Investigating Topic-agnostic Features for Authorship Tasks in Spanish Political Speeches

Silvia Corbara[1], Berta Chulvi Ferriols[2,3], Paolo Rosso[2], and Alejandro Moreo[4]

[1] Scuola Normale Superiore (Pisa, IT) silvia.corbara@sns.it
[2] Universitat Politècnica de València (Valencia, ES)
berta.chulvi@upv.es prosso@dsic.upv.es
[3] Universitat de València (Valencia, ES)
[4] Istituto di Scienza e Tecnologie dell'Informazione, CNR (Pisa, IT)
alejandro.moreo@isti.cnr.it

Abstract. Authorship Identification is the branch of authorship analysis concerned with uncovering the author of a written document. Methods devised for Authorship Identification typically employ *stylometry* (the analysis of unconscious traits that authors exhibit while writing), and are expected not to make inferences grounded on the *topics* the authors usually write about (as reflected in their past production). In this paper, we present a series of experiments evaluating the use of feature sets based on rhythmic and psycholinguistic patterns for Authorship Verification and Attribution in Spanish political language, via different approaches of text distortion used 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 the BETO transformer when the latter is trained on the original text, i.e., when potentially learning from topical information.

Keywords: Authorship Identification · Text distortion · 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 in cases 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 $\{A_1, \ldots, A_m\}$, the goal is to identify the real author of d among the set of candidates; instead, AV can be defined as a binary problem, in which the goal is to infer whether A (the only candidate) is the real author of d or not. While tackling these classification problems, researchers devise methods able to distinguish among the different styles of the authors of interest, often relying on supervised machine learning.

In this article, we evaluate the employment of rhythmic- and psycholinguistic-based features for AV and AA in Spanish. Concretely, we propose to generate new distorted 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 – see Section 3.2). The resulting representations are topic-agnostic strings from which we extract n-grams features. We combine the resulting features with other feature sets that are by now consolidated in the AId field. In order to assess the different effect of our proposed feature sets on the performance, we carry out experiments of *ablation* (in which we remove one feature set from the whole at a time) and experiments of *addition* (in which we test the contribution of one single feature set at a time). Our results seem to indicate that our topic-agnostic features bring to bear enough authorial information as to perform comparably with BETO, the Spanish equivalent to the popular BERT transformer, trained on the original (hence topic-aware) text. The code of the project can be found at: https://github.com/silvia-cor/Topic-agnostic_ParlaMintES.

2 Related Work

The annual PAN shared tasks [1] offer a very good overview of the most recent trends in AId. In the survey by Stamatatos [9], the features that are most commonly used in AId studies are discussed; however, it is also noted 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 [6], or might actively mask topical content via a text distortion approach [10]. As already mentioned in Section 1, in this project we experiment with features capturing the rhythmic and the psycholinguistic traits of the texts, employing a text distortion technique based on syllabic stress or LIWC categories.

The idea of employing rhythmic, or prosodic, features in the authorship field is not a new one. Their most natural use is in studies focused on poetry; nevertheless, they have also been employed in authorship analysis of prose texts. In particular, some researches have studied the application of accent, or stress, for AId problems in English [8]. In the work by Corbara et al. [4], the documents are encoded in sequences of long and short syllables, from which the relevant features are extracted and used for AA in Latin prose texts, with promising results. We aim to investigate the applicability of this idea to Spanish, a language derived from Latin: we thus exploit the concept of *stress*, which gained relevance over the concept of *syllabic quantity* in Romance languages.

Linguistic Inquiry and Word Count (LIWC) [7] is a famous software application for text analysis: its core is composed of a word dictionary where each entry is associated with one or more categories that are related to grammar, emotions, or various cognitive processes and psychological concepts. Nowadays it is a popular tool for the study of the psychological aspect of textual documents, usually by employing the relative frequency of each LIWC category. In the AId field, it has been used for the characterization of a "psychological profile" or a "mental profile mapping" for AA and AV studies [3]. It has also been profitably used for the analysis of speeches regarding the Spanish political debate [5].

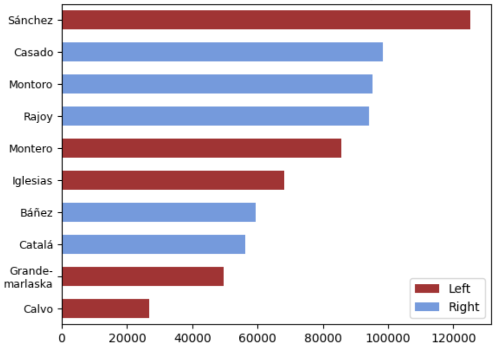
3 Experimental Setting

3.1 Dataset: ParlaMint

For our experiments, we employ the Spanish repository of the *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1* [5] by the digital infrastructure CLARIN, 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 preprocessing steps.

