- ORIGINAL ARTICLE -

Political Alignment Identification: a Study with Documents of Argentinian Journalists

Identificación de alineamiento político: un estudio con documentos de periodistas argentinos

Viviana Mercado[1], Andrea Villagra[1], and Marcelo Errecalde[1,2]

[1]*Laboratorio de Tecnologías Emergentes, Instituto de Tecnología Aplicada, Unidad Académica Caleta Olivia, Universidad Nacional de la Patagonia Austral, Santa Cruz. Argentina*
{vmercado, avillagra}@uaco.unpa.edu.ar

[2]*Laboratorio de Investigación y Desarrollo en Inteligencia Computacional, Universidad Nacional de San Luis, San Luis, Argentina*
merreca@unsl.edu.ar

Abstract

Political alignment identification is an author profiling task that aims at identifying political bias/orientation in people' writings. As usual in any automatic text analysis, a critical aspect here is having available adequate data sets so that the data mining and machine learning approaches can obtain reliable and informative results. This article makes a contribution in this regard by presenting a new corpus for the study of political alignment in documents of Argentinian journalists. The study also includes several kinds of analysis of documents of pro-government and opposition journalists such as the relevance of terms in each journalist class, sentiment analysis, topic modelling and the analysis of psycholinguistic indicators obtained from the *Linguistic Inquiry and Word Count* (LIWC) system. From the experimental results, interesting patterns could be observed such as the topics both types of journalists write about, how the sentiment polarities are distributed and how the writings of pro-government and opposition journalists differ in the distinct LIWC categories.

Keywords: Author Profiling, Exploratory Data Analysis (EDA), Journalist Political Alignment, LIWC, Text Mining

Resumen

La identificación de la alineación/orientación política es una tarea de determinación del perfil del autor que tiene como objetivo identificar el sesgo/orientación política en los escritos de las personas. Como es habitual en cualquier análisis de texto automático, un aspecto crítico aquí es tener disponibles conjuntos de datos adecuados para que los enfoques de minería de datos y aprendizaje automático puedan obtener resultados confiables e informativos. Este artículo hace una contribución a este respecto al presentar un nuevo corpus para el estudio de alineación política en documentos de periodistas argentinos. El estudio también incluye varios tipos de análisis de documentos de periodistas pro-gubernamentales y opositores, como la relevancia de los términos en cada clase de periodista, el análisis de sentimientos, el modelado de temas y el análisis de indicadores psicolingüísticos obtenidos del sistema LIWC. A partir de los resultados experimentales, se pudieron observar patrones interesantes como los temas sobre los que escriben ambos tipos de periodistas, cómo se distribuyen las polaridades de los sentimientos y cómo difieren los escritos de los periodistas pro-gubernamentales y opositores en las distintas categorías de LIWC.

Palabras claves: Análisis Exploratorios de Datos (AED), Determinación del Perfil del Autor, Minería de textos, LIWC, Orientación Política de Periodistas

1 Introduction

Political alignment identification (PAI) in a text or document is a form of *author profiling* (AP), one of the main tasks of *authorship analysis* together with authorship attribution/determination, plagiarism detection and style inconsistency detection [1]. PAI, the same as other AP tasks like the detection of depressed people, paedophiles, cyberbullyers and suicides is a challenging task within the automatic analysis of texts since it involves, in general, the use of representations of texts that capture stylistic and content aspects of their authors. In this context, a particular area within the PAI is that oriented to the study of political orientation in texts written by journalists, and which we will refer from now on as *journalistic texts*. We will consider as journalistic texts that information that a journalist publishes in various media such as a personal blog, an article written in a mass media such as a newspaper or the content expressed in a book of his authorship.

The PAI has been applied to texts generated by regular users of social media such as Twitter [2, 3] although more recently it has been done with the documents produced by journalists [4]. In [5], political speech in Twitter has been analyzed with LIWC during the 2008' German electoral campaign. The same tool, LIWC, was used to determine the psychological state and personality of the candidates for the presidency and vice presidency of the US in the 2004 campaign [6] and the language used by the New York's mayor, R. Giuliani [7] throughout his term. Regarding texts in Spanish language, in [8] the linguistic style of the candidates of the main political parties in the Spanish general elections of 2008 and 2011 is analyzed. On the other hand, the Spanish dictionary of LIWC was applied to analyze the political speech and tweets of the candidates in the elections of Galicia in 2012 [9]. Besides, in [10] is analyzed the institutional speeches of the President of Ecuador, Rafael Correa, from 2007 to 2015, using LIWC to determine the existence of differences in his language style in the periods analyzed.

The previous approaches are related to our work but, as far as we know, there are no PAI studies of journalistic texts in Spanish.[1] In this work, we will make a first approach to the PAI in journalistic texts in Spanish, in particular, of texts generated by Argentinian journalists. The task, in this case, will be to group all the documents of "pro-government" journalists on one side and "opponents" on the other one. In that way, it will allow in the future to visualize it as a binary classification ("pro-government" versus "opponent") problem. Our objectives, in the long run, will be analyse what are the more appropriate document representations and learning algorithms for this task and how it is related to similar studies with journalistic texts written in other languages.

