Future Generation Computer Systems 130 (2022) 59–74

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020

José Antonio García-Díaz [a], Ricardo Colomo-Palacios [b], Rafael Valencia-García [a,*]

[a] Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
[b] Faculty of Computer Sciences, Østfold University College, Halden, Norway

ARTICLE INFO

Article history:
Received 14 April 2021
Received in revised form 28 November 2021
Accepted 15 December 2021
Available online 21 December 2021

Keywords:
Authorship analysis
Author profiling
Authorship attribution
Linguistic features
Natural language processing

ABSTRACT

In general, people are usually more reluctant to follow advice and directions from politicians who do not have their ideology. In extreme cases, people can be heavily biased in favour of a political party at the same time that they are in sharp disagreement with others, which may lead to irrational decision making and can put people's lives at risk by ignoring certain recommendations from the authorities. Therefore, considering political ideology as a psychographic trait can improve political micro-targeting by helping public authorities and local governments to adopt better communication policies during crises. In this work, we explore the reliability of determining psychographic traits concerning political ideology. Our contribution is twofold. On the one hand, we release the PoliCorpus-2020, a dataset composed by Spanish politicians' tweets posted in 2020. On the other hand, we conduct two authorship analysis tasks with the aforementioned dataset: an author profiling task to extract demographic and psychographic traits, and an authorship attribution task to determine the author of an anonymous text in the political domain. Both experiments are evaluated with several neural network architectures grounded on explainable linguistic features, statistical features, and state-of-the-art transformers. In addition, we test whether the neural network models can be transferred to detect the political ideology of citizens. Our results indicate that the linguistic features are good indicators for identifying fine-grained political affiliation, they boost the performance of neural network models when combined with embedding-based features, and they preserve relevant information when the models are tested with ordinary citizens. Besides, we found that lexical and morphosyntactic features are more effective on author profiling, whereas stylometric features are more effective in authorship attribution.

1. Introduction

Online services and products can be customised to our manner by incorporating psychographic traits concerning our personality, attitude, and lifestyle from our digital footprints [1]. For example, travel recommendation systems can offer exciting adventures or relaxing experiences according to the customers' personality. Video games can offer different game experiences based on whether the player is more or less daring. Computational advertising is another field that can benefit from incorporating psychological data by means of incorporating customers' feelings and beliefs [2].

Political ideology can be considered as a psychographic trait that can be used to understand individual and social behaviour, including moral and ethical values as well as inherent attitudes, appraisals, biases, and prejudices [3]. The relationship between personality traits and political ideology was analysed in [4]. From data gathered from 21 countries, the author concluded that political ideology is linked to the big five personality traits. Specifically, conscientiousness was strongly correlated with the right wing, whereas openness to experience and agreeability were notably more correlated to the left wing and, in a minor degree, extraversion. Nevertheless, the author found that the results vary among countries, especially based on the level of prosperity.

Political ideology has a great influence on society. Similar to other psychographic traits, such as personality, our political ideology can guide our day-to-day decisions. However, these decisions are made both consciously and unconsciously, as our ideology is influenced by social and cultural background including religious beliefs as well as cultural and family traditions. For example, in [5] the authors found a correlation between political ideology and the attitude to vaccination campaigns for infectious diseases. Another example is [6], in which the authors measure the association between political party affiliations and environmental concerns. In this sense, direct applications for political ideology

* Corresponding author.
E-mail addresses: joseantonio.garcia8@um.es (J.A. García-Díaz), ricardo.colomo-palacios@hiof.no (R. Colomo-Palacios), valencia@um.es (R. Valencia-García).

https://doi.org/10.1016/j.future.2021.12.011

identification are political forecasting [7] and political micro-targeting (PMT) [8], in which politicians and campaign managers can leverage voter behaviour data and thereby target ad campaigns to be more persuasive to a specific type of voter. It is worth noting that compiling political ideology of citizens involves strong ethical restrictions because affiliation ideas are strongly protected in some countries by personal data protection laws. In this sense, the Facebook–Cambridge Analytica data scandal must be cited. During the 2016 US elections, millions of psychological profiles from Facebook users were inappropriately collected for tailored advertising. This misconduct was harshly criticised and resulted in various economic sanctions [7,9]. Besides the inappropriate use of private information, the clandestine collection of this information can be used to create hoaxes that influence malleable voters and encourages the abstention of the political adversary, resulting in a weakened democracy. Moreover, affiliation to a certain group, such as a religious group or a political party, implies that the individuals share on a great extent the beliefs and identity of the group. In some pernicious environments, social groups can promote self-ideas whereas other points of view are dismissed and ridiculed. This phenomenon is known as the echo chamber effect and causes mutual distrust of the group perceived as "us vs them". Polarisation may lead to the rejection of others' ideas and points of view by using logical fallacies.

As far as our knowledge goes, few works have considered political ideology as a psychographic trait. However, determining traits related to political ideology and affiliation could help to understand social behaviour and, consequently, schedule better social policies. For example, in [10] the authors examined how political polarisation affected the formation of beliefs and their consequences during the COVID-19 pandemic. Their study concluded that confidence in certain actions, such as social distance, was driven primarily by political ideology, but also by age range and geographical area. Similar conclusions are stated in [11]. In their study, the authors assess public policies to control the incidence of COVID-19. They indicate that policies to control COVID-19 were *significantly influenced by political pressures*. Especially during the early stages of the pandemic, certain governments reshaped their policies according to public opinion.

In this work we explore the reliability of incorporating political affiliation traits to authorship analysis tasks. For this purpose, we first compile a corpus composed of Spanish politicians' tweets posted in 2020 including members of the government, senators, deputies, presidents of autonomous communities, mayors, councillors, advisers, and former politicians. Next, we categorised their political spectrum, which is a manner of classifying political ideologies, as a binary classification problem (left vs right wing) and as a multiclass problem (left, moderate left, moderate right, and right wing). Next, we evaluate the PoliCorpus-2020 from an Authorship Analysis (AA) perspective and examine whether political ideology traits can be transferred to average citizens. This study is grounded on the usage of linguistic features that can characterise an author's writing style. The linguistic features are evaluated separately and combined with state-of-the-art embeddings at sentence and word level.

The main contributions of this work can be summarised as:

- **Development of the Spanish PoliCorpus-2020**, based on contributions on Twitter of Spanish politicians during 2020. The political actors are classified by demographic and psychographic traits.
- **Development and evaluation of two AA case studies**. On the one hand, the main study deals with author profiling, focusing on the identification of the political spectrum from a binary and multi-class perspective. Besides, we examine whether the generated models can be transferred by evaluating these methods with a test set composed of journalists that have not been used during training. As a secondary objective, we also evaluate demographic traits such as age range and gender. On the other hand, we conduct a case study on authorship attribution, focused on unveiling the identify of an author basing on their writings.
- **Evaluation of the feature sets**. We evaluate linguistic features, character and word n-grams, and word and sentence embeddings. For this, we test different neural network architectures to determine the accuracy, quality, performance, and the interpretability of the resulting models. We argue that the use of linguistic features specifically designed for Spanish combined with contextual embeddings have not been widely explored yet. We consider that linguistic features that capture stylometric and morphosyntactic clues from writings can enhance the performance and interpretability in author profiling and authorship attribution tasks.

The remainder of this paper is organised as follows. First, Section 2 describes recent works and approaches concerning the identification of political ideology and authorship analysis in the literature and shared tasks. Next, Section 3 contains the development process of the PoliCorpus-2020. Section 4 describes the models and techniques employed in the studies conducted. Section 5 depicts and discusses the results of each experiment and, finally, Section 6 presents summaries of the main findings of this work and provides promising future research directions.

2. State of the art

In this work we examine psychographic traits related to political affiliation. Accordingly, in this section we review recent related work extracting political affiliation from writings (see Section 2.1) and we examine recent work related to AA tasks, paying especial attention to author profiling (Section 2.2).

2.1. Identification of the political ideology

Political ideology, as well as other personality traits, can be inferred from user's digital footprints [12]. Typically, personality prediction has focused on a general framework known as the Big-five personality traits, namely openness, conscientiousness, extroversion, agreeableness, and neuroticism. Different works on personality prediction rely on user profiles and their contributions on social networks. For example, in [13] the authors combined the Big-five traits with dark-triad traits (narcissism, machiavellianism, and psychopathy) to improve cyberbullying detection. In [14], the authors developed a stance-detection system in order to predict whether a user will perform an action or not. Their results indicate that the combination of features regarding personality traits were the most relevant. Although these studies do not mention author profiling directly, these personality traits are obtained from contributions of users in social networks. In [8], the authors deal with political micro-targeting (PMT) that entails compiling personal data on social networks to send them specific political messages. Specifically, the authors focus on determining extraversion, which implies positive traits such as sociability or assertiveness.

As far as our knowledge goes, few researches have been conducted to determine political ideology traits. In [15], the authors examine ideological and organisational affiliation as two independent problems. For this purpose, they examine two corpora related to religious and political documents in Arabic. The first dataset consisted of 552 documents manually labelled with four organisations, whereas the second one consisted of 1485 documents labelled with four ideological streams. The authors

found that the usage of stylistic and content features can be used effectively, achieving near-perfect accuracy. They found a strong correlation between certain keywords that matches certain groups and ideologies but fewer function words. A similar work from [15] was addressed by [16], in which the authors also examined ideological and organisational affiliation from Arabic documents. In this work, the authors took into account statistical features as well. In [17], the authors incorporate traits regarding political preference and income level by creating a corpus labelled by employing distant supervision methods based on users' biography from tweets written in Dutch. The authors collected a total of 17,000 Twitter users. It is worth mentioning that the authors linked the users' accounts with only one political party based only on the textual description. We consider, however, that this strategy can cause that the dataset could have a relevant number of misclassified users because, as far as our knowledge goes, the authors did not consider specific Natural Language Processing (NLP) approaches for understanding users' bio. Therefore, a user may say (s)he dislikes a political party and be wrongly classified as supportive. In [18], the authors examined age range, gender, and political affiliation from Swedish politicians from their speeches in the parliament during 2003 and 2010. They evaluate different feature sets at author-level and document-level with Support Vector Machines (SVM), achieving an accuracy of 81.2% for gender prediction, 89.4% for binary political affiliation, and 78.9% for age range classification. In [19], the authors focused on extracting bias towards political ideology at document-level. To do this, they compile a dataset of 34,737 articles annotated as biased to the left, centre, or right wing and they develop a deep-learning system based on adversarial media and a triplet loss, focused on determining the bias of the documents. The authors evaluate this system with news from media sites that were not in the training split.

Recently, the authors of [20] published a corpus based on political alignment compiled from journalists in Argentina. The corpus is organised as a classification problem, discerning between journalists that are pro-government or opposition. The authors compared topics and linguistic features using Latent Dirichlet Allocation (LDA) and Linguistic Inquiry and Word Count (LIWC). Although this work is the most similar to our proposal, it has important differences. First, the dataset used in the present study operationalises political ideology as for or against, but not along the traditional left and right axes. Second, the number of different accounts is limited, being only 10 journalists, that makes it difficult to draw more representative conclusions.