In order to have a balanced dataset, we select the parties with more than 500 speeches and assign them to the Left or Right wing: PSOE, PSC-PSOE and UP to the former, and PP, PP-Foro and Vox to the latter. We then select for each wing the 5 authors with most speeches in the dataset. We see that the author in this subset with the lowest number of samples (Calvo Poyato) has 142 samples in total; while taking all her samples, we randomly select 142 samples for each other author. We finally end up with 10 authors and 1, 420 samples in total. We show the total number of words for each speaker in Figure 1.

Fig. 1: Total number of words for each speaker



3.2 Feature Extraction: BaseFeatures and Text Encodings

Our focus in this research is to evaluate the employment of rhythm- and psycholinguistic-based features for AId tasks. To this aim, we explore various combinations including other topic-agnostic feature sets commonly used in literature.

As a starting point, we employ a feature set comprised of features routinely used in the AId field, including the relative frequencies of: function words (using the list provided by the NLTK library[6]), word lengths, and sentence lengths. We

[5] https://www.clarin.si/repository/xmlui/handle/11356/1431.
[6] https://www.nltk.org/

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 distortion 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;[7] 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.[8] 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.[9] For each macro-category, we perform a separate text distortion by replacing each word with the corresponding LIWC category tag. 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 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 matches are possible, the one with the longest common prefix is returned.) We show an example of the encodings we are using in Table 1. From a single encoding, we extract the word n-grams in the range $[1, 3]$ and compute the TfIdf weights, which we use as features. We call this feature sets LIWC_GRAM, LIWC_COG, and LIWC_FEELS, respectively.

3.3 Experimental Protocol

We perform experiments in two settings: Authorship Verification (AV) for each author (where each test sample is labelled as belonging to that class/author, or not) and Authorship Attribution (AA) (where each sample is labelled as

[7] https://github.com/linhd-postdata/rantanplan

[8] We employ the Spanish version of the dictionary, which is based on LIWC2007.

[9] We use following categories for each macro-categoy: (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, Tentat, Certeza, Inhib, Incl, Excl, Percept, Ver, Oir, Sentir, NoFluen, Relleno, Ingerir, Relativ, Movim; (iii) Maldec, Afect, EmoPos, EmoNeg, Ansiedad, Enfado, Triste, Asentir, Placer. We avoid employing categories that would repeat information already captured by the POS tags, or topic-related categories such as Dinero or Familia.

Table 1: Example of the encodings employed in the project. (Note there is not a one-to-one correspondence between syllables and stresses since the Rantanplan library caters for linguistic phenomenons across word boundaries, such as synalepha.)

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	PresentverbosEL	w	
LIWC_COG:	w		w	w	MecCog	w	w	MecCog	w	w	w	
LIWC_FEELS:	AfectEmoPos		w	w	w	w	w	w	w	w	w	
STRESS:	$+-+---+--+-+-$											
English translation:	Thank you. There is no need to return what has not existed.											

belonging to one of the 10 classes/authors). We assess the usefulness of the different feature sets by evaluating the performance of a classifier trained using them. In particular, we use 90% of the whole dataset to train the classifier, and evaluate its performance on the remaining 10% test set (the split is done randomly in a stratified way). As evaluation measure, for the AV task we use the well-known F_1 function, and for the AA task we use the *macro-averaged* F_1 (hereafter: F_1^M) and *micro-averaged* F_1 (hereafter: F_1^μ) variants.

We employ a Support Vector Machine (SVM) as learner[10], using the implementation of the SVC module from the `scikit-learn` package.[11] We perform the optimisation of various hyper-parameters: the parameter C, which sets the trade-off between the training error and the margin (we explore the range of values $[0.001, 0.01, 0.1, 1, 10, 100, 1000]$), the kernel function (we explore the following possibilities: *linear*, *poly*, *rbf*, *sigmoid*), and whether the classes weights should be balanced or not. The optimization is computed in a grid-search fashion, via 5-fold cross-validation on the training set. The selected model is then retrained on the whole training set and used for predictions on the test set.

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 ,[12] 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 for each author in Table 2. In the first batch of results, we show the performance of the features 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", subtracting each feature set to the combination of all the feature sets we are exploring (named ALL). These

[10] We also performed preliminary experiments with other learners: SVM showed a remarkably better performance than Random Forest, while no significant differences were noticed between SVM and Logistic Regression.

