Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

IMPACTO DE LA OPINIÓN PÚBLICA SOBRE EL DESARROLLO DE LA ENERGÍA NUCLEAR.

Un análisis de tópicos y sentimientos de las expresiones de los usuarios de la red social Twitter en Argentina.

Autor: ADRIANA A. CRUZ

Mentora: NATALIA SALABERRY

DICIEMBRE 2021

Resumen

Las organizaciones pertenecientes al sector de la energía nuclear deben procurar implementar adecuados procesos de comunicación y divulgación sobre la generación de energía nuclear, que permitan dar viabilidad a los proyectos que fueron relanzados y anunciados por el gobierno nacional argentino durante el año 2021. Sin embargo, existe un déficit en la obtención de información sobre la opinión pública ya que su relevamiento representa un desafío en términos de costos, alcance y oportunidad. El acceso a datos alternativos obtenidos de las redes sociales brinda una oportunidad para tomar conocimiento sobre la percepción pública en el país siempre y cuando sean gestionados adecuadamente. El objetivo general de este trabajo es identificar los tópicos relevantes sobre el desarrollo nuclear en las expresiones de los usuarios de Twitter realizadas en Argentina, que sirvan de insumo para el diseño de una campaña comunicacional en una organización a fin de incrementar la adhesión positiva a los proyectos en curso. Las temáticas predominantes reflejadas son el uso bélico de la tecnología, así como su relación con las energías renovables, el uso en medicina, los riesgos inherentes y las diversas aplicaciones nucleares. El sentimiento principal ha sido neutral, mientras que el porcentaje de tweets positivos ha sido el menor. De esta manera, el plan de comunicación nuclear debe contemplar la difusión de los usos pacíficos de la energía nuclear y la seguridad en la operación en todo su ciclo de vida.

Palabras Clave: Energía nuclear, opinión/percepción pública, datos alternativos, procesamiento de *tweet*s, aprendizaje automático.

Índice

Introducción

La reactivación del plan energético nuclear argentino podría verse interrumpida o demorada en caso de no contar con el apoyo de la población -que participa activamente de audiencias ambientales y plebiscitos-, lo que perjudicaría a la organización[1], y a otras empresas y organismos relacionados a los proyectos de desarrollo, comprometiendo la viabilidad económica de algunas –en especial ligadas al ciclo del combustible nuclear y agua pesada-y/o encareciendo los costos de otras. Para estas organizaciones -altamente interdependientes- obtener datos sobre la percepción pública a través de métodos tradicionales se vuelve costoso y de difícil implementación. Este déficit de información puede ser subsanado, con la obtención de datos alternativos –como los que surgen de las redes sociales- ya que permiten obtener información estratégica para la toma de decisiones permitiendo generar impacto positivo en la opinión pública.

La opinión de la población es relevante durante todo el proceso de construcción y ciclo de vida de los proyectos de energía nuclear. Los proyectos actualmente se enmarcan en la Ley 26.566, aprobada en 2009, y cobran en 2021 un nuevo impulso político, que permitiría avanzar en el desarrollo de la industria nacional. La Jefatura de Gabinete de Ministros[2] ha indicado en el informe al Senado Nro. 129, de junio 2021, el impulso a los proyectos de instalación de la planta de uranio (Formosa), la cuarta (Buenos Aires) y la quinta central nuclear (locación sin definir).

Por otra parte, en el informe Nro. 130 a la Cámara de Diputados de julio 2021, se informó la continuidad de la construcción del reactor modular (SMR) CAREM 25 cuyo impacto a largo plazo podría implicar la instalación de reactores de este estilo en todo el país (CNEA, 2017). Estos proyectos permitirían con perspectiva estratégica, el desarrollo de la planta de enriquecimiento de uranio (Río Negro) y la reactivación de la planta de agua pesada (Neuquén).

Las organizaciones que conforman este ecosistema -involucradas en el desarrollo de estos proyectos- son organismos estatales y empresas públicas que realizan la ingeniería, construcción, operación y mantenimiento de las instalaciones. Distintos procesos, como el ciclo del combustible, requieren especialización técnica que interrelaciona a estas organizaciones en una matriz de insumo-producto. Además, se requiere formación de capital humano

[1] Por razones de confidencialidad no se menciona la organización para la cual se desarrolla este trabajo.

[2] Para mayor información, los informes del Jefe de Gabinete de Ministros al Congreso se publican en https://www.argentina.gob.ar/jefatura/informes-al-congreso.

especializado y servicios de otras industrias como la construcción, siderurgia, química, entre otras.

Ante los recientes anuncios que proponen la reactivación del plan nuclear, se reaviva la necesidad de procurar el apoyo de la comunidad, ya que para su desarrollo pueden ser necesarias audiencias públicas y/o plebiscitos, en instancias municipales, provinciales y/o nacionales. La aceptación pública es un elemento clave por la influencia que ejerce en los diferentes niveles de gobierno que actúan como decisores, impulsores y reguladores en sus diferentes competencias y organismos.

El déficit de información que releve la opinión pública de manera sistémica representa un desafío para estas organizaciones, por sus costos, alcance y oportunidad. Deben procurar adecuados procesos de comunicación y divulgación que permitan minimizar el rechazo a la tecnología ya que el resultado de estos proyectos afecta su viabilidad económica a largo plazo. Entre las herramientas que se utilizan actualmente a efectos de realizar sondeos de opinión pública, se encuentran las redes sociales. La utilización de técnicas de gestión de datos obtenidos de dichas redes permite obtener información relevante para la toma de decisiones de las organizaciones.

Dada esta problemática, el objetivo general de este trabajo es identificar los tópicos relevantes sobre el desarrollo nuclear en las expresiones de los usuarios de Twitter realizadas en Argentina que sirvan de insumo para el diseño de una campaña comunicacional de la organización a fin de incrementar la adhesión positiva a los proyectos en curso. Estos usuarios incluyen tanto al público general, como a los medios de comunicación, instituciones ambientalistas, industriales y gubernamentales, entre otros.

Para poder alcanzar este objetivo, el trabajo se estructura del siguiente modo. El primer apartado expondrá el desconocimiento de la opinión pública que sufre la organización y la posibilidad que los datos no estructurados le brindan para la obtención de información necesaria para ~~el~~ su desarrollo. Se analizarán las características de la información no estructurada que se obtiene de la red social y cómo debe prepararse la organización para el resguardo y aprovechamiento de la misma, entendiendo el marco de control interno que articula para gestión de datos. Se introducirá en los conceptos de metodologías específicas que permiten acceder al valor agregado que encierran.

