



# Rhythmic and Psycholinguistic Features for Authorship Tasks in the Spanish Parliament: Evaluation and Analysis

Silvia Corbara<sup>1(✉)</sup>, Berta Chulvi<sup>2,3</sup>, Paolo Rosso<sup>2</sup>, and Alejandro Moreo<sup>4</sup>

<sup>1</sup> Scuola Normale Superiore, Pisa, Italy  
silvia.corbara@sns.it

<sup>2</sup> Universitat Politècnica de València, Valencia, Spain  
berta.chulvi@upv.es, proso@dsic.upv.es

<sup>3</sup> Universitat de València, Valencia, Spain

<sup>4</sup> Istituto di Scienza e Tecnologie dell'Informazione, CNR, Pisa, Italy  
alejandro.moreo@isti.cnr.it

**Abstract.** Among the many tasks of the authorship field, Authorship Identification aims at uncovering the author of a document, while Author Profiling focuses on the analysis of personal characteristics of the author(s), such as gender, age, etc. Methods devised for such tasks typically focus on the *style* of the writing, and are expected not to make inferences grounded on the *topics* that certain authors tend to write about. In this paper, we present a series of experiments evaluating the use of topic-agnostic feature sets for Authorship Identification and Author Profiling tasks in Spanish political language. In particular, we propose to employ features based on rhythmic and psycholinguistic patterns, obtained via different approaches of text masking that we use to actively mask the underlying topic. We feed these feature sets to a SVM learner, and show that they lead to results that are comparable to those obtained by a BETO transformer, when the latter is trained on the original text, i.e., potentially learning from topical information. Moreover, we further investigate the results for the different authors, showing that variations in performance are partially explainable in terms of the authors' political affiliation and communication style.

**Keywords:** Authorship Analysis · Text masking · Political speech

## 1 Introduction

In the authorship analysis field, Authorship Identification (AId) investigates the true identity of the author of a written document, and it is of special interest when the author is unknown or debated. Two of the main sub-tasks of AId are Authorship Attribution (AA) and Authorship Verification (AV): in the former, given a document  $d$  and a set of candidate authors  $\{A_1, \dots, A_m\}$ , the goal is to identify the real author of  $d$  among the set of candidates; instead, AV is defined as a binary classification task, in which the goal is to infer whether  $A$  (the only

candidate) is the real author of  $d$  or not. On the other hand, Author Profiling (AP) aims at distinguish among classes of authors, in order to investigate the authors’ individual characteristics, which can span from their gender, to their nationality, to their mental state; these studies are especially important in order to find common traits among groups, or to reveal relevant information about a specific author. While tackling these classification problems, the researchers’ goal is to devise methods able to discriminate the different styles of the authors under consideration, often relying on supervised machine learning.

In this article, we evaluate the use of topic-agnostic feature sets for AV, AA and various AP tasks (by gender, by age, and by political affiliation) for Spanish texts. Along with feature sets that are by now consolidated in the authorship field, we introduce rhythm- and psycholinguistics-based feature sets. Concretely, we propose to generate new masked versions of the original text extracting (i) the syllabic stress (i.e., strings of *stressed* and *unstressed* syllables), and (ii) the psycholinguistic categories of the words, as given by the LIWC dictionary (Sect. 3.2). The resulting representations are topic-agnostic strings from which we extract  $n$ -grams features. In order to asses the effect of our proposed feature sets on the performance, we carry out experiments of *ablation* (in which we remove one feature set from the model at a time) and experiments of *addition* (in which we add one single feature set to the model at a time). Our results seem to indicate that our topic-agnostic features bring to bear enough authorial information as to perform on-par with BETO, the Spanish equivalent to the popular BERT transformer, fine-tuned on the original (hence topic-aware) text. The code of the project can be found at GitHub.<sup>1</sup>

This work continues the preliminary experiments presented in the short paper by Corbara et al. [8]. With respect to the previous paper, we present a wider set of experiments, both in terms of size and diversity of the dataset, as well as in terms of the number of tasks. In the current paper, we consider a higher number of authors characterized by a finer-grained political spectrum, and we add experiments of AP. Moreover, we devise a new experimental protocol (Sect. 3.3) and we take an initial step towards interpreting the impact that different feature sets bring to bear in the AV task (Sect. 5).

## 2 Related Work

The annual PAN<sup>2</sup> event presents various shared tasks both for AId and AP, often with challenging settings including the open-set or the cross-domain problems, and thus offers a very good overview of the most recent trends in this field. For example, in the recent 2021 edition [1], the participants tackled an AId task<sup>3</sup> and an AP task (based on the problem of identifying hate speech spreaders). In this

<sup>1</sup> [https://github.com/silvia-cor/Topic-agnostic\\_ParlaMintES](https://github.com/silvia-cor/Topic-agnostic_ParlaMintES).

<sup>2</sup> <https://pan.webis.de/>.

<sup>3</sup> Precisely, the PAN2021 event presented a particular case of AV where the dataset contained pairs of documents, and the aim was to infer whether the two documents shared the same author; we call this task Same-Authorship Verification (SAV).

occasion is was observed that, although the presence of Neural Network (NN) methods increased with respect to past editions, and even though the best results in the AId and AP challenges were obtained by NN methods, simpler approaches based on (character or word)  $n$ -grams and traditional classification algorithms were still competitive; indeed, they were found to outperform NN methods in past editions [18]. In fact, the method by Weerasinghe et al. [27], based on the absolute difference among feature vectors fed to a logistic regression classifier, reached the third (with the large dataset) and fourth (with the small dataset) positions in the overall ranking for the AId task, while in the AP task only 7 out of 66 methods were able to surpass the baseline Support Vector Machine (SVM) fed with character  $n$ -grams.

