Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 75-86 recibido 29-03-2023 revisado 03-05-2023 aceptado 08-05-2023

An approach to lexicon filtering for author profiling

Un enfoque del filtrado de léxico para perfiles de autor

César Espin-Riofrio,[1] **Jenny Ortiz-Zambrano,**[1] **Arturo Montejo-Ráez**[2]
[1]Universidad de Guayaquil, Guayaquil, Ecuador
[2]Universidad de Jaén, Jaén, España
{cesar.espinr, jenny.ortizz}@ug.edu.ec, amontejo@ujaen.es

Abstract: This paper studies the influence of a general Spanish lexicon and a domain-specific lexicon on a text classification problem. Specifically, we address the impact of the choice of lexicons for user modelling. To do so, we identify gender and profession as demographic traits, and political ideology as a psychographic trait from a set of tweets. We experimented with machine learning and supervised learning methods to create a prediction model with which we evaluated our specific lexicon. Our results show that the choice and/or construction of lexicons to support the resolution of this task can follow a given strategy, characterised by the domain of the lexicon and the type of words it contains.
Keywords: Lexicon filtering, user modelling, feature extraction.

Resumen: Este trabajo estudia la influencia de un léxico general del español y un léxico específico del dominio en un problema de clasificación de textos. En concreto, abordamos el impacto de la elección de léxicos para el modelado de usuarios. Para ello, identificamos el género y la profesión como rasgos demográficos, y la ideología política como rasgo psicográfico a partir de un conjunto de tuits. Experimentamos con métodos de aprendizaje automático y aprendizaje supervisado para crear un modelo de predicción con el que evaluamos nuestro léxico específico. Nuestros resultados muestran que la elección y/o construcción de léxicos para apoyar la resolución de esta tarea puede seguir una estrategia determinada, caracterizada por el dominio del léxico y el tipo de palabras que contiene.
Palabras clave: Filtrado de lexicones, modelado de usuario, extracción de características.

1 Introduction

Words are well-described units that provide the link between perception and meaning, so they have been central to developments in computational modelling of language during decades (McClelland and Rumelhart, 1981). The stock of words that speakers can draw on in a language is the lexicon (Clark, 1995). Research uses specific lexicons that have been created to address tasks of a general nature (Sandoval et al., 2022) or about specific domains such as medical (Campillos-Llanos et al., 2021), including the COVID-19 pandemic (Lanza et al., 2021), but also tourism (Moreira, 2021), and about emotional aspects (Roy and Sharma, 2021), among others.

Since the word is the basic unit of a language, improving its representation has a significant impact on various tasks in Natural Language Processing (NLP) (Zhang et al., 2017). By measuring some textual characteristics we can distinguish between texts written by different authors and identify their authorship. Juola and others (2008) define Authorship Attribution as the science of inferring author characteristics from documents written by that author. This task can be seen as a text classification task and it is related to the profiling of the user from the text, i.e. it consists in the identification of user characteristics according to the style used by the user in his or her writings.

Some areas of NLP where the use of lexicons is most relevant are user profiling and authorship attribution (Eder, Rybicki, and Kestemont, 2016). Lexicons have been routinely used for automatic sentiment extraction as a text classification task. The emergence of end-to-end solutions for text classification, such as deep neural networks, has not made lexicons any less useful.

ISSN 1135-5948 DOI 10.26342/2023-71-6

Many of the issues discussed by politicians and the media are so nuanced that even the choice of certain words implies the choice of an ideological position. Erikson and Tedin (2015) define political ideology as the set of beliefs about the proper order of society and how it can be achieved. In terms of groups of individuals

The aim of this paper is to analyse the influence of a specific lexicon versus a general one for text classification, in particular for user profiling, and how to filter and format them. To this end, we address the task of identifying gender and profession as demographic traits, and political ideology as a psychographic trait, from a user's set of tweets, the as a binary (for gender) or multi-class (for the rest) classification problem. We obtain a specific lexicon by filtering according to the influence of different types of word in the prediction of the tags. Thus, we re-experiment with the new filtered lexicon and demonstrate improvements in prediction metrics. We use phraseological and word frequency features as features and supervised learning classification methods for training. The proposed method can be generalised to other tasks related to user profiling and stylometric characterisation. The paper is organized as follows: in Section 2 we present the objectives and questions posed in our research; related research work is studied in Section 3. The dataset used and the lexicons for profile characterisation are described in Section 4. Our approach is detailed in Section 5, and the design of experiments is described in Section 6. In Section 7, we present the results. Discussion of the findings is given in Section 8. Finally, the conclusions of our research are reported in Section 9 along with guidelines for further research.

2 Objectives and contribution

The aim of this work is to determine the impact of lexicons in extracting style features for supervised learning systems on a text classification task. We focus on obtaining clear insights on the following issues:

1. **Generic versus specialised lexicons**. Usually, what is commonly available (even more in stylometry analysis) are general lexicons of frequent words in the target language without considering the specific topic to be addressed. This paper includes the prediction of political ideology as a relevant feature in user profiling, so it is interesting to explore the use of politically oriented lexicons and how this may affect the determination of features within the political domain.
2. **Lexicon size**. To understand the influence of the number of words considered in a lexicon on the performance of the final system, we have found some clues to help us determine the most appropriate size for the task and domain under consideration.
3. **Normalisation of words** We study the importance of the use of words in their inflected form (conjugated, number, gender, etc.) versus their respective normalized form, such as the lemma. The Spanish language is very rich and has profuse morphosyntactic derivations. A canonical representation of words may condense the semantic representation but to the detriment of the grammatical role and thus its possible stylometric value. This paper also analyzes the relevance of the canonical use of the terms.
4. **Influence of grammatical category**. Given that lexicons incorporate terms from a wide range of grammatical categories(*part -of-speech*), we are interested in exploring whether certain categories (verbs, adjectives, adverbs, etc.) have more impact as a source of stylometric features.

