TeamMX at PoliticEs 2022: Analysis of Feature Sets in Spanish Author Profiling for Political Ideology

José Luis Ochoa-Hernández[1,*,†], Yuridiana Alemán[2,†]

[1]Universidad de Sonora (University of Sonora, Blvd. Luis Encinas J, Calle Av. Rosales , Centro, 83000 Hermosillo, Son. México

[2]Tecnologico de Monterrey, Atlixcáyotl 5718, Reserva Territorial Atlixcáyotl, 72453 Puebla, México

Abstract

Natural Language Processing (NLP) is evolving more and more every day and it is becoming a very powerful tool, especially when it works in combination with Machine Learning algorithms, as it is making ventures into areas in which it was not well known, such as automatic programming systems based on the GPT-3 model, the market or sales prediction, even, the risk detection in banking systems on the basis of written exchanges between branch managers or directors of the same bank. The so-called short texts, comments/reviews made on social networks like Twitter, Facebook or Youtube, are becoming relevant in several domains. The corpus provided by the IberLEF 2022 Task - PoliticEs was used for extract political ideology information, it was focused on the identification of the gender, the profession, and the political spectrum from a binary (Left, Right) and multi-class perspective (Left, Right, Moderate-Left and Moderate-Right). Eight methods are proposed, six of them didn't have the expected results, but contributed to the two best ones. We implemented a customized stopwords study for our research in collaboration with experiments such as Best unique words per category, Set-based study, Transition point and others to extract the features, then Random Forest, SVM and Neural Network algorithms with default parameters and the Scikit learn tool were used to identify the categories. Obtaining a Macro F1 value of 0.7984 and the highest value achieved was 0.8270 in the category of Profession.

Keywords

Authorship analysis, Author profiling, Authorship attribution, Linguistic features, Natural language processing

1. Introduction

Natural Language Processing every day becomes a more active and powerful tool [1], as is the case of the GPT-3 model which, to put it in a simply way, writes text as if it were a human and also can program basic video games simply by typing some instructions [2]. Other relevant applications are the revaluation of the price of products or the improve product sales prediction [3], the competitive intelligence [4], election forecasting [5, 6], market prediction [7], sales predictor [8], risk detection in banking systems [9], in the medical field with the patient experienced and perceived outcomes in a hospital across all transitions of care [10] or in the field

IberLEF 2022, September 2022, A Coruña, Spain

*Corresponding author.

†These authors contributed equally.

joseluis.ochoa@unison.mx (J. L. Ochoa-Hernández); yuridiana.aleman@gmail.com (Y. Alemán)

0000-0001-5009-8913 (J. L. Ochoa-Hernández); 0000-0002-0682-003X (Y. Alemán)

of radiology, without excluding other areas, with tasks of text classification, speech recognition, machine translation, and automatic summarization [11], among others. However, an area that is becoming increasingly important is the so-called short texts, e.g. comments/reviews made on social networks Twitter, Facebook or Youtube, in several domains such as product, movie, hospital, tourist place, Political reviews [12] or sentiment analysis, which gives great value to companies that consider the reviews made by users/customers about their products [13]. The analysis of the words in these short messages has much greater relevance, as well as the taxonomy of the sentences plays a very important role [14, 15]. As we know, a comma, can change the meaning of an entire sentence, e.g. ["we don't want to know" vs "no, we want to know"]. Negations are another element that changes the meaning of the texts [16], so the identification of the key element called "entity" is the first thing that must be identified in order to obtain the characteristics that define the feeling of the texts [17].

There is another branch of NLP and sentiment analysis that is becoming increasingly popular: identification or classification of short texts, for example racism identification [18] it would be a binary classification, if it belongs to one domain or another, it would be considered as a multi-class classification, as there are many different domains [19]. There are similar studies that have linked personality traits and political ideology, such as the study by [20], that even if it is not in the field of computer science, are a good indication of the potential that can be obtained if the texts can be classified using computer tools or now known as Machine Learning [21].

In the online paper Political Ideology and Psychology [22] seeks to standardize behavior on the basis of what life should be like and what reality should be like, categorizes into dual elements, e.g. he calls it collectivist or individualist ideology when a group or individual considerations are taken into account respectively. When it comes to referring to capitalism and socialism, it is classified as either right-wing or left-wing ideology and a final category is a function of the type of government, i.e, distinguishing between democracy, populism or dictatorship. Nowadays, being able to predict our political ideology has considerable value to politicians and for ourselves, because the information that we consume day-to-day can influence our daily lives. [23].

Some of these tasks are analyzed in this research. The corpus provided by the IberLEF 2022 Task - PoliticEs competition [24] was used for this purpose, whose goal was to extract political ideology information from short texts, in this case Tweets. For this, an author profiling task was proposed, it was focused on the identification of the gender, the profession, and the political spectrum from a binary (Left, Right) and multi-class perspective (Left, Right, Moderate-Left and Moderate-Right), i.e. being able to identify the political orientation of a text or if you have more pretensions, to be able to identify who said it based on certain characteristics.