2.2. Authorship analysis

Authorship Analysis is a relevant field of study related to Information Retrieval (IR) and NLP focused on retrieving information from people based on their writings [21]. AA has applications in forensic linguistics as it can provide linguistic evidence that unmasks the real author of an anonymous threatening document, determine whether there has been plagiarism between two disputed documents, or determine if a suicide note is real or fake [22]. Marketing is another application of AA because it can categorise demographic traits of customers by analysing their comments and reviews about products and services in social networks. These insights can help companies to adapt their communication style and marketing strategies to provide a more satisfactory customer–company relationship. In the bibliography, most of the traits analysed from the perspective of authorship analysis have focused on demographic traits, such as gender, age, or profession. Recent works have focused on behaviour. For instance, by determining which users of a social network are fake news spreaders [23], whether a person is a celebrity or not [24], the quality of their arguments [25], or their reputation [26].

According to Koppel et al. [15], AA can be subdivided into three tasks: (1) authorship attribution (also known as authorship identification), whose objective is to identify the author of a certain work; (2) authorship verification, whose objective is determining whether the suspected author was the one who wrote a questioned document, given a set of candidate authors and samples of their writing; and (3) author profiling, which deals with the identification of demographic and psychographic traits of the authors to identify groups of people based on their age, gender, educational level, native language or personality [27,28].

Author profiling consists in recognising author's demographic and psychographic traits by analysing their authored texts. Demographic traits are, for instance, gender or age range, whereas psychographic traits studies traits such as educational level, personality, or political ideology. Novel tasks focus on predicting age, gender, and occupation of celebrities based on the profiling of their followers [29]. Other researches have focused on location prediction. For example, in the work described in [30], the authors employed stylistic features from authors' writing style to determine their location. For this purpose, the authors employed a novel term weight scheme to calculate document weights specific to every location area. The benefits of this work is that is a complementary method to guess the authors' location that could improve the reliability of surveillance methods regarding public health. Regarding Spanish language, some works have focused on the development of linguistic resources, such as SpanText [31], a corpus compiled from Spanish documents with a formal style in which the presence of slang, abbreviations or contractions is not common. This dataset has been annotated with the age and gender of the authors. It contains texts from different Spanish-speaking countries and includes different topics and a large variety of authors. Other popular resources regarding author profiling in Spanish are the datasets published at PAN's shared tasks [23,27,32–34]. Note that the techniques employed by the latest PAN' shared-tasks are analysed further in this Section. Another relevant resource regarding author profiling can be found at [35], in which the authors proposed a task for determining the occupation and place of residence of users in a dataset published in Mexican–Spanish. Hate speech identification is another application of author profiling as we observed from [13], regarding cyberbullying detection, or in [36], also focused on abuse detection.

Authorship attribution, on the other hand, consists in establishing the correct link between a set of candidate authors and a set of candidate texts. Authorship attribution can be classified as (1) closed-set, if all the documents were written by authors included in the poll of candidates; or (2) open-set, if any of the documents could have been written by other authors not included in the closed set of candidate authors. A more challenging problem is author clustering, that entails the creation of groups of documents written by the same person [37]. Authorship attribution has been applied widely in forensic tasks. For example, in [38], Rocha et al. performed an open-set authorship attribution on small text samples, focusing on social media with forensic purposes to facilitate the identification of users behind identity concealment mechanisms. In addition, the authors presented a comprehensive review of the existing authorship identification techniques. Emails have also been investigated from an authorship attribution perspective, as nowadays most crimes and scams are performed by e-mail. In this sense, the authors of [39] built an automatic classifier based on SVM by using an analytic hierarchy process (AHP) that involves reshaping different features such as word frequency, sentence structure, and the usage of punctuation signs. They achieved promising results with an accuracy superior to 95%. Within forensics, source code identification has drawn some attention due its applications in some areas such as copyright dispute, ghostwriting detection and preventing cheating in

academia. In this sense, in [40] we can find an approach based on several convolutional neural network architectures combined with TF–IDF n-grams features.

The analysis of recent submissions of the participants in PAN's shared tasks regarding authorship analysis [23,33,41] reveals that the most popular feature sets are term-counting features, such as word or character n-grams. Word embeddings are another popular features employed in PAN's shared tasks. Participants also employ linguistic features including Part-of-Speech (PoS), stylometric parameters such as average text length and the usage of certain punctuation symbols, discourse markers, slang, contractions, misspellings, common intros and outros, and emoticons, among other features. There are also features regarding emotions and personality traits such as Watson Personality Insights by IBM.[1] In addition, some participants evaluate different communication styles that can be categorised as self-revealing, action-seeking, information seeking, and fact-oriented. Methods that rely on percentages of function words as well as syntactic features usually provide good clues for authorship identification [42]. Regarding supervised classifiers, authors employed traditional machine-learning classifiers such as SVMs. As has been noted, the usage of neural networks and word embeddings was still minority, relying mostly on recurrent and convolutional neural networks as well as transformer models based on BERT. The usage of ensembles of classifiers was also popular.

3. PoliCorpus-2020

Twitter is the most popular micro-blogging platform and it is present all over the world. Celebrities, politicians, and companies use it everyday for sharing news, daily experiences and communication campaigns. Among the several characteristics that make Twitter suitable for compiling datasets regarding NLP, it is worth highlighting hashtags and Twitter's public API. On the one hand, hashtags are a means to creating and organising topics in a dynamic way, allowing people to find these topics and discuss them. On the other hand, the Twitter API allows compiling the posts of the users that have an open profile in the network.

We used UMUCorpusClassifier [43] to compile tweets during 2020 from Twitter accounts of politicians in Spain. The accounts were selected primarily from: (1) members of the government of Spain, (2) members of Congress and Senate of Spain, (3) mayors of some important cities in Spain, (4) presidents of the autonomous communities, (5) former politicians, and (6) collaborators affiliated with political parties. In total, we identified 385 different authors and a total of 241,864 tweets excluding retweets. The Spanish politicians published an average of 626.59 tweets during 2020, with a standard deviation of 600.55. We observe that 206 politicians have publications in each month in 2020, and 77.143% of the politicians have publications in, at least, 9 months. Thus, the publications of politicians have been constant throughout 2020.

We labelled each politician with their gender, their year of birth, and their political spectrum on two axes (binary and multiclass). The idea of political spectrum was inspired by the arrangement of the Members of Parliament during the French Revolution. The simplest form of political spectrum is the binary left–right spectrum, dividing into those who supported the revolution sitting on the left, and those who supported the king sitting on the right of the president. Traditionally, left-wing parties emphasises ideas regarding equality, freedom, and internationalism, whereas right-wing parties relies on tradition, nationalism, hierarchy, and duty. However, the binary spectrum is simplistic, and some multiclass alternatives have to categorise political ideologies better. For example, the horse-shoe spectrum considers five positions: central, left, right, far left, and far right, in which extreme positions (far-left wing and the far-right wing) are closer between them but far from central positions. The categorisation of the political parties within the political spectrum was decided based on the self-perceived Spaniard perception.[2,3]

As we observed after reviewing the corpus, not all the tweets were written in Spanish. There are two main reasons for this. First, in Spain some languages other than Spanish are spoken, and politicians use them on certain occasions to empathise with the inhabitants of certain regions in which that language is rooted by tradition. That is the case of co-official languages, such as Catalan, Basque, or Galician. Second, politicians sometimes address their messages to other foreign politicians or citizens of other countries, using the languages spoken on those countries. Accordingly, we identify the language in which each tweet was written by using an approach based on fastText [44,45]. It is worth noting that some of the languages spoken in Spain share words among them, causing false positives and false negatives regarding language identification. To solve this problem, we set a threshold to determine if the prediction of the language is reliable or not. This threshold was set at 75% through trial and error. We observed that nearly 88.95% of the tweets were written in Spanish, followed by Catalan (9.10%), Galician (0.8%), and English (0.6%). We remove non-Spanish tweets due to the notable imbalance among the languages and because the NLP resources to obtain the linguistic features and the pre-trained word embeddings were developed for Spanish. This analysis also revealed that many times politicians share content from news websites without using retweets. As those tweets did not reflect the writing style of the authors, we discarded those that contain mentions to news sites or some linguistic clues, such as the pipe symbol, which is used commonly by news sites to categorise their news.

Next, the most representative tweets per politician were selected. First, we categorised the tweets into twelve bins, according to the month in which the tweet was posted. Second, we ordered each bin by the number of topics that appears in each tweet and their length. Then, we selected proportionally tweets for each bin until we got between 120 and 200 tweets per politician. To obtain the relevant topics, we extracted all the hashtags from the corpus, getting a total of 779 unique hashtags. We reviewed this list manually to merge related hashtags and to create a list of synonyms and similar keywords. The final set contains 15 categories (see Table 1 for a comprehensive list of these categories and some examples of each category). On average, each politician in the PoliCorpus has nearly 200 relevant tweets posted during 2020 with a standard deviation of 30.29.

Next, after discarding non-relevant tweets and authors, the Twitter accounts of the politicians were anonymised by replacing their account with the token *@user{number}*. Other Twitter accounts that were not in the candidate set were encoded as *@user*. Consequently, the author traits cannot be guessed trivially by reading their name and searching information of them on the Internet.

The final step was to arrange the PoliCorpus-2020 to conduct the author profiling and attribution tasks. It is worth noting that for the author profiling task, the politicians from training, validation, and test were independent to prevent that the machine learning approaches learn to identify authors rather than the traits. Accordingly, we selected 166 politicians for training, 52 for validation, and 51 for testing, which results in a total of 269 politicians.

[1] https://www.ibm.com/cloud/watson-personality-insights

[2] https://www.epdata.es/datos/derechas%2Dizquierdas%2Dasi%2Dcalifican%2Despanoles%2Dpp%2Dpsoe%2Dpodemos%2Dciudadanos%2Dvox/253

[3] https://www.statista.com/statistics/1059209/political%2Dideology%2Dof%2Dspaniards/

Table 1
Hashtag distribution.

Topic	Examples
covid19	#covid19, #coronavirus, #covid2020
Traditions	#diadelahispanidad, #semanasanta
Local	#bilbao, #madrid, #murcia
Political activity	#gobierno, #elecciones2020
Social	#diadelamujer, #pinparental, #lgtbifobia
Foreign policy	#brexit, #europa, #china
Political-parties	#pp, #psoe, #gobiernoprogresista
Education	#leycelaa, #vueltaalcole, #leycelaa
Health	#sacapecho, #sanidadpublica
Agriculture	#alimentacion, #agraria, #pesca
Politicians	#ayuso, #pedrosanchez, #pablocasado
Economy	#turismo, #comercio, #pymes
Journalism	#fakenews, #stopbulos, #manipulacionrtve
Climate change	#cambioclimatico, #sosmarmenor
Other issues	#encuestamonarquia, #okupas, #euco

We are aware that one limitation of the PoliCorpus 2020 is that all users are politicians, which makes us consider risk of bias in the neural network models generated. To prevent this, we compiled an extra test dataset from Spanish journalists whose political affiliation could be inferred. This dataset was manually labelled by three annotators of our research team. During the annotation process, we noted that there was a strong consensus regarding which journalists are more in favour of the left or right wing, but less agreement regarding annotating the multiclass political ideology. The 51 journalists have an average number of tweets of 190, with a standard deviation of 43.