In the present article, we contribute to achieve these objectives by presenting a new corpus for the study of political alignment in documents of Argentinian journalists. The study also includes several kinds of analysis of documents of pro-government and opposition journalists such as the relevance of terms in both types of journalists, sentiment analysis [12], topic modelling and the analysis of psycholinguistic indicators obtained from the LIWC system. In that context, our work extends a preliminary report presented in [11] by going deeper in analyzing and comparing the statistics obtained from LIWC, giving a more detailed specification of the sentiment analysis process and presenting a new analysis of the relevance of terms in both types of journalists derived from the information obtained with the PySS3 system [13]. In any case, the analyses carried out in the present article are very introductory and are only intended as an exploratory analysis of the new corpus that can be useful for those interested in developing PAI studies with this collection.

The rest of the article is organized as follows: Section 2 describes the PAI corpus introduced in the present article with statistics and metrics of the whole data set and from each of the two classes involved; Section 3 gives some results obtained from a topic modelling process and a sentiment analysis; Section 4 goes further in analyzing the PAI corpus by taking into account psycholinguistic indicators obtained from the LIWC system. Section 5 finishes this article by giving the main conclusions obtained from our study and some future work.

2 Corpus Description

Our work was focussed on generating a collection of Argentinian journalistic documents obtained from news blogs, online newspapers, books, etc.[2] It consists of 196 documents belonging to 10 journalists: 5 of them that clearly support the actions of the Argentine government in the period 2012 to 2015 and 5 of them that explicitly express themselves against the government in that period. The data set was split into two groups of documents according to the political orientation of the journalists. Thus, 98 documents belonging to the 5 pro-government journalists were selected for the *gov* (pro-government) class and the 98 remaining documents of the opposition journalists were used to build the *oppo* (opposition) class. In that way, a balanced corpus with 2 classes was obtained.

To select the documents some guidelines were taken into account:

- Texts correspond to Spanish documents written by Argentinian journalists.
- Texts refer to different political aspects related to the Argentinian government in the period 2012-2015, such as government actions, politicians' declarations, corruption cases, treatment of laws, etc.
- All the documents contain "formal text", that is to say, they do not present common "informal" aspects of content from social media such as abbreviations, slang expressions, typos, hyperlinks, labels, figures, and emoticons.
- From each journalist, between 18 and 20 documents were taken from his/her personal blog, articles in online newspapers or digital books of his/her authorship.
- Each journalist was clearly identified as pro-government or opponent.
- The same proportion of male and female journalists were kept between both categories.

After collecting the documents, they were manually labeled as belonging to the two above-mentioned

[1] The only exception is the preliminary report [11] extended by the present article.

[2] See `https://github.com/vbmercado/-Corpus_Periodistas_Argentinos` to get access to the collection.

classes *gov*, and *oppo*. Table 1 shows information about how the documents were finally distributed in both classes and what were the sources (online newspaper, blog or digital books) from they were obtained.

Table 1: Distribution of documents in classes and source.

Class	Newspapers	Blogs	Books	Total
gov	50	46	2	98
oppo	60	36	2	98

2.1 Corpus statistics

Before proceeding with a more elaborated analysis of the corpus, some basic statistics were obtained to get some insights about the general characteristics of the documents. First of all, the *number of words* per document was analyzed for each document/article of both classes. Table 2 shows the minimum, maximum, mean and standard deviation values for the number of words in the documents of the *gov*, *oppo* classes and the whole corpus (*gov* + *oppo*).

Table 2: Number of words in documents: minimum, maximum, mean and standard deviation values per class.

Class	Min.	Max.	Mean	St. Dev.
gov	139	36619	1865.18	5031.56
oppo	236	3423	1243.01	733.71
gov + *oppo*	139	36619	1554.09	3608.91

As we can see, although the sizes of the shortest documents (Min) are similar for both classes (139 vs 236), they differ considerably in the longest ones (36619 vs 3423). That can be observed more clearly in Figures 1 and 3 that show the number of words and Figures 2 and 4 with the histograms for the documents in the *gov* and the *oppo* classes. There, for instance, Figure 1 shows that there are some documents in the class *gov* whose sizes exceeds 34000 words. However, as the Figure 2 confirms, most of the documents in the *gov* class do not exceed 5000 words with only a couple of documents (that correspond to books) whose sizes are between 34000 and 36000 words. Figure 4 shows that the *oppo* class has a "smoother" distribution in the number of words of its documents, with sizes that oscillate between 200 and 3500 words approximately.

Regarding the total number of words and the size of the vocabulary of the whole corpus and of each class, we will first introduce the following notation: let $D_{\mathscr{C}}$ be the set of documents belonging to our corpus $\mathscr{C}$; let $D_{\mathscr{G}}$ and $D_{\mathscr{O}}$ be the sets of documents belonging to the pro-government and opposition journalists respectively, $D_{\mathscr{G}} \subset D_{\mathscr{C}}$, $D_{\mathscr{O}} \subset D_{\mathscr{C}}$, $D_{\mathscr{C}} = D_{\mathscr{G}} \cup D_{\mathscr{O}}$, $D_{\mathscr{G}} \cap D_{\mathscr{O}} = \emptyset$. Let $\#_{\mathscr{C}}$ be the total number of words in

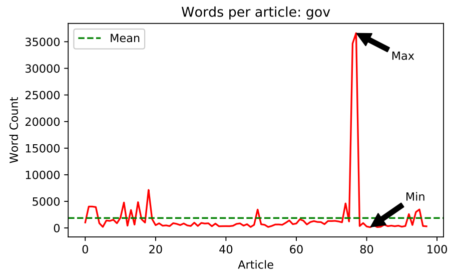

Figure 1: Class *gov*: Word count by article.