The paper is structured as follows, a Related work section where work associated with personality, political ideology and Machine Learning is mentioned, a Data section, where the data used in the competition are specified. The Methodology section specifies the steps followed to obtain our results and the features set used, in Results section, experiments are carried out using two types of tests and average F1 score, also, some experiment of post evaluation phase are mentioned. Finally, Conclusions and future work section shows a resume about experiments and the current work with the data sets.

2. Related work

2.1. Personality Research and Assessment

Personality is a functional pattern consistent with itself, generally consolidated and resistant to change, it determines how that person will respond to a given situation. However, it is able to respond differently to situations that may arise, since these are internalized psychic forms, which do not depend so much on external stimuli or situations. Personality is also shaped by lifestyle, beliefs and motivations and even current worldviews [25]. Therefore, it can be said that personality is a pattern of certain elements such as attitudes, thoughts and recurrent feelings that are generally stable throughout life and which allow a certain degree of predictability as to their *mode of being.*

Personality can therefore be identified and predicted in a computer context, is not simple, is highly complex, but taking into account certain elements or patterns that are embedded in the personality, it can be identifiable, in [26] their study describes a method for assessing personality using an open vocabulary analysis of language extracted from social networks. They compiled the written language from 66,732 Facebook users, applied a questionnaire based on the personality traits described in the Big Five personality traits, then they built a predictive model of personality based on their language.

To be more successful with the customer, is something else that can be achieved with the Personality Assessment, if it is integrated into a recommendation system based on personality traits, for example, offer an extreme tourism excursion to people with these interests or on the contrary, offer a relaxing SPA experience to those people who don't like the extreme activities, the interaction of the two systems can benefit the business sector incorporating the personality criteria into Computational Advertising [27].

2.2. Identification of the political ideology

As discussed above, personality is part of the person, and political ideology, is something that has been formed on the basis of our experiences and knowledge, even, if we don't believe it, the expression of this ideology has become much easier thanks to the interaction that takes place in social networks, interacting with each other by writing, posting, and sharing content, regardless of their location. Representing "one's self" and reflects the actual personality [28, 29].

Several studies have collaborated and made progress, each in their own study area, for example: in the article of [19] the authors described that in the work of [30], the authors developed a stance-detection system in order to predict whether a user will perform an action or not. Their results indicate that the combination of features regarding personality traits were the most relevant. Although these studies do not mention author profiling directly, these personality traits are obtained from contributions of users in social networks. The same applies to the work of [31], the authors deal with political micro-targeting (PMT) that entails compiling personal data on social networks to send them specific political messages. Specifically, the authors focus on determining extra version, which implies positive traits such as sociability or assertiveness.

2.3. Machine Learning

As is known, Machine Learning is concerned with the study of algorithms and statistical methods that are applied in computer systems, which use them to perform a task to the best of their ability, only following the algorithm itself in most cases and there may be parameter-based adjustments for some of them. Machine Learning algorithms are generic (which is why its area of application is infinite, an example [11] where the SVM, Random Forests and Neural Networks are used) and mainly build mathematical models that attempt to find patterns or infer a possible outcome based on the data provided, known as training data, once the algorithm has been trained, test data can be provided which will achieve the intended purpose. There are five types of Machine Learning algorithms: supervised, semi-supervised, active learning, reinforcement and unsupervised learning [32].

An example of works that use the algorithms for the detection of political affiliation can be seen in [33] where the authors examined some features like (gender, political affiliation or range from Swedish politicians from their speeches in the parliament during 2003 and 2010. They evaluate different feature sets at author-level and document-level with Support Vector Machines (SVM), achieving an accuracy of 81.2% for gender prediction, 89.4% for binary political affiliation, and 78.9% for age range classification.

Neural Networks are used in the work of [19] for identify the Psychographic traits based on political ideology, where they test four varieties of the networks: Shallow Neural Networks, Deep Neural Networks, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), mentions that can be transferred to detect the political ideology of citizens and their results indicate that the linguistic features are good indicators for identifying fine-grained political affiliation.

Finally, we present an example of Random Forest applied to the domain of the Ideological and Political Course, first they constructed the training data set using the classification concept which is based on the mathematical model of resource allocation and cost function, in turn compare two other algorithms, KNN and Decision tree, using two training data, one of 200 elements and other of 1000. Their results show that the Random Forest algorithm, has the highest accuracy and has more obvious advantages in small data sets [34].