Thus, our aim of the study is to analyse the influence of the use of a lexicon in a specific attribution domain, in our case a lexicon of political words, in order to determine gender, profession and political ideology (binary and multiclass) on the basis of short texts. Likewise, together with a general lexicon of frequently used words in the Spanish language, to analyse the impact for training and prediction according to the number of words they contain.

We contribute with a lexical analysis for the classification of Spanish texts, and highlight the importance of a specific lexicon related to the prediction topic at hand. Our work provides clues for future research when the use of lexicons for the extraction of style features is considered.

3 Related work

From the 1960s until the late 1990s, research on authorship attribution was dominated by attempts to define traits to quantify writing style, a line of research known as "stylometry" (Holmes, 1994; Savoy, 2020). Tweedie, Singh, and Holmes (1996) define style as a set of measurable patterns that may be unique to an author. Stylometric analysis assumes that style is quantifiable in order to evaluate its distinctive qualities (Neal et al., 2017). Stylometry offers powerful techniques for examining variation in authors' style (Hoover, 2007). Moreover, stylometry can be useful not only for identifying an author, but also for associating styles with certain distinctive features of the author, such as gender or profession (Ikae, Nath, and Savoy, 2019). It can also be used to detect a way of conveying certain messages on social networks, such as political orientation (García-Díaz et al., 2022).

Regarding the use of lexicons, automatic sentiment determination is one of the text classification tasks in the area of PLN in which lexicons have been commonly employed. EuroWordNet (Vossen, 1998), is a multilingual database with word networks of several European languages including Spanish.

Taboada et al. (2011) present a lexicon-based approach to extract sentiment from text by incorporating the semantic orientation of individual words and contextual valence shifters, they describe the Semantic Orientation Calculator (SO-CAL) which they developed. They extract sentiment words (including adjectives, verbs, nouns, and adverbs), use them to compute semantic orientation, and demonstrate that this lexicon-based method works well and is robust across domains and texts.

For automatic analysis of emotions expressed in tweets, specialized lexical resources are necessary. Sidorov et al. (2013) present Spanish Emotion Lexicon (SEL)[1] as a resource for the analysis of emotions in texts, a dictionary marked with probabilities of expressing one of the six basic emotions containing 2,036 words. Moreno-Ortiz and Hernández (2013) perform a lexicon-based sentiment analysis of short texts generated on the social network Twitter in Spanish, carrying out such a performance evaluation with the Sentitext tool[2], a Spanish sentiment analysis system.

Molina-González et al. (2013) present a Spanish resource for Opinion Mining composed of a list of opinion words; SOL (Spanish Opinion Lexicon), with the objective of developing a Spanish lexicon based on one of the most widely used English lexicons for polarity classification, the Bing Liu English Lexicon (Hu and Liu, 2004). They manually revised the lexicon to improve the final word list resulting in iSOL (improved SOL)[3]. Next, they describe eSOLHotel, a new corpus for Sentiment Analysis composed of hotel reviews written in Spanish (Molina-González et al., 2014). They use the corpus to conduct a set of experiments for unsupervised polarity detection using different lexicons.

Davidson et al. (2017) use a crowdsourced hate speech lexicon to collect tweets containing hate speech keywords, tagging a sample of these tweets into three categories: those containing hate speech, only offensive language, and those containing neither. They train a multi-class classifier to distinguish between these different categories. Plaza-del Arco et al. (2018) present the process carried out to generate a new lexicon of intensity of emotions for Spanish, generating a parallel list to the lexicon of intensity of affects for English of (Mohammad, 2017).

A deep learning-based framework for building a sentiment domain lexicon by employing word vector models and deep learning-based classifiers is proposed by (Li et al., 2021), this work proposes a method to create a sentiment-related lexicon in an automated way. Plaza-del Arco et al. (2021) present a Spanish-language corpus for researching offensive language OffendES[4], collects 47,128 Spanish-language comments from young influencers on the social platforms Twitter, Instagram and YouTube, manually tagged in predefined offensive categories. As part of a project on extracting sentiment from text, Taboada (2017) presents the SFU Spanish review corpus[5], a collection of 400 reviews on cars, hotels, washing machines, books, cell phones, music, computers and movies. On the political theme,

[1] https://www.cic.ipn.mx/sidorov/SEL

[2] http://tecnolengua.uma.es/sentitext

[3] http://sinai.ujaen.es/?p=1202

[4] https://github.com/pendrag/MeOffendEs

[5] https://www.sfu.ca/mtaboada/docs/research/SFU_Spanish_Review_Corpus.zip

García-Díaz, Colomo-Palacios, and Valencia-García (2022) publish PoliCorpus 2020[6], a dataset composed of tweets of Spanish politicians published in 2020 including members of government, senators, deputies, presidents of autonomous communities, mayors, councillors, advisors and former politicians, classifying political ideologies as a binary classification problem (left versus right) and as a multiclass problem (left, moderate left, moderate right and right).

