Data & Knowledge Engineering 151 (2024) 102307

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Evaluating Transformers and Linguistic Features integration for Author Profiling tasks in Spanish

José Antonio García-Díaz [a], Ghassan Beydoun [b], Rafel Valencia-García [a,*]

[a] Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo, Murcia, 30100, Murcia, Spain
[b] University of Technology Sydney, 81 Broadway, Ultimo, NSW, Australia

ARTICLE INFO

Keywords:
Author profiling
Authorship analysis
Feature engineering
Natural language processing

ABSTRACT

Author profiling consists of extracting their demographic and psychographic information by examining their writings. This information can then be used to improve the reader experience and to detect bots or propagators of hoaxes and/or hate speech. Therefore, author profiling can be applied to build more robust and efficient Knowledge-Based Systems for tasks such as content moderation, user profiling, and information retrieval. Author profiling is typically performed automatically as a document classification task. Recently, language models based on transformers have also proven to be quite effective in this task. However, the size and heterogeneity of novel language models, makes it necessary to evaluate them in context. The contributions we make in this paper are four-fold: First, we evaluate which language models are best suited to perform author profiling in Spanish. These experiments include basic, distilled, and multilingual models. Second, we evaluate how feature integration can improve performance for this task. We evaluate two distinct strategies: knowledge integration and ensemble learning. Third, we evaluate the ability of linguistic features to improve the interpretability of the results. Fourth, we evaluate the performance of each language model in terms of memory, training, and inference times. Our results indicate that the use of lightweight models can indeed achieve similar performance to heavy models and that multilingual models are actually less effective than models trained with one language. Finally, we confirm that the best models and strategies for integrating features ultimately depend on the context of the task.

1. Introduction

Author Profiling (AP) is a subtask of Authorship Analysis (AA) in which demographic and psychographic traits are inferred from a set of authors' writings. Extracting traits such as gender, nationality, age range, political affiliation, appraisals, or desires — just to a name but a few, has many important applications in marketing and customer service overall, enhancing the user experience. Profiling users can also be used in social networks to identify and address behavior detrimental to the overall benefits of the network, such as detecting bots or spreaders of fake news and hate speech, rather than just focusing on the content itself [1,2].

A Knowledge-Based System (KBS) is a type of computer system that uses expert knowledge and reasoning to solve complex problems in a particular domain. KBSs are used in diverse applications such as medical diagnosis [3], legal advice [4], and crisis management [5]. They mimic human decision-making processes by applying stored knowledge to new situations. KBSs excel at processing large amounts of complex information and making decisions based on specialized knowledge, improving efficiency and accuracy in various fields.

[*] Corresponding author.
E-mail address: valencia@um.es (R. Valencia-García).

https://doi.org/10.1016/j.datak.2024.102307
Received 31 May 2023; Received in revised form 9 April 2024; Accepted 25 April 2024
Available online 29 April 2024

The relationship between AP and KBS is critical to understanding and exploiting the information contained in large textual datasets. Extracting demographic and psychographic traits from author writings, can provide valuable insights for building and managing KBS that understand author preferences, trends, and behaviors. Furthermore, the evaluation of lightweight, monolingual and multilingual Transformer-based models for AP also contributes to the effective design and implementation of KBS based on authorship analysis, improving their ability to organize, manage, and exploit the information contained in large textual datasets.

In the last decade in particular, there has been a growing research interest in AP. The work done in the PAN workshop[1] specializing in AA should be highlighted. PAN has organized several tasks in different languages related to profiling, authorship obfuscation, digital text forensics, or stylometry, among others. Other workshops have also focused on AA tasks, including IberLEF and Evalita, focusing on the Spanish and Italian languages, respectively.

AP is often treated as an Automatic Document Classification (ADC) task, where traits to be identified are construed as a narrow subset of labels. In fact, the same techniques and approaches used in ADC are appropriate for AP. However, the analysis of several other works in AP shows that traditional approaches based on term counting features can improve state-of-the-art approaches based on transformers which have also shown significant improvements in other ADC tasks. One possible explanation for this finding is that there is only a limited number of different authors in a given AP dataset (rather than the hundreds of documents from the same author typically required). Another explanation is that term counting features capture more informative features from smaller datasets. Some datasets may also be biased toward certain keywords or expressions used by authors who compiled the dataset, making the resulting models less generic and less suitable for real-world scenarios.

In recent years, an increasing number of new language models for Spanish have appeared, based on transformers architectures such as BERT and RoBERTa, and based on lightweight models built using distillation techniques or lightweight architectures. In this paper, we revisit some of the most recent AP datasets in Spanish to evaluate their performance in this type of task. We also evaluate different feature integration techniques to combine these models with Linguistic Features (LF), to provide some degree of interpretability in the resulting methods. The feature integration is based on two strategies: Ensemble Learning (EL) and Knowledge Integration (KI). Finally, we compare the results with a baseline of non-contextual sentence embeddings from fastText.

Accordingly, the following Research Questions (RQ) are defined:

- **RQ1**. Determine which language models are most effective for conducting AP in Spanish.
- **RQ2**. Determine whether feature integration improves the performance of separate language models
- **RQ3**. Determine which feature integration techniques improve the performance for conducting AP in Spanish.
- **RQ4**. Determine if the interpretability of the machine learning classifiers can be improved using linguistic features.
- **RQ5**. Determine the performance of the language models in terms of memory and time to perform AP.

Our results show that there is no clear superior language model and that the best performance actually depends on both the task and the available dataset. For instance, MarIA (based on RoBERTa) outperforms BETO in the four evaluated features in PAN which was not observed in PoliticES 2022. We also found that the language models trained on the Spanish dataset outperform multilingual models and that lightweight models can achieve similar performance to large models. Regarding the feature integration, our results show that there is not always a superior performance either, nor a better strategy for combining the features, but using ensemble learning with the highest probability strategy can achieve good results in binary classification tasks. Fourth, regarding the interpretability of the results, we observe that stylometric linguistic features are quite influential in AP and, to a lesser extent, features related to morphology and lexicon are also influential. Finally, in terms of performance evaluation, we observed that the time required for training can be a limiting factor in some scenarios. However, the performance of the lightweight models is very promising and suitable for real scenarios but those models are not as stable as larger models across all tasks.

The rest of the manuscript is organized as follows: In Section 2, the AP state-of-the-art techniques and datasets are described. In Section 3, our proposal and evaluation datasets described in detail. In Section 4, the results are described and discussed. Finally, the conclusions and future research directions are given in Section 5.

2. State-of-the-art

Most of the publicly shared AP challenges are organized on behalf of PAN. Early AP challenge tasks focused on two demographic traits: gender and age range. The most recent challenges however focused social welfare perspective, identifying bots in 2019 [6], fake news spreaders in 2020 [7], hate speech spreaders in 2021 [8], and stereotype spreaders in 2022 [9]. There was also a shared task that focused on profiling celebrities by identifying their occupation, age range, gender, and level of fame. Another task also focused identification of celebrity characteristics, but by tracking the text of their followers rather than the celebrities themselves.

It is common to allow PAN participants to send their results in a specified language. The majority of the above PAN tasks focused still on English and Spanish. In Spanish, most of the shared AP tasks are proposed by IberLEF. For example, the MEX-A3T shared tasks organized in 2018 [10] and 2019 [11] related to the identification of demographic traits: gender, occupation and place of residence. The datasets for these tasks are compiled from Mexican Spanish tweets. In later multimodal editions, images were included. In 2018, the dataset included 5K users, differentiated 8 occupations and 6 birthplaces. Gender and 11 different images to some of the tweets were included in the 2019 dataset. PoliticES 2022 shared task [12], main goal was to extract the political

[1] https://pan.webis.de/shared-tasks.html.

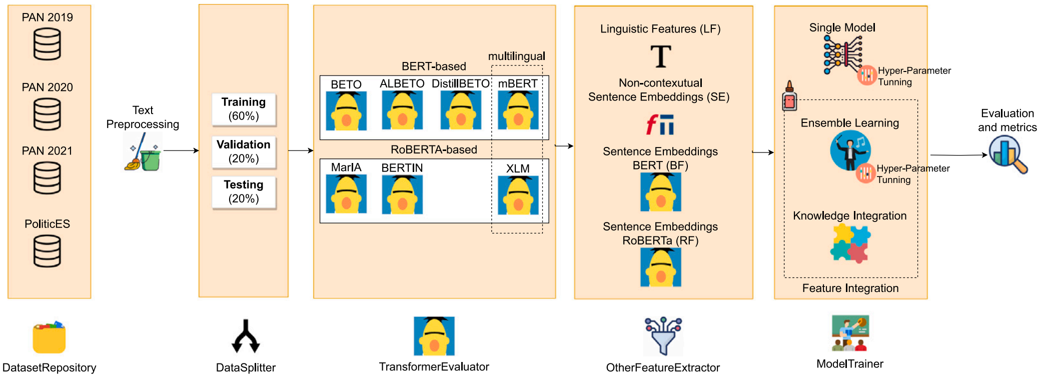

Fig. 1. Pipeline.

ideology from documents. Political affiliation can be construed as a psychographic trait that facilitates the understanding of social behavior. In this work, it is classified as a binary (left vs. right) and multi-class (left vs. right, moderate vs. non-moderate). 14 teams participated in PoliticES 2022, with the majority of approaches based on Transformers.