3. Data

UMUCorpusClassifier data set was used for experiments [19]. It is a corpus in Spanish where each author was labeled using four topics: gender, profession and political spectrum binary and multiclass. Each line of corpus is a tweet, and each user has many tweets. Corpus is divided in two training sets and two test sets. Some metrics of the corpus are showed:

- Initial training: 5,000 tweets, average of 39.96 words per tweet.
- Final training: 37,560 tweets, average of 40.37 words per tweet.
- Practice test: 1,000 tweets, average of 42.69 words per tweet.
- Final test:12,600 tweets, average of 39.84 words per tweet.

The competition goal is to obtain each topic label per user; hence, a joined process was necessary for process all tweets per user. After this process, initial training has 100 users, final

training set has 313 users, practice test has only 20 users and the final test has 105 users. Table 1 shows the number of users and classes per topic.

Table 1
Data set used by topic and number of users

Topic	Class	Initial training	Final training	Practice test	Final test
Gender	Male	53	177	16	69
	Female	47	136	4	36
Profession	Politician	86	251	14	80
	Journalist	14	62	6	25
Politician binary	Right	45	135	15	48
	Left	55	178	5	57
Politician multiclass	Right	14	41	5	17
	Moderate right	31	94	10	31
	Left	19	76	1	21
	Moderate left	36	102	4	36

This process obtained more words per instance: around 200 words per user in training sets and 4,500 words per user in test sets, also, using large texts, the statistical and vocabulary analysis could get better results.

4. Methodology

The goal and the main contribution of this paper is to analyze feature sets, the classifier selection and parameters will be in a second step. For that reason, preliminary experiments were carried on using three algorithms implemented in Scikit learn tool [35]. Random forest and SVM with default parameters and neural network (MLPClassifier) using the follows parameters: hidden layers=(features+classes, 3), activation function=logistic and maximum iterations=500.

The methodology employs text filtering based on stopwords, which were customized for these experiments using an adaptation of the work of [36], stopwords selection was used for all the experiments presented in this paper, not a predefined bag of words used, separating common words like, *cuando, hemos, hay, haber, todo, nuestro, ademas, cierto, porque, todos* and those with statistically high values like: *de, el, que, en, por, a, y, los, es, un*, eliminating all those that do not contribute so much.

These words were obtained following the stopwords study, which includes the following steps:

1. Separating the tweets into categories
2. Processing the tweets employing the Freeling tool (for labelling), which follows the standard NLP procedure [37] and...
3. Applying the stopwords selection algorithm to separate them.

This algorithm follows a few steps, first, we made a statistical study to know the most repeated words, commonly these are the stopwords, then analyze the words and their grammatical category, some categories were identified and finally, all words with these categories were

extracted. An example of these grammatical categories detected are shown (not all of them) and its definition can be found in the EAGLES tag set:

```
ArrayList<String> reglasSinInteresDos = new ArrayList<String>();
reglasSinInteresDos.add("Z");//1er, 25, etc.
reglasSinInteresDos.add("Fz");//@, \#
reglasSinInteresDos.add("P0");//se, me
reglasSinInteresDos.add("PP");//me, lo, yo, nos
reglasSinInteresDos.add("PD");//eso, esos, esto, aquello
reglasSinInteresDos.add("DI");//un, unos, una, unas
reglasSinInteresDos.add("DP");//su, tu, mi, nuestro
reglasSinInteresDos.add("CC");//y
reglasSinInteresDos.add("CS");//que
reglasSinInteresDos.add("SP");//de, con, por, para
reglasSinInteresDos.add("VS");//son, eran, es
reglasSinInteresDos.add("VA");//ha, han, hayan, hemos, habeis, hay
```

An example of this words of the Binary Class is shown (stopword, frecuency):

el	22242	.	17673
de	12283	"	8451
,	7560	que	6914
a	6883	y	6426
él	5289	en	5106
ser	4087	uno	3552
haber	2469	...	

For features selection process, three types of attributes were used: words, lemmas and parts of speech. Also, the frequency per attribute and tf-idf metric were calculated. We worked with eight experiments to obtain the features sets used, some of them obtaining better results than others and some leading to others, they are explaining as follows:

1. Statistical Study

This experiment consisted of a simple statistical study of the most repeated words per category belonging to the main class, i.e. the Gender topic, which consisted of two classes Male and Female. The study had two variants, the full text and the text without stopwords. Likewise, the statistics were carried out in two ways: 1) using the words in their normal/original mode and 2) in their lemmas mode, reducing the statistics considerably. Results were high in Gender and Profession topic, but they have many features.

2. Part of Speech Tagging

A statistical experiment was carried out looking at the grammatical categories of these words, which we believed would help us to differentiate the class. Spacy tool [38] was used for obtaining categories. The results were high only for the Profession topic, but not high enough compared to other sets.