El segundo apartado, retomará en profundidad los conceptos de carga, limpieza y transformación de los datos; y mostrará el impacto de su aplicación en la base de *Tweet*s

(también llamados "Tuits", por la castellanización realizada por la RAE[3]) que fueron recopilados desde el 18 de agosto al 22 de noviembre 2021 referidos a la temática nuclear, a través de la API (*Application Programming Interface*) de Twitter con conexión a través del lenguaje *Python* (Van Rossum & Drake, 2010*)*. Los algoritmos de minería de textos permitirán dar lugar a la captura de los datos y a procesar los tweets para posteriormente realizar un análisis exploratorio través de técnicas de procesamiento del lenguaje natural. Por último, en el tercer apartado, los algoritmos de aprendizaje automático (*Machine Learning)* permitirán obtener resultados sobre el modelado de tópicos y el análisis de sentimiento, donde se sintetizan los elementos que deben considerarse para la elaboración del plan de comunicación y su seguimiento.

[3] La Real Academia Española define Tuit como "Mensaje digital que se envía a través de la red social Twitter y que no puede rebasar un número limitado de caracteres (Del Inglés: Tweet)" en https://dle.rae.es/tuit?m=form.

Impacto de los datos alternativos en la gestión de la organizacional

La comunicación en las redes sociales constituye un nuevo campo desde donde la organización puede obtener datos alternativos que le permitan conocer más acerca de los intereses y cuestiones presentes en la opinión pública. Este apartado presentará el análisis y características de las comunicaciones en Twitter, a partir de obtener datos alternativos que requieren un tratamiento especial y que deben ser gestionados a través de metodologías específicas que permiten acceder al valor agregado que encierran. Se accede de este modo a uno de los objetivos de este trabajo que es identificar las características de los datos no estructurados provistos por la red social Twitter, para enmarcarlas en las metodologías de análisis de una organización que requieren una adecuada gestión y control interno. Los pasos en que se organiza esta sección abarcan la comprensión de la organización ante la falencia de la información tradicional, la caracterización de los datos alternativos provistos por Twitter, los desafíos que plantean a la gestión y control interno de la organización y las metodologías necesarias para su aprovechamiento.

1.1. La organización ante la falencia de información tradicional

Existe poca información sobre la opinión que la población tiene respecto al desarrollo de la energía nuclear en el país. En el congreso americano de la Asociación Internacional de Protección Radiológica (IRPA) celebrado en México en 2006, Martín Chahab expresaba algunas ideas sobre las características de la opinión pública respecto a la utilización de la energía nuclear, anticipando una falta de apoyo de la población a partir de motivaciones ambientales. Elaboró, además, una idea de cómo favorecer la postura hacia a la tecnología nuclear:

> La apertura del marco perceptivo de las personas habría que generarla 'no en oposición a las creencias existentes' sino tratando de ganar espacios mentales alternativos en los sujetos. En este espacio habría que crear y luego fortalecer nuevas imágenes y símbolos que favorezcan la opción nuclear con un vínculo positivo y armonioso con el medio ambiente. (Chahab, 2006. p. 1)

El autor, luego de mencionar la carencia de estudios al respecto, referencia un trabajo estadístico de Greenpeace de 2006 a partir del cual se reflejaba que el 60% de las 600 personas encuestadas consideraban que las centrales nucleares son una fuente peligrosa de energía. Entre

otros resultados se resaltaba que 8 de cada 10 encuestados creía que las centrales nucleares contaminan el medio ambiente.

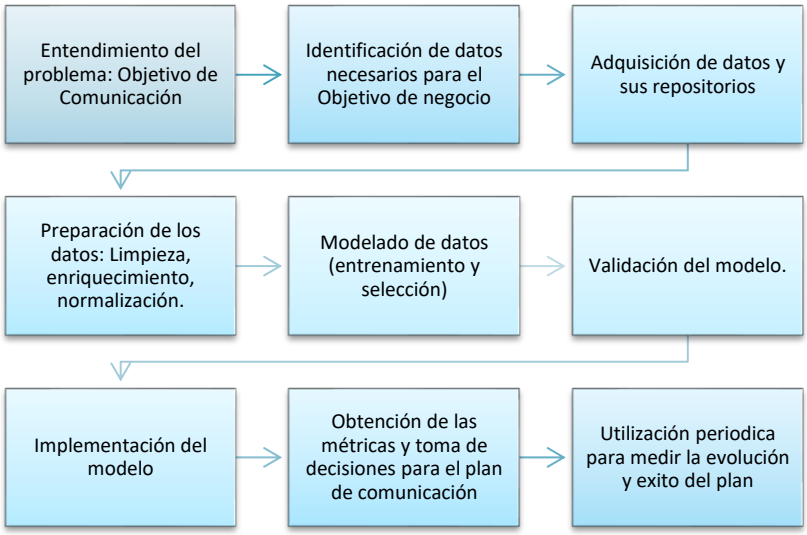
Ese año se realizó en Argentina la segunda encuesta de alcance nacional sobre temas de popularización de la ciencia (SECYT, 2007). Esta encuesta incluyó, por primera y única vez, un capítulo de la energía nuclear abarcando tópicos tales como: el conocimiento sobre la producción de energía nuclear en el país (68% no conocía el plan de reactivación y un 60% no sabía o creía que no se produce energía nuclear en el país) y los usos reconocidos (donde se destacó la medicina). Otras variables que analizó fueron la confianza en distintos actores sociales como fuente de información (predominando científicos -52%- y organizaciones ambientales -12%-), la percepción de la capacidad argentina en la materia (resultando "poco destacada" con el 45%), el riesgo percibido en el desarrollo (el 49,4% lo considera riesgoso pero controlable) y la aceptación social para el desarrollo nuclear (41% a favor, 30% en contra).

Al analizar estos resultados, Polino, C. y Fazio, M. E. (2009) han destacado que las instituciones científicas y las organizaciones del sector nuclear han realizado poca sensibilización sobre la opinión pública. Mencionan como factor la actitud de confidencialidad y el halo de misterio o secreto con que se maneja la industria poco afecta a aplicar políticas de comunicación social. Concluyen que esta actitud visibiliza sus consecuencias en la encuesta nacional.