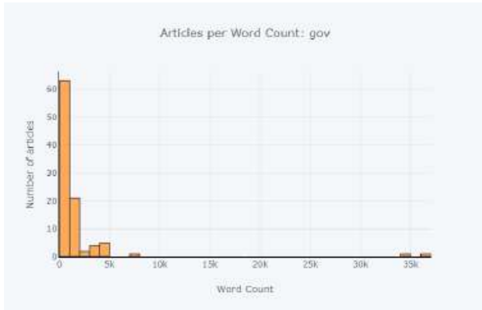

Figure 2: Class *gov*: Histogram by word count.

our corpus $\mathscr{C}$, with similar meanings for $\#_{\mathscr{G}}$ and $\#_{\mathscr{O}}$. Besides, let $\mathscr{V}_{\mathscr{C}}$, $\mathscr{V}_{\mathscr{G}}$, and $\mathscr{V}_{\mathscr{O}}$ be the *vocabularies*[3] of $D_{\mathscr{C}}$, $D_{\mathscr{G}}$, and $D_{\mathscr{O}}$, respectively. Table 3 gives some statistics related to these collection of documents.

Table 3: Statistics on the documents of the whole corpus ($D_{\mathscr{C}}$), documents of pro-government ($D_{\mathscr{G}}$) and opposition ($D_{\mathscr{O}}$) journalists.

$\#_{\mathscr{C}}$	280343
$\lvert\mathscr{V}_{\mathscr{C}}\rvert$	24323
$\#_{\mathscr{G}}$	167844
$\lvert\mathscr{V}_{\mathscr{G}}\rvert$	18497
$\lvert\mathscr{V}_{\mathscr{G}}\rvert/\#_{\mathscr{G}}$	0.11
$\#_{\mathscr{O}}$	112499
$\lvert\mathscr{V}_{\mathscr{O}}\rvert$	13146
$\lvert\mathscr{V}_{\mathscr{O}}\rvert/\#_{\mathscr{O}}$	0.11
$\lvert\mathscr{V}_{\mathscr{G}} \cap \mathscr{V}_{\mathscr{O}}\rvert$	7320
$\lvert\mathscr{V}_{\mathscr{G}} / \mathscr{V}_{\mathscr{O}}\rvert$	11177
$\lvert\mathscr{V}_{\mathscr{O}} / \mathscr{V}_{\mathscr{G}}\rvert$	5826

One of the first things we can observe from Table 3 is that although when the number of words in the whole corpus is high ($\#_{\mathscr{C}} = 280343$), the number of *distinct* words (the size of the *vocabulary*) is relatively small ($\lvert\mathscr{V}_{\mathscr{C}}\rvert = 24323$). That differs from the size of vocabularies in texts from social media which usually are bigger. A possible cause of this is that writings in social media are usually informal and prone to have abbreviations

[3] The vocabulary of a collection of documents is the set of *distinct* words that appear in that collection.

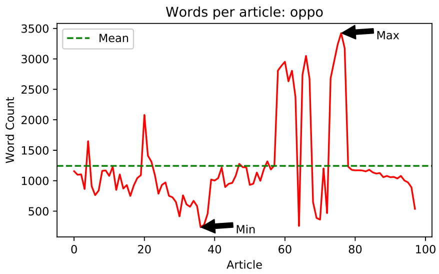

Figure 3: Class *oppo*: Word count by article.

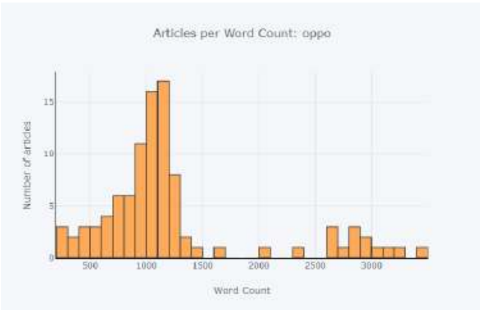

Figure 4: Class *oppo*: Histogram by word count.

and typos, increasing in that way the number of distinct words. Another interesting datum from Table 3 is that the vocabulary of pro-government journalists is considerably bigger than the one of opposition journalists ($|\mathscr{V}_{\mathscr{G}}| = 18497$ and $|\mathscr{V}_{\mathscr{O}}| = 13146$). One cause of this might be that due to the greater number of words in the documents of pro-government journalists ($\#_{\mathscr{G}} > \#_{\mathscr{O}}$) we will probably have a greater number of distinct words. For this reason, as vocabulary richness estimation it is frequently used the ratio between the size of vocabulary and the number of words in the collection. In that case, we can see that those metrics for pro-government ($|\mathscr{V}_{\mathscr{G}}|/\#_{\mathscr{G}}$) and opposition ($|\mathscr{V}_{\mathscr{O}}|/\#_{\mathscr{O}}$) journalists are the same.

