Universidad Austral de Chile

Facultad de Ciencias de la Ingeniería
Escuela de Ingeniería Civil en Informática

RECOPILACIÓN, ANÁLISIS Y VISUALIZACIÓN DE DATOS DE MEDIOS DE PRENSA INTERNACIONALES, PERIOSDISTAS Y FUENTES DE INFORMACIÓN

Proyecto para optar al título de
Ingeniero Civil en Informática

PROFESOR PATROCINANTE:
MATTHIEU PIERRE VERNIER
DR. EN INFORMÁTICA

CO-PATROCINANTE
FABIÁN ALEXIS RUIZ FLORES
ING. CIVIL EN INFORMÁTICA

PROFESOR INFORMANTE
LUIS RAMÓN CÁRCAMO ULLOA
DR. EN PERCEPCIÓN, COMUNICACIÓN Y TIEMPOS

ELEAZAR ZACARÍAS VÁSQUEZ FUENTEALBA
VALDIVIA – CHILE
2022

ÍNDICE

ÍNDICE DE TABLAS

Tabla **Página**

ÍNDICE DE FIGURAS

Figura **Página**

RESUMEN

La recopilación, análisis y visualización de datos relacionados con medios de prensa, periodistas y fuentes de información, son parte de una ampliación relacionada con una línea investigativa perteneciente al grupo Sophia2. Este nuevo enfoque generado, que contó con retroalimentación y aportes de revisores de una revista internacional, implementa nuevas ideas y herramientas con el objetivo de generar un aporte real a este grupo de trabajo.

Mediante el desarrollo de diversos scripts, se generaron módulos que permitieron la recopilación y procesamiento de información relacionada con los medios de prensa. Todo esto se realizó implementando diversos métodos de extracción de datos y utilizando recientes técnicas de procesamiento de texto.

La implementación de este desarrollo contó con una arquitectura de software, perteneciente a este grupo de investigadores, que soportó la integración de los módulos generados. Este acoplamiento permitió y permite actualmente la automatización de los procesos de recopilación y procesamiento de datos. Además, la integración de una base de datos en esta arquitectura ayudó a generar el almacenamiento y la disponibilidad de los datos obtenidos.

Finalmente, se realizó una etapa de análisis y visualización con el objetivo de identificar aspectos generales de la información obtenida. Por lo tanto, este trabajo no solo aporta nuevos enfoques de desarrollo, sino que también entrega resultados preliminares que pueden ser la base de futuras investigaciones.

ABSTRACT

The collection, analysis and visualization of data related to press media, journalists and information sources are part of an extension related to a research line belonging to the Sophia2 group. This new approach, which received feedback and contributions from reviewers of an international journal, implements new ideas and tools with the aim of generating a real contribution to this working group.

Through the development of several scripts, modules were generated that allowed the collection and processing of information related to the press media. All this was done by implementing various data extraction methods and using recent text processing techniques.

The implementation of this development relied on a software architecture, belonging to this group of researchers, which supported the integration of the modules generated. This coupling allowed and currently allows the automation of data collection and processing processes. In addition, the integration of a database in this architecture helped to generate the storage and availability of the data obtained.

Finally, an analysis and visualization stage was carried out in order to identify general aspects of the information obtained. Therefore, this work not only provides new development approaches, but also delivers preliminary results that can be the basis for future research.

1. INTRODUCCIÓN

1.1. Contexto: Sophia2, un grupo de investigación interdisciplinar para analizar el pluralismo en los medios desde la Lingüística computacional y el Machine Learning

El presente proyecto se enmarca en el proyecto Fondecyt de Iniciación n° 11190714 titulado "Sophia2: Métodos basados en Lingüística Computacional y *Machine Learning* para analizar el pluralismo en los medios de prensa" desarrollado en la Facultad de Ciencias de la Ingeniería (FCI) de la UACh entre diciembre 2019 y diciembre 2023. En este proyecto colaboran un grupo de investigadores y estudiantes de la FCI y de la Facultad de Filosofía y Humanidades de la UACh.

Desde una perspectiva general, el proyecto Sophia2 busca crear indicadores del pluralismo en los medios de prensa basándose en algoritmos de Lingüística Computacional y *Machine Learning*. El grupo de investigadores tiene como visión proporcionar nuevas herramientas que permitan concientizar el pluralismo en los medios y activar el pensamiento crítico de los ciudadanos.

Desde 2016, este grupo desarrolla softwares y protocolos científicos para analizar fenómenos sociales y cognitivos vinculados al consumo y a la producción de noticias de prensa en Chile. En particular, podemos citar el trabajo de tesis de Magíster en Informática realizado por Fabian Ruíz (2020) (Ruiz, 2020), y en fase de publicación en la revista científica *Digital Journalism*, cuyo objetivo principal consistía en diseñar y probar un método computacional para caracterizar la evolución de los sesgos de género en la prensa chilena basándose sobre el algoritmo *Dynamic Topic Models* (Blei & Lafferty, 2006). Este método fue probado con un caso de estudio sobre 369,860 noticias de la prensa chilena entre 2016 y 2019. Permitió obtener indicadores de sesgos de género similares a los obtenidos por la UNESCO y el *Global Media Monitoring Project* de manera más escalable. Por ejemplo, permitió caracterizar automáticamente que el discurso mediático chileno tiende a presentar las mujeres como actrices o políticas y los hombres de manera más diversa (futbolistas, políticos, expertos, líderes de opinión, líderes artísticos, etc.).

1.2. Origen del proyecto: una primera investigación sobre las fuentes de información y los periodistas realizada en 2019-2020 que requiere ampliarse

Una de las últimas investigaciones tuvo como resultado un manuscrito no publicado titulado: "*Do woman exist in Chilean?: Gender of Journalist and sources*", en donde se hace un análisis de los medios de prensa, periodistas y fuentes de información (personas citadas en las noticias, ver ejemplo en Tabla 1) en chile en temas de género y que plantea y responde preguntas como:

- ¿Quién escribe las noticias en Chile?

- ¿Quiénes son las fuentes citadas en los medios en Chile?
- ¿Existe una relación entre el género de los periodistas y el género de las fuentes que citan?

Tabla 1. Ejemplos de fuentes de información

Extractos de noticias	**Fuentes de información**
"La presidenta de la Convención Constituyente (CC), **Elisa Loncón**, envió un oficio a la Cámara de Diputados para solicitar ayuda…" (biobiochile.cl, 2021)	Elisa Loncón
"…cuenta que sigue sin anunciarse la renovación de **Lionel Messi** seis días después de que acabara su contrato, el presidente **Joan Laporta** ha querido mandar un mensaje de tranquilidad…" (mundodeportivo.com, 2021)	Lionel Messi Joan Laporta

Para responder esas preguntas se utilizó un método cuantitativo basado en un enfoque de ciencias sociales computacionales. Este método combina técnicas de *web scraping* y Procesamiento del Lenguaje Natural (NLP) para recopilar y procesar los datos. Para los análisis se utilizaron más de 12 mil artículos publicados de noticias correspondiente a 16 medios de prensa digitales, abarcando un periodo de tiempo de aproximadamente 2 meses. Los resultados que lograron obtener mostraron importantes sesgos de género, por ejemplo: el 62% de las noticias fueron escritas por hombres y el 79% de las fuentes citadas eran hombres. Todos estos análisis fueron enfocados para poner sobre la mesa resultados en concreto y que estos puedan ser sujetos de discusiones para concientizar sobre estos temas a los estudiantes y profesionales de esta área en específico.

Este trabajo fue enviado en la revista científica *Feminist Media Studies* el I semestre 2020 y se recibieron comentarios que apuntan a ampliar la investigación (ver Anexo 1). En particular, los dos revisores mencionan el interés del método computacional utilizado ("Algo a destacar de este artículo, que no se enfatiza lo suficiente en el texto, es el método utilizado para recolectar y analizar los datos", "Crear un algoritmo, ponerlo a disposición para su reutilización y proponer nuevas formas de extraer corpus para análisis es muy valioso para el campo de estudio y otros investigadores."), pero recomiendan que el *dataset* final incluya más medios de prensa para ser más representativo ("algo que también falta en el artículo es incluir una perspectiva teórica del Sur Global y de América Latina en particular"), y más metadatos sobre los periodistas y fuentes de información para enriquecer el análisis ("el conjunto de datos final se queda corto, ya que no incluye suficientes variables, más allá del género y el medio de comunicación, para permitir un análisis más profundo", "se podrían incluir variables como el tema a tratar, la profesión de la fuente, la edad del periodista o el tiempo de permanencia en la profesión, la visibilidad de la noticia dentro del medio, entre otros factores, que permitiría una comprensión más amplia del fenómeno.").

El presente trabajo toma su origen en modo de darle la oportunidad al equipo Sophia2 de ampliar esta línea investigativa. Se busca seguir las principales recomendaciones recibidas por parte de los revisores de la revista *Feminist Media Studies*.

1.3. ¿Cómo ampliar la investigación del equipo Sophia2?

Para ampliar la investigación del equipo Sophia2 se considera tres ejes de mejoramiento:

1. Se considera una cobertura mucho más amplia. En lugar de solo de analizar 16 medios de prensa de Chile, este proyecto integra 150 medios de prensa, abarcando medios de prensa de distintos continentes (América, Europa y Oceanía) incluyendo diferentes idiomas (inglés, español, portugués, italiano y francés) y culturas.
2. Se enriquece los metadatos describiendo las fuentes de información y periodistas. El estudio anterior sólo incluía el nombre y el género como variables, cuando este proyecto toma en cuenta ocho variables que se pueden extraer automáticamente con técnicas de *web scrapping* y tratamiento automático del lenguaje: nombre, género, edad, profesión, nacionalidad, temáticas, popularidad y visibilidad mediática. Estas variables se describen más en detalle en el capítulo 3.
3. Se aumenta el periodo de análisis. El estudio anterior analizaba un mes de noticias de prensa (julio 2019), cuando este proyecto analiza varios años de datos (2018-2021).

Este trabajo de ampliación implica una complejidad mucho mayor al estudio anterior en término de ciencia de la computación y de ingeniería de software. El cambio de escala es importante entre recopilar noticias de 16 medios de prensa de un mes y recopilar noticias de 150 medios de prensa de 4 años. Este cambio requiere diseñar e implementar una arquitectura de software adaptada para soportar una fuerte carga computacional. Estos aspectos no son directamente parte de este proyecto de título. Para estos objetivos, el ingeniero Fabian Ruiz desarrolla la arquitectura de software llamada Caleuche. Sin embargo, este proyecto de título necesita poder integrarse con esta arquitectura.

En particular, este proyecto apunta a integrar scripts para “scrappear” 150 medios de prensa, scripts para extraer metadatos desde las noticias de prensa y Wikipedia y scripts para almacenar todos los datos recopilados en una base de datos relacional. Estos scripts deben integrarse con la arquitectura Caleuche. Detallamos estos aspectos en el capítulo 3.

1.4. Objetivos

1.4.1. Objetivo General

Este proyecto busca ampliar la investigación del equipo Sophia2 implementando una serie de scripts para recopilar, analizar y visualizar datos de medios de prensa, periodistas y fuentes de información. Los scripts deben integrarse con la arquitectura de Software Caleuche que desarrolla el equipo Sophia2.

1.4.2. Objetivos Específicos

1. Recopilar datos de medios de prensa internacionales integrando “scrappers” en la arquitectura Caleuche.
2. Implementar scripts para identificar el autor de las noticias y clasificarlo según distintas variables (por ejemplo: género, popularidad, temática, etc.).
3. Implementar scripts para identificar las fuentes de información y clasificarlas según distintas variables (por ejemplo: género, popularidad, edad, profesión, etc.).
4. Extraer hallazgos, a través de estadísticas y visualizaciones de los datos, que faciliten el análisis de las desigualdades de géneros en los medios de prensa.

2. MARCO TEÓRICO

2.1. Web Scraping: definición, aspectos legales y éticos

Data scraping o raspado de datos es una técnica en la que un programa informático extrae datos de una manera legible por humanos (datos estructurados) procedentes de otros programas (por lo general, datos no estructurados). Cuando esta extracción se realiza sobre datos contenidos en páginas web (formatos HTML), esto se conoce como *web scraping* (ver Figura 1). El programa utilizado para este proceso accede directamente a la web (*world wide web*) usando el protocolo de transferencia de hipertexto (HTTP) o por un *web browser* (navegador web). Por otro lado, esta es una técnica que puede ser realizada manualmente por un usuario mediante algún software, pero comúnmente este término hace referencia a procesos automatizados implementados mediante un *bot* o *web crawling*.

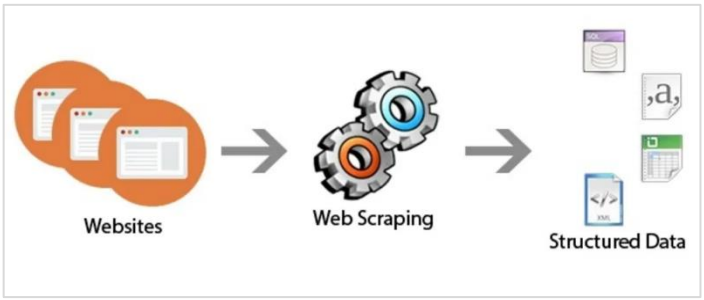


Figura 1. Representación de *web scraping*

Web crawling o comúnmente llamado Araña, es un *bot* el cual su propósito es indexar sitios webs y para ello navega sistemáticamente por la web. Este tipo de rastreador a menudo se enfrentan a problemas porque consumen recursos en los sistemas que visitan, por lo cual existen diversos mecanismos para detectarlos y bloquearles el acceso.