Estos estudios no han sido actualizados debido a la dificultad de realizar nuevos con alcance adecuado y a un costo razonable, y con la suficiente amplitud para tomar decisiones respecto a cada provincia. Como parte del entendimiento del problema, entre las causas de la poca sensibilización publica ya mencionada, se encuentra la poca información disponible sobre la percepción social. Una forma de resolver este problema, es mediante la utilización de datos alternativos. Los datos alternativos, comprenden datos semi y no estructurados que pueden ser gestionados para la comunicación organizacional mediante herramientas de *Big Data*, donde se aprecian características especiales de volumen, velocidad, variedad y veracidad.

En la comunicación corporativa su gestión requiere establecer objetivos de comunicación (función), generar los repositorios que contemplen las características especiales de los datos, permitiendo realizar análisis descriptivos y de diagnóstico a fin de obtener una evaluación que otorgue valor añadido y fiabilidad a la organización (Miquel Segarra, 2020). Siguiendo esta idea de proceso, la figura 1 muestra en forma esquemática los pasos teóricos que conllevará para la organización la implementación de un proyecto que se desarrolle basada en este tipo de datos.

Figura 1:
Proceso Organizacional para la Implementación de un Modelo que Aproveche Datos Alternativos.



Nota: El proceso es iterativo, por lo cual que cada etapa puede requerir volver a la anterior. Y una vez terminado el ciclo, aparecer nuevos requerimientos de negocio que justifiquen cambios al modelo. Gráfica de elaboración propia.

Hasta aquí se ha cumplido con una primera aproximación del problema organizacional, es decir, se ha cumplido con la primera casilla de la gráfica anterior, dando inicio al proceso para la obtención de métricas para un plan comunicacional. Para poder avanzar en el desarrollo, es necesario conocer la fuente de datos alternativa que se utilizará y con esta caracterización volver al ámbito de la organización para asentar diferentes aspectos necesarios a tener en cuenta para su implementación.

1.2. Caracterización de datos alternativos de Twitter

En los últimos años, las redes sociales acapararon la atención del público, que se volcó en diversa medida a expresar sus opiniones a través ellas. Ante la masividad en el uso, numerosos estudios utilizaron información de estas redes con fines variados como actos electorales (Jungherr 2016), investigaciones de salud pública (Paul M.J y Dredze M., 2011), predicciones de valor de mercados (Ranco et al, 2015). De estas fuentes se verifica que los investigadores han optado por la red social Twitter, dado que su principal modo de comunicación es el texto y permite realizar procesos de minería de datos.

En Argentina la red social cuenta con 6,1 millones de usuarios activos, ocupando el puesto 17 en el mundo en octubre de 2021 (Statista Research Department, *2021*). Twitter permite

acceder a la los *tweet*s por medio de una API[4] que la misma red social brinda a desarrolladores de apps e investigadores académicos. Con esta facilidad, se pueden consultar *tweets* relacionados a ciertas palabras claves elegidas por un período de 10 días. El formato que se obtiene son registros JSON[5], que tienen la característica de estar constituido por objetos/estructuras y listas ordenadas de valores. La estructura JSON es fácil de leer por parte de humanos y máquinas (Warner, 2020). Cada objeto es un conjunto de pares nombre/valor que comienza y termina con símbolos de llave ({ }). Donde cada nombre es seguido por un símbolo de dos puntos (:). Cada par se separa entre sí con comas.

Por ejemplo, un tuit de la red social comienza de la siguiente manera:

Figura *2*

Extracto de Tuit Extraído de la Red Social.

```
      {
       "created_at": "Sun Oct 17 23:03:07 +0000 2021",
       "id": 1449873639130779653,
       "id_str": "1449873639130779653",
       "full_text": "CAMBIO CLIM\u00c1TICO: \u00bfLA ENERGIA NUCLEAR, ALIADA CONTRA
EL CAMBI... https://t.co/Afo2zxnTB5",
```

Cuadro de elaboración propia, a partir de un tweet recopilado.

En el código precedente se observan cuatro pares de objetos, los nombres *"created_at", "id", "id_str" y "full_text"* se refieren registros/estructuras del objeto tuit, y a continuación de los dos puntos un predicado, o valor asignado a cada nombre. Aunque los datos de texto son considerados "no estructurados", debido a la dificultad que implica la interpretación de la estructura gramatical, morfológica y semántica de las oraciones del lenguaje, este tipo de formato que brinda JSON permite hablar de datos "semi" estructurados.

Los datos semi estructurados contienen un esquema que existe en forma auto-descriptiva (Buneman, 1997). Como se ha visto en el extracto, cada dato es antecedido por un nombre descriptivo y la utilización de llaves (y corchetes para listas de elementos), proporciona una jerarquía de información que facilita su interpretación. Si bien cada predicado asociado a un nombre no tiene restricciones sobre el largo de su contenido, el esquema permite establecer la

[4] API: *application programming interface.* Se trata del conjunto de llamadas a ciertas bibliotecas que ofrecen acceso a servicios de información para ser utilizada por otro software.

[5] JSON: *JavaScript Object Notation.* Es un formato de texto ligero para el intercambio de datos. Si bien es independiente del lenguaje utiliza un estándar de JavaScript (ECMA-262 3a Edición - Diciembre de 1999). Más información en https://www.json.org/json-es.html.

relación entre los datos, facilita la navegación por ellos (aun cuando sean distintos en cada registro), habilita la realización de consultas (*querys*) a partir de la estructura existente y -por último- permiten evolucionarlos a datos estructurados si es necesario.

Entre las características de la información semi-estructurada se encuentra que brinda cierta facilidad para evolucionar el esquema, en caso que sea necesario agregar nueva información no contemplada en la definición existente de cada dato (Tajima, 2013). Al igual que el formato XLM el formato no es rígido y la capacidad de administrar nuevos componentes de la estructura permite evolucionar los datos sin que los desarrollos previos sean invalidados. Por ello, si Twitter posteriormente cambia la estructura, los procesos no serían interrumpidos o no serían difíciles de adaptar a la nueva definición.

La información semi-estructurada permite construir estructuras anidadas. En el caso de los tweets, esto se observa con la información del usuario, que es una estructura inserta dentro da cada tuit. Asimismo, es posible tabular la información que se obtiene, a efectos de manejarla amigablemente. Al ser un formato de texto simple, permite guardar grandes volúmenes de registros con poco costo de almacenamiento. La API, además de una limitación temporal a 10 días, incluye una limitación a la cantidad de registros por cada consulta. Cabe destacar el campo "full-text" que se ve en el ejemplo anterior y en el apéndice 1. En la visualización del campo se observa truncado por la utilización del intérprete *Colaboratory* (Colab) para ejecutar las consultas (Google Research, 2017), pero son textos que pueden contener hasta 280 caracteres (sin contar menciones). Los textos de este campo resultan -per se- datos no estructurados.