In classical machine learning algorithms, the choice of the features to employ is crucial. In his survey, Stamatatos [24] discusses the features that are most commonly used in the authorship field; however, he also notes that features such as word and character  $n$ -grams might prompt methods to base their inferences on topic-related patterns rather than on stylometric patterns. In fact, an authorship classifier (even a seemingly good one) might end up unintentionally performing topic identification if domain-dependent features are used [2]. In order to avoid this, researchers might limit their scope to features that are clearly topic-agnostic, such as function words or syntactic features [15], or might actively mask topical content via a text-masking approach [14, 25]. Continuing the first experimentation [8], in this project we focus our attention on features capturing the rhythmic and the psycholinguistic traits of the texts, by employing a text-masking technique based on syllabic stress and LIWC categories.

The idea of employing rhythmic features in the authorship field is not a new one. Their most natural use is in studies focused on poetry [19], although they have also been employed in authorship analysis of prose texts. In the work by Plecháč [23], the role of accent, or stress, is studied for AId problems in English; in the research by Corbara et al. [9], the documents are encoded as sequences of long and short syllables, from which the relevant features are extracted and used for AA in Latin prose texts, with promising results. Since Spanish derives from Latin, we aim to investigate the extent to which similar considerations apply to the Spanish language as well. In this case, we exploit the concept of *stress*, which gained relevance over the concept of *syllabic quantity* in Romance languages.

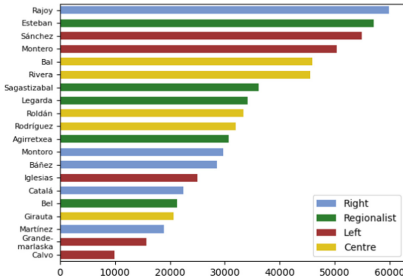
Linguistic Inquiry and Word Count (LIWC) [21] is a well-known software application for text analysis. LIWC is built around a word dictionary where each entry is associated with one or more categories related to grammar, emotions, or other cognitive processes and psychological concepts. Nowadays, LIWC has become a popular tool for the study of psychological aspects of textual documents, usually by employing the relative frequency of each LIWC category. It has been profitably used for the characterization of a “psychological profile” or a “mental profile mapping” for authorship studies [4, 13], and also for the analysis of speeches regarding the Spanish political debate [11]. In a similar vein, García-Díaz et al. [12] designed UMUTextStats, a LIWC-inspired tool, and studied its application to AA and various AP tasks (gender, age range, and political spectrum) on a dataset of Spanish political tweets.

### 3 Experimental Setting

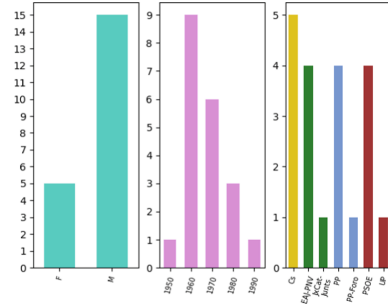
#### 3.1 Dataset: ParlaMint

In this project, we employ the Spanish repository (covering the years 2015–2020) of the *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1* by the digital infrastructure CLARIN,<sup>4</sup> which contains the annotated transcriptions of many sessions of various European Parliaments. Because of their declamatory nature, between the written text and the discourse, these speeches seem particularly suited for an investigation on rhythm and psycholinguistic traits. Apart from lowercasing the text, we did not apply any further pre-processing steps.

In order to have a balanced dataset, we select the parties with more than 300 speeches in the dataset and assign them to the Left, Right, Centre, or Regionalist<sup>5</sup> wing. In particular, we assign PSOE and UP to the Left, PP and PP-Foro to the Right, EAJ-PNV and JxCat-Junts to the Regionalist wing, and only the Ciudadanos (Cs) party to the Centre. We then delete all the speeches that have less than 50 words, and for each wing we select the 5 authors with most speeches in the dataset. The minimum number of samples per author is 70 (Bal Francés), while the maximum is 467 (Sánchez Pérez-Castejón). We randomly select 50 samples for each author to compose the training set, keeping all the remaining samples as test instances. We thus obtain 1000 training samples and 3048 test samples in total. Figure 1 reports the total number of words per author in the training set, divided by political wing, while Fig. 2 reports the distribution of authors by gender, age, and political party.



**Fig. 1.** Total number of words for each speaker in the training set, grouped by political wing.



**Fig. 2.** Number of speakers for each category in gender, age, and political party.

<sup>4</sup> <https://www.clarin.si/repository/xmlui/handle/11356/1431>.

<sup>5</sup> Regionalist parties aim for more political power for regional entities.

<sup>6</sup> Note that we use the decade of birth as representation of age group. We assign the closest decade label to each author’s birth; for example, an author born in 1984 is assigned the label ‘1980’, while an author born in 1987 is assigned the label ‘1990’.

### 3.2 Feature Extraction: BaseFeatures and Text Encodings

Our focus in this research is to evaluate the employment of rhythm- and psycholinguistics-based features for AId and AP tasks. To this aim, we explore various combinations of feature sets, including other topic-agnostic feature sets commonly used in literature. In particular, we follow the same feature-extraction approach as in the preliminary experiments [8].

As a starting point, we employ a feature set comprised of features routinely used in the authorship field, including the relative frequencies of: function words (using the list provided by the NLTK library<sup>7</sup>), word lengths, and sentence lengths. We set the range of word (sentence) lengths to  $[1, n]$ , where  $n$  is the longest word (sentence) appearing at least 5 times in the training set. We call this feature set BASEFEATURES.