Table 2 shows the statistics of the author profiling configuration and how the demographics and psychographic traits are distributed for the politicians and journalists. Note that we do not include the gender and age range of the journalists, as we are interested on the evaluation of the psychographic traits. We can observe that age range is the most unbalanced trait, as there were fewer young politicians between the ages of 25 and 34 and fewer older politicians over 65. The other traits present a slight imbalance. The proportion is 42.01% vs 57.99% regarding gender, that is to say, between female and male politicians; 54.27% vs 45.72% regarding binary political spectrum, that is, between the left vs the right wing; and 20.82%, 33.46%, 30.85%, and 14.50% regarding multiclass political spectrum (between left, moderate left, moderate right, and right wing). The journalists dataset was also imbalanced, with a proportion of 64.51% regarding journalists more akin to the left, and a proportion of 39.22%, 21.57%, 25.49%, and 13.72% regarding multiclass political spectrum (left, moderate left, moderate right, and right wing).

For the author attribution task, we used the same tweets and politicians that are in the training set of the author profiling task, namely 109 tweets per 165 politicians accounts. Next, we selected 80 more tweets from each politician in the training set, 40 for validating and 40 for testing. None of these 80 tweets per politician are included in any of the previous splits mentioned above.

The PoliCorpus-2020 has been released to be used for the scientific community.[4] It has been formatted in a similar way to which it is done in the corpora used in some of the PAN shared-tasks, containing a file with ground truth data, that is to say, the anonymised set of authors, including their age range, gender, and political spectrum and, for each task, a file containing the Twitter IDs organised by users. The tweets compiled from the Spanish journalists are also included. It is worth noting that the public version of the corpus contains only the IDs of the tweets rather than the text itself. This decision was made based on Twitter guidelines[5] because it enables the authors of the tweets to maintain their rights about the content they published on the Internet.

Table 2
Distribution into demographics and psychographic traits for the author profiling task with the PoliCorpus 2020 and the Journalist dataset.

Trait	Class	Total	Train	Val	Test
Politicians					
Gender	female	113	67	23	23
	male	156	99	29	28
Age	25–34	28	21	1	6
	35–49	126	80	23	23
	50–64	104	57	26	21
	over 65	11	8	2	1
Spectrum (binary)	left	146	88	31	27
	right	123	78	21	24
Spectrum (multiclass)	left	56	37	12	7
	m-left	90	51	19	20
	m-right	83	54	15	14
	right	39	23	6	10
Journalists					
Spectrum (binary)	left	31	–	–	31
	right	20	–	–	20
Spectrum (multiclass)	left	20	–	–	20
	m-left	11	–	–	11
	m-right	13	–	–	13
	right	7	–	–	7

4. Materials and methods

In this work, we perform two AA tasks on the PoliCorpus-2020: author profiling and authorship attribution. The method employed to carry out our proposal can be summarised as follows (see Fig. 1): First, through the DataLoader module we obtained the set of authors and their writings for each experiment and trait. Second, through the Text pre-processing module the texts were cleaned. Third, the feature sets were extracted. Fourth, we conducted a feature selection process in order to simplify the linguistic features by keeping only the most discriminatory. Fifth, the Splitter module divides the corpus into training, validation and testing according to the task. Sixth, the validation dataset was used for evaluating the best deep-learning models and the feature sets. Last, the final model was evaluated on the testing dataset.

4.1. Dataset loader module

The Dataset Loader module enabled the retrieving of the PoliCorpus-2020. It can obtain the documents organised by user, in which all tweets are returned individually, or by user, in which all the tweets of the same author are merged. As for the tweets posted by Spanish journalists, the module works in a similar manner.

4.2. Text pre-processing module

To solve the authorship attribution and the author profiling tasks, we handled different types of feature sets, including word and character n-grams, linguistic features and different forms of embeddings. In this sense, we performed a common pre-processing step, in which we (1) remove hyperlinks; (2) lowercase the texts; (3) remove digits including numbers, phones, dates, or hours; (4) expand hashtags, acronyms, and SMS language; (5) remove punctuation symbols and quotations; and (6)

[4] https://pln.inf.um.es/corpora/politics/policorpus-2020.rar

[5] https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

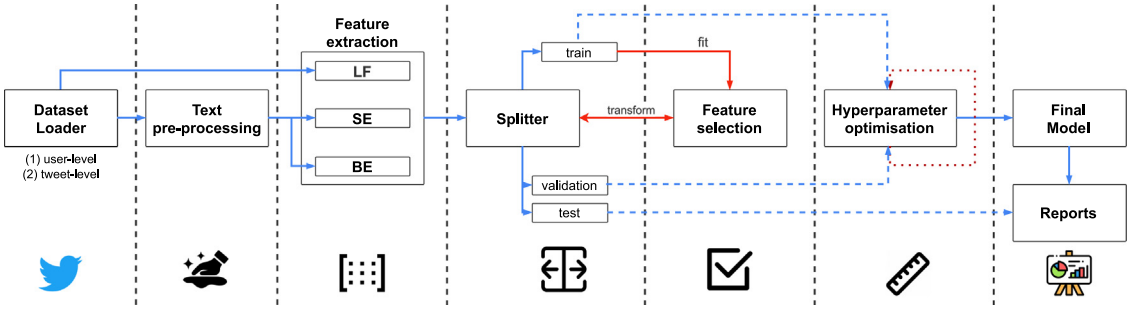


Fig. 1. The pipeline of our proposal.

remove white spaces and break-lines. It is worth noting that the pre-processed tweet and its original version are used to obtain the linguistic features. For example, the original text is used to obtain linguistic features regarding correction and style.

4.3. Feature extraction

In this work we evaluate different feature sets that can be categorised as linguistic features, statistical features based on n-grams, and embedding-based features. Next, each feature set is described, as well as the strategy by means of which they were obtained and why they were considered. We also evaluated the combination of these feature sets in different ways.

4.3.1. N-gram based features (NG)

Word and character n-grams are popular features concerning IR tasks that represent a text as a vector containing the frequency of certain keywords and phrases. The major drawbacks of n-gram-based approaches are that they are computationally expensive, as they tend to have thousands of features; and they are context-less, as an n-gram does not account for its surrounding n-grams, making them weak against linguistic phenomena such as homonymy. In addition, n-grams capture content words, which can cause the model to lose generality when the texts from training are focused on certain topics.

We extracted word and character n-grams using TF–IDF. For word n-grams we combined the unigrams, bigrams, and trigrams, whereas for the character n-grams we combined sequences between 2 and 7 character length without word boundaries. We merged both feature sets into one, and then we applied Latent Semantic Analysis (LSA) for dimensionality reduction to obtain a vector of 100 components. For this, we use the method applied by [46] in the PAN shared task of 2018, as they provided very good results and we want to establish a robust method as baseline. These features were obtained by tweet level, and then we averaged those vectors by user. As for the Spanish journalists, we used the same vocabulary extracted from the PoliCorpus 2020 training set.

4.3.2. Linguistic features (LF)

LF are features that represent the variety of a language from multiple perspectives, including register, jargon, part-of-speech, and figurative language among others. LF have been widely used for conducting authorship analysis, reflecting features that are hard to be captured by other means. For example, the raising of an author's voice expressed by means of uppercase letters would require the usage of cased word n-grams and embeddings. However, due to the curse of dimensionality, it would be preferable in other cases to keep uncased n-grams in order to generalise the usage of certain keywords better. In this sense, LF capture this phenomenon more easily.

LF were obtained with UMUTextStats, a tool inspired in LIWC [47], designed for Spanish from scratch by our research group. UMUTextStats has been already evaluated for conducting automatic text classifications in some domains such as infodemiology [48] and misogyny identification [49].

According to Tausczik and Pennebaker [47], the words and expressions employed in writing communication fall into two broad categories: (1) content words, which convey the content of the message, and (2) style (or function) words, which reflect how people communicate. In the last decade, the efforts of the scientific community have focused on shift from hand-crafted features to reusable features like term-counting features and word embeddings [21]. However, these kinds of features tend to be larger, so they are more difficult to be interpreted and reused. Moreover, the challenge of the domain requires larger and varied datasets to learn to build robust models, which is a problem for languages other than English.

The current version of UMUTextStats compiles a total of 365 linguistic features that fall into the following categories:

- **Phonetics (PHO)**. They include features such as expressive lengthening, a linguistic device that consists in repeating some of the letters of a word for emphasis [50].
- **Morphosyntax (MOR)**. These features are divided into three major subcategories. First, PoS-based features, that includes adverbs, adjectives, determiners or pronouns, to name but a few. Second, features that capture components of the words including stems and affixes. Third, features that capture the grammatical gender and number of words. It is worth noting that Spanish is a highly inflected language. These inflections denote multiple syntax and semantic meanings and capture the communication style of an author.
- **Correction and style (CAS)**. These features include orthographic, stylistics, and performance errors. Orthographic errors capture wrong use of Spanish accentuation, sentences that start in lowercase or misspellings. Stylistic errors capture sentences that starts with cardinal numbers or sentences that start with the same word. Performance errors detect duplicated words or wrong use of punctuation symbols such as dots after exclamatory and interrogative clauses or two consecutive commas or dots. Performance errors also capture redundant and common errors.
- **Semantics (SEM)**. These features include lexicons concerning (1) onomatopoeia, (2) euphemism and dysphemism, and (3) synecdoche.
- **Pragmatics (PRA)**. These features capture (1) the use of figurative language (hyperboles, idiomatic expressions, rhetorical questions, verbal irony, understatements, metaphors and similes), (2) discourse markers used for structuring the conversation regarding connectors, reformers, argumentative clauses, and conversational-bookmarks; and (3) typical courtesy forms for greetings or condolences, among others.

- **Stylometry (STY)**. These features are composed of statistics regarding stylometry that include length of the text, lexical diversity of the users by using the type-token ratio (TTR) standard, number of words and syllables, number of sentences, number of words in uppercase, readability formulas or punctuation symbols.
- **Lexis (LEX)**. These features include lexicons of different domains that tend to capture the purpose of the message. They include lexicons of concrete concepts such as animals, weapons, jobs, crime, money, health, and ingesting, as well as features that capture abstract concepts that include social and complex ideas such as achievement, risk, and cognitive processes.
- **Psycho linguistic processes (PLP)**. These features include positive, negative, and neutral sentiments and attitudes.
- **Register (REG)**. These features capture offencive language, informal speech, or the usage of learned words.
- **Social media (SOC)**. These features capture the degree in which users make use of terminology related to social networks. They include the usage of hashtags, mentions, and hyperlinks, as well as lexicons with terminology specific to social networks.

To generate a linguistic profile per each politician, we first extracted the linguistic features for each tweet and then averaged those features per politician.

4.3.3. Embeddings-based features

Embeddings are efficient and dense representation of characters, words, sentences, or documents, in which semantically similar items tend to have similar encoding representation. These representation techniques have some benefits over features based on character or word n-grams. First, considering computational efficiency, neuronal networks are more efficient with dense vectors rather than sparse vectors. Second, word embeddings can be trained by using unsupervised methods over large corpora and translate this knowledge to more specific tasks, such as sentence classification resulting in models that converge faster and tend to generalise better. Third, word embeddings can be aware of out-of-vocabulary words and can cluster new words based on the context in which they are used. In this work, we evaluate two fixed-length sentence-level representation learned from FastText and BERT, as well as two non-fixed word embeddings representations based on pretrained word embeddings and contextual word embeddings. Moreover, some authors have explored the linguistic properties that sentence and document-level vectors are capable of encoding [51]. Next, the sentence and word embeddings employed are described.