In addition to the above forums, there is in the literature some works that deal with AP. For example, in 2022, [13] the authors evaluate some deep learning approaches in AP in documents written in Spanish or English. They found that it is possible to achieve and outperform state-of-the-art models with less than half the training samples. In [14], the authors proposed a system based on a novel fusion strategy for transformers, consisting in a fusion of word embeddings and stylistic features of authors using a self-attention. They evaluated their proposal in Spanish and English in several AP tasks, including author profiling and plagiarism and authorship analysis. The improvements obtained ranged from 2.5% to 3.5%.

In comparison to existing approaches, we focus in this work on the evaluation of novel language models to perform AP tasks in Spanish. Due to the high computational requirements of Transformers-based language models compared to traditional approaches [15], we focus on measuring the performance in terms of reliability and training and inference time. We include in this work different types of language models including heavy models, lightweight models, Spanish and multilingual. Accordingly, to answer the proposed research questions, we select different datasets from those described above. We discard those that are not available in Spanish and those that are older than 2019. The reader can find more details about these tasks in Section 3.2.

3. Materials and methods

In this section, we describe our methodology and the techniques used (see Section 3.1). We also describe the datasets used for the evaluation (see Section 3.2).

3.1. System architecture

The system architecture in our proposal is shown in Fig. 1. It works as follows: First, the `DatasetRepository` module is responsible for selecting one of the datasets to be evaluated. This includes the latest versions of PAN available in Spanish and the PoliticES 2022 dataset. Second, the datasets are then pre-processed and normalized. Third, the `TransformerEvaluator` module next identifies the most appropriate models based on the Transformers architecture for AP. These include monolingual and multilingual transformers, and lightweight and heavyweight versions of the transformers. They are organized in two main sub-modules: BERT and RoBERTa, and they identify the best candidate for each. Fourth, the remaining of the feature sets are then extracted, including LF from UMUTextStats [16], non-contextual sentence embeddings from FastText (SE), and two sentence embeddings from the transformers — one based on the best BERT (BF) and another based on the best RoBERTa (RF). Fifth and last step, the `ModelTrainer` module evaluates several deep learning architectures. These models use only one feature set as input (single model) or several. If multiple feature sets are evaluated, two strategies are considered: KI, where all features are used in a multi-input neural network, and EL, where the outputs of all single machine learning models are combined using different strategies. I what follows, we provide details on each of the modules of the system and their functionalities.

3.1.1. Text preprocessing

The goal of the text preprocessing module is to cleanse and normalize all documents in the datasets. Basically, all social media content such as hyperlinks, hashtags, and mentions are removed from the documents. Numbers are also replaced with a fixed token, and expressive elongations used as an emphasis markers is also removed. Finally, misspellings are corrected using ASPELL. For LF extraction, we keep different versions of each document. First, a normalized version is used to extract Part-of-Speech (PoS) features. Second, the original version of each document is used to extract specific LF, such as expressive lengthening, social media jargon and correction, and style.

3.1.2. Contextual embeddings evaluator

The main advantage of models based on transformers is that they extract contextual embeddings, i.e. the embedding of each token varies according to the surrounding words. This is computed using an attention method. The main advantage of the transformers is that they can handle polysemy [17].

These models are typically available as Large Language Models (LLMs). LLMs are typically available trained from unsupervised tasks such as Mask Language Modeling (MLM) or Next Sentence Prediction (NSP), making transformers a form of transfer learning, as these models can be fine-tuned to solve other tasks such as ADT or question answering.

The first proposed transformer architecture was Bidirectional Encoder Representations from Transformers (BERT) [18], by Google. BERT takes into account the specific context of each word by scanning the entire document. Thus, it generates a different representation for the same word depending on its context. BERT learns the word embeddings using two strategies: The first is MLM and this consists in randomly masking several words in the training dataset and learning to predict the masked words. The second is NSP and in this BERT concatenates two sentences and the model must determine whether the two sentences are consecutive.

Several of these transformer-based architectures are available for Spanish, including models based on BERT, RoBERTa and lightweight versions based on ALBERT and distillation techniques. In this step, our system evaluates all these models separately and then selects the best model for each architecture. Thus, we can evaluate whether BERT and RoBERTa are complementary or not. Next, we describe the transformers evaluated in this paper.

- **BETO** [19], is a Spanish language model based on BERT. BETO has been trained on Spanish Unannotated Corpora.
- **multilingual BERT** [18]. This version of BERT is trained on documents written in more than 100 languages from Wikipedia. The model is trained using MLM.
- **DistilBETO** [20]. This model is a distilled version of BETO [19]. Therefore, their architectures are similar. However, distilBETO has neither token-type embeddings nor the pooler layer, and has only 6 hidden layers.
- **ALBETO** [20] is another pre-trained model based on BERT, but trained in an ALBERT fashion [21], which is a more efficient strategy for training BERT in which all parameters are shared among the various layers of the model.

RoBERTA is another transformer-based architecture. It is based on several improvements to the BERT training process. First, RoBERTa does not use NSP during training. Second, RoBERTa dynamically masks tokens during the training epochs. Since our focus is Spanish documents, we evaluate the following language models based on RoBERTa:

- **MarIA** [22] is a Spanish pre-trained model based on the RoBERTa architecture. MarIA has been pre-trained with content crawled from the National Library of Spain with almost 600 GB.
- **BERTIN** [23]. This model is a training from scratch of RoBERTa, trained on the Spanish texts from mC4, compiled from the public Common Crawl web scrape. This dataset contains *about 416 million samples and 235 billion words in about 1 TB of uncompressed data*.
- **XLM-RoBERTa** [24]. This is a multilingual version of RoBERTa, pre-trained with about 2.5 TB of data from the 100 different languages of CommonCrawl. It is trained using masked language modeling, which randomly masks the 15% of the tokens in the input.

Benchmarking of the language models is done as follows: For each model, we obtain its embeddings using the normalized version of each document. For a comparison, we then perform a hyperparameter tuning step, training 10 models per model. The tuned hyperparameters are: (1) the number of epochs, between 1 and 5; (2) the weight decay, between 0.0 and 0.3; (3) two batch sizes, 8 and 16 (it is worth noting that we only test these two batch sizes due to hardware limitations); (4) the number of warm-up steps, with values of 0, 250, 500 and 1000, and (5) the learning rate, with values between 1e–5 and 5e–5. The performance of each model is ranked using the custom validation split. The results of this process are presented in Section 4.1.

3.1.3. Feature engineering

The next step in our system is to extract four feature set

- **Linguistic Features (LF)**. These features are extracted using the UMUTextStats [16]. LF captures a wide variety of LF, including PoS, figurative language, pragmatics, register, jargon, and stylometry, among others. In ADT, and especially in AA, the results of LF are satisfactory because they reflect features that are difficult to capture by other means. For example, the raising of an author's voice expressed by capital letters would require the use of cased word n-grams and cased embeddings [25]. However, in some scenarios it is preferable to keep uncased n-grams and embeddings to better generalize the use of certain keywords. In this sense, the LFs capture this phenomenon more easily. For example, linguistic features have proven effective in detecting fake news [26] using surface information features, part of speech, discursive analysis, and readability indices.
- **Non-Contextual Sentence Embeddings (SE)**. We extract non-contextual sentence embeddings (SE) from the Spanish pre-trained model of fastText [27]. These embeddings are pre-trained using Wikipedia and the Common Crawl datasets. This feature set consists of a vector of length 300 for each document. fastText averages these vectors from the individual word vectors, adding an extra `EOS` token to indicate the end of the sentence. To compute the user-level SE, we average all sentence embeddings from each tweet of each user. Note that these features are not used in the feature integration. This decision is made because contextual sentence embeddings are more efficient, and non-contextual sentence embeddings cannot provide interpretability to the model in the way that LF can.

- **Contextual Sentence Embeddings from BERT (BF) and RoBERTa (RF)**. As explained in Section 3.1.2, we choose one model from BERT and another from RoBERTa architectures. We compute the embeddings at the sentence level because it is easy to combine with the LF and because of the length restrictions of transformers. The sentence embeddings are obtained from the [CLS] token, as suggested in [28]. Both BF and RF are embeddings of length 768. This gives us a representation of each tweet as a fixed-length vector of length 768. To get the user-level embeddings, we average all the sentence embeddings from the same author.

3.1.4. Model trainer

The goal of the model trainer is to obtain the best possible neural network to solve each AP task. We divide these neural networks into three categories: single-input neural networks, knowledge integration (also known as multi-input neural networks), and ensemble learning. Single-input neural networks are used to evaluate the performance of each feature set separately. Whereas knowledge integration (KI) and ensemble learning (EL) are used to evaluate the integration of feature sets. As explained in Section 3.1.3, the feature sets to be integrated are the LF with the contextual sentence embeddings from BERT and RoBERTa.

Regardless of the strategy, all neural networks are based on Multi-Layer Perceptrons (MLP). This architecture is chosen because all feature sets are at the user level, i.e., they are the average of the features obtained for each font from each user. Since these features are fixed in size, they are not suitable for training convolutional or recurrent networks, since the spatial and temporal information of natural language is implicitly stored in the data.