We also employ a text-masking approach, where we replace each word in the document with the respective Part-of-Speech tag (we exploit the POS annotation already available in the ParlaMint dataset). From the encoded text, we then extract the word  $n$ -grams in the range  $[1, 3]$  and compute the TfIdf weights, which we use as features. We call this feature set POS.

We follow a similar approach to extract the rhythm of the discourse, i.e., we convert the document into a sequence of stressed and unstressed syllables, using the output of the RANTANPLAN library;<sup>8</sup> from this encoding, we extract the character  $n$ -grams in the range  $[1, 7]$  and compute the TfIdf weights as features. We call this feature set STRESS.

Similarly, in order to encode the psycholinguistic dimension of the document, we employ the LIWC dictionary.<sup>9</sup> We define three macro-categories from a subset of the LIWC category tags, representing (i) grammatical information, (ii) cognitive processes or actions, and (iii) feelings and emotions.<sup>10</sup> For each macro-category, we perform a separate text masking by replacing each word with the

<sup>7</sup> <https://www.nltk.org/>.

<sup>8</sup> <https://github.com/linhd-postdata/rantanplan>.

<sup>9</sup> We employ the Spanish version of the dictionary, which is based on LIWC2007.

<sup>10</sup> We use the following categories: (i) YO, NOSOTRO, TUUTD, ELELLA, VOSUTDS, ELLOS, PASADO, PRESENT, FUTURO, SUBJUNTIV, NEGACIO, CUANTIF, NUMEROS, VERBYO, VERBTU, VERBNOS, VERBVOS, VERBOSEL, VERBELLOS, FORMAL, INFORMAL; (ii) MECCOG, INSIGHT, CAUSA, DISCREP, ASENTIR, TENTAT, CERTEZA, INHIB, INCL, EXCL, PERCEPT, VER, OIR, SENTIR, NOFLUEN, RELLENO, INGERIR, RELATIV, MOVIM; (iii) MALDEC, AFECT, EMOPOS, EMONEG, ANSIEDAD, ENFADO, TRISTE, PLACER. We avoid employing categories that would repeat information already captured by the POS tags, or topic-related categories (e.g., DINERO, FAMILIA).

**Table 1.** Example of the encodings employed in this project. Note there is not a one-to-one correspondence between syllables and stresses due to linguistic phenomena across word boundaries (e.g., synalepha), which RANTANPLAN accounts for.

Original text:	Gracias	No	hay	que	restituir	lo	que	no	ha	existido			
POS:	NOUN	PUNCT	ADV	AUX	SCONJ	VERB	PRON	PRON	ADV	AUX	VERB	PUNCT	
LIWC.GRAM:	w		NEGACIO	PRESENT	w	w	ELELLA	w	NEGACIO	PRESENT	VERBOS	EL	w
LIWC.COG:	w		w	w	MecCOG	w	w	MecCOG	w	w			w
LIWC.FEELS:	AFFECT	EMO	POS	w	w	w	w	w	w	w			w
STRESS:	+ - + - - - + - - + - + -												
English translation:	Thank you. There is no need to return what has not existed												

corresponding LIWC category tag.<sup>11</sup> From a single encoding, we extract the word  $n$ -grams in the range  $[1, 3]$  and compute the TfIdf weights as features. We call these feature sets LIWC\_GRAM, LIWC\_COG, and LIWC\_FEELS, respectively. We show an example of all the encodings in Table 1.

### 3.3 Experimental Protocol

We perform AId experiments in two settings: Authorship Verification (AV) for each author (where each test sample is labelled as belonging to that author, or not) and Authorship Attribution (AA) (where each sample is labelled as belonging to one of the 20 authors). We perform AP experiments by labelling each sample based on the gender, age group, political wing or political party of the author it belongs to. We assess the effect of the different feature sets by evaluating the performance of a classifier fed with them. As evaluation measure, for the AV task we use the well-known  $F_1$  function, and for the AA and AP tasks we use the macro-averaged  $F_1$  ( $F_1^M$ ) and micro-averaged  $F_1$  ( $F_1^\mu$ ) variants.

We employ SVM as learner;<sup>12</sup> the implementation we employ in this study is the SVC module from the `scikit-learn` package.<sup>13</sup> We perform the optimisation of various hyper-parameters: the parameter  $C$ , which sets the trade-off between the training error and the margin ( $[0.001, 0.01, 0.1, 1, 10, 100, 1000]$ ), the kernel function (*linear*, *poly*, *rbf*, *sigmoid*), and whether the classes weights should be balanced or not. The optimization is computed in a grid-search fashion,

<sup>11</sup> Formally, LIWC can be seen as a map  $m : w \rightarrow C$ , where  $w$  is a word token and  $C \subset \mathcal{C}$  is a subset of the psycholinguistic categories  $\mathcal{C}$ . Given a macro-category  $M \subset \mathcal{C}$ , we replace each word  $w$  in a document by the categories  $m(w) \cap M$ . If  $|m(w) \cap M| > 1$ , then a new token is created which consists of a concatenation of the category names (following a consistent ordering). If  $m(w) \cap M = \emptyset$ , then  $w$  is replaced with the ‘w’ symbol. (Note that some entries in LIWC have the suffix truncated and replaced with an asterisk ‘\*’, e.g., *president\**; the asterisk is treated as a wildcard in the mapping function, and in case more than one match is possible, the match with the longest common prefix is returned).

