), 241–259.

Cite this article: Deutsch C and Paraboni I (2023). Authorship attribution using author profiling classifiers. *Natural Language Engineering* **29**, 110–137. https://doi.org/10.1017/S1351324921000383