3. Personally Selected Features

Another experiment was realized including a words study that was considered relevant in a manual mode, thanks to the previous Simple Statistical Study, for example: it was decided to choose the number of mentions of text *@user*, words tagged with *Hashtags (#)*, mentions of *[politicalParty]*, mentions of *quantities or numbers*, names of *countries, cities, towns o communities*, *proper nouns*, *abbreviations*, *misspellings*, number of *"no"* or negations , mentions of *"gobierno"* or *institutions*, direct mentions of the word *"derecha"* or *"izquierda"*.

This list was joined with others certain elements per tweet, such as the *tweet length* , the *number of words per tweet*, the number of *words in capitals*, the number of *words in lowercase*, the number of *arrows*, the number of the words *"tb", "q", "d"*, however the results were lower in all topics.

4. Best Unique Words per Category

Another study that provided better results was to select the most mentioned words uniquely in each category (not the most repeated). The 65 best words of each category were chosen and with these words the Gender topic was detected, something similar was done with the rest of the categories, the results were good. Example for gender:

Male category: regióndemurcia, véase, soñadores, aplicación, acabó, responder, artista, borbón, significa, bucle, etc.

Female category balears, gitanas, elegido orgullosa, gitanos, pedrosánchez, gitano, pedrosanchez, feijóo, compañer, etc.

5. Set-based Study

The study that gave the best results consisted of the model based on Set Theory. [39, 40], using the Complement property, which says *"The complement of a set A asks for all the elements that aren't in the set but are in the universal set"*, this means that we keep the complement of each of the classes, thus discarding the total number of words that could be in other categories.

6. Correlation Analysis

A Pearson's correlation method was analyzed to obtain the best correlation indices between each attribute and the class. Weka tool [41] was used to obtain the attributes list and several ranges were used: the best 50, 100, 200, 500 and 800 correlation indices with lemmas and words, represented by frequency and tf-idf metric. Results were high for Gender and Profession topic, but moderate for Political topics.

7. Transition Point

It is the application of Zipf law [42] (the product between range and word frequency is constant). Languages have two characteristics: unification and diversification, the first is to use very common words (high frequency), and the second is to use specific terms. According to this analysis, the most important words in a text are not the high frequency, if a vocabulary is divided in high frequency words and low frequency words, the most important is the last of the first group. The experiments were executed with different ranges of words, using the 30, 50, 100, and 200 nearest words to transition point. This set got high results in all experiments.

8. Average Analysis

The higher values for word average were used as attributes. In these experiments, several set of features were obtained, using a minimal average value: 0.02, 0.1. 0.5 and 1. Values are small due to the large number of instances in corpus. The values of the documents were frequency and the tf-idf metrics. This set got high results in all experiments.

5. Results

Two kind of test sets were used for experiments: practice and final test (Table 1). Following this order, first, the analysis with practice test and all feature sets are presented. In the second subsection, just the feature sets with the best results in the practice test were used for the final test analysis.

5.1. Using practice test

Figure 1 shows the best Macro F1 score per feature set (feature sets are described in methodology section). Set 4 only was implemented for Gender topic, thus, the figure shows zero in the rest of topics.

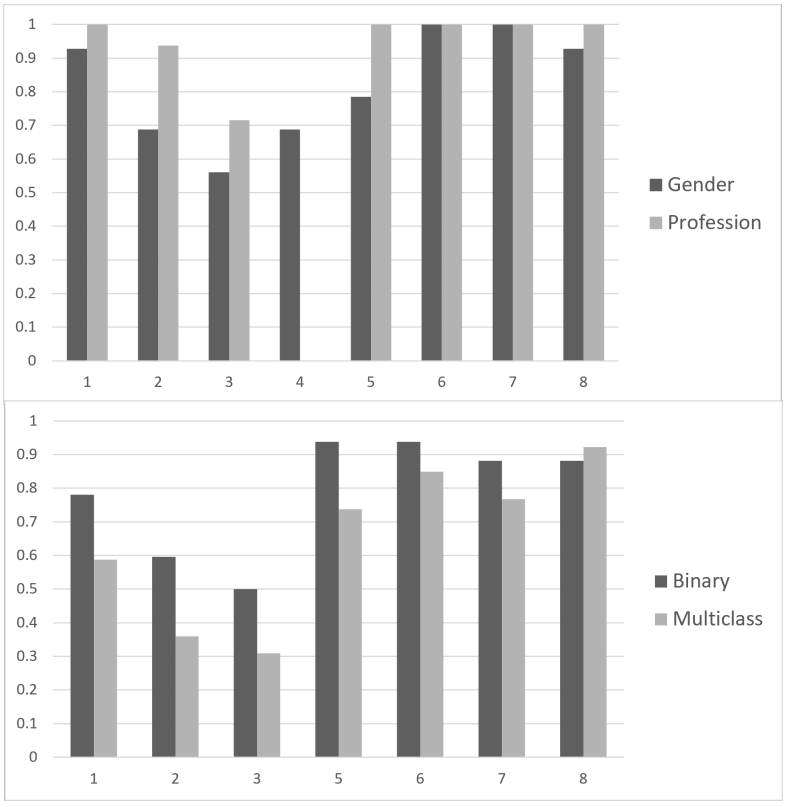

Figure 1: Best Macro F1 by topic and features set.

