ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

Análisis Exploratorio de los Datos para la Identificación de la Alineación Política de Periodistas Argentinos

Exploratory Data Analysis for Political Alignment Identification of Argentinian Journalists

Viviana Mercado[1], Andrea Villagra[1], Marcelo Errecalde[1, 2]
vmercado@uaco.unpa.edu.ar, avillagra@uaco.unpa.edu.ar, merreca@unsl.edu.ar
[1] Unidad Académica Caleta Olivia - Universidad Nacional de la Patagonia Austral
Instituto de Tecnología Aplicada - LabTEm- Laboratorio de Tecnologías Emergentes
Ruta N°1 Acceso Norte – Caleta Olivia- Santa Cruz - Argentina
[2] Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Universidad Nacional de San Luis – San Luis

Recibido: 05/08/2020. Aceptado: 17/11/2020

RESUMEN

En la Minería de Datos, un área que ha ganado interés es la *determinación del perfil del autor*, que identifica patrones compartidos por un grupo de personas y aborda problemas de clasificación de los usuarios de la Web de acuerdo a edad, género, orientación política, etc. En este informe, se tomará como caso de estudio la tarea de *identificación de la alineación política*, la cual tiene como objetivo determinar a partir de los escritos de las personas el sesgo político. En este tipo de campo, un aspecto clave es disponer un conjunto de datos adecuados para que los procesos de minería de datos y aprendizaje automático puedan obtener resultados confiables. Para ello, se realiza la experimentación sobre un corpus para el estudio de la alineación política en documentos de periodistas argentinos. El estudio incluye varios tipos de análisis de documentos de periodistas opositores y progubernamentales, como ser la relevancia de los términos en cada clase de documentos, el análisis de sentimientos, el modelado de tópicos y el análisis de indicadores psicolingüísticos obtenidos del sistema LIWC. En los resultados experimentales, se pueden observar patrones interesantes, como por ejemplo los tópicos sobre los que escriben ambos tipos de periodistas, cómo se distribuyen las polaridades de los sentimientos y cómo los escritos de los periodistas progubernamentales y periodistas opositores difieren en las distintas categorías de LIWC.

Palabras clave: Minería de Texto; Análisis Exploratorio de los Datos (AED); Alineamiento Político de Periodistas; Análisis de Sentimientos; LIWC.

ABSTRACT

A field that is gaining interest in Data Mining is *author profiling*, the identification of patterns shared by a group of people. It includes classification problems of Web's users according to their age, gender, and political alignment, among others. This report addresses the *political alignment identification* problem, an author profiling task that aims at identifying political bias/orientation in people's 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 methods can obtain reliable and informative results. Thus, the experimental work uses a new

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

corpus for the study of political alignment in documents of Argentinean 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 modeling 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: Text Mining; Exploratory Data Analysis (EDA); Journalist Political Alignment; Sentiment Analysis; LIWC.

1. INTRODUCCIÓN

A partir de la disponibilidad de volúmenes inmensos de información en la Web, se refuerza cada día más el rol de la Minería de Datos (MD) como una herramienta fundamental para hacer un uso adecuado y ventajoso de esta información. En particular, un área que comienza a ganar creciente interés es la *determinación del perfil del autor* (DPA). La DPA, un sub-campo del área más general conocida como *análisis de autoría* (AA), es un tema muy importante de investigación principalmente por sus potenciales (y actuales) aplicaciones en problemas de seguridad nacional e inteligencia, lingüística forense, análisis de mercados e identificación de rasgos de personalidad, entre otros. Otro sub-campo del AA denominado *atribución de autoría* (ATA), consiste en la atribución de un texto de autoría desconocida a uno de un conjunto de autores potenciales. Si bien existen diversas herramientas para trabajar en MD, DPA y la ATA, presentan usualmente el problema que están dispersas, escritas en lenguajes y plataformas diferentes y, en muchos casos, como en el análisis de información textual, no están disponibles para el idioma español.

La *identificación del alineamiento político* (IAP) en un texto o documento es una forma de DPA, una de las tareas principales de AA junto con la atribución/determinación de autoría, detección de plagio y detección de inconsistencia de estilo. La IAP, de igual forma que otras tareas de DPA, como la detección de personas con depresión o con diferentes rasgos de personalidad, pedófilos y suicidas, es una tarea desafiante dentro del análisis automático de textos, ya que implica, en general, el uso de representaciones de textos que capturan aspectos estilísticos y de contenido de sus autores. En este contexto, un área particular dentro del IAP es aquella que está orientada al estudio de la orientación política en textos escritos por periodistas, y al que nos referiremos de ahora en más como textos periodísticos. Consideraremos como textos periodísticos aquellos que un periodista publica en diversos medios como un blog personal, un artículo escrito en un medio de comunicación como un periódico o el contenido expresado en un libro de su autoría. La IAP se ha aplicado a textos generados por usuarios habituales de redes sociales como Twitter (Cohen, R. and Ruths, D., 2013) aunque más recientemente se ha hecho con los documentos producidos por periodistas (Lazaridou, K. and Krestel, R., 2016). Sin embargo, estos textos han sido escritos principalmente en inglés o en otros idiomas.

En este artículo presentamos un primer acercamiento a la IAP en textos periodísticos en español, en particular, de textos generados por periodistas argentinos. La tarea, en este caso, consiste en agrupar todos los documentos de periodistas "progubernamentales" por un lado y

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

"opositores" por el otro. De esa manera, permitirá en el futuro visualizarlos como un problema de clasificación binaria ("progubernamentales" versus "opositores"). Además, se introduce un corpus para el estudio de la alineación política en los documentos de los periodistas argentinos. El estudio también incluye varios tipos de análisis de documentos de periodistas progubernamentales y opositores, como el análisis de indicadores psicolingüísticos obtenidos del sistema LIWC (Pennebakerm J.W. et. al. , 2015), el análisis de sentimientos y el modelado de tópicos obtenido con métodos como LDA (en inglés, Latent Dirichlet Allocation) (Blei, D., et. al, 2003).

En la Sección 2 presentamos el estado del arte, describiendo trabajos relacionados. En la Sección 3 describimos y analizamos el corpus. En la Sección 4 presentamos análisis basados en LIWC y LDA. Finalmente, en la Sección 5 comentamos los resultados y trabajos futuros.