Another are where general lexicons has been used is when determining lexical complexity; the more common (frequent) is the vocabulary used, the easier its understanding. General lexicons have been found useful even when integrated in end-to-end solutions like deep neuronal networks in lexical complexity detection. For example, Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez (2022) review the use of word embeddings and compare them to a broader list of lexical-level linguistic features. They use tuned Transformers-based models run on the pre-trained models BERT (Devlin et al., 2018), XLM-RoBERTa (Conneau et al., 2019) and RoBERTalarge-BNE (Gutiérrez Fandiño, Armengol Estapé, and others, 2022) on the different Spanish datasets with various regression algorithms.

4 Data

There is debate about the use of different lexical units in research (Laufer and Cobb, 2020). Word types have to be taken into account when compiling word lists or lexicons (Webb, 2021). It is also essential for lexical profiles of texts and corpora, which indicate learning objectives in a specific domain. Also, machine learning methods tend to draw on more data to create better prediction models.

We will use data collections from evaluation campaigns for the specific task of text classification, establishing a point of comparison with the results obtained by previous participants.

4.1 Bechmarking collection

We have selected the PoliticEs task - Spanish Author Profiling for Political Ideology task organised in the IberLEF 2022 evaluation campaign (García-Díaz et al., 2022). We use the dataset provided for this task, which was collected between 2020 and 2021 from Twitter accounts of politicians and political journalists in Spain using the UMU-CorpusClassifier (García-Díaz et al., 2020), from selected users whose political affiliation can be identified according to the party to which the politicians belong or the editorial line of the newspapers where the journalists write. Each author is anonymized and tagged with their gender (male, female), and their political spectrum on two axes: binary (left, right) and multiclass (left, left_moderate, right, right_moderate). It is composed of messages from about 400 different users with at least 120 tweets. There are two datasets, a training and a test dataset (80% and 20% respectively), which are independent to prevent machine learning approaches from identifying authors instead of features. The dataset is an extension of PoliCorpus 2020 (García-Díaz, Colomo-Palacios, and Valencia-García, 2022).

We participated in the IberLEF 2022 campaign in the PoliticEs task. As "SINAI" team, we presented an approach with features of frequently used Spanish words (*citation removed for blind review*) using a combination of a Transformer model (Vaswani et al., 2017) with traditional machine learning methods. Twenty teams submitted results (García-Díaz et al., 2022) and, although we ranked 16th, our results are not far off given the simplicity of our method, highlighting the importance of using a lexicon of frequently used Spanish words, Table 1. This paper covers a deeper analysis on the use of lexicons with several strategies to optimize its integration in text classification tasks.

Among the most outstanding works were LosCalis (Carrasco and Rosillo, 2022) with a system based on Transformers. They combine BETO (Canete et al., 2020) and MarIA (Gutiérrez Fandiño, Armengol Estapé, and others, 2022) to extract document-level features together with a Multilayer Perceptron (MLP) for tag decoding. The NLP-CIMAT team (Villa-Cueva et al., 2022) proposed PolitiBETO, a pre-trained BETO model in the political domain that predicts test data at the tweet level, merges these predictions through a majority vote to determine the tags of a given author based on his or her tweets. Thirdly, Alejando Mosquera (Mosquera, 2022) explores the use of a regularised

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

Team	Average Macro F1	gender	profession	ideology_binary	ideology_multiclass
LosCalis	0.90226 (01)	0.90287 (01)	0.94433 (01)	0.96162 (01)	0.80023 (04)
NLP-CIMAT	0.89096 (02)	0.78484 (06)	0.92125 (03)	0.96148 (02)	0.89628 (01)
Alejandro Mosquera	0.88918 (03)	0.82671 (03)	0.93345 (02)	0.95152 (03)	0.84504 (03)
SINAI	0.72147 (16)	0.78571 (05)	0.75395 (15)	0.78469 (15)	0.56154 (16)

Table 1: PoliticEs official ranking according to F1 metrics.

L2 Logistic Regression (LR) model based on n-grams of words and characters together with readability features.

4.2 Lexicons

We use the lexicon of frequently used Spanish words from the Corpus de Referencia del Español Actual (CREA)[7] referenced by the Real Academia Española (RAE). As its name suggests, it is a list of the most frequently used words in the Spanish language, from which we have incrementally taken the first 1,000 words for our experimentation (Table 2).

1	de	26	sus	51	mi	76	vida
2	la	27	le	52	porque	77	otro
3	que	28	ha	53	qué	78	después
4	el	29	me	54	sólo	79	te
5	en	30	si	55	han	80	otros
6	y	31	sin	56	yo	81	aunque
7	a	32	sobre	57	hay	82	esa
8	los	33	este	58	vez	83	eso
9	se	34	ya	59	puede	84	hace
10	del	35	entre	60	todos	85	otra
11	las	36	cuando	61	así	86	gobierno
12	un	37	todo	62	nos	87	tan
13	por	38	esta	63	ni	88	durante
14	con	39	ser	64	parte	89	siempre
15	no	40	son	65	tiene	90	día
16	una	41	dos	66	él	91	tanto
17	su	42	también	67	uno	92	ella
18	para	43	fue	68	donde	93	tres
19	es	44	había	69	bien	94	sí
20	al	45	era	70	tiempo	95	dijo
21	lo	46	muy	71	mismo	96	sido
22	como	47	años	72	ese	97	gran
23	más	48	hasta	73	ahora	98	país
24	o	49	desde	74	cada	99	según
25	pero	50	está	75	e	100	menos

Table 2: CREA Lexicon, showing the first 100 frequently used words in Spanish.