Set 1 gets high results in Gender and Profession topics, but all words were used, in Political topics (binary and multiclass) this set did not obtain high results as in the first topics. Sets 6, 7 and 8 obtained high results in the four topics (0.768 - 1.000), even set 6 and 7 obtained 1.0 in Gender and Profession topic. Set 3 get lower results in all topics, with a Macro F1=0.3 for Political multiclass topic, also, it is important emphasize that it is the best result for this set, including different classifiers and representations (frequency and tf-idf).

Table 2
Obtained results using test data

Topic	Num	Feature set	Representation	Classifier	Features	Macro F1
Gender	1	Transition Point	Tf-Idf (lemmas)	Random forest	4,755	1.0000
	2	Correlation Analysis	Tf-idf (words)	SVM	200	1.0000
	3	Statistical Study	Tf-idf (lemmas)	Random forest	6,300	0.9283
	4	Correlation Analysis	Tf-idf (lemmas)	SVM	200	0.9283
	5	Correlation Analysis	Tf-idf (lemmas)	Random forest	800	0.9283
	6	Correlation Analysis	Frequency (lemmas)	Random forest	500	0.9283
	7	Correlation Analysis	Tf-idf (lemmas)	Random forest	800	0.9283
	8	Transition Point	Frequency (words)	Random forest	273	0.9283
	9	Correlation Analysis	Tf-idf (words)	SVM	500	0.9283
	10	Correlation Analysis	Frequency (words)	Random forest	500	0.9283
Profession	1	Statistical Study	Tf-Idf (lemmas)	Random forest	6,300	1.0000
	2	Transition Point	Tf-idf (lemmas)	Random forest	680	1.0000
	3	Transition Point	Tf-idf (lemmas)	Random forest	723	1.0000
	4	Transition Point	Frequency (lemmas)	Random forest	723	1.0000
	5	Transition Point	Tf-idf (lemmas)	Random forest	400	1.0000
	6	Transition Point	Frequency (lemmas)	Random forest	400	1.0000
	7	Transition Point	Frequency (lemmas)	SVM	4755	1.0000
	8	Transition Point	Tf-idf (lemmas)	Random forest	606	1.0000
	9	Correlation Analysis	Tf-idf (lemmas)	Random forest	500	1.0000
	10	Correlation Analysis	Tf-idf (lemmas)	Random forest	800	1.0000
Ideology binary	1	Correlation Analysis	Frequency (lemmas)	Random Forest	100	0.9373
	2	Correlation Analysis	Frequency (lemmas)	Neural network	100	0.9373
	3	Set-based Study	Tf-idf (words)	SVM	3,515	0.9373
	4	Set-based Study	Tf-idf (words)	Random Forest	3,515	0.9373
	5	Set-based Study	Frequency (words)	Random Forest	3,515	0.9373
	6	Transition Point	Tf-idf (lemmas)	Random Forest	901	0.8810
	7	Correlation Analysis	Tf-idf (lemmas)	Random Forest	100	0.8810
	8	Correlation Analysis	Frequency (lemmas)	Random Forest	200	0.8810
	9	Correlation Analysis	Tf-idf (lemmas)	Random Forest	500	0.8810
	10	Transition Point	Frequency (words)	Random Forest	591	0.8810
Ideology multiclass	1	Average Analysis	Frequency (words)	Neural network	100	0.9222
	2	Average Analysis	Frequency (lemmas)	Neural network	103	0.8827
	3	Transition Point	Tf-idf (lemmas)	Random forest	723	0.8485
	4	Transition Point	Frequency (words)	Random forest	591	0.7972
	5	Set-based Study	Tf-idf (words)	Random forest	8,189	0.7679
	6	Transition Point	Tf-idf (words)	Random forest	320	0.7411
	7	Transition Point	Tf-idf (lemmas)	Random forest	494	0.7375
	8	Correlation Analysis	Tf-idf (words)	SVM	500	0.7375
	9	Transition Point	Tf-idf (lemmas)	Random forest	4,755	0.7134
	10	Transition Point	Frequency (lemmas)	Random forest	336	0.7126

Table 2 shows the best results by topic. This specifies general parameters (words, lemmas, features set), representation, used classifier, number of features and Macro F1 metric. Better results were obtained using random forest in three topics: Gender, Profession and ideology

binary. In ideology multiclass topic, neural networks obtain the best Macro F1. In Macro F1 column, there are not significant difference in results, only in ideology multiclass, where the best results are between the first and the last experiment.