2. TRABAJOS RELACIONADOS

La MD, la DPA y la ATA son áreas de investigación científica muy activas, sin embargo, surgen algunos inconvenientes cuando son aplicadas a problemas concretos de la vida real.
A diferencia de los estudios de laboratorio, donde es usual disponer de datos recolectados y procesados a priori, listos para ser analizados, el proceso de extracción de conocimiento (en inglés KDD, por *Knowledge Discovery in Data*) (Kurgan, L. and Musilek, P. , 2006), (Fayyad, U. et. al., 1996) involucrados en resolver problemas prácticos concretos requiere de varias etapas y herramientas para la recopilación de información, pre-procesamiento y extracción de características, análisis y visualización. El problema es que, usualmente, estas herramientas están dispersas, escritas en lenguajes y plataformas diferentes y, en muchos casos, como en el análisis de información textual, no están disponibles para el idioma español. Por lo tanto, realizar una experiencia concreta sobre uno o varios problemas particulares (como la DPA y la ATA) utilizando una plataforma de este tipo, permitirá ganar experiencia que podrá servir no sólo en problemas de Minería de Textos y de la Web como los involucrados en este artículo, sino en otras tareas de análisis futuros que involucran otros datos arbitrarios como, por ejemplo, imágenes, videos e información de sensores.

El AA (Stamatatos, 2009), se enfoca en la clasificación automática de textos basándose fundamentalmente en las elecciones estilísticas de los autores de los documentos, e incluye distintas tareas de análisis como, por ejemplo: a) la atribución de autoría, b) la verificación de autor, c) la detección de plagios, d) la determinación del perfil del autor y e) la detección de inconsistencias estilísticas. Los enfoques predominantes en esta área están basados en el aprendizaje automático supervisado. Es decir, estos enfoques derivan, a partir de un conjunto de datos etiquetados (conjunto de entrenamiento), un proceso inductivo de aprendizaje/entrenamiento y un clasificador, que puede generalizar sus predicciones a otros datos no observados previamente. Desde el momento que la disponibilidad de volúmenes de información en la Web es inmensa, se distingue cada día más el rol del AA como una herramienta fundamental para hacer un uso adecuado y ventajoso uso de esta información. Esto ha quedado plasmado en un incremento de Workshops y Competencias Internacionales específicos de esta temática y en particular, la DPA identificando patrones compartidos por un conjunto de personas, abordando problemas de clasificación por edad y género (Peersman, C. et. al., 2011), (Schler, J.et. al., 2006), (Argamon, S., et. al, 2005) nacionalidad, personalidad (Celli, F. et. al., 2014), (Mairesse, F., et. al., 2007), orientación política (Abooraig, R., et. al., 2014), (Conover, M., et. al., 2011), (Malouf, R. and Mullen, T., 2007), entre otros.

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

La IAP se ha aplicado a textos generados por usuarios de redes sociales como Twitter (Cohen, R. and Ruths, D., 2013), (Conover, M. D., et. al, 2011) aunque recientemente se ha realizado con documentos producidos por periodistas (Lazaridou, K. and Krestel, R., 2016). En (Tumasjan, A.,et. al, 2010) el discurso político en Twitter se analizó con LIWC durante la campaña electoral alemana de 2008. También la herramienta LIWC, se utilizó para determinar el estado psicológico y personalidad de los candidatos a la presidencia y vicepresidencia de los Estados Unidos en la campaña de 2004 (Slatcher, R.B.,et. al., 2007) y el lenguaje utilizado por el alcalde de Nueva York, R. Giuliani (Pennebaker, J. and Lay, T., 2002) a lo largo de su mandato. En cuanto a textos en lengua española, en (Carrera-Fernández,M.J., et.al., 2014) se analiza el estilo lingüístico de los candidatos de los principales partidos políticos en las elecciones generales de 2008 y 2011. Por otro lado, se aplicó el diccionario español de LIWC para analizar el discurso político y los tweets de los candidatos en las elecciones de Galicia en 2012 (Fernández-Cabana,M., et. al, 2014). Además, en (Rúas-Araújo, J., et. al, 2017) se analizan los discursos institucionales del presidente de Ecuador, Rafael Correa, desde 2007 hasta 2015, utilizando LIWC para determinar la existencia de diferencias en su estilo de lenguaje en los períodos analizados.

En cuanto al modelado de tópicos utilizando LDA, éste ha tomado gran importancia en diferentes disciplinas, por ejemplo, en temas como Redes Sociales, Ingeniería de software, Criminología, también en áreas de Geografía, Medicina / biomedicina, Ciencia lingüística, y en particular, en Ciencia Política se encuentran trabajos planteados por (Cohen, R. and Ruths, D., 2013) donde analizan la precisión de diferentes técnicas estándares para inferir la orientación política de usuarios de Twitter. En (Greene, D. and J.P. Cross., 2015) se analizan las interacciones políticas en el Parlamento Europeo (PE) para detectar tópicos latentes en los discursos legislativos a lo largo del tiempo. El contenido del discurso se analiza utilizando modelado dinámico de tópicos, basado en dos capas de factorización matricial. Los resultados sugieren que la agenda política del PE ha evolucionado significativamente con el tiempo, afectada por la estructura del comité del Parlamento que reacciona a eventos exógenos con un impacto significativo en lo que se debate en el Parlamento. En (Preoţiuc-Pietro, D., et. al., 2017) se examina la ideología política de los usuarios de Twitter utilizando una escala de siete puntos que permite identificar usuarios políticamente moderados y neutrales, grupos que son de particular interés para los científicos y encuestadores políticos. Los resultados identifican diferencias tanto en la inclinación política como en el compromiso y la medida en que cada grupo utiliza palabras clave políticas. Además, se demuestra cómo mejorar la precisión de la predicción ideológica explotando las relaciones entre los grupos de usuarios.

En este contexto, podemos decir que los enfoques anteriores están relacionados con nuestro trabajo, pero hasta donde sabemos, no hay estudios de IAP de textos periodísticos en español.

3. DESCRIPCION Y ANÁLISIS DEL CORPUS

Para generar el corpus se utilizó una colección de documentos periodísticos argentinos obtenidos de blogs de noticias, periódicos en línea, libros, etc. Este corpus consta de 196 documentos pertenecientes a 10 periodistas: 5 de ellos que respaldan claramente las acciones del gobierno argentino en el período 2012 a 2015 y 5 de ellos que se expresan explícitamente contra el gobierno en ese período. El conjunto de datos se dividió en dos grupos de documentos según la orientación política de los periodistas. De esta manera, se seleccionaron 98 documentos pertenecientes a los 5 periodistas progubernamentales para la clase *progubernamental* y los 98 documentos restantes de los periodistas de la oposición se usaron

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

para construir la clase *oposición* (Mercado, V., et . al., 2019). De esa forma, se obtuvo un corpus equilibrado con 2 clases.

Para seleccionar los documentos se tuvieron en cuenta algunas pautas:

- Los textos corresponden a documentos en español escritos por periodistas argentinos.
- Los textos se refieren a diferentes aspectos políticos relacionados con el gobierno argentino en el período 2012-2015, tales como acciones gubernamentales, declaraciones de políticos, casos de corrupción, tratamiento de leyes, etc.
- Todos los documentos contienen "texto formal", es decir, no presentan aspectos comunes "informales" del contenido de las redes sociales, como abreviaturas, expresiones de argot, errores tipográficos, hipervínculos, etiquetas, figuras y emoticonos.
- De cada periodista, se tomaron entre 18 y 20 documentos de su blog personal, artículos en periódicos en línea o libros digitales de su autoría.
- Cada periodista estaba claramente identificado como *progubernamental* (gob) u *oponente* (opo).
- La misma proporción de periodistas masculinos y femeninos se mantuvo entre ambas categorías.

