What Can Go Wrong in Authorship Profiling: Cross-Domain Analysis of Gender and Age Prediction

Hongyu Chen[1] and **Michael Roth**[2] and **Agnieszka Falenska**[1,2]
[1]University of Stuttgart, Interchange Forum for Reflecting on Intelligent Systems
[2]University of Stuttgart, Institute for Natural Language Processing
{hongyu.chen, agnieszka.falenska}@iris.uni-stuttgart.de
michael.roth@ims.uni-stuttgart.de

Abstract

Authorship Profiling (AP) aims to predict the demographic attributes (such as gender and age) of authors based on their writing styles. Ever-improving models mean that this task is gaining interest and application possibilities. However, with greater use also comes the risk that authors are misclassified more frequently, and it remains unclear to what extent the better models can capture the bias and who is affected by the models' mistakes. In this paper, we investigate three established datasets for AP as well as classical and neural classifiers for this task. Our analyses show that it is often possible to predict the demographic information of the authors based on textual features. However, some features learned by the models are specific to datasets. Moreover, models are prone to errors based on stereotypes associated with topical bias.

1 Introduction

Authorship Profiling (AP) aims to identify authors' demographic characteristics through their writing style. In recent years, this task has polarized the NLP community. On the one side, researchers emphasize the potential of AP for computational social science applications, where predicting who wrote given texts can enrich analyses of data that lacks explicit demographic information (Morales Sánchez et al., 2022; Deutsch and Paraboni, 2023). Such additional automatically predicted attributes could allow for uncovering demographic patterns in societal trends, political ideologies, or cultural shifts. The automatic prediction of such attributes may also be helpful to other practical applications, such as forensics, abuse detection, and marketing (Mukhopadhyay et al., 2021; Bugueño and Mendoza, 2020; Mishra et al., 2018; Abdul-Mageed et al., 2019). As a result, the majority of work on AP is motivated by these practical applications and focuses primarily on improving model performance (Cheng et al., 2009; Pardo and Rosso, 2016; Soler-Company and Wanner, 2018; Fabien et al., 2020, among others).

On the other side, researchers are alarmed by the potential societal harm that AP models can cause. Firstly, these tools come with the risk of privacy breaches and the dangers of using authors' features without their consent (Emmery et al., 2022; Larson, 2017). Secondly, the AP tasks and datasets commonly understate complexity of how demographic characteristics relate to the language production. For example, gender, one of the most frequently predicted demographic traits, is often analyzed in isolation from other related features like age (HaCohen-Kerner, 2022) and oversimplified (Koolen and van Cranenburgh, 2017). AP models traditionally treat gender as a binary variable and lack reflection on the spectrum of gender identities, potentially leading to reinforcing stereotypes and misrepresentations (Dev et al., 2021). Finally, misclassifying people can lead to feelings of exclusion, negatively affecting individuals' self-esteem and confidence (Fosch-Villaronga et al., 2021).

To move forward, it is essential to reach a consensus regarding the circumstances necessitating the deployment of AP models. Fundamental to this process is a thorough understanding of what these models learn, what type of biases they capture, and who is affected by their errors. To this end, this paper examines the core assumption underlying the majority of research motivated by the practical applications of AP: that demographically related signals are *comparable across datasets*. With a focus on gender and age – two demographic features that are strongly interrelated – we explore the extent to which writing styles are consistent and transferable across datasets. Our work centers around three core research questions:

1. *What is the accuracy of standard classifiers for gender and age prediction, and to what extent does*

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 150–166

it change in cross-domain applications?

We train classical and neural classifiers on two well-established datasets from two domains: online conversations and blog posts, and two languages: English and Spanish. Our findings indicate that neural classifiers have only a modest advantage when predicting gender and age. Moreover, the performance of all classifiers drops close to the majority baseline in cross-domain applications (§5).

2. *Are the writing styles of authors consistent across datasets and languages?*

We perform a statistical analysis of authors' writing styles to uncover that gender and age differences found in one dataset are not fully reproducible within another. The finding is consistent across domains as well as languages (§6).

3. *How do topics affect AP performance?*

Finally, we ask what information the AP models capture. We find that while topical signs alone are inadequate for effectively modeling demographic features, they influence models' behavior: misclassifications appear commonly in topics predominantly addressed by one demographic group (§7).

The contributions of our paper are twofold. Firstly, we provide methodological insights into AP classifiers, challenging the practical usefulness of these tools, especially in cross-dataset settings. Secondly, we add empirical evidence to the discussion on the need to take the AP results with caution. Otherwise, the potential risks of marginalizing and misrepresenting certain demographic groups are disregarded, leading to biases and discrimination (Zuiderveen Borgesius et al., 2018).

2 Bias Statement

Our work examines the behavior of AP models, focusing on how models predict gender and age across domains. We specifically define *gender bias* as a notable difference in prediction of an author's gender based on topic preference or writing style. In parts, such differences can be explained by the underlying training data used by AP models. For example, Bamman et al. (2014) observe that male authors tend to use named entities at a higher rate in their writing compared to female authors, a phenomenon also related to topic choices that are rich in named entities, such as specific hobbies or career paths. Despite such insights, approaches to author profiling sometimes rely only on style-based features and overlook topical differences. We examine the impact regarding gender bias by testing how likely AP models mispredict gender when authors write about topics typically associated with another gender. For example, male authors discussing shopping-related topics may be mispredicted as female. This indicates that the model picks up on topics stereotypically associated with one gender and performs inadequately when authors from another gender engage with those same topics, which may cause representational harms (Blodgett et al., 2020). Moreover, biases of AP models can easily be misinterpreted as general differences in gender or age, leading to an issue of reinforcing stereotypes.

Our work is grounded in the belief that uncovering biases is crucial for developing equitable NLP applications. We acknowledge as a main limitation that all data used in this work assumes a binary gender framework. Therefore, our analyses may not fully capture the complexities and nuances of gender identity. Future work should aim to include more inclusive and representative data to better understand and address gender bias in AP models.

3 Related Work

Previous work can be roughly divided into three categories: work on the task of authorship profiling itself (§3.1), sociolinguistic studies of stereotypes and gender differences (§3.2) as well as efforts to model or counteract (topical) biases (§3.3).