Generally, better parameters are to use lemmas, but the features set are different in each topic: for gender topic correlation analysis get better results using tf-idf metric. The best result is with transition point, but almost all the vocabulary is necessary (4,755 features) to obtaining it. Instead, SVM algorithm just need the 200 words with the higher Pearson's correlation with the class. In Profession topic, lemmas and transition point get better results using only 400 features, vocabulary obtain high results too, but the number of features is more than 6,000. In ideology binary and multiclass, there are some results using words and lemmas with several feature sets (correlation, hand craft analysis and transition point) and simple frequency and tf-idf. Macro F1 is lower than first topic, but the number of features is lower too.

5.2. Final test

Final test was used for competition experiments. In this step, each topic was analyzed separately: The best experiments using initial test by topic was replicated for label the final test, in other words, for every experiment had been used a different classifier and different feature set, considering its behavior in previous experiments (results in Table 2). Table 3 shows the competition results. Ten proposals were sent, the best of them obtained 0.7984 in Macro F1, which positions us in the 8th better result (**TeamMX**).

Table 3
Competition results

General ranking	Team name	Macro F1	Gender F1	Profession F1	Ideology binary F1	Ideology multiclass F1
1	LosCalis	0.9022	0.903 (1)	0.944 (1)	0.967 (1)	0.800 (4)
2	NLP-CIMAT-GTO	0.8909	0.785 (6)	0.921 (3)	0.961 (2)	0.896 (1)
3	Alejandro Mosquera	0.8891	0.827 (3)	0.933 (2)	0.951 (3)	0.845 (3)
4	$CIMAT_2021$	0.8797	0.837 (2)	0.895 (5)	0.941 (4)	0.845 (2)
5	TeamHalBERT	0.8253	0.726 (13)	0.898 (4)	0.922 (5)	0.756 (6)
6	Bernardo Sigüenza	0.8199	0.792 (4)	0.850 (8)	0.913 (6)	0.725 (8)
7	I2C	0.7999	0.744 (11)	0.867 (7)	0.862 (9)	0.726 (7)
8	**TeamMX**	**0.7984**	**0.782 (7)**	**0.827 (11)**	**0.821 (11)**	**0.763 (5)**
9	andrei.manea	0.7869	0.738 (12)	0.883 (6)	0.902 (7)	0.624 (12)
	...					
20	BASELINE	0.5112	0.576 (19)	0.432 (18)	0.596 (19)	0.441 (19)

These metric were obtained with the follows classifiers and parameters:

- Gender: The best 200 Pearson's correlation words, using tf-idf metric and SVM classifier (experiment 2, Table 2).
- Profession: Transition point analysis with lemmas using tf-idf metric and random forest classifier (experiment 2, Table 2).
- Ideology binary: Set based study using tf-idf metric and SVM classifier (experiment 3, Table 2).
- Ideology multiclass: Average analysis with lemmas using frequency and neural networks (experiment 2, Table 2).

Results are lower in competition than initial test, but the best metrics were obtained with less features (except in ideology binary topic, where 3,515 features were necessary). Using the initial test set, results of 1.0 were obtained, but in the competition, the results were between 0.76 and 0.82. In both cases, the multiclass ideology obtained the lowest results, but the gender issue obtained lower results than expected.

Finally, for increasing Macro F1, the best results in evaluation phase were joined and sent for post evaluation phase, obtaining 0.8238. Table 4 shows the initial experiment for a few teams.

Table 4
Post evaluation phase results (data obtained on May 22th, 2022)

Team name	Macro F1	Gender F1	Profession F1	Ideology binary F1	Ideology multiclass F1
hiramcp	0.8429	0.746 (2)	0.833 (1)	0.961 (1)	0.831 (1)
TeamMX	**0.8238**	**0.792 (1)**	**0.827 (2)**	**0.913 (2)**	**0.763 (2)**
BASELINE	0.5112	0.576 (3)	0.432 (3)	0.596 (3)	0.441 (3)

This result was obtained with the follows classifiers and parameters:

- Gender: The best 500 Pearson's correlation words, using simple frequency and random forest classifier (experiment 10, Table 2).
- Profession: Transition point analysis with lemmas using tf-idf metric and random forest classifier (experiment 2, Table 2).
- Ideology binary: The best 100 Pearson's correlation lemmas, using simple frequency and neural networks classifier (experiment 2, Table 2).
- Ideology multiclass: Average analysis with lemmas using frequency and neural networks (experiment 2, Table 2).

6. Conclusions and future work

In this paper, a initial analysis for selecting features was applied in Spanish author profiling. Some conclusions obtained from these experiments are mentioned below:

- A unique feature set is not enough for classify four topics in the corpus, the best technique is to analyze each topic separately.
- Test sets are different, then, the feature sets and classifier have not the same behavior.
- Correlation analysis and transition point sets got better results in the final test.
- Set-based study does not appear in the first topics, but in binary ideology it obtains high values.

