UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGIA

Modelo para la minería de textos en el sector periodístico

MÁSTER EN LETRAS DIGITALES

TRABAJO DE FIN DE MÁSTER

Curso 2019-2020

AUTORA

Lleó Pérez-Abadín, Julia

TUTOR

Riesco Rodríguez, Adrián

Departamento de Sistemas Informáticos y Computación

Madrid, septiembre 2020

Calificación: 9,5

A mis padres, por todos los sacrificios que han hecho para permitirnos estudiar y convertirme en la persona que soy hoy.

Gracias.

Modelo para la minería de textos en el sector periodístico

Julia Lleó Pérez-Abadín

RESUMEN

El siguiente trabajo de fin de Máster presenta un modelo de análisis basado en tecnologías de la minería de textos y su aplicación en el estudio de un caso concreto dentro del sector periodístico: la evaluación de las modificaciones de audiencia que durante el mes de marzo provoca la situación del estado de alarma en España.

Con ello se consigue la automatización de exploraciones lingüísticas como la medición estadística de diferentes fenómenos, el modelado automático de tópicos y la extracción de sentimiento. Dicho modelo se presenta en forma de repositorio abierto de GitHub para futuras aplicaciones.

Palabras clave: minería de textos, PLN, periodismo digital, automatización, análisis de datos, análisis de sentimiento, modelado de tópicos.

ABSTRACT

The following Master's Thesis consists in the presentation of an analysis model based on text mining technologies and its application in the study of a specific case of study within the journalism sector: the evaluation of the audience modifications caused by the lockdown in Spain during March 2020.

Automatic linguistic explorations are achieved, such as the statistical measurement of different phenomena, the topic modeling automatization and an approach to sentiment analysis. This model is shared as an open GitHub repository for future applications.

Keywords: text mining, NLP, digital journalism, automatization, data analysis, sentiment analysis, topic modeling.

Índice general

Índice de figuras

Índice de ilustraciones

Introducción

El mes de marzo de 2020 y la expansión del virus Covid-19 en España serán recordados por los medios de comunicación digitales como un hito histórico, y es que, si el arma letal de este enemigo invisible fue su desconocimiento, la población buscó su defensa en la información desde el refugio del confinamiento.

El repentino y espectacular crecimiento del tráfico de audiencia en periódicos digitales alcanza un aumento del 77% en tan solo la primera semana de confinamiento[1]. Para entonces, ya era mundialmente conocida una nueva enfermedad causada por coronavirus, y desde que China advierte a la OMS sobre el brote de Covid-19 el 31 de diciembre de 2019, la inquietud y desconfianza en la información recibida no hace más que aumentar. Ante esta situación excepcional, y la sobrecarga informativa no verificada que comienza a circular por la Red, los medios tradicionales recuperan la confianza de sus lectores y consiguen sumar algunos nuevos.

Aunque con diferentes resultados, toda la prensa digital se enfrenta a un reto inhóspito: la urgencia por informar viene de la mano del ingente caudal de información que a escala global se filtra, haciendo que la cobertura dada a la situación en cada diario juegue un papel crucial en la selección de los lectores. Entre los principios éticos que comparten todos estos medios se encuentran la necesidad de separar la información y la opinión, acudir a fuentes fiables, verificar los hechos, no publicar contenido audiovisual inapropiado o respetar la intimidad de las personas sobre las que se informa.

De lo que no cabe duda es de que, dado el contexto, su conducta resulta crucial, o así lo entiende la Organización Mundial de la Salud cuando asegura que la falta de una comunicación adecuada conduce a la pérdida de confianza y reputación, impactos económicos y, en el peor de los casos, pérdida de vidas (World Health Organization, 2020).

En el siguiente trabajo proponemos un modelo de análisis de mercado aplicado al sector del periodismo mediante técnicas de minería de texto. Con él,

[1] Datos extraídos de la firma Comscore, empresa dedicada a la medición del comportamiento de audiencias digitales.

automatizaremos el estudio de algunos de los factores que influyen en la elección del medio por el que la población decide informarse, y sacaremos algunas conclusiones del incremento de audiencias y suscripciones generalizado durante el mes de marzo. El mismo modelo será publicado en forma de repositorio de código en la plataforma GitHub para futuras investigaciones.

Antes de comenzar, cabe recordar que a pesar del trimestre de bonanza que parece vivir la industria, el desplome publicitario, la precarización de su trabajo y la paradoja entre los récords de audiencia vividos y su falta de monetización no contrarresta el impacto que la Covid-19 ha tenido para todo el sector.

Estado de la cuestión

Ante la doble naturaleza de los elementos que participan en nuestra investigación, se ha optado por una división interna de contenidos en este apartado que nos ayude a contextualizar: una vertiente de tipo teórica, delimitada por el estudio del periodismo digital nacional hasta su repercusión en nuestros días, y otra de tipo analítica, que atañe al desarrollo de técnicas de procesamiento automatizado de textos y contextualización del campo de estudio.

2.1. Periodismo digital en España

En los siguientes puntos contextualizamos los pasos previos al asentamiento de las TIC en nuestro país, el desarrollo de un nuevo tipo de periodismo: el periodismo digital, y una breve mención a la cobertura que anteriores crisis sanitarias han tenido en los medios globales y locales.

2.1.1. Las ciencias de la comunicación en España

El interés por el análisis de carácter científico de la producción periodística en nuestro país se remonta a los últimos tiempos del régimen franquista, impulsado por la institucionalización universitaria de este tipo de estudios. Según Jones, la producción española se sumía antes de este cambio en "una debilidad teórica evidente, una censura político-ideológica explícita y una primacía por parte del «aparato del Estado»" (Jones, 1997, pág. 104).

Nos diferenciábamos entonces del avance proferido por el resto de los países del occidente europeo, en los que la investigación en comunicación irrumpe tras la derrota de las facciones fascistas de la Segunda Guerra Mundial y goza de reconocimiento académico ya en los años cincuenta[2]. En un nuevo contexto institucional en nuestro país,

[2] En el monográfico publicado por la revista *Anàlisi. Quaderns de comunicació i cultura*, con el título de «La recerca europea en comunicació social», podemos encontrar trabajos sobre las exposiciones de

la autonomía intelectual y organizativa permite que la *doctrina española de la información* de inspiración fascista (de Moragas i Spà, 1981, págs. 224-225) comience a ver su fin.

Con la llegada de la democracia termina una etapa de autarquía y cerrazón, y España comienza a progresar en todos los órdenes: político, social, económico y, también, comunicativo. Dentro de este último nivel, se comienza a ver esta apertura con la creación de grupos multimedia (Prisa, Zeta o Godó); la desregulación de la televisión con su consecuente política de competencia entre emisoras (estatales y ahora también regionales); o el espectacular crecimiento de la industria publicitaria y relaciones públicas.

Pero al tiempo que el país se somete a una compleja revisión de valores, en el mundo de finales de los ochenta comienza la llamada carrera por la explosión de la comunicación (Breton & Proulx, 1989), y se consolida la identidad de la comunicación como disciplina a nivel internacional.

La explosión del negocio de la comunicación demanda la formación de profesionales de la comunicación, lo que genera una mayor profesionalización en las facultades de Ciencias de la Información de entonces. En menos de quince años se produce un impresionante incremento en la oferta de estudios de la Comunicación, que quintuplica su presencia en los centros educativos y crea nuevas titulaciones como la propia de Periodismo (Rodríguez Serrano & Gil Soldevilla, 2018).

En resumen, esta mirada al pasado nos sirve para ver cómo el camino hacia el asentamiento de los medios de comunicación, que en otros países se desarrolló por un avance paulatino hasta su consolidación, en nuestro país ha sido: forzado (por las presiones de mercado), precipitado (en apenas quince años se trata de dar solución al volumen de demanda), importado (el impulso por cortar con los tiempos de dictadura propició una abrupta internacionalización y permitió una forzada acogida del mensaje globalizado sin apenas filtro) y, en general, sin tiempo para una correcta maduración. A todo esto, se debe sumar un hito que cambia irrevocablemente nuestra historia y la

antecedentes históricos de la investigación comunicativa actual en países europeos entre los que se incluye España (cfr. Parés i Maicas, 1997).

historia de la comunicación: la popularización de las tecnologías de la información y la comunicación (TIC).

2.1.2 Periodismo en línea

Una de las consecuencias de la avenencia entre informática y comunicación es el surgimiento de un nuevo modelo de periodismo: uno que se vale, para su producto final, de la distribución digital. En nuestro estudio optaremos por el término "periodismo digital" para referirnos a esta nueva propuesta que tanta polémica despierta en torno a su nomenclatura, poniendo en valor que, aunque todos los trabajos acaben siendo procesados por código binario, no significa que todas las rutinas periodísticas clásicas hayan evolucionado para el nuevo medio.

Recordemos que para hacer las TIC posibles se ha debido asentar en nuestras vidas el uso de Internet, red de redes y autopista de la información (Wolton, 2000), después de que la Agencia de Proyectos de Investigación Avanzada (ARPA) realizara la primera conexión entre computadores de núcleos diferentes en el año 1969, y Berners-Lee definiera la Red en 1990, referida habitualmente como sinécdoque de Internet.

Abreu Sojo (Sojo, 2003) destaca dos tendencias principales de este tipo de periodismo: una claramente negativa, como ha sido la orientación sensacionalista caracterizada por la constante búsqueda de información alarmista sin censura, y otra más positiva, que utiliza el nuevo medio para dar cabida a voces que no han tenido una representación directa en el mundo del periodismo *analógico*. Entre los mejores ejemplos de esta última vía, podemos recordar cómo *El Periódico de Catalunya* ofreció una cobertura más profunda que sus competidores durante la epidemia del virus ébola tras contactar con científicos especialistas en el tema a través de foros de discusión de Internet.

El cambio de paradigma que la integración de nuevos puntos de vista supone, hace que la función clásica de «agenda-setting», teoría que postula que los medios de comunicación tienen una gran influencia en determinar qué posee interés informativo y cuánto espacio e importancia se les da, sea cuestionada. El nuevo modelo de

producción avanza para Harvey (Harvey, 1990) de un modelo fordista (de división del trabajo y producción en serie) a uno basado en la flexibilidad y la elaboración simbólica como fuerza productiva. Con ello las instituciones de cualquier tipo, que solían beneficiarse del control de flujos informativos, deben ahora evolucionar hacia un modelo de *accountability*[3]. Y es que como lo define Castells (Castells, 1996), el determinismo tecnológico debe ser interpretado como un falso problema, ya que la tecnología es ya la sociedad: esta herramienta no determina por sí misma el cambio histórico o la evolución social, pero sí ofrece la capacidad a la sociedad de transformarse.

En "Mapa de los cibermedios de España en 2018: análisis cuantitativo" (Salaverría Aliaga, del Pinar Martínez-Costa, & Breiner, 2018), presentado en el XXIV Congreso de la Sociedad Española de Periodística, se identificaban ya un total de 3.431 cibermedios en marzo de 2018, de los que 3.065 se consideran activos.

De este número, la investigación (ver figura 1) resalta que un tercio de las cifras lo componen medios nativos digitales, es decir, nacidos en la propia red y que no encuentran su correlato en la impresión física.

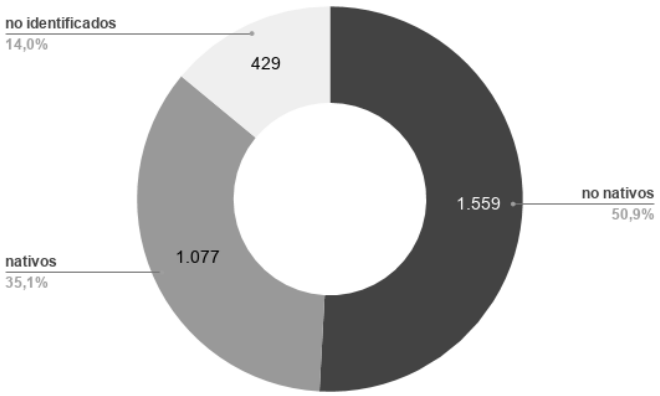

Figura 1. Cibermedios activos en España en 2018, según su origen (N=3.065) en (Salaverría Aliaga, del Pinar Martínez-Costa, & Breiner, 2018)

El proyecto también indaga en las temáticas que cubren estos medios, donde se observa una aventajada mayoría de cobertura de información general (el 68,5% de las publicaciones). Del 31,5% restante, que llamaríamos medios especializados, se revela

[3] Concepto ético que refiere la necesidad de transparencia y asunción de responsabilidad por parte de organizaciones públicas o privadas.

cómo el apartado Salud no se encuentra ni entre los 10 temas específicos más compartidos.

2.1.3 Crisis sanitaria

La última alerta de pandemia declarada por la Organización Mundial de la Salud tuvo lugar en el año 2009 debido a un brote de gripe A H1N1. El nuevo virus, nunca visto en personales o animales, fue identificado en Estados Unidos propagándose rápidamente por el mundo debido al contexto de globalización que impera en nuestra sociedad. En España, la crisis más reciente fue en el año 2014 después de que una profesional sanitaria se contagiara con un brote aislado del virus del Ebola (EVE).

En total, en los últimos años se cuentan hasta 11 epidemias o pandemias de importancia a nivel mundial, de entre las cuales 9 han sido causadas por virus (Esparza, 2016). Dos de los virus que se identifican como focos, SARS y MERS, pertenecen a la familia de los coronavirus.

La OMS (Organización Mundial de la Salud) sentencia en su boletín *Sixth Futures Forum on Crisis Communication* que salud, crisis y comunicación están íntimamente relacionadas. Todas las crisis de salud son también crisis de comunicación (World Health Organization, 2004). En otra comunicación más reciente de la OMS, se especifica además cómo esta comunicación debe afrontarse, en situaciones de crisis, con información puntual y correcta, empatía, esperanza y confianza pública en las autoridades y políticas efectivas (World Health Organization, 2013).

Si hacemos un recorrido por la cobertura que crisis sanitarias anteriores han recibido en los medios descubrimos que, en situaciones de emergencia, como la provocada por la gripe A, evidenciaban un alarmismo elevado. Señalan Rossmann, Meyer y Schulz que, en ese momento, tanto la prensa de calidad como la considerada de tipo sensacionalista exageraron la información del virus H1N1, enfatizando los

riesgos, presentando la información de manera dramatizada y hasta amplificando la amenaza (Rossmann, Meyer, & Schulz, 2018). En la prensa española también se evidenció este alarmismo, especialmente en los primeros meses de su llegada a España, siendo comparada incluso con crisis sanitarias anteriores como la pandemia de gripe de 1918. Este alarmismo vino, además, acompañado con cierto sensacionalismo en el momento de difundir datos personales y clínicos de algunos afectados de la enfermedad, vulnerando las normas de confidencialidad del historial clínico de los pacientes y su derecho a la intimidad (Costa-Sánchez & López-García, 2020).

En su análisis de la cobertura periodística del ébola, Monjas-Eleta y Gil-Torres también señalan el alarmismo y carácter sensacionalista como elemento principal del tratamiento informativo, basando su estudio en la imagen y en la personalización de la información, búsqueda de fuentes alternativas y léxico alarmista[4] (Monjas-Eleta & Gil-Torres, 2017).

2.2 Procesamiento del lenguaje natural

Dado que nuestro análisis textual utiliza técnicas del área del Procesamiento del Lenguaje Natural (PLN), más concretamente, aplicadas a la minería de textos, se vuelve preciso definir el estado de la cuestión de esta rama de la inteligencia artificial.

2.2.1. ¿Qué es el Procesamiento del Lenguaje Natural?