Después de recopilar los documentos, se etiquetaron manualmente como pertenecientes a las dos clases mencionadas anteriormente gob y opo. La Tabla 1 muestra información sobre cómo se distribuyeron los documentos en ambas clases y cuáles fueron las fuentes (periódico en línea, blog o libros digitales) es decir, de donde se obtuvieron.

Clase	Periódicos	Blogs	Libros	Total
gob	50	46	2	98
opo	60	36	2	98

Tabla 1 : Distribución de documentos en clases y fuente.

La Tabla 2 muestra información sobre el número de palabras en los documentos (mínimo, máximo, medina, y desviación estándar) por clase.

Clase	Mínimo	Máximo	Mediana	Desviación Estándar
gob	139	36619	1865,18	5031,56
opo	236	3423	1243,01	733,71
gob + opo	139	36619	1554,09	3608,91

Tabla 2: Número de palabras en los documentos: mínimo, máximo, mediana y desviación estándar por clase.

Como podemos ver en la Tabla 2, aunque los tamaños de los documentos más cortos (mínimos) son similares para ambas clases (139 frente a 236), difieren considerablemente en los más largos (36619 frente a 3423). Cabe aclarar que, en ambas clases se ha distribuido la misma cantidad de libros que contienen mayor fuente de vocabulario que se representa en la Figura 1.

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

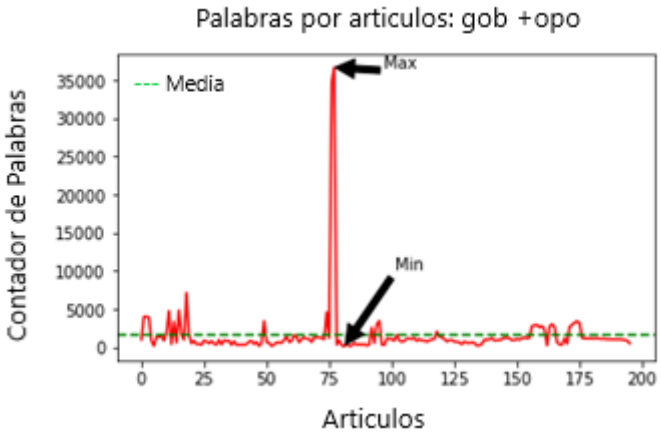

Figura 1 Número de Palabras en los documentos en las clases progubernamental y oposición.

Respecto a las palabras que aparecen en la colección y en cada una de las clases, en la Tabla 3, lo primero que se puede observar es que, aun cuando el número de palabras en todo el corpus es alto (*#C = 280343*), el número de palabras distintas (el tamaño del vocabulario)[1] es relativamente pequeño ($|V_C|$ = *24323*). Eso difiere del tamaño de los vocabularios en los textos de las redes sociales que generalmente son más grandes. Una posible causa de esto es que los escritos en las redes sociales suelen ser informales y propensos a tener abreviaturas y/o errores tipográficos, aumentando así el número de palabras distintas. Otro dato interesante es que el vocabulario de los periodistas progubernamentales es considerablemente mayor que el de periodistas de la oposición ($|V_G|$ = 18497 versus $|V_O|$ = 13146). Una de las causas de esto podría ser que debido a la mayor cantidad de palabras en los documentos de los periodistas progubernamentales *#G > #O*) probablemente tendremos una mayor cantidad de palabras distintas. Por esta razón, como estimación de la riqueza de vocabulario se usa con frecuencia la relación entre el tamaño del vocabulario y el número de palabras en la colección. Además, podemos ver que estas métricas para periodistas progubernamental ($|V_G|$ / *#G*) y opositores ($|V_O|$ / *#O*) son iguales.

#C	280343
$\|V_C\|$	24323
#G	167844
$\|V_G\|$	18497
$\|V_G\|$ /*#G*	0,11
#O	112499
$\|V_O\|$	13146
$\|V_O\|$ /*#O*	0,11

Tabla 3 Estadísticas sobre los documentos del corpus completo y de cada clase

Un análisis que suele ser informativo es medir la "relevancia" de los términos en el corpus según alguna métrica específica. Un enfoque es estimar la importancia de un término de acuerdo al peso que recibiría en un esquema particular de representación del documento, como *tf-idf*.. Este esquema (*tf-idf*) es un modelo ponderado comúnmente utilizado para problemas de recuperación de información y es no supervisado en el sentido de que cuando pondera un término en un documento, no tiene en cuenta ninguna información sobre la clase a

[1] El vocabulario de una colección de documentos es el conjunto de palabras distintas que aparecen en esa colección.

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

la que pertenece el documento. Por ejemplo, en nuestro corpus, si tomamos como términos los uni-gramas, los términos con el valor *tf-idf* más alto son: “colegio”, “años”, “comisión”, “madre”, “dijo”, “perón”, “día”, “plata”, “Dos”, “decía”, “chica”, “mujeres”, “dados”, “mamá”, “después”, “flaco”, “casa”, “Alicia”, “Néstor” y “cristina”. En la Figura 2 se muestra los uni-gramas de palabras, donde se puede observar en color más claro (rojo) aquellas que representan características con valores *tf-idf* más bajos y en color oscuro (azul) los que representan valores más altos para el corpus.