Dado que la legislación sobre el *web scraping* varia de un país a otro y en algunos casos no está lo suficientemente claro, hay que tener muchas consideraciones para poder abordar e implementar soluciones basadas en este enfoque. Por ejemplo, Han y Anderson (2020) exponen lo importante de tener claro que los datos protegidos por los derechos de autor y su utilización con fines comerciales se pueden considerar como algo ilegal, esto puede generar problemas que incluso coloquen en riesgo todo el trabajo realizado. Además, estos autores recalcan el tener en cuenta que las leyes actuales están cambiando constantemente y por eso, es recomendable tener un asesoramiento legal en estos casos.

Además de las cuestiones legales, se debe tener en consideración las limitaciones éticas a la hora de recopilar datos para investigaciones o estudios. Massimino (2016) argumenta la importancia de tener en cuenta que cualquier entidad u organización puede verse afectada por el uso del *web scraping*, por ejemplo, el hecho de que con esta técnica se permitan realizar peticiones de automatizadas y rápidas sobre un host, puede ocurrir que se exceda el umbral de ancho de banda y hacer que los servidores no respondan. Por eso, este autor comenta algunas indicaciones a seguir como limitar las solicitudes por

segundos, ejecutar sobre horas en las que existe menor demanda o leer y tener en consideración el Protocolo de Exclusión de Robots (REP). Esta información se puede extraer regularmente de los mismos sitios webs.

2.2. ¿Qué técnicas de tratamiento automático del lenguaje se puede utilizar para extraer datos sobre las fuentes de información y periodistas?

2.2.1. Preprocesamiento

Con la finalidad de recopilar datos de fuentes de información, se hace necesario identificar previamente a personas. La extracción de nombres de entidades en un texto, particularmente nombres de personas, es posible mediante soluciones existentes como la que proporciona *spaCy*. Esta es una librería gratuita de código abierto para el procesamiento del Lenguaje Natural (NLP) desarrollada en Python. Uno de los enfoques que provee esta librería para resolver esta tarea, está relacionado con lo que llaman Tuberías Entrenadas (*Trained Pipelines*) (Figura 2ver Figura 2). Esta es una secuencia de procesos encadenados entre sí, que recibe como parámetro de entrada un flujo de datos (texto) y que devuelve el resultado de haber procesado esta información.

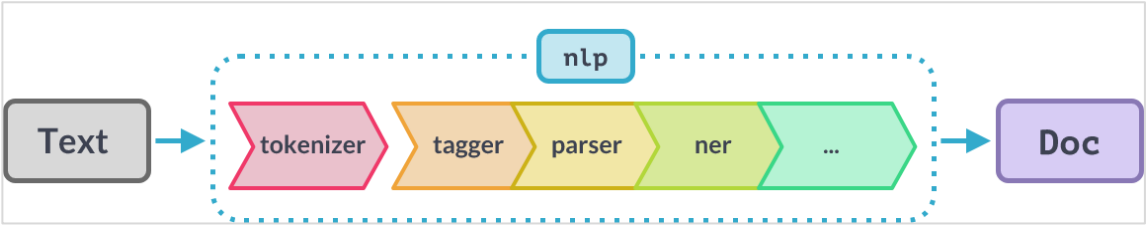


Figura 2. Representación de Tuberías Entrenadas

Reconocimiento de Entidades Nombradas (*Named Entity Recognition*, NER) es una de las implementaciones integradas y facilitadas dentro de estos *pipelines*. Este proceso detecta y etiqueta entidades con nombres dado un texto. Una entidad con nombre se define como un "objeto del mundo real" al que se le asigna un nombre (ver ejemplo Figura 3). Esta solución detecta múltiples tipos de entidades y una de ellas es *person* (nombres de personas). Esta asignación se realiza utilizando modelos estadísticos, entrenados con datos etiquetados permitiendo la predicción de estas entidades.

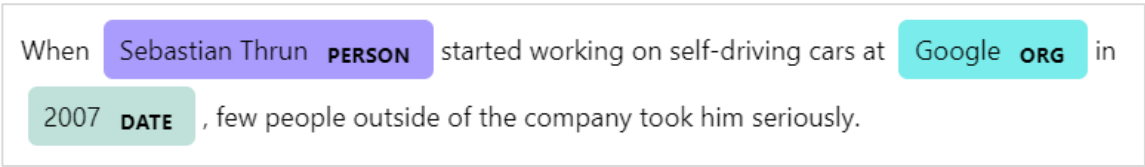


Figura 3. Ejemplo de Reconocimiento de Entidades Nombradas con spaCy

Otra librería de Python que proporciona una solución a esta tarea es la llamada Transformers. Esta librería trabaja de manera bastante similar a spaCy en la identificación de entidades. Su diferenciación nace en el uso de arquitecturas de propósito general como

BERT, *GTP-2*, *RoBERTa* y otras, conocidas como Transformadores (Vaswani et al., 2017). Esto son modelos de aprendizaje profundo (*Deep Learning*) utilizados principalmente en el área del procesamiento del lenguaje natural, planteando nuevos paradigmas con mejores resultados

A inicios del año 2021, el equipo de desarrolladores de spaCy lanzó la versión 3.0 de la biblioteca. En esta nueva versión, se incluyen nuevas tuberías con modelos de Transformers pre entrenados, esto supone mejores rendimientos en las tareas que estos pueden realizar.

Generalmente, es posible encontrarse con diversos problemas al realizar esta tarea. No importando el modelo utilizado, es importante tener en consideración que no se obtendrán resultados exactamente precisos. En spaCy, los modelos mejor catalogados giran aproximadamente en un 90% de precisión. Por otro lado, a partir del etiquetado pueden surgir a futuro ambigüedades o errores en análisis proveniente de alcances de nombres entre distintas personas.

2.2.2. Question-Answering con Transformers

Otra tarea relevante es recopilar información pública de personas y de fuentes de libre acceso. Un enfoque para realizar esta tarea consiste en utilizar modelos Transformers y lo que se denomina *Question-Answering*.

El Transformador es un modelo de aprendizaje profundo (Vaswani et al., 2017) que utiliza el mecanismo de Atención (*Attention*), que de una manera simplista consiste en mejorar partes importantes de los datos de entrada y desvanecer el resto. Este nuevo tipo de arquitecturas han tenido un aumento en implementaciones con modelos pre entrenados. Mejor rendimiento y capacidad han provocado el reemplazo de las típicas arquitecturas utilizadas en NLP, como eran los modelos de Redes Neuronales Recurrentes.

Entre las tareas que se habían planteado como desafío en NLP, se puede encontrar la comprensión lectora. Por ejemplo, si se hiciera una pregunta sobre un párrafo ¿Cómo podría responderse esa pregunta? Una solución a esto fue desarrollada por el grupo de NLP de la Universidad de Stanford, mediante la creación de dos conjuntos de datos SQUAD y SQUAD 2.0. Estos son conjuntos de "comprensión lectora", consisten en preguntas planteadas por *crowd workers* en un conjunto de artículos de Wikipedia, donde la respuesta a cada pregunta es un segmento de texto del pasaje de lectura correspondiente. Además, estos conjuntos disponen de preguntas que no pueden ser respondidas dado el contexto (por diseño), con la idea de permitir la formación de sistemas que puedan admitir que no saben la respuesta.

Huggingface, comunidad de desarrollo donde es parte la librería Transformers, proporciona modelos Transformers pre entrenados que han sido afinados en diferentes conjuntos de datos, entre ellos SQUAD, permitiendo realizar el *Question-Answering*.

El uso de estos modelos es muy simple, solo se necesita un texto y una pregunta como entrada. Obviamente, las precisiones de respuesta de estos modelos están condicionados en gran medida a sí la pregunta está acorde al contexto del texto del ingresado.

2.3. ¿Cuáles son los trabajos más similares a este trabajo?

Diversos trabajos plantean enfoques similares a este, distinguiéndose ciertas características más comunes o distintas que a continuación se detallan.

Madrid-Morales (2020) presentó algunas herramientas disponibles para la recopilación sistemática y automatizada, el almacenamiento y análisis de noticias presentadas de maneras digitales en algunos medios de África. Su objetivo era estudiar los patrones de los contenidos de estas noticias. En esa línea, propuso su propio enfoque de cuatro pasos utilizando paquetes desarrollados en el lenguaje de programación de código abierto R. Este proceso consiste en primer lugar en la recopilación de los datos utilizando la técnica de *web scraping*; más adelante considera alguna herramienta para el análisis de texto como: quanteda (Benoitet al. 2018), tidytext (Silge & Robinson, 2016) o tm (Feinerer, Hornik & Meyer, 2008) perteneciente al procesamiento de datos; para el análisis de datos presenta métodos computacionales como: *Dictionary-based methods* (Grimmer & Stewart, 2013), *Supervised machine learning* (Boumans & Trilling, 2016), *Unsupervised machine learning* (Welbers, Van Atteveldt, & Benoit, 2017); finalmente, utiliza el entorno robusto y versátil ggplot (Wickham, 2016) para la visualización de los datos.

Otro trabajo interesante realizado por Eshbaugh-Soha y McGauvran (2017), aborda la desigualdad de ingresos a través de la cobertura de noticias en los Estados Unidos. La extracción de los datos se realiza utilizando la técnica de *web scraping* y para el análisis del texto utilizan Leximancer 3.1, software de análisis de texto con un enfoque en encontrar conceptos de alto nivel dentro de un contexto.

Bhagat, Mishra, Dixit y Chang (2021) realizaron una comparación de los discursos entre diferentes noticias y blogs con respecto a la positividad o negatividad de los discursos relacionados con el Covid-19. Para esto, utilizaron para la extracción de datos *web scraping* y para la parte del análisis utilizaron la técnica de análisis de sentimientos (Medhat, Hassan & Korashy, 2014).

Como resumen de estos trabajos, se puede decir que, a pesar de los distintos enfoques de estudio o investigación que se abordan, las técnicas de cómo estos se desarrollan siguen una línea bastante común. Esto se puede ver primeramente en cómo se recopilan los datos (uso del *web scraping*), los enfoques de procesamiento (relacionados con el procesamiento del lenguaje natural) y los resultados encontrados. En contraparte, existen ciertas diferencias en cuanto a este trabajo. En primer lugar, estos trabajos solo se enfocan en un idioma base, por lo cual se podría decir que es una limitación en cuanto al alcance de recopilación de información y en efecto de los mismos resultados. Por este motivo, este trabajo busca diferenciarse de modo que incluso se logren capturar diferencias (si es que la hubiera) culturales en los resultados, por ejemplo, diferencias entre países

hispanohablantes y anglosajones. Además, aquí se contempla un cantidad y diversidad de datos muy superior (ordenes de millones), lo cual en teoría permitiría reflejar la realidad de lo que se encuentre con mejor precisión.

3. ARQUITECTURA DE SOFTWARE DE SOPHIA2

3.1. Diagrama general de la arquitectura de software de Sophia2

¿Cuáles son los principales componentes de la arquitectura de Sophia2? La arquitectura Sophia2 se compone de una serie de servidores (ver Figura 4) que interactúan entre sí y que permiten la ejecución de los diferentes proyectos asociados a Sophia2.

Sophia2-Control, servidor dedicado a levantar cada uno de los servidores principales de Sophia2. Caleuche-prod, servidor en él que se ejecutan múltiples servicios de *scraping*. Estos servicios se conectan con Sun (servidor que contiene la base de datos) para guardar los datos recopilados. Babylonia-prod, conjuntos de máquinas dedicadas al procesamiento de los datos (Sun). Sun-backup-prod, servidor de backup de la base de datos.