Tradicionalmente se data el año 1950, con la publicación de *Computing machinery and intelligence* (Turing, 1950) y su test de Turing, el momento en que se afianzaría una nueva subdisciplina computacional llamada Procesamiento del Lenguaje Natural. Desde ese momento, son muchas las definiciones y aplicaciones que de esta rama de convergencia entre lingüística e ingeniería se han explorado.

[4] El mayor grado de alarmismo se alcanza, en esta situación, con la utilización de imágenes de la afectada, su marido y su perro por parte de todos los diarios nacionales (Gou-Núñez, 2017).

G. Chowdhury, uno de los teóricos que se ha aventurado a definirlo, nos dice que: “Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things.” (Chowdhury, 2003, pág. 51). Bird, Klein y Loper amplían esta definición para mostrar las dos formas de entender el Procesamiento del Lenguaje Natural, una basada en modelos abstractos matemáticos y otra más cercana a la lingüística y aplicaciones del lenguaje:

> At one extreme, it could be as simple as counting word frequencies to compare different writing styles. At the other extreme, NLP involves “understanding” complete human utterances, at least to the extent of being able to give useful responses to them (Bird, Loper, & Klein, 2009, pág. ix).

Lo que constatan estas definiciones es el punto de enlace que debe existir entre disciplinas como la informática, las ciencias de la información, la lingüística, las matemáticas, la inteligencia artificial, robótica y psicología para poder afrontar la tarea de comprender los procesos de comunicación humana y poder replicar estos en una máquina. La justificación de que estos campos trabajen conjuntamente recae en el hecho de que el lenguaje natural se distingue de otros lenguajes artificiales por su riqueza (tanto en vocabulario como en construcciones), flexibilidad (encontramos múltiples excepciones a casi cualquier regla), ambigüedad (la ambivalencia de significado por contexto se expresa tanto en palabras como en frases), indeterminación (consintiendo referencias y elipsis) y multiplicidad de interpretaciones del sentido final del texto (Verdejo Maillo, 1994, pág. 5).

Las principales aplicaciones que ha tenido el PLN desde sus orígenes han sido, según lo recoge Llisterri (2019): la traducción automática, la verificación y corrección automática de textos y el tratamiento de la información (recuperación de información, resumen automático, extracción y búsqueda automática, etc.).

Dentro de lo que Llisterri considera tratamiento de la información, Liddy (Liddy, 1998) y Feldman (Feldman, 1999) distinguen siete niveles independientes que operan conjuntamente en la extracción del significado:

- Nivel fonético y fonológico, que se ocupa de la pronunciación y su variación.
- Nivel morfológico, centrado en partículas mínimas de significado dentro de una palabra, llamadas morfemas.
- Nivel léxico, que estudia el lexicón mental y las funciones sintácticas en que opera este.
- Nivel sintáctico, encargado de la gramática y la organización oracional.
- Nivel semántico, que aborda el significado de expresiones lingüísticas.
- Nivel de discurso, orientado al análisis del discurso, es decir el estudio del discurso hablado y escrito como forma de uso de la lengua.
- Nivel pragmático, el más problemático y epicentro de discusiones en PLN, enfocado al modo en que un contexto externo puede interferir en la interpretación de significado interno de un texto.

2.2.2 Minería de textos

La minería de textos es una de las ramas de investigación del procesamiento de textos más demandada en los últimos años, y trata de ser un término paraguas situado dentro del campo de las ciencias de la computación y estadística.

Podemos definir a la minería de textos como el proceso de analizar grandes cantidades de datos textuales a través de los avances que el campo del Procesamiento del Lenguaje Natural permite para el hallazgo de patrones en grandes volúmenes de textos. Es decir, su objetivo es el descubrimiento y extracción de algún aspecto pertinente dentro de una colección de datos semiestructurados o desestructurados. Tal y como muestra la figura 2, se trata de un campo multidisciplinario que utiliza técnicas de la recuperación de información, minería de datos, aprendizaje automático, estadística, etc.

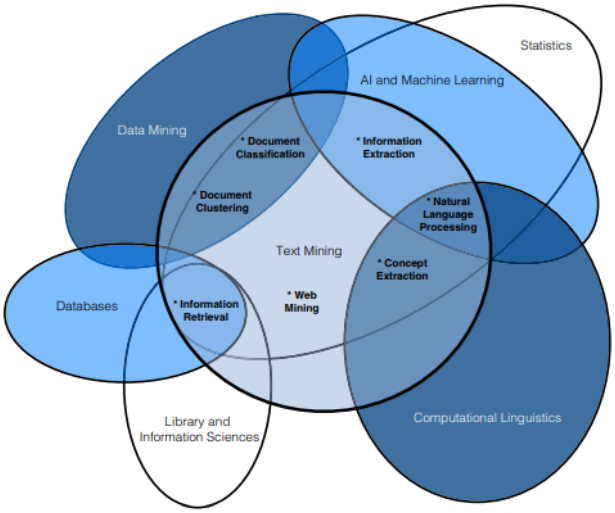

Figura 2. Diagrama de Venn extraído de (G. Miner, 2012)

El procedimiento básico que se esconde tras cada tarea de minería de textos es la computación del lenguaje, de forma que los algoritmos matemáticos trabajan junto a la lingüística permitiendo definir modelos estadísticos para el análisis de contenido.

En las tareas de minería de textos las principales fuentes de recolección de datos suelen ser los blogs, páginas especializadas en reseñas o sitios de *microblogging* como Twitter, ya que se precisa recabar una gran colección de materiales de texto. Este último constituye la fuente principal de las actuales investigaciones, debido a la enorme cantidad de información semi-estructurada en formato *JSON* recuperable tanto en flujo directo como por el archivo histórico de la red social.

Por su parte, gran parte de la computación científica requerida para la minería de datos a gran escala se implementa actualmente en archivos ejecutables desarrollados en Java, R o Python. Este último es el más utilizado, ya que su sintaxis no tipada permite gran flexibilidad. Además, gracias a la gran popularidad que ha ganado este lenguaje en los últimos años[5], que lo convierten hoy en día en un estándar dentro de muchas comunidades científicas, se han desarrollado incontables bibliotecas que facilitan y maximizan el alcance de la minería de datos, entre las que destacan Pandas, NumPy, Matplotlib y SciPy.

[5] *Data is Beautiful* ha realizado en el anterior año una animación que muestra la evolución de la popularidad de los principales lenguajes de programación entre los años 1965-2019 según encuestados de Estados Unidos: https://www.youtube.com/watch?v=Og847HVwRSI . (Data is Beautiful, 2019).

A la hora de realizar un análisis de minería de textos, el flujo de trabajo habitual recorre los siguientes pasos, ilustrados en la figura 3:

- Colección de datos desestructurados de tipo texto.
- Operaciones de preprocesamiento y limpieza para la eliminación de anomalías. Este proceso nos permite asegurarnos de que se está capturando correctamente la esencia del texto.
- Procesamiento de los datos limpios.
- Extracción de información relevante para su análisis.
- Interpretación de los datos obtenidos.

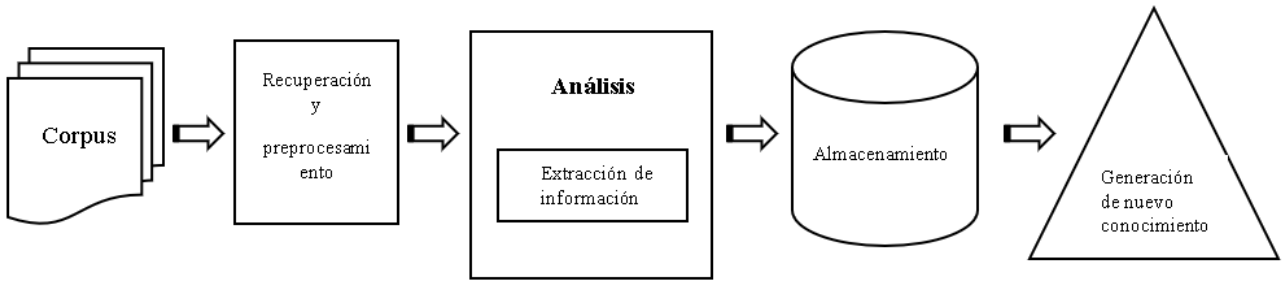

Figura 3. Flujo de trabajo habitual en la minería de textos

Una de las técnicas que mayor popularidad goza dentro de la minería de textos es el análisis de sentimiento o *sentiment analysis.* El objetivo de esta técnica es analizar el texto de forma que se puedan deducir cargas emocionales dentro de él. Para conseguir esto, es preciso disponer de *lexicons* que procesen, reconozcan y evalúen dichos sentimientos (Leetaru, 2011) y determinen cargas que pueden ir desde lo negativo a lo positivo. En relación con este tema, es importante recordar que las técnicas de *opinion mining* y *sentiment analysis,* generalmente confundidas, están diseñadas como aplicaciones distintas (Kechaou, Ben-Ammar, & Alimi, 2013), orientándose la primera a la detección de polaridad y la segunda al reconocimiento de emociones. No obstante, por su gran similitud, suelen ser utilizadas como sinónimos (Cambria, Schuller, Liu, & Wang, 2013), y así lo haremos en este estudio.

Por último, el modelado de tópicos o *topic modeling* es otra de las aplicaciones en auge de la minería de texto. Esta técnica tiene como objetivo la extracción de tópicos automáticos a partir del análisis de un vasto corpus de texto de los que extrae

conocimiento sobre temas, relaciones temporales y patrones en la clasificación (Arora, y otros, 2013). El procedimiento habitual para realizar el modelado parte de esta extracción abstracta de temas, a los que más tarde pone nombre el investigador, para terminar por agrupar los documentos según su pertenencia. Internamente, la tarea selecciona las palabras más frecuentes del corpus (indicativo de su pertenencia a un cierto tema) y observa su presencia en los diferentes documentos. El modelo más simple de *topic modeling* es el Latent Dirichlet Allocation (LDA), que cuenta con patrones implícitos y conjuntos de términos que se pueden llamar temas (Blei, Topic modeling and digital Humanities, 2012).

En definitiva, como señalan Hearst (Hearst, 1999) y Kodratoff (Kodratoff, 1999), la minería de textos se puede definir como un conjunto de herramientas relacionadas con el análisis del lenguaje que nos permite crear nuevo conocimiento a través de la relación del contenido de varios textos.

Objetivos y motivaciones

Como se ha adelantado en la introducción, la prensa digital ha actuado como eslabón fundamental en una crisis sanitaria mundial que llega a España con la noticia del primer infectado en La Gomera el 31 de enero. La necesidad de explicar a la ciudadanía lo que está sucediendo, los riesgos de la nueva enfermedad, y la necesidad de hacerla partícipe en la solución, convierten a todos los sistemas de comunicación, y en especial el periodismo, en un importante aliado para la gestión política, social, institucional y sanitaria. Como lo resume Fernando Vega, "la prensa hoy reafirma su autoridad como plataforma confiable de información y opinión. En tiempos de Coronavirus, Fake News en WhatsApp, cuarentenas e incertidumbre, los medios golpean la mesa para recuperar ese terreno aparentemente perdido" (Vega, 2020).

El objetivo de esta investigación pasa por adentrarse en la cobertura que los diez periódicos mejor posicionados tras el incremento de tráfico durante la crisis sanitaria. Concretamente, enfocaremos nuestro estudio en el mes de marzo, por ser este un punto de inflexión en el desconcierto de la ciudadanía. Recordemos que, en apenas treinta y un días, se suceden comunicados no verificados de la situación en Italia por WhatsApp, se celebra una polémica manifestación por el Día de la Mujer, se extiende un primer bulo que amenaza con cerrar la capital española, y a partir del día 14 se declara un estado de alarma que impone el confinamiento de los ciudadanos.

Si bien las estadísticas del consumo de informativos digitales a final de este mes se presentan en la figura 4.

Figura 4. Fuente: José Manuel Rodríguez vía Twitter

Detrás se oculta un traslado de usuarios entre periódicos digitales y un aumento del 34% de usuarios únicos de diarios de carácter generalista (según datos de Comscore). Estas modificaciones de audiencia entre los meses de febrero y marzo también son calculadas por el medidor consensuado por el sector periodístico, como se muestra en la figura 5.

Figura 5. Fuente: José Manuel Rodríguez vía Twitter

De la comparación de estas dos mediciones podemos extraer conclusiones como las siguientes:

- El medio **eldiario.es** gana un 61,8% de usuarios entre el mes de febrero y marzo, encabezando una lista de variaciones que en segundo lugar presenta al periódico **OkDiario**, con una respectiva subida del 41%. Ambos periódicos se posicionan en

octavo y noveno lugar en la relación de diez periódicos más leídos en España durante el mes de marzo.

- Los diarios **El Mundo**, **El Periódico**, **El País** y **20 Minutos** se benefician de un crecimiento de entre el 30,1% y el 38,6% de sus usuarios, con cifras bastante parejas dada las circunstancias descritas de aumento general de lectores digitales.
- **El Confidencial** y **La Vanguardia** se posicionan como últimos en la lista de captación de nuevos lectores, con un aumento del 23,5% y 23,4% respectivamente. No obstante, en el posicionamiento general de audiencia del mes de marzo, La Vanguardia continúa liderando con 28,4 millones de usuarios; 1,1 millón más que el segundo medio más leído, El Mundo.

Ante el inusual comportamiento del incremento de lectores, surge la pregunta que guía nuestra investigación: qué explicación puede tener esta desnivelación entre los datos totales de posicionamiento de los diez periódicos digitales más influyentes y el rumbo que toma la adscripción de lectores en el mes de marzo.

Para ello, utilizaremos herramientas del campo del Procesamiento del Lenguaje Natural en el desarrollo de un modelo de minería de textos y exploración lingüística a través del lenguaje Python, que más tarde aplicaremos al corpus de noticias seleccionadas del mes de marzo de 2020. Con los resultados extraídos, esperamos poder sacar algunas conclusiones que expliquen este comportamiento.

En nuestro caso, no se desarrollarán nuevos algoritmos para el análisis, sino que se hará acopio de bibliotecas para el desarrollo de estudios de PLN en Python y métodos de Análisis de sentimiento y Modelización del tópicos. Se definirá, no obstante, un plan de acción que pase por la aplicación de fórmulas orientadas específicamente para nuestro análisis. Posteriormente, los datos obtenidos serán revisados para la extracción de conclusiones de forma manual.

Dicho análisis no resulta algo novedoso dentro de los estudios de comunicación, si bien la originalidad del trabajo consiste en la implementación de un sistema automatizado que permite enfrentar el análisis de contenido de una gran cantidad de datos. El mismo modelo se encuentra publicado en el repositorio de *GitHub*

https://github.com/JuliaLleoPA/Periodismo-con-TextMining con las debidas indicaciones para el provecho de cualquiera que quiera emprender una investigación propia.

Metodología

Para la definición del modelo de análisis, se ha desarrollado un programa implementado en lenguaje Python y lanzado con la aplicación web JupyterLab. Se ha escogido este entorno por su flexibilidad en el control de libretas o *notebooks* y sencillez en la codificación y pruebas.

Una vez seleccionados los medios, hemos procedido con el flujo de trabajo habitual para las tareas de análisis textual definida en la *figura 3. Flujo habitual de trabajo e la minería de textos* (ver figura 3). Los resultados por consola se han visualizado y extraído con las bibliotecas Plotly y Matplotlib de Python, así como con tablas Excel y otras herramientas visuales.