<sup>12</sup> We also carried out preliminary experiments with Random Forest (RF) and Logistic Regression (LR). SVM showed a remarkably better performance than RF, while no significant differences were noticed between SVM and LR.

<sup>13</sup> <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>.

via 5-fold cross-validation on the training set. The best model is then retrained on the whole training set and is used to make predictions on the test set samples.

We apply a feature selection approach, since the number of different features generated using the LIWC encodings tends to be very high.<sup>14</sup> We keep only the 10% most important features (employing  $\chi^2$  as score measure of importance) for each feature set derived from the LIWC encodings.<sup>15</sup>

Finally, we also compare the results obtained with the aforementioned features with the results obtained by a method trained on the original text (hence, potentially mining topic-related patterns). To this aim, we employ the pre-trained transformer named ‘BETO-cased’, from the Huggingface library [6],<sup>16</sup> with the learning rate set to  $10^{-6}$  and the other hyper-parameters set as default. We fine-tune the model for 50 epochs on the training set.

## 4 Results

We show the results of the AV experiments in Table 2. In the first batch of results, we show the performance of the feature sets in the experiments “by addition”, using the BASEFEATURES set as a baseline. In the second batch of results, we report the experiments “by ablation”, where we subtract each feature set from the combination of all the feature sets we are exploring (named ALL). These results are obtained using a SVM learner. Finally, we also report the results obtained using the BETO transformer. Moreover, we perform the non-parametric McNemar’s paired test of statistical significance between the results obtained using our best SVM configuration and the results obtained using BETO [6], for each of the authorship tasks. We take 0.05 as the confidence level.

BETO obtains the best result in 10 out of 20 cases, 5 of which are statistically significant; conversely, the SVM classifier obtains the best performance 10 out of 20 cases, 7 of which are statistically significant. Thus, we might consider the performance of the two methods comparable, even though SVM does not exploit any topic-related information (as the BETO transformer instead could do). Focusing on the SVM results, we observe that the best-performing feature set is often (in 10 out of 20 cases) the one combining BASEFEATURES and POS, confirming that the syntactic encoding is a good indicator of style. The other best-resulting feature sets are mostly different “ablations” of the ALL set. In particular, it seems that the LIWC\_FEELS features are rather detrimental, since the configuration ALL - LIWC\_FEELS yields the best results in 5 cases. In line with our preliminary results [8], we observe severe fluctuations in performance across the authors, with the best result and relative merits of each of

<sup>14</sup> Indeed, LIWC\_GRAM, LIWC\_COG and LIWC\_FEELS create the highest number of features in our experiments, ranging from 3000 to more than 20000.

<sup>15</sup> The selection is always carried out in the training set. During the 5-fold cross-validation optimization phase, feature selection is carried out in the corresponding 80% of the training set used as training.

<sup>16</sup> <https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased>. This model obtained better results than the ‘uncased’ version in preliminary experiments.

**Table 2.** Results of the AV experiments. \*

	Martínez	Sagastizabal	Rodríguez	Legarda	Agirretxea	Girauta	Esteban	Rajoy	Sánchez	Catalá	Montero	Báñez	Iglesias	Rivera	Roldán	Bel	Bal	Calvo	GMarlaska	Montero
BaseFeatures	.616	.626	.272	.580	.277	.100	.510	.420	.528	.397	.468	.671	<b>.482<sup>†</sup></b>	.526	.161	.599	.203	.227	.362	.479
+ POS	<b>.742<sup>†</sup></b>	.761	.524	.672	.560	<b>.188<sup>†</sup></b>	<b>.516<sup>†</sup></b>	<b>.515<sup>†</sup></b>	<b>.640<sup>†</sup></b>	<b>.537<sup>†</sup></b>	<b>.540</b>	.717	.444	<b>.654<sup>†</sup></b>	.400	.609	<b>.449</b>	<b>.481<sup>†</sup></b>	.425	.453
+ STRESS	.675	.618	.293	.586	.359	.086	.458	.414	.464	.373	.488	.645	.389	.557	.136	.662	.171	.204	.321	.381
+ LIWC.GRAM	.529	.517	.091	.621	.329	.060	.365	.165	.535	.277	.379	.539	.319	.443	.080	.460	.047	.189	.295	.431
+ LIWC.COG	.538	.503	.092	.640	.374	.070	.281	.259	.508	.367	.423	.603	.313	.323	.091	.506	.205	.167	.290	.456
+ LIWC.FEELS	.549	.408	.089	.521	.277	.051	.273	.229	.425	.214	.366	.543	.345	.402	.138	.483	.068	.131	.327	.359
ALL	.706	.646	.371	.650	.589	.081	.338	.492	.618	.503	.395	<b>.748</b>	.437	.622	.289	.677	.189	.383	.478	.401
- BaseFeatures	.724	<b>.781</b>	.372	<b>.734</b>	.524	.046	.362	.447	.514	.258	.328	.719	.307	.534	.349	.594	.206	.318	.423	.305
- POS	.599	.415	.229	.543	.403	.078	.291	.348	.504	.347	.356	.657	.403	.463	.237	.568	.090	.209	.409	.346
- STRESS	.723	.655	.441	.629	.545	.036	.379	.477	.602	.489	.409	.745	.437	.636	.310	.654	.189	.394	<b>.480</b>	.447
- LIWC.GRAM	.709	.568	.392	.552	.560	.131	.341	.449	.631	.518	.454	.689	.418	.604	.212	<b>.703<sup>†</sup></b>	.200	.268	.420	.501
- LIWC.COG	.692	.568	.381	.708	.533	.088	.348	.469	.625	.441	.456	.707	.421	.587	.184	.652	.267	.329	.386	.384
- LIWC.FEELS	.710	.611	<b>.526<sup>†</sup></b>	.572	<b>.674</b>	.056	.363	.513	<b>.640<sup>†</sup></b>	.487	.435	.747	.454	.629	<b>.471</b>	.683	.175	.353	.408	<b>.570<sup>†</sup></b>
Beto.base.cased	.836	.798	.314	.771	.632	<b>.247</b>	.352	.757	.388	.729	.610	.800	.437	.460	.468	.601	.381	<b>.494</b>	<b>.664</b>	.426