On political words, Sánchez-Junquera, Ponzetto, and Rosso (2020) present a corpus of tweets from politicians' accounts of the main political parties during the Spanish elections of 10 November 2019 (10N Spanish elections). The authors performed a semi-automated annotation process of themes and feelings/emotions, and provided a preliminary qualitative analysis of the dataset on different topics addressed in the election campaign. From this corpus, we extract the most frequently used words in these texts, thus creating a list of 300 words of political use which we call ELECC (Table 3). We then used this lexicon together with the CREA lexicon to determine word frequency characteristics.

The different versions of the Spanish political lexicon, original and filtered, are hosted at Spanish-election-lexicon[8].

1	vox	26	fuerza	51	educación	76	niños
2	españa	27	libertad	52	simpatizantes	77	reforma
3	psoe	28	mañana	53	madrid	78	laboral
4	gobierno	29	electoral	54	europa	79	casa
5	sánchez	30	debate	55	minuto	80	plan
6	gracias	31	quiero	56	millones	81	derogar
7	país	32	pedro	57	puedes	82	trabajadores
8	españoles	33	ciudadanos	58	seguirlo	83	problemas
9	campaña	34	torra	59	impuestos	84	hablando
10	barcelona	35	sociales	60	ungobiernocontigo	85	medios
11	10n	36	única	61	derecho	86	apoyo
12	acto	37	derechos	62	voto	87	ortega
13	cataluña	38	quieren	63	democracia	88	votantes
14	personas	39	noche	64	encuentro	89	murcia
15	directo	40	irene	65	valientes	90	real
16	política	41	único	66	alternativa	91	ilusión
17	entrevista	42	nacional	67	político	92	convivencia
18	partido	43	oviedo	68	elecciones	93	vía
19	años	44	mitin	69	señor	94	miles
20	seguir	45	domingo	70	hablar	95	votapsoe
21	gente	46	presidente	71	unidas	96	abascal
22	10nvotasolucions	47	futuro	72	catalunya	97	sede
23	aversivoyaserdeup	48	montero	73	español	98	frente
24	ley	49	propuestas	74	socialista	99	congreso
25	españasiempre	50	familias	75	mensaje	100	pensiones

Table 3: ELECC Lexicon, showing the top 100 words in political usage.

5 Approach

We start from the hypothesis of the usefulness of lexicons to determine the user profile and how decisive is the choice of a lexicon adapted to the specific domain related to the object under study. We experimented with the extraction of profile features such as gender and profession, and semantics for political ideology, in order to establish and differentiate the impact of the lexicons used.

We carried out our experimental approach in the following way: Starting from a general lexicon of frequently used words of the Spanish language, we train classifier models incrementally as a function of the number of words used. We add a specific lexicon of political words to establish another point of comparison in conjunction with the general lexicon.

[7] https://corpus.rae.es/lfrecuencias.html

[8] https://github.com/cespinr/Spanish-election-lexicon

Now, we analyse how determinant both lexicons are on the outcome by correlating the variables with the prediction. We continue by exploring the lexicons with their words in their inflected or lemma form and check again which lexicon is more influential. Having established which lexicon is more determinant, we go deeper with it in its grammatical word forms and filter out which forms have more weight on the studied features. Now, we can obtain a new lexicon even more specific and determinant according to the domain treated, in our case the political one. At each step we obtain prediction metrics for a final sequential comparison.

In more detail, our model extracts style features by analysing the CREA lexicons of frequently used words in the Spanish language and the ELECC lexicon of political words. Taking each word from the lexicons we determine its frequency of occurrence in the analysed texts by computing a frequency ratio per 1000. In addition, we compute other phraseological features such as MeanWordLen, LexicalDiversity, MeanSentenceLen, StdevSentenceLen, MeanParagraphLen and DocumentLen, based on the Stylometry library[9]. This set of features has no variations throught our experiment. Thus, we obtain the final feature vector used for training different classical machine learnnig algorithms: Logistic Regression (LR), Ramdom Forest (RF), Decision Tree (DT), Multilayer Perceptron (MLP) and Gradient Boosting (GB), together with an ensemble voting classification method. Thus, we have our experimental approach to filter a lexicon for a specific domain by training and evaluating different classifiers for the prediction of gender, profession, binary_ideology and multiclass_ideology, Figure 1.

With the most influential lexicon we generate a version with the inflected types or lemma of the words ("siguiendo" – "seguir"). We also estimate the grammatical categories by means of POS, i.e. whether they are verb, adjective, noun, etc. We relate again this versions with F1 score metric using Pearson correlation to determine more precisely the types of words that have the greatest influence on the performance of the system. We thus establish a new filtered ELECCNEW lexicon with the most important grammatical forms of words, Fig. 2.