Finally, it is worth to note that the intersection between pro-government and opposition vocabularies ($|\mathscr{V}_{\mathscr{G}} \cap \mathscr{V}_{\mathscr{O}}|$) is very small compared to the vocabulary of the whole corpus ($|\mathscr{V}_{\mathscr{C}}|$). That means that many words are used by a group and not by the other one, and the other way around. This point can be easily observed in Table 3 where the number of words used by pro-government and not by opposition journalists ($|\mathscr{V}_{\mathscr{G}} \setminus \mathscr{V}_{\mathscr{O}}|$) is 11177 and $|\mathscr{V}_{\mathscr{O}} \setminus \mathscr{V}_{\mathscr{G}}|$ = 5826. For instance, some words of $\mathscr{V}_{\mathscr{G}} \setminus \mathscr{V}_{\mathscr{O}}$ are "milicos", "globitos", "latinoamericana", "egoísmo", and "ultraderecha" and some words of $\mathscr{V}_{\mathscr{O}} \setminus \mathscr{V}_{\mathscr{G}}$ are "monárquica", "tribunera", "dilapidado", "negociaron", "hitlerismo", and "avaricia".

Another analysis that is usually informative is to measure the "relevance" of the terms in the corpus according to some specific metric. For instance, an approach is estimating the importance of a term according to the weight that it would receive in a particular document representation scheme, such as tf-idf. This scheme (tf-idf) is a weighted model commonly used for information retrieval problems. It is an unsupervised model in the sense that when weighting a term in a document, it does not take into account any information about the class that document belong to; for instance, in our corpus, if we take as terms the word uni-grams, the terms with the highest tf-idf value are: "colegio", "años", "comisión", "madre", "dijo", "perón", "día", "plata", "dos", "decía", "chica", "mujeres", "dice", "mamá", "después", "flaco", "casa", "alicia", "néstor", and "cristina". Taking as terms words 2-grams, the terms with the highest tf-idf value are: "próximo gobierno", "cinco años", "santa fe", "derechos humanos", "clase media", "muerte néstor", "años después", "día siguiente", "néstor kirchner", "muchas veces", "provincia buenos", "primera vez", "cristina dijo", "procurador general", "gils carbó", "cristina fernández", "buenos aires", "santa cruz", "néstor cristina" and, "río gallegos". Finally, "magdalena ruiz guiñazú", "economía axel kicillof", "ministro economía axel", "asignación universal hijo", "josé pablo feinmann", "joaquín morales solá", "gobernador provincia buenos", "da mucha bronca", "triple crimen general", "derechos humanos cidh", "cristina fernández kirchner", "mil millones dólares", "manuel abal medina", "juan manuel abal", "madres plaza mayo", "comisión interamericana derechos", "aumento mínimo imponible", "ciudad buenos aires", "interamericana derechos humanos", and "provincia buenos aires" are the terms with the highest tf-idf values when word 3-grams are used as terms.

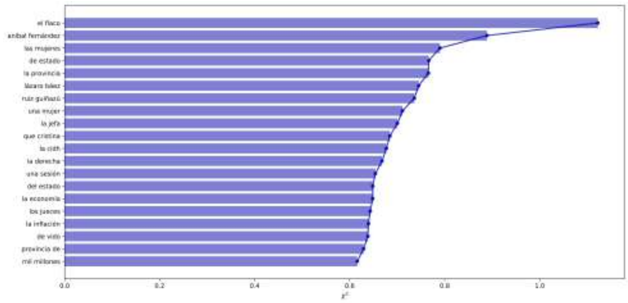

Figure 5: Top 20 word 2-grams according to χ^2 scores.

A second approach to measure the importance of terms, is considering *supervised* metrics that capture the importance of each term concerning its class/category, such as χ^2 and *information gain*, among others. They are usually used in *feature selection* processes to determine what are the most informative features to be preserved in the document representation. Here, we will calculate χ^2 scores for all the terms consisting in word 2-grams and the top 20 are shown in Figure 5.

A third alternative that usually shows interesting information about the importance of features, is to obtain

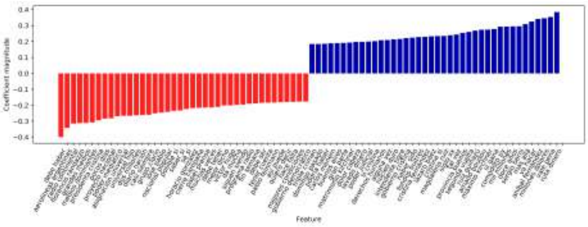

Figure 6: Largest and smallest coefficients of logistic regression trained on tf-idf features.

the *coefficients* of a model learned in a training stage of a prediction task with the whole corpus. In that case, those coefficients reflect how the learned model *weights each feature* for that task. In that way, we can look at the largest coefficients, and see which words these correspond to. For instance, Figure 6 shows a bar chart with the 25 largest and 25 smallest coefficients of a logistic regression model, with the bars showing the size of each coefficient. Word 2-grams are used as terms and the negative coefficients on the left belong to terms that according to the model are indicative of pro-government alignment, while the positive coefficients on the right belong to terms that according to the model indicate an article written by an opposition journalist. Most of the terms are quite intuitive, in the sense that reflect aspects very recognizable in texts from both political alignment, such as “grandes medios”, “medios dominantes”, and “nacional popular” indicating pro-government journalists, while “ruta dinero”, “lázaro baez” and, “lavado dinero” indicate opposition documents.

An analysis similar to the one carried out with the learned coefficients by a method like logistic regression can be achieved by using a novel classification system named *SS3*. SS3 is a supervised learning model for text classification that naturally supports incremental classification/learning and provides tools for interpretable/explainable decisions [14]. It is a system that allows dealing with data streams and in fact, has become the-state-of-the-art in several early risk detection (ERD) problems on social media [15]. Beyond its effectiveness in this type of problems, SS3 provides support for interpretable (white box) decision making with a friendly interactive environment for experimentation[4] and Python implementation [13].