Regarding sentence embeddings (SE, SBE), we followed two approaches based on contextual and non-contextual embeddings. On the one hand, the non-contextual sentence embeddings (SE) were obtained by using FastText using the Spanish models [52]. The pre-trained model of Spanish have been trained from Common Crawl and Wikipedia but adjusted with the PoliCorpus 2020 during training. SE assemble a single vector of dimension 300 for a sequence of the individual word embeddings. The process in which sentence vectors are calculated from FastText is the average of their word vectors plus the `EOS` token. Therefore, we followed the same approach, calculating the sentence embeddings per tweet and averaging them per user. On the other hand, we evaluated Sentence BERT transformers (SBE). One of the disadvantages of non-contextual word embeddings is that they do not handle polysemy, as the word representation is the same regardless of the context in which the word is used. Contextual word embeddings based on transformers make use of an attention-method that outputs the embeddings vector based on the context word. Bidirectional Encoder Representations from Transformers (BERT) [53] is a pretrained model developed by Google that takes into account the specific context of each word in a sentence. For this, BERT could read the entire sequence in a single step taking into account the context of a word from its surrounding words (previous and subsequent). Thus, it outputs different embeddings for the same word according to its context. SBE were obtained with BETO, a Spanish BERT model [54], that was trained from the Spanish Unannotated Corpora. To get the BERT-based embeddings at sentence-level, we fed each tweet into BETO and took the encoding from the classifier token (`[CLS]`). This gave us a representation of each tweet as a fixed-length vector [55] of length 768. After calculating SBE per tweet, we averaged them by politician, as we do with NG, LF or SE.

For the word embedding features, we also evaluated contextual (BERT) and non-contextual pretrained word embeddings (PWE). First, to learn the PWE we rely on Spanish pretrained word embeddings models from fastText, GloVe, and Word2Vec. One of the benefits of word embeddings over sentence embeddings is that they allow for the evaluation of other neural network architectures, such as convolutional and recurrent neural networks, which exploits properties of human language such as spatial and temporal dimensions. On the one hand, Convolutional Neural Networks (CNNs) contain layers with convolution filters that can be used to learn local features and generate intermediate features from a higher order. For example, a CNN can learn words whose meaning is different from the one of those words separately. In addition, CNNs are more effective at guessing polysemic words, as their meaning can be understood by looking at the surrounding words. Recurrent Neural Networks (RNNs), on the other hand, make the most of the temporal dimension, which means that they can exploit information regarding the position of the words within a sentence. Specifically, in this work we evaluated Bidirectional Gated Recurrent Units (BiGRU), a specific type of RNN based on two gates. Second, we used BERT to handle contextual word embeddings. Similar to SBE, we relied on BETO because it is adapted to Spanish. It is worth noting that BERT has a maximum length restriction of 512 tokens. Therefore, we followed an approach similar to the one described in [56], consisting in training the network at tweet level and then averaging the results per politician based on the mode of the predictions.

4.4. Feature selection

The next step involved selecting the most relevant linguistic features. Succinctly, the process can be described as follows. First, we normalised each feature independently into a range [0, 1] using a `MinMaxScaler`, as there were features measured on different scales. Prior to this step, we ensured that these features did not contain outliers. Next, we applied a feature selection process by obtaining the Mutual Information (MI) between each feature with the target class to determine their inter-dependency. MI was fitted over the training dataset and feature selection was applied individually for each trait (gender, age range, and political spectrum both binary and multiclass) in the author profiling task and for each author in the authorship attribution task. Next, we discarded those linguistic features for which MI fell below the first quartile (Q_1).

4.5. Splitter module

The Splitter module is responsible for extracting the training, development, and testing datasets. As explained previously in Section 3, the splits depend on the task. For author profiling, the dataset was composed by 269 politicians divided into 166 for training, 52 for validation, and 51 for testing. These politicians

are the same regardless of the demographic or psychographic traits evaluated. As for the Spanish journalists, we have 51 users that are used only for testing. For the authorship attribution task, the training split is the same split of the author profiling task, but the validation and testing splits are composed of 80 different tweets from each of these politicians, in a proportion of 50–50 for validation and testing, that is to say, 40 tweets for validation and 40 tweets for testing for each politician.

4.6. Hyper-parameter optimisation

For each experiment we conducted a hyperparameter tuning to find the best neural network architecture according to the macro-averaged F1-score. This stage evaluates different batch sizes, dropout ranges, activation functions, and neural network architectures. For all feature sets except for pre-trained word embeddings (PWE), we relied on multilayer perceptrons (MLP) that included shallow neural networks, composed by one or two hidden layers and keeping the same number of neurons in each layer (brick shape), and deep neural networks, with a number of hidden layers between 3 and 8 and in which the number of neurons vary according to the following shapes: brick, funnel, long funnel, diamond, rhombus, and triangle. In case of PWE, we included two more architectures: a CNN, in which we evaluated different kernel sizes, and an RNN based on BiGRU. In addition, we evaluated three different Spanish pre-trained word embeddings with PWE: fastText, Glove, and Word2Vec. We fixed the learning rate to $10e^3$ with a time-based scheduler, the number of epochs to 1000, as we used an early stopping mechanism, and Adam as optimiser.

Table 3 contains a list of the hyperparameters evaluated. As explained in Section 4.3, NG, LF, SE, and SBE operate at user level, that is to say, they have one vector per politician, whereas for BERT and PWE operate at tweet level, that is to say, there is one vector per tweet, and the final results are averaged using the mode of the predictions. Therefore, we adapted the batch size hyperparameter to larger values when training at tweet level to ensure that there is a representative number of instances in each batch.

We also evaluated combinations of some of the feature sets in pairs. For this purpose, we built only a neural network with the functional API of Keras by using multiple independent input layers. Regardless of the architecture (MLP, CNN, or RNN), each feature set is restricted to its own network architecture as described above. Then, the last hidden layer of each feature set is combined and connected to the final output layer. As regards BERT, we proceeded as follows. For BERT in isolation, we fine-tuned BETO with the HuggingFace's trainer during 3 epochs, 500 warm up steps, a weight decay of 0.01, and a batch size of 16. For BETO and the LF, we fed the `CLS` tokens and the LF features into a deep-neural network with Keras and performed the hyperparameter optimisation stage from scratch.

5. Results and analysis

This section is organised according to the AA tasks carried out. First, Section 5.1 describes the demographic traits regarding gender and age range, and the psychographic traits regarding political spectrum (binary and multiclass) for the politicians. Furthermore, we evaluate the generated models for the psychographic traits with the Spanish journalist dataset. Second, Section 5.2 describes a closed-set authorship attribution. The best results from both tasks are ranked with the macro F1-score that is to say, the harmonic mean between Precision and Recall of each class. Results also include the F1-score for each class and the weighted F1-score to compare the results since we address different balance among the classes.

Table 3
Hyperparameter options for the neural networks architectures evaluated.

Parameter	Ranges
Shared hyperparameters	
Batch size	[4, 8, 16] (userlevel) [128, 256, 512] (tweetlevel)
Dropout	[False, 0.1, 0.2, 0.3]
Neurons per layer	[8, 16, 48, 64, 128, 256, 512, 1024]
Shallow neural networks	
Activation	[linear, relu, sigmoid, tanh]
Numbers of layers	[1, 2]
Shape	[brick]
Deep neural networks	
Activation	[sigmoid, tanh, selu, elu]
Numbers of layers	[3, 4, 5, 6, 7, 8]
Shape	[funnel, rhombus, longfunnel, brick, diamond, triangle]
Convolutional neural networks	
Activation	[sigmoid, tanh, selu, elu]
Numbers of layers	[1, 2]
Shape	brick
kernel size	[3, 5, 7]
Recurrent neural networks	
Bidirectional	[True, False]
Activation	[sigmoid, tanh, selu, elu]
Numbers of layers	[1, 2]
Shape	[brick]
kernel size	[3, 5, 7]

5.1. Author profiling

For author profiling, as has been stated in Section 3, the dataset is divided as follows: 166 politicians' profiles for training, 52 politicians' profiles for evaluating the models, and 51 politicians' profiles for testing in a ratio near to 60-20-20. Besides, for the evaluation of the psychographic traits, we included a total of 51 Spanish journalists whose political ideology is inferred from the models generated with the PoliCorpus 2020. This section is divided into demographic (see Section 5.1.1) and psychographic (see Section 5.1.2) traits.

5.1.1. Demographic traits evaluation

Two demographic traits are evaluated in this Section, the results being shown in Table 4 and described below.

Regarding gender identification, the best result is obtained with PWE using BiGRU, achieving a macro F1-score of 72.022%. Combined with the LF, PWE using BiGRU obtains exactly the same performance. A similar behaviour is observed with PWE using CNN, which achieves exactly the same performance regardless the usage of LF. Only PWE with an MLP improves their results when combined with LF (from 66.447% to 70.641%). LF also improves the macro F1-score of SBE (from 65.826% to 70.543%) and BETO (from 69.118% to 71.727%). However, the addition of LF decreases slightly the results achieved by SE (from 66.615% to 64.692%).

On the other hand, it can be observed that all the feature sets and neural network architectures are more reliable for the *male* class, this difference being more remarkable in some combinations, such as BETO, BETO+LF, or PWE with MLP. However, the combination of the LF with other feature sets reduces this difference among labels. For example, the difference between *male* and *female* in SBE is 10.721%, but only 2.31% when SBE is combined with LF. This reduction can also be observed when adding the LF to the PWE (MLP), and SE, but to a lesser degree. These results suggest that the incorporation of LF into machine learning models is beneficial in the majority of cases for the identification of demographic traits.

Table 4
Author profiling based on demographic traits of gender and age range.