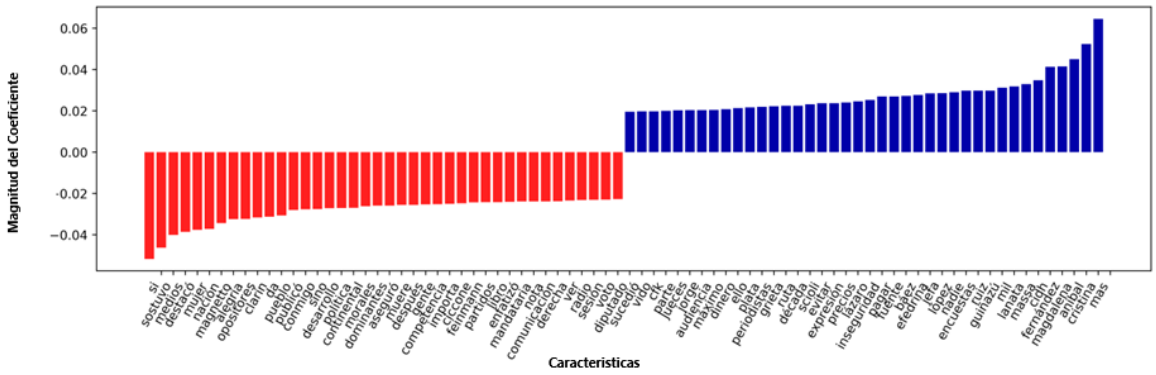

Figura 2 Uni-gramas de palabras

Tomando bi-gramas de palabras, los términos con el valor más alto de *tf-idf* son: “próximo gobierno”, “cinco años”, “santa fe”, “Derechos humanos”, “clase media”, “muerte néstor”, “años después”, “néstor Kirchner”, “muchas veces”, “provincia buenos”, “primera vez”, “Cristina dijo”, “procurador general”, “gils carbó”, “cristina fernández”, “buenos aires “,” “santa cruz”, “néstor cristina” y “rio gallegos”. Finalmente, “magdalena ruiz guiñazú”, “economía axel kicillof”, “ministro economía axel”, “asignación hijo universal”, “josé pablo feinmann”, “joaquın morales solá”, “gobernador provincia buenos”, “da mucha bronca”, “triple crimen general”, “derechos humanos cidh”, “cristina fernández kirchner”, “mil millones de dólares”, “manuel aval medina”, “juan manuel aval”, “madres plaza mayo”, “comisión interamericana derechos”, “aumento mínimo imponible”, “ciudad buenos aires”, “interamericana derechos humanos”, y “provincia buenos aires” son los términos con el valores más altos de *tf-idf* cuando se usan tri-gramas de palabras como términos. En la Figura 3 se presentan los bi-gramas de palabras donde, al igual que en la gráfica de la Figura 2, en color más claro (rojo) se presentan aquellos bi-gramas con valores de *tf-idf* más bajos y en color oscuro (azul) los valores más altos para el corpus.

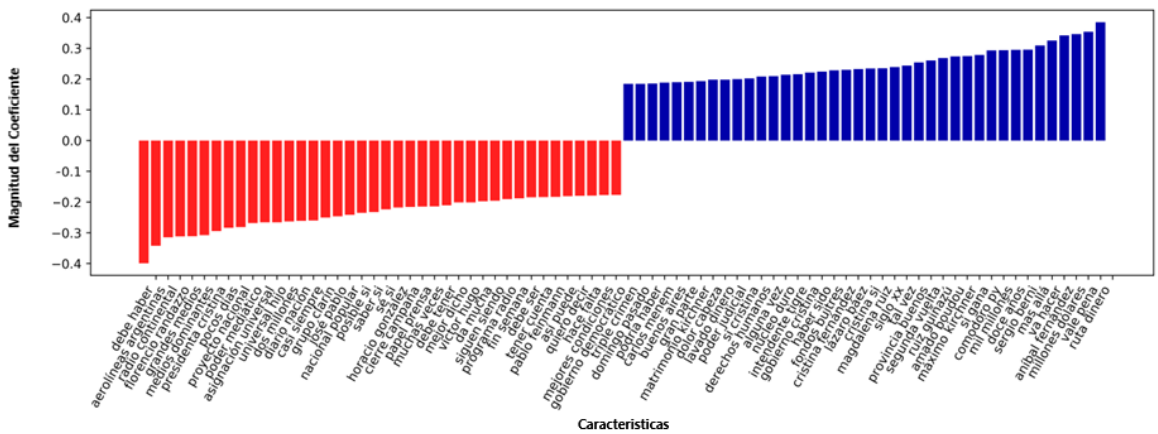

Figura 3 Bi-gramas de palabras

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

Otro enfoque para medir la importancia de los términos, considerado supervisado, captura la importancia de cada término con respecto a su clase/categoría, como estimada por métodos como χ^2 (Chi cuadrado) y ganancia de información, entre otros. Estas suelen utilizarse como función de evaluación en procesos de selección de atributos que determinan cuáles son las características más informativas que deben preservarse en la representación del documento. Aquí, se calculan los valores χ^2 para todos los términos que consisten en bi-gramas de palabras y los 20 principales, se muestran a continuación en la Figura 4. Podemos observar que "el flaco" es el de mayor valor de χ^2 mientras que "mil millones" es el que obtiene el menor, evidenciando que a mayor valor es mayor la relevancia del término.

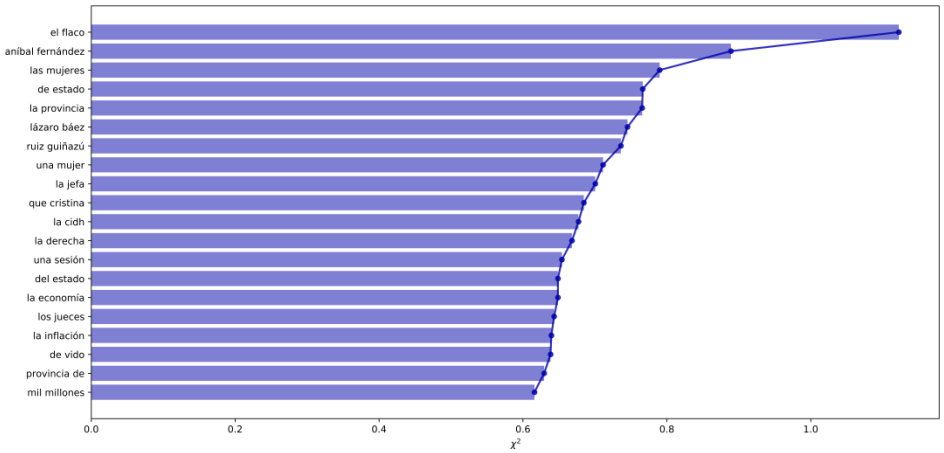

Figura 4 Chi Cuadrado bi-gramas de palabras.

4. ANALISIS BASADO EN LIWC Y MODELADO DE TOPICOS

LIWC es una herramienta desarrollada por el psicólogo estadounidense Pennebaker (Pennebaker, J. and Lay, T., 2002) y se ha utilizado en varios estudios relacionados con aspectos psicológicos de los individuos. LIWC calcula las proporciones de ciertas palabras en el texto que coinciden con cada una que se incluyen dentro de dimensiones del lenguaje (hasta 90 características de texto según la versión). En nuestro estudio, utilizamos las más recientes versiones de LIWC, LIWC2015 (Pennebaker, J.W, et. al., 2001), (Pennebakerm,J.W., et. al., 2015).