Las fases de ejecución del proyecto han sido:

I. Definición de los objetivos y metas para nuestro caso de estudio concreto. Para su definición, se llevó a cabo un proceso de investigación previa que atendiera tanto al interés de la elección del tema como el provecho que la investigación podía arrojar. De forma paralela, se indagó en el estado de la cuestión del análisis textual por técnicas de PLN (§ESTADO DE LA CUESTIÓN) y se realizaron pequeñas pruebas antes de escoger la opción Notebook de JupyterLab.
II. El segundo paso consistió en concretar los textos que conformarían nuestro corpus de estudio acorde a las preguntas que guían la investigación y recuperarlos para su análisis mediante un proceso de recuperación de información.
III. Una vez obtenido el corpus, comenzamos el proceso de desarrollo de las tareas de minería de texto, como son la limpieza, almacenamiento de los datos en un nuevo formato, creación de una matriz lingüística y diferentes exploraciones de estos datos que se han considerado pertinentes para dar respuesta a la investigación.
IV. Además de tareas de tipo estadístico dentro del análisis, se midió la polaridad de los distintos periódicos en el total de sus publicaciones, y se compuso un sistema

de modelado de tópicos que discriminara los temas de interés de la prensa a lo largo del mes de marzo.

V. Por último, se han examinado los datos obtenidos y extraído conclusiones que permitan caracterizar la cobertura de los diferentes medios y ayuden a explicar la elección de los lectores.

4.1 Elección del corpus

La elección del corpus es, dentro de cualquier programa de análisis, una tarea compleja y determinante que debe estar basada en una clara especificación de los objetivos de investigación. Las dos principales complicaciones de esta tarea se concentran en la elección del tamaño del corpus y su variedad.

Gracias a la implementación de soluciones del ámbito del PLN, cuestiones como los recursos de tiempo disponibles dejan de ser uno de los principales escollos en la selección. En nuestro modelo de análisis, además, la tipología de este corpus deberá ceñirse a materiales de tipo textual disponibles como recurso web.

Ya que nuestra investigación está relacionada con el género periodístico, cabe tener en cuenta que los periódicos digitales más importantes de España suelen contar con un buscador integrado que recupere los artículos publicados en línea, aunque este contenido solo se pueda rastrear hasta el inicio de la andadura del medio en formato digital y en numerosos casos bajo pago de suscripción. Estas hemerotecas suponen un gran aliado tanto para la selección (por las búsquedas avanzadas que nos permiten precisar) como para la obtención del recurso URL de aquellas publicaciones que sean de nuestro interés.

Cabe tener en cuenta las modificaciones que el aspecto natural de las noticias en los medios tradicionales sufre en el nuevo entorno digital. Por ejemplo, la profesora Concha Edo (2003) resalta las alteraciones en la propia estructura de la noticia desde el formato convencional de titular-entradilla-cuerpo, en el que la entrada y el cuerpo se encontraban en una misma página, al esquema actual en que los medios web diluyen

las demarcaciones espaciales y la rivalidad por la captación de atención se convierte en el motor de cada publicación.

Las principales partes que configuran actualmente una noticia en internet, y que podríamos seleccionar como objetos de estudio son:

- El titular. En el nuevo entorno digital, este titular pierde en parte su objetivo de delimitador del tema y pasa a competir por despertar el interés en la noticia. Si este titular cumple su objetivo y resulta atrayente, podremos acceder a ella por otras secciones, buscadores, redes sociales o agregadores.
- La entradilla de la noticia. Tradicionalmente, esta entradilla presentaba un resumen de los sucesos importantes de la noticia, pero actualmente se considera una estrategia más dentro del juego del *clickbait*[6]. La clásica fórmula de las 5W con que se diseñaba este apartado se ve superada por una exigencia de brevedad por la que los periodistas se ven obligados "a utilizar menos palabras, a decidirse por un lenguaje más directo y a prescindir de algunos datos (…) que dan como resultado una presentación farragosa de difícil digestión" (Edo, 2003, pág. 95).
- El cuerpo informativo. El cuerpo en prensa digital debe ser más corto que en prensa escrita, recomendándose no superar los cinco párrafos y no más de dos niveles hipertextuales. También en este apartado Edo ofrece consejos de carácter general como "utilizar frases cortas, utilizar habitualmente un vocabulario sencillo, pero correcto y con los términos adecuados, recurrir a verbos en forma activa o no repetir tópicos desgastados" (Edo, 2003), además de destacar que aquellos lectores que han llegado a este nivel de la noticia, es porque están interesados en el tema y confían en el medio como fuente de información.
- Otros elementos de la noticia. En esta sección se incluye cualquier soporte multimedia o hipermedia que pueda servir al periodista para construir la noticia con éxito. Ya que no se trata de contenido textual, el análisis de estos elementos nunca debe ser el objetivo de un estudio de minería de texto.

[6] El neologismo *clickbait* describe los contenidos en Internet que utilizan triquiñuelas sensacionalistas y engañosas en sus titulares y miniaturas para atraer la mayor proporción de *clics* posibles.

4.2 Obtención del corpus

Las funciones descritas en este apartado se corresponden con el archivo 1. Conseguimos los datos.ipynb del repositorio.

El término *Web scraping* o, como se suele traducir al español, raspado de datos web, es una técnica de recuperación de información web que consiste en la automatización de la navegación y extracción de información. Al software programado para rastrear se le llama *spider* o *crawler*, pero los resultados más flexibles y certeros se pueden conseguir mediante su implementación en un lenguaje de programación. En nuestro caso, al trabajar con Python y poder hacer uso de las facilidades que sus múltiples bibliotecas ofrecen, recurrimos a la biblioteca HTTP de Python request y BeautifulSoup para definir nuestra función automatizada.

La forma de proceder en el momento de *scrapear* funciona como una petición a servidor del código de una URL siguiendo e l protocolo HTTP habitual. Lo que vamos a hacer con Python es automatizar esta petición y recibir el código fuente de nuestra página y, gracias a las funciones BeautifulSoup, analizar el etiquetado HTML para extraer las partes que nos interesen. Aunque en nuestro caso de estudio (§CASO DE ESTUDIO) veremos las problemáticas encontradas, la estructura de nuestra función se muestra en la ilustración 2. En esta función llamada "`url_periodico`" el parámetro seleccionado será el conjunto de url que definamos en una clase diferente. Un ejemplo de cómo definir esta clase se encuentra en Ilustración 1,

```
In [ ]: urls = ['https://www.elmundo.es/ciencia-y-salud/salud/2020/03/10/5e677bfdfdddff2b9e8b4577.html']
```

Ilustración 1. Ejemplo de la clase url

Dentro de la función se siguen algunos pasos, como se explica dentro de las anotaciones de la ilustración 2:

```
In [ ]: # Importamos nuestras dos librerías previamente instaladas
        import requests
        from bs4 import BeautifulSoup
```

```
In [ ]: # Función general para hacer web scraping

        def url_periodico(url):

            '''Recupera el texto de la URL facilitada por una petición HTTP'''
            pagina = requests.get(url).text

            '''BeautifulSoup acepta dos argumentos: el marcado actual, y el parser que se quiere usar.
            En nuestro caso escogemos el parser "lxml", ya que también funciona para versiones antiguas de Python y es muy rápido.
            Para ello, es preciso instalarlo antes en nuestro equipo'''
            soup = BeautifulSoup(pagina, "lxml")

            '''Esta operación recupera el texto dentro de las etiquetas <p> de clase "article-text" (recordemos que la clase se her
            text = [p.text for p in soup.find_all('p')]

            print(text)

            return text
```

Ilustración 2. Extraída del repositorio de GitHub: función de web scraping

El resultado de aplicar la fórmula `url_periodico` a la lista de `urls` definidas (en este caso, solo una muestra del periódico El Mundo), se muestra en la Figura 6:

```
['Portada', 'Los últimos trabajos publicados apuntan que el coronaviru
s SARS-CoV-2 tiene mayor supervivencia de la que se estimaba, que pued
e contagiarse a más distancia y no sólo por vía aérea', 'El coronaviru
s puede sobrevivir en el aire durante al menos 30 minutos y difundirse
hasta 4,5 metros, es decir, más lejos de la "distancia de seguridad" r
ecomendada por las autoridades sanitarias de todo el mundo, según un e
studio publicado por científicos chinos. Además, otro trabajo apunta q
ue la vía aérea no sería la única vía de transmisión. Las heces podría
n ser otro vehículo de contagio. ', 'La supervivencia del coronavirus
fuera del organismo es una de las cuestiones que más se están estudian
do. El hecho de que permanezca "durante días" en las superficies donde
caen las gotas respiratorias infectadas aumenta el riesgo de contagio
simplemente por tocar y luego llevarse la mano a la cara. La Organizac
ión Mundial de la Salud (OMS) señala en su web que a tenor de los estu
dios disponibles, este coronavirus "puede persistir en las superficies
durante algunas horas o hasta varios días". ', 'La variación depende d
e condiciones como el tipo de superficies, la temperatura o la humedad
del ambiente. Según el nuevo estudio, por ejemplo, a alrededor de 37 g
rados centígrados puede sobrevivir de dos a tres días en materiales co
mo vidrio, tela, metal, plástico o papel.', "Un trabajo reciente publi
cado en 'Journal of Hospital Infection' apuntaba hasta nueve días de s
upervivencia fuera del organismo. Se trataba de una revisión de la lit
eratura científica disponible sobre distintos coronavirus realizada po
r expertos de la Universidad de Leibniz (Alemania). Al parecer, esta c
apacidad de resistencia la mostraba el SARS o el MERS en algunas super
ficies como la cerámica, el caucho, el metal, el cristal o el plástico
. La pregunta era si estos resultados se podían extrapolar al SARS-CoV
-2. ", 'Los expertos consultados por este periódico recordaban que la
información es muy variada. Así como algunos estudios indicaban una pe
rmanencia de 48 horas, otros apuntaban muchos días. No obstante, lo qu
e este informe remarcaba es que cabía la posibilidad de este coronavir
us sobreviviera más tiempo de lo que se pensaba en superficies no poro
sas. ', 'Y efectivamente, el nuevo estudio concluye que en materiales
como vidrio, tela, metal, plástico o papel puede sobrevivir tres días
si se encuentra a una temperatura de 37 grados. Ya en anteriores inves
tigaciones se había observado que las condiciones de mayor riesgo incl
uye una temperatura de menos de 30 grados, con humedades relativas sup
```

```
eriores al 50% y en superficies de plástico. ', 'En cuanto a la distan
cia de 4,5 metros, de ser cierto, desafía completamente al consejo de
las autoridades de salud de todo el mundo de que las personas deben pe
rmanecer separadas a una "distancia segura" de uno a dos metros, señal
a el periódico South China Morning Post.', 'A falta de estudiar más es
ta cuestión para comprender y conocer mejor el comportamiento del coro
navirus SARS-CoV-2 fuera del cuerpo humano,conviene recordar la import
ancia de lavarse las manos con frecuencia con agua con jabón al menos
durante 20 segundos y limpiar con productos estándar como la lejía, qu
e inactiva al virus en cinco minutos después del contacto. ', 'Otra de
las cuestiones que aún no están claras tiene que ver con las rutas de
transmisión, que no están del todo establecidas. La revista \'Journal
of the American Medical Association\' publicaba esta semana la experie
ncia de unos microbiólogos del Centro Nacional de Enfermedades Infecci
osas de Singapur que indica que la expansión ambiental del virus es am
plia y que parece realizarse no solo a través de las gotitas respirato
rias sino también por las heces. Los autores de ete análisis sugieren
que "el entorno es un medio potencial de transmisión y apoya la necesi
dad de una adherencia estricta a la higiene ambiental y de manos".', '
Los virólogos realizaron esta prueba en las habitaciones de tres pacie
ntes que estaban aislados por el coronavirus, a raíz de la comunicació
n de varios brotes de infección nosocomial comunicados desde China.',
'Ninguno de los pacientes tenía neumonía, aunque sí presentaban las ví
as respiratorias superiores afectadas. Dos de ellos tenían síntomas mo
derados que incluían tos y fiebre, mientras que otro presentaba una ca
ra más suave de la enfermedad, y prácticamente solo tenía tos.', 'Prec
isamente la habitación de este último paciente se analizó durante vari
os días antes de que se realizara la limpieza rutinaria diaria.', 'Los
investigadores hallaron que casi el 90% de todos los muebles y enseres
, incluyendo ventanas, suelo, lámparas y sillas, daban positivo para e
l virus. Además el 60% de la superficie de los inodoros también alberg
aban al patógeno.', 'En cambio, al analizar las habitaciones de otros
dos pacientes aislados también por el coronavirus después de haberse l
impiado, no encontraron presencia del patógeno, lo que indica, según e
scriben los autores, que "las medidas descontaminantes actuales son su
ficientes".', 'Conforme a los criterios de', 'El director de El Mundo
selecciona las noticias de mayor interés para ti.\n', 'Parece que la e
pidemia tiene posibilidades de empezar a remitir. Nosotros creemos que
España no va a tener, como mucho, más allá de algún caso diagnosticado
. Esperemos que no haya transmisión local. Si la hay, será transmisión
muy limitada y muy controlada." DON SIMON, 31 de ENERO']
```

Figura 6. Ejemplo del resultado de la fórmula de recuperación

Para poder conocer cuáles son y cómo referenciar el contenido de etiquetas HTML que nos interesan, es necesario haber inspeccionado antes la página.

En la función de recuperación mostrada en la Ilustración 2, hemos guardado en una lista las variables de tipo cadena de texto que se han recuperado por cada medio. A continuación se ha creado otra lista con los nombres de los periódicos, y para almacenar este raspado, hemos utilizado el módulo Pickle de Python, que nos permite representar nuestros objetos como cadenas de bytes, es decir, sin necesidad de realizar ninguna conversión especial. Estos archivos pickle nos permitirán, gracias a su capacidad de serialización, almacenar los resultados para próximos programas Python.

```
In [ ]: # Importamos el módulo previamente instalado
        import pickle

In [ ]: #Guardamos los archivos

        '''Fíjate en que a cada variable dentro de la lista "periodico" le corresponda su nombre'''

        for i, c in enumerate(nombrePeriodicos):

            '''Previamente hemos creado una carpeta llamada "corpus" donde guardaremos nuestros objetos serializados'''
            '''El indicador wb nos sirve para crear el archivo de escritura en modo binario'''

            with open("corpus/" + c + ".txt", "wb") as file:
                pickle.dump(periodico[i], file)
```

Ilustración 3. Extraída del repositorio de GitHub: cómo guardar archivos pickle

Más adelante, cuando queramos recurrir a estos archivos, usaremos un método similar (`pickle.load`), y el modo de apertura será "`rb`", es decir, solo lectura en modo binario.

4.3 Limpieza del corpus

Las funciones descritas en este apartado se corresponden con el archivo 2. Limpieza.ipynb del repositorio.

Antes de organizar nuestros datos en los estándares de texto que usaremos para realizar nuestros análisis (un corpus limpio y una matriz de términos), el texto recuperado deberá pasar por algunos procesos de limpieza que nos aseguren estar trabajando con texto pertinente para nuestros fines.

Para poder empezar a pensar en métodos de limpieza, siempre es preciso realizar una minuciosa revisión del texto obtenido para definir los problemas a los que nos enfrentaremos en esta etapa de limpieza.

Definir los objetivos y límites de la limpieza dependerá en cualquier caso de los propósitos que tenga nuestro corpus. Según para lo que se vaya a utilizar, nos interesará ser más minuciosos en la limpieza, recuperando un resultado de menor tamaño, pero de mayor relevancia (común en los análisis de contenido), o preferiremos conservar algunos de los rasgos distintivos del texto (técnica que utiliza el análisis estilístico, entre otros). En nuestro caso, hemos optado por hacer un barrido profundo que únicamente conserve piezas clave que usaremos para realizar análisis del texto, de sentimiento y

modelado de tópicos. Algunos de los problemas de depuración más comunes en cualquier limpieza de texto son los siguientes:

- Marcas de etiquetado que se hayan podido quedar.
- Errores tipográficos u ortográficos.
- Signos de puntuación.
- Envoltorios artificiales de espaciado.
- Presencia de números que requieran manipulación.
- Marcadores de sección que nos interese eliminar.