A key component in the SS3 system is that in charge of estimating the relative importance of terms found in texts, denoted *gv* (or *global value*). In a few words, the function $gv(w,c)$ allows valuing *words* in relation to *categories*. More specifically, *gv* takes a word w and a category c and outputs a number in the interval $[0,1]$ representing the degree of *confidence* with which w is believed to *exclusively* belong to c. There, $gv(w,c) = v$ is read as “w has a global value of v in c” or, alternatively, “the *global value* of w in c is v”.

[4]https://github.com/sergioburdisso/pyss3

An interesting aspect empirically corroborated in [14] is that the global value correctly captures the *significance* and *discriminating* power of words by giving high scores to mid-frequency words that result to be the most informative.

In that context, in the present article, we computed the $gv(w,c)$ values for both categories $c = gov$ and $c = oppo$ and we considered as terms w to the sets of uni, bi and, three-grams of words obtained with each category. Then, we selected the top 50 terms (according to the *gv* score) of each category and we could observe some interesting patterns described below.

In pro-government journalists (class *gov*) the most informative uni-grams has to do with singular first person pronouns (“me”, “mi”, “yo”), people like “nestor”, “peron”, “alicia”, “gladis”, “valeria”, and different words like “mujer”, “mujeres”, “libro”, “flaco”, “pueblo”, “día”, “derecha”, “padre”, “peronista”, “cara”, “militantes”, “clase”, “fuerza”, “chica”, “nota”, “peronistas”, “madre”, and “voz”. With bi-grams, the most informative terms were those included the word “que” (“que me”, “que era”, “que había”, “mas que”, “que te”), the word “era” (“que era”, “era una”, “era la”, “no era”, “era el”, “era un”), and different bi-grams like “la casa”, “los dos”, “y cristina”, “nestor y”, “el flaco”, “una mujer”, “casa de”, “el peronismo”, “un tipo”, “la derecha”, “las mujeres”, “la cara”, “la plaza”, “a nestor”, “de nestor”, “el dia”, “la politica”, “los grandes”, “el pueblo”, “rio gallegos”, and “su vida”. Finally, three-grams representative of this class were “nestor y cristina”, “la casa de”, “de la plata”, “de la historia”, “de su vida”, “en este pais”, “por primera vez”, “no se puede”, “de los medios”, “la casa rosada”, “la muerte de”, “los grandes medios”, “a la politica”, “de la democracia”, “de la vida”, “de la rua”, and “en la plaza”.

On the other hand, journalists considered as the opposition to the government (class *oppo*) the informative uni-grams were those related to the field of justice (“juez”, “fiscal”, “jueces”, “corte”, “fiscales”, “judicial”, “penal”), people like “scioli”, “massa”, “daniel”, “anibal”, “baez”, “sergio”, “lazaro”, “lanata” “lopez” and “cfk” and different words like “estado”, “dolares”, “mil”, “inflacion”, “jefa”, “corrupcion”, “argentinos”, “futuro”, “dinero”, “precios”, “pagar”, “presidenta”, “funcionarios”, “supuesto”, “aumento”, “banco”, “control”, “inseguridad”, “delito”, and “efedrina’. As examples of informative bi-grams, we can mention “la provincia”, “de buenos”, “la jefa”, “la justicia”, “la economia”, “la presidenta”, “la inflacion”, “el juez”, “la causa”, “la corte”, “el futuro”, “la corrupcion”, “nivel de”, “la camara”, “la constitucion”, and different names of pepole like “daniel scioli”, “maximo kirchner”, “anibal fernandez”, “sergio massa”, and “lazaro baez”. Finally, three-grams representative of this class were “la provincia de”, “de buenos aires”, “la jefa de”, “jefa de estado”, “frente para la”, “para la victoria”, “de la campora”, “de la economia”, “de cristina fernandez", "la falta de", and "libertad de expresion".

3 Topic Modelling and Sentiment Analysis

Topic modeling is an umbrella term describing a class of text analysis methods whose task is assigning each document to one or multiple *topics*, usually without supervision. A good example of this is news data, which might be categorized into topics like "politics", "sports", "finance", and so on. Intuitively, a topic is a group of words that appear together frequently. In that context, "topics" obtained by a topic modeling process might not be what we would normally call a topic in everyday speech. In other words, the obtained groups (topics) might or might not have a semantic meaning clearly identifiable by a person. Often, when people talk about topic modeling, they refer to one particular decomposition method called *Latent Dirichlet Allocation* (LDA).

LDA is a generative probabilistic model for discrete data collections, such as text collections. LDA represents documents as a mixture of different topics or themes; each topic consists of a set of words that maintain some semantic link between them. The words, in turn, are chosen based on a probability.

The process of selecting topics and words is repeated to generate a document or a set of documents. As a result, each generated document is made up of different topics [16].

Using a generative model in reverse, LDA analyzes the set of documents to find the most likely set of topics possibly addressed in a document. We can consider LDA as a tool that generates groups of similar words, such as LIWC; but unlike LIWC, LDA automatically generates groups of words (topics). Also, the topics of LDA are not tagged, and their content is different depending on the corpus where LDA is trained. In summary LDA not only tries to find group of words that appear together frequently. It also requires that each document can be understood as a "mixture" of a subset of the topics.