To obtain single input neural networks, we use hyperparameter tuning. This process allows us to determine the best configuration of parameters for each model. The best model is selected based on the macro F1 score over a custom validation set from the training data of 20% of the instances. For each neural network in the MLP, we evaluate the number of layers and the number of neurons per layer. We divide the MLP into two groups. Shallow neural networks consist of one or two hidden layers and the same number of neurons per layer. Deep neural networks consist of more than two hidden layers and a different number of neurons per layer arranged in different shapes, namely brick, funnel, long funnel, triangle, rhombus, and diamond.[2] Other parameters evaluated are (1) the activation function that connects each hidden layer to the next; (2) the dropout mechanism that is used to avoid overfitting. This process randomly removes some of the neurons during training. We configure this parameter to evaluate a value between 10% and 30%. We also evaluate without using the dropout mechanism checking for the batch size and the learning rate. All models were trained using a time-based learning rate scheduler (see Eq. (1)).

$$lr * 1/(1 + (lr/epochs) * epoch) \tag{1}$$

To obtain the KI, we repeat the process described for single-input neural networks, but train from scratch a multi-input neural network that combines the LF, BF, and RF. This allows the neural network to learn during training how to combine the features and to identify which ones are best suited for each sample. In this network, each feature set is used as input to different hidden layers, and then all hidden layers are concatenated before the output of the last layer.

To obtain the EL, which consists in generating the output of the network combining the outputs or probabilities of each model, we evaluate four different ways of combining the individual networks: (1) using the mode of predictions (hard voting) (2) using weighted mode, in which the influence of each model depends on its performance with the validation split (soft voting) (3) using average probabilities, where we average the probabilities of each neural network, and (4) using highest probability, where the final output is decided by the network with greater confidence.

3.2. Datasets

The Table 1 summarizes the last AP challenge tasks organized by PAN. In the following subsections, the most recent records and a brief summary of the overview of each task are given.

3.2.1. PAN author profiling 2019

The Bots and Gender Profiling 2019 [6] determines whether a Twitter profile is a real user or a bot and, if the user is human, to determine its gender. This challenge was proposed in Spanish and English, with the proposals being ranked by the average of both languages.

Due to the scope of our work, we only focus on results obtained for Spanish of which the best is achieved by Pizarro [35]. Their proposal is based on traditional word n-gram features and Support Vector Machines (SVM) with hyperparameter tuning, improving modern approaches based on deep learning. The organizers of the challenge also provided two similar baselines, based on character and word n-grams. The character n-gram baseline performed best, but slightly slower than Pizarro (89.72% accuracy for bot identification and 72.89% accuracy for gender prediction).

Looking at the overall results of the 55 participants, we can see that the average results are reasonably high: 84.080% accuracy for bot identification and 70.170% accuracy for gender identification.

[2] Inspired by https://github.com/autonomio/autonomio.

Table 1
Summary of the AP shared task in PAN in Spanish comparing gender and age range scores.

Year	Topic	Has spanish	# of spanish authors
2013 [29]	General	Yes	90,860
2014 [30]	Social media	Yes	1,960
”	Blogs	Yes	158
”	Twitter	Yes	294
2015 [31]	Twitter	Yes	228
2016 [32]	Twitter	Yes	370
2017 [33]	Twitter	Yes	7,000
2018 [34]	Twitter	Yes	5,200
2019 [6]	Bots	Yes	2,400
2020 [7]	Fake news	Yes	500
2021 [8]	Hate-Speech	Yes	400
2022 [9]	Stereotypes	No	–

3.2.2. PAN author profiling 2020

This AP shared task from 2020 [7] determines whether an author is a fake news spreader. The dataset from this shared task contains Spanish and English authors and it was compiled from Twitter. The dataset is balanced and contains 100 tweets from a total of 500 authors (250 of which are fake news spreaders).

This contest is scored by accuracy, and both languages were scored separately and combined. A total of 54 participants submitted a proposal to solve the Spanish challenge, achieving an average accuracy of 73.29%, with a standard deviation of 0.06. The best result for the Spanish part was achieved by Pizarro [36] with an accuracy of 82%. Pizarro et al. use on a Support Vector Machine classifier and character and word n-gram features encoded with TF–IDF. They worked as a user level because it combined all documents from the same authors. They then preprocessed documents by replacing emoticons and digits and converting some text to its lowercase form. They also performed a hyperparameter evaluation to decide: (1) values from the features, such as the n-gram range and the maximum and minimum frequency of the n-grams, and (2) SVM parameters, such as the loss function or C-value. This step was performed by using fold cross-validation ($n = 10$).

3.2.3. PAN author profiling 2021

The AP shared task from 2021 [8] determines which authors are hate speech spreaders. The dataset from this shared task includes Spanish and English authors compiled from Twitter. The organizers compiled 200 tweets per author. In addition, they implemented several baselines based on TF–IDF features and neural networks based on LSTM.

The best result, in both in Spanish and English, was achieved by [37], which reached an accuracy of 85% for the Spanish, in which a shallow convolutional neural network (CNN) was trained to perform the classification task. The authors also worked at the user level, reading each XML file containing all the writings of the same author and inputting them into the first layer of their deep neural network. They then split the texts into nearly 4000 character n-grams. The next layer generated a word embedding of length 100 for each character n-gram. The next step was to apply several 1-dimensional convolutional filters and then use a mean pooling layer to merge the features. The neural network was trained using binary cross entropy as the loss function.

3.2.4. PoliticEs 2022

The PoliticEs 2022 dataset [12] is an extension of the Spanish PoliCorpus 2020 [38]. It was compiled from Twitter of tweets from politicians and journalists in Spain. The dataset is annotated with gender, profession and political ideology (binary and multiclass) of the users. The dataset contains more than 120 tweets from a total of 400 users. A total of 14 teams participated in this shared task. Most of their proposal were based on transformers, but also included traditional machine learning algorithms and other techniques such as data augmentation [39], adversarial attacks [40], or ensemble learning [41].

3.2.5. Dataset statistics

To perform a hyperparameter optimization step, we remove some users from the training splits for each dataset and trait. Table 2 describes the total number of users for each dataset, the number of features and the number of labels. It can be observed that the number of users is significantly larger in 2019 PAN shared task, with 4800 users half of whom were real users. All PAN shared tasks are binary other than for the PoliticES 2022 dataset, which has a strong imbalance, especially for profession and political ideology in a multiclass perspective.

4. Results and discussion

This section is organized to answer previously formulated RQs. First, Section 4.1 answers the question which language models are best for performing AP tasks. Second, Section 4.2 analyzes whether particular features improve AP and what techniques are required to do so. Third, Section 4.3 focuses on the interpretability of the models, focusing on LF. Finally, Section 4.3 measures the performance of the evaluated language models in terms of size and training and inference time.

We measure the performance of the language models using the macro-averaged precision, the recall score and the macro F1 score. Precision measures the accuracy of positive predictions and represents the ratio of correctly predicted positive instances to

Table 2
Number of users for each trait of the PAN and PoliticES datasets.

PAN shared tasks									
2019 (Nature)					2019 (Gender)				
label	train	val	test	total	label	train	val	test	total
Bot	1203	297	900	2400	Female	608	142	450	1200
Human	1212	288	900	2400	Male	604	146	450	1200
Total	2415	585	1800	4800	Total	1212	288	900	2400
2020 (Fake news spreaders)					2021 (Hate-speech spreaders)				
label	train	val	test	total	label	train	val	test	total
Yes	124	26	100	250	Yes	80	20	50	150
No	125	25	100	250	No	81	19	50	150
Total	249	51	200	500	Total	161	39	100	300
PoliticES 2022									
Gender					Profession				
label	train	val	test	total	label	train	val	test	total
Female	116	21	36	173	Journalist	42	20	25	87
Male	136	41	69	246	Politician	210	42	80	332
Total	252	62	105	419	Total	252	62	105	419
Political ideology (binary)					Political ideology (multiclass)				
label	train	val	test	total	label	train	val	test	total
Left	147	31	57	235	Left	66	10	21	97
					Left moderate	81	21	36	138
Right	105	31	48	184	Right moderate	70	24	31	125
					Right	35	7	17	59
Total	252	62	105	419	Total	252	62	105	419

the total number of predictions. The recall score measures the model's ability to capture all positive instances and represents the ratio of correctly predicted positives to the total number of actual positives. The F1 score provides a balanced measure considering both precision and recall scores, calculated as the harmonic mean of the two. We average these metrics using the macro strategies to produce a balanced performance measure across all labels. It is worth noting that: (1) not all evaluated datasets are balanced, as is the case for all PoliticES 2022 features; (2) political ideology across four axes is a multi-classification problem; and (3) in some binary classification problems, all classes such as gender or occupation detection need to be considered equally important.

4.1. RQ1. Language models evaluation

In this subsection, we report our results obtained using different state-of-the-art language models based on Transformers. The evaluated language models include two architectures (BERT-based and RoBERTa-based), heavyweight and lightweight models based on ALBERT and Distillation, and Spanish and multilingual models. All models are trained at document-level using hyperparameter tuning, due to the maximum length restriction of Transformer's models. The range of the hyperparameters evaluated was described in Section 3.1.2. Finally, as described in the methodology (see Section 3.1), we select two models from these experiments, one based on each of the architectures, BERT and RoBERTa.