El enfoque de limpieza escogido ha sido el uso de la biblioteca de expresiones regulares para Python (re), y algunos de los métodos para asegurar la limpieza de nuestro corpus son los siguientes:

```
In [ ]: def primeraLimpieza(text):
            '''Primero vamos a separar algunas palabras se han unido a otras en el proceso de raspado'''
            text = re.sub(r'([a-z])([A-Z])', r'\1 \2', text)

            '''También es común que se conserven las marcas del cambio de línea'''
            text = re.sub('\n', ' ', text)

            '''Y de tabulación'''
            text = re.sub('\t', ' ', text)

            '''Eliminamos la puntuación con la constante de Python, que se encarga de: !"#$%&amp;'()*+,-./:;&lt;=&gt;?@[\]^_`{|}~'''
            text = re.sub('[%s]' % re.escape(string.punctuation), ' ', text)

            '''Añadimos alguna puntuación que no recoge la constante'''
            text = re.sub('[’‘“”—«»¡¿▪]', ' ', text)

            '''Por último, eliminamos el exceso de espaciado que hemos estado generando en nuestra limpieza'''
             text = " ".join(text.split())

            return text
```

Ilustración 4. Extraída del repositorio de GitHub: función de limpieza con algunos recursos generales

Este método re.sub retoma la cadena modificando el primer patrón definido (en caso de encontrarlo) por un espacio. Al final de la función nos ocupamos de estos espacios de más generados.

Si además esperamos conseguir una matriz lingüística precisa, deberemos aplicar una segunda fase de limpieza como la siguiente:

```
In [ ]: def segundaLimpieza(text):

            '''Transformamos el texto a minúscula'''
            text = text.lower()

            '''También palabras que contengan números en su interior'''
            text = re.sub('\w*\d\w*', ' ', text)

            '''Por último, eliminamos el exceso de espaciado que hemos estado generando en nuestra limpieza'''
             text = " ".join(text.split())

            return text
```

Ilustración 5. Extraída del repositorio de GitHub: función de limpieza con tareas más específicas

4.4 Creación de corpus y matriz

Las funciones descritas en este apartado se corresponden con el archivo 3. Creación del corpus y la matriz.ipynb del repositorio.

Después de haber limpiado el texto, necesitamos preparar el formato de nuestros datos para poder utilizarlos más tarde. En este caso, prepararemos y almacenaremos los datos en forma de corpus y matriz.

Cuando hablamos de corpus nos referimos a la definición que el término adopta dentro de la lingüística de corpus, es decir, una colección de textos en soporte informático. Tradicionalmente, en el campo de la lingüística se ha utilizado el término en un sentido más amplio para referir cualquier conjunto de materiales escritos u hablados compilados según criterios lingüísticos definidos para su investigación. En la actualidad, no obstante, los procesos computacionales y tecnológicos hacen que este concepto pase a estar altamente ligado a la informática.

En este caso, el corpus recopilado para su análisis posterior son el conjunto de noticias seleccionadas por cada periódico, debidamente limpiadas y almacenadas como cadena de texto. Como nos interesa realizar un análisis comparativo, manteniendo el total de noticias en un mismo archivo, se ha previsto como solución al conflicto de cómo diferenciar el periódico de pertenencia mediante un marco de datos del paquete Pandas. Este paquete proporciona estructuras de datos tabulares con columnas de tipo heterogéneo con la posibilidad de incorporar etiquetas en cada columna y fila. El tipo de objeto escogido para nuestra representación es una tabla en dos dimensiones DataFrame, basada, como todos los objetos Pandas, en la extensión Numpy.

Para poder componer debidamente nuestro corpus, de forma que podamos controlar a qué periódico pertenece cada conjunto de noticias, el método definido se detalla en la ilustración 6:

```
In [ ]: #Importamos el paquete previamente instalado
        import pandas as pd

        '''Personalizamos la representación'''
        pd.set_option('max_colwidth',200)

        '''Habremos guardado nuestros datos en un diccionario que contenga: Clave (nombre del periódico) - Valor(cadena de texto limpio
        corpus_df = pd.DataFrame.from_dict(diccionario).transpose()

        '''Llamamos a la columna Noticias'''
        corpus_df.columns = ['Noticias']

        '''Con este método Pandas ordena las etiquetas con sus debidos ejes'''
        corpus_df = corpus_df.sort_index()

In [ ]: # Guardamos nuevamente nuestra serialización pickle
        corpus_df.to_pickle("corpus.pkl")
```

Ilustración 6. Extraída del repositorio de GitHub: almacenar los datos del corpus

Para nuestros futuros análisis también necesitaremos crear una bolsa de palabras que contabilice el número de apariciones de cada característica lingüística de los elementos o *token.*

Esta bolsa de palabras es un elemento común dentro del campo del PLN, y consiste en un modelo de representación del vocabulario utilizado mediante una matriz en la que cada columna represente un token y se contabilice el número de apariciones de este.

Para poder hacer esta conversión de texto a números, tendremos que hacer una vectorización. Este proceso seguirá los siguientes pasos:

- *Tokenización* del texto.
- Creación de un vocabulario.
- Codificación del documento.

Para nuestra investigación utilizaremos la clase CountVectorizer de la biblioteca de aprendizaje automático Scikit-learn. De este modo, se contará la frecuencia de tokens y se creará una matriz que devuelva un vector codificado con la longitud del vocabulario y un número entero con las veces que cada palabra ha aparecido en el documento.

Antes de proceder a crear la matriz, conviene eliminar aquellas palabras sin significado por sí solas, como artículos, pronombres, preposiciones e incluso algunos verbos modales o auxiliares, que modifican o acompañas a otras que sí lo tienen. Este tipo de términos se conoce, dentro del campo de la lingüística computacional, con el nombre de *stop words*, y para nuestro modelo hemos creado un conjunto de 395

palabras nulas que acompañan a los programas en el repositorio con el nombre de palabras_vacias.txt. Aunque según cada caso de estudio estas se deban aumentar según el foco de estudio, en la ilustración 7 se muestra un buen punto de partida.

```
In [ ]: #Importamos el módulo codecs, que nos ahorrará problemas con la codificación y decodificación de nuestro documento

        import codecs

        with codecs.open('palabras_vacias.txt', encoding='utf-8') as f:
            stopwords = f.read().splitlines()
```

```
In [ ]: # Utilizamos CountVectorizer para crear la matriz, excluyendo las palabras definidas como vacías

        from sklearn.feature_extraction.text import CountVectorizer

        '''Especificamos la exclusión de nuestra lista de palabras vacías'''
        cv = CountVectorizer(stop_words = stopwords)

        '''La función transform convertirá a un vector nuestra columna de texto'''
        matriz_cv = cv.fit_transform(corpus_df.Noticias)

        '''Expresamos su resultado de nuevo en una tabla DataFrame'''
        df_matriz = pd.DataFrame(matriz_cv.toarray(), columns=cv.get_feature_names())
        df_matriz.index = corpus_df.index
```

```
In [ ]: # De nuevo, guardamos para posterior uso

        df_matriz.to_pickle("df_matriz.pkl")
```

```
In [ ]: #También guardamos nuestro elemento CountVectorizer para posterior uso en la detección de temas automática

        pickle.dump(cv, open("cv.pkl", "wb"))
```

Ilustración 7. Extraída del repositorio de GitHub: creación de la matriz de términos

4.5 Análisis del texto

Las funciones descritas en este apartado se corresponden con el archivo 4. Análisis del texto.ipynb del repositorio.

Las operaciones con que trabajaremos computacionalmente con nuestra matriz vectorizada nos permitirán modelar abstracciones estadísticas de diferentes características lingüísticas encontradas en el texto.

Los objetivos que esperamos conseguir en esta fase serán la medición de palabras más frecuentes por periódico, que delaten los temas preferentes en sus informaciones; la riqueza léxica total por medio, que indirectamente revele la variedad de aspectos destacados en la selección específica de noticias; y, por último, la media de palabras por artículo, que nos permita sacar conclusiones sobre el nivel de cobertura concedida a la pandemia.

4.5.1. Obtención de las palabras más comunes

Este cálculo lo podemos realizar con la propia matriz de términos. Para conseguirlo, solo deberemos especificarle que, por cada medio representado, nos devuelva los términos con el número de repeticiones igual o superior al definido, tal y como se muestra en la ilustración 8.

```
In [ ]: # Leemos la matriz de términos
import pandas as pd

matriz = pd.read_pickle('df_matriz.pkl')
```

```
In [ ]: # Por ejemplo, seleccionemos las 10 palabras más comunes
'''Buscamos recoger tanto el nombre de la palabra como el número de repeticiones,
así que crearemos un diccionario con esa estructura'''
top10 = {}

'''Para simplicar el posterior bucle, alteraremos el orden normal de representación de
la matriz(cada columna corresponde a un término, y cada fila a un periódico). Con el método transpose conseguimos que
cada periódico componga una columna y las filas pasen a ser el recuento de apariciones de cada palabra'''
matriz = matriz.transpose()
'''Recorreremos cada periódico'''
for c in matriz.columns:
    '''La función .head nos devolverá el número especificado de elementos (10) con mayor valor'''
    top = matriz[c].sort_values(ascending=False).head(10)
    '''Ordenamos nuestro diccionario para que la clave sea el nombre del diccionario y el valor sea una lista
    con tuplas que contengan ('término', nº de repeticiones)'''
    top10[c]= list(zip(top.index, top.values))
```

Ilustración 8. Extraída del repositorio de GitHub: cuenta de las 10 palabras más comunes por medio

4.5.2. Riqueza léxica del medio

Para este objetivo, operaremos con la extensión para vectores y matrices Python Numpy, biblioteca que cuenta con funciones matemáticas de alto nivel y nos permite operar con la matriz de forma sencilla. Como hemos comentado anteriormente, Pandas y nuestros DataFrame están construidos sobre estas bibliotecas también.

Para ello, primero definimos una lista vacía que iremos rellenando con las apariciones. A continuación, escribimos un bucle `for` que corra las columnas de nuestra matriz por cada periódico, de forma que un contador hará la suma de veces en que el valor en las celdas de esas columnas (que corresponden al total de términos identificados en la matriz de todos los medios)[7] tienen al menos una ocurrencia, es decir, que el valor de la celda es diferente a 0. Por último, introducimos los resultados como elementos en la lista `lista_apariciones`.

[7] Un ejemplo de matriz lingüística se encuentra en la ilustración 19 del siguiente apartado.

```
In [ ]: # Creamos una lista con las palabras que tienen una aparción por lo menos

lista_apariciones = []
for periodico in matriz.columns:
    '''El [0] refiere el índice del DataFrame al que nos estamos refiriendo'''
    apariciones = matriz[periodico].to_numpy().nonzero()[0].size
    lista_apariciones.append(apariciones)
```

Ilustración 9. Extraída del repositorio de GitHub: medición de la riqueza léxica

4.5.3. Media de palabras por artículo

Accederemos al total de palabras de cada medio dividido entre el número de artículos recogidos. Para ello, medimos el total de palabras en cada periódico, recorriendo la matriz con la función de suma, de una forma similar a como lo hemos hecho anterior función, aunque en este caso utilizaremos el método de suma de Python. En la ilustración 10 podemos ver este resultado.

```
In [ ]: #Calculamos el total de palabras por periódico
total_palabras = []
for word in matriz.columns:
    total = sum(matriz[word])
    total_palabras.append(total)
```

Ilustración 10. Extraída del repositorio de GitHub: suma del total de palabras por medio

Habiendo definido manualmente una lista con elementos de tipo `int` (de tipo *integer*, número entero) del total de artículos por periódico recogidos, podemos pasar a calcular la media de palabras. Para ello, se han definido nuevas columnas de nuestro marco de datos DataFrame. Exactamente, se han definido tres nuevas columnas:

- Una que recoja el número de palabras utilizada por cada medio
- Otra que recoja el número de artículos por periódico
- Una última llamada 'Media' que contendrá la operación de cálculo de la media (es decir, la división del número de palabras por los artículos por periódico)

Se ha decidido operar por columnas del DataFrame ya que dos listas como nuestras variables `total_list` (`array` con valores `int` del total de palabras) y `total_articles` (`array` con valores `int` del total de artículos), no pueden presentarse como operandos en una división en Python. En la Ilustración 11 se muestra un ejemplo de cómo se procedería al cálculo:

```
In [ ]: data['Palabras usadas por cada periódico'] = total_list
        data['Artículos recogidos por periódico'] = total_articles
        data['Media'] = data['Palabras usadas por cada periódico'] / data['Artículos recogidos por periódico']
```

Ilustración 11. Extraída del repositorio de GitHub: media de palabras por periódico

4.6 Modelado automático de tópicos

Las funciones descritas en este apartado se corresponden con el archivo 5. Modelado automático de tópicos.ipynb del repositorio.

Cuando hablamos de modelado de tópicos nos referimos a la detección automática de posibles abstracciones de "temas" en una colección de documentos mediante un modelo estadístico. Existen diferentes técnicas para este proceso, siempre basadas en el aprendizaje no supervisado, como son *Correlated Topic Model* (Blei & Lafferty, 2007) o el *Biterm Topic Model* (Yan X., 2013). En nuestro caso, escogeremos el algoritmo LDA, ya que ha sido específicamente formulado para tipos de datos textuales y es, actualmente, la técnica más explotada en los análisis de Procesamiento de Lenguaje Natural.

En 2003 se presenta este nuevo modelo estadístico generativo para el modelado de documentos partiendo de una bolsa de palabras, con el nombre de *Latent Dirichlet Allocation* (LDA) (David M Blei, 2003). La ecuación que define este proceso estadístico quedaba como se indica en la figura 7:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn},\beta) \right) d\theta_d$$

Figura 7. Ecuación LDA extraída de (David M Blei, 2003)

Se trata, en suma, de un modelo probabilístico generativo de tópicos. El corpus que se proporciona se representará como una combinación aleatoria de tópicos latentes entre los que cada tópico estará caracterizado por una distribución de probabilidades sobre el conjunto fijo de palabras (Hernández & Navarro, 2015).

A la hora de trabajar con el algoritmo, recurrimos al modelo LDA de la biblioteca gensim. Para hacer uso de la implementación de este modelo en el paquete gensim,

debemos contar con una bolsa de palabras previa (en nuestro caso, la matriz). Para hacerlo necesitaríamos:

- Crear una matriz dispersa a partir de la nuestra.

```
In [ ]: # Vamos a hacer una matriz dispersa con nuestra matriz

cuenta_dispersas = scipy.sparse.csr_matrix(matriz)
```

Ilustración 12. Extraída del repositorio de GitHub: Ejemplo de la creación de una matriz dispersa para el modelado de tópicos

- Generar un corpus gensim propio con esa matriz.

```
In [ ]: #A continuación creamos el corpus gensim

corpus = matutils.Sparse2Corpus(cuenta_dispersas)
```

Ilustración 13. Extraída del repositorio de GitHub: Ejemplo de la creación de un corpus gensim

- Recoger los identificadores de las palabras presentes en nuestro contador de vectores.

```
In [ ]: #Para nuestro LDA también necesitaremos el objeto CountVectorizer definido en la etapa 3.Creación del corpus y la matriz

cv = pickle.load(open("cv.pkl", "rb"))
id2word = dict((v, k) for k, v in cv.vocabulary_.items())
```

Ilustración 14. Extraída del repositorio de GitHub: Ejemplo del contador necesario para nuestro LDA

- Llamar a las funciones del modelo especificando algunos parámetros.

```
In [ ]: #Lo ponemos en marcha

'''Debemos especificar el número de topics que creemos puede haber y el número de veces que recorrerá el texto.
Estos dos valores podemos variarlos hasta obtener un resultado relevante'''
lda = models.LdaModel(corpus=corpus, id2word=id2word, num_topics=4, passes=100)
lda.print_topics()
```

Ilustración 15. Extraída del repositorio de GitHub: Ejemplo de los parámetros especificados en nuestro LDA

- Observar la relación que automáticamente se genera entre temas valorados y nuestros corpus cargados.
- Repetir los últimos dos pasos hasta concluir un resultado significativo, y poner nombre a estos temas “descubiertos”.