Feature set	Architecture	Gender				Age range					
		$F1_{FEMALE}$	$F1_{MALE}$	$F1_{WGT}$	$F1_{MACRO}$	$F1_{25-34}$	$F1_{35-49}$	$F1_{50-64}$	$F1_{OVER65}$	$F1_{WGT}$	$F1_{MACRO}$
NG	MLP	66.6667	73.6842	70.5194	70.1754	26.6667	23.8095	26.3158	28.5714	25.2710	26.3409
LF	MLP	60.0000	74.1935	67.7925	67.0968	–	28.5714	52.6316	**66.6667**	35.8642	36.9674
SE	MLP	65.3061	67.9245	66.7437	66.6153	**58.8235**	52.0000	32.2581	–	43.6541	35.7704
SBE	MLP	60.4651	71.1864	66.3513	65.8258	40.0000	47.6190	40.9091	–	43.0261	32.1320
PWE	MLP	57.8947	75.0000	67.2859	66.4474	–	60.0000	16.0000	–	33.6471	19.0000
PWE	CNN	62.2222	70.1754	66.5887	66.1988	–	61.9718	16.6667	–	34.8108	19.6596
PWE	BiGRU	68.1818	75.8621	72.3984	**72.0219**	–	50.0000	58.6207	–	27.1552	**46.6870**
BETO	BERT	58.8235	79.4118	70.1269	69.1176	–	53.9683	31.2500	–	37.2063	21.3046
LF+SE	MLP	64.0000	65.3846	64.7602	64.6923	–	50.0000	54.0541	25.0000	45.2968	32.2635
LF+SBE	MLP	**69.3878**	71.6981	70.6562	70.5429	–	50.0000	**66.6667**	40.0000	50.7843	39.1667
SE+SBE	MLP	65.2174	71.4286	68.6275	68.3230	–	37.5000	42.8571	33.3333	35.2124	28.4226
LF+PWE	MLP	63.1579	78.1250	71.3751	70.6414	–	**62.9630**	58.5366	–	52.4984	30.3749
LF+PWE	CNN	62.2222	70.1754	66.5887	66.1988	–	58.0645	36.3636	–	41.1592	23.6070
LF+PWE	BiGRU	68.1818	75.8621	72.3984	**72.0219**	–	40.9091	50.9804	–	39.4411	22.9724
BETO+LF	BERT	62.8571	**80.5970**	**72.5967**	71.7271	–	62.0690	62.5000	33.3333	**54.3808**	39.4756

Next, we analysed the LFs that have more correlation with gender traits. For this, we averaged the LFs per class (*male*, *female*) to obtain the MI per class (see Fig. 2 (left)). We observed that the most discriminatory feature is related to whether dates are written in textual mode or with digits. As can be observed, topics regarding *female social groups* is another relevant linguistic feature. This feature includes word and expressions to refer to female family members, such as grandmothers, mothers, aunts, or daughters. This feature may appear in discussion topics related to feminism, abortion, or education. Our analysis indicates that these topics were mostly referred by female politicians. Similarly, the usage of feminine personal pronouns is also relevant and it appears more often in tweets posted by female politicians. Other relevant linguistic features are related to morphosyntax, as we can observe differences on a special type of suffixes called verbalisers, which includes words such as *asesinar* (to kill), transforming the word *asesino* (killer) into a verb. However, we observed that, except for a couple of exceptions mentioned above, there are not highly discriminating linguistic features in terms of *male* and *female*.

Next, before presenting the results regarding age range identification, it is worth remembering that age range is heavily imbalanced as there are fewer politicians in the age ranges of 25–34 and over 65 (see Table 2). Similar to the gender trait, the best result is obtained with PWE trained with BiGRU achieving a macro F1-score of 46.687%. Note that the macro F1-score penalises those models that are not able to classify any instances of a specific class. Despite this fact, NG obtains limited results even though it is the only feature set capable of correctly identifying politicians in all age ranges. When observing the weighted F1-score, which is less influenced by minority classes, the best results are achieved with BETO combined with LF (54.381%), being able to identify correctly most of the politicians aged between 35 and 64. Similarly to the gender trait, the addition of LF to SBE improves the results (from 32.132% to 39.167% of macro F1-score). However, it reduces the results of SE (from 35.77% to 32.264% of macro F1-score). The combination of SE and SBE is not beneficial either (28.423% of macro F1-score). The combination of LF with PWE is beneficial for the weighted F1-score (from 33.647% to 52.498% with MLP, from 34.811% to 41.159% with CNN, and from 27.155% to 39.441% with BiGRU), but it decreases the macro F1-score of BiGRU (from 46.687% to 22.972%). Regarding the analysis of each age range individually, only fixed sentence embeddings (SE and SBE) and n-grams (NG) are able to classify the age range between 25 and 34 and only n-grams, LF, and the combination of LF with SBE, BE, and BETO are able to classify politicians over 65. None of the PWE (in isolation or combined with LF) are able to classify any older politicians.

Next, we analysed the most discriminatory LF based on age range (see Fig. 2 (right)) and we observed that there are many features related to morphosyntax and verbs, including the percentage of simple subjunctive verbs, verbs in indicative, and verbs in present indicative. Also from the morphosyntactic category, it is also relevant the number of personal pronouns classified by gender, which is more common in younger politicians. Another relevant feature is the usage of colloquialisms from the register category. This feature is common among younger politicians and older politicians, being less frequent in middle-aged subjects. Besides, other relevant features include topics related to countries and languages that appear in territorial policy issues.

In summary, we observed that morphosyntax is the most relevant linguistic category for determining demographic traits. However, their importance for gender and age range prediction appears in different degrees. The only morphosyntactic feature that seems to be equally important is the psycho-linguistic process negativity, which suggests that negative statements can help to discern among gender or age range. One surprising fact is the lack of stylometric features regarding demographic traits. We only observed a few exceptions, such as the usage of dashes, more common among female politicians.

5.1.2. Psychographic traits evaluation

In this section we evaluate psychographic traits regarding political ideology from a binary perspective (left vs right wing) and multiclass that consisted in four labels, namely left, moderate-left, moderate-right, and right wing. We first trained and evaluated the model with the PoliCorpus 2020 dataset, and next we evaluated if the results can be transferred to with the Journalist dataset. First, the results for politicians are shown in Table 5.

Regarding binary political spectrum, we can observe that all feature sets and neural network architectures behave similarly regardless of the political wing. Besides, it can be noticed a high quality leap among fixed sentence features (NG, LF, SE, SBE) with respect to the contextual and non-contextual word embedding features (PWE and BETO). PWE (MLP) (regardless of whether they are combined with LF or not) achieves the best result with a macro F1-score of 98.036%. The results achieved by CNN and RNN combined with LF are slightly inferior (94.118% and 88.194%, respectively). BETO and LF also achieves almost perfect results, with a macro F1-score of 98.027%. Out of the features evaluated in isolation, we observe than LF and SBE achieves the worst results (70.543% and 68.651% of macro F1-score respectively) whereas the results achieved by SE are very high (88.231%). This finding suggests that SBE loses relevant information regarding political ideology identification if we compare the performance

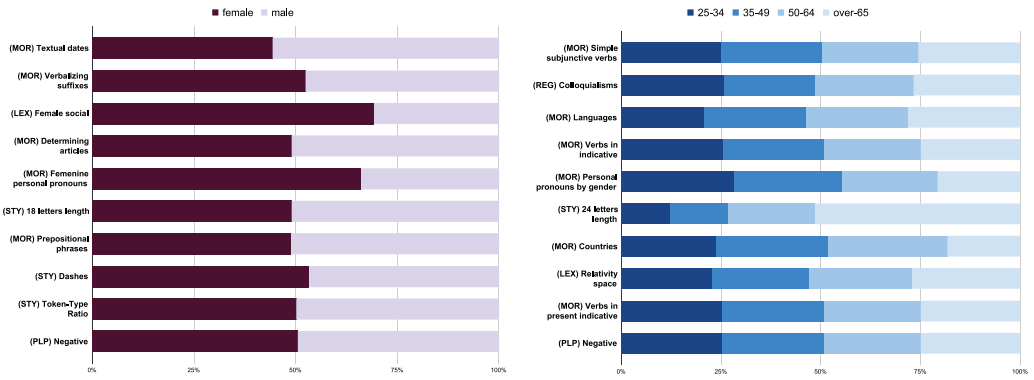


Fig. 2. Differences among linguistic features over gender (left) and age range (right).

Table 5
Author profiling based on psychographic traits of political ideology with the politicians' test split (binary and multiclass).

Feature set	Architecture	Binary				Multi-class					
		$F1_{LEFT}$	$F1_{RIGHT}$	$F1_{WGT}$	$F1_{MACRO}$	$F1_{LEFT}$	$F1_{M-LEFT}$	$F1_{M-RIGHT}$	$F1_{RIGHT}$	$F1_{WGT}$	$F1_{MACRO}$
NG	MLP	79.1667	81.4815	80.2560	80.3241	20.0000	44.4444	10.5263	22.2222	27.4212	24.2982
LF	MLP	69.3878	71.6981	70.4750	70.5429	66.6667	74.2857	60.0000	81.8182	70.7953	70.6926
SE	MLP	88.0000	88.4615	88.2172	88.2308	62.5000	77.7778	74.2857	66.6667	72.5436	70.3075
SBE	MLP	69.2308	68.0000	68.6516	68.6154	58.8235	72.7273	58.3333	70.5882	66.4482	65.1181
PWE	MLP	98.1132	**97.9592**	98.0407	**98.0362**	76.9231	92.3077	84.8485	82.3529	84.1080	86.1965
PWE	CNN	96.1538	96.0000	96.0814	96.0769	82.3529	89.4737	86.6667	82.3529	86.3295	85.2116
PWE	BiGRU	92.0000	92.3077	92.1448	92.1538	80.0000	87.1795	81.2500	75.0000	82.1782	80.8574
BETO	BERT	96.2963	95.8333	96.0784	96.0648	92.3077	**95.2381**	86.6667	82.3529	89.9564	89.1413
LF+SE	MLP	69.2308	68.0000	68.6516	68.6154	58.8235	77.7778	48.2759	60.0000	63.5918	61.2193
LF+SBE	MLP	88.8889	87.5000	88.2353	88.1944	40.0000	82.9268	60.0000	57.1429	65.6856	60.0174
SE+SBE	MLP	92.3077	92.0000	92.1629	92.1538	71.4286	75.6757	66.6667	53.3333	68.2388	66.7761
LF+PWE	MLP	98.1132	**97.9592**	**98.0407**	**98.0362**	80.0000	89.4737	**87.5000**	82.3529	86.2354	84.8317
LF+PWE	CNN	94.1176	94.1176	94.1176	94.1176	72.7273	90.4762	83.8710	88.8889	85.9156	83.9908
LF+PWE	BiGRU	87.5000	88.8889	88.1536	88.1944	**93.3333**	89.4737	86.6667	**94.7368**	**90.2649**	**91.0526**
BETO+LF	BERT	**98.1818**	97.8723	98.0362	98.0271	80.0000	90.0000	83.8710	75.0000	84.0038	82.2177

with non-contextual sentence embeddings (SE) and contextual word embeddings (BETO).

In case of multiclass political spectrum classification, the best result is achieved with the combination of PWE and LF using a BiGRU with a macro F1-score of 91.053%. Note that this is the only experiment in which we observe that the combination of LF with PWE is beneficial, improving from 80.857% to 91.053%. The most surprising fact regarding multiclass political spectrum is the drop of the reliability of NG, dropping from 80.324% macro F1-score (binary) to 24.298% (multi-class). We calculated the confusion matrix (not shown) and observed that NG misclassifies the left with the moderate left wing, the moderate right with the left wing and, to a lesser degree, the right wing with the moderate wings. This finding suggests that n-grams are not suitable for conducting a fine-grained distinction of political ideology. However, this large difference between binary and multiclass does not appear in other feature sets. SE only decreases from 88.231% to 70.308%, SBE from 68.615% to 85.118%, and the PWE features decreases from an almost perfect classification (98.063%) to a macro F1-score 86.196%. The only feature set that achieves better results in multiclass than in binary classification are LF (72.725% of macro F1-score), achieving a 90% F1-score over the right class and only obtaining limited results over the left class.