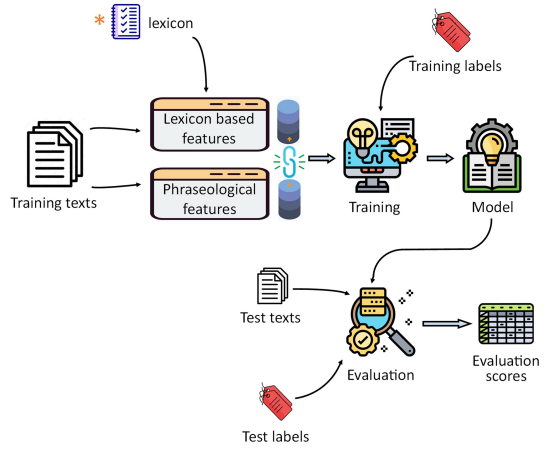

Figure 1: Experimental approach.

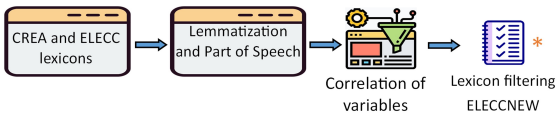

Figure 2: Lexicon filtering process.

Now we have a new point of analysis with this new ELECCNEW lexicon, we train again with the classifiers and method proposed and compare these results with the first ones obtained.

6 *Experiments design*

Our experimentation proceeds as follows:

- Exploration with the original lexicons; first only with the frequently used words for CREA Spanish and then in conjunction with the ELECC lexicon of political words in incremental combinations of 100 by 100, based on the analysis of word frequency and phraseological features. In this way, we observe the behaviour of the lexicons and their importance on prediction.
- Determination of the importance of the use of the words in their inflected form versus their respective normal form. Knowing which lexicon is the most decisive, we use Freeling[10] to obtain the lemma of the words and unify their different grammatical variants.
- Analysis of the grammatical categories of the lexicon from the parts of

[9] https://github.com/jpotts18/stylometry

[10] https://nlp.lsi.upc.edu/freeling/

speech (POS) with the Spacy[11] pipeline *es_core_news_sm*, counting the number of occurrences of each type.

- Filtering a new political lexicon by determining the correlation of words and their influence on prediction.
- New training with filtered lexicon to obtain resulting performances.

7 Results

We then have several points of assessment to present your results.

The results of evaluating the original ELECC lexicon together with CREA in incremental word combinations of 100 by 100, are shown in Table 4. Also, Figure 3 shows where the best results were obtained for each label according to the lexicon combinations.

CREA	ELECC	gender	profession	ideology binary	ideology multiclass
100	100	0.7122	0.8798	0.7673	0.5934
100	200	0.7268	0.8785	0.8220	0.6238
100	300	0.7463	0.8885	0.8429	0.6702
200	100	0.7207	**0.9149**	0.8005	0.5898
200	200	0.7015	**0.9149**	0.8429	0.6697
200	300	0.7207	0.8945	0.8533	0.6311
300	100	0.7037	0.8872	0.7673	0.5927
300	200	0.7292	0.9060	0.8113	0.6623
300	300	0.7324	0.8667	0.8325	0.6524
400	100	0.7486	0.8561	0.8113	0.5942
400	200	0.7679	0.8647	0.8533	0.6414
400	300	0.7890	0.8770	0.8728	0.6501
500	100	0.7817	0.8959	0.8128	0.6168
500	200	0.7572	0.9048	0.8429	0.6239
500	300	0.7268	0.8972	0.8325	0.6320
600	100	0.7785	0.8699	0.8220	0.6329
600	200	0.7744	0.8945	0.8337	0.6376
600	300	0.7378	0.8770	0.8128	0.6572
700	100	0.7918	0.8872	0.7806	0.5858
700	200	0.7634	0.8840	**0.8635**	0.6629
700	300	0.7548	0.8945	0.8220	0.6520
800	100	0.7324	0.8734	0.8220	0.6533
800	200	0.7548	0.8647	**0.8635**	0.6704
800	300	0.7486	0.8476	0.8533	0.6431
900	100	0.7853	0.8753	0.8220	0.6376
900	200	0.7208	0.8929	0.8325	0.6941
900	300	0.7679	0.8734	0.8533	0.6531
1000	100	0.7420	0.8581	0.8220	0.6276
1000	200	0.7160	0.8770	0.8429	0.6162
1000	300	**0.8349**	0.8667	0.8220	**0.6743**

Table 4: F1 weighted average for each label combining the CREA and ELECC lexicons.

We have, Table 5 , the influence of ELECC and CREA on label prediction by measuring the statistical relationship (Pearson correlation coefficient) between the number of CREA and ELECC words with the weighted mean F1 obtained for each label. The strong correlation (above 0.5) of ELECC on the ideology_binary and ideology_multiclass labels is clearly visible. CREA has a moderate influence on gender. From here, we continue the deeper analysis of ELECC.

	gender	profession	ideology binary	ideology multiclass
CREA	**0.4403**	-0.4216	0.2203	0.2808
ELECC	0.0845	-0.0363	**0.5887**	**0.5803**

Table 5: Pearson correlation of CREA and ELECC words on prediction of each tag.

Table 6 shows the prediction using the ELECC words in their inflected or lemma form, with 263 words making up this new list.