Para cada archivo de texto, LIWC2015 genera aproximadamente 90 variables de salida como una línea de datos en un archivo de salida. Esta información incluye el nombre del archivo y el recuento de palabras, 4 variables resumen de lenguaje (pensamiento analítico, influencia, autenticidad y tono emocional), 3 categorías de descriptores generales (palabras por oración, porcentaje de palabras objetivo capturadas por el dictado y el porcentaje de palabras en el texto que tienen más de seis letras), 21 dimensiones lingüísticas estándar (por ejemplo, porcentaje de palabras en el texto que son pronombres, artículos, verbos auxiliares, etc.), 41 categorías de palabras de construcciones psicológicas (p. ej., afecto, cognición, procesos biológicos, impulsos), 6 categorías de intereses personales (p. ej., trabajo, hogar, actividades de ocio), 5 marcadores de lenguaje informal (acuerdos, rellenos, palabrotas, jergas) y 12 categorías de puntuación (puntos, comas, etc.). Las propiedades de los documentos generados por LIWC2015 se han utilizado como representaciones de documentos en varios estudios y también para analizar cómo estas medidas difieren entre artículos de diferentes clases.

En el presente informe, primero identificamos cuáles son las características en la que existen diferencias estadísticamente significativas entre ambas clases y luego se muestra información

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

más detallada sobre algunas de ellas. Dado que la distribución de los valores de las características no se conoce y no se puede suponer nada al respecto, se utiliza, al igual que en trabajos similares con LIWC (Pennebakerm J.W., et.al, 2015), el test (no paramétrico) de Wilcoxon de rango con signo con un valor $p<0,05$ para significancia estadística. La hipótesis nula (H_0) que se intenta refutar es que no hay una relación estadísticamente significativa entre el valor de la media de una característica que pertenece a la clase progubernamental y el valor medio de la misma característica perteneciente a la clase opositora. En ese contexto, se determinaron diferencias estadísticas significativas en 34 categorías de LIWC. Por ejemplo, los periodistas progubernamentales muestran un mayor uso de verbos, adverbios, primera persona del singular (“yo”, “mi”, “mío”), procesos sociales (“compañero”, “hablar”, “ellos”) y, procesos perceptuales (“mirar”, “escuchar”, “sentir”). Por otro lado, los periodistas opositores hacen un mayor uso de palabras con una longitud mayor a 6 letras y hacen más referencias a expresiones relacionadas con dinero (“dinero”, “efectivo”, “adeudar”). Por ejemplo, la Figura 5 muestra diagramas de caja comparativos de periodistas progubernamentales y opositores para una característica de LIWC con diferencias estadísticamente significativas.

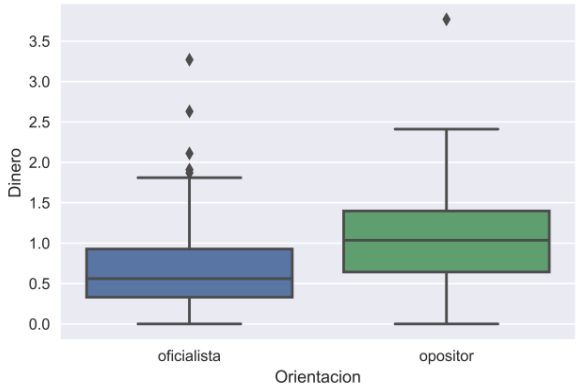

Figura 5 Blox-plot periodistas progubernamentales versus opositor ante la categoría Dinero.

El modelado de tópicos es un término general que describe una clase de métodos de análisis de texto cuya tarea es asignar cada documento a uno o múltiples tópicos, generalmente sin supervisión. Un buen ejemplo de esto son los datos de noticias, que podrían clasificarse en tópicos como “política”, “deportes”, “finanzas”, etc. Intuitivamente, un tópico es un grupo de palabras que aparecen juntas con frecuencia. En ese contexto, “tópicos” obtenidos por un proceso de modelado de tópicos podría no ser lo que normalmente llamaríamos un tópico en el habla cotidiana. En otras palabras, los grupos obtenidos (tópicos) pueden o no tener una semántica claramente identificable por una persona. A menudo, cuando la gente habla sobre modelado de tópicos, se refieren a un particular método de descomposición llamado LDA.

LDA es un modelo probabilístico generativo para colecciones discretas de datos, como colecciones de texto. LDA representa documentos como una mezcla de diferentes tópicos; cada tópico consta de un conjunto de palabras que mantienen algún vínculo semántico entre ellos. Las palabras, a su vez, se eligen en función de una probabilidad. Se repite el proceso de selección de tópicos y palabras para generar un documento o un conjunto de documentos. Como resultado, cada documento generado está compuesto de diferentes tópicos (Blei, D.,et. al., 2003).

Podemos considerar a la LDA como una herramienta que genera grupos de palabras similares, tales como LIWC; pero a diferencia de LIWC, LDA automáticamente genera grupos de palabras (tópicos) sin existir categorías pre-definidas. Además, los tópicos de LDA no están etiquetados y su contenido es diferente dependiendo del corpus donde se entrena LDA. En

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

resumen, LDA no solo trata de encontrar un grupo de palabras que aparecen juntos con frecuencia, sino que también requiere que cada documento puede entenderse como una "mezcla" de un subconjunto de los tópicos.
Como ejemplo, aplicando LDA a documentos progubernamentales y documentos de opositores y estableciendo el número de tópicos a 100, se obtienen varios tópicos con significados intuitivos.
La Tabla 4 muestra las primeras 20 palabras de algunos de esos tópicos, dos de los documentos progubernamentales y dos de los documentos opositores. También, se puede observar que los tópicos progubernamentales tienen que ver con derechos de las mujeres (Tópico # 9) y planes de seguridad social (Tópico # 49), mientras que los tópicos de oposición están relacionados con la comunicación, medios y periodistas (Tópico # 46) y la relación entre el culto oficial argentino y el papa y algunos políticos (Tópico # 91).

gob	opo
Tópico #9: voto mujeres décadas femenino incluso ciclo siglo luchas derecho quiera derrota banderas fitzgerald capital feministas siglos diez políticamente evita consigue	**Tópico #46** radio mitre canal censura lanata tn intento sabemos intervenir oyentes diario adecuación marcelo pánico convertirse puesta usureros colegas clarín clarin
Tópico #49: auh asignación pobreza fondos plan implicó reparación ése cfk octubre previsionales diputados pasaba Narváez proyectos dirigencia impulso región decreto corporaciones	**Tópico #91:** papa francisco iglesia ayuda mirada hombres página uca bergoglio cuervo vaticano guillermo michetti sienten larroque carrió explican quedaron opositores alegría

Tabla 4 Algunos tópicos de periodistas progubernamentales vs opositores.

Para la aplicación del método LDA, se utilizó el software disponible en la librería gratuita *gensim*[2,3]. Este es un paquete genérico en Python para modelado de tópicos, indexación de documentos y recuperación de similitudes. Existe una implementación de este modelo en el entorno de desarrollo *JupyterLab*[4] que tiene soporte para la aplicación web *Jupyter Notebook.* Asimismo, esta última, provee una manera simple de codificación y pruebas, modelado estadístico, simulación numérica, etc. en el lenguaje mencionado.