[11] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[12] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased. This model obtained better results than the 'uncased' version in preliminary experiments.

results are obtained using a SVM learner. Finally, we report the results obtained using the BETO transformer. Even though BETO obtains better results in 6 out of 10 cases, the fact that our proposed model obtains a comparable performance, without the aid of topic-related information, is highly promising. It is also interesting that we observe markedly different results depending on the author considered, both regarding the highest F_1 value and, more importantly, which feature combinations achieve the highest performance. In fact, some feature sets seem to work very well for certain authors, while being detrimental for others (e.g., the LIWC_FEELS set, while being counterproductive in the case of Rajoy and Montero, greatly helps the evaluation in the case of Sánchez and Montoro). We hypothesize the demographic or political group each single author belongs to might be responsible for some of the differences in the results we have observed; we leave these considerations for future work. Nevertheless, the combination of many feature sets seems to usually lead to better performance.

We show the results of the AA experiments in Table 3. We proceed in the same way as for the report of the AV experiments (Table 2). In these experiments, the ALL features combination employing the SVM learner obtained the best results, even outperforming BETO. Moreover, every feature set causes a drop in the performance if taken out from the ALL combination. However, in the experiments by addition, the individual feature set appears to have little impact, especially in the case of STRESS and LIWC_FEELS.

We perform a non-parametric McNemar's paired test of statistical significance between the results obtained using our best SVM configuration and the results obtained using BETO, for each of the authorship tasks. The test is carried out by converting the predictions of the two methods into binary values, where 1 stands for a correct prediction and 0 stands for a wrong prediction. The test indicates the differences in performance are not statistically significant at a confidence level of 95% in most cases (the only exception being the AV experiment for Calvo). This brings further evidence that the (topic-agnostic) features we propose in this work yield comparable results to a transformer trained on the original (topic-aware) text.

5 Government vs Opposition

In a final experiment, we test if the AV classification performance behaves differently depending on whether the speaker's speeches come from a period when their political party was part of the government, or instead was part as the opposition. To do this, we employ the speeches by the current Spanish Prime Minister, Pedro Sánchez Pérez-Castejón, who in the present dataset has 70 speeches dating back when he was in the opposition and 72 speeches since he has been in the government, hence making a rather balanced comparison. We thus perform the same AV experiment for the author as in Table 2, but only considering his speeches while he was either in the government or in the opposition as positive samples. The results are reported in Table 4.

Understandably, given the smaller number of positive samples, the general performance declines, except for the feature set + POS and for the BETO clas-

Table 2: Results for AV (divided in left-wing and right-wing speakers).
The best result for the SVM methods is in bold, while the best result in general is in italic; the same format applies for the other tables as well.

Method	**Sánchez**	**Iglesias**	**Montero**	**GMarlaska**	**Calvo**
BaseFeatures	0.444	0.606	0.526	0.556	0.478
+ POS	0.571	**0.667**	0.667	**0.783**	0.429
+ STRESS	0.261	0.538	0.571	0.581	**0.488**
+ LIWC_GRAM	0.133	0.453	0.452	0.474	0.450
+ LIWC_COG	0.250	0.296	0.500	0.533	0.439
+ LIWC_FEELS	0.467	0.500	0.444	0.462	0.311
ALL	***0.692***	0.571	0.625	0.636	0.444
- BaseFeatures	0.636	0.538	0.417	0.545	0.429
- POS	0.667	0.457	0.500	0.643	0.474
- STRESS	0.636	0.606	0.667	0.600	0.485
- LIWC_GRAM	0.640	0.529	0.667	0.600	0.387
- LIWC_COG	0.560	0.500	0.625	0.571	0.452
- LIWC_FEELS	0.316	0.645	**0.690**	0.645	0.483
BETO_base_cased	0.286	*0.741*	*0.741*	*0.800*	*0.667*