3.1 Authorship Profiling

The earliest automated AP task was performed on a subset of British National Corpus (BNC) using a combination of function words and n-grams of POS tags as features (Koppel et al., 2002) . Later work focused on English blog posts, where gender prediction was addressed with improved feature selection and machine learning methods (Mukherjee and Liu, 2010, inter alia). According to a recent survey, accuracy for gender prediction varies across publications from 52% to 91% (HaCohen-Kerner, 2022). Authors suggest that this large variance might be caused by different factors, including text genres, age groups, and types of applied classifiers. For example, Ceccucci et al. (2013) find that female authors compose longer text messages, but this finding does not seem to generalize to blog posts. Regarding literary texts, a recent finding by Lettieri et al. (2023) suggests that women tend to employ more positive words than men, also imply-

ing a correlation between sentiment and the authors' gender. In general, however, there is little consistency regarding high-accuracy phenomena for gender prediction, suggesting that differences in terms of online writing could largely be dependent on the respective datasets. For instance, Alvarez-Carmona et al. (2015) achieve accuracy of 91% using liblinear SVM on PAN15 datasets, but the number drops to 81.72% on a Twitter dataset (Pizarro, 2019). Therefore, by identifying which features and models do (not) generalize across datasets, we address a major gap in existing research.

3.2 Sociolinguistic Analyses and Gender

AP builds directly on the stylometry, sociolinguistics, and theoretical issues in demographic differences in writing (Koolen and van Cranenburgh, 2017; Xia, 2013). Empirically, gender has been shown to be a main characteristic for categorization (Rudman and Glick, 2021) and linguistic differences have been observed across various datasets and domains (Leech et al., 1992; Baker, 2014; Argamon et al., 2003, 2007), ranging from scientific articles (Bergsma et al., 2012), political discussions (Hu and Kearney, 2021), to contemporary fiction (Dahllöf, 2023). Though prominent, these differences cannot be simply attributed to gender alone, as the contexts in which people communicate often limit their language use (Baker, 2014). Cameron (1997) critiques the traditional view of gender as a fixed characteristic that explains behaviors. She advocates for understanding gender as something that needs to be explained in its own right, suggesting that gender is constructed, performed, and enacted in social contexts rather than being a natural, unchangeable attribute that determines how individuals act. However, this does not mean to deny the existence of gender differences, but rather to provide more insights on proceeding with such types of differences related to languages with more caution (Koolen and van Cranenburgh, 2017; Liu et al., 2021). Because what comes along with such differences is the issue of oversimplification and stereotyping (Bing, Janet and Bergvall, 1998). For AP models, the interpretation of the correlations between demographic groups and style/content features is beneficial for researchers to learn the potential pattern that a model might learn. One should however be careful to avoid over-generalization.

3.3 Topical Bias

Topical bias is another contextual factor that affects profiling demographic differences in writing. Works in authorship attribution and authorship verification have pointed out that topical preference will lead to errors when the topics shifts (Hu et al., 2023). Similarly in AP, it was demonstrated that the choice of topics by female and male authors can exhibit significant differences (Verhoeven et al., 2017). For instance, women tend to gravitate toward themes of relationships and connections, while men tend to focus on topics related to politics and hierarchy (Bischoping, 1993). Measures proposed to mitigate the effects of topics include topic-independent features (Litvinova et al., 2018), topic-debiased representations (Hu et al., 2023), and explicitly considering errors made by authorship attribution models regarding topics (Altakrori et al., 2021). Though this work does not focus on topic debiasing, we also include an analysis on interactions between topics and demographic predictions by AP models.

4 Data

We use two well-explored datasets of texts annotated with self-reported gender and age of their authors – PAN13 and BLOG. We select datasets that are fundamental to the AP research field: PAN13, used in the first PAN-AP shared task, and BLOG, used in the earliest work for studying gender effect on texts).

PAN13 originates from a shared task on plagiarism detection, authorship verification, and authorship identification (Rangel et al., 2013). It includes conversations in two languages: English (referred to as **PAN13-EN**), comprising a total of 283,240 conversations and Spanish (**PAN13-ES**), with 90,860 conversations. The dataset includes a variety of topics to reflect real-world usage and complexity, with an emphasis on everyday language in social media.

In the data preparation step, we exclude posts from authors who pretend to be minors.[1] Both English and Spanish datasets come with the training and test split. To ensure a comparable analysis across languages, we downsample the training part of the English dataset so that it has the same number of samples as Spanish (we do not alter test parts). Table 5 in Appendix A gives data statistics.

[1] Information comes from the names of the files.

	Gender			Age		
	PAN13-EN	PAN13-ES	BLOG	PAN13-EN	PAN13-ES	BLOG
Majority	0.50	0.50	0.50	0.36	0.56	0.33
LR	**0.60**	**0.67**	**0.76**	**0.56**	**0.67**	**0.69**
DT	0.54	0.56	0.63	0.52	0.55	0.52
RF	0.58	0.59	0.65	0.52	0.61	0.56
NB	0.53	0.55	0.61	0.43	0.44	0.53
BERT	0.61	0.72	0.76	0.59	**0.68**	0.67
RoBERTa	**0.64**	**0.71**	**0.79**	**0.65**	0.67	**0.76**
XLNet	0.60	–	0.77	0.64	–	0.72

Table 1: Accuracy for gender and age prediction on test data (averages from six models trained with different random seeds, standard deviation in Appendix A, Table 6). Best white- and black-box classifiers are bolded.

BLOG is the Blog Authorship Corpus (Schler et al., 2006), that was constructed in August 2004 from `blogger.com`, including a total of 71,000 blogs and 681,284 posts. Each post is annotated with a date, blogger's ID, self-provided gender ('female', 'male'), age, industry, and zodiac sign.

The corpus does not include a pre-defined training and test split. Therefore, we first randomly divide it into 80/20 split. Secondly, since BLOG includes whole articles and not single conversation inputs, its data points are much longer than in PAN13. Therefore, to make these two datasets more comparable, we downsample the training part of BLOG to have approximately the same number of words as in PAN13-ES (keeping full articles intact). We ensure that all the datasets are balanced regarding the gender of the authors. We convert the ages in BLOG to the same categories as in PAN13: '10s' (13-17), '20s' (23-27), and '30s' (33-47).

For both of the datasets, we group and concatenate posts from the same author and take such concatenated texts as our data points. Moreover, we eliminate URLs in the preprocessing of the texts.

5 Gender and Age Prediction

We start from answering what is the accuracy of standard classifiers when predicting gender and age for the given text.

5.1 Method

We test classifiers that are straightforward to implement, making them popular choices for predicting the demographics of authors. We categorize these classifiers into two groups: white-box and black-box models (Loyola-González, 2019). White-box models, like logistic regression, offer easy-to-understand interpretations of results, appealing to researchers who prioritize insight into the decision-making process of their models (Morales Sánchez et al., 2022; Rudin, 2019). On the other hand, black-box models, typically including neural networks, are often regarded as more effective but harder to interpret.

White-box We follow the white- and black-box classifier selection outlined in Jang et al. (2023). Concretely, for white-box classifiers, we use Logistic Regression (**LR**), Random Forest (**RF**), Decision Tree (**DT**), and Naive Bayes (**NB**). We implement them using the scikit-learn library (Pedregosa et al., 2011) with default hyperparameters.[2] For these models, each text is represented as a vector of (lower-cased) word-based tf-idf scores.