El método LDA con *gensim*, ofrece distintos parámetros para definir cómo entrenar y generar el modelo de tópicos. En primer lugar, requiere definir la *cantidad de tópicos (k)* a buscar. Luego, es necesario establecer los valores de *alfa* y *beta* que utilizará el modelo.
A continuación, en la Tabla 5 se presentan los parámetros y valores utilizados: a) *Cantidad de tópicos* (*k*) con valor 20; b) Hiperparámetros (*Alfa y Beta*) que afectan la densidad de tópicos, el valor de ambos es igual a 1/ k; c) número de documentos a ser utilizados en cada iteración de entrenamiento (*Chuncksize)* su valor es 5; d) verificaciones de actualización de los parámetros del modelo *(Update_every)* con valor 100*; y* e) cantidad de iteraciones por el corpus durante el entrenamiento (*Passes)* es igual a 25. Estos valores son los utilizados en la literatura y con los cuales luego de varios experimentos se han obtenido mejores resultados.

[2] https://radimrehurek.com/gensim/
[3] https://en.wikipedia.org/wiki/Bag-of-words_model.
[4] https://jupyter.org/

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

Parámetro	
Cantidad de tópicos	k=20
Hiperparámetro	*Alfa, Beta* =1/k
Chunksize	5
Update_every	100
Passes	25

Tabla 5 Valores de parámetros para gensim.

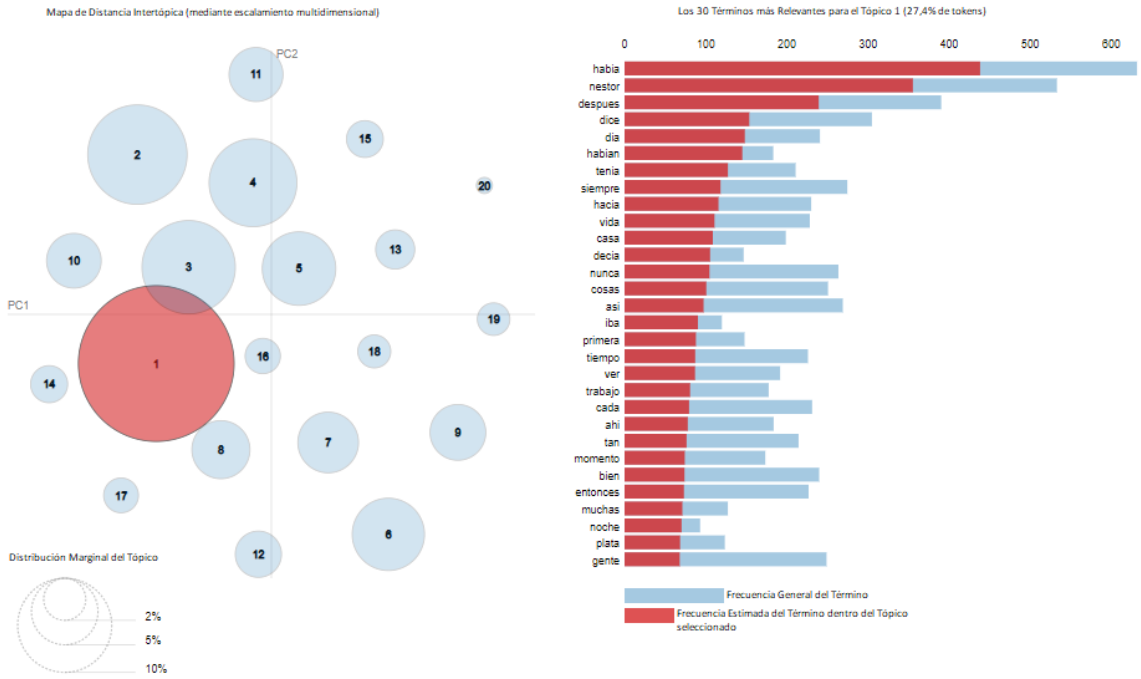

Figura 6 Visualización obtenida para términos más relevantes del Tópico 1.

En la Figura 6 se muestra una visualización obtenida con el modelo LDA utilizando la librería *pyLDAvis*. En la parte izquierda de la gráfica cada burbuja representa un tópico. Cuanto más grande la burbuja, se puede decir que es más predominante ese tópico. Para este caso en particular, se han tomado los 30 términos más importantes. Un buen modelo de tópicos es aquel que tiene burbujas bastante grandes y que no se solapan, además se puede observar que hay otras burbujas dispersas en los cuadrantes del gráfico.

Un modelo con muchos tópicos seguramente tendrá burbujas pequeñas, ubicadas en una misma región del gráfico y con muchos casos de solapamiento, además si observáramos las mismas palabras clave repetidas en muchos de los tópicos, sería probablemente una señal de que el parámetro *k* es muy grande, y no es lo que sucede en este caso. La parte derecha de la gráfica muestra las palabras clave más importantes que forman el tópico seleccionado, en este caso Tópico 1, con un 27,4% de *tokens*: “había”, “Néstor”, “después”, “dice”, entre otros. Podemos observar un único caso de solapamiento hacia el cuadrante inferior del gráfico (burbujas 1 y 3 con frontera en la burbuja 8). En cuanto a las palabras claves, están claramente definidas en un tópico que podríamos destacarlo de los restantes, se trata del pasado del verbo “haber”. Por otro lado, otro caso más disperso a la derecha (burbuja 20) y de tamaño menor a los restantes, se puede inferir que es debido al menor número de veces que se utiliza en los documentos.

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

Otro análisis habitual de los textos en un corpus está enfocado en aspectos afectivos del contenido y es conocido como análisis de sentimientos (en inglés *Sentiment Analysis*, SA). SA, también llamado minería de opinión, es el campo de estudio que analiza las opiniones, los sentimientos, las evaluaciones de las personas, actitudes y emociones hacia las entidades tales como productos, servicios, organizaciones, individuos, problemas, eventos, tópicos y sus atributos (Liu, B., 2012).
SA es un área de investigación muy activa en sí misma que ha sido abordada con diferentes técnicas como los algoritmos de aprendizaje automático supervisado y los métodos basados en léxicos, entre otros. Aunque la mayoría de los recursos, y herramientas para SA corresponden al lenguaje inglés, hay un creciente interés en su aplicación para el lenguaje español con eventos científicos específicamente dedicados a esta tarea (Diaz-Galiano, M.C. et. al., 2019).
SA puede abordar tres diferentes niveles: documento, oración y aspecto o entidad siendo el último nivel probablemente el más desafiante.

Para el análisis exploratorio de este corpus, se pueden presentar algunas tareas básicas de SA. Uno de ellos es determinar la polaridad de cada documento promediando la polaridad de sus palabras componentes. Una herramienta para esta tarea es *TextBlob*[5] que calcula la polaridad de sentimiento en el rango de [-1; 1] donde 1 significa sentimiento positivo y -1 significa sentimiento negativo. De esa manera, es habitual mostrar algunos artículos con el sentimiento más alto/más bajo o incluso cercano a neutral (cero) puntaje de polaridad o dar alguna distribución de los artículos según sus valores de polaridad.
A continuación, se muestran las distribuciones de polaridad de sentimiento de ambas clases de periodistas, es decir, progubernamental (Figura 7) y opositores (Figura 8).