Los datos no estructurados no son -en primera instancia- fáciles de filtrar, de navegar o de utilizar para búsquedas. En el caso del campo de texto tampoco implican un almacenamiento costoso. Para analizar estos datos, primero la organización deberá someterlos a un proceso de transformación que permita luego ingresarlo a los modelos de análisis.

Este preprocesamiento plantea en el caso de los tweets ciertas dificultadas que vale la pena anticipar. En primera instancia son textos cortos, limitados a no más de 280 caracteres, por lo que la información de contexto que aportan es mínima. En segunda instancia el lenguaje no es uniforme: contiene abreviaturas, errores de ortografía, emoticones, jergas locales, y onomatopeyas. Dadas estas características de formato y extracción de los tweets para conformar una base de datos, la organización deberá responder con una arquitectura de gestión del dato y un control interno adecuado para implementar los modelos de análisis.

1.3. Desafíos de la gestión y control de datos alternativos en la Organización

En la página web de la misma red social indica se especifican los términos y condiciones del servicio que prestan, y donde se indica que:

> Los datos de Twitter son únicos y se extraen a partir de datos de la mayoría de las otras plataformas sociales porque reflejan información que los usuarios deciden compartir de forma pública. Nuestra plataforma de API ofrece acceso amplio a los datos de Twitter que los usuarios han decidido compartir con el mundo. (Twitter, 2021)

Esto incluye como datos públicos en los *Tweet*s, la información de perfil (incluida la biografía de un usuario y la ubicación declarada), el nombre para mostrar (que puede o no ser real) y el nombre de usuario. Los usuarios al ingresar a la plataforma aceptan los términos de servicio que incluyen la difusión de los *Tweets* a través de la plataforma, a través de las APIS y otras integraciones. Para las organizaciones, la información brindada por Twitter tiene un atractivo más a los ya mencionados. La red suele asociarse a usuarios disconformes, o a reclamos que brindan información útil para las empresas.

El costo de la adquisición de los datos es también un factor a considerar (J.P.Morgan, 2017). En este caso la red social puede brindar información limitada con cuentas de desarrollador gratuitas (similar a la de investigación que se obtuvo para este trabajo) o una cuenta de empresa que tiene la opción de elegir un plan con un costo en función a la cantidad de *request* mensuales, pero permite un mejor acceso cómo históricos de 30 días, más *tweets* por solicitud y enriquecimiento de los metadatos. Según el J.P.Morgan al considerar el costo de los conjuntos de datos, se debe considerar el costo del tiempo invertido en analizarlos sobre todo ante la posibilidad que no sean útiles en un modelo en producción. Esto requerirá del escrutinio de la calidad (completitud, *outliers*, metodogías) cuya responsabilidad comenzarán en la primera línea de control interno. Esto implica la adaptación de las estructuras tradicionales para el tratamiento de datos, ya que la obtención de ventajas competitivas dependerá de la habilidad de lidiar con las características de volumen, velocidad y variedad, en un proceso de análisis, implementación y control continuo (Salaberry, 2019).

De acuerdo a las necesidades de la organización se deberá considerar si el acceso a estos datos se realizará en tiempo real o con accesos periódicos. La gestión organizacional, en cuanto a la arquitectura a implementar y a la agilidad en el procesamiento de los datos se verán impactadas por esta decisión.

La organización debe elegir en qué tipo de base de datos se realizará el resguardo y explotación de la información. Actualmente los tweets se obtienen en formato JSON, que es semi-estructurado. Twitter puede cambiar el esquema del formato JSON cuando lo desee, por lo cual, para asegurar la compatibilidad con otras versiones es más eficiente resguardar la estructura autodescriptiva vigente en cada registro. Los datos obtenidos se guardarán en formato original (crudo), por si es necesario reprocesar las tareas, o se decide utilizarlos de otra forma. Con esta base ya podrían realizarse consultas básicas como ser contar palabras y encontrar así las más utilizadas, analizar los usuarios más activos, etc.

Las bases de datos relacionales (SQL) tienen ventajas tales como la persistencia de datos, la concurrencia, la integración con varias aplicaciones, la estandarización de su uso en las diferentes tecnologías existentes producida por tener mayor historia y un amplio abanico de opciones. Sin embargo, su estructura de tupla y algebra relacional impiden tener datos con anidaciones de registros, y su escalamiento se produce en forma vertical (Sarasa, 2019). El mismo autor menciona fortalezas de las bases NoSQL como la productividad en el desarrollo de aplicaciones, la capacidad de manejar datos a gran escala, su implementación como clúster dándole escalabilidad horizontal, y la falta de esquemas fijos que permiten adaptarse mejor al entorno dinámico de las redes sociales. Entre los modelos comerciales de bases NoSQL se encuentran Cassandra, Mongo Db, Dynamo, Redis, Apache Hbase, Riak , Oracle Nosql, Neo4j.

Como hablamos de datos en formato JSON, el tipo de base NoSQL que naturalmente presentan los datos como un objeto de estas características son las bases de Documentos, ya que son eficientes e intuitivas.

A partir de esta base histórica y acumulativa de la información cruda, se procede al preprocesamiento que se les aplica a los registros y permite enriquecer y estructurar para resguardar una base SQL (base Silver) que pueda ser accesible para análisis del usuario final, en este caso, el personal de relaciones públicas o institucionales de la organización. Por último, los resultados de la aplicación de las técnicas cuantitativas resultan en la tercera base (base Gold), que concentrará datos estadísticos, métricas y el histórico de los resultados obtenidos.

Respecto al procesamiento se hace necesario la aplicación de segregación de ambientes, mínimamente entre un entorno de desarrollo, prueba y otro productivo a fin de preservar la integridad de los datos en cada entorno, y no correr riesgos de impactos indebidos en la base en producción.