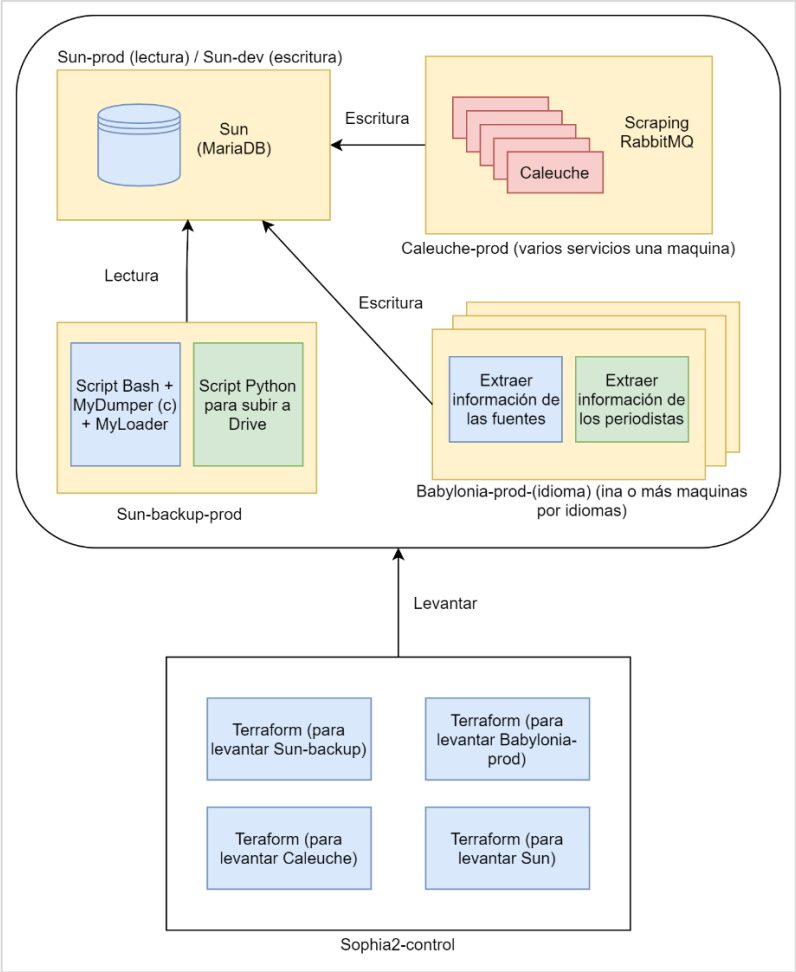


Figura 4. Arquitectura de software Sophia2

¿Cuáles son los componentes involucrados en mi trabajo? Los componentes de esta arquitectura que están involucrados en el desarrollo e implementación de este trabajo son: Sun, Caleuche y Babylonia.

3.2. Base de datos de Sophia2: Sun

¿Cuál es el modelo de la base de datos Sun? El modelo de datos relacional que actualmente existe en Sophia2 se puede observar en la Figura 5. Este modelo considera una serie de tablas que serán fundamentales para el desarrollo de este trabajo.

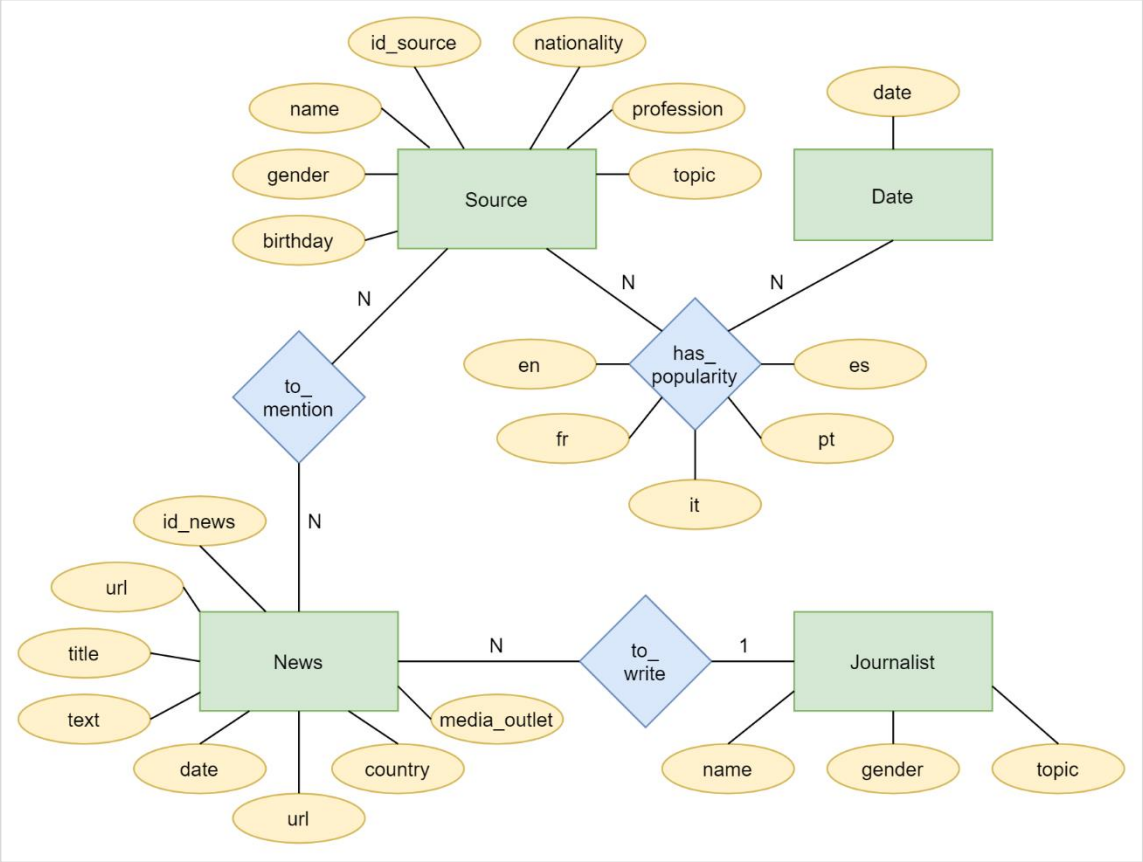


Figura 5. Base de datos de Sophia2

¿Cuáles tablas están involucradas en mi trabajo? Las tablas involucradas son las siguientes:

- News, donde se guardarán todas las noticias recopiladas
- Journalist, donde se guardará la información de los periodistas
- Sources, donde se guardará la información de las fuentes de información

3.3. Servicio de scraping de datos: Caleuche

Caleuche es una serie de servicios asociados al *scraping*. Cada servicio está diseñado para integrar los distintos scripts de *scraping* que se desarrollarán en este trabajo. Su propósito general es controlar y programar extracciones de datos (noticias) cada cierto periodo de tiempo.

3.4. Servicio de extracción de información de las fuentes de información y periodistas: Babylonia

Babylonia, es un servicio relacionado al procesamiento de los datos almacenados. Este servicio cuenta con una serie de máquinas (por idioma), las cuales integrarán los scripts de procesamiento de datos que se desarrollarán. Su propósito es procesar la información cada vez que el servicio Caleuche finalice.

4. DATOS Y MÉTODOS

4.1. Desarrollo de *scrappers* de medios de prensa

4.1.1. ¿Cuáles son los medios de prensa que se quiere *scrapear*? ¿Por qué?

Antes de iniciar con el desarrollo de scripts de *scraping* y la recopilación de noticias, se hizo necesario definir formalmente cuales son condiciones para la búsqueda y selección de los medios que finalmente son parte de este trabajo. Estas condiciones tienen que ver con:

- Medios más importantes en el país o región, considerando América (Norte, Centro y Sur), Europa y Oceanía.
- Medios que contemplen el nombre del periodista en sus artículos de noticia (en todos o en su gran mayoría).
- Medios que contengan noticias de a lo menos 2 años hacia atrás (contando del 2018 aproximadamente).
- Medios escritos en los siguientes idiomas: español, inglés, francés, portugués e italiano.

La importancia de un medio se sustenta en el *Digital News Report 2020*, un estudio anual encargado por el *Reuters Institute for the Study of Journalism* dedicado a comprender cómo se consumen las noticias en múltiples países del mundo (ver ejemplo en Figura 6). A través de diversas encuestas, se ponderan o resalta (gráficamente) los medios de prensa más consumidos.

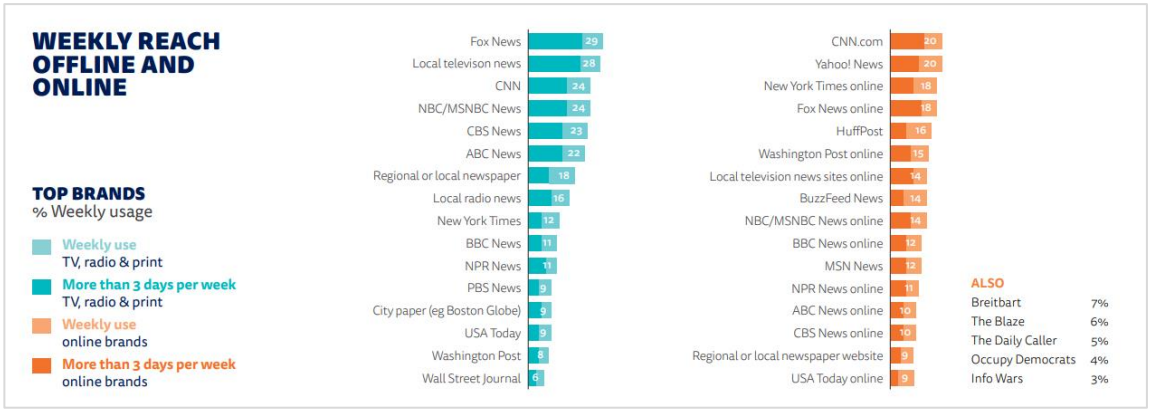


Figura 6. Consumo de noticias en los Estados Unidos. *Digital News Report 2020*

El estudio anterior contempla un número limitado de países en comparación a este trabajo. Por eso, se utilizó el sitio web *Prensa Escrita* (*www.prensaescritas.com*) como nueva fuente para completar el número total de medios contemplados. Este sitio proporciona una gran cantidad de enlaces de medios de prensa, distribuidos por regiones o subregiones alrededor del mundo (ver Figura 7).

Figura 7. Página inicio de la página web Prensa Escrita

A través de estas dos fuentes citadas, se identificaron y seleccionaron los medios que son parte de este trabajo. Las técnicas y métodos utilizados para la recopilación de las noticias se muestran en 3.1.3. El listado completo de los medios se puede visualizar en el Anexo B.

Esta necesidad de requerir medios trascendentales se sostiene en que los datos recopilados y analizados provengan de fuentes confiables, dándole un respaldo o una base en la cual se puedan sustentar los hallazgos encontrados y no así, de fuentes que se podrían catalogar como orígenes de noticias falsas (*Fake news)*, que podrían alterar o hacer dudar de los resultados.

4.1.2. ¿Cuál es la distribución de los medios por países y continentes?

La distribución de los medios de prensa se equilibró mediante dos grandes regiones, asociadas por su relación sociocultural: la primera compuesta por Norteamérica, Oceanía y Europa con un total de 79 medios; la segunda compuesta por Sudamérica y Centroamérica con los 88 medios restantes.

4.1.2.1. Distribución de medios de Norteamérica

Norteamérica, compuesta por un total de 18 medios (ver Tabla 2). En relación con los Estados Unidos, no se consideran los medios que provienen de origen latino (escritos en español). En Canadá, es posible encontrar medios escritos en francés e inglés.

Tabla 2. Distribución de medios de prensa en América del Norte

País	Cantidad de medios
Canadá	8
Estados Unidos	10

4.1.2.2. Distribución de medios de Centroamérica

En Centroamérica y considerando a México por su mayor cercanía o afinidad con el resto de estos países, se contabilizan 49 medios (ver Tabla 3).

Tabla 3. Distribución de medios de prensa en Centroamérica

País	Cantidad de medios
México	10
Panamá	3
El Salvador	5
Costa Rica	5
Cuba	5
Guatemala	5
Honduras	4
Nicaragua	4
Puerto Rico	3
República Dominicana	5

4.1.2.3. Distribución de medios de Sudamérica

Sudamérica (ver Tabla 4), contabilizando todos los países sin incluir territorios como Surinam, Guyana o Guyana Francesa, cuenta con un total de 39 medios.

Tabla 4. Distribución de medios de prensa en Sudamérica

País	Cantidad de medios
Brasil	5
Chile	5
Perú	6

Argentina	4
Bolivia	2
Uruguay	5
Paraguay	2
Ecuador	1
Venezuela	4
Colombia	5

4.1.2.4. Distribución de medios de Europa

En Europa (ver Tabla 5), se consideraron los países que cumplen con el requisito del idioma antes mencionado, contabilizando 49 medios en total. Se podría decir que estos son los países más importantes, salvo excepciones como Alemania u Holanda.

Tabla 5. Distribución de medios de prensa en Europa

País	Cantidad de medios
España	8
Francia	10
Irlanda	8
Reino Unido	8
Portugal	7
Italia	8

4.1.2.5. Distribución de medios de Oceanía

Por último, en Oceanía se contabilizan 12 medios (ver Tabla 6). En las Islas Pacíficas se consideran medios de países como: Islas Cook, Papúa Nueva Guinea, Polinesia Francesa y Samoa y Vanuatu.

Tabla 6. Distribución de medios de prensa en Oceanía

País	Cantidad de medios
Australia	4
Nueva Zelanda	2

Islas Pacificas	6

4.1.3. ¿Qué técnica(s) se utilizaron para el scraping?

La recopilación de noticias contó con el desarrollo de 150 scripts (uno por cada medio seleccionado) desarrollados en el lenguaje de programación Python, implementando *webs scraping* más el uso del lenguaje XPath. Estos scripts siguen una estructura base similar, pero varían ciertos parámetros dependiendo de cómo es la estructura (HTML) de las páginas de estos medios.