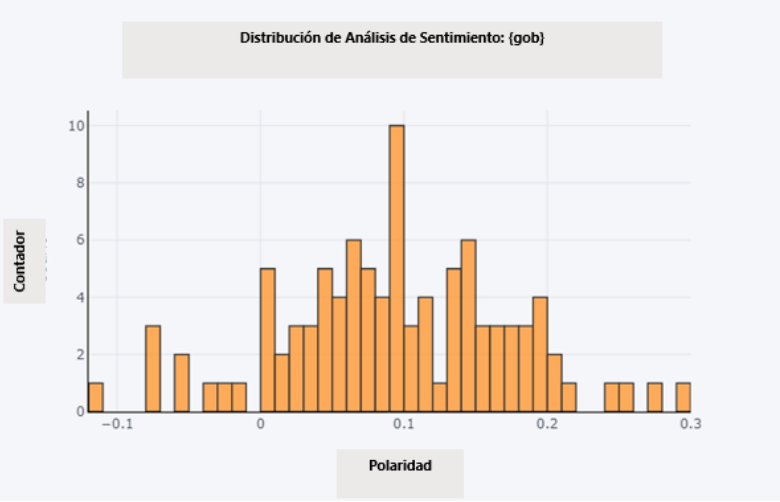

Figura 7 Histograma de artículos progubernamentales.

Esos gráficos se obtuvieron aplicando *TextBlob* por separado, traduciendo (del español al inglés) versiones de los documentos en ambas clases de periodistas.

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

ICT-UNPA-261-2020
ISSN: 1852-4516
Aprobado por Resolución N° 0540/20-R-UNPA
http://doi.org/10.22305/ict-unpa.v12.n3.749

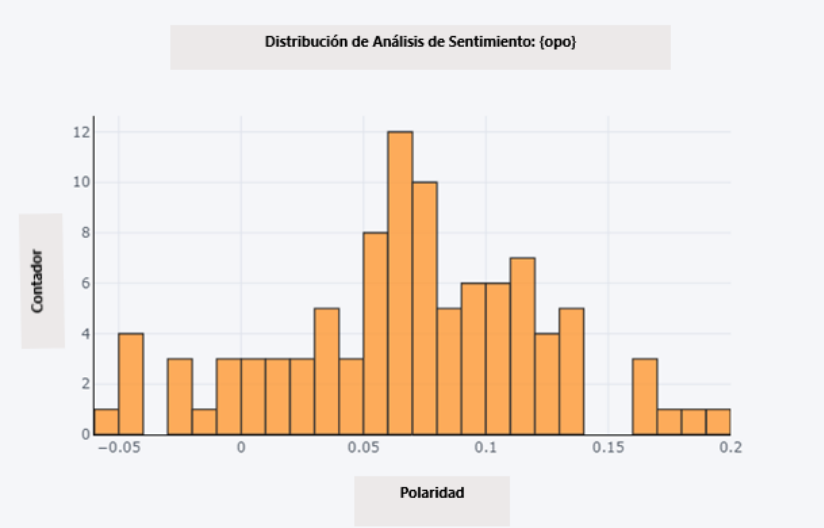

Figura 8 Histograma de artículos opositores.

Se puede observar una mayor frecuencia de los artículos progubernamentales con puntajes más altos (alrededor de 0,1) mientras que los artículos de periodistas opositores obtienen valores más pequeños (alrededor 0,05). Además, el puntaje positivo más alto alcanzado por periodistas opositores (0,2) es superado por varios artículos progubernamentales. Por otro lado, el puntaje más bajo global (negativo) también lo obtienen los artículos periodísticos progubernamentales (menos de -0,1) que indican estos artículos muestran el mayor rango de variación en puntajes de polaridad.

5. CONCLUSIONES

Este trabajo utiliza un corpus para la identificación de alineación política en periodistas argentinos que, hasta donde sabemos, es la primera colección con esas características. En ese contexto, se lleva a cabo un análisis exhaustivo de ese corpus, que incluye el estudio de las estadísticas, modelado de temas y análisis de sentimientos y una comparación de textos basados en las categorías de LIWC. En nuestra opinión, los datos presentados con este corpus y el análisis exploratorio en ambas clases de periodistas, es una contribución científica interesante para aquellos investigadores que trabajan en la creación de perfiles de autor en general y en la identificación de alineamiento político con textos de periodistas argentinos, en particular. Consideramos que nuestro estudio permitió obtener información sobre las características principales que son distintivas para identificar ambos tipos de periodistas y pueden ser útiles para identificar características en modelos de representación de documentos para predicción de tareas basadas en enfoques de aprendizaje automático.

Como resultado de este análisis, algunos patrones interesantes fueron identificados que revelan diferencias evidentes entre los escritos de periodistas progubernamentales y opositores. Por ejemplo, temas identificados en periodistas progubernamentales tienen que ver con los derechos de las mujeres, deuda con fondos buitres y planes de seguridad social mientras los temas de oposición están más relacionados con los medios de comunicación y los periodistas, algunos eventos relacionados con casos de corrupción y la relación entre el culto argentino, el Papa y algunos políticos.

En cuanto a los términos específicos que se identifican como relevantes para caracterizar ambas clases, se podría decir que pronombres singulares en primera persona, nombres coloquiales para referirse al presidente y al ex presidente (“Néstor”, “el flaco”, “néstor y cristina”) y discurso más ideológico / politizado (“militantes", “La derecha”, “la política”, “el pueblo”, “el peronismo”, “Los grandes medios”) parecen ser predominantes en periodistas

progubernamentales. Por otro lado, expresiones relacionados con la justicia (“juez”, “fiscal”, “la corte"), dinero, supuestos defectos del gobierno (“corrupción", “inflación”,“ inseguridad”, “libertad de expresión”), más referencias formales / distantes al presidente (“cfk”, “la presidenta”, “jefa de estado”, ”de cristina fernández”) y el uso de nombres propios de personas vinculadas a los casos de corrupción parecen ser indicativos de periodistas de la oposición. También se observaron algunas diferencias en cómo las polaridades de sentimientos se distribuyen, y eso fue más evidente en el análisis con el sistema LIWC donde determinamos diferencias estadísticas significativas en 34 Categorías de LIWC.

Como trabajo futuro, planeamos usar los diferentes tipos de información obtenida en el presente trabajo en la representación de documentos para tareas supervisados (clasificación) y tareas no supervisadas (agrupamiento). Por lo tanto, la idea es utilizar características/tópicos basados en LIWC y LDA en tareas de clasificación de textos y compararlos con enfoques clásicos (bolsa de palabras) y con enfoques más recientes como redes neuronales profundas con incrustaciones de palabras.