Method	**Rajoy**	**Catalá**	**Báñez**	**Casado**	**Montoro**
BaseFeatures	0.545	0.571	0.846	0.529	0.357
+ POS	0.640	0.706	0.889	0.643	0.514
+ STRESS	0.533	0.483	0.774	0.500	0.387
+ LIWC_GRAM	0.636	0.296	0.786	0.455	0.480
+ LIWC_COG	0.526	0.273	0.786	0.421	0.381
+ LIWC_FEELS	0.519	0.500	0.720	0.414	0.516
ALL	0.714	0.647	0.923	**0.720**	0.595
- BaseFeatures	0.714	0.621	0.846	**0.720**	0.667
- POS	0.581	0.629	0.720	0.667	0.552
- STRESS	0.667	0.688	***0.963***	0.692	***0.647***
- LIWC_GRAM	0.583	**0.727**	0.880	0.667	0.556
- LIWC_COG	0.741	0.667	0.833	0.600	0.579
- LIWC_FEELS	***0.828***	0.647	0.923	0.667	0.564
BETO_base_cased	0.800	*0.889*	0.839	*0.889*	0.615

Table 3: Results for AA

Method	F_1^M	F_1^μ
BaseFeatures	0.584	0.585
+ POS	0.653	0.655
+ STRESS	0.521	0.528
+ LIWC_GRAM	0.558	0.563
+ LIWC_COG	0.610	0.620
+ LIWC_FEELS	0.500	0.500
ALL	***0.718***	***0.718***
- BaseFeatures	0.648	0.648
- POS	0.625	0.634
- STRESS	0.676	0.676
- LIWC_GRAM	0.668	0.669
- LIWC_COG	0.665	0.662
- LIWC_FEELS	0.685	0.683
BETO_base_cased	0.683	0.697

Table 4: Results for Government vs Opposition

Method	**Government**	**Opposition**
BaseFeatures	0.308	0.286
+ POS	0.250	**0.615**
+ STRESS	0.308	0.316
+ LIWC_GRAM	0.381	0.200
+ LIWC_COG	0.296	0.333
+ LIWC_FEELS	0.188	0.296
ALL	0.250	0.400
- BaseFeatures	0.000	0.222
- POS	0.250	0.154
- STRESS	0.250	0.182
- LIWC_GRAM	0.250	0.333
- LIWC_COG	***0.444***	0.400
- LIWC_FEELS	0.250	0.250
BETO_base_cased	0.222	*0.727*

sifier, both when considering only the opposition speeches. More generally, it seems to be slightly easier to classify the author when they are in the opposition, probably because the role allows and demands a more personal and sharp language. Nevertheless, the generally small differences might denote a communication that remains largely stable regardless of the political position. In future work, we plan to better understand the possible relations between the differences in rhetorical style and the variance in performance we have observed in the + POS and BETO methods.

6 Conclusion and Future Work

In this research, we investigate the extent to which rhythmic and psycholinguistic features sets, obtained via a text distortion approach, are useful for AId in Spanish language, tackling both AV and AA tasks using a dataset of political speeches. We show that such features perform comparably to a BETO transformer fine-tuned with the non-distorted texts (hence potentially learning from topic-related information). Moreover, we see that the combinations of different topic-agnostic feature sets are in general fruitful, although the effect of the single feature set changes considerably depending on the specific author.

In future work, we are interested in analysing the different performances obtained in our experiments, and in further studying a possible explanation for the variance in the results. Moreover, we are aware of the present limitations of the LIWC-based representation, since we currently do not attempt to disambiguate the polysemous words. Refining this aspect, while also developing an effective feature selection strategy, might improve the overall classification results.

7 Acknowledgment

The work by Silvia Corbara has been carried out during her visit at the Universitat Politècnica de València and has been supported by the AI4Media project, funded by the European Commission (Grant 951911) under the 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., Chulvi, B., Peña Sarracén, G.L.D.L., Kestemont, M., Manjavacas, E., Markov, I., Mayerl, M., Potthast, M., Rangel, F., Rosso, P., et al.: Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 419–431. Springer (2021)
2. Bischoff, S., Deckers, N., Schliebs, M., Thies, B., Hagen, M., Stamatatos, E., Stein, B., Potthast, M.: The importance of suppressing domain style in authorship analysis. arXiv:2005.14714 (2020)
3. Boyd, R.L.: Mental profile mapping: A psychological single-candidate authorship attribution method. PloS one **13**(7) (2018)
4. Corbara, S., Moreo, A., Sebastiani, F.: Syllabic quantity patterns as rhythmic features for latin authorship attribution. arXiv:2110.14203 (2021)
5. 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 de Psicología **44**(2), 169–184 (2014)
6. 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)
7. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015. Tech. rep. (2015)
8. Plecháč, P.: Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns. Digital Scholarship in the Humanities **36**(2), 430–438 (2021)
9. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology **60**(3), 538–556 (2009)
10. Stamatatos, E.: Masking topic-related information to enhance authorship attribution. Journal of the Association for Information Science and Technology **69**(3), 461–473 (2018)