CREA	ELECC lemma	gender	profession	ideology binary	ideology multiclass
100	100	0.6132	0.8581	0.7584	0.5506
100	200	0.6971	0.8885	0.7915	0.5654
100	263	0.6987	0.8753	**0.8635**	0.5221
200	100	0.7043	0.8872	0.7559	0.5552
200	200	0.6696	0.8785	0.7559	0.5376
200	263	0.6839	0.8785	0.8325	0.5634
300	100	0.7099	0.8411	0.7356	0.6066
300	200	0.7207	0.8770	0.7294	0.5643
300	263	0.7263	0.8753	0.7444	0.6048
400	100	**0.7766**	0.8770	0.7785	0.5692
400	200	0.7572	0.8770	0.8325	0.5510
400	263	0.7350	0.8929	0.7896	0.6006
500	100	0.7634	0.8857	0.8429	0.6189
500	200	0.7486	0.8667	0.8005	0.5905
500	263	0.7451	0.8667	0.8005	0.6557
600	100	0.7400	0.8667	0.7915	0.5935
600	200	0.7207	0.8496	0.7716	0.6407
600	263	0.6952	0.8428	0.8220	0.6493
700	100	0.6690	**0.9137**	0.8005	0.6352
700	200	0.7268	0.8770	0.8022	0.6286
700	263	0.7524	0.8840	0.7785	0.6197
800	100	0.7422	0.8734	0.8220	0.6643
800	200	0.7037	0.8496	0.7896	0.6012
800	263	0.7698	0.8667	0.8533	0.5786
900	100	0.7268	0.8929	0.7896	**0.6687**
900	200	0.7634	0.8753	0.8429	0.6541
900	263	0.7324	0.8857	0.8418	0.6542
1000	100	0.7437	0.8840	0.8113	0.6387
1000	200	0.7314	0.8840	0.8522	0.6381
1000	263	0.7437	0.8647	0.8205	0.6574

Table 6: F1 weighted average F1 for each label using the inflected form or lemma of the words ELECC.

With the PoS grammatical categories of ELECC words, we continue filtering this lexicon according to the influence of each grammatical category, Table 7, we again use Pearson.

	gender	profession	ideology binary	ideology multiclass
ADV	-0.0178	0.0984	**0.6901**	**0.6589**
NOUN	0.0832	-0.0344	**0.5912**	**0.5825**
CCONJ	-0.0178	0.0984	**0.6901**	**0.6589**
PRON	0.1642	-0.1612	0.3296	0.3462
SPACE	0.0486	0.0134	**0.6462**	**0.6289**
AUX	0.1317	-0.1069	0.4656	0.4708
PROPN	0.1144	-0.0800	**0.5185**	**0.5184**
VERB	0.0825	-0.0333	**0.5927**	**0.5838**
ADP	0.1642	-0.1612	0.3296	0.3462
ADJ	0.0727	-0.0196	**0.6106**	**0.5991**

Table 7: Correlation of ELECC grammatical types on each label.

We are left with the grammatical types

[11] https://spacy.io

Figure 3: F1 weighted avg using the CREA and ELECC lexicons.

ADV, NOUN, CCONJ, PROPN and VERB, which show a strong correlation with the prediction labels, highlighted in Table 7. Thus, we have obtained a new filtered ELECCNEW lexicon of 250 words with which to proceed to the final training and prediction, see Table 8.

CREA	ELECC NEW	gender	profession	ideology binary	ideology multiclass
100	100	0.7524	0.8798	0.8022	0.5931
100	200	0.7074	0.8539	0.8005	0.5596
100	250	0.6922	0.8840	0.8325	0.6377
200	100	0.7400	**0.9248**	0.7532	0.5535
200	200	0.7539	0.8429	0.8312	0.5864
200	250	0.7057	0.8514	0.7806	0.6113
300	100	0.7400	0.8454	0.7326	0.6077
300	200	0.6971	0.8840	0.7356	0.5692
300	250	0.7364	0.8734	0.7471	0.6176
400	100	0.7572	0.8840	0.8113	0.6035
400	200	0.7122	0.8476	0.7806	0.5717
400	250	0.7100	0.8753	0.8429	0.6300
500	100	**0.7785**	0.8872	0.8205	0.5887
500	200	0.6885	0.8561	**0.8533**	0.6475
500	250	0.7420	0.8392	0.8113	0.6399
600	100	0.7378	0.8598	0.8022	0.6042
600	200	0.7400	0.8667	0.8337	0.6804
600	250	0.7099	0.8647	**0.8533**	0.6745
700	100	0.7400	0.9048	0.8113	0.6405
700	200	0.7314	0.8647	0.8233	0.6731
700	250	0.7122	0.8734	0.8113	0.6285
800	100	0.7437	0.8734	0.8325	0.6633
800	200	0.7229	0.8840	**0.8533**	0.6698
800	250	0.7698	0.8712	0.8429	0.6918
900	100	0.7378	0.8840	0.8220	0.6772
900	200	0.7634	0.8734	0.8429	0.6547
900	250	0.7766	0.8625	0.8429	0.6552
1000	100	0.7400	0.8753	**0.8533**	0.6129
1000	200	0.6885	0.8734	0.8325	0.6478
1000	250	0.7333	0.8539	0.8429	**0.7006**

Table 8: F1 weighted average for each label with filtered ELECCNEW lexicon.

Finally, in Figure 4, we have the individual behaviour of the LR, RF, DT and MLP classifiers used for training, now with the new ELECCNEW lexicon.

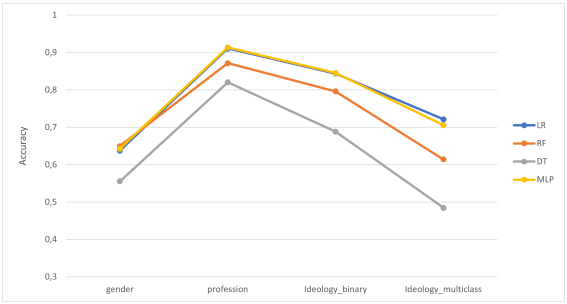

Figure 4: Performance of classifiers using CREA and ELECCNEW.