4.1.3.1. Aspectos generales

El script se basa en dos clases principales:

- *Crawler*, clase que se encarga de obtener URLs de noticias dado un enlace SEED.

  Este enlace, hace referencia a la página de un tema específico en un medio (por ejemplo: "*https://www.latercera.com/canal/mundo/*", donde el tópico es "mundo"). En este tipo de páginas (ver Figura 8), es posible visualizar el listado de las ultimas noticias relacionadas con esa materia. Usualmente, es posible encontrar estos enlaces sobre los títulos o imágenes de la noticia.

Figura 8. Sección internacional del medio La Tercera

  Es común que los medios dividan sus noticias en múltiples tópicos (por ejemplo: política, deportes, nacional, internacional, economía, social, etc.). Debido a esto, por cada medio se obtuvieron múltiples SEED (correspondientes a estos tópicos más comunes) para la recopilación de sus enlaces.

- La segunda llamada *Scraper*, permite recopilar los datos (título, fecha y texto) de esas noticias a partir de los enlaces obtenidos de la clase anterior.

4.1.3.2. Herramienta de desarrollo del navegador

La herramienta de desarrollo de los navegadores, son una extensión de ellos mismos. Estos permiten interactuar con los elementos de las páginas webs, así como otras múltiples características (por ejemplo, ver el código fuente de los sitios).

4.1.3.3. XML *Path Language* (XPath)

XPath es un lenguaje de consultas que permite seleccionar nodos dentro de un documento *XML*. Una expresión *XPath* es una cadena de texto que representa el recorrido en el árbol del documento.

4.1.3.4. Elaboración de expresiones XPath

Tanto para la clase *Crawler* (por enlaces de noticias) como para la clase *Scraper* (por el título, la fecha y el texto de las noticias) es necesario la elaboración de expresiones XPath por cada medio de prensa.

La elaboración de este tipo de expresiones, requieren de un análisis inicial para entender cómo se estructuran estas páginas webs. Utilizando la herramienta de desarrollo del navegador es posible visualizar y analizar esta estructura (código fuente) en tiempo real.

Por ejemplo, en la Figura 9 se muestra parte del código fuente del medio peruano La Primera (sección de noticias del mundo). En este caso, se requiere identificar dónde están los enlaces de noticias dentro del árbol de nodos. Por la figura, se observa que el enlace está ligado al título de la noticia

Figura 9. Código fuente de la sección internacional del medio La Primera

A partir de haber analizado este código fuente, es posible generar el árbol de nodos (etiquetas HTML) de esa página (ver en la Figura 10). La etiqueta *div* identificada con la

clase *td-ss-main-content* es el contenedor donde se colocan todas las noticias. Cada noticia está contenida en un bloque *div* con clase *td-block-row*. Siguiendo el árbol, se encuentra la etiqueta que hace referencia al título de la noticia (*h3*) que a su vez contiene el enlace buscado (etiqueta *a*).

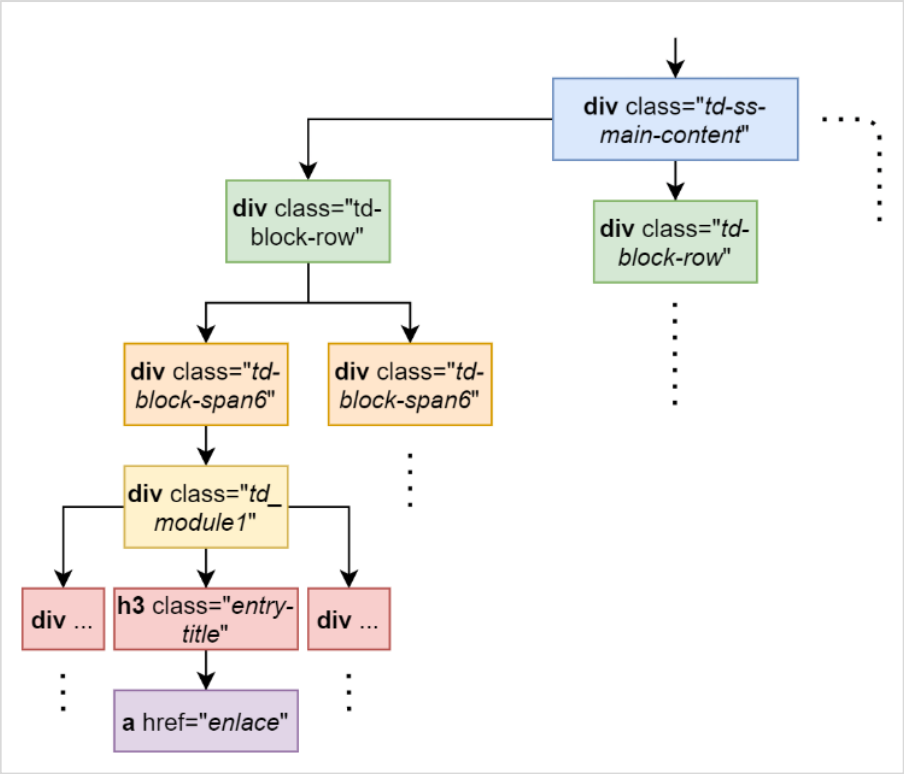


Figura 10. Árbol de nodos HTML pagina La Primera. Seccion internacional

En la figura 11 se encuentra la expresión XPath generada a partir del análisis anterior. Cómo es posible ver en esta expresión, no es necesario buscar en todo el contenido del documento. Por medio del doble carácter *// (slash)* se enfoca la búsqueda a partir de un cierto elemento (*td-ss-main-content*). A partir de este punto se podría seguir el árbol como se muestra en la figura anterior *(//div/div/div/div/h3/a/@href*), pero como existen caminos que lleven a enlaces que no sean noticias, se pueden volver a realizar un enfoque directamente a la etiqueta *h3* con el nombre de clase específico (*entry-title td-module-title*). Con el carácter @ (arroba) se selecciona el valor del atributo *href*, en este caso el enlace de noticia.

```
"//div[@class='td-ss-main-content']//h3[@class='entry-title td-module-title']/a/@href"
```

Figura 11. Expresión XPath para recopilar enlaces del medio La Primera de Perú

Esta expresión selecciona todos los elementos que coincidan con esta estructura. Como cada medio regularmente enlaza sus noticias dentro de bloques específicos, se pueden extraer la mayoría de sus enlaces de noticias.

Realizando este tipo análisis recién mencionado, se pueden generar expresiones XPath que permite encontrar los enlaces de noticias de cualquier medio (sitio web). También, por ese mismo método es posible identificar los elementos adicionales que se buscan recopilar por medio de la clase *Scraper*.

4.1.3.5. Clase *Crawler*

Generalmente, se utiliza el navegador para acceder a páginas webs (realizar consultas HHPT) (ver Figura X). Cuando esto ocurre, el servidor envía hacia el navegador la respuesta (HTML, imágenes, etc.), este procesa la respuesta y muestra el contenido.

Utilizando la librería *requests* de Python, la clase *Crawler* replica el ejemplo anterior y envía una consulta HTTP GET al servidor, pasando como parámetro la página SEED (ruta) (ver Figura 12). En este caso, la respuesta recibida es almacenada en una variable declarada en la clase.

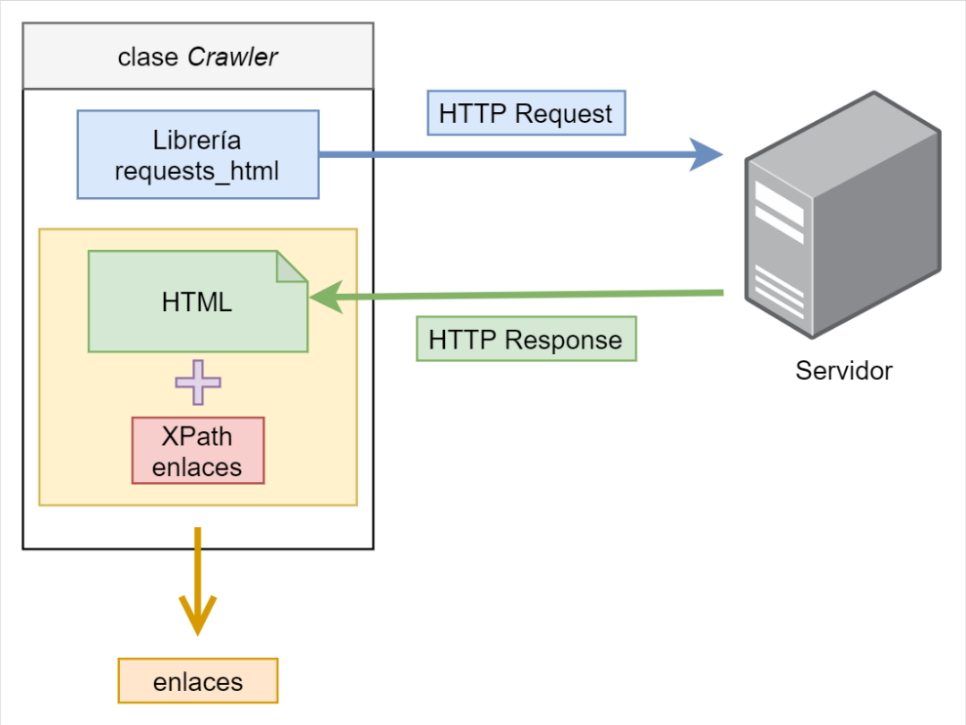


Figura 12. Diagrama de funcionamiento de la clase *Crawler*

En el siguiente paso, se extraen los enlaces de noticias desde objeto almacenado (HTML) utilizando la expresión XPath elaborada. En este punto final, se tiene un listado con enlaces esperando a ser procesadas por la siguiente clase.

Estas páginas "SEED" comúnmente tienen un número limitado de noticias a la vista, considerando que un medio tiene cientos de miles de páginas en su historia. Por lo cual, se hizo necesario iterar sobre esta para tratar de obtener todas las noticias. Regularmente

estas páginas contienen botones de “Cargar más” o “números de páginas”, lo que permite obtener casi la totalidad de estas noticias modificando levemente la URL “SEED” de acuerdo con lo que estos botones realizan.

4.1.3.6. Clase *Scraper*

La clase *Scraper* realiza un proceso similar a la clase anterior (consulta tipo GET y obtención del HTML de la página) (ver Figura 13), en este caso se recibe la URL de noticias como parámetro de entrada.

Para la obtención de los datos, se crean 3 expresiones Xpath encargados de la búsqueda y selección de cada uno de los datos a extraer.

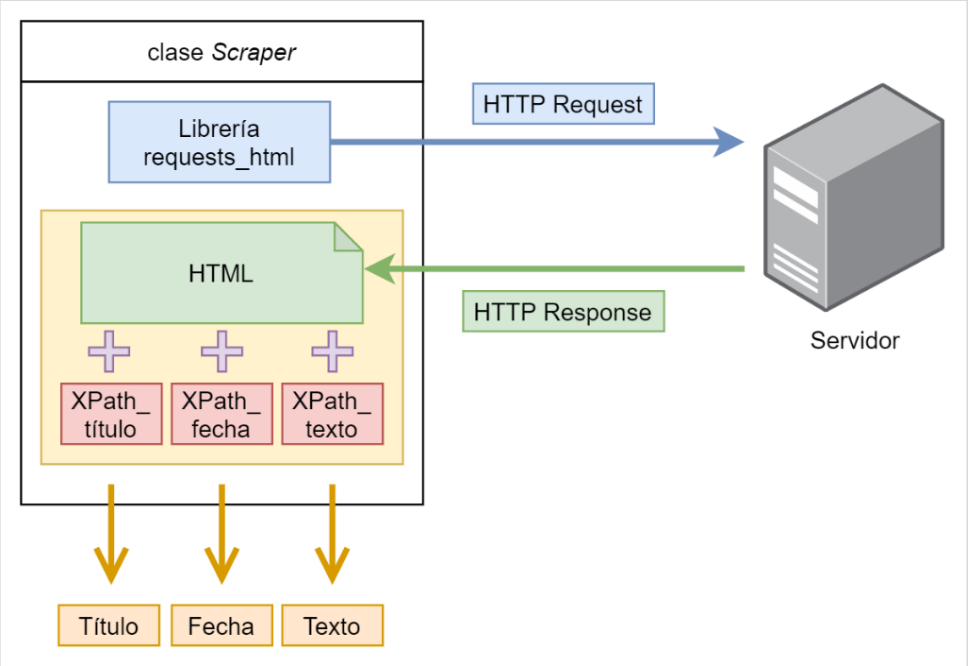


Figura 13. Diagrama de funcionamiento de la clase *Scraper*

4.1.4. ¿Cuáles son los principales problemas encontrados y solucionados?

Uno de los problemas que se encontró tiene relación con la recopilación de los enlaces de noticias. Por las formas dinámicas de algunas páginas web, el método de modificar cierta parte de las URLs no servía. Los que se realizó consistió en analizar el tráfico de red al momento de presionar el botón, a través de la pestaña de red de la herramienta de desarrollo del navegador, y detectar las respuestas de API de esta página o aplicación web. Esta respuesta comúnmente consistía en la devolución de un objeto tipo JSON con la información de las noticias más antiguas. Así, por medio de esta URL de la API, se volvió al método en la cual se modifica cierto parámetro de la URL para encontrar los enlaces más antiguos.