Results of each language model applied to the PAN shared tasks are shown in Table 3. These results are reported with the official test set published for each challenge. As can be seen, MarIA achieved the best results. However, the degree to which MarIA outperforms BETO varies from task to task. In 2019 (Nature), the difference is a 0.167. In 2019 (Gender), it is 7.074%. In 2020 (Fake news spreaders), it is 1.424%, and in 2021 (Hate-speech spreaders) the difference is 7.24%. This illustrates the poor reliability of multilingual and distilled models for 2019 (Nature), in which only BETO and MarIA achieved decent results. These limited results are not observed when we evaluate our model using the custom validation set in other tasks. Moreover, the precision and recall of the models obtained for 2019 using distilled and multilingual models are similar. Since the label distribution is balanced, i.e. the model does not always predict the same class, there is likelihood of a missing data problem. The results also show that there are significant differences between the training and validation splits, where multilingual and distilled models evidently fail to generalize. Another noteworthy finding is that DistilBETO and ALBETO outperform BETO in 2019 (Gender), 2020 (Fake news spreaders), and 2021 (Hate-speech spreaders). These results all indicate that the improvements of these architectures outperform the performance of models based on BERT, while still relying on lightweight architectures. Finally, comparing multilingual and language-specific models, we find that multilingual BERT (mBERT) outperforms XLM in all tasks.

We perform a document-level error analysis to examine the nature of false predictions made by MarIA and BETO in all the PAN shared tasks evaluated. We do not observe any significant differences between the BERT and RoBERTa architectures used. In what follows, we discuss the main findings per trait. First, with respect to the problem of distinguishing between real users and bots (2019 Nature), we observe that there are more false predictions of real users misclassified as bots. Most of the misclassified users are those who use the social network as a showcase for their work, including photographs, paintings, poems, or song lyrics. Other legitimate

Table 3
Comparison of transformer-based features with the 2019, 2020, and 2021 PAN shared tasks.

	Precision	Recall	f1-score	Precision	Recall	f1-score
	2019 (Nature)			2019 (Gender)		
BETO	88.461	88.389	88.383	65.731	65.000	64.589
DistilBETO	51.709	50.389	38.516	66.601	66.222	66.028
ALBETO	47.674	48.778	41.883	65.779	65.778	65.777
MarIA	**88.624**	**88.556**	**88.550**	**71.770**	**71.667**	**71.633**
BERTIN	60.996	58.556	56.121	61.631	61.444	61.289
mBERT	62.853	60.000	57.650	67.669	67.333	67.178
XLM	50.910	50.833	49.777	68.247	67.889	67.731
	2020 (Fake news spreaders)			2021 (Hate-speech spreaders)		
BETO	80.012	80.000	79.998	77.083	76.000	75.758
DistilBETO	80.012	80.000	79.998	78.180	78.000	77.965
ALBETO	80.503	80.500	80.500	81.012	81.000	80.998
MarIA	**82.042**	**81.500**	**81.422**	83.013	**83.000**	**82.998**
BERTIN	78.733	78.500	78.456	74.155	74.000	73.958
mBERT	81.453	81.000	80.931	**83.119**	**83.000**	82.985
XLM	80.303	80.000	79.950	77.098	77.000	76.979

users use the network to post headers and hyperlinks to personal projects and blogs. We also observe a significant presence of hashtags among legitimate users such as `quotes`, and the presence of emoji of musical notes and instruments. The information gain of the LFs of the misclassified subset indicates that relevant features are the number of sentences (e.g. larger tweets), the use of hyperlinks (more common in bots), sentences starting with lowercase letters (more common in bots), and lexical items related to organizations, places, and people but to different degrees. For example, for the misclassified documents, we observe that more bots contain topics related to people. Second, with respect to the distinction between male and female users (2019 Gender), there are misclassifications for both genders. We observe that many of the misclassifications are related to tweets used as hyperlinks to news sites. Third, with respect to identifying fake news spreaders (2020), we find that several false classifications of users who are really fake news spreaders use a starting phrase such as *Mira lo que he compartido*.[3] Fourth, regarding the identification of hate speech spreaders (2021), we observe that many tweets misclassified as non-hate speech spreaders contain offensive language but not necessarily hate speech, which may indicate some inconsistencies in the annotation guidelines of the dataset. Regarding the linguistic features of misclassified tweets, we did not observe any relevant differences, except for a large number of hashtags in messages written by hate speech spreaders and the use of similes (a figurative language device) in non-hate speech spreaders.

The results of each model applied to the PoliticES 2022 task are shown in Table 4. Again, these results are obtained using the official test split. In contrast to PAN (see Table 3), the best results are not obtained with MarIA. MarIA does not actually give the best result for any of the traits evaluated. In the case of gender, XLM gives the best f1 score (84.268%), but all models evaluated achieve competitive results. For the trait "profession", both BETO and ALBETO give the same performance. In the case of the political ideology trait (left–right leaning), BERTIN achieves the best performance with an f1 score of 94.243%. From a multiclass perspective, BETO model achieves the best performance, with an f1 score of 79.341%. For all the other features, excluding gender, we observe that the performance of DistilBETO and ALBETO are similar to that of BETO. This is consistent with the results obtained with the PAN datasets. Comparing mBERT and XLM, the two models have similar performance in only two cases: occupation and political ideology, but XLM outperforms mBERT in gender identification and multiclass political ideology.

Next, we report the errors of the document-level error analysis applied to each of the PoliticES 2022 features. First, with respect to gender identification, we observe that the information gain of the LF over the misclassified subset differed between BETO and MarIA. The main differences between males and females of the misclassified tweets by BETO are related to stylometric features, including number of words and sentences. We also find differences in morphosyntactic features, such as the use of suffixes and common nouns (more for users tagged as male) and personal pronouns. However, MarIA's misclassified tweets contain differences in features related to pragmatics and lexis, including the use of similes, general lexis, and the use of mentions, both of which are more common in tweets written by women. Specific lexis related to performance are more common in men's tweets. Second, with respect to the identification of profession, we find that problematic tweets for BETO to discriminate between politicians and journalists are related to family topics — more common in politicians' tweets, and forms of politeness — more common in journalists' tweets. MarIA has difficulties with the stylometric and morphosyntactic features of the tweets, including readability, the number of quoted expressions and the lexicon related to timidity. Lastly, regarding the identification of political ideology (binary and multiclass), neither BETO nor MarIA have relevant linguistic features in the subset of misclassified tweets, except for slight differences in sentences ending with more than one exclamation mark, usually used by left-wing users to give more emphasis, and the use of similes excerpts used in the moderate right.

Given the above results, we can draw the following conclusions regarding RQ1 (Which language models are more appropriate for AP in Spanish):

[3] In English: Look what I've shared.

Table 4
Comparison of transformer-based features with the PoliticES 2022.

	Precision	Recall	f1-score	Precision	Recall	f1-score
	Gender			Profession		
BETO	84.659	**83.152**	83.811	**89.773**	**92.875**	**91.167**
DistilBETO	**87.875**	80.495	82.671	84.903	87.625	86.119
ALBETO	78.721	75.483	76.605	**89.773**	**92.875**	**91.167**
MarIA	78.076	80.133	78.694	87.853	88.875	88.346
BERTIN	84.333	81.099	82.323	87.338	90.250	88.643
mBERT	80.602	78.261	79.178	85.745	91.000	87.825
XLM	87.500	82.548	**84.268**	86.325	88.250	87.220
	Political ideology (binary)			Political ideology (multiclass)		
BETO	93.998	92.873	93.213	**82.373**	**78.569**	**79.341**
DistilBETO	93.998	92.873	93.213	77.737	69.492	69.814
ALBETO	92.530	90.789	91.225	68.175	69.460	67.829
MarIA	93.254	91.831	92.222	79.258	74.191	76.001
BERTIN	**94.243**	**94.243**	**94.243**	73.700	71.721	72.245
mBERT	92.502	92.160	92.296	76.985	65.566	67.075
XLM	92.344	92.654	92.364	76.712	75.257	74.495

- None of the architectures tested is a clear winner. While models based on RoBERTa outperform models based on BERT in all tasks proposed in the PAN data, models based on BERT outperform RoBERTa in all tasks proposed in the PoliticES data (except for political ideology in binary). This suggests that the best language model actually depends on the task and the dataset.
- The performance of distilled BETO and ALBETO is similar to that of BETO. However, none of the models is superior in all observed tasks.
- The performance of multilingual models is limited compared to models trained for Spanish. This is also consistent with the results obtained for other languages such as French [42] or Finnish [43].

4.2. RQ2 and RQ3. Feature integration

RQ 2 and 3 seek to determine whether feature integration is beneficial for improving performance and the best strategy for combining those features. In this work, the features obtained from the language models (one from BERT and one from RoBERTa) are combined with LF. As commented in Section 3.1.4, two strategies are evaluated: KI and EL.