\* The best result for SVM is in **bold**, while the best overall result is in *italic*; statistical significance is indicated with a † in the best SVM result.

the feature sets being strongly dependent on the author under consideration; for example, the feature set STRESS appears to be beneficial for authors like Agirretxea and Bel, while the same feature set seems to be detrimental for authors like Sagastizabal, Grande-Marlaska, and Montero.

We show the results of the AA experiments in Table 3. In these experiments, the ALL - STRESS and the ALL - LIWC.FEELS feature combinations employing the SVM learner obtain the best results, both outperforming BETO in a statistically significant sense. In fact, consistently with previous results [8], the feature sets LIWC.FEELS and STRESS exhibit a constant disturbance effect.

We show the results of the AP experiments in Table 4. While BETO and SVM do not show remarkable differences for *gender* prediction, SVM excels in the other tasks, with statistical significance in the case of *age* and *political wing*. Consistently with what observed for the AA and AV case, the POS feature set shows a clear proficiency, while the contrary can be said for LIWC.FEELS.

Overall, these experiments allow us to draw some interesting conclusions regarding the features we study for authorship analysis in the Spanish language: on the one hand, TfIdf-weighted n-grams computed on POS-tags encodings are effective for multiple tasks and settings; on the other hand, the feature sets STRESS and LIWC.FEELS tend to fare poorly. Interestingly enough, the combination of multiple topic-agnostic feature sets proved to fare comparably to, and to outperform in some cases, a state-of-the-art neural network that has full access to topic-related information.

## 5 Post-hoc Analysis of the AV Results

Given the differences in performance spotted in the AV results among authors, we further analyse the system’s behaviour in order to outline a suitable explanation for such variances. To this aim, we resort to a series of tools for data analysis: the one-way ANOVA test (Sect. 5.1) and the Spearman test applied to the Analytic



**Table 3.** Results for the AA experiment. \*

	AA	
	$F_1^M$	$F_1^P$
BaseFeatures	.401	.444
+ POS	.570	.620
+ STRESS	.392	.436
+ LIWC_GRAM	.430	.480
+ LIWC_COG	.446	.493
+ LIWC_FEELS	.348	.394
ALL	.580	.631
- BaseFeatures	.545	.599
- POS	.435	.485
- STRESS	<b>.585</b>	<b>.638<sup>†</sup></b>
- LIWC_GRAM	.565	.615
- LIWC_COG	.562	.613
- LIWC_FEELS	<b>.585</b>	<b>.635<sup>†</sup></b>
Beto_base_cased	.417	.471

**Table 4.** Results for the AP experiments. \*

	Gender		Age		Wing		Party	
	$F_1^M$	$F_1^P$	$F_1^M$	$F_1^P$	$F_1^M$	$F_1^P$	$F_1^M$	$F_1^P$
BaseFeatures	.720	.802	.428	.478	.599	.631	.547	.563
+ POS	.751	.828	.545	.592	.685	.715	<b>.642</b>	<b>.681</b>
+ STRESS	.705	.803	.402	.459	.581	.613	.539	.561
+ LIWC_GRAM	.696	.788	.421	.469	.601	.636	.508	.536
+ LIWC_COG	.716	.812	.441	.511	.619	.653	.492	.502
+ LIWC_FEELS	.669	.777	.366	.434	.525	.554	.492	.531
ALL	.736	<b>.854</b>	.551	.589	.690	.719	.604	.648
- BaseFeatures	.709	.843	.485	.540	.650	.681	.552	.614
- POS	.700	.825	.411	.469	.640	.669	.498	.544
- STRESS	.732	.850	.553	.594	<b>.707</b>	<b>.736<sup>†</sup></b>	.610	.658
- LIWC_GRAM	<b>.754</b>	<b>.854</b>	.522	.573	.677	.709	.611	.637
- LIWC_COG	.736	.838	.522	.564	.687	.717	.610	.637
- LIWC_FEELS	.725	.831	<b>.573</b>	<b>.609<sup>†</sup></b>	.706	<b>.737<sup>†</sup></b>	.613	.650
Beto_base_cased	.762	.847	.337	.420	.666	.698	.574	.662

\* The best result for SVM is in **bold**, while the best overall result is in *italic*; statistical significance is indicated with a † in the best SVM result.

Thinking Index (ATI), the Categorical-versus-Narrative Index (CNI), and the Adversarial Style Index (ASI) (Sect. 5.2).

### 5.1 One-Way ANOVA Test for Political Groups

The one-way ANOVA is a parametric test used to check for statistically significant differences in any outcome among groups that are under one categorical variable. We employ this test in order to see if, by grouping the speakers by categories (wing, party, gender, or age), statistically significant differences in performance emerge from the adoption of certain sets of features in the AV task. We use 0.05 as confidence level, and we check that the assumptions for the test (independence, normality and homogeneity of variance) are met. We only consider groups with more than one member, e.g., when grouping by age, we do not consider the groups corresponding to decades 1950 and 1990, since each group would have only one member, Montoro and Rodríguez, respectively.