Black-box The black-box classifiers use the transformer-based language models supported by the Hugging Face Transformers library (we refer to Table 3 in Appendix A for details).[3] For English, we experiment with **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019), and **XLNet** (Yang et al., 2019).[4] For Spanish, we use the BERT adaptation by Cañete et al. (2020) and RoBERTa adaptation by De la Rosa et al. (2022). All classifiers underwent fine-tuning for a duration of five epochs, employing a learning rate of 1e-5, a weight decay factor of 0.01, and a batch size of 16.

5.2 In-Domain Results

Table 1 provides gender and age prediction results from the white-box (top) and black-box (bottom) models. First of all, it is evident that almost all classifiers outperformed the majority baseline, with

[2]Code and data selection will be released with this paper.

[3]`https://huggingface.co/`

[4]We do not include BertAA (Fabien et al., 2020), i.e., BERT model for authorship attribution, because it was fine-tuned on BLOG, which we did use for our analysis.

	Gender		Age	
	PAN13	BLOG	PAN13	BLOG
Majority	0.50	0.50	0.33	**0.42**
LR	**0.51**	**0.57**	0.39	0.40
DT	0.46	0.50	0.44	0.31
RF	0.51	0.50	0.47	0.34
NB	0.50	0.50	0.42	0.35
BERT	0.53	0.65	**0.50**	0.38
RoBERTa	**0.55**	**0.69**	0.37	0.38
XLNet	**0.55**	0.68	0.40	0.40

Table 2: Cross-dataset accuracy (averages from six models trained with different random seeds, standard deviation in Appendix A, Table 4) for gender and age prediction on English test datasets. Best white- and black-box classifiers are bolded.

the only exception of the DT model for age prediction in the PAN13-ES dataset.[5] Among the white-box classifiers, LR stands out as the best, consistently surpassing the other models by a significant margin. In the black-box category, RoBERTa achieves the best results.[6] However, the advantage that RoBERTa holds over LR is relatively modest (at most 0.09 for age prediction in PAN13-EN), prompting the question if the loss of inherent interpretability coming with LR is worth the slight performance gain.

When analyzing the performance across all three datasets, an interesting pattern can be noticed – accuracy on PAN13-EN is uniformly lower than on its Spanish counterpart, PAN13-ES. Similarly, the accuracy on PAN13-ES is consistently lower than on the BLOG dataset, positioning BLOG as the "easiest" dataset for the classifiers to handle. The differences in performance across languages (PAN13-EN and PAN13-ES) and domains (PAN13-EN vs. BLOG) are substantial, raising questions about the underlying factors behind them.

5.3 Cross-Domain Results

Before we go to the analysis of style differences, we conduct a preliminary cross-dataset[7] experiment, in which we train a model on PAN13-EN and test it on BLOG and vice versa. The outcomes are presented in Table 2. As expected, the accuracy of all the models decreased compared to the in-domain results (cf. Table 1). Regarding gender prediction, certain trends observed previously persist: LR and RoBERTa remain the best classifiers, with RoBERTa keeping its advantage over LR. Moreover, accuracy on BLOG is still higher than on PAN13-EN. However, the practical usability of any of these classifiers is debatable. Although nearly all models outperformed the majority baseline, this improvement is often minuscule. The most promising results come from black-box classifiers applied to BLOG, suggesting that some gender-related signals effectively transfer from PAN13-EN. This could be due to the larger training datasets improving cross-domain gender prediction accuracy — as seen with PAN13-EN compared to BLOG— although previous evidence suggests this is not always the case (Dias and Paraboni, 2020).

For age prediction, a slightly different picture can be observed. Apart from three exceptions (DT, RF, and BERT applied to PAN13-EN), none of the classifiers exceeded the majority baseline. Notably, in the context of BLOG, no model successfully transferred age-related features from PAN13-EN. These findings underscore a critical point: while certain stylistic elements do vary across datasets, the features that remain consistent are insufficient to enable classifiers to effectively generalize.

6 Demographics vs. Style

The findings from the previous section highlight differences in classifier performance across datasets. Next, our objective is to uncover the underlying causes behind these differences. Given that the accuracy of AP models is often linked to the demographically influenced writing styles of authors (see Section 3.2), our first analysis involves examining our datasets from the perspective of their style.

6.1 Method

We investigate the explanatory power of style-related variables in predicting demographic characteristics. Our analysis focuses on two dependent variables (DVs): binary gender (male/female) and age (10s, 20s, 30s). The independent variables (IVs) include word-based features derived from the datasets (see below). Additionally, we incorporate the other demographic feature as an IV – for instance, when gender serves as the DV, age is included as an IV, and vice versa. The analysis is conducted using only the training parts of datasets.

[5]The variation in the majority baseline results comes from the datasets being balanced with respect to gender but not age.

[6]Our best white- and black-box classifiers align with the findings of Jang et al. (2023), who evaluated the same models for figurative language recognition.

[7]We consider PAN13 and BLOG are datasets from two different domains: PAN13 includes conversational posts from social media and BLOG includes individual blog posts of longer length

Feature extraction We extract word-level features from five categories. For the description of all single features in each category, we refer to Falk and Lapesa (2022) (see Appendix in Falk and Lapesa (2022) for details)

- Surface (6 features) including characteristics like token length, average character count per word, and number of syllables per word.
- Syntactic (6 features) involving metrics such as the proportion of fine-grained part-of-speech tags within each post, including personal pronouns, auxiliaries, and named entities.
- Textual complexity (14 features) encompassing diverse measures of lexical diversity, lexical sophistication, and readability.
- Sentiment and polarity (20 features) including emotional indicators like joy and fear.

For the extraction of surface and syntactic features, we used scripts from Falk and Lapesa (2022). For textual complexity and sentiment features, we employed SEANCE (Crossley et al., 2017), TAALED (Kyle et al., 2021), and TAALES (Kyle et al., 2018). Given that these tools are designed only for English, their application was limited to the PAN13-EN and BLOG datasets. As a result, we extracted a total of 46 features[8] for these two datasets and 12 features for PAN13-ES.

We use `stats` and `nnet` packages from R and two types of models: a binomial logistic regression for gender as DV and a multinomial logistic regression for age as DV. To compare across English datasets (PAN13-EN and BLOG), we load the model with all 46 features. Comparison across languages (PAN13-EN, BLOG and PAN13-ES) is performed with 12 common features. Additionally, Appendix A.1 provides details on the best combination of features for each dataset.

6.2 Data Analysis

Figures 1 and 2 present significant features that correlate with gender as the controlled variable. We focus our discussion on the results from the models on gender prediction as a case of our methodology. We obtained similar findings in terms of significant features for age, which are shown in Figures 5 and 6 in Appendix A.