4.7 Análisis de sentimiento

Las funciones descritas en este apartado se corresponden con el archivo 6. Análisis de sentimiento.ipynb del repositorio.

Por último, exploraremos nuestro texto mediante técnicas de análisis de sentimiento. Este tipo de análisis extrae información de la subjetividad presente en una serie de textos o documentos, y, como hemos avanzado en el estado de la cuestión (§ESTADO DE LA CUESTIÓN), suele estar centrado en la extracción de emociones, discriminando entre tres valores fundamentales: positivo, negativo y neutro.

Nuestro objetivo en este apartado no será demasiado osado, pero sí altamente significativo. Extraeremos un sentimiento general para cada uno de los medios que conforman nuestro corpus.

La gran distancia de recursos existentes para modelos en lengua inglesa y española nos han privado en este apartado del uso de paquetes tan completos como TextBlob o las funciones de análisis de sentimiento de word2vec, dos de las herramientas más comunes y eficientes en el mundo del PLN.

Por este motivo, se ha optado, tras una pequeña exploración previa, por utilizar el proyecto sentiment-analysis-spanish de Hugo J. Bello (Bello). Esta biblioteca utiliza redes neuronales convencionales para predecir la detección de sentimiento en idioma español. El modelo ha sido entrenado con 800.000 valoraciones de usuarios a través de las plataformas de *eltenedor, Decathlon, tripadvisor, filmaffinity* y *ebay*, y alcanza un 88% de precisión.

Su uso no será tan sencillo como los anteriores pasos, pues además de instalar el paquete sentiment-analysis-spanish, se requiere contar con las bibliotecas Keras y Tensorflow en nuestro dispositivo, de forma que cuando lo importemos a nuestra libreta Jupyter se mostrará algo parecido a:

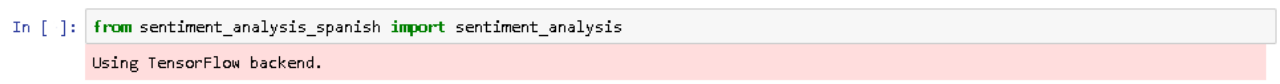

Ilustración 16. Extraída del repositorio de GitHub: llamada a la biblioteca sentiment-analysis-spanish

En este caso, retomaremos el corpus limpio creado en los primeros pasos, y aplicaremos la función de análisis de sentimiento a cada uno de los periódicos, de forma que extraigamos una valoración general a través del método `sentiment()`, que nos devolverá una valoración que va desde el 0 (más negativo) a 1 (más positivo).

In [ ]:
```
# Recordemos que tenemos por corpus un marco DataFrame de Pandas

corpus_df = pd.read_pickle('corpus.pkl')
```

In [ ]:
```
#Iniciamos la función se análisis de sentimiento

sentimiento  = sentiment_analysis.SentimentAnalysisSpanish()

'''Definimos qué medio queremos investigar apuntándolo en el DataFrame'''
medio = data.Noticias.loc["Nombre del periódico"]

sentimiento = sentimiento.sentiment(medio)
```

Ilustración 17. Extraída del repositorio de GitHub: análisis de sentimiento de nuestros periódicos

Caso de estudio

A continuación, hacemos una prueba del modelo definido en el apartado de metodología (§METODOLOGÍA) y compartido en el repositorio de *GitHub* `https://github.com/JuliaLleoPA/Periodismo-con-TextMining` aplicado al caso de estudio: la inspección de publicaciones de los diez periódicos generalistas más populares en el mes de marzo.

5.1. Elección del corpus

Según se ha fijado en la Justificación y objetivos de la investigación (§JUSTIFICACIÓN Y OBJETIVOS DE LA INVESTIGACIÓN), nuestro objetivo de análisis son las publicaciones del seguimiento de la crisis sanitaria de los medios digitales ABC, El Mundo, El País, El Periódico, 20minutos, La Vanguardia, El Confidencial, El Español, eldiario y OkDiario entre los días 1-31 de marzo de 2020.

Al enfrentarnos a una tarea de análisis de contenido[8] centrado en los efectos de la pandemia, no se ha considerado recoger todas las publicaciones del mes, sino que se ha realizado una selección previa que permitiera filtrar aquellas directamente relacionadas con el tema y resultaran pertinente, dejando de lado otro tipo de noticias que generaran ruido y anularan los frutos de la investigación.

Para realizar este primer filtrado, acudimos a las hemerotecas digitales que cada medio guarda en su portal web (excepto en el caso de Ok Diario, cuya exploración se realiza directamente con la herramienta de búsqueda del portal web), y se procuró la presentación de las siguientes palabras clave bien en los titulares bien en los cuerpos de noticia:

[8] Jaime Andréu Abela define el análisis de contenido como una técnica de interpretación de textos cuyo contenido, debidamente interpretado, nos abre las puertas al conocimiento de diferentes aspectos y fenómenos de la vida social. Lo característico de esta técnica de investigación recae en que se trata de una técnica que combina la observación y producción de datos, y la interpretación o análisis de los datos (Andréu Abrela, 2002).

- De materia sanitaria**:** Se definen las palabras clave de búsqueda perfilando una temática principalmente sanitaria y científica. Estas son: "**coronavirus**" y "**COVID-19**".
- De materia social**:** Se seleccionan los términos que atienden a la experiencia social durante el confinamiento. Se escogen: "**confinamiento**" y "**cuarentena**".
- De materia política nacional o internacional**:** Se escogen palabras que seleccionen noticias de interés político. Se seleccionan: "**crisis**" y "**pandemia**".

La elección de estas palabras se lleva a cabo tras un estudio de las secciones en que nuestros periódicos digitales dividen sus portales, así como sucesivas pruebas hasta dar con la cantidad de términos justa que recopile el mayor número de noticias relevantes para nuestro caso. En el momento de selección de la palabra "crisis", se previó la posibilidad de que esta apareciera asociada a temáticas no pertinentes para nuestra investigación como el desarrollo de la crisis de refugiados no relacionadas con su situación ante la pandemia. Por este motivo se realizó un segundo filtrado manual que excluyera las noticias no vinculadas a nuestro tema de estudio.

Dentro de cada publicación, hemos optado por el análisis del cuerpo informativo de la noticia. Esta elección se toma en un contexto en que diferentes medios se encuentran cubriendo sucesos similares y la preferencia de los lectores se puede rastrear desde el tratamiento de la información dentro de la noticia. Escoger este elemento nos permitirá tanto hacer generalizaciones sobre el nivel de cobertura de la pandemia como estudiar los diferentes enfoques con que se enfrenta el seguimiento de eventos.

En la figura 8 recogemos una lista con la referencia a las hemerotecas digitales que nos sirvieron para hacer la selección y el número de publicaciones recogidas en cada una. No se especifica la palabra clave por la que fueron recuperadas las noticias, ya que en muchas ocasiones estas aparecían duplicadas en un mismo recurso:

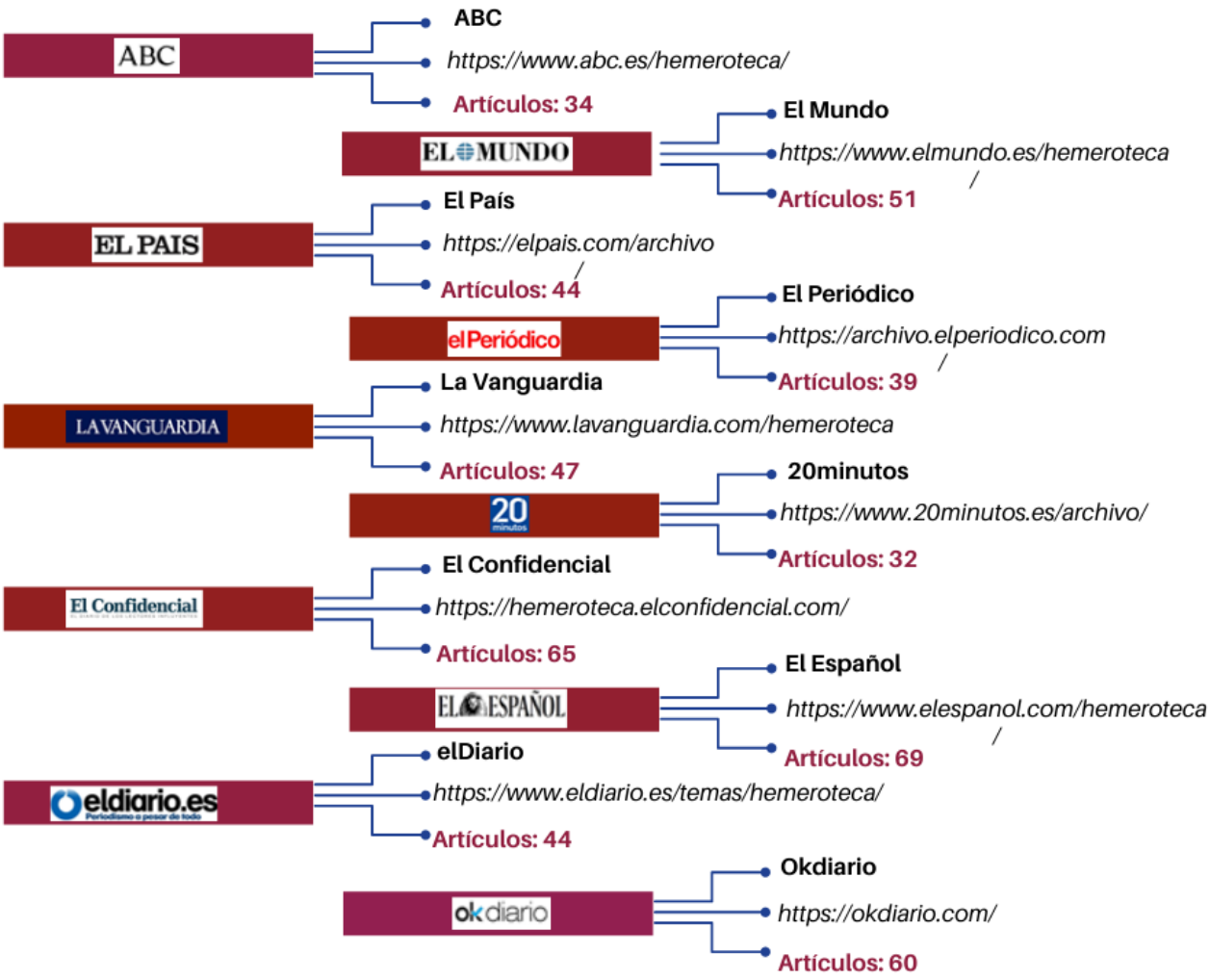

Figura 8. Relación de medios analizados con las direcciones de sus hemerotecas

5.2. Obtención del corpus

Para la obtención del cuerpo informativo de las noticias seleccionadas, se ha definido una función diferente por periódico examinado. Este paso se ha dado ya que, a la hora de recoger el texto perteneciente al cuerpo de cada publicación, ha sido preciso referenciarlos identificando los párrafos (etiqueta <p> en lenguaje HTML) asociados a la clase correspondiente por la Hoja de estilo en cascada (CSS) del portal.

Por ejemplo, así quedó nuestra función definida para el raspado el periódico 20minutos y automatización de la petición HTTP que recupere la etiqueta definida:

```python
# Raspado de texto del periódico 20minutos
def p20_minutos(url):
  '''Recorre el código HTML de https://www.20minutos.es/ y recupera los párrafos (etiqueta <p>) dentro de la clase "article-text"'''
  pagina = requests.get(url).text

  '''BeautifulSoup acepta dos argumentos: el marcado actual, y el parser que se quiere usar. En nuestro caso escogemos el parser "lxml", ya que también funciona para versiones antiguas de Python y es muy rápido'''
  html = BeautifulSoup(pagina, "lxml")

  '''Esta operación recupera el texto dentro de las etiquetas <p> pertenecientes a la clase "article-text"'''
  texto = [p.text for p in html.find(class_="article-text").find_all('p')]
  print(url)
  return texto
```

Función 1. Muestra de la función de web scraping

En total, estas fueron las clases identificadas como contenedoras de las etiquetas <p> con cuerpo informativo según cada medio:

Figura 9. Relación clases CSS para el llamado de la etiqueta párrafo

Siguiendo el resto de pasos, explicados en nuestro modelo, hemos obtenido una carpeta que contiene los diez archivos de texto plano correspondientes al conjunto de noticias escogidas por medio.

5.3. Limpieza del corpus

Para el proceso de limpieza se ha comenzado haciendo una previsualización de nuestros datos cargados. Para ello, hemos procedido creando un marco DataFrame, elemento de la biblioteca Pandas que nos permite trabajar con una hoja de datos en forma de tabla, pudiendo acceder o alterar los elementos de esta. Además, este tipo de hojas ofrecen la ventaja de permitir archivar su contenido en formato csv.

Para generar este marco de datos podemos proceder de distintas formas: expresando los datos iniciales en formato diccionario de Python o importando una tabla csv que ya contenga los datos. En nuestro caso, hemos creado un diccionario en el momento de cargar los datos iniciales almacenados con el módulo Pickle, de forma que la clave contenga el nombre del periódico y el valor almacene el conjunto de cadena de texto correspondiente al contenido de sus noticias. Hemos creado nuestro marco de datos siguiendo los siguientes pasos:

```
# Creamos el marco de datos de Pandas para visualizar nuestro diccionario
import pandas as pd

'''Especificamos cómo queremos que se represente'''
 pd.set_option('max_colwidth',200)

'''Por defecto los nombres de periódicos se colocarán como columnas, así que los transformamos a índices de fila'''
DataFrame = pd.DataFrame.from_dict(dict_raspado).transpose()

'''Le ponemos nombre a la columna de datos para poder referirnos a ella'''

DataFrame.columns = ['Noticias']

'''Ordenamos estas alfabéticamente'''
DataFrame = DataFrame.sort_index()
```

Función 2. Creamos un marco de datos con Pandas

A la hora de revisar este material podemos hacerlo de distintas formas:

- Desde la propia consola: llamando al nombre de las claves definidas en el diccionario cargado, o especificando la columna del DataFrame que nos interesa ver.
- De forma local: otra de las ventajas de estos marcos de datos es su facilidad de conversión a elementos CSV, permitiéndonos descargar los datos en nuestro ordenador y analizarlos en el editor apropiado.

Durante la exploración de los datos raspados nos encontramos con problemas no contemplados en nuestro modelo como:

- Presencia de distintos de emojis y emoticonos.
- Referencias a cuentas de Twitter y etiquetas.
- Enlaces hipermedia (principalmente a páginas web y fotografías de Twitter).
- Gazapos lingüísticos.