Table 5 shows the results obtained for the PAN's datasets. According to 2019 (Nature), LF outperform the rest of the features trained in isolation. In this case, LF are useful because they can capture stylometric variables and content typically posted in social media environments related to hyperlinks or hashtags. However, the performance of the LF in other tasks is limited. In this task, only EL based on highest probability outperforms results of LF. With respect to 2019 (Gender), none of the feature integration strategies outperforms the results achieved by MarIA (RF). It is worth noting that RF generally outperforms both LF and SE. To detect limitations of LF and SE, we observe that the performance with the custom validation set is very significant, and both feature sets obtain scores higher than 90%. This means that models based on LF or SE do not learn to generalize and rather behave more like a random classifier. With respect to KI, the result is limited compared to that obtained with MarIA (RF). It is similar to the EL strategies evaluated, being the best strategy with the highest probability. The results of feature integration are limited due to the low performance of LF and SE. Regarding 2020 (fake news spreaders), RF is the best performer all individual feature sets. Feature integration based on KI is beneficial for detecting fake news spreaders. However, none of the EL strategies outperforms RF separately. For 2021 (hate speech spreaders), the best result from individual feature sets is obtained with MarIA (RF). In this task, unlike 2020 (Fake news spreaders), feature integration is not beneficial due to the limited performance of SE. However, both KI and all EL strategies achieve very similar results.

The best results obtained for the language models evaluated separately or in combination with other features are limited compared to the official results obtained in the PAN joint task. In 2019, Pizarro [35] achieved an accuracy of 93.3%, while our EL based on the highest probability achieved an accuracy of 91.78% distinguishing between humans and bots, but an important degradation of performance from an accuracy of 81.72% to 71.23% achieved with embeddings based on MarIA (RF) in the gender prediction task. Thus, classical features based on n-grams and traditional machine learning methods, such as SVM, seem to better capture discriminating features between men and women. In the English variant of the 2019 challenge [6], the accuracy obtained by the participants is similar between Spanish and English for identifying whether the writings are from bots, but slightly worse in Spanish than in English. Observing the results obtained in the common tasks of 2020 and 2021, the best systems in the official competition achieved similar accuracy to that we obtained in our experiments: 82% with KI on the Spanish part of the dataset in 2020 and 83% with RF embeddings in 2021. If we compare the results obtained in the official shared tasks ranking, the accuracy obtained in English is generally more limited than in Spanish.

In order to analyze the limited performance in detecting the multiclass political spectrum compared to the official results, we extract the confusion matrix of the BETO model (see Fig. 2). We observe that the model is effective in terms of precision and recall for each ideology, reaching a perfect precision and a recall of 70.58% when detecting right wing users. The lowest precision is

Table 5
Comparison of feature sets, separately or combined, with the 2019, 2020, and 2021 PAN shared tasks.

Feature set	Precision	Recall	f1-score	Precision	Recall	f1-score
	2019 (Nature)			2019 (Gender)		
LF	89.993	89.778	89.764	55.750	53.556	48.657
SE	87.806	87.500	87.475	52.058	51.778	50.079
BF	86.772	85.778	85.681	66.800	66.333	66.098
RF	88.182	87.944	87.926	**71.260**	**71.222**	**71.209**
KI	87.535	87.333	87.316	70.576	70.444	70.397
EL (MODE)	89.775	89.444	89.422	68.420	67.778	67.495
EL (MEAN)	89.587	89.444	89.435	68.648	68.222	68.040
EL (WEIGHTED)	89.775	89.444	89.422	68.420	67.778	67.495
EL (HIGHEST)	**92.032**	**91.778**	**91.765**	71.023	70.889	70.842
Feature set	2020 (Fake news spreaders)			2021 (Hate-speech spreaders)		
LF	75.161	75.000	74.960	73.770	73.000	72.780
SE	78.503	78.500	78.499	63.550	58.000	53.209
BF	80.012	80.000	79.998	77.083	76.000	75.758
RF	**82.116**	**82.000**	81.984	**83.013**	**83.000**	**82.998**
KI	82.051	**82.000**	**81.993**	77.083	76.000	75.758
EL (MODE)	80.527	80.500	80.496	78.373	77.000	76.718
EL (MEAN)	81.000	81.000	81.000	77.083	76.000	75.758
EL (WEIGHTED)	80.527	80.500	80.496	78.373	77.000	76.718
EL (HIGHEST)	80.048	80.000	79.992	77.083	76.000	75.758

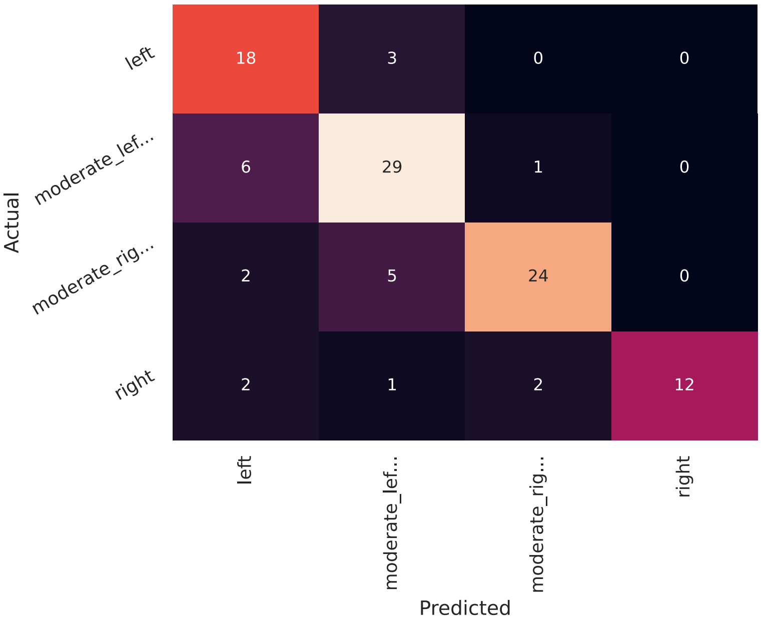

Fig. 2. Confusion matrix for the political ideology (multiclass) by BETO in PoliticES 2022.

64.28% when detecting left wing users. The confusion matrix also shows that no relevant misclassifications are made, i.e. between the left and the right, since the most frequently mistaken users are between the left and the moderate left.

Table 6 shows the results obtained for each feature in the PoliticES 2022 data to evaluate feature integration. There are two traits where the best results are obtained with feature integration: Gender and Political Ideology (binary). For gender, the best result is obtained with the Highest Probability strategy, reaching an f1 score of 86.515%, outperforming the best single model, BETO (BF). For political ideology (binary), the best feature combination strategy is KI, which outperforms EL. For the remaining features, occupation and political ideology (multiclass), the best overall scores are achieved by BETO (BF). For both traits, the best accuracy is obtained with feature integration: KI for occupation and EL for political ideology (multiclass).

Comparing the results of the evaluated language models and the official PoliticES 2020 ranking, we can see that the results are somewhat more limited: In fact, the winning team LosCalis [44], achieved the best macro average F1 score of three of the tasks, namely gender prediction (90.287%), occupation recognition (94.443%) and political ideology with a binary perspective (96.162%). The NLP-CIMAT team achieved the best macro averaged F1 score of 89.629% in political ideology with a multiclass perspective. The best results of the evaluated models and feature combination strategies are 86.515% for gender recognition (3.772%

Table 6
Evaluation of the feature sets, separately or in combination, with the PoliticES 2022.

Feature set	Precision	Recall	f1-score	Precision	Recall	f1-score
	Gender			Profession		
LF	61.823	60.024	60.317	78.382	85.250	80.361
SE	70.769	71.739	71.131	80.240	86.500	82.321
BF	84.659	83.152	83.811	89.773	**92.875**	**91.167**
RF	78.076	80.133	78.694	87.853	88.875	88.346
KI	79.173	81.522	79.808	**90.409**	91.500	90.936
EL (MODE)	83.368	82.428	82.857	88.818	90.875	89.776
EL (MEAN)	87.467	84.601	85.753	87.000	91.625	88.915
EL (WEIGHTED)	83.368	82.428	82.857	88.818	90.875	89.776
EL (HIGHEST)	**89.935**	**84.662**	**86.515**	81.250	87.125	83.326
Feature set	Political ideology (binary)			Political ideology (multiclass)		
LF	86.941	86.239	86.459	68.785	59.278	60.488
SE	81.378	81.469	80.951	71.386	72.242	71.637
BF	93.998	92.873	93.213	82.373	**78.569**	**79.341**
RF	93.254	91.831	92.222	79.258	74.191	76.001
KI	**95.286**	**95.121**	**95.195**	77.917	69.988	69.874
EL (MODE)	93.998	92.873	93.213	**82.680**	78.289	78.927
EL (MEAN)	93.998	92.873	93.213	79.172	73.911	75.130
EL (WEIGHTED)	93.998	92.873	93.213	81.356	77.099	77.873
EL (HIGHEST)	93.998	92.873	93.213	75.579	66.666	68.326

below), 91.167% (3.276% below), 95.195% (0.967% below), and 79.341% (10.287% below). Hence, the main difference between the LosCalis and NLP-CIMAT teams is that both performed some kind of data augmentation steps. LosCalis included more tweets from Spanish politicians and that the new profiles do not include a fixed number of tweets. NLP-CIMAT extended BETO with a new dataset from a dataset focused on sentiments on Twitter and documents related to social media and politics.

Given the above results, we draw the following conclusions to answer RQ2 and RQ3 (regarding the appropriateness of applying feature integration strategies to AP tasks and which techniques are most appropriate):

- The integration of features does not always improve the overall accuracy. Only in 4 of the 8 tasks evaluated, the best f1 score is obtained with one of the two techniques evaluated. In 2 cases, the best accuracy is obtained with feature integration.
- When feature integration proved beneficial, the two strategies evaluated (KI and EL) reached a tie. EL is the best strategy in 2019 (nature, PAN) and gender (PoliCorpus), while KI is the best strategy in 2020 (fake news spreaders, PAN) and political ideology (binary, PoliCorpus).
- The best strategy in EL is *highest probability* which gives the best result in all cases where EL is most effective.