Style across datasets Comparing the two English datasets in Figure 1, we first notice that the two plots clearly differ. While for PAN13-EN, significant features (dark markers) can be seen across all categories, for BLOG, they group mostly in the bottom three. Analyzing individual categories, differences can be spotted already in the surface features (second from the bottom). For example, the percentage of syllables per word (`syll_per_word`), Gunning fog index (`gunningFog`), and Flesch reading ease (`flesch`)[9] indicate significant correlations with the 'male' category, similar to previous findings about "women tend[ing] to compose longer texts than men" (Xia, 2013). However, this significance is observed only in the BLOG dataset. Regarding syntactic features, only the use of auxiliaries and the presence of named entities emerge as significant factors across both datasets. In contrast, the frequency of subordinate conjunctions appears only in BLOG and adjectives only in PAN13-EN. Finally, we find that PAN13-EN contains a greater number of significant features from the categories of sentiment, and text complexity, compared to BLOG. For features such as the "certainty component", our finding of it correlating more with the 'female' category, is aligned with previous evidence that women tend to have more positive sentiment in texts than men do (Lettieri et al., 2023).

Style across languages As shown in Figure 2, surface and syntactic features emerge as distinctive attributes associated with gender in both English and Spanish PAN13 datasets. However, a closer inspection reveals nuanced variations in the contributions of these features between the two languages. For instance, in PAN13-EN, female authors tend to use more adjectives, whereas in PAN13-ES, this trend is reversed, correlating more with male authors. Other syntactic features, such as auxiliaries, are significantly correlated with male authors in PAN13-EN but exhibit no discernible effect on either female or male authors in PAN13-ES. Similarly, adverbs in PAN13-ES are highly associated with female authors, while there is no such association in PAN13-EN.

[8] Due to the limited capacity of TAALED processing long posts, there are 4 features we did not manage to extract for both datasets. These features are: McD_CD_AW, Sem_D_AW, content_poly and hyper_verb_noun_Sav_Pav from the textual complexity category (lexical sophistication).

[9] Flesch Reading Ease (Flesch, 1948) and Gunning Fog Index are two readability metrics measuring a combination of information involving the length of the sentences or words, and the number of complex words. Unlike lexical diversity and sophistication features relying on the variants of token ratio, these scores are sensitive to the length of texts, thus they are classified as surface features.

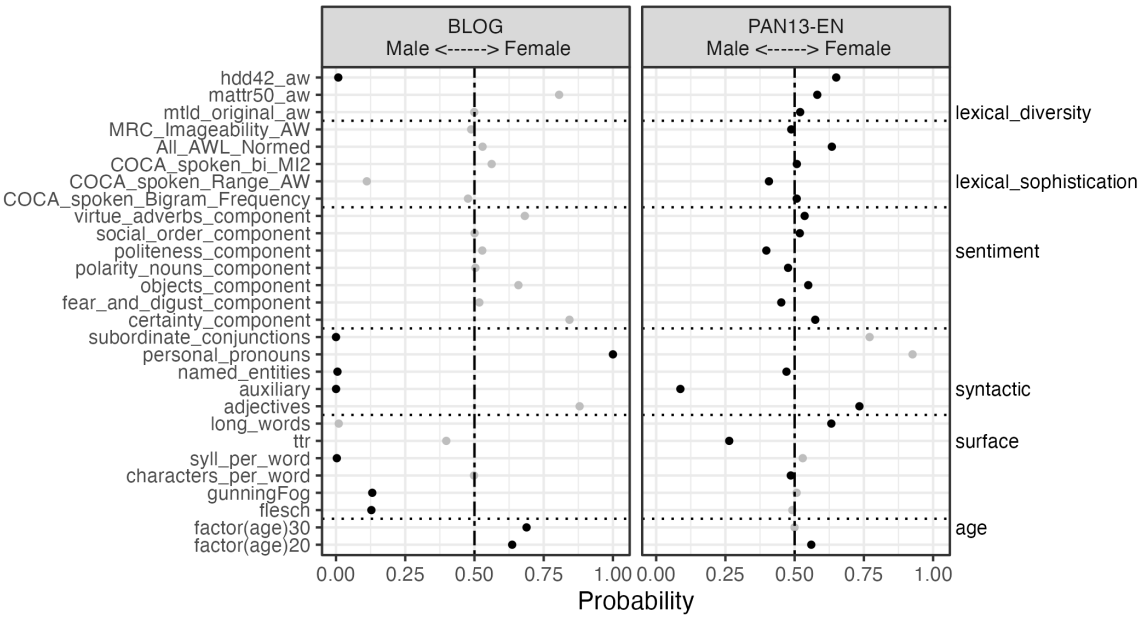


Figure 1: Significant ($p < 0.05$) features for gender as DV; model selection used all five categories of features. Labels on the left are feature names; right are group names. Light gray markers show non-significant features.

In summary, we conclude that gender-related style signals are inconsistent across our selected domains and languages.

7 Demographics vs. Topics

As explained in Section 3.3, topics are the second type of information frequently considered to influence classifiers' ability to predict demographic features of authors. Thus, in this section, we analyze topic-based differences in our data and their influence on the classifiers' errors. This section focuses exclusively on the PAN13-EN dataset, which we identified as the most challenging for the classifiers in Section 5.

7.1 Method

To extract topics, we use BERTopic (Grootendorst, 2022) with the default parameters. Specifically, we assign one topic to each post in the training data. To ensure coherent content in topics, we constrain the topic number to 100, covering 75,895 posts. All the remaining texts were assigned the default topic -1, which BERTopic designates for outliers. We exclude these texts from the analysis.

7.2 Data Analysis

Figure 3a shows the five most common topics for different demographic groups (male vs. female, 20s vs. 30s) in the PAN13-EN dataset as well as the corresponding numbers of articles.[10] The topic of website and marketing (`label 0`) emerges as the most commonly addressed across all groups. The second ranked topic concerns shoes and handbags (`label 1`) for all groups except for males in their 20s, for whom love and god (`label 2`) is ranked second. Apart from order, the top-5 topics within each age group are the same across gender.

Larger differences can be observed across age groups: While `labels 2` and `4` appear only in the top-5 topics for authors in their 20s, the topics home and furniture (`label 3`) and weight and fat (`label 6`) are in the top-5 only for authors in their 30s. In other words, male and female authors show a relatively strong interest in fashion, love, religion and/or friends in their 20s. However, interests differ across age groups, with other interests being more important in the 30s, independently of gender. Our data is not longitudinal, meaning that while we can identify the topic difference across age populations, we are unable to track the evolving interests of specific individuals over time.

7.3 Error Analysis

Having determined the distribution of topics, we investigate their influence on AP classifiers. Concretely, we perform an error analysis of the RoBERTA classifier, the best-performing model from Section 5. To not compromise our test sets,

[10] We exclude the 10s group for data sparsity reasons.