Finalmente, reorganizaremos los documentos del corpus analizado en el presente trabajo según los autores de estos documentos. De esa manera, tendremos diez clases diferentes (una para cada periodista) y se abordará como una tarea de atribución de autoría.

Un punto interesante, en este caso, será determinar cómo se incrementa la dureza de esta tarea cuando la atribución de autoría está limitada a los periodistas de la misma orientación política.

6. AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto 29/B225 “Soluciones inteligentes para el desarrollo urbano sostenible”. Los autores agradecen a la Universidad Nacional de la Patagonia Austral.

7. REFERENCIAS BIBLIOGRÁFICAS

ABOORAIG, R., ALWAJEEH, A., AL-AYYOUB, M., and HMEIDI, I. (2014). On the automatic categorization of arabic articles based on their political orientation. *Proc. of the Third International Conference on Informatics Engineering and Information Science (ICIEIS2014).*

ARGAMON, S., DHAWLE, S., KOPPEL, M., and PENNEBAKER, J. (2005). Lexical predictors of personality type. *Joint Annual Meeting of the Interface and the Classification Society of North America.*

BLEI, D., NG, A. and JORDAN,M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3.*, 993-1022.

CARRERA-FERNÁNDEZ, M. J., GÚARDIA-OLMOS, J., and PERÓ-CEBOLLERO, M. (2014). Linguistic style in the mexican electoral process: Language style matching analysis. *Revista Mexicana de Psicología, 31(2),* 138-152.

CELLI, F., LEPRI, B., BIEL, J., GATICA-PEREZ, D., RICCARDI, G., and PIANESI, F. (2014). The workshop on computational personality recognition 2014. *Proceedings of the ACM International Conference on Multimedia, MMA'14, pages 1245-1246, New York, NY, USA. ACM.*

CHUANG, J., MANNING, C. D., AND HEER, J. (2012). Termite: Visualization Techniques forAssessing Textual Topic Models. *AVI '12: Proceedings of the International Working Conference on Advanced Visual Interfaces.*, 74-77.

COHEN, R. and RUTHS, D. (2013). Classifying political orientation on twitter: It's not easy. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.*

CONOVER, M., GONÇALVES, B., RATKIEWICZ, J., FLAMMINI, A., and MENCZER, F. (2011). Predicting the political alignment of twitter users. *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom).*

DIAZ-GALIANO, M. C., VEGA, M. G, CASASOLA, E., CHIRUZZO, L., CUMBRERAS, M. A. G., MARTINEZ-CAMARA, E., MOCTEZUMA, D., RAEZ, A. M., CABEZUDO, M. A. S., TELLEZ, M., GRAFF, M., and MIRANDA-JIMENEZ, S. (2019). Overview of tass 2019: One more further for the global spanish sentiment analysis corpus. *IberLEF@SEPLN*.

FAYYAD, U., and PIATETSKY-SHAPIRO, G., and SMYTH, P. (1996). From data mining to knowledge discovery: an overview. Advances in knowledge discovery and data mining. *American Association for Artificial Intelligence*, 1-34.

FERNÁNDEZ-CABANA, M., RÚAS-ARAÚJO, J. and ALVES-PÉREZ, M. T. (2014). 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)*, 169-184.

GREENE, D. and J. P. CROSS. (2015). Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis. *Proceedings of the ACM Web Science Conference. ACM*.

HOUVARDAS, J. and STAMATATOS, E. (2006). N-gram feature selection for authorship identification. *Artificial Intelligence: Methodology, Systems, and Applications. Springer*, 77–86.

KURGAN, L. and MUSILEK, P. (2006). A survey of knowledge discovery and data mining process models. *Knowledge Engineering Review., 1*(21), 1-24.

LAZARIDOU, K. and KRESTEL, R. (2016). Identifying political bias in news articles. *TCDL Bulletin, vol. 12*.

LIU, B. (2012). Sentiment Analysis and Opinion Mining. *Morgan & Claypool Publishers*.

MAIRESSE, F., WALKER, M. A., MEHL, M. R., and MOORE, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *JAIR, 30*, 457-500.

MALOUF, R. and MULLEN, T. (2007). Graph-based user classification for informal online political discourse. *Proceedings of the 1st Workshop on Information Credibility on the Web (WICOW), At Miyazaki, Japan.*

MERCADO, V. , VILLAGRA, A. and ERRECALDE, M. . (2019). Exploratory analysis of a new corpus for political alignment identification of argentinian journalists. *Actas del XXV Congreso Argentino de Ciencias de la Computacion (CACIC 2019)*, 507–516.

PEERSMAN, C., DAELEMANS, W., and VAN VAERENBERGH, L. (2011). Predicting age and gender in online social networks. *Proceedings of the 3rd international workshop on Search and mining user-generated contents. SMUC 11.*, 37-44.

PENNEBAKER, J. and LAY, T. (2002). Language use and personality during crises: analyses of mayor rudolph giuliani's press conferences. *Journal of Research in Personality., 36*, 271-282.

PENNEBAKER, J. W., BOOTH, R. J. and FRANCIS, M. E. (2001). Linguistic inquiry and word count (liwc). *[Software]*.

PENNEBAKERM J. W. , BOYD, R. L., JORDAN, K., and BLACKBURN, K. (2015). The development and psychometric properties of LIWC2015. *University of Texas at Austin, Austin, TX*.

PREOȚIUC-PIETRO, D., LIU, Y, HOPKINS, D., and UNGAR, L. (2017). Beyond binary labels: political ideology prediction of twitter users. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics., 1. Long Papers*, 729–740.

RÚAS-ARAÚJO, J., ALVES-PÉREZ, M., and FERNÁNDEZ-CABANA,M. (2017). Comunicación, lenguaje y política: Análisis de los discursos institucionales del presidente de ecuador, rafael correa (2007-2015), con la herramienta liwc. *Razón y Palabra, 20 (4-95)*, 591-607.

SCHLER, J., KOPPEL, M., ARGAMON, S., and PENNEBAKER, J. (2006). Effects of age and gender on blogging. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 199-205.

SIEVERT, C. and SHIRLEY, K.E. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces.*, 63-70.

SLATCHER, R. B., CHUNG, C. K., PENNEBAKER, J. W. and STONE, L. D. (2007). Winning words: individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality, 41*, 63-75.

STAMATATOS, E. (2009). A survey of modern authorship atrribution methods. *Journal of the American Society for Information science and technology., 60(3)*, 538-556.

STAMATATOS, E. (2011). Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology, 62(12)*, 2512–2527.

TUMASJAN, A., SPENGER, T. O., SANDNER, P. G. and WELPE, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI conference on Weblogs and Social Media (ICWSM)*, 178-185.