4.3. RQ4. Interpretability

In this section we focus on the interpretability of the LFs. First, we compute the information gain of each LF for the PAN (see Fig. 3) and PoliticES 2022 (see Fig. 4) datasets.

In terms of PAN 2019 (nature), there are strong differences between the bot and human classes. For example, the use of interjection, demonstrative personal pronouns, and exclamatory sentences is more common among humans. But the use of hyperlinks and hashtags is more common among bots. For 2019 (gender), the differences are more difficult to discern. The most informative features are error-related, such as misspellings. They can also be stylistic, such as the use of hyphens or very short words (two characters long). Regarding PAN 2020 (fake news spreaders), due to the subject matter, common written messages from fake news spreaders is about human body parts (e.g. messages about hand washing) and professions (e.g. doctors or politicians). This dataset also includes more farewell messages, suggesting that fake news spreaders intend to end a conversation in a polite way to appear more friendly. Regarding PAN 2021 (hate speech spreaders), there are less obvious differences compared to fake news spreaders. In fact, hate-speech spreaders make more use of figurative forms, such as hyperboles. The presence of demonyms, which typically refer to people from other regions and direct hatred towards specific groups, is also found to be relevant. Connectors, which are a type of discourse marker, are another relevant feature to connect different ideas within the same speech.

From PoliticES 2022 (see Fig. 4), the most significant features for gender identification are the following: number misspelled words, the number of words in uppercase, morphological features, (e.g. the percentage of pronouns, prepositions, and verbs) and lexis concerning human beings. The most relevant features for distinguishing male and female are related to stylometry and readability, which suggest that the length of sentences and the average length of words are different. The use of social media elements, such as the percentage of hashtags, is also quite relevant. As expected, the percentage of hashtags is much higher for journalists than for politicians. The remaining relevant features are related to stylometry. In terms of political ideology, the most relevant features to distinguish between the left and the right are morphological and stylometric features (including the average length of the texts), the percentage of feminine words and words in the singular. The use of pronouns is also relevant. From a multiclass perspective, political ideology also depends on stylometric and morphological features, but also includes the percentage of misspelled words and features related to psychological processes, such as the percentage of generic positive words.

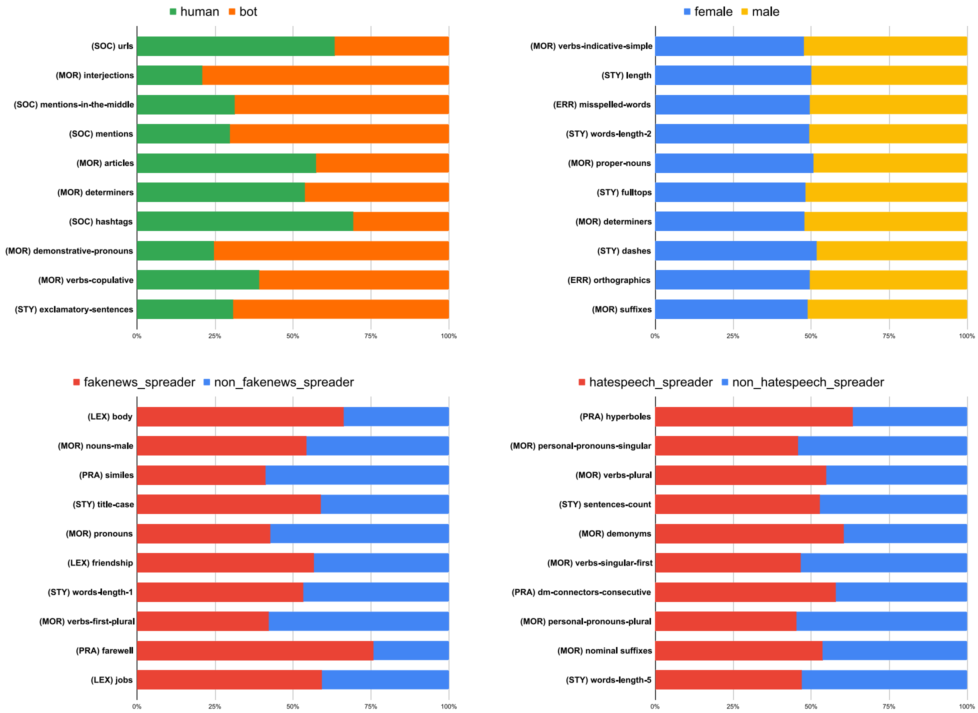

Fig. 3. Information gain of the PAN shared tasks.

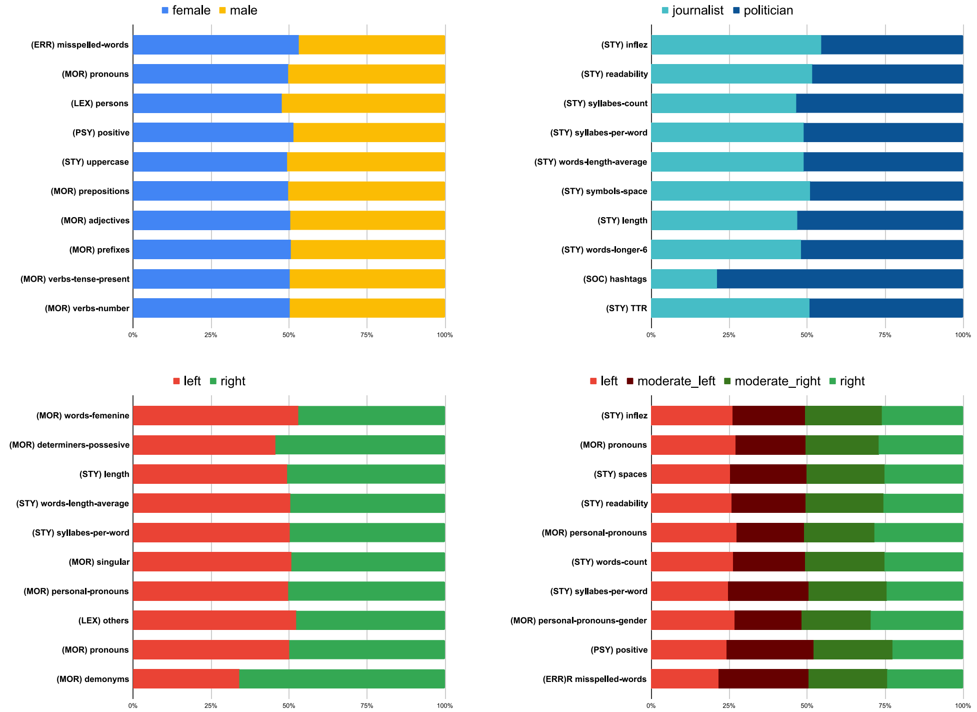

Fig. 4. Information gain of the PoliticES 2022 dataset.

Table 7
Comparison of training time in hours of the fine-tuning process, inference time in minutes of the test split, and size (gigabytes) of the best model (in gigabytes) for each LLM with the 2019, 2020, and 2021 PAN shared tasks.

LLM	Size	Training	Inference	Size	Training	Inference
	2019 (Nature)			2019 (Gender)		
ALBETO	0.13	14.23	16.363	0.13	9.59	7.251
BERTIN	1.39	20.09	15.132	1.39	11.10	7.580
BETO	1.23	23.48	13.707	1.23	13.78	5.670
DistilBETO	0.75	12.41	7.401	0.75	4.79	3.106
MarIA	1.39	12.54	15.130	1.39	7.81	7.568
mBERT	1.87	12.54	12.836	1.87	12.04	5.644
XLM	3.11	19.53	13.822	3.11	12.04	6.170
	2020 (Fake news spreaders)			2021 (Hate-speech spreaders)		
ALBETO	0.13	1.61	0.700	0.13	1.02	0.409
BERTIN	1.39	1.38	0.778	1.39	1.94	0.821
BETO	1.23	1.73	0.711	1.23	1.76	0.391
DistilBETO	0.75	0.81	0.443	0.75	0.73	0.271
MarIA	1.39	1.69	0.776	1.39	1.55	0.821
mBERT	1.87	1.37	0.650	1.87	0.98	0.391
XLM	3.11	3.32	0.738	3.11	1.62	0.526

Table 8
Comparison of training time in hours of the fine-tuning process, inference time in minutes of the test split, and size (Gigabytes) of the best model (in gigabytes) for each LLM with the PoliticES 2022 shared task for each trait.