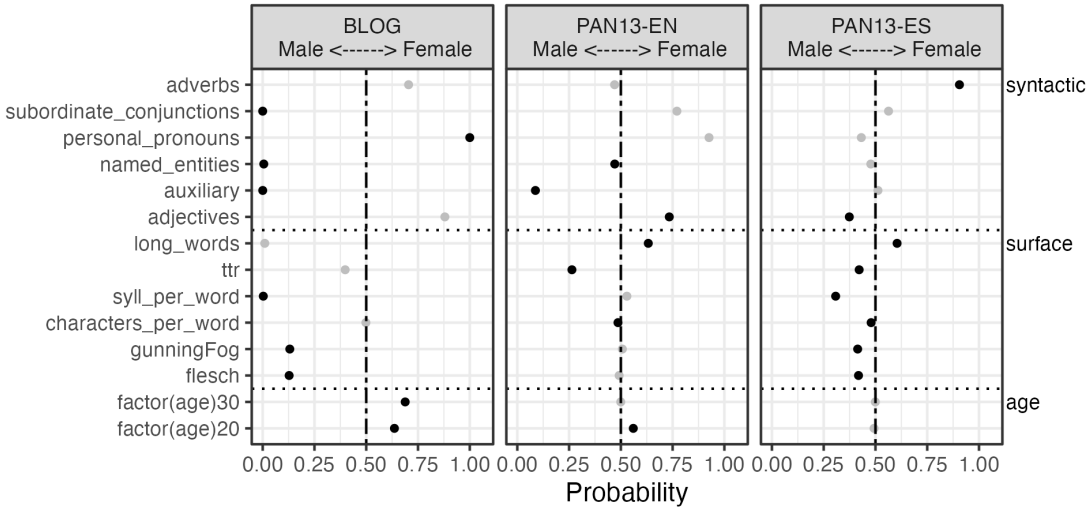


Figure 2: Significant ($p < 0.05$) features for gender as DV; model selection used only surface and syntactic features. Labels on the left are feature names; right are group names. Light gray markers show non-significant features.

we perform this analysis on the training sets, for which we collect model predictions on gender and age by 5-fold jackknifing. Figure 3b shows absolute and relative errors counts for the discussed groups and topics (for a full breakdown of results by gender and age, see Appendix A, Table 8).

As expected, the most frequent topics in the whole dataset—websites and marketing (`label 0`), shoes and handbags (`label 1`), and love, life, and Jesus (`label 2`)—are also the ones with the highest error counts. Interestingly, a clear pattern can be observed when comparing the distribution of topics against the prediction errors. Topics that are more frequent for one gender, such as shoes and handbags (`label 1`) and home and furniture (`label 3`) for females, or greetings and friendship (`label 4`) for males, tend to be underpredicted. This result can be interpreted as a potential stereotype in the model: Men writing about shoes, handbags, or furnishings will be more frequently mispredicted as women, while women writing about games as men.

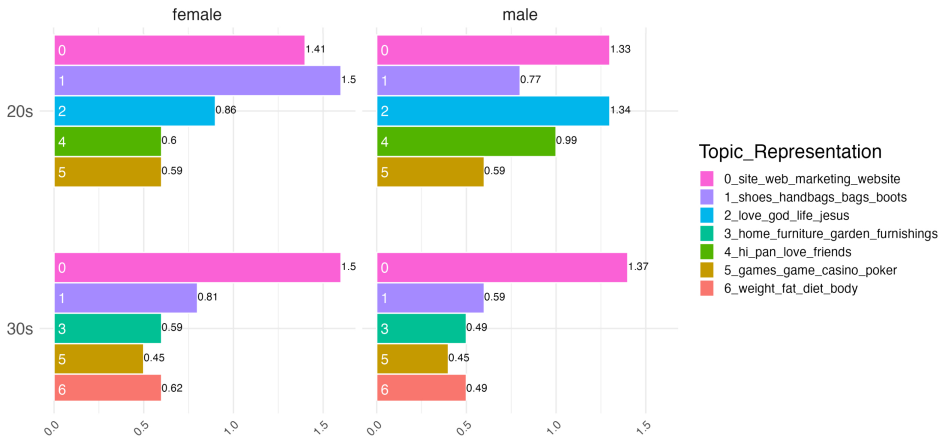
We observe an additional pattern for topics such as websites and marketing (`label 0`) and love, life, and Jesus (`label 2`). The same general rule applies: Topics more frequent for one gender lead to a higher error rate in identifying another gender. However, errors in these topics appear for both genders, accompanied by a comparable age distribution. This pattern indicates that the topical signals alone are inadequate for effective modeling. Individuals of different genders can discuss similar subjects and are also equally susceptible to being incorrectly classified in such discussions.

8 Conclusion

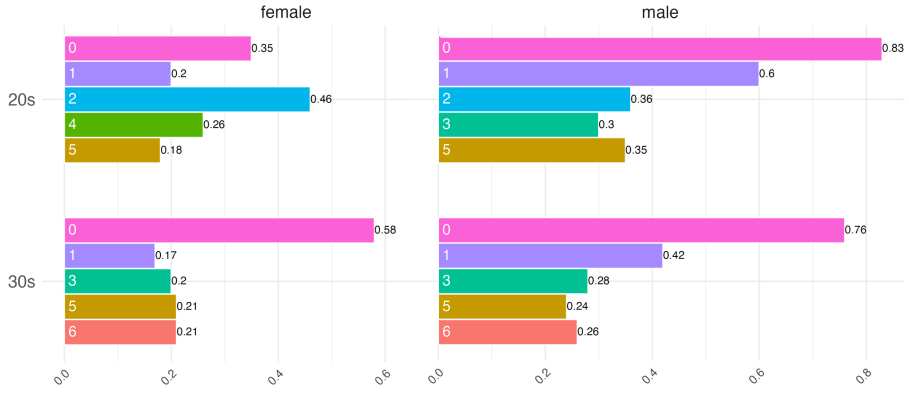
Gender is a complex attribute, and linguistic signals can be very blurry to distinguish among demographic groups (Liu et al., 2021). AP tasks with binary gender classification tend to oversimplify these nuances, potentially reinforcing stereotypes and misrepresentations. In this work, we revisited the authorship profiling task to understand what bias AP models capture and where they make mistakes[11]. We started by demonstrating that standard classifiers achieve relatively low accuracy in predicting authors' gender and age, with varying performance across datasets. Our feature analysis revealed that these differences might be attributed to altering demographically related signals. While the results confirmed that surface and syntactic features significantly correlate with the demographics of authors, surprisingly, the strength and direction of these correlations vary across datasets, irrespective of whether they are in the same or different languages. Moreover, the signals that are consistent across datasets are insufficient for a successful transfer of models between them. Finally, we show that a strong signal for classifiers is the topic of the text. However, classifiers that base their decisions more on the content and not style can exhibit biased behaviors, making mistakes in topics stereotypically associated with a particular interaction between gender and age, causing representational harm.