In order to check whether there were errors between opposite political ideologies, we obtained the confusion matrix (see Table 6) of the best model (PWE with BiGRU and LF). We can observe that only two left-wing politicians were wrongly classified as moderate right, whereas only one right-wing politician was classified as moderate-right. None of the politicians from non-moderate positions were classified on its opposite spectrum, the majority of wrong classifications taking place among moderate postures.

Table 6
Confusion matrix of the multiclass political spectrum from the PoliCorpus-2020 with the combination of LF and PWE with BiGRU.

	LEFT	M-LEFT	M-RIGHT	RIGHT
LEFT	7	0	0	0
M-LEFT	1	17	2	0
M-RIGHT	0	1	13	0
RIGHT	0	0	1	9

The MI of the ranked 10 best linguistic features for both binary and multiclass are shown in Fig. 3. On the one hand, regarding binary political spectrum, we can observe that politicians from the right-wing employed more words and terms related to religion and demonyms. The usage of negative statements is also remarkable. With respect to the left-wing parties, there is an increase in the usage of qualifying adjectives and lexical linguistic features related to the spatial dimension that include verbs such as *abrir* (to open), *colocar* (to put), *rodear* (surround); nouns such as *horizonte* (horizon); and spatial orientations (left, right, up, down) that, due to polysemy, they can also refer to political adversaries. As mentioned above, regarding the performance leap between sentence and word embeddings, we can conclude that averaging embeddings to conform sentence-embeddings could lead to loss of morphosyntactic information. It can be also noticed that there is a strong imbalance in the usage of words longer than 24 characters from left-wing politicians, but this fact is due to the usage of specific hashtags during the COVID pandemic. As explained in Section 4.2, hashtags were expanded in order to keep their meaning. To do this, we split camel case hashtags into pieces. However, some hashtags that were completely composed by all lowercase letters remained the same. On the other hand,

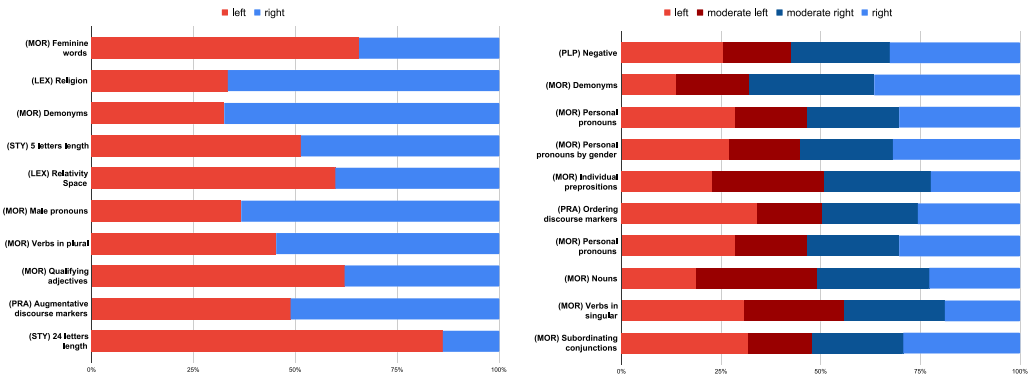


Fig. 3. Differences among linguistic features over author profiling: political spectrum binary (left) and multiclass (right).

regarding multiclass political spectrum, we observe that there are larger differences between the left and moderate left wings than in the right and moderate right. For example, the usage of demonyms is slightly higher for the moderate left wing than for the left wing, but almost the same between the right and the moderate right. The usage of personal pronouns based on grammatical gender is higher in the left wing than in the moderate left wing. This fact is also observed in the right wing as compared to the moderate right, which suggests than non moderate positions discuss more topics related to women. Regarding the usage of discourse markers, we find that the ones used for ordering are relevant for multiclass ideological classification, being more frequent in the left wing.

In order to compare how LF change among the LF categories and psychographic traits, Fig. 4 contains two polar charts for each psychographic trait analysed regarding each LF category: errors, lexis, morphosyntax, phonetics, pragmatics, psycho-linguistics, register, semantics, and social media. Each linguistic category contains the average of their LF, grouped by class. With respect to the error category, in multiclass political spectrum we observe a slight difference between the right and moderate right and the left and moderate left. Regarding morphosyntax, and similar to errors, there is a difference in the multiclass political spectrum in which politicians from the left wing stand out from the rest. Phonetics is the linguistic category that presents the largest differences among traits. This linguistic category is mainly based on the usage of expressive lengthening, and indicates that it is a relevant feature for the multiclass political spectrum. Pragmatics and psycho-linguistic categories show differences in binary and multiclass political spectrum. Register is a relevant feature to discern among multiclass political spectrum, as it helps to discern between moderate and non-moderate political parties. As for semantic features, it can be observed than there are wider differences in the multiclass political spectrum. The usage of social media is relevant in the multi-class political spectrum but not for the binary political spectrum. This category presents major usage by politicians from the right wing. However, politicians from the non-moderate left wing are the ones that make fewer use of hashtags or hyperlinks.

As explained for the dataset collection (see Section 3), one possible bias is that the models generated could be biased due to the fact that all the subjects were politicians. In order to check whether the results can be transferable, we obtained the political ideology from Spanish journalists using the generated models from the politicians training split. The results are shown in Table 7.

As can be seen in Table 7, PWE with BiGRU is the model that degrades the most when it is evaluated with a different type of user, shifting from 92.154% to 68.52%. The combination of LF and PWE with GRU affects the performance of the model even more, achieving a macro F1 score of 53.782%. In addition, we can observe that the left wing is the most affected label. We assume, therefore, that the dependencies learnt from BiGRU among the temporal dimension are biased to the left-wing politicians and cannot be transferable. Nevertheless, the limited results achieved by BiGRU are not observed with other neural networks that use PWE. CNN and MLP perform reasonably well even if evaluated with users with another profession. Another feature set that affects its performance is NG. We assume, therefore, that the grams learned from these models are biased to the politician profession per se instead of the political ideology. This fact was not surprising, as n-gram features can be dominant with content words from conversations and may not capture the true political ideology. Besides, the results of SBE are worth highlighting, as it could seem that their results are limited as compared with other feature sets for the rest of the politicians. However, their performance does not degrade when applied to the inference of the political ideology of average users, achieving a similar performance with LF and SE. In case of the multiclass political spectrum, the most affected model is LF. NG also achieves limited results, but similar to the ones achieved by the politicians. However, LF improves the results of SE (from 19.898% to 24.259%) and SBE (from 51.613% to 63.420%). These results indicate that, although LF is insufficient used in isolation, it can complement other feature sets based on embeddings. Moreover, LF also improves methods based on word embeddings, as we can observe from PWE (48.804% to 63.420% with MLP, from 48.804% to 57.684% with CNN, and from 44.384% to 53.365% with BiGRU) and BETO (from 59.772% to 64.142%).

5.2. Authorship attribution

In this experiment we evaluate the reliability of the PoliCorpus 2020 and the feature sets to identify authors basing on writings from the political domain. For this purpose, we used the same 166 politicians from the training dataset of the author profiling task, but adding 80 new tweets to each politician for validation and testing in a proportion of 50–50. Another important fact is that, unlike the author profiling task in which the results were displayed by user, the results from the authorship attribution task are obtained at document level.

The results from the authorship attribution task are shown in Table 8. They indicate that the combination of LF and BETO achieves the best results, with a macro F1 average of 29.336%. In contrast, the usage of NG and PWE is very limited. NG, on the one hand, achieves a macro F1-score of 8.094%; thus, it is not feasible to categorise individual tweets with word and character n-grams, as the resulting neural network models cannot differentiate well between the authors of political texts. The usage of PWE, on the other hand, also achieves limited results, highlighting the limited performance of BiGRU (1.541%). In contrast, fixed sentence

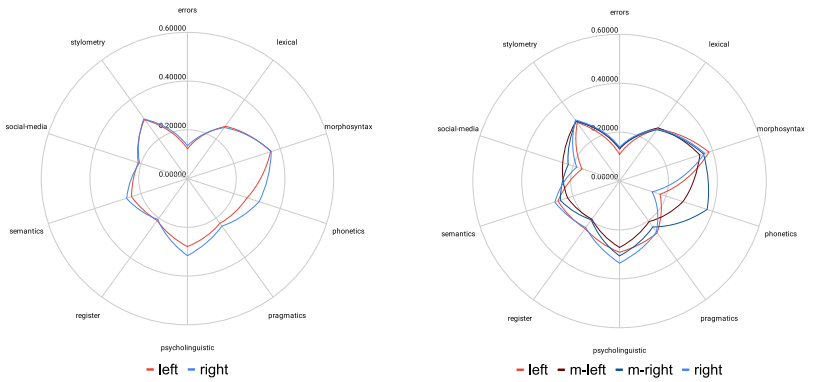


Fig. 4. Polar chart of the Linguistic Features per political-spectrum binary (left) and multi-class (right).

Table 7
Evaluation of the author profiling based on psychographic traits of political ideology (binary and multiclass) over journalists.

Feature set	Architecture	Binary				Multi-class					
		$F1_{LEFT}$	$F1_{RIGHT}$	$F1_{WGT}$	$F1_{MACRO}$	$F1_{LEFT}$	$F1_{M-LEFT}$	$F1_{M-RIGHT}$	$F1_{RIGHT}$	$F1_{WGT}$	$F1_{MACRO}$
NG	MLP	53.846	52.000	53.122	52.923	8.696	38.710	25.000	25.000	21.563	24.351
LF	MLP	67.925	65.306	66.898	66.615	20.000	27.273	20.000	28.571	28.863	25.818
SE	MLP	69.231	68.000	68.748	68.615	47.826	11.765	20.000	–	26.391	19.898
SBE	MLP	76.190	61.538	70.445	68.864	79.070	57.143	41.667	28.571	57.875	51.613
PWE	MLP	88.889	82.051	86.207	85.470	**86.207**	52.632	59.259	22.222	57.876	51.613
PWE	CNN	84.746	79.070	82.520	81.908	69.565	41.667	47.619	36.364	53.397	48.804
PWE	BiGRU	66.667	70.370	68.119	68.519	48.649	44.444	40.000	44.444	44.960	44.384
BETO	BERT	87.879	77.778	83.918	82.828	75.000	57.143	62.500	44.444	63.768	59.772
LF+SE	MLP	73.684	66.667	70.932	70.175	45.161	34.483	17.391	–	29.581	24.259
LF+SBE	MLP	71.186	60.465	66.982	65.826	32.432	26.667	13.333	50.000	28.732	30.608
SE+SBE	MLP	53.333	63.158	57.186	58.246	48.980	26.667	28.571	20.000	34.987	31.054
LF+PWE	MLP	**92.308**	**86.486**	**90.025**	**89.397**	75.000	57.143	60.000	**61.538**	65.477	63.420
LF+PWE	CNN	84.211	80.000	82.559	82.105	71.698	37.500	60.000	**61.538**	59.946	57.684
LF+PWE	BiGRU	47.619	63.333	55.476	53.782	76.923	50.000	61.538	25.000	60.068	53.365
BETO+LF	BERT	89.855	78.788	85.515	84.321	80.000	**63.158**	**68.966**	44.444	**68.675**	**64.142**

Table 8
Results from the authorship attribution task.