LLM	Size	Training	Inference	Size	Training	Inference
	Gender			Profession		
ALBETO	0.13	1.42	0.797	0.13	2.28	0.797
BERTIN	1.39	1.47	0.536	1.39	1.43	0.538
BETO	1.23	2.12	0.663	1.23	1.96	0.661
DistilBETO	0.75	0.84	0.412	0.75	0.97	0.411
MarIA	1.39	1.17	0.535	1.39	2.65	0.536
mBERT	1.87	1.86	0.557	1.87	1.48	0.555
XLM	3.11	1.35	0.624	3.11	2.28	0.627
	Political ideology (binary)			Political ideology (multiclass)		
ALBETO	0.13	1.37	0.799	0.13	2.10	0.798
BERTIN	1.39	1.15	0.536	1.39	2.31	0.537
BETO	1.23	1.32	0.659	1.23	2.50	0.659
DistilBETO	0.75	0.95	0.411	0.75	1.01	0.409
MarIA	1.39	2.08	0.536	1.39	1.25	0.537
mBERT	1.87	1.43	0.556	1.87	1.99	0.556
XLM	3.11	3.27	0.625	3.11	1.57	0.626

In light of the above results, we derive the following conclusions to answer RQ4 (regarding the interpretability of the models):

- The presence of hyperlinks and hashtags are relevant features for distinguishing bots from humans. However, these features are not particularly relevant for identifying harmful users (spreaders of fake news or hate speech).
- Stylometrics is the linguistic category most relevant to AP tasks. Most of these features are related to document length and average word length.
- Morphology is another linguistic marker relevant to conducting AP. However, the specific LF is different for each scenario. Such markers include interjections, articles, determiners, suffixes, or verbs.
- Features related to lexis, which are used to identify the topic of the conversation, are also relevant in some scenarios.
- Psycholinguistic processes, which include features related to emotion and sentiment analysis, are not particularly relevant in AP.

4.4. RQ5. Performance

In this section, we evaluate the performance in terms of memory use and time required to train and evaluate the test split of each of the evaluated LLMs.

All experiments are ran on a cluster equipped with a GeForce RTX 3090 with 24 GB of RAM, and are queued using SLURM Workload Manager. We reserve the GPU and an AMD EPYC 7413 24-core processor, where we reserve 10 CPU cores.

Table 7 shows the memory usage of the best model, the hours required to fine-tune each LLM at the document level and the minutes required to evaluate the set split. Note that there are large differences between the datasets. However, the amount of memory required is the same because it does not depend on the size of the dataset. But the training time required varies greatly depending on the size of the dataset. The PAN 2019 datasets require larger training, with some of the models requiring almost a day of training such as the heavy models BETO and BERTIN. MarIA requires significantly less training time, although it is also a heavy model and is based on RoBERTa (like BERTIN). The lightweight models, ALBETO and DistilBETO require similar training times in 2019 to detect whether the tweet is from a human or a bot. However, for gender detection, the training time is reduced from 14 h for ALBETO to 4.49 h for DistilBETO. For the PAN 2020 and 2021 datasets, the training times are drastically reduced due to the reduced size of the datasets, as the 2019 dataset is about 2400 different profiles to the 500 users in 2020 and 400 in 2021. The training times for the PAN 2020 and 2021 datasets are also reduced depending on the model, but in general all models require about one hour for fine-tuning. In terms of inference time, DistilBETO is the fastest model. ALBETO, the other lightweight model, required the largest inference time in 2019 (Nature) and generally large inference times in the remaining tasks, except in 2021 (Hate-speech spreaders). The comparison of BETO, MarIA and BERTIN, as the three Spanish LLMs, shows that BERTIN is the slowest model in terms of inference. Among the multilingual models, the slowest inference times are obtained with XLM.

Next, Table 8 compares the performance in terms of model size in gigabytes, the training time for fine-tuning in hours, and the inference time of the test split for the four traits of PoliticES 2022. Similar to PAN, and as expected, the size of the models does not vary too much after pre-training. However, the training times depend on the task, despite the fact that the size of the dataset is constant. This is due to the dynamic selection of hyperparameters during tuning, which affects the batch size or the number of epochs. Summing up all the training times, profession is the feature that took most to find the best model for, with a little more than 13 h, and the fastest training time is for gender detection. Political ideology (multiclass) did not reach the maximum, with 12.73 h, although it was the only multiclass experiment. Regarding the lightweight models, there are important differences in the training times in ALBETO, which took 1.37 h for the binary multiclass, but 2.28 h for profession. DistilBETO, on the other hand, shows more stable training times regardless of the feature. In terms of inference time, DistilBETO is the fastest and ALBETO is the slowest model. It should be noted that all these experiments used exactly the same dataset, so the inference time is more equal among the features. The heavy models based on the RoBERTa architecture (BERTIN and MarIA) are faster than BETO based on the BERT architecture. However, the multilingual BERT achieved faster inference times.

In light of the above results, we derive the following conclusions to answer RQ5 (on the performance of the models):

- The time required to train an AP model can be a barrier in certain scenarios, especially in datasets where there are many users and many writings from them.
- The performance of lightweight models, especially those based on distillation, can significantly improve training and inference times.
- ALBETO uses less memory than DistilBETO, but training and inference times are longer.
- The cost of disk space for multilingual models is a factor to consider, although it does not affect the longer training and inference times.
- Language model memory after fine-tuning does not depend on the size of the dataset.

5. Conclusions and further work

In this work, we investigate the performance of novel language models based on transformers for performing AP tasks in Spanish. These models include multilingual and language-specific models, as well as lightweight models based on distillation. The integration of these models with other feature sets such as LF is also evaluated. Our results indicate that the best language model and feature integration improves the overall performance of AP, but this depends on the task. Our recommendation is to evaluate different language models and combinations to get the best result. Our experiments also show that the use of distilled models achieves similar results to the large models, making them suitable for industry and real-time applications. However, multilingual models achieve limited results compared to the other models. In terms of feature integration, the two strategies evaluated seem to produce identical results. Going forward, our results have significant implications for the design and implementation of KBS. By leveraging the insights gained from AP, such systems could improve their ability to effectively organize and retrieve information effectively, ultimately improving the user experience and decision-making processes.

As promising lines of future research to improve AP in Spanish, we will focus on the compilation of multimodal datasets that include other sources such as audio and video content typically shared by users in social networks. Another interesting line of research will be to combine AP with emotion detection and to identify depression and suicide risk. It is worth noting that Twitter's privacy policy prohibits profiling users based on *sensitive personal information*. This includes political affiliation and beliefs as well as racial or ethnic origin. Thus, a line of research on authorship analysis can also take into account the privacy of the data collected and its use to detect malicious users who spread fake or hate speech in order to improve the overall social benefits of social networks. Of course, any further research direction should ensure data anonymization and preserve users ownership rights over their own texts. With all their clear benefits, we believe that AP models should always be used as a guide for human decision makers, rather than a decision-making tool that makes full decisions such as canceling a user's account. In addition, future research could explore the integration of demographic and psychographic traits to improve item recommendation systems based on review data [45] or user preferences [46], thereby increasing personalization and relevance for end users.

CRediT authorship contribution statement

José Antonio García-Díaz: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Ghassan Beydoun:** Writing – review & editing, Writing – original draft, Visualization, Resources, Investigation, Data curation. **Rafel Valencia-García:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way of making Europe.

References

[1] G. Faye, W. Ouerdane, G. Gadek, S. Gahbiche, S. Gatepaille, A novel hybrid approach for text encoding: Cognitive attention to syntax model to detect online misinformation, Data Knowl. Eng. 148 (2023) 102230.

[2] I. Dimitriadis, G. Dialektakis, A. Vakali, CALEB: A conditional adversarial learning framework to enhance bot detection, Data Knowl. Eng. 149 (2024) 102245.

[3] M.G. Ayadi, H. Mezni, R. Alnashwan, H. Elmannai, Effective healthcare service recommendation with network representation learning: A recursive neural network approach, Data Knowl. Eng. 148 (2023) 102233.

[4] G. Sukanya, J. Priyadarshini, Modified Hierarchical-attention network model for legal judgment predictions, Data Knowl. Eng. 147 (2023) 102203.

[5] D. Grissa, E. Andonoff, C. Hanachi, Discovering and evaluating organizational knowledge from textual data: Application to crisis management, Data Knowl. Eng. 148 (2023) 102237.

[6] F.M.R. Pardo, P. Rosso, Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling in Twitter, in: L. Cappellato, N. Ferro, D.E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, in: CEUR Workshop Proceedings, vol. 2380, CEUR-WS.org, 2019, pp. 1–7, URL: http://ceur-ws.org/Vol-2380/paper_263.pdf.

[7] F.M.R. Pardo, A. Giachanou, B. Ghanem, P. Rosso, Overview of the 8th author profiling task at PAN 2020: Profiling fake news spreaders on Twitter, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, in: CEUR Workshop Proceedings, vol. 2696, CEUR-WS.org, 2020, pp. 1–18, URL: http://ceur-ws.org/Vol-2696/paper_267.pdf.

[8] F. Rangel, G.L.D. la Peña Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling hate speech spreaders on Twitter task at PAN 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - To - 24th, 2021, in: CEUR Workshop Proceedings, vol. 2936, CEUR-WS.org, 2021, pp. 1772–1789, URL: http://ceur-ws.org/Vol-2936/paper-149.pdf.

[9] R.O. Bueno, B. Chulvi, F. Rangel, P. Rosso, E. Fersini, Profiling irony and stereotype spreaders on Twitter (IROSTEREO). overview for PAN at CLEF 2022, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - To - 8th, 2022, in: CEUR Workshop Proceedings, vol. 3180, CEUR-WS.org, 2022, pp. 2314–2343, URL: http://ceur-ws.org/Vol-3180/paper-185.pdf.