From this analysis, we have not found any significant difference employing the grouping by gender or by age. However, we have found that the  $F_1$  results display significant differences for multiple feature sets and for BETO if grouped by political party, and especially so if grouped by political wing. The results are reported in Table 5. The Tukey’s Honestly Significant Difference (HSD) [26], a statistical test that compares all possible pairs of means among various result groups, reveals that: i) when grouping by political wing, the significant difference always occurs between the Centre and the other wings, ii) when grouping by political party, the significant difference occurs between Cs and EAJ-PNV, and between Cs and PP. If we focus our attention to the features BASEFEATURES + LIWC\_COG (the only SVM feature setting giving rise to statistically significant differences in performance when the groups are generated by wing, and also when the groups are generated by political party), it turns out that authors belonging to the Centre/Cs obtained  $F_1$  scores significantly lower than authors from other

groups (Fig. 3). In fact, the Cs is a relatively new political party (it was founded in 2006) and its members have been in various different parties before joining it; that could have lead to a certain difficulty in creating a specific personal style.

## 5.2 Spearman Coefficient Applied to Style Indices

The Analytic Thinking Index (ATI), introduced by Pennebaker et al. [22] and formerly named Categorical Dynamic Index, is a unit-weighted score computed from grammatical categories derived from LIWC. This measure is based on the observation that the use of articles and prepositions is associated with a more abstract thinking, while the use of pronouns, auxiliary verbs, conjunctions, adverbs, and negations is associated with a more intuitive and narrative style. The score is obtained by adding the percentages of the former, and subtracting the percentages of the latter. Thus, a positive score denotes a more analytic thinking, while a negative score denotes a more intuitive one. This score has been used to analyze long-term trends in political language in EEUU [17].

The Categorical-versus-Narrative Index (CNI), introduced by Chulvi et al. [7] and inspired by the study of Nisbett et al. [20], is similar in nature to ATI. The score is obtained by adding the percentages of nouns, adjectives, and prepositions, and subtracting the percentages of verbs, adverbs, and personal pronouns. Like ATI, a higher score of CNI denotes a language more focused on the exposition of abstract concepts, while a lower score denotes a language more prone to narration and storytelling.

The Adversarial Style Index (ASI), also proposed by Chulvi et al. [7], is a ratio representing how much an author refers directly to the political adversary in a confronting manner in political debates. The adversarial genre has been vastly studied in parliamentary and election debates, both in the English [5] and Spanish context [3]. The score is the ratio between the sum of the percentages of LIWC categories TuUTD and VosUTDS (2nd person singular and plural Spanish pronouns), and the sum of the percentages of LIWC categories Yo, NOSOTRO (1st person singular and plural Spanish pronouns), TuUTD and VosUTDS.

We show the ATI, CNI and ASI scores computed for each author on the entire dataset in Fig. 4, 5 and 6, respectively.

We employ these measures to quantify the extent to which the AV performance correlates to certain styles of communication. To this aim, in Table 6 we show, for each of the indices, the Spearman correlation coefficient ( $r$ ) between the classification scores and the authors' index scores. We see that BETO displays the strongest positive correlation with respect to ATI and CNI, followed closely by the BASEFEATURES + LIWC\_COG feature set. This seems reasonable, since ATI and CNI hinge upon abstract explanations and concepts, which are captured by LIWC\_COG, while a big portion of the training of BETO is comprised of sources like Wikipedia, legislative texts and talks. Interestingly, the correlations between  $F_1$  scores and the psycholinguistic indexes obtained by each author is found to be statistically significant for more feature sets combinations for CNI (8) than for ATI (1). We hypothesize that this might be due to the fact that ATI, unlike CNI, captures a degree of formality that is rather

**Table 5.** ANOVA p-values on the  $F_1$  results on the AV experiments when grouped by wing or by party.

\*\*

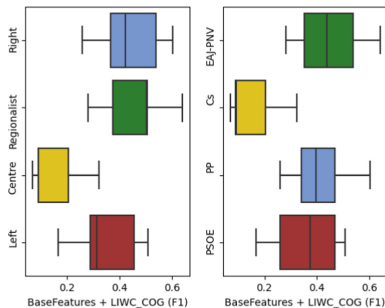
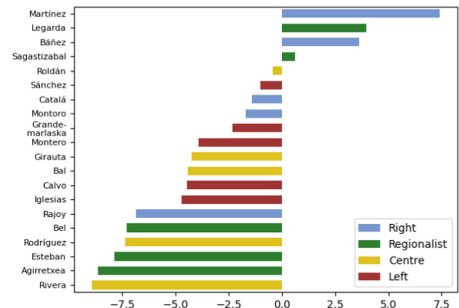
	Wing	Party
BaseFeatures	.026	.085
+ POS	.078	.192
+ STRESS	.014	.060
+ LIWC_GRAM	.023	.054
+ LIWC_COG	.008	.031
+ LIWC_FEELS	.042	.117
ALL	.045	.136
- BaseFeatures	.065	.135
- POS	.045	.133
- STRESS	.071	.185
- LIWC_GRAM	.054	.159
- LIWC_COG	.042	.136
- LIWC_FEELS	.192	.399
Beto_base_cased	.001	.008

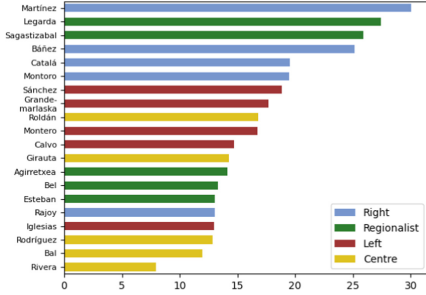
\*\* Statistically significant values are in bold.