La organización con un marco de control interno adecuado, respeta el modelo llamado “De las tres Líneas” (originariamente llamado “De las tres líneas de defensa”) elaborado por el

Instituto Interno de Auditores[6], basa su idea que el gobierno de una organización define estructuras y procesos orientadas a establecer las acciones para lograr los objetivos, a supervisar la organización, y dar aseguramiento y asesoramiento por parte de la función de auditoría interna independiente. Define entonces a las tres líneas, siendo la primera aquella especializada en la cadena de valor sustantiva, es decir aquellas áreas directamente relacionadas a los productos y servicios que la organización brinda. La segunda línea esta más asociada a las funciones de apoyo y de gestión de riesgo a la segunda línea. Cabe aclarar que las "líneas" no son elementos per se estructurales, sino una definición útil de roles (The IIA, 2020)

La primera línea definirá los requisitos de negocio en detalle, alineados a las decisiones estratégicas respecto a mejorar la percepción pública sobre la energía nuclear. Validará los datos de entrada en cuanto a que sean apropiados y completos; así como los resultados cumplen los niveles de precisión esperados. Monitoreará su propia performance, y las necesidades de cambio de modelo. En la segunda línea, el sector de IT actuará validando permisos, arquitecturas y las definiciones de sistemas. Mantendrá también el registro del log de procesamiento en toda la cadena implementada, y la administración de la base de datos. Otros departamentos como las áreas administrativas controlarán que las contrataciones cumplan los requisitos definidos, y se obtengan a precio de mercado. El departamento de Capital Humano habilitará la gestión de capacitación sobre las habilidades y formación adecuadas controlando los requerimientos de la primera línea.

La tercera línea (auditoría interna), revisará los procesos de acuerdo a su evaluación de riesgo y a lo actuado por las otras líneas. Para esta estimación considerará los requisitos de *compliance* sobre los datos, la complejidad del modelo, si el mismo ha sido adquirido pre-entrenado o desarrollado internamente, su compatibilidad con otros procesos empresariales, la frecuencia de los cambios en el modelo, la documentación de requerimientos, tecnologías y desarrollo, la utilización de las métricas para acciones concretas de comunicación, las, el uso adecuado de los ambientes segregados de desarrollo, pruebas y producción y la efectividad de los controles generales de IT (Sammy, 2018). Desde su función de consultoría, auditoría interna puede además proponer activamente ideas de mejoras a partir de su propio conocimiento y *expertise* en riesgos, control y gobierno para agregar valor y mejorar las operaciones de la organización.

---

[6] The IIA es el instituto de auditores internos, organización cita en Estados Unidos que establece estándares mundiales relacionados a gobierno, riesgos y control interno. Certifica además las competencias individuales de los expertos en el tema.

1.4. Recapitulación sobre las metodologías para analizar datos no estructurados

Las técnicas de minería de textos, llamadas de procesamiento de lenguaje natural (PLN o NPL por sus siglas en inglés), procuran obtener datos estructurados a partir de textos libres que permitan analizarlos bajo el contexto de *Big Data*, y aplicar metodologías de aprendizaje automático para extraer el potencial de los datos y convertirlos en información. Entre estas metodologías, para ser consideradas por el plan de comunicación, es momento de introducir el concepto de modelado de tópicos y de análisis de sentimiento, que se abordarán en el segundo apartado y que usan como *features* las palabras que componen los documentos.

El modelado de tópicos se refiere a descubrir los principales temas que componen una colección de documentos no estructurados. Así se puede categorizar una colección según los temas descubiertos (Blei, 2012). En este caso, se aplicará la técnica a cada tuit, que serán los documentos de la colección, con el desafío evidente de la brevedad de sus contenidos. Los modelos son conjuntos probabilísticos, que buscan no sólo esta clasificación sino también como se interrelacionan. Existen también algoritmos de análisis dinámico que evalúan la evolución de los temas en el tiempo, permitiendo detectar tendencias. Existe un amplio abanico de metodologías aplicables: el análisis probabilístico de semántica latente –PLSA-(Hofmann, 2001), la asignación latente de Dirichlet –LDA- (Blei & Jordan, 2003), el modelo de tópicos bitérmino –BTM- (Cheng et al 2014), entre otros.

Por su parte el análisis de sentimiento permite polarizar los documentos se acuerdo a si la connotación semántica de las palabras (o conjunto de ellas) representan ideas positivas o negativas sobre los conceptos principales de los tweets. Dependiendo el modelo elegido, pueden considerarse una tercera opción: la neutral (Agarwal et al, 2011). Incluso, puede procurarse abarcar calificaciones más específicas, o que distingan cierta intensidad. Por ejemplo, algo puede ser positivo, pero puede ser “bueno” o “excelente”, o en el caso contrario un texto puede dar una polaridad “mala” o “espantosa”. El algebra del algoritmo asignará valores numéricos para estas categorías y buscar una categorización a partir de ellas (Jo, 2018).

La idea de realizar minería de textos para ser utilizada por la organización tiene como antecedentes los estudios sobre la opinión antinuclear que incluyen la revisión del contenido de noticias digitales sobre la actividad nuclear (Burscher, Vliegenthart, & Vreese, 2016), utilizando análisis de conglomerados. Sus conclusiones indican que la utilización en especial de los títulos y los encabezados permitió discriminar con mayor precisión los elementos de controversia sobre la opinión pública. Por su parte, con otra metodología, también se ha

utilizado redes neuronales para el análisis de sentimiento de los tweets en inglés sobre energía nuclear (Liu & Na, 2018), definiendo características necesarias para el armado de una base de datos centrándose en la recopilación de tweets (de enero a marzo 2018) y el pre-procesamiento de los textos, elementos fundamentales de la gestión de datos alternativos.

A modo de conclusión, en primera instancia se ha establecido las necesidades de información de la organización, que costosas de obtener por métodos tradicionales, puede optar por datos obtenidos a través de la red social. Estos datos alternativos, en formato no estructurado, contienen un campo de texto que, en su estado original, no aporta valor a la compañía. Una vez realizada la ingesta de datos, las organizaciones deben procurar procesos de limpieza para consolidar la calidad de los datos en cuanto en términos de unicidad, oportunidad, validez, precisión y consistencia. También se hacen necesarias tareas de transformación como enriquecimiento, normalización, y otros, según los objetivos planteados. Mencionados los tipos de análisis que pueden desarrollarse para lograr los objetivos propuestos los procesos especiales serán detallados en el apartado siguiente, a la par que se desarrolla el análisis de los tweets recopilados.

Procesamiento de los Tweets sobre energía nuclear

La obtención de los Tweets se realiza a través de la API de Twitter con conexión a través del lenguaje *Python*. Las técnicas de procesamiento del lenguaje natural permitirán convertir a los datos a la estructura necesaria para ser procesados por las metodologías de análisis que permitan la aproximación a las necesidades de la organización. En primer lugar se realizará la obtención de la base de tweets, y la determinación aproximada de la ubicación geográfica a partir de los registros de los usuarios. Segundo, se procederá a la limpieza y transformación de los datos para prepararlos para su procesamiento. Por último, ya con los datos preparados se explora su contenido desde una perspectiva cuantitativa, para realizar definiciones necesarias para la implementación de los modelos y reajustar el proceso de limpieza si fuera necesario.