With the evidence above, we emphasize that us-

[11]The datasets and experimental code for this work are available at https://github.com/HongyuChen2022/AP-task



(a) percentage of posts in top 5 frequent topics (%)



(b) percentage of posts with wrongly predicted gender in top 5 frequent topics (%)

Figure 3: Topics in PAN13-EN; numbers on the right show each bar's share of the total dataset.

ing and interpreting results even from AP classifiers that include only features for gender/age prediction necessitates caution, accounting for both the domains and the models' behavior. Similarly to other NLP classification tasks, AP models aim to learn dataset-specific patterns. These patterns, once learned, are then applied to predict information about new texts. However, as we showed, dataset-specific patterns do not reflect general demographic differences. Therefore, practically applying AP models to new data results in decisions that are either based mostly on stereotypes or that have very low accuracy. Therefore, in use cases that require AP models, it is important to understand the differences between the training and application datasets. Moreover, white-box classifiers that are easy to interpret are the better choice for the prediction methods.

Limitations

Methodologically, our work provides a new perspective on the authorship profiling task and its model behavior for gender/age prediction. We emphasize the importance of examining the relations between dataset-specific patterns and general demographic differences.

However, our work would benefit from exploring more extensive datasets and a broader range of languages. Our experiments are limited to English and Spanish, as they are the two most common languages analyzed in authorship profiling tasks for gender and age prediction (HaCohen-Kerner, 2022). Meanwhile, some of our feature categories are limited to the English datasets. Future research should extend beyond surface and syntactic features across languages. Also, the existing datasets we rely on treat gender as a binary variable (male and female), and age is restricted to only three ranges (10s, 20s and 30s). These restrictions drastically limit the insights of our analyses as well as the models' ability to handle more nuanced variations.

Furthermore, mitigating the identified biases and limitations in AP models requires incorporating strategies such as domain adaptation, reducing topic bias, and creating more robust and generaliz-

able features. Exploring these strategies in future work will enhance the robustness and fairness of AP models, contributing to their practical value and ethical application. Future work could also expand to state-of-the-art Large Language Models that perform very well in related tasks and that are potentially capable of representing features that generalize across datasets. Whether these steps will lead to AP models that are more accurate, fairer and ethically sound remains an open question that needs to be addressed in future work.

Acknowledgements

This work is supported by the Ministry of Science, Research, and the Arts, Baden-Württemberg through the project IRIS3D (Reflecting Intelligent Systems for Diversity, Demography, and Democracy, Az. 33-7533-9-19/54/5). Work by the second author was funded by the DFG Emmy Noether program (RO 4848/2-1). We would like to thank the anonymous reviewers for their valuable feedback. We also thank our colleagues Neele Falk and Pema Gurung for the experimental setup and many conversations on this work.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Arun Rajendran, AbdelRahim Elmadany, Michael Przystupa, and Lyle Ungar. 2019. Sentence-Level BERT and Multi-Task Learning of Age and Gender in Social Media. *arXiv preprint arXiv:1911.00637*.

Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin C. M. Fung. 2021. The topic confusion task: A novel evaluation scenario for authorship attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4242–4256, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Miguel A. Alvarez-Carmona, Adrian Pastor Lopez-Monroy, Manuel Montes y Gómez, Luis Villaseñor-Pineda, and Hugo Jair Escalante. 2015. Inaoe's participation at pan'15: Author profiling task. In *Conference and Labs of the Evaluation Forum*.

Shlomo Argamon, Russell Horton, Mark Olsen, and Sterling Stuart Stein. 2007. Gender, race, and nationality in black drama, 1850-2000: mining differences in language use in authors and their characters. *Digital Hum*, pages 8–10.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text & talk*, 23(3):321–346.

Paul Baker. 2014. *Using corpora to analyze gender*. A&C Black.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, Montréal, Canada. Association for Computational Linguistics.

M Bing, Janet and Victoria L Bergvall. 1998. The question of questions: Beyond binary thinking. In Jennifer Coates, editor, *Language and Gender: A Reader*, pages 496–510. Blackwell, Oxford.

Katherine Bischoping. 1993. Gender differences in conversation topics, 1922–1990. *Sex roles*, 28:1–18.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Margarita Bugueño and Marcelo Mendoza. 2020. Learning to detect online harassment on twitter with the transformer. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 298–306. Springer.

Deborah Cameron. 1997. Theoretical debates in feminist linguistics: Questions of sex and gender. *Gender and discourse*, 1:21–36.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Wendy Ceccucci, Alan Peslak, SE Kruck, and Patricia Sendall. 2013. Does gender play a role in text messaging? *Issues in Information Systems*, 14(2):186.

Na Cheng, Xiaoling Chen, Rajarathnam Chandramouli, and KP Subbalakshmi. 2009. Gender Identification from E-mails. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pages 154–158. IEEE.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2017. Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49:803–821.

Mats Dahllöf. 2023. Author gender and text characteristics in contemporary swedish fiction. *Language and Literature*, page 09639470231223533.

Javier De la Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Maria Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Caio Deutsch and Ivandré Paraboni. 2023. Authorship attribution using author profiling classifiers. *Natural Language Engineering*, 29(1):110–137.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rafael Dias and Ivandré Paraboni. 2020. Cross-domain author gender classification in brazilian portuguese. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1227–1234.

Chris Emmery, Ákos Kádár, Grzegorz Chrupała, and Walter Daelemans. 2022. Cyberbullying classifiers are sensitive to model-agnostic perturbations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2976–2988, Marseille, France. European Language Resources Association.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for Authorship Attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Neele Falk and Gabriella Lapesa. 2022. Reports of personal experiences and stories in argumentation: datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5553, Dublin, Ireland. Association for Computational Linguistics.

R Flesch. 1948. A new readability yardstick journal of applied psychology 32: 221–233.

Eduard Fosch-Villaronga, Adam Poulsen, Roger Andre Søraa, and BHM Custers. 2021. A little bird told me your gender: Gender inferences in social media. *Information Processing & Management*, 58(3):102541.

Maarten R. Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794.

Yaakov HaCohen-Kerner. 2022. Survey on profiling age and gender of text authors. *Expert Systems with Applications*, 199:117140.

Lingshu Hu and Michael Wayne Kearney. 2021. Gendered tweets: Computational text analysis of gender differences in political discussion on twitter. *Journal of Language and Social Psychology*, 40(4):482–503.