4.1.5. ¿Cuáles son las limitaciones? ¿Qué medios no se pudieron scrapear y por qué?

Existieron puntos que se tuvieron que abordar y que tienen relación a la exclusión de ciertos medios por motivos encontrados tanto en la búsqueda como en el desarrollo de los scripts.

Un aspecto fundamental de este proceso era identificar el nombre del periodista. Por otro lado, se pudo notar que existen medios que hacen referencia solo a la línea editorial o simplemente no aportan esta información requerida. Este hecho, implicó que un cierto número de medios, que podríamos catalogar "importantes", se excluyeran y por consiguiente no fueran parte de este trabajo. Esta situación se nota con un realce mayor en Sudamérica, donde se pueden citar los casos de Ecuador, Paraguay y Bolivia, países con menos de 3 medios que aportan este dato, y que dejó como resultado una distribución dispar de medios.

Otro punto por lo que se apartaron medios, tiene relación con que no se podían recopilar datos de noticias antiguas. Generalmente, estos solo mostraban noticias de un cierto periodo de tiempo hacia atrás (horas o días) o el tiempo desde la primera noticia publicada no era lo suficientemente antigua como era requerido.

Por otro lado, también existen otros medios que requieren cuentas personales, ya sea de carácter gratuito o que están sujetas a un servicio de suscripción, para tener acceso al contenido noticioso (ver ejemplo en Figura 14). Este hecho descrito, implica una mayor complejidad para acceder a los datos y hacer un gasto financiero, lo cual no se contempla por ahora.

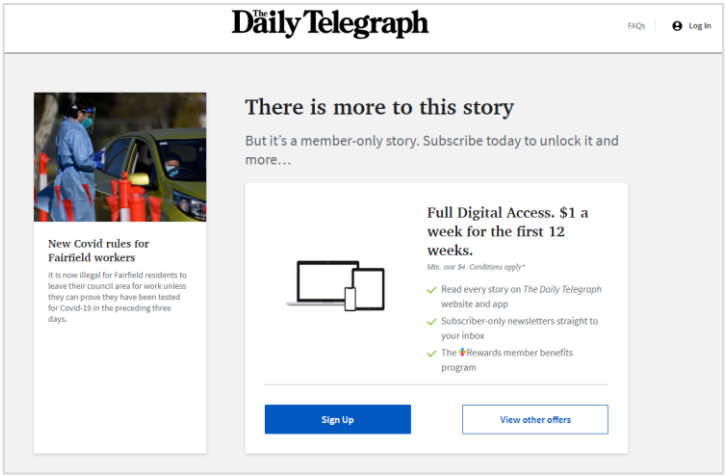


Figura 14. Noticia que requiere suscripción. The Daily Telegraph´

4.2. Extracción de datos sobre las fuentes de información

4.2.1. Preprocessing (Spacy + NER)

La extracción de nombres de personas, dentro de un texto, se realizó utilizando algunos componentes de la librería de Python Spacy (NER). Por ese motivo, se probaron diversos

modelos de esta librería y se seleccionaron aquellos que proporcionaban una relación entre las métricas de precisiones (ver Tabla 7).

Tabla 7. Información de los modelos de Spacy

Idioma	Nombre de modelo	Métricas (NER)		
		Precision	Recall	F-score
ingles	en_core_web_lg	0.85	0.85	0.85
español	es_core_news_lg	0.90	0.90	0.90
portugués	pt_core_news_lg	0.90	0.90	0.90
italiano	it_core_news_lg	0.88	0.88	0.88
frances	fr_core_news_lg	0.84	0.84	0.84

En la Figura 15 se puede observar el funcionamiento del extractor de nombres de personas. Este script (por idioma) recibe el texto (artículo de noticia) y por medio del modelo se obtienen las entidades. En la etapa de procesamiento se realiza un filtrado para obtener solo los datos requeridos, eliminando cualquier tipo de duplicaciones. El resultado es una lista de nombres citados en la noticia.

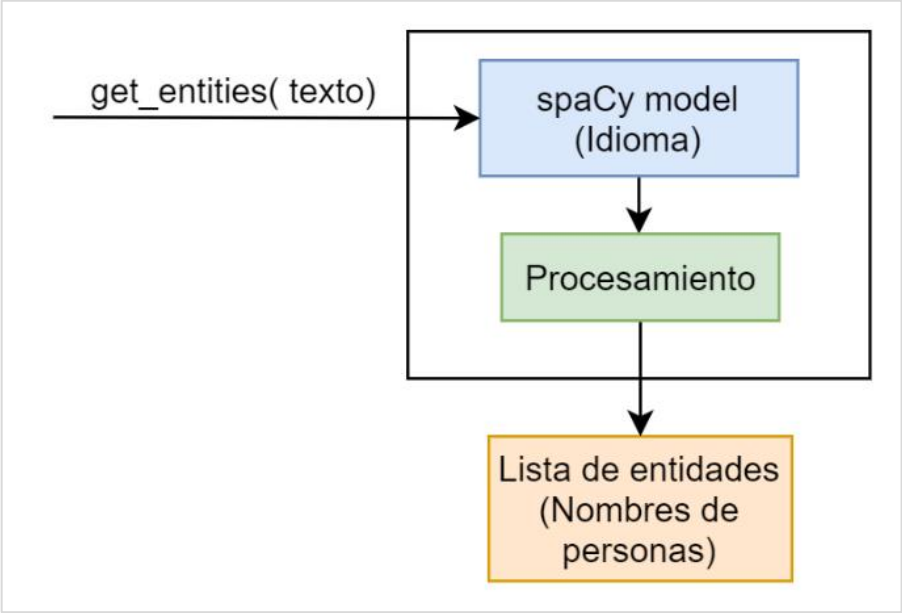


Figura 15. Diagrama de extracción de nombres de personas

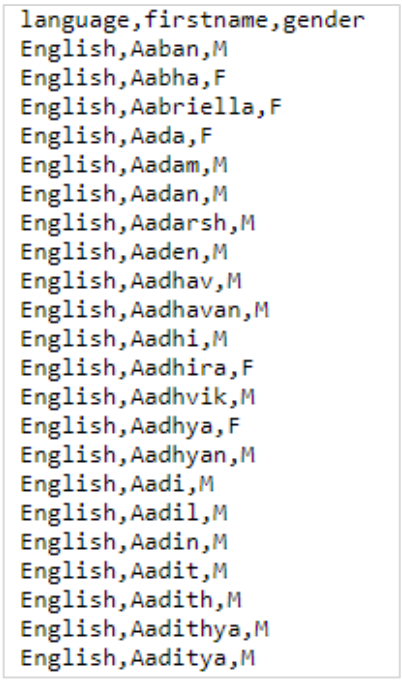
4.2.2. Clasificación de género con diccionarios de nombre

La clasificación del género de los nombres consistió en la búsqueda de grandes conjuntos de datos, asociados a registros de organizaciones gubernamentales (ver Tabla 8), con el objetivo de crear un diccionario global (*dataset*) que actúe como un clasificador. Estos conjuntos de datos requerían tener como mínimo: el nombre de personas y su género.

Tabla 8. Propósito y fuentes de los datos de nombres utilizados

Idioma	Propósito de dataset	Fuente
Ingles	Nombres populares de bebés de EE. UU.	Administración del seguro social de los EE. UU (SSA).
Español	Análisis y estudios demográficos de España	Instituto Nacional de Estadística de España (INE).
portugués	Clasificación por género en nombres brasileños, con base en datos del CENSO DE 2010	Instituto Brasileño de Geografía y Estadística (IBGE).
Italiano	Registro de administradores locales y regionales (1985-2014)	Departamento de Asuntos Internos y Territoriales del Ministerio del Interior de Italia (DAIT).
Francés	Registro de nombres de bebés franceses desde 1900	Instituto Nacional de Estadística y Estudios Económicos de Francia (INSEE).

Para esto, se extrajeron los datos de las fuentes antes mencionada y a través de una etapa de procesamiento de estos datos, donde se eliminaron datos innecesarios o se formateo información contenida, se generó un diccionario (CVS) considerando como columnas *language*, *first name* y *gender* (ver Figura 16).



```
language,firstname,gender
English,Aaban,M
English,Aabha,F
English,Aabriella,F
English,Aada,F
English,Aadam,M
English,Aadan,M
English,Aadarsh,M
English,Aaden,M
English,Aadhav,M
English,Aadhavan,M
English,Aadhi,M
English,Aadhira,F
English,Aadhvik,M
English,Aadhya,F
English,Aadhyan,M
English,Aadi,M
English,Aadil,M
English,Aadin,M
English,Aadit,M
English,Aadith,M
English,Aadithya,M
English,Aaditya,M
```

Figura 16. Nombres clasificados por su género. Extracto del archivo CSV

El conjunto de datos cuenta aproximadamente con 200 mil nombres de personas (ver Tabla 9). Además, aproximadamente se tiene la misma cantidad de nombres masculinos y femeninos resultando en un conjunto equilibrado.

Tabla 9. Resumen de nombres por género e idioma

Idiomas	**Nombres**	
	M	**F**
ingles	12.917	17.543
español	25.154	24.179
portugués	23.952	26.791
italiano	23.923	8.914
francés	15.304	18.377
Subtotal	101.250	95.804
Total	197.054	

Para realizar la clasificación, se desarrolló un script que recibe como parámetros el idioma origen del texto y el nombre de la persona (ver Figura 17). Esta función carga el dataset

mencionado y busca si existe alguna correspondencia. En caso de que la información ingresada se encuentre, se retorna el valor que se encuentre en la columna gender (M o F). De no encontrarse, se retorna un valor nulo.

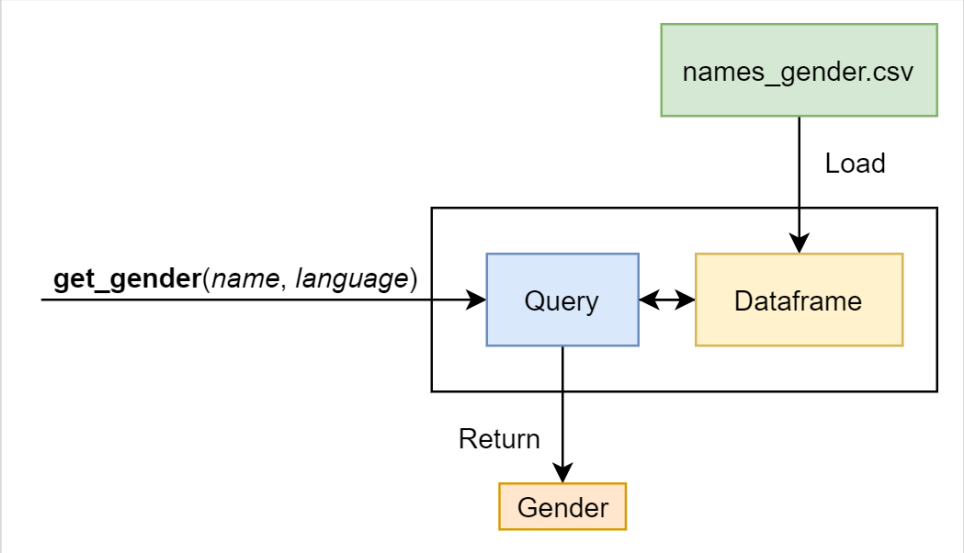


Figura 17. Diagrama de función que retorna el género de un nombre

4.2.3. Script de llamada a la API de Wikipedia

Extraer y recopilar datos de personas (fuentes de información), como se verá en el punto siguiente (3.2.4), implica identificar una fuente (texto) donde buscar tal información. Una enciclopedia que abarca gran parte de los contenidos que existen en internet es Wikipedia.

Generalmente, artículos asociados a personas en Wikipedia proporcionan datos personales de estas mismas. Es común que dentro del resumen (primeros párrafos) se encuentren los elementos que en este trabajo se requieren identificar y recopilar (año de nacimiento, nacionalidad, profesión) (ver ejemplo Figura 18).

Figura 18. Artículo de persona en Wikipedia

Considerando lo anterior, se desarrolló un script en Python que utiliza la librería *requests* (mencionada en 3.1.3) para consumir la API de Wikipedia (ver Figura 19). Este script realiza una petición HTTP tipo GET al punto de entrada a la API, adicionando algunos parámetros necesarios para este caso específico:

https://es.wikipedia.org/w/api.php?titles=Lionel_Messi&action=query&format=json&prop=extracts&exchars=1200&exlimit=20&explaintext&exintro

Entre los parámetros importantes están:

- titles: Aquí se pasa el nombre de la persona buscada. Ejemplo: "*Lionel_Messi*".
- format: Indicamos el formato de entrega de la información (*json*)
- prop: Indicamos qué información se requiere del artículo. En este caso se especifica el extracto (*extracts)*.

Además, es posible cambiar el idioma base de Wikipedia ("es", "en", "it", "fr" o "pt") en esta URL, permitiendo recopilar artículos en múltiples idiomas de origen.