2.1. Obtención y localización de los tweets de los usuarios

A fin de suplir la falencia de información sobre la opinión pública que tienen las organizaciones del sector nuclear, se ha definido la necesidad de relevar las redes sociales como un medio alternativo de obtener datos que permitan la toma de decisiones sobre el plan de comunicación necesario. A través de la API de Twitter de uso público y gratuito con accesos otorgado por la red social con propósitos de realizar este estudio, es factible realizar consultas a través de una conexión en lenguaje *Python* (Van Rossum & Drake, 2010*)* utilizando la librería *tweepy* de J. Roesslein (2020).

El acceso proporcionado permite obtener en cada consulta los tweets de los últimos 10 días que contengan determinadas palabras. En este caso se utilizaron las palabras: “nuclear","Nuclear","radiactivo","Atucha","atomic","energianuclear","uranio","nucleares","fisión","fision","fusion","Hualong","Candu","CAREM","Chernobyl" y "Fukushima". La elección es claramente temática y se evitó otros términos que podrían causar ambigüedades y traer mensajes no pertinentes al estudio (ej: energía y combustible). También se contemplaron algunos de los términos sin tildes, atento al uso común de la red.

A fin de explicar estos términos brevemente, cabe mencionar que las palabras “nuclear”, y “atomic” (esta última acortada, para que la API traiga todas sus variantes) referidas al tipo generación de energía, “Atucha” y “Candu” refieren a las centrales actuales: la primera es el nombre coloquial de las plantas en la localidad de Lima, Buenos Aires; la segunda, la tecnología de la planta sita en Embalse, Córdoba. “Fisión” y “Fusión” son reacciones nucleares que liberan la energía almacenada en el núcleo de un átomo, la primera es la que se utiliza actualmente en

las Centrales Nucleares, la segunda en cambio forma parte de numerosos experimentos alrededor del mundo. Por su parte “Hualong” es el nombre de la tecnología China del proyecto de IV Central nuclear y “CAREM” es el nombre del proyecto de reactores modulares que está en construcción en Lima. Por último, “Chernobyl” y “Fukushima”, son las locaciones donde sucedieron los dos accidentes nucleares más recientes de gravedad de los cuales la sociedad pudiera hacer mención. La palabra ‘Uranio’ hace referencia al ciclo del combustible nuclear, y así se podría captar cuestiones relacionadas al proyecto de la nueva planta de procesamiento a radicarse en Formosa.

Cabe destacar que, al realizarse consultas con periodicidad menor a 10 días, el servicio puede proporcionar registros de tweets ya obtenidos anteriormente. Por tal motivo al cargarse la totalidad de los tweets en la base, se eliminaron los duplicados considerando el campo de identificación de tuit, el usuario y la fecha de emisión del tuit. El primer archivo Excel grabado fue el día 19 de agosto e incluye tweets del día 12 del mismo mes, y el último archivo fue gravado el día 22 a las 24 horas GMT-3 (Argentina). La obtención total de tweets al eliminar duplicados conformó 6.339 registros. Luego se eliminaron los registros de un usuario en particular, que por la intensidad de posteos y la temática abarcada podría desvirtuar el análisis, por lo que se conformó en 5.437 tweets.

Un tema particularmente sensible es la geolocalización de lo Tweets. La red social permite en sus consultas solicitar un punto geográfico central y un radio de consulta para obtener los tweets. A fines de abarcar el territorio argentino, se posicionó este punto central en las coordenadas Latitud: -38.430387, longitud -64.880758, con radio de 1800Kms, por lo que esto incluyó países limítrofes y se excluyó Antártida. Los tweets se solicitaron solo en idioma español, y se rescataron los campos 'id', 'fecha', 'usuario', 'tweet_text', 'geo', 'coordenadas', "truncado", ‘locacion', 'seguidores', 'en respuesta a', 'retweets', 'favorita' para conformar archivos excels en cada bajada de datos. Las columnas ‘geo’ y ‘coordenadas’ no trajeron resultados, por lo que los datos requirieron programar una asignación especifica a partir del campo ‘Locación’ que es el completado voluntariamente por el usuario en su perfil.

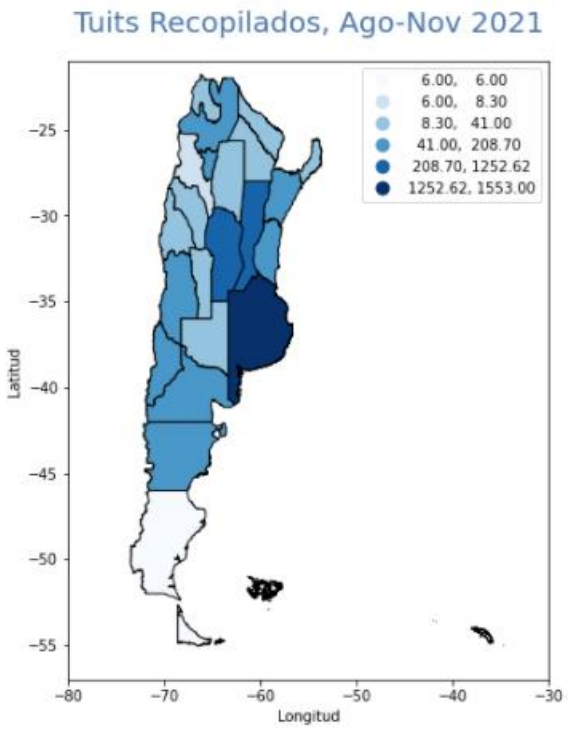
Con estos datos, se procedió a realizar la asignación geográfica a partir de los datos obtenidos en el campo locación. Esto requirió una serie de pasos para realizar la mejor aproximación posible de país y provincia a partir de los datos consignados. En el primer paso extrajeron todas las locaciones consignadas conformando una base temporaria a partir de los registros únicos y así realizar el procesamiento sobre la menor cantidad de registros posible. El segundo paso, se crearon los campos país y provincia, completándolos con el texto “SIN

DATO". La primera asignación fue buscando en los textos, los nombres de los países y asignando el campo "país" para una primera discriminación entre tweets argentinos y extranjeros.