**Table 6.** Spearman  $r$  correlation between the  $F_1$  values in the AV experiments and the indices values. \*\*

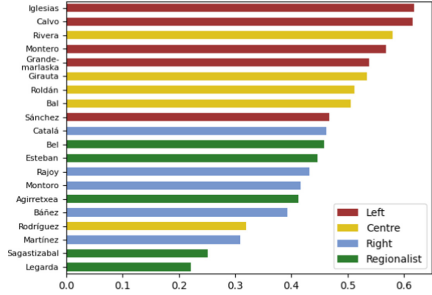
	ATI		CNI		ASI	
	$r$	p-value	$r$	p-value	$r$	p-value
BaseFeatures	.335	.148	.432	.057	-.444	.050
+ POS	.296	.205	<b>.460</b>	<b>.041</b>	<b>-.683</b>	<b>.001</b>
+ STRESS	.284	.225	.411	.072	<b>-.508</b>	<b>.022</b>
+ LIWC_GRAM	.403	.078	<b>.558</b>	<b>.011</b>	-.429	.059
+ LIWC_COG	<b>.489</b>	<b>.029</b>	<b>.621</b>	<b>.003</b>	<b>-.489</b>	<b>.029</b>
+ LIWC_FEELS	.438	.054	<b>.543</b>	<b>.013</b>	-.400	.081
ALL	.335	.148	<b>.483</b>	<b>.031</b>	<b>-.459</b>	<b>.042</b>
- BaseFeatures	.251	.286	.389	.090	<b>-.642</b>	<b>.002</b>
- POS	.360	.119	<b>.471</b>	<b>.036</b>	-.402	.079
- STRESS	.314	.177	<b>.463</b>	<b>.040</b>	<b>-.487</b>	<b>.029</b>
- LIWC_GRAM	.271	.248	.442	.051	-.435	.056
- LIWC_COG	.319	.171	<b>.453</b>	<b>.045</b>	<b>-.543</b>	<b>.013</b>
- LIWC_FEELS	.211	.373	.331	.154	<b>-.493</b>	<b>.027</b>
Beto_base_cased	<b>.544</b>	<b>.013</b>	<b>.675</b>	<b>.001</b>	<b>-.517</b>	<b>.020</b>

common in parliamentary speeches, hence preventing meaningful differences to emerge. Moreover, while 8 feature sets show a positive correlation with CNI, as many feature sets show a negative correlation with ASI, where BASEFEATURES + POS holds the strongest correlation. For comparison, we show the plot of both CNI and ASI correlated with the results for the ALL feature set in Fig. 7 and Fig. 8 respectively. The opposite nature of the two indices is understandable, since a more adversarial style would naturally be less abstract and more focused on events and narration. Indeed, some studies, both regarding Question Time in English [10, 16] and face-to-face Spanish political debates [3], noted that adversarial speeches in the parliamentary context present certain repeating oratory patterns. This could explain why the present features, and in particular the POS set, perform worse on speakers with higher ASI, who are likely to use common syntactic patterns. Conversely, it is easier to recognize speakers with a more abstract communication style.

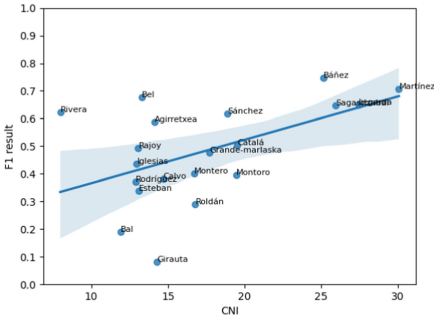
**Fig. 3.** AV results for the BASEFEATURES + LIWC\_COG feature set, divided by political wing and party.**Fig. 4.** Analytical Thinking Index values for each author.



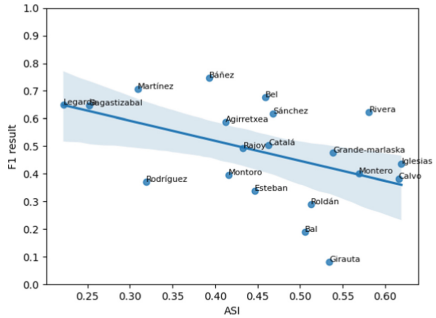
**Fig. 5.** Categorical-versus-Narrative Index values for each author.



**Fig. 6.** Adversarial Style Index values for each author.



**Fig. 7.** Correlation among AV results for the ALL feature set and CNI values.



**Fig. 8.** Correlation among AV results for the ALL feature set and ASI values.

## 6 Conclusion and Future Work

In this research, we investigate the extent to which topic-agnostic features, and in particular rhythmic and psycholinguistic feature sets obtained via a text-masking approach, are useful for AId and AP tasks in the Spanish language, using a dataset of political speeches. We show that such features perform comparably to a BETO transformer fine-tuned with the non-masked texts (hence potentially learning from topic-related information) in all such tasks. Moreover, we conduct a series of statistical analysis, showing that the different results for the various authors in the AV task are at least partially linked with the political affiliation of the author and their communication style, with a positive correlation with abstract and categorical style, and a negative correlation with adversarial style.

In future work, we aim to extend this study to other forms of political communication (e.g., tweets [12]), and to further explore the relation between the author’s profile and the classification performance with a more comprehensive statistical analysis.

**Acknowledgment.** The research work by Silvia Corbara was carried out during her visit at the Universitat Politècnica de València and was supported by the AI4MEDIA project, funded by the EU Commission (Grant 951911, H2020 Programme ICT-48-2020).