8 *Discussion*

We used CREA to explore the relative frequency of words in profile discovery (Espin-Riofrio, Ortiz-Zambrano, and Montejo-Ráez, 2022). The best prediction was obtained for profession with F1 of 0.9060, followed by ideology 0.8418, gender 0.7853 and ideology 0.6494. These values, compared to the official PoliticEs results, are very competitive and close to the state of the art.

The addition of the original ELECC lexicon, in combination with CREA, shows that the more ELECC words used, the better the prediction of political ideology, highlighting the importance of building specific lexicons when trying to identify thematically narrow characteristics of individuals. Improvements were also obtained in predicting gender and profession.

Original ELECC is more influential in de-

termining political ideology, while CREA is more influential in determining gender, as can be seen from the correlation between lexicons and label prediction.

When comparing the prediction, using the inflected form or lemma of the ELECC words, with the prediction using the original ELECC lexicon, the results are very similar, even gender has some decrease. Therefore, it does not seem that the use of a canonical form of the lexicons results in a relevant improvement in the performance of the systems.

ELECC has a strong positive association for ideology_binary and ideology_multiclass, highlighting the grammatical forms ADV, NOUN, CCONJ, PROPN, VERB and ADJ, we filter these and get the filtered lexicon ELECCNEW with which we get better performance on almost all tags. We show the sequential results of our experiments, which can be seen in Table 9.

A marked improvement in the prediction results using the ELECCNEW filtered lexicon is observed, thus outperforming the results presented by the SINAI team in the IberLEF 2022 PoliticES task (Table 10).

Regarding the classifiers used, LR and MLP presented better accuracy than RF and DT, all with better results in profession prediction.

9 Conclusions and future research

We have established that the choice of the type of lexicon to be used in a profile identification task is decisive. There are aspects of the profile that are more closely related to domain and semantics, and others to style. General lexicons like CREA (in the case of Spanish) are a resource that we can find useful in the identification of issues related to style or with inherent characteristics of the author such as gender and profession. But, when target information about content, such as political ideology, it is clear that the ELECCNEW lexicon adapted to the domain related to the object of that profile aspectec, politics, is more appropriate.

In addition, we have seen how certain grammatical categories are more useful for providing certain information (such as determinants with gender) and that the relevance of these categories also depends on the level of specialisation of the lexicon. This gives us clues as to what both CREA and ELECC can provide, either with words that carry the semantics of the content of the text being analysed, or words that have to do with the person's writing style.

We contribute to the community with the ELECCNEW filtered political lexicon and its original and lemma versions of the most influential words.

We will continue experimenting with new methods, such as word type, word family, phrasing, etc., for lexical filtering and their influence on prediction according to the lexical profile of the text under investigation. We will analyze the use of several Transformer-based pretrained classifiers to model the context and relationship between words in a text, and compare them with the results obtained here. To extend our results, we will also use other Spanish datasets such as those proposed in the PAN workshops. In addition, considering other languages will help us to check the consistency of our results across tasks and languages.

Acknowledgements

This work has been partially supported by projects Big Hug (P20_00956, PAIDI 2020) and WeLee (1380939, FEDER Andalucía 2014-2020) both funded by the Andalusian Regional Government, and projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government.

References

Campillos-Llanos, L., A. Valverde-Mateos, A. Capllonch-Carrión, and A. Moreno-Sandoval. 2021. A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making*, 21(1):1–19.

Canete, J., G. Chaperon, R. Fuentes, et al. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Carrasco, S. S. and R. C. Rosillo. 2022. Loscalis at politices 2022: Political author profiling using beto and maria. In *Proceedings of the Iberian Languages Eval-*

Lexicon used	gender	profession	ideology_binary	ideology_multiclass
only CREA	0.7853	0.9060	0.8418	0.6494
only ELECC	0.6487	0.8454	0.8429	0.6223
ELECC + CREA	0.8349	0.9149	0.8635	0.6743
ELECC lemmas + CREA	0.7766	0.9137	0.8635	0.6687
ELECCNEW + CREA	**0.7785**	**0.9248**	**0.8533**	**0.7006**

Table 9: Results of sequential experiments.

Team	gender	profession	ideology_binary	ideology_multiclass
LosCalis	0.90287 (01)	0.94433 (01)	0.96162 (01)	0.80023 (04)
NLP-CIMAT	0.78484 (06)	0.92125 (03)	0.96148 (02)	0.89628 (01)
Alejandro Mosquera	0.82671 (03)	0.93345 (02)	0.95152 (03)	0.84504 (03)
SINAI	0.78571 (05)	0.75395 (15)	0.78469 (15)	0.56154 (16)
with ELECCNEW lexicon	**0.7785**	**0.9248**	**0.8533**	**0.7006**

Table 10: Results with new ELECCNEW lexicon vs. those obtained in IberLEF 2022 PoliticES. F1 score.

uation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruna, Spain.

Clark, E. V. 1995. *The lexicon in acquisition*. Number 65. Cambridge University Press.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Davidson, T., D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eder, M., J. Rybicki, and M. Kestemont. 2016. Stylometry with r: a package for computational text analysis. *The R Journal*, 8(1).