Luego se cargó una base normalizada de provincias argentinas, y se buscó específicamente si los usuarios habían declarado sus provincias. Como esta comparación se realizó en forma literal, y no cubría otros niveles geopolíticos, se utilizó en un siguiente paso el servicio de normalización y enriquecimiento de datos geográficos de Argentina (Sampietro, 2018). El servicio es una API REST, la misma funciona bajo un esquema público y abierto que no requiere autenticación, si bien su cuota es limitada (DNDA, 2020). Para utilizar esta API se utilizaron funciones elaboradas por el equipo de Datos de la Nación Argentina para la PyCoNar 2018 que son públicas su GitHub [7]. El servicio permitió iterar búsquedas a nivel provincia (para normalizar los casos de escrituras abreviadas que no hubieran sido captados por la comparación "literal" ya realizada), a nivel departamento y a nivel localidad para extraer la provincia a la que pertenecen. Como puede verse en la siguiente gráfica, la mayoría de los tweets fueron geolocalizados en la Provincia de Buenos Aires.

[7] La PyCon (Python Conference) es una convención anual para la discusión y promoción del lenguaje de programación Python. Desde 2009 se realiza también en Argentina, donde adquirió el nombre PyConAr. Es gratuita y abierta a todo público. El Github donde las funciones publicadas por el gobierno se encuentran en: https://github.com/datosgobar/taller-georef-pyconar-2018/blob/master/pyconar.ipynb

Figura 3:
Asignación geográfica provincial de los Tweets Argentinos.



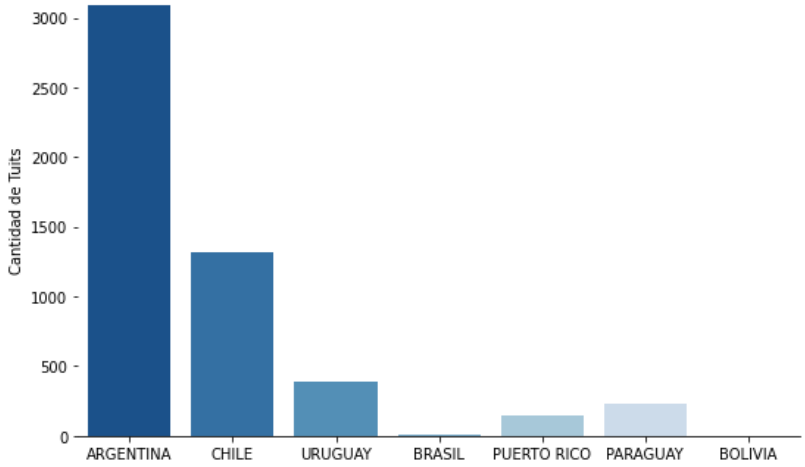
Mapa de elaboración propia en Python con Capas SIG (shapefile) del Instituto Geográfico Nacional

Del análisis de cómo se impactaban estos cambios en la base de tweets, un remanente quedaba sin identificar ya que el usuario consignaba el nombre de la ciudad, estos casos fueron tratados manualmente para los casos más repetitivos. Una posible problemática es la homonimia de departamentos y localidades. Para el caso, la función utilizada resguarda el primer resultado devuelto por la API. Esto podría influir en el sesgo de los datos a ser asignados a la provincia de Buenos Aires. Otra situación está dada por CABA, si los usuarios solo consignaron “Buenos Aires” o alguna abreviatura, no pudieron ser discriminados como de la ciudad. Aun así, persistieron en la base final datos que pudieron ser identificados por país, pero no a nivel provincia y casos que no pudieron ser asignados, que en una revisión visual de estos remanentes en general eran provenientes de ciudades o pueblos extranjeros o sin dato alguno.

Esta segregación por país fue guardada en una planilla de cálculo con diferentes solapas. Si bien la asignación país/provincia no es perfecta, es una aproximación razonable y viable de ser reproducida de forma automática cada vez que se realiza una nueva consulta. La cantidad de tweets de cada país lleva ciertos interrogantes, que pueden desprenderse de la figura a continuación. El principal es plantear ¿Por qué países con poco desarrollo de la industria nuclear como Chile (que únicamente tiene dos reactores de experimentación, de los cuales solo uno se encuentra activo) o Uruguay (que tiene prohibido el uso por ley), que tienen una población

notablemente menor que Argentina tienen una cantidad considerable de registros? Esta información podría ameritar algún estudio posterior, sobre todo bajo la perspectiva de comercio internacional que tiene el proyecto CAREM25.

Figura 4:
Asignación de los tweets a países al final del proceso de geolocalización.



Gráfica de distribución elaborada en Python con la librería Seaborn (Waskom, 2021)

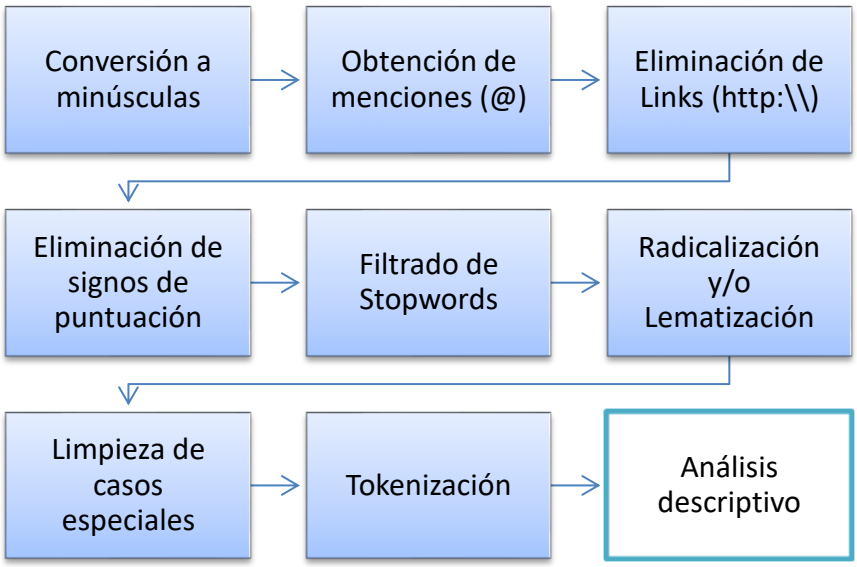
Compilados los tweets y segregados geográficamente, podemos a continuación comenzar la limpieza y transformación de los textos a través de técnicas de procesamiento de lenguaje natural que faciliten ser tratados. Posteriormente avanzaremos en el análisis exploratorio a fin de prever las dificultades para realizar el modelado de tópicos y el análisis de polaridad en el apartado siguiente.