For future work, a depth analysis for the sets of Correlation analysis, Transition point and set-based Study with better results is necessary to increase Macro F1. Also, to experiment with different parameters for SVM and random forest classifier. Best unique word per category only was applied to Gender topic, but its behavior could be better using a large training a different tool to obtain lemmas.

References

[1] R. G. [Online], 2022, The Power of Natural Language Processing (harvard business review), URL: https://hbr.org/2022/04/the-power-of-natural-language-processing.

[2] M. Zhang, J. Li, A commentary of GPT-3 in MIT Technology Review 2021, Fundamental Research 1 (2021) 831–833. URL: https://www.sciencedirect.com/science/article/pii/S2667325821002193. doi:https://doi.org/10.1016/j.fmre.2021.11.011.

[3] N. Archak, A. Ghose, P. G. Ipeirotis, Show Me the Money! Deriving the Pricing Power of Product Features by Mining Consumer Reviews, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 56–65. URL: https://doi.org/10.1145/1281192.1281202. doi:10.1145/1281192.1281202.

[4] S. Collovini, P. N. Gonçalves, G. Cavalheiro, J. Santos, R. Vieira, Relation extraction for competitive intelligence, in: International Conference on Computational Processing of the Portuguese Language, Springer, 2020, pp. 249–258.

[5] E. Taylan, Numbers in politics: Comparative quantitative analysis & modeling in foreign policy orientation and election forecasting, Master's thesis, Department of International Relations, İhsan Doğramacı Bilkent University, 2019.

[6] J. W. Park, Election Prediction with Twitter Data via NLP and Machine Learning Algorithms: Tweets, User Description, and Sentiment Analysis, Ph.D. thesis, , 2022.

[7] K. Xiong, X. Ding, L. Du, T. Liu, B. Qin, Heterogeneous graph knowledge enhanced stock market prediction, AI Open 2 (2021) 168–174. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000243. doi:https://doi.org/10.1016/j.aiopen.2021.09.001.

[8] Y. Liu, X. Huang, A. An, X. Yu, ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 607–614. URL: https://doi.org/10.1145/1277741.1277845. doi:10.1145/1277741.1277845.

[9] C. Nopp, A. Hanbury, Detecting Risks in the Banking System by Sentiment Analysis, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 591–600. URL: https://aclanthology.org/D15-1071. doi:10.18653/v1/D15-1071.

[10] M. Khanbhai, L. Warren, J. Symons, K. Flott, S. Harrison-White, D. Manton, A. Darzi, E. Mayer, Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care, International Journal of Medical Informatics 157 (2022) 104642. URL: https://www.sciencedirect.com/science/article/pii/S1386505621002689. doi:https://doi.org/10.1016/j.ijmedinf.2021.104642.

[11] J. W. Luo, J. J. Chong, Review of Natural Language Processing in Radiology, Neuroimaging Clinics of North America 30 (2020) 447–458. URL: https://www.sciencedirect.com/science/article/pii/S1052514920300563. doi:https://doi.org/10.1016/j.nic.2020.08.001, machine Learning and Other Artificial Intelligence Applications.

[12] S. Ahmed, A. Danti, Effective Sentimental Analysis and Opinion Mining of Web Reviews Using Rule Based Classifiers, in: H. S. Behera, D. P. Mohapatra (Eds.), Computational