Erikson, R. S. and K. L. Tedin. 2015. *American public opinion: Its origins, content and impact*. Routledge.

Espin-Riofrio, C., J. Ortiz-Zambrano, and A. Montejo-Ráez. 2022. Sinai at politices 2022: Exploring relative frequency of words in stylometrics for profile discovery. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruna, Spain.*

García-Díaz, J. A., Á. Almela, G. Alcaraz-Mármol, and R. Valencia-García. 2020. Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Lenguaje Natural*, 65:139–142.

García-Díaz, J. A., R. Colomo-Palacios, and R. Valencia-García. 2022. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130:59–74.

García-Díaz, J. A., S. M. Jiménez-Zafra, M.-T. M. Valdivia, F. García-Sánchez, L. A. Ureria-López, and R. Valencia-García. 2022. Overview of politices 2022: Spanish author profiling for political ideology. *Procesamiento del Lenguaje Natural*, 69.

Gutiérrez Fandiño, A., J. Armengol Estapé, et al. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Holmes, D. I. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

Hoover, D. L. 2007. Corpus stylistics, stylometry, and the styles of henry james. *Style*, 41(2):174–203.

Hu, M. and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Ikae, C., S. Nath, and J. Savoy. 2019. Unine at pan-clef 2019: Bots and gender task. In *CLEF (Working Notes)*.

Juola, P. et al. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334.

Lanza, C., A. Folino, E. Pasceri, and A. Perri. 2021. Lexicon of pandemics: A semantic analysis of the spanish flu and the covid-19 timeframe terminology. *Journal of Documentation*.

Laufer, B. and T. Cobb. 2020. How much knowledge of derived words is needed for reading? *Applied Linguistics*, 41(6):971–998.

McClelland, J. L. and D. E. Rumelhart. 1981. An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review*, 88(5):375.

Mohammad, S. M. 2017. Word affect intensities. *arXiv preprint arXiv:1704.08798*.

Molina-González, M. D., E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Urena-López. 2014. Cross-domain sentiment analysis using spanish opinionated words. In *Natural Language Processing and Information Systems: 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings 19*, pages 214–219. Springer.

Molina-González, M. D., E. Martínez-Cámara, M.-T. Martín-Valdivia, and J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Moreira, G. L. 2021. El léxico del turismo en los diccionarios de español (the tourism lexicon in spanish dictionaries). *Terminàlia*, pages 27–38.

Moreno-Ortiz, A. and C. P. Hernández. 2013. Lexicon-based sentiment analysis of twitter messages in spanish. *Procesamiento del lenguaje natural*, 50:93–100.

Mosquera, A. 2022. Alejandro mosquera at politices 2022: Towards robust spanish author profiling and lessons learned from adversarial attacks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruna, Spain. D. Moctezuma, and V. Muniz-Sánchez.*

Neal, T., K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36.

Ortiz-Zambrano, J., C. Espin-Riofrio, and A. Montejo-Ráez. 2022. Transformers for lexical complexity prediction in spanish language. *Procesamiento del Lenguaje Natural*, 69:177–188.

Plaza-del Arco, F. M., M. D. Molina-González, S. M. Jiménez-Zafra, and M. T. Martín-Valdivia. 2018. Lexicon adaptation for spanish emotion mining. *Procesamiento del Lenguaje Natural*, 61:117–124.

Plaza-del Arco, F. M., A. Montejo-Ráez, L. A. Urena-López, and M. Martín-Valdivia. 2021. Offendes: A new corpus in spanish for offensive language research. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1096–1108.

Roy, G. and S. Sharma. 2021. Analyzing one-day tour trends during covid-19 disruption–applying push and pull theory and text mining approach. *Tourism Recreation Research*, 46(2):288–303.

Sánchez-Junquera, J., S. P. Ponzetto, and P. Rosso. 2020. A twitter political corpus of the 2019 10n spanish election. In *International Conference on Text, Speech, and Dialogue*, pages 41–49. Springer.

Sandoval, L. G. M., A. P. Quimbaya, C. E. C. Gutiérrez, J. F. G. Pachón, and D. F. V. Ramírez. 2022. Comparación de métodos de análisis de sentimientos en comunidades de habla hispana. *Encuentro Internacional de Educación en Ingeniería*.

Savoy, J. 2020. *Machine learning methods for stylometry*. Springer.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Trevino, and J. Gordon.

2013. Empirical study of machine learning based approach for opinion mining in tweets. In *Advances in Artificial Intelligence: 11th Mexican International Conference on Artificial Intelligence, MICAI 2012, San Luis Potosí, Mexico, October 27–November 4, 2012. Revised Selected Papers, Part I 11*, pages 1–14. Springer.

Taboada, M. 2017. SFU Review Corpus — Maite Taboada.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Tweedie, F. J., S. Singh, and D. I. Holmes. 1996. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30(1):1–10.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Villa-Cueva, E., I. González-Franco, F. Sanchez-Vega, and A. P. López-Monroy. 2022. Nlp-cimat at politices 2022: Politibeto, a domain-adapted transformer for multi-class political author profiling. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruna, Spain.*

Vossen, P. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

Webb, S. 2021. The lemma dilemma: How should words be operationalized in research and pedagogy? *Studies in Second Language Acquisition*, 43(5):941–949.

Zhang, M., Y. Liu, H. Luan, and M. Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970.