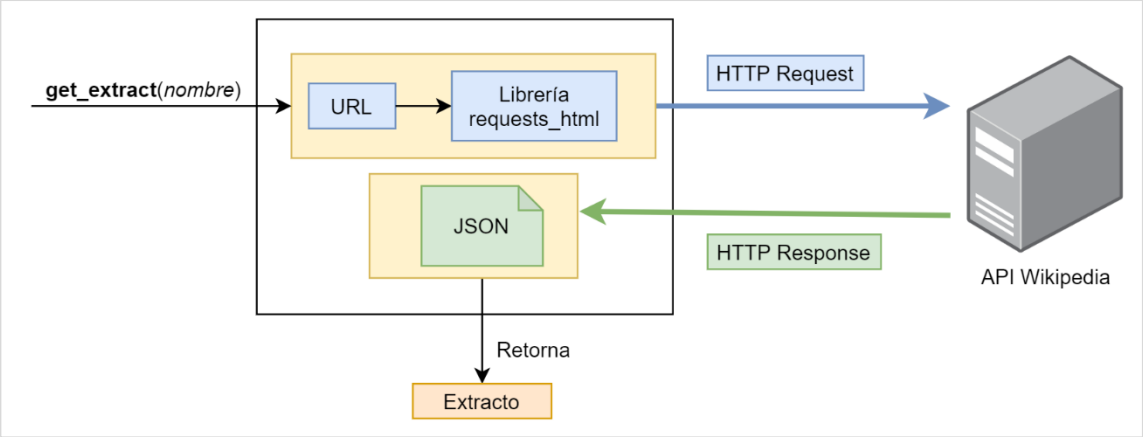


Figura 19. Diagrama funcionamiento de consumo de la API de Wikipedia

Al ejecutar esta petición, desde la API se devuelve el objeto con la información solicitada (ver Figura 20). Manipulando este objeto, se extrae y se almacena la información buscada (*extract*).


```
{
    "batchcomplete": "",
    "query": {
        "normalized": [
            {
                "from": "Lionel_Messi",
                "to": "Lionel Messi"
            }
        ],
        "pages": {
            "185205": {
                "pageid": 185205,
                "ns": 0,
                "title": "Lionel Messi",
                "extract": "Lionel Andr\u00e9s Messi Cuccittini (Rosario, Santa Fe; 24 de junio de 1987), conocido como Leo Messi, es un futbolista argentino que juega
            }
        }
    }
}
```


Figura 20. Respuesta de la llamada a la API de Wikipedia

4.2.4. Question-Answering con Transformers

Habiendo obtenido una fuente (extracto de Wikipedia) de donde extraer datos de personas, el siguiente paso es obtener dicha información. Para esto, se hizo uso Qustion-Answering, tareas proporcionadas por algunos de los modelos Transformers pertenecientes a la plataforma HuggingFace.

Identificar cúales modelos cumplen con las condiciones dadas (permiten realizar la tarea mencionada), requirió una búsqueda directa desde el sitio web que proporciona HuggingFace (*https://huggingface.co/models*). Por cada idioma, se identificaron y probaron los modelos que son parte de este paso (ver Tabla 10).

Tabla 10. Modelos de Transformers para Question-Answering

Idioma	Nombre de modelo	F-Score
ingles	DistilBERT base cased distilled SQuAD	87.1
español	BETO (Spanish BERT) + Spanish SquAD2.0	86.0781
portugués	Bert base Portuguese cased finetuned on SquAD	
italiano	Italian BERT fine-tuned on SquAD	74.16
francés	Camembert base squadFR fquad piaf	

Se desarrolló un script el cual hace uso de estos modelos (ver Figura 21). Haciendo uso de un pipeline proporcionado por la librería, se carga el modelo indicando la tarea que debe resolver ("Question-Answering").

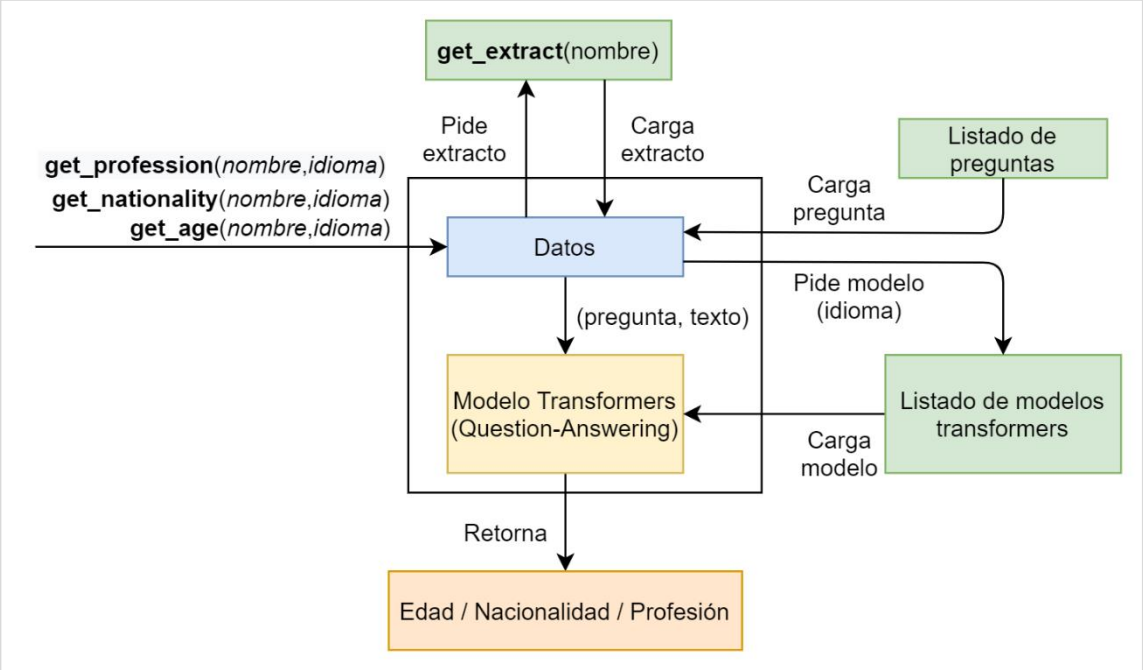


Figura 21. Diagrama de la extracción de profesión, nacionalidad y edad de entidades

Para el uso, es necesario contar con dos argumentos. El primero es el contexto, el extracto (3.2.3) y la pregunta que se quiere realizar. Para lo último, se definieron una serie de

preguntas por idiomas (ver Tabla 11), que están relacionadas con identificar la información que se está buscando.

Tabla 11. Preguntas de entrada para los modelos Question-Answering

Idioma	Pregunta a realizar
ingles	What year was she or he born?
	What is her or his profession?
	What is her or his nationality?
español	¿Cuál es su profesión?
	¿En qué año él o ella nació?
	¿Cuál es su nacionalidad?
portugués	O que é ela ou a sua profissão?
	Em que ano nasceu ele ou ela?
	Qual é sua nacionalidade?
italiano	Qual è la sua professione?
	In che anno è nato?
	Qual è la tua nazionalità?
francés	Quelle est sa profession ?
	En quelle année est-il né ?
	Quelle est sa nationalité ?

4.2.5. Popularidad de personas

Para la obtención de la popularidad de las personas, se utilizó como referencia las visitas en sus páginas de Wikipedia. Por eso, se generó un módulo que utiliza una librería llamada *pageviewapi*. Esta librería retorna una lista de visitas por día de cualquier página existente en Wikipedia.

Así, este módulo recibe como parámetro el nombre de la persona y mediante esta librería se obtienen las visitas diarias, las cuales se promedian (por mes) para obtener un valor referencial.

4.2.6. Nombres de periodistas

La identificación de los periodistas, son un elemento que se contempló en la etapa de *scraping* de las noticias. Por eso, mediante el mismo proceso por el que se obtuvieron datos anteriores como el texto, título y fecha de noticias (mencionado en este mismo capítulo), se logró extraer el nombre del autor del texto.

5. RESULTADOS

5.1. Noticias de prensa

5.1.1. Resumen de la recopilación de noticias

Hasta esta fecha, la recopilación de noticias de prensa (ver en la Tabla 12) contempla en la base de datos más de 17 millones de noticias. Este conjunto de datos proviene de medios de 30 países distintos aproximadamente. Como adicional, en este resumen se consideran las noticias de medios incluido en el trabajo enteramente citado y otros medios adicionales. El conjunto total se divide en 216 medios diferentes, siendo Francia e Italia los países que más medios aportan a esta colección, Reino unido como el país que más noticia aporta y finalmente, Ecuador como el país del que menos noticias fueron recopiladas (de solo un medio).

Tabla 12. Cantidad de noticias por país

País	Cantidad de medios	Cantidad de noticias
Reino Unido	10	1826226
Canadá	8	1415989
Estados Unidos	8	1328939
España	10	1228464
Chile	7	1098397
Brasil	7	1095610
Irlanda	9	1010282
Argentina	7	795753
Nueva Zelanda	3	786350
República Dominicana	4	760244
Francia	12	678972
China	1	659437
Nigeria	2	584257
Australia	8	541364

Italia	11	528599
Perú	4	523554
Costa Rica	4	419698
México	11	310153
Nicaragua	4	303184
Colombia	4	190707
Guatemala	5	185441
Portugal	7	177454
Cuba	5	173541
Uruguay	5	135869
El Salvador	5	126144
Venezuela	3	96348
Honduras	2	78296
Panamá	3	43777
Puerto Rico	3	43624
Vanuatu	1	31981
Islas Cook	1	28441
Papua Nueva Guinea	1	22416
Samoa Americana	1	20419
Paraguay	1	12357
Polinesia Francesa	1	9864
Ecuador	1	8051
Total		17280202

5.1.2. Cantidad de noticias por idioma

La distribución de las noticias por el idioma (ver Figura 22) tuvo como resultado que, aproximadamente el 45% de las noticias fueron escritas en inglés. Como se puede analizar con la tabla anterior, los países anglosajones aportan las mayores cantidades de noticias. El español, idioma con el 38% aproximado de noticias, consecuencia de la gran cantidad de medios de países de habla hispana incluidos. Con porcentajes mucho menores se encuentran el portugués, francés e italiano. Estos idiomas son menos utilizados (cantidad de países) y era de esperarse estos bajos porcentajes.

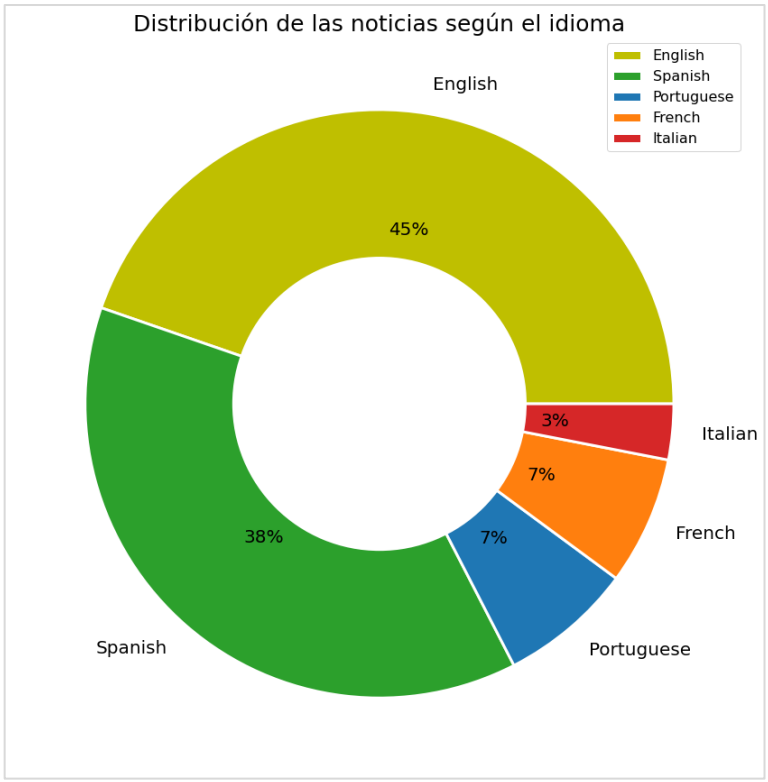


Figura 22. Distribución de noticias por idioma

5.1.3. Cantidad de noticias según el año

Un aspecto interesante de analizar es la evolución de la digitalización de los medios escritos. En la Figura 23, se muestra la evolución en el tiempo de la cantidad de noticias publicadas por año. En entre los años 2000 y 2020, existe un claro aumento, consistente en el tiempo, de noticias publicadas en general. Este resultado podría interpretarse o de algún modo dar un indicio del proceso tecnológico de esta época. Colocando números, en el año 2000 la cantidad de noticias no supera las 100 mil. Ya para el 2020, este número supera con creces los 2 millones.