Intelligence in Data Mining—Volume 1, Springer India, New Delhi, 2016, pp. 171–179.
[13] B. Liu, Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies 5 (2012) 1–167. URL: https://doi.org/10.2200/S00416ED1V01Y201204HLT016. doi:10.2200/S00416ED1V01Y201204HLT016. arXiv:https://doi.org/10.2200/S00416ED1V01Y201204HLT016.
[14] A. Yadav, D. K. Vishwakarma, Sentiment analysis using deep learning architectures: a review, Artificial Intelligence Review 53 (2019) 4335 – 4385. doi:https://doi.org/10.1007/s10462-019-09794-5.
[15] Z. Wang, H. Wang, Understanding Short Texts, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Association for Computational Linguistics, Berlin, Germany, 2016. URL: https://aclanthology.org/P16-5007.
[16] S. M. Jiménez-Zafra, N. Cruz-Díaz, M. Taboada, M. Martín-Valdivia, Negation detection for sentiment analysis: A case study in Spanish, Natural Language Engineering 27 (2020) 1–24. doi:10.1017/S1351324920000376.
[17] K. Priyanka, S. Janakiraman, M. Deva Priya, Aspect level Sentimental Analysis of Opinion Mining – A Review, Materials Today: Proceedings (2021). URL: https://www.sciencedirect.com/science/article/pii/S2214785321012657. doi:https://doi.org/10.1016/j.matpr.2021.02.183.
[18] A. Field, S. L. Blodgett, Z. Waseem, Y. Tsvetkov, A Survey of Race, Racism, and Anti-Racism in NLP, CoRR abs/2106.11410 (2021). URL: https://arxiv.org/abs/2106.11410. arXiv:2106.11410.
[19] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on Spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74. URL: https://www.sciencedirect.com/science/article/pii/S0167739X21004921. doi:https://doi.org/10.1016/j.future.2021.12.011.
[20] M. Fatke, Personality Traits and Political Ideology: A First Global Assessment, Political Psychology 38 (2017) 881–899. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/pops.12347. doi:https://doi.org/10.1111/pops.12347. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/pops.12347.
[21] C. Stachl, F. Pargent, S. Hilbert, G. M. Harari, R. Schoedel, S. Vaid, S. D. Gosling, M. Bühner, Personality Research and Assessment in the Era of Machine Learning, European Journal of Personality 34 (2020) 613–631. URL: https://doi.org/10.1002/per.2257. doi:10.1002/per.2257. arXiv:https://doi.org/10.1002/per.2257.
[22] P. [Online], 2020, Ideología política y psicología: Relación entre ideología, sistemas políticos y comportamiento, URL: https://paradigma40pec.com/ideologia-politica-y-psicologia-relacion-entre-ideologia-sistemas-politicos-y-comportamiento/.
[23] B. Baumgaertner, J. E. Carlisle, F. Justwan, The influence of political ideology and trust on willingness to vaccinate, PLOS ONE 13 (2018) 1–13. URL: https://doi.org/10.1371/journal.pone.0191728. doi:10.1371/journal.pone.0191728.
[24] J. A. García-Díaz, S. M. Jiménez-Zafra, M. T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticEs 2022: Spanish Author Profiling for Political Ideology, Procesamiento del Lenguaje Natural 69 (2022).
[25] [Online], 2022, Personalidad, URL: https://concepto.de/personalidad/.

[26] G. Park, H. Schwartz, J. Eichstaedt, M. Kern, M. Kosinski, D. Stillwell, L. Ungar, M. Seligman, Automatic Personality Assessment Through Social Media Language, Journal of personality and social psychology 108 (2015) 934. doi:10.1037/pspp0000020.

[27] J. T. Yun, C. M. Segijn, S. Pearson, E. C. Malthouse, J. A. Konstan, V. Shankar, Challenges and Future Directions of Computational Advertising Measurement Systems, Journal of Advertising 49 (2020) 446–458. URL: https://doi.org/10.1080/00913367.2020.1795757. doi:10.1080/00913367.2020.1795757. arXiv:https://doi.org/10.1080/00913367.2020.1795757.

[28] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff, S. D. Gosling, Facebook Profiles Reflect Actual Personality, Not Self-Idealization, Psychological Science 21 (2010) 372–374. URL: https://doi.org/10.1177/0956797609360756. doi:10.1177/0956797609360756. arXiv:https://doi.org/10.1177/0956797609360756, pMID: 20424071.

[29] G. Seidman, Self-presentation and belonging on Facebook: How personality influences social media use and motivations, Personality and Individual Differences 54 (2013) 402–407.

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

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

[32] P. Langley, Human and machine learning, Machine Learning 1 (1986) 243–248.

[33] M. Dahllöf, Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability, Literary and linguistic computing 27 (2012) 139–153.

[34] X. Lei, Resource Sharing Algorithm of Ideological and Political Course Based on Random Forest, Mathematical Problems in Engineering 2022 (2022) 1–8. doi:10.1155/2022/8765166.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[36] J. L. Ochoa Hernández, Learning morphosyntactic patterns for multiword term extraction, Scientific Research and Essays 6 (2011). doi:10.5897/SRE11.1299.

[37] L. Padró, E. Stanilovsky, Freeling 3.0: Towards Wider Multilinguality, in: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), ELRA, Istanbul, Turkey, 2012.

[38] [Online], 2022, Industrial-Strength Natural Language Processing IN PYTHON, URL: https://v2.spacy.io/.

[39] P. K. Maji, R. Biswas, A. R. Roy, Soft set theory, Computers & Mathematics with Applications 45 (2003) 555–562.

[40] P. Zhu, Q. Wen, Operations on Soft Sets Revisited, Journal of Applied Mathematics 2013 (2013). doi:10.1155/2013/105752.

[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explor. Newsl. 11 (2009) 10–18. URL: https://doi.org/10.1145/1656274.1656278. doi:10.1145/1656274.1656278.
[42] C. J. V. Rijsbergen, Information Retrieval, 2nd ed., Butterworth-Heinemann, 1979.