[10] M.Á. Álvarez-Carmona, E. Guzmán-Falcón, M. Montes-y Gómez, H.J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, A. Rico-Sulayes, Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in mexican spanish tweets, in: Notebook Papers of 3rd Sepln Workshop on Evaluation of Human Language Technologies for Iberian Languages (Ibereval), Seville, Spain, volume 6, 2018, pp. 1–28.

[11] M.E. Aragón, M.A.A. Carmona, M. Montes-y Gómez, H.J. Escalante, L.V. Pineda, D. Moctezuma, Overview of MEX-A3T at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets., in: IberLEF@ SEPLN, 2019, pp. 478–494.

[12] J.A. García-Díaz, S.M. Jiménez Zafra, M.T. Martín Valdivia, F. García-Sánchez, L.A. Ureña López, R. Valencia García, Overview of politices 2022: Spanish author profiling for political ideology, Procesamiento del Lenguaje Nat. 69 (2022) 265–272.

[13] M. Chinea-Rios, T. Müller, G.L. De la Peña Sarracén, F. Rangel, M. Franco-Salvador, Zero and few-shot learning for author profiling, in: International Conference on Applications of Natural Language To Information Systems, Springer, 2022, pp. 333–344.

[14] R. López-Santillán, L.C. González, M. Montes-y Gómez, A.P. López-Monroy, When attention is not enough to unveil a text's author profile: Enhancing a transformer with a wide branch, Neural Comput. Appl. (2023) 1–20.

[15] M. Polignano, M. de Gemmis, G. Semeraro, Contextualized BERT sentence embeddings for author profiling: The cost of performances, in: Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part IV 20, Springer, 2020, pp. 135–149.

[16] J.A. García-Díaz, P.J. Vivancos-Vicente, A. Almela, R. Valencia-García, UmuTextStats: A linguistic feature extraction tool for spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 6035–6044.

[17] D. Yenicelik, F. Schmidt, Y. Kilcher, How does BERT capture semantics? A closer look at polysemous words, in: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 2020, pp. 156–162.

[18] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/n19-1423.

[19] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, PML4DC at ICLR 2020 (2020).

[20] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and DistilBETO: Lightweight spanish language models, 2022, arXiv preprint arXiv:2204.09145.

[21] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, 2019, CoRR URL: http://arxiv.org/abs/1909.11942. arXiv:1909.11942.

[22] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C.P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Nat. 68 (2022) 39–60, URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.

[23] J.D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, BERTIN: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Nat. 68 (2022) 13–23, URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2019, CoRR arXiv:1911.02116. URL: http://arxiv.org/abs/1911.02116.

[25] A. Chiorrini, C. Diamantini, A. Mircoli, D. Potena, Emotion and sentiment analysis of tweets using BERT, in: EDBT/ICDT Workshops, volume 3, 2021, pp. 1–7.

[26] E. Puraivan, R. Venegas, F. Riquelme, An empiric validation of linguistic features in machine learning models for fake news detection, Data Knowl. Eng. 147 (2023) 102207.

[27] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, 2018, CoRR URL: http://arxiv.org/abs/1802.06893. arXiv:1802.06893.

[28] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3980–3990, http://dx.doi.org/10.18653/v1/D19-1410.

[29] F.M.R. Pardo, P. Rosso, M. Koppel, E. Stamatatos, G. Inches, Overview of the author profiling task at PAN 2013, in: P. Forner, R. Navigli, D. Tufis, N. Ferro (Eds.), Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013, in: CEUR Workshop Proceedings, vol. 1179, CEUR-WS.org, 2013, pp. 1–13, URL: http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-RangelEt2013.pdf.

[30] F.M.R. Pardo, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, W. Daelemans, Overview of the author profiling task at PAN 2014, in: L. Cappellato, N. Ferro, M. Halvey, W. Kraaij (Eds.), Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014, in: CEUR Workshop Proceedings, vol. 1180, CEUR-WS.org, 2014, pp. 898–927, URL: http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-RangelEt2014.pdf.

[31] F.M.R. Pardo, F. Celli, P. Rosso, M. Potthast, B. Stein, W. Daelemans, Overview of the 3rd author profiling task at PAN 2015, in: L. Cappellato, N. Ferro, G.J.F. Jones, E. SanJuan (Eds.), Working Notes of CLEF 2015 - Conference and Labs of the Evaluation Forum, Toulouse, France, September 8-11, 2015, in: CEUR Workshop Proceedings, vol. 1391, CEUR-WS.org, 2015, pp. 1–40, URL: http://ceur-ws.org/Vol-1391/inv-pap12-CR.pdf.

[32] F.M.R. Pardo, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, B. Stein, Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations, in: K. Balog, L. Cappellato, N. Ferro, C. Macdonald (Eds.), Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum, ÉVora, Portugal, 5-8 September, 2016, in: CEUR Workshop Proceedings, vol. 1609, CEUR-WS.org, 2016, pp. 750–784, URL: http://ceur-ws.org/Vol-1609/16090750.pdf.

[33] F.M.R. Pardo, P. Rosso, M. Potthast, B. Stein, Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017, in: CEUR Workshop Proceedings, vol. 1866, CEUR-WS.org, 2017, pp. 1–26, URL: http://ceur-ws.org/Vol-1866/invited_paper_11.pdf.

[34] F.M.R. Pardo, P. Rosso, M. Montes-y-Gómez, M. Potthast, B. Stein, Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in Twitter, in: L. Cappellato, N. Ferro, J. Nie, L. Soulier (Eds.), Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018, in: CEUR Workshop Proceedings, vol. 2125, CEUR-WS.org, 2018, pp. 1–38, URL: http://ceur-ws.org/Vol-2125/invited_paper_15.pdf.

[35] J. Pizarro, Using N-grams to detect bots on Twitter, in: CLEF (Working Notes), 2019, pp. 1–10.

[36] J. Pizarro, Using N-grams to detect fake news spreaders on Twitter., in: Working Notes for CLEF 2020 Conference, Online, volume 2696, 2020, pp. 1–8.

[37] M. Siino, E. Di Nuovo, I. Tinnirello, M. La Cascia, Detection of hate speech spreaders using convolutional neural networks., in: CLEF (Working Notes), 2021, pp. 2126–2136.

[38] J.A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020, Future Gener. Comput. Syst. 130 (2022) 59–74.

[39] C.G. Holgado, A. Sinha, HalBERT at PoliticEs 2022: Are machine learning algorithms better for author profiling? in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, a Coruna, Spain. D. Moctezuma, and V. Muniz-SáNchez, 2022, pp. 1–13.

[40] A. Mosquera, Alejandro Mosquera at PoliticEs 2022: Towards robust spanish author profiling and lessons learned from adversarial attacks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, a Coruna, Spain. D. Moctezuma, and V. Muniz-SáNchez, 2022, pp. 1–8.

[41] E. Villa-Cueva, I. González-Franco, F. Sanchez-Vega, A.P. López-Monroy, NLP-CIMAT at politices 2022: Politibeto, a domain-adapted transformer for multi-class political author profiling, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, a Coruna, Spain, 2022, pp. 1–13.

[42] L. Martin, B. Muller, P.J.O. Suárez, Y. Dupont, L. Romary, É.V. de la Clergerie, D. Seddah, B. Sagot, CamemBERT: a tasty french language model, in: ACL 2020-58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7203–7219.

[43] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: BERT for finnish, 2019, arXiv preprint arXiv:1912.07076.

[44] S.S. Carrasco, R.C. Rosillo, LosCalis at PoliticEs 2022: Political author profiling using BETO and maria, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, a Coruna, Spain, volume 3202, CEUR-WS.org, 2022.

[45] J. Liu, T. Li, Z. Yang, D. Wu, H. Liu, Fusion learning of preference and bias from ratings and reviews for item recommendation, Data Knowl. Eng. (2024) 102283.

[46] D. Vandic, L.J. Nederstigt, F. Frasincar, U. Kaymak, E. Ido, A framework for approximate product search using faceted navigation and user preference ranking, Data Knowl. Eng. 149 (2024) 102241.

José Antonio García-Díaz received the B.Sc., M.Sc., and Ph.D. (2022) in computer science from the University of Murcia, Espinardo, Spain. He is a member of the TECNOMOD (Knowledge Modeling, Processing and Management Technologies) Research Group. His research interests include Natural Language Processing and infodemiology.

Ghassan Beydoun received the degree in computer science and the Ph.D. degree in knowledge systems from the University of New South Wales. He is currently a Professor of information systems with the School of Information, Systems, and Modeling, University of Technology Sydney. He has authored more than 150 papers in international journals and conferences. His research projects are sponsored by the Australian Research Council, the NSW State Government, and the private sector. He investigates the best uses of models in developing methodologies for distributed intelligent systems. His other research interests include cloud adoption, disaster management, decisions support systems, and their applications.

Rafael Valencia-García received the B.E., M.Sc., and Ph.D. degrees in Computer Science from the University of Murcia, Espinardo, Spain. He is currently a Full Professor with the Department of Informatics and Systems, University of Murcia. His main research interests are natural language processing, Semantic Web and recommender systems. He has participated in more than 35 research projects. He has published over 150 articles in journals, conferences, and book chapters, 50 of them in JCR-indexed journals. He is the author or coauthor of several books. He has been guest editor of five JCR-indexed journals (CSI, IJSEKE, JRPIT, JUCS, SCP).