De estos elementos, se ha optado por conservar las referencias a cuentas y etiquetas de la red social Twitter, por considerarse que constituyen muestras válidas de los temas y personas de actualidad por los que se ha interesado cada medio. Para el resto de los conflictos, se ha diseñado una segunda fase de limpieza común previa a la creación del corpus lingüístico y de la matriz a través de métodos de sustitución *regex*.

Se ha conservado un seguimiento de todas las cuestiones resueltas en la tabla recogida en el primer anexo (§ ANEXOS).

Tras esta limpieza, se han vuelto a comprobar los datos para asegurarnos no habernos dejado ningún otro asunto que entorpezca el acceso al texto en el futuro.

5.4. Creación de corpus y matriz

La obtención del corpus ha seguido los pasos propuesto en el modelo sin conflictos, con la posibilidad de incluir una nueva columna en nuestro marco Pandas donde especificar algunos detalles de las publicaciones. El resultado obtenido presenta la siguiente estructura de tabulación:

	Noticias	Información del periódico
20 Minutos	Al menos 5 mayores han fallecido en lo que va de marzo en residencias de ancianos públicas y privadas de toda España, en el marco de la crisis pr...	Publicaciones del periódico 20 minutos durante el mes de marzo de 2020
ABC	El Obispado de Menorca retirará el agua bendita de las iglesias de la isla por el coronavirus El Obispado de Menorca ha emitido este jueves u...	Publicaciones del periódico ABC durante el mes de marzo de 2020
El Español	'El Loco' Gatti se encuentra en la capital de España aquejado de coronavirus. Este lunes saltó la noticia de su llegada a una clínica para someter...	Publicaciones del periódico El Español durante el mes de marzo de 2020
El Mundo	El coronavirus se ha cobrado la vida de una trabajadora de Correos de 51 años que reliazaba tareas de atención al cliente y reparto en la localida...	Publicaciones del periódico El Mundo durante el mes de marzo de 2020
El País	Boris Johnson y su ministro de Sanidad, Matt Hancock están infectados por el coronavirus y recluidos en sus hogares con síntomas leves. Ambos han ...	Publicaciones del periódico El País durante el mes de marzo de 2020
El Periódico	Sanitarios del Hospital Clínic de Barcelona devuelven a la ciudadanía los aplausos, el 24 de marzo. MARTA PÉREZ (EFE) La falta de test...	Publicaciones del periódico El Periódico durante el mes de marzo de 2020
El confidencial	Tras anunciar este sábado una nueva cifra récord de fallecidos diarios por el coronavirus y rozar ya los 0 muertos, el Gobierno italiano ha decid...	Publicaciones del periódico El Confidencial durante el mes de marzo de 2020
La Vanguardia	Actualización: 7 de julio de : La OMS no descarta la transmisión aérea del coronavirus 6 de julio de : Científicos instan a la OMS a cons...	Publicaciones del periódico La Vanguardia durante el mes de marzo de 2020
OkDiario	El presidente del Gobierno, Pedro Sánchez, deja en la cuneta a los autónomos para frenar la crisis económica del coronavirus. Así lo ha afirmado...	Publicaciones del periódico OkDiario durante el mes de marzo de 2020
eldiario	El entrenador catalán Pep Guardiola ha donado un millón de euros para comprar material sanitario contra el coronavirus. La aportación del...	Publicaciones del periódico eldiario durante el mes de marzo de 2020

Ilustración 18. Ejemplo del corpus en un marco DataFrame

En el caso de la matriz, además de la lista de palabras vacías (stop words) definido en el apartado de Metodología, añadiremos algunas más adaptadas para nuestro caso de estudio. Esta decisión se toma después de comprobar cómo cierto número de palabras se encuentran entre las más recurrentes en más del 70% de nuestras fuentes. Por ejemplo, el hecho de que un término como “coronavirus” sea demasiado utilizado en todos los periódicos, resta valor específico y nos dificulta hacer exploraciones más concretas, por lo que podemos pasar a considerarlo una “palabra vacía”. Del resultado de hacer estos cálculos extraemos que la lista de palabras que podrían dificultar nuestro estudio y su correspondiente media de aparición por noticia:

Figura 10. Palabras vacías definidas para nuestro caso

Con esta actualización en la variable de `palabras_vacias`, pasamos a calcular nuestra matriz según la fórmula del modelo. En total, la bolsa de palabras generada tiene 23304 columnas (es decir, términos diferentes identificados) y 10 filas correspondientes a los periódicos contenedores. En ellas, como ya explicamos, se contabilizan el número de apariciones de término por medio, con un aspecto similar al siguiente:

	abad	abaitua	abajo	abalanzado	abanca	abandera	abandona	abandonado	abandonados	abandonan	...	últimas	último
20 Minutos	2	0	0	0	0	0	0	0	0	0	...	5	5
ABC	0	0	0	0	2	0	0	0	0	0	...	18	11
El Español	0	0	0	0	0	1	0	3	0	1	...	18	11
El Mundo	0	0	0	2	0	0	0	0	0	0	...	48	18
El País	0	0	0	0	0	0	1	1	1	0	...	11	17
El Periódico	0	0	1	0	0	0	0	1	0	0	...	10	3
El confidencial	0	0	0	0	0	0	1	5	0	0	...	35	25
La Vanguardia	0	0	0	0	0	0	0	1	0	0	...	5	7
OkDiario	0	2	0	0	0	0	0	0	0	0	...	10	7
eldiario	0	0	0	0	0	0	0	0	1	0	...	48	9

10 rows × 23304 columns

Ilustración 19. Muestra de una visualización de nuestra matriz

5.5. Análisis del texto

Llegados a este punto, si la limpieza ha sido exhaustiva y correcta, los resultados de nuestras funciones no necesitarán mayores adaptaciones. Dividimos a continuación los tres procesos de exploración.

5.5.1. Palabras más comunes por medio

En el anexo 2 (§ ANEXOS) se muestra la solución al cálculo de frecuencia de los diez términos más repetidos por medio. A partir de ese resultado, también hemos calculado la media de aparición de estas palabras por noticia. El resultado obtenido lo podemos expresar con una bolsa de palabras a través de la biblioteca Python Matplotlib:

Figura 11. Nube de palabras por medio

5.5.2. Riqueza léxica de cada medio

Aplicada la fórmula de medición de riqueza léxica definida en el modelo, obtenemos el siguiente resultado:

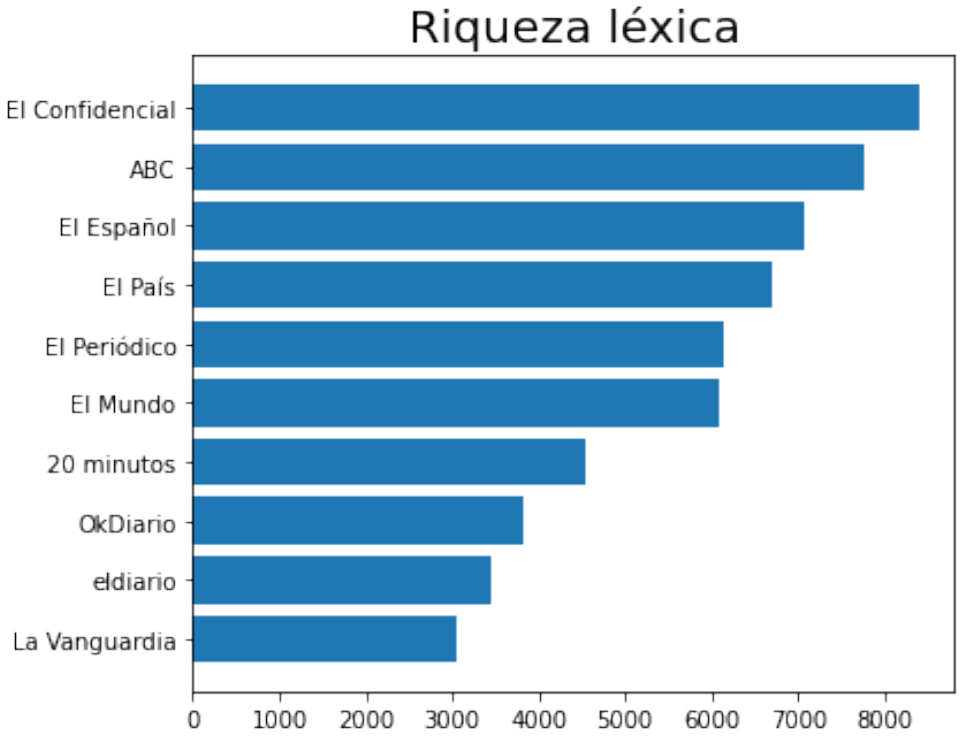

Figura 12. Cálculo de la riqueza lingüística

En ella se muestra cómo El Confidencial lidera la riqueza en la expresión y medios como OkDiario, eldiaro o La Vanguardia se ciñen a una menor variedad y una mayor insistencia de palabras o expresiones.

5.5.3. La media de palabras total por cada periódico

El cálculo de la media de palabras tampoco tuvo problemas aplicando la fórmula diseñada. En su definición, se anotó el número de publicaciones por medio, y el resultado extraído tiene la siguiente forma:

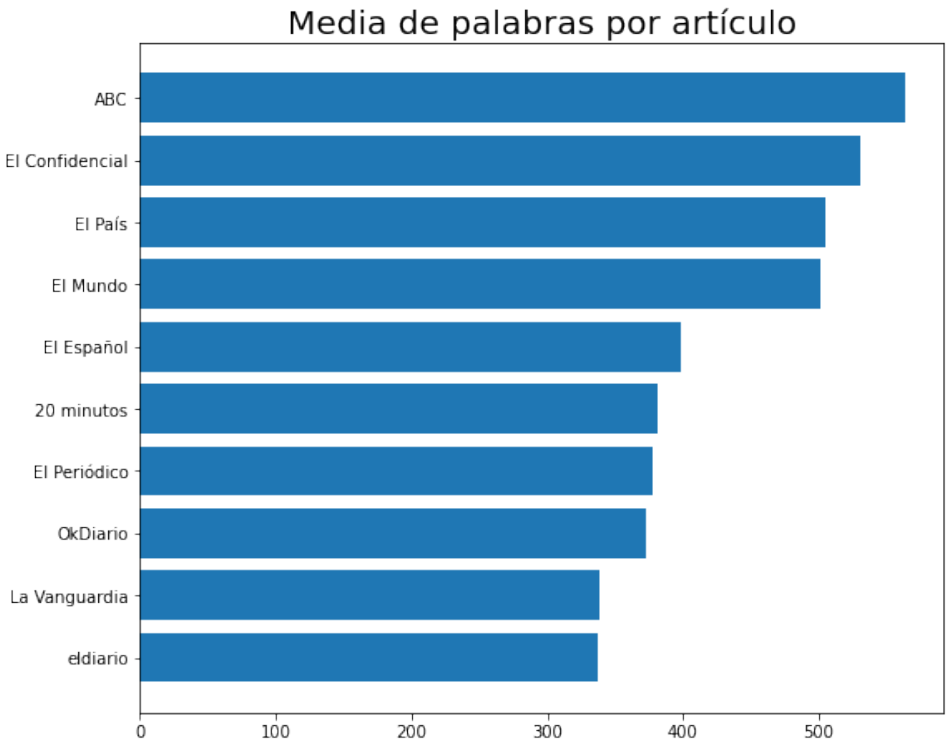

Figura 13. Cálculo de la media de palabras por artículo

Aunque la media general no varía demasiado entre todos los medios (se mantiene fija entre las 300 y 550), sí se destaca cómo algunas comunicaciones tienden a ser más breves que otras.

5.6. Modelado automático de tópicos

En este nuevo apartado, el reto principal consistió en identificar la mejor optimización de los parámetros: número de tópicos esperados y número de vueltas que debía recorrer el algoritmo. Finalmente, los resultados más significativos se obtuvieron con 1.000 vueltas, y un número de tópicos inferior a 5.

En nuestra búsqueda se tuvo presente que el número de tópicos con que partía nuestra investigación eran tres, los mismos que guiaron nuestra selección inicial de datos: una temática de contenido social, otra sanitaria, y otra de política internacional.

De entre estas pruebas, hemos seleccionado la opción de cuatro tópicos por su relevancia semántica. A continuación, aparecen representadas las relaciones que el modelo LDA ha previsto:

A partir de estas conexiones hemos pasado a dar nombre a cada temática y definir su contenido, así como calculado, con la segunda parte de la fórmula definida en el apartado de Metodología (§Metodología), la temática más sobresaliente en cada medio:

Conexiones detectadas con número de tópicos: 4									
número	cierre	ministerio	contagios	prensa	informado	nacional	ministro	contagio	autoridades
montero	moncloa	ministra	contagios	cese	ministro	líder	illa	contagio	prensa
evitar	tiempo	medida	número	sanitaria	ministerio	cierre	sanitario	contagios	mayor
sistema	sanitario	población	tiempo	última	número	seguir	evolución	mayor	actuación

Figura 14. Entidades detectadas por tópico

Figura 15. Identificación de las temáticas por medio

Como se observa, la temática más cubierta en los periódicos estudiados ha sido la situación y avance del estado de alarma y su correspondiente confinamiento del total de la población. A pesar de que este decreto no fue anunciado hasta el 14 de marzo, medio mes de cobertura ha sido suficiente para los periódicos 20minutos, ABC, El mundo, El Periódico y La Vanguardia lo conviertan en su tema estrella.

5.7. Análisis de sentimiento

Por último, el modelo de análisis de sentimiento aplicado a nuestro caso de estudio devolvió el resultado de la valoración de cada medio entre una escala de 0 a 1, donde 0 se corresponde con el mayor grado de carga subjetiva negativa, 0.5 indicaría una carga subjetiva neutra y 1 una carga subjetiva positiva.

Para poder proceder con cualquier método de detección automática de sentimiento, es necesario identificar un lexicón previamente que actuará como identificador de subjetividad. En el caso del paquete sentiment-analysis-spanish, este lexicón se construye a través de un corpus, donde las características propias de este corpus se utilizan como semilla inicial para la ampliación del lexicón (McKeown, 1997), (Kanayama, 2006). En el repositorio del programa podemos encontrar el corpus de opiniones negativas y positivas que se utilizaron para el entrenamiento del modelo[10]. Entre la documentación del proyecto también se puede ver cómo en las pruebas de evaluación del modelo, que obtienen un 88% de precisión, una de comprobaciones es precisamente una entradilla del periódico ABC publicada a 31 de marzo de 2020 en relación con el avance de la pandemia en nuestro país[11].

Los resultados de nuestra medición de sentimiento por periódico obtienen los siguientes resultados:

[10] Disponible en el enlace: https://github.com/sentiment-analysis-spanish/sentiment-analysis-model-neural-network/tree/master/data/json_bundle_reviews.