Xinyu Hu, Weihan Ou, Sudipta Acharya, Steven HH Ding, Ryan D'Gama, and Hanbo Yu. 2023. Tdrlm: Stylometric learning for authorship verification by topic-debiasing. *Expert Systems with Applications*, 233:120745.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): version 2.0. *Behavior research methods*, 50:1030–1046.

Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2):154–170.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Geoffrey Leech et al. 1992. 100 million words of english: the british national corpus (bnc). *Language research*, 28(1):1–13.

Giada Lettieri, Giacomo Handjaras, Erika Bucci, Pietro Pietrini, and Luca Cecchetti. 2023. How male and female literary authors write about affect across cultures and over historical periods. *Affective Science*, 4(4):770–780.

Tatiana Litvinova, Pavel Seredin, Olga Litvinova, and Olga Zagorovskaya. 2018. Identification of gender of the author of a written text using topic-independent features. *Pertanika Journal of Social Sciences & Humanities*, 26(1).

Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. 2021. The authors matter: Understanding and mitigating implicit bias in deep text classification. *arXiv preprint arXiv:2105.02778*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A Robustly Optimized Bert Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Octavio Loyola-González. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Damián Morales Sánchez, Antonio Moreno, and María Dolores Jiménez López. 2022. A white-box sociolinguistic model for gender detection. *Applied Sciences*, 12(5).

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA. Association for Computational Linguistics.

Debajyoti Mukhopadhyay, Kirti Mishra, Kriti Mishra, and Laxmi Tiwari. 2021. Cyber Bullying Detection Based on Twitter Dataset. In *Machine Learning for Predictive Analysis: Proceedings of ICTIS 2020*, pages 87–94. Springer.

Francisco Manuel Rangel Pardo and Paolo Rosso. 2016. On the impact of emotions on author profiling. *Inf. Process. Manag.*, 52:73–92.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Juan Pizarro. 2019. Using n-grams to detect bots on twitter. In *CLEF (Working Notes)*.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the Author Profiling Task at PAN 2013. In *CLEF conference on multilingual and multimodal information access evaluation*, pages 352–365. CELCT.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Laurie A Rudman and Peter Glick. 2021. *The social psychology of gender: How power and intimacy shape gender relations*. Guilford Publications.

J Schler, M Koppel, S Argamon, and JW Pennebaker. 2006. Effects of Age and Gender on Blogging in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, volume 1.

Juan Soler-Company and Leo Wanner. 2018. On the Role of Syntactic Dependencies and Discourse Relations for Author and Gender Identification. *Pattern Recognition Letters*, 105:87–95.

Ben Verhoeven, Iza Škrjanec, and Senja Pollak. 2017. Gender profiling for Slovene Twitter communication: the influence of gender marking, content and style. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 119–125, Valencia, Spain. Association for Computational Linguistics.

Xiufang Xia. 2013. Gender differences in using language. *Theory & Practice in Language Studies*, 3(8).

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Frederik Zuiderveen Borgesius et al. 2018. Discrimination, Artificial Intelligence, and Algorithmic Decision-making. *línea], Council of Europe*.

A Appendix

EN	BERT	bert-base-uncased
EN	RoBERTa	roberta-base
EN	XLNet	xlnet-base-cased
ES	BERT	bert-base-spanish-wwm-uncased
ES	RoBERTa	bertin-roberta-base-spanish

Table 3: Hugging Face models used for black-box classifiers.

	Gender		Age	
	PAN13	BLOG	PAN13	BLOG
Majority	0.001	0.000	0.204	0.000
LR	0.004	0.000	0.001	0.001
DT	0.003	0.003	0.004	0.011
RF	0.003	0.005	0.024	0.019
NB	0.000	0.000	0.000	0.000
BERT	0.023	0.012	0.017	0.012
RoBERTa	0.003	0.017	0.008	0.006
XLNet	0.002	0.016	0.013	0.007

Table 4: Standard deviation for results in Table 2.

A.1 Best Model Selection

fig. 4 describes the syntax we use in R for two types of regression models being applied to PAN13-EN and BLOG datasets:a binomial logistic regression classifier (LR) for gender classification and a multinomial logistic regression classifier (MLR) for age classification.

table 9 shows the Bayesian Information Criterion (BIC) metrics we use to determine the model that best fits the data. Two scenarios for models are assessed: models with our four groups of features and also with gender/age information controlled; and models with four groups of features only. Lower BIC scores indicate a more favorable fit. For PAN13-EN, model 5 (gender prediction, controlled for age) with a BIC score of 102978 and model 5 (age prediction, controlled for gender) with 113900 are selected. In the case of PAN13-ES, adding gender/age information of authors does not improve the scores of models as much as in the case of PAN13, and also due to limited feature groups available for PAN13-ES, we have selected model 2 (gender prediction) with a BIC score of 104489 and model 2 (age prediction), with 119938. For BLOG, though exposed with more feature options, only surface and syntactic features give models the lowest BIC scores, where model 2 (gender prediction) and model 2 (age prediction, gender controlled) emerged as the preferred choices, with BIC scores of 5467 and 8115, respectively.

```
glm(Gender ~ Group A, family = 'binomial
    ', data = PAN13/Blogs)
glm(Gender ~ Group A + Group B, family =
     'binomial', data = PAN13/Blogs)
...
glm(Gender ~ Group A + Group B + Group C
     + Group D + Group E, family = '
    binomial', data = PAN13/Blogs)
```

```
multinom(Age ~ Group A,family = '
    multinomial', data = PAN13/Blogs)
multinom(Age ~ Group A + Group B,family
    = 'multinomial', data = PAN13/Blogs)
...
multinom(Age ~ Group A + Group B + Group
     C + Group D + Group E,family = '
    multinomial', data = PAN13/Blogs)
```

Figure 4: Binomial logistic regression for gender (top) and multinomial logistic regression for age (bottom).

	PAN13-EN	
	train	test
male	37,949/24,477,667	12,648/5,696,380
female	37,949/28,233,153	12,711/7,075,832
10s	2,500/1,969,032	1776/1,094,296
20s	42,598/26,476,213	9175/2,988,055
30s	30,800/24,477,667	14,408/8,689,861
Total	75,898/52,922,912	25,359/12,772,212

	PAN13-ES	
	train	test
male	37,950/10,311,857	4,080/991181
female	37,950/9,420,533	4,080/877,135
10s	2,500/411,742	288/56,518
20s	42,600/10,363,481	4,608/1,042,463
30s	30,800/8,957,167	3,264/769,335
Total	75,900/19,732,390	8,160/1,868,316

	BLOG	
	train	test
male	2,096/15,451,310	1,931/12,986,336
female	2,094/15,502,898	1,931/14,161,124
10s	1,400/7982322	1,648/8424833
20s	1,398/11,419,916	1,616/12,906,272
30s	1,392/11,551,970	598/5,816,355
Total	4,200/30,888,893	3,862/27,147,460

Table 5: Statistics of the data used for analysis: number of files (authors) / number of words.