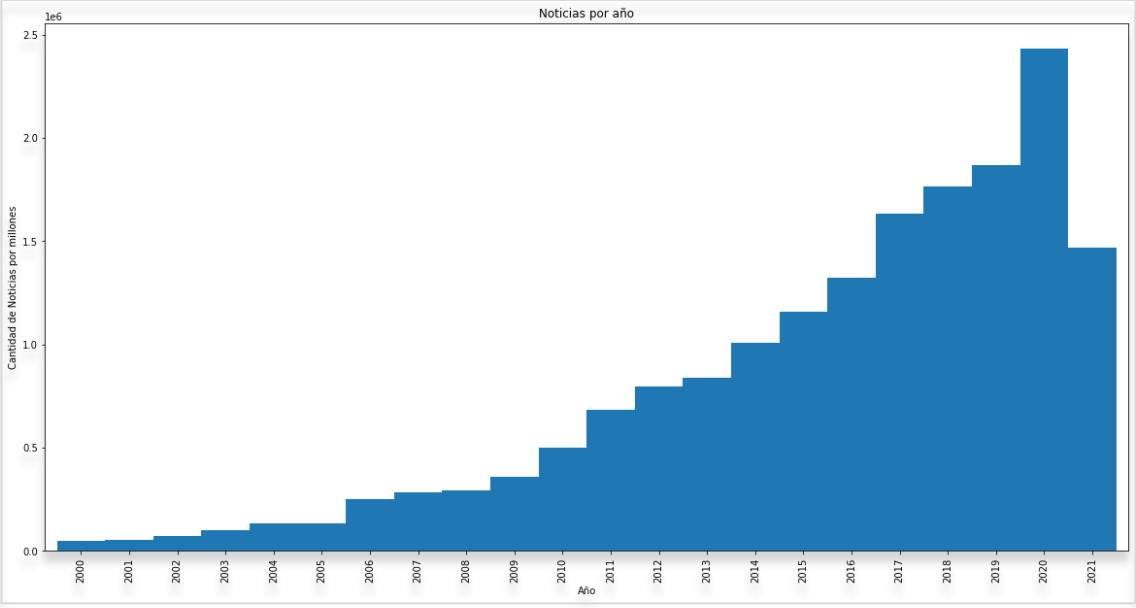


Figura 23. Cantidad de noticias por año

5.2. Información sobre las fuentes de información

5.2.1. Resumen relacionado a las fuentes de información

La extracción de fuentes de información contó con aproximadamente 5 millones de registros (ver Tabla 13). De ese total, se logró clasificar correctamente un casi un 80% de estos datos por género. Como resultado, se puede sintetizar que las fuentes de información del género masculino, con más de la mitad del total, son más expuestos o visibles comunicacionalmente. Por otro lado, el género femenino no logra ni la mitad de visibilidad en comparación con el género opuesto.

Tabla 13. Cantidad de fuentes según su género

Género	Cantidad	Porcentaje
Femenino	1088983	23.1%
Masculino	2580256	54.8%
Nulo	50453	1.1%
Sin Información	989306	21%
Total	4708998	100%

Como detalle, existe un porcentaje no menor de datos que no fueron clasificados, esto se debe principalmente a que:

- No existe el nombre en el diccionario de datos

- El dato no tiene relación con el nombre de una persona (errores en el procesamiento).

5.2.2. Fuentes de información según el año de su nacimiento

En la Figura 24, se muestra la pirámide de edad (en relación con su año de nacimiento) de las fuentes de información por cada género. En cuanto a los varones, es posible analizar que se puede notar que las personas más citadas están en el rango de 170 y 460 años. A su vez, las mujeres más citadas entre los 200 y 350 años, un rango mucho más pequeño y proporcionalmente con menores registros en cuanto a la cantidad de personas.

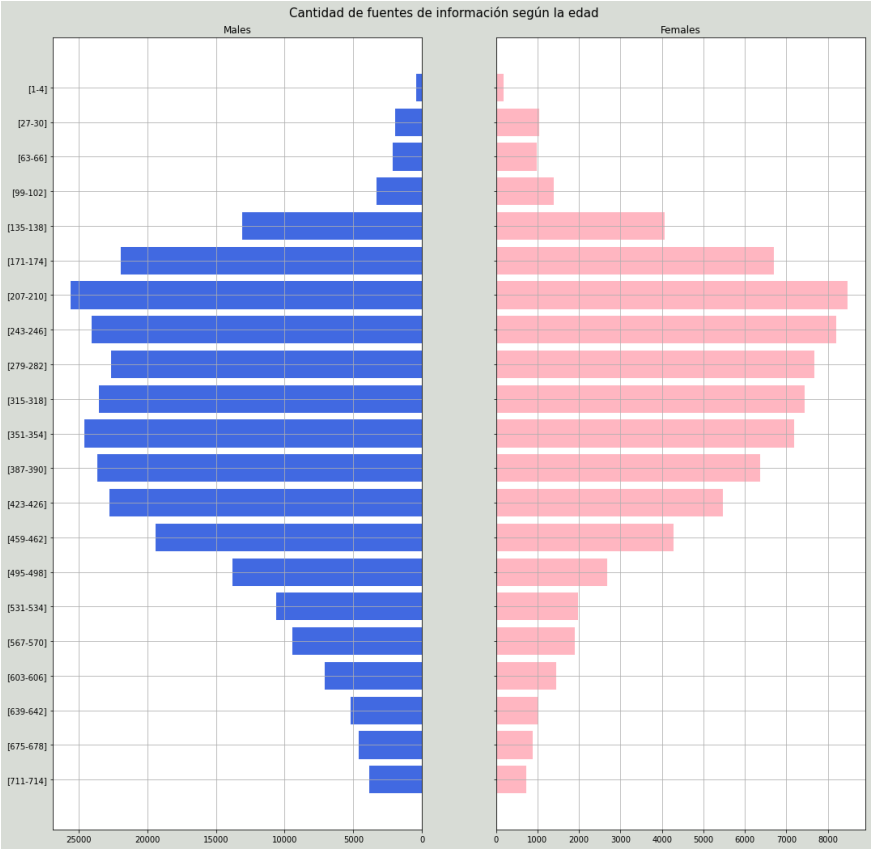


Figura 24. Cantidad de fuentes de información según el año de nacimiento

5.2.3. Profesiones más frecuentes de las fuentes de información

Abordando otros tipos de datos, nos encontramos con las profesiones de estas fuentes de información. De la tabla 14, se puede notar que las personas de género masculino con profesiones relacionadas con la política y el deporte son las más visibilizadas en los medios. En el caso de las mujeres, el ser actriz o política les proporciona una mayor en ese sentido.

Tabla 14. Profesiones más frecuentes por género (Top 10)

Profesiones por género			
Hombres	**Cantidad**	**Mujeres**	**Cantidad**
Footballer	17119	actress	4735
Politician	9116	politician	3032
Actor	7008	journalist	1628
Coach	5302	writer	1433
Futbolista	5247	footballer	1406
Journalist	3356	actriz	1291
Lawyer	3113	author	1149
businessman	2713	singer	922
Writer	2713	artist	836
Nulos	2202202	Nulos	980553

5.2.4. Nacionalidades más frecuentes de las fuentes de información

Las nacionalidades más frecuentes encontradas (ver Tabla 15), se relacionan de algún modo con la cantidad de noticias por idioma. Como se analizó anteriormente, los medios anglosajones tenían los mayores aportes de noticias. Este hecho también se puede relacionar con que las nacionalidades más frecuentes encontradas provienen de este tipo de lugares, siendo las personas de Estados Unidos, con un gran porcentaje, las fuentes de información más utilizadas.

Tabla 15. Nacionalidades más frecuentes (Top 10)

Nacionalidad	**Cantidad**
American	94393
English	23007
British	21496
Australian	19261
Canadian	17466

Italiano	15303
Español	10938
Chinese	9234
Irish	8870
Nulos	4163282

5.2.5. Fuentes de información más mencionadas

En cuanto a las fuentes de información más mencionadas (ver Tabla 16), es posible observar que mayoritariamente los políticos (jefes de estados) son las personas que tienen mayor repercusión en los medios y que son considerados como fuentes de información.

Tabla 16. Personas más mencionadas (Top 9)

Nombre	Cantidad
Donald Trump	218831
Barack Obama	78619
Pedro Sánchez	68289
Boris Johnson	65780
Martina Manfredi	58481
Nicolás Maduro	58214
Joe Biden	50376
Mariano Rajoy	49139
Emmanuel Macron	46529

5.2.6. Popularidad de fuentes de información

De los datos recopilados, se pueden realizar análisis como el que aparece en la Figura 25. La relación de popularidad entre los 2 candidatos a la presidencia de Chile (2021). Aquí se pueden observar los momentos de más relevancia en ese periodo de tiempo.

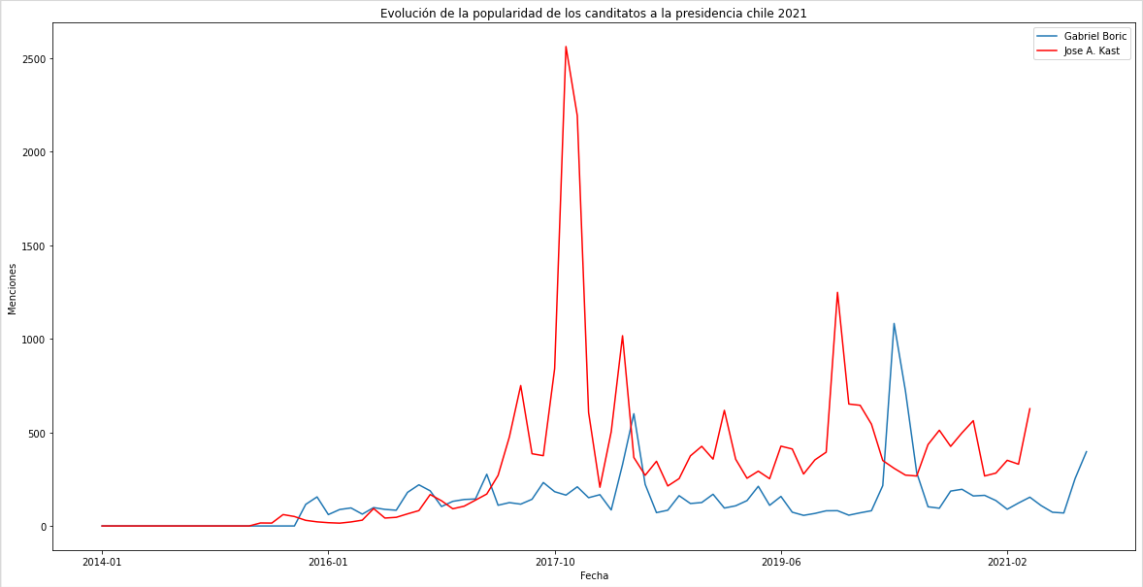


Figura 25. Ejemplo de evolución de popularidad de fuentes de información

6. CONCLUSIÓN

6.1. Una ampliación importante

El trabajo realizado logró ampliar significativamente lo realizado en la investigación anterior. En términos generales, el desarrollo e implementación de scripts y la contemplación de toda una arquitectura de software dedicada (como soporte), permitió y permite actualmente generar flujos de recopilación y procesamiento de datos que se podrían considerar procesos casi industriales (millones de datos). Los resultados de estos flujos aplicados permitieron realizar una exploración más exhaustiva y precisa de los elementos de interés, reflejando de mejor forma la realidad.

La retroalimentación y nuevas consideraciones aportadas por los revisores de la revista científica, ayudó también a discutir e implementar desde otros enfoques el desarrollo del trabajo. Este matiz aportado externamente, proporciona un mayor respaldo al trabajo realizado y a las conclusiones obtenidas.

6.2. Algunas limitaciones en la calidad de los datos

A pesar de que los métodos aplicados para extraer y procesar los datos son bastantes avanzados, no se puede olvidar que todavía no es posible asegurar una confiabilidad absoluta. Como se explicó en los capítulos anteriores, los modelos utilizados no son 100% precisos y esto trae como consecuencia la recopilación de datos “basura” o ambiguos, lo que conlleva para más adelante una etapa de post procesamiento.

Los modelos Question-Anwering reciben el texto de una página (de una persona) de Wikipedia. A pesar de que la información que en este trabajo se recopila (año de nacimiento, profesión, etc.), en ciertos casos esta información no se encuentra en el formato deseado. En ese mismo sentido, una profesión similar puede ser escrita de distintas maneras (por ejemplo: “jugador de fútbol”, “futbolista”) o puede ser el caso de que una persona tenga múltiples profesiones (por ejemplo: “Político, Cientista y abogado”). Estas problemáticas implican que, al momento de clasificar este tipo de datos, los conjuntos resultantes sean mucho mayores a los que en teoría debieran obtenerse.

Otro elemento limitante es la clasificación por género. El hecho de utilizar diccionarios implica la no clasificación de elementos menos frecuentes. Por ejemplo, un nombre poco frecuente (nombres extranjeros) o escritos levemente diferente a los comunes (“Christian” en vez de “Cristian”) podrían no encontrarse en este tipo de diccionario. También, se podrían generar ambigüedades en nombres que utilizan ambos géneros (Andrea, nombre femenino en Latinoamérica, pero, masculino en Italia.).