[11] Disponible en https://github.com/sentiment-analysis-spanish/sentiment-analysis-model-neural-network/blob/master/sentiment_regression_neural_network/evaluate_model.py

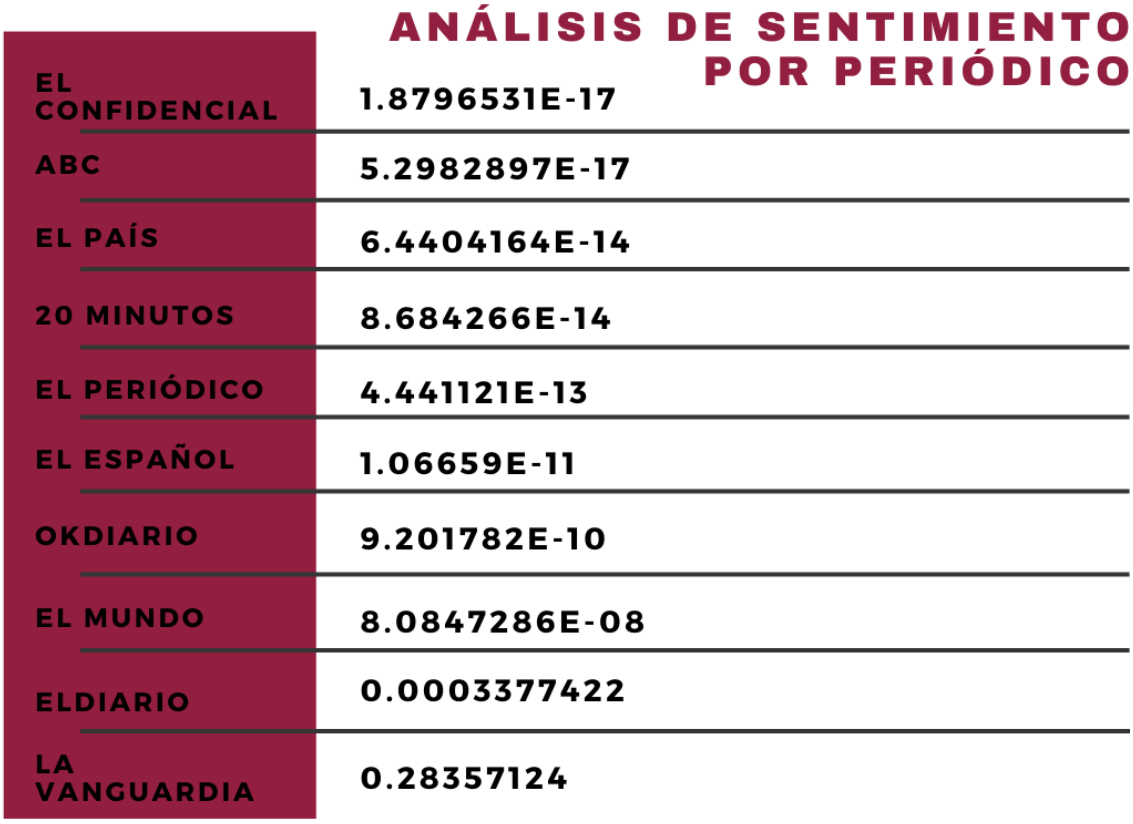

Figura 16. Extracción del sentimiento por periódico

La drástica tendencia al valor 0 (el mayor grado de carga subjetiva negativa) se puede observar de forma más visual en la siguiente tabla:

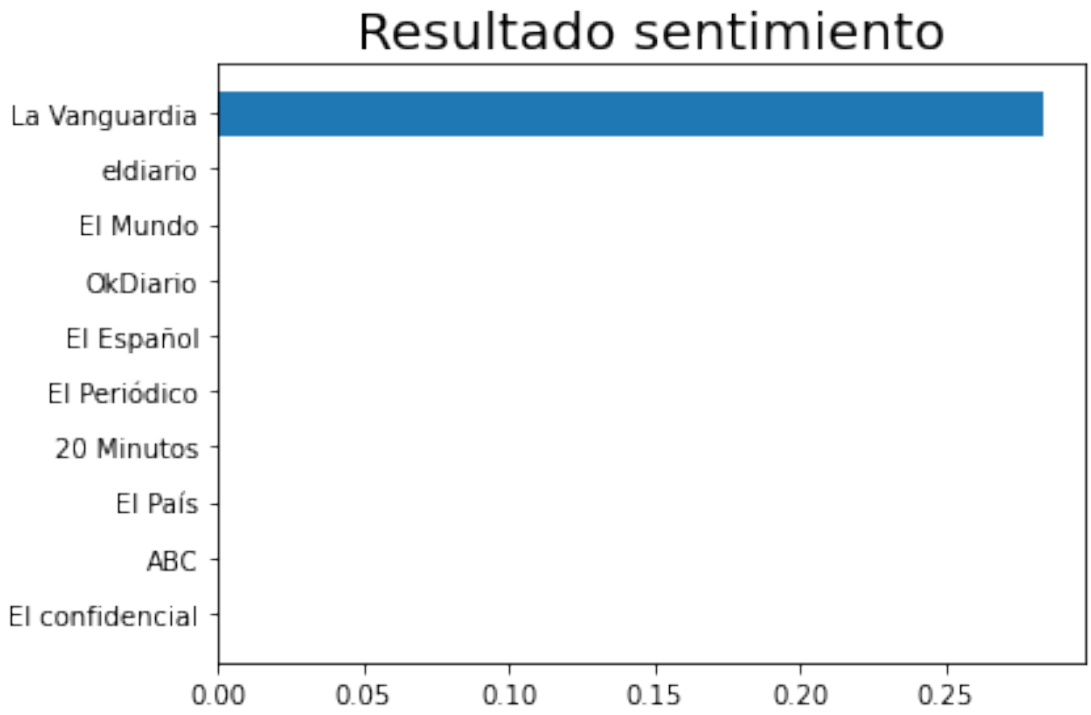

Figura 17. Visualización en gráfica de los resultados de sentimiento

Una de las posibles conclusiones ante estos resultados es que, en una situación de crispación política, incomprensión y amenaza, los valores extraídos tienden, en todo caso, a ser negativos. A excepción del periódico La Vanguardia, que sube hasta los 0.28 puntos, el resto de los diarios muestran valores rozando el mínimo de negatividad del sistema.

Para explicar estos datos recurrimos a los tres tipos de actores identificados por Wilson (Wilson, 2005) en la expresión de opinión en nuestro idioma:

- Variables de negación o afirmación: se trata de expresiones polares con carga semántica negativa o positiva. Por ejemplo, elementos como “no”, “siempre”, “sin”, “tampoco”.
- Modificadores de variable u oracionales: modificadores de polaridad con carga negativa o positiva, como “grande”, “por supuesto”, “en absoluto”.
- Variables de modificación de polaridad: palabras que actúan con carga negativa o positiva en un determinado contexto, pero no contienen una carga polar intrínseca, como pueden ser “esperar” o “rápido”.

Este último elemento resulta el más engañoso y, dado que la tecnología utilizada identifica los valores de polaridad a través de un lexicón de críticas en páginas web de servicios, algunas de las mediciones pueden no haber sido concluyentes en nuestro caso. Por ejemplo, en las publicaciones analizadas, los términos “positivo” y “negativo” permutan sus significados, de forma que una herramienta entrenada con valoraciones como sentiment-analysis-spanish no resulta completamente eficaz, pero sí nos permite aproximarnos al tono imperante de las publicaciones.

Interpretación de los datos

Con los resultados obtenidos tras la aplicación de nuestro modelo de automatización de análisis textual, podemos ya dar paso a la interpretación del tema que guía nuestra investigación: el drástico aumento y traslación de lectores entre los diez periódicos digitales más leídos en España en el mes de marzo de 2020.

Si recordamos los datos recogidos por el medidor de audiencias digitales Comscore (mostrados en las figuras 4 y 5), en referencia al mes de marzo, nos encontrábamos con que, en total, son cuatro los periódicos que experimentan cambios severos distinguiéndose de la tendencia del mes a un aumento controlado de lectores:

1. **Eldiario**: por su rotundo aumento de lectores (61,8%), escalando a la primera posición en la lista de variaciones de marzo.
2. **OkDiario**: con una subida del 41%, siendo el segundo medio con mayor aumento.
3. **La Vanguardia**: con un aumento moderado, que sin embargo no le impide encabezar de nuevo la lista de medios digitales generalistas más leídos.
4. **El Confidencial**: su poco crecimiento de lectores, en comparación a la subida en otros medios, le hacen descender en la lista general.

Si ponemos el foco en estos cuatro medios, y revisamos los datos obtenidos en nuestro caso de estudio, podemos extraer algunas conclusiones generales como las que se presentan a continuación.

Nuestra nube de palabras y el cálculo de los diez términos más frecuentes por medio generados en la primera fase de análisis de texto revelan cómo los dos periódicos triunfantes en los cambios de audiencia de marzo se especializan en temas de política nacional y económica respectivamente.

En el caso de OkDiario la información recopilada del mes de marzo tiende a tratar asuntos de política interna, en las que destacan las numerosas referencias al actual presidente y vicepresidente del Gobierno. Estas referencias, además de revelar un interés por las políticas de contención del virus, parecen interesarse por los casos de ministros, consejeros o senadores bajo amenaza de contagio durante el mes.

Figura 18. Nube de palabras del periódico OkDiario

En el caso de eldiario, la información recopilada tanto desde el cómputo de frecuencia como la nube de palabras revelan un interés por el sector económico ligado a una apuesta por la información de datos. También destaca su atención a la situación de empresas y laboral de la población española.

Figura 19. Nube de palabras del periódico eldiario

Por su parte, podemos ver cómo los temas más destacados de una exploración completa del medio La Vanguardia descubren un mayor seguimiento a cuestiones internacionales como el avance de la pandemia en Italia, primer foco europeo de la pandemia de coronavirus desde la identificación del primer contagiado el 28 de febrero. Otro de los intereses del medio parece ser las cuestiones económicas.

Figura 20. Nube de palabras del periódico La Vanguardia

Por otro lado, la variedad en la expresión, que observamos en la *figura 12. Cálculo de la riqueza lingüística* (ver figura 12) de la medición de la riqueza lingüística de los medios, parece actuar en contra de lo esperado: es precisamente la menor variedad expresiva la técnica utilizada por los medios mejor posicionados.

En completa oposición al periódico El Confidencial (principal afectado en el nuevo posicionamiento), se encuentran los favorecidos OkDiairo, eldiario, y La Vanguardia. Esta tendencia revela cómo la información clara de los hechos, con la creación de tópicos y reiteración de referencias y términos de actualidad, mejora la recepción de las comunicaciones en un periodo de impaciencia y confusión.

También en este sentido, en la *figura 12. Cálculo de la media de palabras por artículo* (§ figura 12) vemos cómo los comunicados breves y directos parecen ser los vencedores en la captación de lectores. Con una media entre las 300 y 400 palabras, el contenido se vuelve más atractivo y aumenta la difusión en las publicaciones.

Por su parte, el análisis de sentimiento, aunque ya concluimos no del todo ajustado a nuestros fines en la fase de estudio, por haber sido entrenado sobre un corpus de valoraciones que poca luz pueden arrojar para fines periodísticos, sí nos ha permitido confirmar una atmósfera de negatividad en la redacción de toda la prensa estudiada del mes. En este sentido, Jones *et al.* (Jones, Waters, Holland, Bevins, & Iverson, 2010) estudian la forma en que una advertencia de miedo generalizada en situaciones en que solo existe un riesgo hipotético, como ocurre en este mes preludio de las primeras consecuencias del virus en nuestro país, favorecen una confrontación más liviana llegados los peores escenarios.

Como lo resuelven Carmen Costa-Sánchez y Xosé López-García, en los estadios previos a una pandemia los mensajes deben enfocarse a incrementar el conocimiento de la enfermedad e informar de los comportamientos de protección para reducir el riesgo de su transmisión (Costa-Sánchez & López-García, Comunicación y crisis del coronavirus en España. Primeras lecciones, 2020, pág. 4). Este estudio, además, muestra claramente el cambio de paradigma en las comunicaciones en diferentes canales desde una etapa "precrisis", que alentaba a conservar la calma y equiparaba la enfermedad a otra gripe estacional, al cambio radical en el mensaje a partir del mes de marzo y el drástico incremento de infectados en el país. Es en este momento cuando, al tener que alertar a la población contra los mensajes tranquilizadores anteriores, esta reacciona con miedo provocando excesos como el desabastecimiento de supermercados, la vuelta a la prensa tradicional (por confiar en la mayor veracidad de sus mensajes) y, en general, el aumento en el consumo de información. Sin duda, nuestro análisis de sentimiento sí revela esta tendencia a la dramatización y alarma en las publicaciones.

Por último, la aproximación a los tópicos detectados y destacados en cada medio vuelve a mostrar similitudes en la transmisión de los acontecimientos por los dos periódicos líderes en crecimiento de audiencia. En conclusión, otra de las claves que podría haber atraído a muchas de las personas alarmadas y, más tarde, confinadas, es la atención concedida a la política nacional y crisis económica.

A este respecto, recordamos que las conexiones lingüísticas identificado como tema propio por nuestro modelo LDA incluía diferentes nombres de agentes gubernamentales que, en toda situación de crisis, se vuelven los actantes más importantes. La decisión de proporcionar actualizaciones continuas de sus comunicaciones públicas ha servido a estos medios para aumentar la credibilidad en la información proporcionada.

Conclusiones

En conclusión, con nuestro modelo hemos podido explorar una gran cantidad de datos textuales en muy poco tiempo y con relativamente pocos recursos. Gracias a este acercamiento hemos podido sacar algunas conclusiones generales que nos han ayudado a dar respuesta a la pregunta con que partía nuestra investigación: los cambios de audiencia que la situación de incertidumbre provocada por la crisis sanitaria y posterior confinamiento de la población habían provocado en los hábitos de consumo de prensa digital. Todo ello ha sido posible gracias a, en primer lugar, la revolución digital en el sector periodístico, que abre la puerta a exploraciones de minería de textos en gran cantidad de sus publicaciones, y los logros del campo del Procesamiento del Lenguaje Natural, que facilitan la exploración de la información textual aplicada a todo tipo de ámbitos.

Para la consecución de nuestros fines, hemos definido un modelo específico de exploración del ámbito periodístico y publicado el mismo para futuros usos de análisis con datos diferentes. Dentro de este, algunas de las tareas han resultado más concluyentes, como la medición estadística de diferentes fenómenos a partir de la elaboración de una matriz lingüística o el modelado automático de tópicos de tipo LDA. El análisis de sentimiento, otro de los objetivos fijados, no se valora como solución definitiva por no haber sido creado *ad hoc* para nuestros propósitos, presentando problemáticas tan notorias como el cambio en la semántica que términos como “positivo” y “negativo” tienen en estas circunstancias, apareciendo en considerables ocasiones en los corpus y que llegan a alterar la precisión de la tecnología utilizada.

Aunque se trata de un modelo no demasiado osado, resulta una buena muestra de los beneficios que las tecnologías informáticas de exploración textual ofrecen a campos de negocio como la medición de audiencia dentro del periodismo digital. Entre sus futuras mejoras se considera la implementación de limpieza por *stemming* o acortamiento de los términos de la matriz a su raíz, y una mayor exploración en los recursos de análisis de sentimiento, que hasta ahora continúan siendo escasos para su aplicación al idioma español.

Bibliografía

Andréu Abrela, J. (2002). *Las técnicas de análisis de contenido: una revisión actualizada.* Fundación Centro de Estudios Andaluces.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., . . . Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. *30th Intl conf on machine learning*, 280-288.

Bello, H. J. (s.f.). https://github.com/sentiment-analysis-spanish/sentiment-analysis-model-neural-network.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with Python:.* California: O'Reilly Media, Inc.

Blei, D. M. (2012). Topic modeling and digital Humanities. *Journal of digital humanities, 2*(1), 8-11. Obtenido de http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Blei, D. M., & Lafferty, J. D. (2007). *A correlated topic model of science.* Princeton: University and Carnegie Mellon University.

Breton, P., & Proulx, S. (1989). *La explosión de la comunicación.* Barcelona: Civilización Ediciones.

Cambria, E., Schuller, B., Liu, B., & Wang, H. (2013). Knowledge-based approaches to concept-level sentiment analysis. *IEEE intelligent systems, 28*(2), 12-14.