The research work by Paolo Rosso was partially funded by the Generalitat Valenciana under DeepPattern (PROMETEO/2019/121).

## References

1. Bevendorff, J., et al.: Overview of PAN 2021: authorship verification, profiling hate speech spreaders on Twitter, and style change detection. In: Candan, K., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 419–431. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-85251-1\\_26](https://doi.org/10.1007/978-3-030-85251-1_26)
2. Bischoff, S., et al.: The importance of suppressing domain style in authorship analysis. [arXiv:2005.14714](https://arxiv.org/abs/2005.14714) (2020)
3. Blas-Arroyo, J.L.: ‘Perdóneme que se lo diga, pero vuelve usted a faltar a la verdad, señor Gonzalez’: form and function of politic verbal behaviour in face-to-face Spanish political debates. *Discour. Soc.* **14**(4), 395–423 (2003)
4. Boyd, R.L.: Mental profile mapping: a psychological single-candidate authorship attribution method. *PLoS One* **13**(7), e0200588 (2018)
5. Bull, P., Wells, P.: Adversarial discourse in Prime Minister’s questions. *J. Lang. Soc. Psychol.* **31**(1), 30–48 (2012)
6. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained BERT model and evaluation data. In: PML4DC at ICLR 2020 (2020)
7. Chulvi, B., Rosso, P., Molpeceres, M.A., Sánchez-Junquera, J., Rodrigo, M.: Us and them: immigrant’s stereotypes and language style on political parliamentary speeches (under revision) (2022)
8. Corbara, S., Chulvi, B., Rosso, P., Moreo, A.: Investigating topic-agnostic features for authorship tasks in Spanish political speeches. In: Rosso, P., Basile, V., Martínez, R., Mètais, E., Meziane, F. (eds.) NLDB 2022. LNCS, vol. 13286, pp. 394–402. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-08473-7\\_36](https://doi.org/10.1007/978-3-031-08473-7_36)
9. Corbara, S., Moreo, A., Sebastiani, F.: Syllabic quantity patterns as rhythmic features for Latin authorship attribution. [arXiv:2110.14203](https://arxiv.org/abs/2110.14203) (2021)
10. Fenton-Smith, B.: Discourse structure and political performance in adversarial parliamentary wuestioning. *J. Lang. Polit.* **7**(1), 97–118 (2008)
11. Fernández-Cabana, M., Rúas-Araújo, J., Alves-Pérez, M.T.: Psicología, lenguaje y comunicación: análisis con la herramienta LIWC de los discursos y tweets de los candidatos a las elecciones gallegas. *Anuario Psicol.* **44**(2), 169–184 (2014)
12. García-Díaz, J.A., Colomo-Palacios, R., Valencia-García, R.: Psychographic traits identification based on political ideology: an author analysis study on Spanish politicians’ tweets posted in 2020. *Futur. Gener. Comput. Syst.* **130**, 59–74 (2022)
13. Gaston, J., et al.: Authorship attribution vs. adversarial authorship from a LIWC and sentiment analysis perspective. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 920–927. IEEE (2018)
14. van der Goot, R., Ljubešić, N., Matroos, I., Nissim, M., Plank, B.: Bleaching text: abstract features for cross-lingual gender prediction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Volume 2: Short Papers, pp. 383–389 (2018)

15. Halvani, O., Graner, L., Regev, R.: TAVeer: an interpretable topic-agnostic authorship verification method. In: Proceedings of the 15th International Conference on Availability, Reliability and Security (ARES 2020), pp. 1–10 (2020)
16. Harris, S.: Being politically impolite: extending politeness theory to adversarial political discourse. *Discour. Soc.* **12**(4), 451–472 (2001)
17. Jordan, K.N., Sterling, J., Pennebaker, J.W., Boyd, R.L.: Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proc. Natl. Acad. Sci.* **116**(9), 3476–3481 (2019)
18. Kestemont, M., et al.: Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF (Working Notes). CEUR Workshop Proceedings, vol. 2125. CEUR-WS.org (2018)
19. Neidorf, L., Krieger, M.S., Yakubek, M., Chaudhuri, P., Dexter, J.P.: Large-scale quantitative profiling of the old English verse tradition. *Nat. Hum. Behav.* **3**(6), 560–567 (2019)
20. Nisbett, R.E., Peng, K., Choi, I., Norenzayan, A.: Culture and systems of thought: holistic versus analytic cognition. *Psychol. Rev.* **108**(2), 291 (2001)
21. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015. Technical report (2015)
22. Pennebaker, J.W., Chung, C.K., Frazee, J., Lavergne, G.M., Beaver, D.I.: When small words foretell academic success: the case of college admissions essays. *PLoS One* **9**(12), e115844 (2014)
23. Plecháč, P.: Relative contributions of Shakespeare and Fletcher in Henry VIII: an analysis based on most frequent words and most frequent rhythmic patterns. *Digit. Scholarsh. Humanit.* **36**(2), 430–438 (2021)
24. Stamatatos, E.: A survey of modern authorship attribution methods. *J. Am. Soc. Inform. Sci. Technol.* **60**(3), 538–556 (2009)
25. Stamatatos, E.: Masking topic-related information to enhance authorship attribution. *J. Am. Soc. Inf. Sci.* **69**(3), 461–473 (2018)
26. Tukey, J.W.: Comparing individual means in the analysis of variance. *Biometrics*, pp. 99–114 (1949)
27. Weerasinghe, J., Singh, R., Greenstadt, R.: Feature vector difference based authorship verification for open world settings. In: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum. CEUR-WS.org (2021)