As an example, applying LDA to the pro-government and opposition documents and setting the number of topics to 100, several topics with intuitive meaning are obtained. Table 4 and Table 5 show the first 20 words of some of those topics, three of the pro-government documents (Table 4) and three of the opposition documents (Table 5). There, it can be observed that pro-government topics have to do with women's rights (***Topic #9***), Argentine debt with vulture funds (***Topic #27***) and social security plans (***Topic #49***), while opposition topics are related to communication media and journalists (***Topic #46***), some events related to what was popularly known as "the cause of the ephedrine" (***Topic #89***), and the relationship between the official Argentine cult and the Pope and some politicians (***Topic #91***).

Table 4: Some topics of pro-government journalists.

Topic #9: voto mujeres décadas femenino incluso ciclo siglo luchas derecho quiera derrota banderas fitzgerald capital feministas siglos diez políticamente evita consigue
Topic #27: fondos deuda buitre final tema ministro documento kicillof soberana necesidad procesos comunicado litigiosidad previsibilidad reestructuración anexo conformes soberanas deudas líderes
Topic #49: auh asignación pobreza fondos plan implicó reparación ése cfk octubre previsionales diputados pasaba narváez proyectos dirigencia impulso región decreto corporaciones

Another usual analysis of the texts in a corpus is the one focussed in *affective* aspects of the content and known as *sentiment analysis* (SA). Sentiment analysis, also called *opinion mining*, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [17]. SA is a very active research area in its own that has been addressed with different techniques such as supervised machine learning algorithms and lexicon-based methods, among others. Although most of the works, resources, and tools for SA correspond to the English language, there is a growing interest in its application for Spanish with scientific events specifically dedicated to this task [18]. SA can be addressed at three different levels: *document*, *sentence* and *aspect* or *entity* being the last level probably the most challenging.

Table 5: Some topics of opposition journalists.

Topic #46 radio mitre canal censura lanata tn intento sabemos intervenir oyentes diario adecuación marcelo pánico convertirse puesta usureros colegas clarín clarin
Topic #89: aníbal desmentida triple kilos crimen efedrina quilmes melnyk clara importación agosto negocio granero morsa pérez nombres indispensable junto lanatta quizá
Topic #91: papa francisco iglesia ayuda mirada hombres página uca bergoglio cuervo vaticano guillermo michetti sienten larroque carrió explican quedaron opositores alegría

Sentiment analysis is an important and complex problem in its own whose exhaustive treatment would exceed the space limit of the present article. However, to contribute to the exploratory analysis of the corpus, some basic SA tasks can be presented. One of them is determining the polarity of each document by averaging the polarity of its component words. A popular tool for this task is **TextBlob**[5] that calculates sentiment polarity in the range of $[-1, 1]$ where 1 means *positive* sentiment and -1 means *negative* sentiment. In that way, it is usual to show some articles with the highest/lowest or even close to neutral (zero) sentiment polarity score or give some distribution of the articles according to their polarity scores. Due to space restrictions, we will only analyze the last alternative presenting the sentiment polarity distributions of both, pro-government (Figure 7), and opposition (Figure 8) journalists. Those graphics were obtained by applying TextBlob by separate to the translated (from Spanish to English) versions of the documents in both classes of journalists.

There, it can be observed a greater frequency of pro-government articles on higher scores (around 0.1) than the one shown by the opposition journalists (around 0.05). Besides, the highest positive score achieved by opposition journalists (0.2) is surpassed by several pro-government articles. The other way around, the global lowest (negative) score is also obtained by pro-government journalists (less than -0.1) indicating that pro-government articles show the greatest variation range in polarity scores.

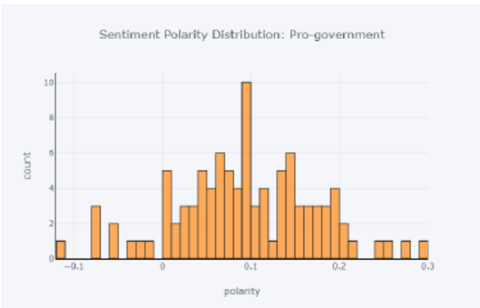

Figure 7: Histogram of pro-government articles.

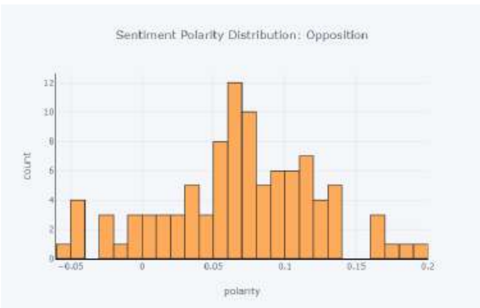

Figure 8: Histogram of opposition articles.

[5] https://textblob.readthedocs.io/en/dev

4 LIWC-based Analysis

LIWC is a tool developed by the American psychologist J. Pennebaker and colleagues [19] and has been used in several studies related to the psychological aspects of individuals. LIWC calculates the proportions of certain grammatical, lexical, and semantic markers, as well as markers belonging to other categories (up to 90 text features depending on the version). In our study, we used the most recent version of LIWC, LIWC2015 [20]. For each text file, LIWC2015 generates approximately 90 output variables as one line of data to an output file. This data record includes the file name and word count, 4 summary language variables (analytical thinking, clout, authenticity, and emotional tone), 3 general descriptor categories (words per sentence, percent of target words captured by the dictionary, and percent of words in the text that are longer than six letters), 21 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.), 41 words categories tapping psychological constructs (e.g., affect, cognition, biological processes, drives), 6 personal concern categories (e.g., work, home, leisure activities), 5 informal language markers (assents, fillers, swear words, netspeak), and 12 punctuation categories (periods, commas, etc.).