Castells, M. (1996). *he Rise of the Network Society, The Information Age: Economy, Society, and Culture.* Oxford: Blackwell.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology, 37*, 51–89.

Costa-Sánchez, C., & López-García, X. (2020). Comunicación y crisis del coronavirus en España. Primeras lecciones. *El profesional de la información, 9*(3), 1-14.

Costa-Sánchez, C., & López-García, X. (2020). Comunicación y crisis del coronavirus en España. Primeras lecciones. *El profesional de la información, 29*(3), 1-14.

Data is Beautiful. (19 de octubre de 2019). Most Popular Programming Languages 1965 - 2019. EEUU. Recuperado el 2 de agosto de 2020, de https://www.youtube.com/watch?v=Og847HVwRSI

David M Blei, A. Y. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*, 993–1022.

de Moragas i Spà, M. (1981). *Teorías de la comunicación. Investigaciones sobre.* Barcelona: Gustavo Gili.

Edo, C. (2003). *Periodismo informativo e interpretativo: el impacto de Internet en la noticia, las fuentes y los géneros.* Comunicación Social Ediciones y Publicaciones.

Esparza, J. (2016). Epidemias y pandemias virales emergentes: ¿Cuál será la próxima? *Investigación clínica, 57*(3), 231-235.

Feldman, S. (1999). NLP meets the jabberwocky. *Online, 23*, 62-72.

G. Miner, D. D. (2012). The Seven Practice Areas of Text Analytics. En D. D. G. Miner, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (págs. 29-41). Har/Dvdr: Academic Press.

Gou-Núñez, M. (2017). *Crisis de salud en los medios de comunicación : El ébola en España. Análisis de los diarios La Vanguardia, El Mundo y El País.* Univeritat Autònoma de Barcelona, Grau en Periodisme).

Harvey, D. (1990). *The Condition of Postmodernity. An Enquiry into the Origins of Cultural Change.* Oxford: Blackwell.

Hearst. (1999). Untangling Text Data Mining. *Proc. of ACL'99: The 37th Annual Meeting of the Association for Computational.* University of Maryland.

Hernández, D., & Navarro, B. (2015). Una aproximación a la recomendación de artículos científicos según su grado de especificidad, Pr*ocesamiento del Lenguaje Natural, 55*, 91-98.

Jones, D. E. (1997). Investigació sobre comunicació social a l'Espanya de les autonomies. *Anàlisi. Quaderns de comunicació i cultura*, 101-120.

Jones, S. C., Waters, L., Holland, O., Bevins, J., & Iverson, D. (2010). Developing pandemic communication strategies: Preparation without panic. *Journal of business research, 63*(2), 126-132.

Kanayama, H. &. (2006). Fully Automatic Lexicon Expansion for DomainOriented Sentiment Analysis. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (págs. 355–363). Association for Computational Linguistics.

Kechaou, Z., Ben-Ammar, M., & Alimi, A. (2013). A multi-agent based system for sentiment analysis. *International journal on artificial intelligence tools, 22*(2), 1-28.

Kodratoff. (1999). Knowledge Discovery in Texts: A Definition. *Proc. of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99)*, (págs. 16-29).

Leetaru, K.-H. (2011). *Data mining methods for the content analyst: An introduction to the computational analysis of informational center.* New York: Routledge.

Liddy, E. (1998). Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science, 24*, 14-16.

Llisterri, J. (15 de diciembre de 2019). *Aplicaciones del procesamiento del lenguaje natural*. Obtenido de http://liceu.uab.cat/~joaquim/language_technology/NLP/PLN_aplicaciones.html

McKeown, V. H. (1997). Predicting the Semantic Orientation of Adjectives. *Proc. of the 35th ACL Proc. of the 18th Conf. on Computational Linguistics*, 299-305.

Monjas-Eleta, M., & Gil-Torres, A. (2017). Comunicación institucional y tratamiento periodístico de la crisis del ébola en España entre el 6 y el 8 de octubre de 2014. *Revista de comunicación, 6*(1), 97-121.

Parés i Maicas, M. (1997). La recerca europea en comunicació social. *Anàlisi.Quarderns de comunicació i cultura, 21*, 21-234.

Rodríguez Serrano, A., & Gil Soldevilla, S. (2018). *Investigar en la era neoliberal. Visiones críticas sobre la investigación en comunicación en España.* Universitat Pompeu Fabra : Universitat Autònoma de Barcelona, Servei de Publicacions : Universitat de València, Servei de Publicacions.

Rossmann, C., Meyer, L., & Schulz, P. J. (2018). The mediated amplification of a crisis: Communicating the A/H1N1 pandemic in press releases and press coverage in Europe. *Risk analysis, 38*(2), 357-375.

Salaverría Aliaga, R., del Pinar Martínez-Costa, M., & Breiner, J. (2018). Mapa de los cibermedios de España en 2018: análisis cuantitativo. *Revista Latina de Comunicación Social*, 1034-1053.

Sojo, C. A. (2003). *El periodismo en Internet.* Universidad Central de Venezuela: Fondo editorial de Humanidades y Educación.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, New Series, 59*(236), 433-460.

Vega, F. (24 de marzo de 2020). *https://www.comscore.com/lat/Prensa-y-Eventos/Blog/Los-medios-tradicionales-recuperan-poder-y-credibilidad-con-la-pandemia-provocada-por-el-Coronavirus.* Obtenido de https://www.comscore.com/: https://www.comscore.com/esl/Prensa-y-Eventos/Blog/Los-medios-tradicionales-recuperan-poder-y-credibilidad-con-la-pandemia-provocada-por-el-Coronavirus

Verdejo Maillo, M. F. (1994). Procesamiento del lenguaje natural: fundamentos y aplicaciones. *UNED, Curso de Verano.*

Wilson, T. W. (2005). Recognizing contextual. *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 347-354.

Wolton, D. (2000). *Internet, ¿y después?. Una Teoría Crítica Sobre los Nuevos "Media".* Barcelona: Gedisa.

World Health Organization. (2004). *Sixth Futures Forum on Crisis Communication.* Obtenido de https://www.euro.who.int/__data/assets/pdf_file/0004/90535/E85056.pdf

World Health Organization. (2013). *Health and environment: communicationg the risks.* Obtenido de https://www.euro.who.int/__data/assets/pdf_file/0011/233759/e96930.pdf

World Health Organization. (2020). *Risk communication and community engagement readiness and response to coronavirus disease (Covid-19).*

Yan X., G. J. (2013). *A Biterm Topic Model for Short Texts.* CAS Beijin: Institute of Computing Technology.

Anexos

ANEXO 1

CONFLICTOS RESUELTOS EN UNA SEGUNDA FASE DE LIMPIEZA PERSONALIZADA

En la siguiente tabla se recogen las cuestiones resueltas en la fase de limpieza personalizada para nuestro caso de estudio.

EMOJIS	HIPERVÍNCULOS	EXTRAÍDO DEL MEDIO	GAZAPOS LINGÜÍSTICOS
	pic.twitter.com IXu EA3EHu5	ABC	explicandome
	pic.twitter.com 8n6D0Pap EL	ABC	videconferencia
	pic.twitter.com NJSd R2o EST	ABC	ncohe
	pic.twitter.com W0KTaqirrj	ABC	permanante
	pic.twitter.com r ZRls Y2Mf	ABC	infeciones
	pic.twitter.com m Pa VNz IHMz	ABC	níumero
	pic.twitter.com c Nw Z7z HU2X	ABC	desifectado
	pic.twitter.com Iv AKFc Oq9q	ABC	auentan
	pic.twitter.com 1Jt Ira4 .36	ABC	explcado
	pic.twitter.com r ZRls Y2Mf5	ABC	osbtante
	pic.twitter.com IKpagpq Gpu	El Español	reisgo
	pic.twitter.com 8ZYVnkh QSr	El Español	miemo
	pic.twitter.com 9Te6a FP0Ri	El Español	percepcion
	pic.twitter.com Tgu WH6Blij	El Español	pdieramos
	pic.twitter.com VJRQpj AAl8	El Español	prohibe
€	pic.twitter.com ejw FYwl LKJ	El Español	cornavirus
$	pic.twitter.com z W0GLl SCq2	El Español	madrido
	pic.twitter.com ud XCb Z2apq	El Español	encuentan
	pic.twitter.com dbn Uzi43i R	El Español	manentenemos
	pic.twitter.com AOy Ll7z0r FComo	El Español	proque
	pic.twitter.com 1Ixr Be30g5	El Español	cotagio
	pic.twitter.com cj JT89Od ND	El Español	producieron
	pic.twitter.com E9rwp9o D3I	El Español	desemparada
	https: t.co LVha Yrd IGE pic.twitter.com zmfnae VZL5	El Español	interveniendo
	pic.twitter.com IKpagpq Gpu	El Español	detatallado
	https: t.co Hi J1c Pd Eyc	El Español	principaes
	https: t.co ddq Ft2P6Nv	El Español	prevenención
	https: t.co TXmpyp Fpo2	El Español	electricadad

	pic.twitter.com 9Te6a FP0Ri	El País	cuerentena
	pic.twitter.com d MIJiv GXl I	El País	indetificado
	https: t.co 4jqgfc UHno	El País	entreteniento
	pic.twitter.com b3h8dns Qe A	El Periódico	funciomiento
	https: t.co Nw2b Ssjwho pic.twitter.com v Rjs E5BBk PA	El Periódico	contínuamente
	https: t.co z2yd4Rlw XO pic.twitter.com Bqy Vd0QQxb	El Periódico	insitencia
	pic.twitter.com gl QKg4MYl V	El Periódico	auque
	pic.twitter.com Umwiv ANw SV	El Periódico	econóimica
	pic.twitter.com l5bh WXDi Ut	El Periódico	excecpcionales
	pic.twitter.com k Zk Eadmes8	El Periódico	autónomomas
	pic.twitter.com j N3v DB678h	El Periódico	covd
	pic.twitter.com p Tg Jzt UXqo	El Periódico	
	https: t.co s P3t Aj YSm1	El Periódico	
	https: t.co S54n Hwax Lr pic.twitter.com e AB4QPVthi	El Periódico	
	https: t.co TIF6DD	El Periódico	
	https: t.co p1qs OYVADw	El Periódico	
	https: t.co FFB1eyn US7	El Periódico	
	https: t.co 8q Gj JQJEJF	El Periódico	
	https: t.co xra7FDBV83	El Periódico	
	https: t.co S54n Hwax Lr pic.twitter.com e AB4QPVthi	El Periódico	
	https: t.co JODY3WB36Q#coronavirus	El Confidencial	
	https: t.co Wf6ph S2wa G pic.twitter.com Zm U3Nealhr	El Confidencial	
	https: t.co AEc MZTRwyz	El Confidencial	
	https: t.co Ps OXAlh Ohq	El Confidencial	
	pic.twitter.com Xlp Cv Fh0mv	El Confidencial	
	pic.twitter.com TFs Tlhbr Im	El Confidencial	
	pic.twitter.com 67r S4XBE5J	El Confidencial	
	pic.twitter.com i Ls Gu7YXb8	El Confidencial	
	pic.twitter.com 95z R4Gw2vg	El Confidencial	
	pic.twitter.com Jpr No NA2r6	El Confidencial	
	pic.twitter.com 07n SL8V1pv	El Confidencial	
	pic.twitter.com 9Te6a FP0Ri	El Confidencial	
	pic.twitter.com 67r S4XBE5J	El Confidencial	
	pic.twitter.com i Ls Gu7YXb8	El Confidencial	
	pic.twitter.com 95z R4Gw2vg	El Confidencial	
	pic.twitter.com Jpr No NA2r6	El Confidencial	
	pic.twitter.com 07n SL8V1pv	El Confidencial	
	pic.twitter.com Xy R3Zfz Q5u	El Confidencial	
	pic.twitter.com cs Y0Dc	El Confidencial	
	pic.twitter.com yk6d tf	El Confidencial	
	https://t.co/hr3GM8rGHS\n	elDiario	
	https: t.co hr3GM8r GHS	elDiario	
	pic.twitter.com/h19Ub0D02A\n	elDiario	

ANEXO 2

CÁLCULO DE LOS DIEZ TÉRMINOS MÁS REPETIDOS POR MEDIO

En la siguiente tabla se recogen los diez términos más frecuentes por medio, así como su total de apariciones en el corpus por periódico y la media correspondiente de frecuencia por noticia.

Periódico	Palabras	Apariciones	Media de apariciones por noticia
20Minutos	marzo	28	0,875
	casa	27	0,84375
	residencias	26	0,8125
	confinamiento	22	0,6875
	comunidad	22	0,6875
	ancianos	19	0,59375
	semana	19	0,59375
	salir	17	0,53125
	sánchez	16	0,5
	mayores	16	0,5
ABC	sanidad	68	2
	italia	65	1,911764706
	sábado	61	1,794117647
	informa	61	1,794117647
	china	61	1,794117647
	comunidad	58	1,705882353
	simón	57	1,676470588
	marzo	53	1,558823529
	horas	52	1,529411765
	hospital	51	1,5
El Español	casa	79	1,144927536
	iglesias	51	0,739130435
	sanidad	48	0,695652174
	pasado	48	0,695652174
	alimentos	45	0,652173913
	sociales	45	0,652173913
	bien	44	0,637681159
	semanas	41	0,594202899
	euros	40	0,579710145
	italia	40	0,579710145
El Mundo	domingo	174	3,411764706
	horas	125	2,450980392

	covid	111	2,176470588
	sanidad	110	2,156862745
	sánchez	109	2,137254902
	informa	108	2,117647059
	sábado	100	1,960784314
	comunidad	93	1,823529412
	número	93	1,823529412
	ministerio	84	1,647058824
	generalitat	84	1,647058824
El País	enfermedad	56	1,272727273
	países	54	1,227272727
	confinamiento	53	1,204545455
	china	51	1,159090909
	comunidad	48	1,090909091
	semana	46	1,045454545
	epidemia	43	0,977272727
	millones	42	0,954545455
	mundo	40	0,909090909
	sistema	40	0,909090909
El Periódico	información	97	2,487179487
	confinamiento	40	1,025641026
	casa	40	1,025641026
	barcelona	28	0,717948718
	pacientes	25	0,641025641
	empleo	24	0,615384615
	síntomas	22	0,564102564
	jornada	21	0,538461538
	hospital	21	0,538461538
	enfermedad	21	0,538461538
El Confidencial	sánchez	133	3,41025641
	sanidad	112	2,871794872
	comunidad	93	2,384615385
	sábado	92	2,358974359
	ciudadanos	84	2,153846154
	iglesias	81	2,076923077
	medida	80	2,051282051
	cierre	74	1,897435897
	enfermedad	74	1,897435897
	consejo	73	1,871794872
La Vanguardia	economía	39	0,829787234
	millones	35	0,744680851
	deuda	34	0,723404255
	italia	31	0,659574468
	pasado	29	0,617021277

	epidemia	29	0,617021277
	confianza	26	0,553191489
	semanas	25	0,531914894
	confinamiento	25	0,531914894
	casa	25	0,531914894
OkDiario	sánchez	87	1,45
	iglesias	79	1,316666667
	sanidad	57	0,95
	ejecutivo	55	0,916666667
	pedro	48	0,8
	marzo	45	0,75
	pablo	40	0,666666667
	ministros	40	0,666666667
	sociales	37	0,616666667
	pasado	36	0,6
eldiario	información	120	2,727272727
	mundo	105	2,386363636
	boletín	70	1,590909091
	hazte	70	1,590909091
	respuestas	67	1,522727273
	datos	63	1,431818182
	eldiario	61	1,386363636
	minuto	59	1,340909091
	vida	55	1,25
	epidemia	54	1,227272727