6.3. Potencial de ampliación de la investigación

En términos generales, la recopilación de datos extraídos de internet se podría decir que no tiene limitaciones. Como se pudo observar, existe un claro aumento o interés en la digitalización de contenidos noticiosos por parte de los medios de prensa. La cantidad de información o datos relacionados son cada vez más frecuentes y disponibles para el interés público. Estos hechos permiten ampliar la investigación no solo considerando otros países del mundo (incluyendo nuevos idiomas), sino que además se podría incluir ciertas regiones mucho más específicas. Por ejemplo, recientemente el grupo Sophia2 participó en un proyecto específico donde se incluyeron datos de medios (20) de la región de los Lagos.

Además, los constantes avances relacionados con procesamiento de texto no solo permitirían mejorar la calidad de datos en bruto (actualizando o integrando nuevos modelos de procesamiento de texto) en un futuro, sino que además se podría mejorar el rendimiento en la extracción (considerando modelos más rápidos). En ese mismo sentido, la arquitectura de software utilizada está diseñada y preparada para integrar dichas mejoras de manera de poder escalar en el tiempo.

7. DESARROLLO PENDIENTE

Como desarrollo pendiente, queda la terminación del proceso de *scraping* de los nombres de los periodistas. Una vez finalizado este proceso, la obtención de datos como el género, se puede realizar utilizando los mismos módulos generados para las fuentes de información. Otros metadatos adicionales buscados, se pueden obtener solo generando consultas a la base de datos.

8. REFERENCIAS

Benoit K., Watanabe K., Wang H., Nulty P., Obeng A., Müller S. & Matsuo A. (2018). *quanteda: An R package for the quantitative analysis of textual data*. Journal of Open-Source Software, 3(30), 774.

Bhagat, K. K., Mishra, S., Dixit, A., & Chang, C. Y. (2021). *Public Opinions about Online Learning during COVID-19: A Sentiment Analysis Approach*. Sustainability, 13(6), 3346.

Blei, D. M. and Lafferty, J. D. (2006). *Dynamic topic models.* In Proceedings of the 23rd International Conference on Machine Learning, pages 113–120.

Boumans, J. W., & Trilling D. (2016). *Taking stock of the Toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars*. Digital Journalism 4 (1): 8–23.

Eshbaugh-Soha, M., & McGauvran, R. J. (2017). *Presidential Leadership, the News Media, and Income Inequality*. Political Research Quarterly, 71(1), 157–171.

Feinerer I., Hornik K. & Meyer D. (2008). *Text Mining Infrastructure in R*. Journal of Statistical Software, 25(5), 1–54

Grimmer, J., & Stewart B. M. (2013). *Text as data: The promise and pitfalls of automatic content analysis methods for political texts*. Political Analysis 21 (3): 267–297.

Han, S., & Anderson, C. K. (2020). *Web Scraping for Hospitality Research: Overview, Opportunities, and Implications*. Cornell Hospitality Quarterly, 62(1), 89–104.

Madrid-Morales, D. (2020). *Using Computational Text Analysis Tools to Study African Online News Content*. African Journalism Studies, pages 1–15.

Massimino, B. (2016). *Accessing online data: Web-crawling and information-scraping techniques to automate the assembly of research data*. Journal of Business Logistics, 37(1), 34–42.

Medhat, W., Hassan, A. & Korashy, H. (2014). *Sentiment analysis algorithms and applications: A survey*. Ain Shams Eng. J. 2014, 5, 1093–1113.

Ruiz, F. (2020). *Evolution of the gender biases in the news media: a computational method using dynamic topic model* [Tesis del Magíster en Informática]. Universidad Austral de Chile.

Silge J. & Robinson D. (2016). *tidytext: Text Mining and Analysis Using Tidy Data Principles in R*, Journal of Open-Source Software, 1(3), 37.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). *Attention is all you need*. CoRR.

Welbers, K., Van Atteveldt W., & Benoit K. (2017). *Text Analysis in R.* Communication Methods and Measures 11 (4): 245–265.

Wickham H. (2016*). ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. ISBN 978-3-319-24277-4

ANEXO

Anexo A: Correo electrónico de respuesta de Feminist Media Studies

Feminist Media Studies - Decision on Manuscript ID RFMS-2020-0048
Feminist Media Studies <onbehalfof@manuscriptcentral.com> 6 de julio de 2020, 14:39
Responder a: CarterCL@cardiff.ac.uk
Para: mvernier@inf.uach.cl
06-Jul-2020
Dear Dr Vernier:
Your manuscript entitled "Do women exist in Chilean press?: Gender of Journalists and Sources" which you submitted
to Feminist Media Studies, has been reviewed. The reviewer comments are included at the bottom of this letter.
Our reviewers have raised substantial concerns, and therefore your paper cannot be accepted for publication in
Feminist Media Studies in its current form. However since they also find much to merit in the paper, I would be willing
to reconsider if you wish to undertake major revisions and re-submit, addressing the referees' concerns.
Please note that resubmitting your manuscript does not guarantee eventual acceptance, and that your resubmission
will be subject to re-review before a decision is rendered.
You will be unable to make your revisions on the originally submitted version of your manuscript. Instead, revise your
manuscript using a word processing program and save it on your computer. Once you have revised your manuscript, go to https://mc.manuscriptcentral.com/rfms and login to your Author
Centre. Click on "Manuscripts with Decisions," and then click on "Create a Resubmission" located next to the
manuscript number. Then, follow the steps for resubmitting your manuscript.
Because we are trying to facilitate timely publication of manuscripts submitted to Feminist Media Studies, your revised
manuscript should be uploaded within 9 months, by 02-Apr-2021.
I look forward to receiving your resubmission by the date noted above.
Sincerely,
Cindy
Cynthia Carter & Isabel Molina Guzman
Co-Editor, Feminist Media Studies

CarterCL@cardiff.ac.uk, imolina@illinois.edu
Reviewer(s)' Comments to Author:
Reviewer: 1
Comments to the Author
The article is an interesting piece that allows us to better understand the situation of Chilean journalism in regard
gender issues, in particular in topics such as authorship, sources cites and the relationship between gender of
journalists and gender of the sources that they cited. The topic is relevant to the scope of a publication such as FMS
and partly account for a significant research gap in this country. However, the following revisions might further
strengthen the paper.
1. I am not sure if the study reflects the reality of the entire Chilean press. The way the authors reports their
findings implies that data that they analyzed represents all the media in Chile. The title, the RQs and the findings are
presented using the expression "the news in Chile" or the "Chilean press".

However, as the author recognized, the study is based on 16 digital outlets out of 500 different media channels.
According to Mellado et al., (2011), there were about 4,283 active journalists working on different media channels in
the 4 most important regions of Chile almost a decade ago. This figure may be higher now with the increase of digital
channels in the last decade. The study only analyzed the work of 158 journalists.
The time frame is also an issue. The study only covers two months in a specific period of time (June-July). So, the
information of most of the year remain uncover.
Therefore, the authors should reframe the way that they present their findings. They should refer to the specific
sample of the study and not the entire Chilean news media.
2. One of the main purposes of research is to advance theory. This is particularly relevant in high quality journal
such as FMS. Therefore, something that is missing from the article is more discussion about the theoretical
implications of the findings. On page 18, in the first paragraph there is a reference to Nicolson´s theory related to work

spaces and gender equality. However, this is a small reference to a topic that deserve much greater analysis from the
abundant theorizing of fields such as critical theory, feminist theory and gender theory.
3. Finally, something that is also missing in the article is to include a theoretical perspective from the Global South
and from Latin American, in particular. There have been a strong criticism of research projects/publications that frame
their studies/projects from authors/ideas that come exclusively from the Global North (Biegel, 2013, 2014;
Chakravartty et al., 2018; Merton, 2018, 2019; Hanitzsch, 2019; Mosbah-Natanson & Gingras, 2014; Waisbord &
Mellado &, 2014). It is implicit that the rigorous and serious knowledge is only occurs in Westerns countries (Mignolo,
2010).
90% of the authors and studies cited in the article belong to researchers from Western countries. This could be
questioned in a study that focuses on the reality of a country from the Global South. I think the article might be further
improve if the authors try to include not only studies, but also theoretical ideas and assumptions that come from Latin
American authors. Theoretical concepts such as patriarchalism, Marianism or ideas that come from authors related to
feminism and decolonization (Catherine Walsh, María Lugones, Ochy Curiel, Rita Segato and Silvia Rivera
Cusicanqui), could be also important in the article.
Reviewer: 2
Comments to the Author
Gender inequality in the workplace is a significant concern. Especially within the newsroom, since inequality in these
spaces has direct implications for women's representation in the media, which renders the object of this research a
highly relevant issue. However, the work submitted presents some problems that prevent me from giving it the green
light for publication, as described below:
1. The revision of the literature is scarce. Moreover, the authors stated that there is no existing literature exploring the
relationship between the journalist's gender and the source's gender. However, this issue has been addressed

previously by several academics (Craft & Wanta, 2004; Zeldes & Fico, 2009; Ross, 2007). Armstrong (2004) found
that "the influence of the gender of the reporter has a bearing on the gender of sources used within stories" (p. 149).
According to the author, females in the byline are a significant predictor of females appearing within news stories. This
seems an essential precedent for this study, which was not taken into account. The lack of a proper literature review
results in research questions that are more confirmatory than argumentative, negatively impacting the final result for
the paper.
2. The argumentation in the paper is too weak for a Q1 journal, and as mentioned before, it has more of a
confirmatory aspect. The novelty element that this research brings to the field of study is not clearly stated. The article
indicated that the study's conclusion or main contribution is that "Through the results of this study, it has been
revealed that there is no gender equality in the Chilean press." However, and according to the literature review in the
paper, this is a fact previously proven by statistical data and other reports. Also, it is essential to explain why a more
profound study to the Chilean case can help to further the knowledge of the phenomenon that could be applied
beyond the country concerned, which makes the presence of the research on an international journal more adequate.
3. Something to highlight from this paper - which is not emphasized enough in the text- is the method used to collect
and analyze the data. Creating an algorithm, make it available for reuse, and propose new ways of extracting corpus
for analysis is very valuable for the field of study and other researchers. However, the final dataset falls short, as it
does not include enough variables, beyond gender and media outlet, to allow for a more in-depth analysis. Based on
previous studies, one could include variables such as the topic to be covered, the profession of the source, the age of
the journalist or the time spent in the profession, the visibility of the news within the medium, among other factors,
which would allow a broader understanding of the phenomenon.

In sum, the study is relevant and adequate for the journal and of great interest for scholars of the field, but it needs
further work. I want to encourage the author to keep working on the subject. There is no doubt that the paper will
benefit significantly from further review of the literature that allows them to build better arguments and add some
variables that help us better understand the issue. The research needs to be a little more ambitious in its arguments,
and it could become a valuable contribution to the field. I hope to see a more structured version of this study further in
time and published in this or any other journal of the same importance.

Anexo B: Listado de medios de prensa incluidos

ABC
El español
okdiario
Público
Larazon.es
El correo españa
El independiente
El periódico
Thesun
Dailymail
The scottish sun
Wales online
Mirror
Daily star
The guardian
My london news
Leparisien
Lefigaro
Lacroix
Madeinmarseille
Francesoir
Humanite
Liberation
L'independant

Lyoncapitale
Lavoixdunord
La stampa
Ilfattoquotidiano
Fanpage
Gionale di sicilia
Lives sicilia
Il secolo XIX
La sicilia
Il mattino di padova
Publico
Cmjornal
Jornali
A noticia
Jornal do alarge
Linhas de elvas
Correo do minho
Irish independent
Irish time
Thejournal
Irish sun
Dublin live
Breaking news
CorkBeo
Limerick Leader
News.com
Nine
The age
Herald sun
New herald
Odt
Cook islands news
Post courier
The national
La depeche de tahiti
Samoa news
Daily post
CTV
Global news

Huffington post
CBC
The globe and mail
Toronto start
L'actualité
Le soleil
Fox news
Abcnews
Nprnews
Washington post
Huffpost
CBS
Mercuri news
Fenver post
Houston chronicle
Los angeles times
BBC news
Grupo formula
Uno tv
La jornada
La razón
Diario basta
El punto critico
24 horas
Noticias en la mira
Eje central
Via país
Letra p
Los andes
Diario alfil
O globo
Ilha noticias
Estado de minas
La razón
Opinión
El comercio
Diario correo
Perú 21
La primera

El perfil
Extra
Caras y caretas
El acontecer
Ecos regionales
Rivera mi cuidad
Diario cambio
ABC
Noticiario digital
Diario el vistazo
Periódico del delta
Desde abajo
Minuto 30
Bolivarense
La piragua
Diario del norte
Diario los andes
El digital Panamá
Destino Panamá
En segundos
El Salvador
La prensa grafica
Diario colatino
El Salvador Times
Verdad digital
Nación
La republica
La teja
El mundo
El país
Granma
Juventud Revende
Sierra maestra
Adelante
Ahora
Al día
Publinews
Soy 502
Republica

Guatemala
Tiempo
En alta voz
Paradigma
El país
La prensa
Confidencial
Articulo 66
Nicaragua Investiga
Es noticia
Presencia
Periódico el oriental
Hoy
El caribe
El nuevo Diario
Al momento
Diario Visión