	Gender			Age		
	PAN13-EN	PAN13-ES	BLOG	PAN13-EN	PAN13-ES	BLOG
Majority	0.001	0.000	0.000	0.000	0.000	0.126
LR	0.000	0.001	0.001	0.001	0.000	0.001
DT	0.002	0.004	0.004	0.001	0.003	0.004
RF	0.012	0.019	0.030	0.023	0.016	0.027
NB	0.000	0.000	0.000	0.000	0.000	0.000
BERT	0.004	0.001	0.003	0.009	0.004	0.006
RoBERTa	0.003	0.007	0.005	0.005	0.003	0.005
XLNet	0.008	–	0.005	0.005	–	0.006

Table 6: Standard deviation for results in Table 1.

	Gender			Age		
	PAN13-EN	PAN13-ES	BLOG	PAN13-EN	PAN13-ES	BLOG
Majority	0.50	0.50	0.50	0.56	0.56	0.33
LR	0.58	0.66	0.74	0.61	0.67	0.70
DT	0.54	0.56	0.61	0.54	0.55	0.53
RF	0.58	0.63	0.71	0.60	0.63	0.63
NB	0.55	0.55	0.61	0.46	0.45	0.52
BERT	0.58	0.71	0.76	0.60	0.68	0.65
RoBERTa	0.59	0.70	0.79	0.62	0.67	0.73
XLNet	0.58	–	0.75	0.62	–	0.71

Table 7: Accuracy for gender and age prediction on 5-fold evaluation of training datasets.

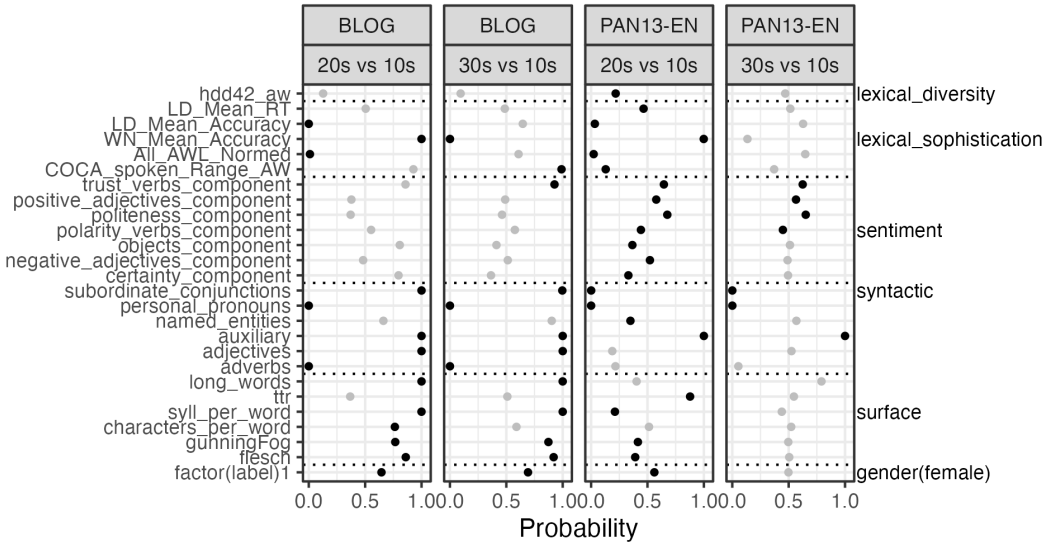


Figure 5: Significant ($p < 0.05$) features for age as DV; model selection used all five categories of features.

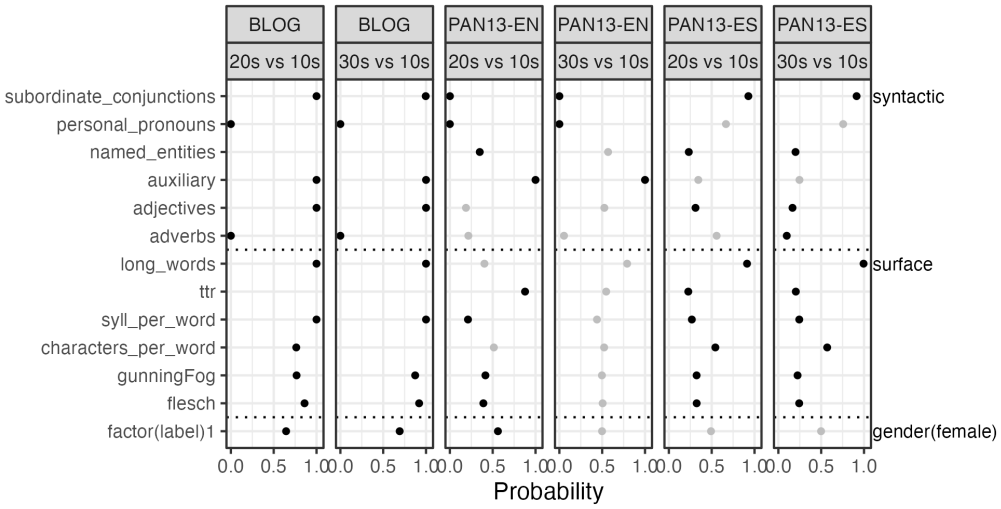


Figure 6: Significant ($p < 0.05$) features for age as DV; model selection used only surface and syntactic features.

		10s		20s		20s		
		correct	error	correct	error	correct	error	Total
female	PAN13-EN	860	390	14,793	6,506	9,529	5,871	37,949
	PAN13-ES	925	325	14,959	6,341	10,621	4,779	37,950
	BLOG	564	136	540	159	553	144	2,096
male	PAN13-EN	461	**789**	11,393	**9,906**	7,937	**7,463**	37,949
	PAN13-ES	836	414	15,003	6,297	10,521	4,879	37,950
	BLOG	518	182	574	125	553	142	2,094

Table 8: Correct and error cases in predicting gender by RoBERTa. Highest number of errors in each column bolded.

	Gender				
	m1	m2	m3	m4	m5
PAN13-EN	104,300	104,341	103,941	103,804	103,195
	104,274	104,314	103,850	103,672	102,978
BLOG	5,581	5,467	5,617	5,702	5,717
	5,541	5,467	5,558	5,643	5,658
PAN13-ES	104,629	104,489			
	Age				
	m1	m2	m3	m4	m5
PAN13-EN	118,048	117,952	116,506	115,272	114,132
	118,022	117,952	116,416	115,139	113,900
BLOG	8,349	8,177	8,473	8,640	8,626
	8,304	8,115	8,411	8,578	8,626
PAN13-ES	120,134	119,938			

Table 9: BIC score for gender and age prediction model 1 to model 5.