Properties of documents generated by LIWC2015 have been used as document representations in several studies [21] and also to analyze how these measures differ between texts of different classes [22]. This last approach will be the one used in the present article. We will first identify what are the features/characteristics in which there are statistical differences between both classes and then we will show more detailed information on some of them. Since the distribution of the feature values is not known and we cannot make any assumption about it, we used, the same as similar works with LIWC features [23], the (non-parametric) Wilcoxon signed-rank test for comparing paired data samples with a p-value < 0.05 for statistical significance. The null hypothesis ($H0$) that we are trying to refute is that there is no statistically significant relationship between the mean value of a feature belonging to the pro-government class and the mean value of the same feature belonging to the opposition class. In that context, we determined significant statistical differences in 34 LIWC categories. For instance, pro-government journalists show greater use of *verbs*, *adverbs*, *first-person singular* (“yo”, “mi”, “mio”), *social* processes (“compañero”, “hablar”, “ellos”) and, *perceptual* processes (“mirar”, “escuchar”, “sentir”). Opposition journalists, on the other hand, make a higher use of words with a *length* > 6 letters and make more references to expressions related to *money* (“dinero”, “efectivo”, “adeudar”). As an example, Figure 10 shows comparative boxplots of pro-government and opposition journalists for 2 LIWC categories with statistically significant differences: Perceptual processes and Money.

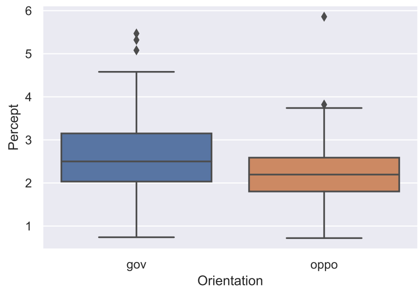

Figure 9: Comparative box-plots for Perceptual processes categories.

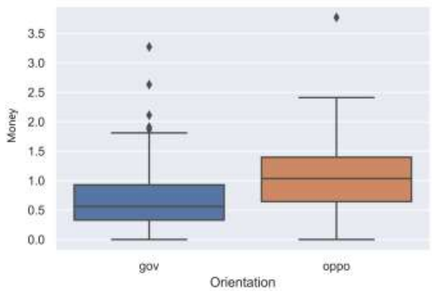

Figure 10: Comparative box-plots for Money categories.

As can be seen, some of the categories identified as statistically significative for both classes (like "first-person singular" and "Money") include different words that were already detected as informative in the previous analysis carried out with SS3 that associated some first-person singular pronouns with pro-government journalists and terms related to money with the opposition.

5 Conclusions and Future Work

This article introduces a new corpus for political alignment identification of Argentinian Journalists that, as far as we know, is the first collection with those characteristics. In that context, a comprehensive analysis of that corpus has been carried out which included the study of the corpus statistics, topic modelling and sentiment analysis, and a comparison of texts based on LIWC categories. In our opinion, the presented data set and the exploratory analysis carried out on both types of journalists is an interesting scientific contribution for those researchers working in author profiling in general and political alignment identification of Argentinian journalists in particular. We consider that our study allowed gaining insight about the main characteristics that are distinctive to identify both types of journalists and can be useful for identifying relevant features in document representation models for predictive tasks based on machine learning approaches.

As a result of this analysis, some interesting patterns were identified that reveal evident differences between the writings of pro-government and opposition journalists. For instance, topics identified in pro-government journalists have to do with women's rights, Argentine debt with vulture funds and social security plans while opposition topics are more related to communication media and journalists, some events related to corruption cases, and the relationship between the official Argentine cult and the Pope and some politicians.

Regarding the specific terms that were identified as relevant to characterize both classes, it could be seen that first-person singular pronouns, colloquial names to refer to the president and the former president ("néstor", "el flaco", "néstor y cristina") and more ideological/politicised discourse ("militantes", "la derecha", "la política", "el pueblo", "el peronismo", "los grandes medios") seem to be predominant in pro-government journalists. On the other hand, expressions related to the justice ("juez", "fiscal", "la corte"), money, alleged government flaws ("corrupción", "inflación", "inseguridad", "libertad de expresión"), more formal/distant references to the president ("cfk", "la presidenta", "jefa de estado", "de cristina fernandez") and the use of proper names of people linked to corruption cases seem to be indicative of opposition journalists.

Some differences were also observed in how their sentiment polarities are distributed, and that was more evident in the analysis with the LIWC system where we determined significant statistical differences in 34 LIWC categories.

As future work, we plan to use the different types of information obtained in the present work in the representation of documents for supervised (classification) and, non-supervised (clustering) tasks. Thus, the idea is using LIWC-based and LDA/topics-based features in text classification tasks, and comparing them against classical (bag of words) and more recent approaches like deep neural-networks with word embeddings.

Finally, we will reorganize the documents of the corpus analyzed in the present work according to the *authors* of these documents. In that way, we will have ten different classes (one for each journalist) and the task will be addressed as an *authorship attribution* task. An interesting point, in this case, will be determine how the hardness of this task is incremented when the authorship attribution is constrained to journalists of the same political orientation.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

ME and AV conceived the idea, VM and ME conducted the experiments. All authors analyzed the results, wrote and revised the manuscript.