Feature set	Architecture	$F1_{macro}$
NG	MLP	8.0939
LF	MLP	18.6417
SE	MLP	18.9682
SBE	MLP	20.8305
PWE	MLP	11.9486
PWE	CNN	8.1058
PWE	BiGRU	1.5146
BETO	BERT	27.2605
LF+SE	MLP	26.2711
LF+SBE	MLP	26.2557
SE+SBE	MLP	21.9318
LF+PWE	MLP	21.2380
LF+PWE	CNN	15.8582
LF+PWE	BiGRU	3.7828
BETO+LF	BERT	**29.3361**

embeddings (SE, SBE) and LF are more reliable for authorship attribution. On the one hand, LF achieves a macro F1-score of 18.642%, similar to SE (18.968% of macro F1-score) and slightly inferior to contextual SBE (20.831% of macro F1-score). We observed that the combination of LF with embeddings improves the overall performance of the author attribution. LF boost SE (from 18.968% to 26.271%), SBE (from 20.831% to 26.256%), and BETO (from 27.261% to 29.336%). These results suggest that the LF provides linguistic evidence that is not possible to capture by any form of the embeddings involved in this research.

As can be observed for information gain (see Fig. 5), the LFs related to stylometry (STY) are the most discriminatory ones, but some of them with a high correlation, as happens with number of words, length of tweets, and number of syllables. Orthographic and misspelled errors are other relevant features to discern among authors by correction and style (COR). The usage of augmentative suffixes, used for emphasis, and the number of positive words are also relevant. It can be observed that the discriminatory categories of LF for the authorship attribution task are different from the author profiling task. On the one hand, linguistic features related to lexical variety and morphological features are more relevant to discern between demographic traits such as gender or age range, and psychographic traits such as political ideology. On the other hand, those features related to stylometry are more relevant for authorship attribution.

6. Conclusions and further work

The focus of this work lies on determining digital footprints through psychographic traits based on political affiliation. Specifically, two major tasks regarding authorship analysis have been conducted. On the one hand, from an author profiling perspective, psychographic traits such as political affiliation and demographic traits have been studied. On the other hand, we have evaluated authorship attribution based on partially anonymised documents. Both tasks are grounded on the usage of interpretable linguistic features, embedding-based features, and n-grams. These features have been evaluated in isolation and combined into pairs with different neural network architectures, including shallow, deep, convolutional, and recurrent neural networks.

The main contribution of this work is the release of the PoliCorpus-2020, a dataset compiled from Twitter accounts of Spanish politicians during 2020, as well as the release of the source code of the project in https://github.com/Smolky/FGCS-

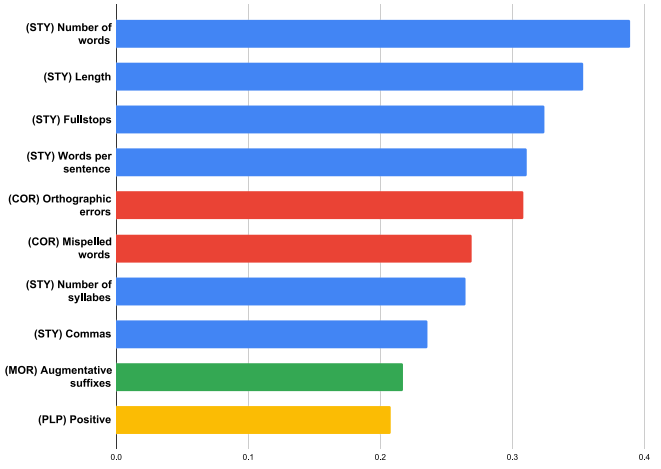


Fig. 5. Top 10 linguistic features ranked by mutual information for the authorship attribution task.

political-ideology-2021.[6] We want to emphasise that this methodology for compiling the corpus can be applied to conduct similar studies in other countries, as we indicate methods for automatically compiling and labelling the datasets. However, one of the main limitations is that if all users share the same profession, the resulting models can be somehow biased and, therefore, it is difficult to guarantee that the models learned can be transferred to the rest of the citizens. Although we have included an evaluation of these models with accounts of journalists, whose ideology can also be inferred, this validation is still incomplete. In our understanding, the deep complexity of evaluating these methods with a representative sample of the society is that political ideology is something internal for each one, so it does not make sense for external annotators to tell which political ideology of a person is, for each one individually. We believe that a more equitable approach would involve the request of volunteers to allow the researchers to consult their texts and to check their political ideology.

During our evaluation, we have found that LF are effective for conducting authorship analysis tasks. Specifically, the combination of linguistic features with embedding-based features generally boosts the results achieved separately. This behaviour was observed with all feature combinations of the author attribution task and the combination of LF with contextual embeddings (SBE and BETO) in the author profiling task. This finding indicates that LF and contextual embeddings extract complementary information. We have also found that the linguistic features for conducting author profiling and authorship analysis are different. The analysis of the results of each trait and the authorship attribution task revealed that demographic traits are more closely related to lexical and morphosyntactic features, whereas authorship attribution is more related to stylometric ones. However, our results suggest that the best feature set combination depends on the task, as we have observed that the combination of LF with SE or SBE outperforms the word embeddings in the authorship attribution task, PWE achieves better results for binary political spectrum and transformers with LF for multiclass political spectrum. We have also found that the usage of n-gram features are less useful when the context of the tweets changes, as we observed an important drop on the performance when we evaluated average users. Moreover, we find that they are also less effective when dealing with fine-grained political affiliation. In contrast, the usage of BERT combined with LF has provided the overall best results in all tasks.

It is worth noting the resources employed, both time and memory, to train the neural networks. We observed significant differences among the feature set and the neural network architecture. The models based on the feature sets that compact the information on a fixed size, such as LF or any form of sentence embeddings (SE, SBE), are able to complete an epoch of training in a few seconds. However, networks that employed an embedding layer with CNN o RNN increased significantly the resources needed. The training time for a single epoch of a convolutional neural network was around 20 s, but with BiGRU the time required varied between 384.5 and 700 s, depending on the hyperparameters and the complexity of the neural network.[7] This excess was due to the high number of parameters for training, reaching more than 15 million parameters in shallow networks composed with only one hidden layer plus the embedding layer. In case of neural networks based on transformers, the training time reached about 3 h for completing 3 epochs, including the time required for the evaluation of the validation dataset. Due to the high computational demand, we limited the number of hyperparameters evaluated with transformers, as well as with convolutional and recurrent neural networks. Moreover, we are aware that the results could be more robust if nested cross-validation is applied. However, we consider that this approach complicates the replication of these results and the comparison with further methods.

In order to build interpretable models, we consider that linguistic features and neural networks are an interesting research direction regarding NLP. Moreover, we have relied on authors' writings without taking into account contextual features like when and how users interact with each other in social networks. On this premise, these kinds of features could provide more insights regarding social behaviour. In the same line, it is possible to incorporate other sources of information regardless of texts and explore the analysis of personality and affiliation traits from a multi-modal perspective by analysing the multimedia content shared in the tweets.

Regarding the techniques employed, as future work we will evaluate the reliability of applying ML ensembles as a means to combine different feature sets. Regarding the dataset, one limitation of the PoliCorpus-2020 is that it only comprises writings by politicians and journalists. A promising research direction, consequently, is to evaluate the automatic classifiers developed in this work with random users and evaluate whether the same linguistic features apply to determine their political ideology. Another research direction is that we have limited this study to the Spanish language discarding those tweets written in English, Italian, France, Basque or Catalan to determine if being a polyglot could have an impact on demographic and psychographic traits.

Finally, we consider that the PoliCorpus'2020 is a valuable source to carry out research related to the identification of Hate Speech spreaders on social networks. We argue that understanding affiliation traits could also help to prevent hate speech, by detecting profiles on social networks that spread offencive messages to vulnerable groups. There is previous research focused on determining what kind of messages are hate speech in political communication, such as the works described at [57], as well as proposed taxonomies for identifying stereotypes about hate-speech and stereotypes towards immigrants, such as in the works described in [58] and [59], respectively. In this sense, we propose to evaluate the taxonomy described in [58] on the

[6] This repository is private right now. We will make it public in case of acceptance of the paper. Meanwhile, the scripts are shared with the reviewers in the following link: https://pln.inf.um.es/corpora/politics/source-code.rar.

[7] The models were trained with an Intel(R) Xeon(R) Gold 5218 CPU @ 2.30 GHz, 64 cores, and 500 Megabytes of RAM.

PoliCorpus'2020 to compare the results. We also consider extracting personality traits from politicians and compare them according to their political ideology to observe differences among the Big Five personality traits, namely openness, conscientiousness, agreeableness, extraversion, or neuroticism [60]. Another promising research direction is to extract emotions about the topics identified in the PoliCorpus'2020 and compare them to past events, such as the Spanish dataset concerning the 2019 10N Spanish elections [61] and datasets beyond the political scope [62] to check whether the results are transferable.

CRediT authorship contribution statement

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

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.

Acknowledgements

This paper is part of the research project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/ AEI/10.13039/501100011033. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado industrial programme.

References

[1] C. Stachl, F. Pargent, S. Hilbert, G.M. Harari, R. Schoedel, S. Vaid, S.D. Gosling, M. Bühner, Personality research and assessment in the era of machine learning, Euro. J. Personal. 34 (5) (2020) 613–631.

[2] J.T. Yun, C.M. Segijn, S. Pearson, E.C. Malthouse, J.A. Konstan, V. Shankar, Challenges and future directions of computational advertising measurement systems, J. Advert. 49 (4) (2020) 446–458, http://dx.doi.org/10.1080/00913367.2020.1795757, arXiv:https://doi.org/10.1080/00913367.2020.1795757.

[3] B. Verhulst, L.J. Eaves, P.K. Hatemi, Correlation not causation: The relationship between personality traits and political ideologies, Am. J. Polit. Sci. 56 (1) (2012) 34–51.

[4] M. Fatke, Personality traits and political ideology: A first global assessment, Polit. Psychol. 38 (5) (2017) 881–899.

[5] B. Baumgaertner, J.E. Carlisle, F. Justwan, The influence of political ideology and trust on willingness to vaccinate, PLoS One 13 (1) (2018) e0191728.

[6] S.M. Cruz, The relationships of political ideology and party affiliation with environmental concern: A meta-analysis, J. Environ. Psychol. 53 (2017) 81–91, http://dx.doi.org/10.1016/j.jenvp.2017.06.010, URL https://www.sciencedirect.com/science/article/pii/S027249441730083X.

[7] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artif. Intell. Rev. (2019) 1–27.

[8] B. Zarouali, T. Dobber, G. De Pauw, C. de Vreese, Using a personality-profiling algorithm to investigate political microtargeting: assessing the persuasion effects of personality-tailored ads on social media, Commun. Res. (2020) 0093650220961965.

[9] S.C. Matz, R.E. Appel, M. Kosinski, Privacy in the age of psychological targeting, Curr. Opin. Psychol. 31 (2020) 116–121, http://dx.doi.org/10.1016/j.copsyc.2019.08.010, URL https://www.sciencedirect.com/science/article/pii/S2352250X19301332, Privacy and Disclosure, Online and in Social Interactions.