Acknowledgements

The authors acknowledge the support of Universidad Nacional de la Patagonia Austral. The last author also acknowledges the constant support afforded by the Universidad Nacional de San Luis.

References

[1] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[2] R. Cohen and D. Ruths, "Classifying political orientation on twitter: It's not easy!," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[3] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom)*, pp. 192–199, 10 2011.

[4] K. Lazaridou and R. Krestel, "Identifying political bias in news articles," *Bulletin of the IEEE TCDL*, vol. 12, 2016.

[5] A. Tumasjan, T. O. Spenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment.," *Proceedings of the Fourth International AAAI conference on Weblogs and Social Media (ICWSM), Washington, S. 178-185.*, (2010).

[6] R. B. Slatcher, C. K. Chung, J. W. Pennebaker, and L. D. Stone, "Winning words: individual differences in linguistic style among u. s. presidential and vice presidential candidates. j," *ournal of Research in Personality, vol. 41, pp.63-75.*, (2007).

[7] J. W. Pennebaker and T. C. Lay, "Language use and personality during crises: analyses of mayor rudolph giuliani's press conferences.," *Journal of Research in Personality, vol. 36, pp.271-282.*, (2002).

[8] M. J. Carrera-Fernández, J. Guárdia-Olmos, and M. Peró-Cebollero, "Linguistic style in the mexican electoral process: Language style matching analysis," *Revista Mexicana de Psicología, 31(2), 138-152.*, (2014).

[9] M. Fernández-Cabana, J. Rúas-Araújo, and M. T. Alves-Pérez, "Psicología, lenguaje y comunicación: análisis con la herramienta liwc de los discursos y tweets de los candidatos a las elecciones gallegas de 2012," *Anuario de Psicología, 44(2), pp.169-184.*, (2014).

[10] J. Rúas-Araújo, M. Alves-Pérez, and M. Fernández-Cabana, "Comunicación, lenguaje y política: Análisis de los discursos institucionales del presidente de ecuador, rafael correa (2007-2015), con la herramienta liwc. communication, language and politics: Analysis of the institutional discourses of the president of ecuador, rafael correa (2007-2015), with the tool liwc.," *Razón y Palabra. vol.20, nro. 4-95, pp.591-607*, (2017).

[11] V. Mercado, A. Villagra, and M. Errecalde, "Exploratory analysis of a new corpus for political alignment identification of argentinian journalists," in *Actas del XXV Congreso Argentino de Ciencias de la Computación (CACIC 2019)*, pp. 507–516, 2019.

[12] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Information Sciences*, vol. 311, pp. 18 – 38, 2015.

[13] S. G. Burdisso, M. Errecalde, and M. M. y Gómez, "Pyss3: A python package implementing a novel text classifier with visualization tools for explainable ai - arxiv 1912.09322," 2019.

[14] S. G. Burdisso, M. Errecalde, and M. M. y Gómez, "A text classification framework for simple and effective early depression detection over social media streams," *Expert Systems with Applications*, vol. 133, pp. 182 – 197, 2019.

[15] S. G. Burdisso, M. L. Errecalde, and M. Montes y Gómez, "Unsl at erisk 2019: a unified approach for anorexia, self-harm and depression detection in social media," in *Working Notes of the Conference and Labs of the Evaluation Forum - CEUR Workshop Proceedings*, vol. 2380, 2019.

[16] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research 3*, no. 993-1022, 2003.

[17] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[18] M. C. Díaz-Galiano, M. G. Vega, E. Casasola, L. Chiruzzo, M. Á. G. Cumbreras, E. Martínez-Cámara, D. Moctezuma, A. M. Ráez, M. A. S. Cabezudo, E. S. Tellez, M. Graff, and S. Miranda-Jiménez, "Overview of tass 2019: One more further for the global spanish sentiment analysis corpus," in *IberLEF@SEPLN*, 2019.

[19] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "Linguistic inquiry and word count (liwc)," *[Software].*, (2001).

[20] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, *The development and psychometric properties of LIWC2015*. University of Texas at Austin, Austin, TX, 2015.

[21] M. P. Villegas, D. Funez, M. J. Ucelay, L. Cagnina, and M. Errecalde, "Lidic - unsl's participation at erisk 2017 : Pilot task on early detection of depression.," in *Working Notes of the Conference and Labs of the Evaluation Forum - CEUR Workshop Proceedings*, vol. 1866, 2017.

[22] T. A. Litvinova, P. V.Seredin, O. A. ALitvinova, and O. V.Romanchenko, "Identification of Suicidal Tendencies of Individuals Based on the Quantitative Analysis of their Internet Texts," *Computación y Sistemas*, vol. 21, pp. 243 – 252, 06 2017.

[23] J. Rúas, M. Fernández, and I. Puentes, “Aplicación de la herramienta liwc al análisis del discurso político. los mítines de los candidatos en las elecciones al parlamento de galicia de 2012.,” in *Actas del 2do Congreso Nacional sobre Metodología de la Investigación en Comunicación*, pp. 47–64, 2013.

Citation: V. Mercado, A. Villagra and M. Errecalde. *Political Alignment Identification: a Study with Documents of Argentinian Journalists*. Journal of Computer Science & Technology, vol. 20, no. 1, pp. 43–52, 2020.
DOI: 10.24215/16666038.20.e05.
Received: March 3, 2020 **Accepted:** May 17, 2020.