[10] C. Makridis, J.T. Rothwell, The real cost of political polarization: Evidence from the COVID-19 pandemic, 2020, Available At SSRN 3638373.

[11] C.A. Tisdell, Economic, social and political issues raised by the COVID-19 pandemic, Econ. Anal. Policy 68 (2020) 17–28, http://dx.doi.org/10.1016/j.eap.2020.08.002, URL http://www.sciencedirect.com/science/article/pii/S0313592620304082.

[12] D. Azucar, D. Marengo, M. Settanni, Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis, Personal. Individ. Differ. 124 (2018) 150–159, http://dx.doi.org/10.1016/j.paid.2017.12.018, URL https://www.sciencedirect.com/science/article/pii/S0191886917307328.

[13] V. Balakrishnan, S. Khan, T. Fernandez, H.R. Arabnia, Cyberbullying detection on twitter using Big Five and Dark Triad features, Personal. Individ. Differ. 141 (2019) 252–257, http://dx.doi.org/10.1016/j.paid.2019.01.024, URL https://www.sciencedirect.com/science/article/pii/S0191886919300364.

[14] F.R. Gallo, G.I. Simari, M.V. Martinez, M.A. Falappa, Predicting user reactions to Twitter feed content based on personality type and social cues, Future Gener. Comput. Syst. 110 (2020) 918–930, http://dx.doi.org/10.1016/j.future.2019.10.044, URL https://www.sciencedirect.com/science/article/pii/S0167739X19304091.

[15] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, J. Am. Soc. Inf. Sci. Technol. 60 (1) (2009) 9–26.

[16] R. Abooraig, S. Al-Zu'bi, T. Kanan, B. Hawashin, M. Al Ayoub, I. Hmeidi, Automatic categorization of Arabic articles based on their political orientation, Digit. Invest. 25 (2018) 24–41.

[17] R.G. van Dalen, L.R. Melein, B. Plank, Profiling dutch authors on twitter: Discovering political preference and income level, Comput. Linguist. Netherlands J. 7 (2017) 79–92.

[18] M. Dahllöf, Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability, Lit. Linguist. Comput. 27 (2) (2012) 139–153, http://dx.doi.org/10.1093/llc/fqs010, arXiv:https://academic.oup.com/dsh/article-pdf/27/2/139/2749255/fqs010.pdf.

[19] R. Baly, G. Da San Martino, J. Glass, P. Nakov, We can detect your bias: Predicting the political ideology of news articles, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, EMNLP '20, 2020, pp. 4982–4991.

[20] V. Mercado, A. Villagra, M. Errecalde, Political alignment identification: a study with documents of Argentinian journalists, J. Comput. Sci. Tech. 20 (1) (2020) e05.

[21] S.H.H. Ding, B.C.M. Fung, F. Iqbal, W.K. Cheung, Learning stylometric representations for authorship analysis, IEEE Trans. Cybern. 49 (1) (2019) 107–121, http://dx.doi.org/10.1109/TCYB.2017.2766189.

[22] A. Almela, G. Alcaraz-Mármol, A. García-Pinar, C. Pallejá, Developing and analyzing a spanish corpus for forensic purposes, Linguist. Evid. Secur. Law Intell. 3 (2019).

[23] F.M. Rangel, 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.

[24] M. Wiegmann, B. Stein, M. Potthast, Overview of the celebrity profiling task at PAN 2019, 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. 402–416, URL http://ceur-ws.org/Vol-2380/paper_246.pdf.

[25] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, M. Potthast, Semeval-2019 task 4: Hyperpartisan news detection, in: J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S.M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, Association for Computational Linguistics, 2019, pp. 829–839, http://dx.doi.org/10.18653/v1/s19-2145.

[26] E. Amigó, J.C. de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. de Rijke, D. Spina, Overview of RepLab 2014: Author profiling and reputation dimensions for online reputation management, in: E. Kanoulas, M. Lupu, P.D. Clough, M. Sanderson, M.M. Hall, A. Hanbury, E.G. Toms (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings, in: Lecture Notes in Computer Science, vol. 8685, Springer, 2014, pp. 307–322, http://dx.doi.org/10.1007/978-3-319-11382-1_24.

[27] F.M. Rangel, 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. 352–365, URL http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-RangelEt2013.pdf.

[28] P. Juola, Industrial uses for authorship analysis, Math. Comput. Sci. Ind. (2015) 21–25.

[29] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al., Overview of PAN 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on Twitter, and style change detection, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2020, pp. 372–383.

[30] G.S. Reddy, T.M. Mohan, T.R. Reddy, Author profiling approach for location prediction, in: First International Conference on Artificial Intelligence and Cognitive Computing, Springer, 2019, pp. 389–395.

[31] M.P. Villegas, M.J. Garciarena Ucelay, M.L. Errecalde, L. Cagnina, A Spanish text corpus for the author profiling task, in: XX Congreso Argentino de Ciencias de la ComputaciÓN, Buenos Aires, 2014, pp. 621–630.

[32] F.M. Rangel, 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.

[33] F.M. Rangel, 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. 1–36, URL http://ceur-ws.org/Vol-2380/paper_263.pdf.

[34] J. Bevendorff, B. Chulvi, G.L.D. la Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2021: Authorship verification, profiling hate speech spreaders on Twitter, and style change detection, in: K.S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, in: Lecture Notes in Computer Science, vol. 12880, Springer, 2021, pp. 419–431, http://dx.doi.org/10.1007/978-3-030-85251-1_26.

[35] M.A.A. Carmona, E. Guzmán-Falcón, M. Montes-y-Gómez, H.J. Escalante, L.V. nor Pineda, V. Reyes-Meza, A.R. Sulayes, Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in mexican spanish tweets, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J.C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-Located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, in: CEUR Workshop Proceedings, vol. 2150, CEUR-WS.org, 2018, pp. 74–96, URL http://ceur-ws.org/Vol-2150/overview-mex-a3t.pdf.

[36] P. Mishra, M. Del Tredici, H. Yannakoudakis, E. Shutova, Author profiling for abuse detection, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1088–1098.

[37] M. Potthast, F. Rangel, M. Tschuggnall, E. Stamatatos, P. Rosso, B. Stein, Overview of PAN'17, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2017, pp. 275–290.

[38] A. Rocha, W.J. Scheirer, C.W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A.R. Carvalho, E. Stamatatos, Authorship attribution for social media forensics, IEEE Trans. Inf. Forensics Secur. 12 (1) (2016) 5–33.

[39] Q. Zheng, X. Tian, M. Yang, H. Wang, The email author identification system based on support vector machine (SVM) and analytic hierarchy process (AHP), IAENG Int. J. Comput. Ence 46 (2PT. 141-263) (2019) 178–191.

[40] M. Abuhamad, J. su Rhim, T. AbuHmed, S. Ullah, S. Kang, D. Nyang, Code authorship identification using convolutional neural networks, Future Gener. Comput. Syst. 95 (2019) 104–115, http://dx.doi.org/10.1016/j.future.2018.12.038, URL https://www.sciencedirect.com/science/article/pii/S0167739X18315528.

[41] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, B. Stein, Overview of the cross-domain authorship verification task at PAN 2020, 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. 372–383, URL http://ceur-ws.org/Vol-2696/paper_264.pdf.

[42] K. Luyckx, W. Daelemans, The effect of author set size and data size in authorship attribution, Lit. Linguist. Comput. 26 (1) (2011) 35–55.

[43] J.A. García-Díaz, A. Almela, G. Alcaraz-Mármol, R. Valencia-García, UmuCorpusClassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks, Procesamiento Del Leng. Nat. 65 (2020) 139–142, URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6292.

[44] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 427–431.

[45] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, FastText.zip: Compressing text classification models, 2016, arXiv preprint arXiv:1612.03651.

[46] S. Daneshvar, D. Inkpen, Gender identification in Twitter using N-grams and LSA: notebook for PAN at CLEF 2018, 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, URL http://ceur-ws.org/Vol-2125/paper_213.pdf.

[47] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (1) (2010) 24–54.

[48] J.A. García-Díaz, M. Cánovas-García, R. Valencia-García, Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America, Future Gener. Comput. Syst. 112 (2020) 641–657, http://dx.doi.org/10.1016/j.future.2020.06.019, URL http://www.sciencedirect.com/science/article/pii/S0167739X2030892X.

[49] J.A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings, Future Gener. Comput. Syst. 114 (2021) 506–518, http://dx.doi.org/10.1016/j.future.2020.08.032, URL http://www.sciencedirect.com/science/article/pii/S0167739X20301928.

[50] E. Fersini, E. Messina, F.A. Pozzi, Expressive signals in social media languages to improve polarity detection, Inf. Process. Manage. 52 (1) (2016) 20–35.

[51] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, M. Baroni, What you can cram into a single vector: Probing sentence embeddings for linguistic properties, in: ACL 2018-56th Annual Meeting of the Association for Computational Linguistics, Vol. 1, Association for Computational Linguistics, 2018, pp. 2126–2136.

[52] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, European Language Resources Association (ELRA), Miyazaki, Japan, 2018, URL https://aclanthology.org/L18-1550.

[53] 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.

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

[55] 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.

[56] C. Zhang, M. Abdul-Mageed, BERT-based arabic social media author profiling, 2019, arXiv preprint arXiv:1909.04181.

[57] S. Jaki, T.D. Smedt, Right-wing german hate speech on Twitter: Analysis and automatic detection, 2019, arXiv:1910.07518, URL http://arxiv.org/abs/1910.07518.

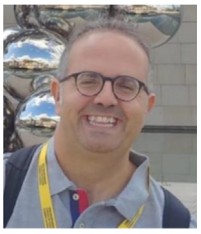
[58] J. Sánchez-Junquera, B. Chulvi, P. Rosso, S.P. Ponzetto, How do you speak about immigrants? Taxonomy and StereoImmigrants dataset for identifying stereotypes about immigrants, Appl. Sci. 11 (8) (2021) 3610.

[59] J. Sánchez-Junquera, P. Rosso, M. Montes, B. Chulvi, et al., Masking and BERT-based models for stereotype identication, Procesamiento Leng. Nat. 67 (2021) 83–94.

[60] A. Giachanou, E.A. Ríssola, B. Ghanem, F. Crestani, P. Rosso, The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers, in: International Conference on Applications of Natural Language To Information Systems, Springer, 2020, pp. 181–192.
[61] J. Sánchez-Junquera, S.P. Ponzetto, P. Rosso, A Twitter political corpus of the 2019 10n Spanish election, in: International Conference on Text, Speech, and Dialogue, Springer, 2020, pp. 41–49.
[62] 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.

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

Ricardo Colomo-Palacios Ricardo Colomo-Palacios is a Full Professor at the Computer Science Department of the Østfold University College, Norway. Formerly he worked at Universidad Carlos III de Madrid, Spain. His research interests include applied research in information systems, software project management, people in software projects, business software, software and services process improvement and web science. He received his Ph.D. in Computer Science from the Universidad Politécnica of Madrid (2005). He also holds a MBA from the Instituto de Empresa (2002). He has been working as Software Engineer, Project Manager and Software Engineering Consultant in several companies including Spanish IT leader INDRA. Prof. Dr. Colomo-Palacios is also an Editorial Board Member and Associate Editor for several international journals.



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).