2.2. Técnicas de preparación de los datos alternativos

Los tweets están escritos por usuarios, en lo que llamamos el "lenguaje natural", el uso cotidiano del español para expresarse con todas sus particulares locales y personales. La limpieza y preparación de los datos, difiere sutilmente para los dos objetivos de análisis planteados. Para el modelado de tópicos nos interesa eliminar palabras sin valor conceptual y buscar las grandes ideas tras los textos, en cambio para el análisis de sentimiento, cualquier expresión o incluso los emojis y emoticones podrían aportar información sobre el tono y polaridad dado al tuit. A fin de preparar los tweets para el análisis de tópicos, buscamos disminuir la variabilidad de formas de expresión y uniformar la escritura eliminando ruidos al

análisis subsiguiente (Kulkarni & Shivananda, 2021). Esto conlleva una serie de pasos, que anticipamos en la siguiente gráfica y detallamos a continuación.

Figura 5:
Pasos para la preparación de los tweets a fin de realizar un análisis descriptivo.



Gráfica de elaboración propia

Conversión a minúsculas todos los textos: Nos permite que puedan reconocerse como iguales palabras que por iniciar la oración o el deseo del usuario de dar énfasis, se procesarían como si fueran diferentes.

Obtener menciones: Los usuarios suelen arrobar a otros usuarios ya sea a modo de respuesta, ya sea para involucrarlos en el tuit de alguna forma, o para citarlos. Por este motivo si bien se mantienen en los textos las menciones (sin la arroba). Al obtener la lista se determinó que las 1704 menciones se distribuían en 1000 usuarios (únicos) y un promedio de 1,74 menciones por usuario, lo que indica que hay pocos líderes relevantes sobre estos temas. El usuario más referido (con 34 menciones) fue @operadornuclear, un divulgador español que es operador en una planta nuclear de ese país y autor del libro 'la energía nuclear salvará el mundo'. Los usuarios de agencias de noticias que se destacan son @noticiasde (31 menciones), @lanacion (12), @infobae (10) y @clarincom (8), entre otras. Aparecen entre las dominantes las entidades del sector con @cnea (24 menciones), @ibalseiro (11), @nucleoelectrica (8), @iaeaorg (5) e @invapargentina (5). Otros usuarios mencionados son políticos argentinos, periodistas, y otros influencers. Para mostrar la dispersión 878 de los 1000 usuarios eran mencionados 1 o 2 veces y solo 20 en 8 o más oportunidades.

Eliminar hipervínculos: los Links que los usuarios agregan ya sea para referenciar páginas web con la ampliación de los temas que están comentando, o que originan una respuesta o viralización de temas son eliminados. Como los links son en un formato corto (y codificado) no aportan palabras que puedan dar valor a ninguno de los análisis que se emplearan más adelante.

Eliminar símbolos de puntuación: si bien en el lenguaje natural permiten establecer la entonación de los escritos, a efectos de su tratamiento como *Big Data*, podrían prestar a confusiones. También se eliminan otros códigos de textos y caracteres especiales (ej: '&gt,' que representa el símbolo "mayor qué"). Los emoticones (basados en signos de puntuación) y -últimamente- los emojis (basados en pequeños gráficos) han ganado popularidad porque son un atajo para ciertas expresiones, que los hacen útiles para análisis de sentimiento (Kwartler, 2017), por lo que se eliminan para el modelado de tópicos pero serán utilizados en el análisis de sentimientos, ya que pueden convertirse a palabras.

Eliminar Stopwords: Existen palabras que no son relevantes a los contenidos del texto, por lo que son removidas para mayor eficiencia (Jo, 2018). El listado de palabras es internalizado a través de un archivo o una librería y todas aquellas que se encuentren en la lista son eliminadas. En este caso se utilizó la lista provista por el corpus de la librería NLTK en español (Bird, Loper & Klein, 2009).

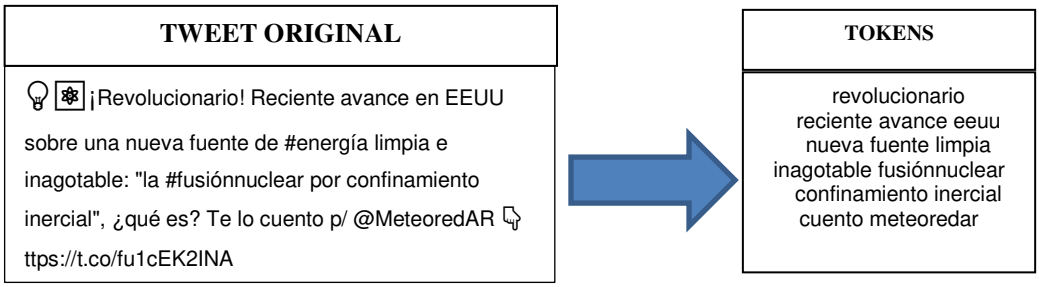
Radicalización (*stemming*) vs. Lematización: Las palabras tienen variantes por su conjugación, por su género, o número. También se encuentra variedad en la riqueza del idioma que encuentra sinónimos para expresar las mismas ideas o conceptos. El método de obtener la raíz de las palabras permite acortar las conjugaciones de las mismas, pero tiene la desventaja que le quita legibilidad a los resultados cuando se trata del idioma español. Para ejemplificar esto, al encontrarse las palabras "salvó", "salvará", "salvaría" al acortar la raíz solo se mantendría para todos los casos la leyenda "salv": todos los verbos y gerundios acortados de esta forma dificultan la lectura. En cambio, la lematización busca sinónimos y para este tipo de conjugaciones, elige quedarse con un verbo en infinitivo, para el ejemplo dado: "salvar". Por esta diferencia, a efectos de este trabajo se prefirió lematizar a través de la librería spaCy (Honnibal et al, 2020).

Luego de correr este proceso, se corrigieron manualmente algunas lematizaciones que fueron identificadas como erróneas y espacios excedentes. Como las acciones de limpieza y transformación son retroalimentadas por el análisis exploratorio y por los procesos de aprendizaje automático que se aplican, también se incluye una etapa de eliminación temas que

no aporten valor a la temática buscada. Por ejemplo, al analizar por cual palabra clave fueron obtenidos los tweets encontramos la distribución de grafica a continuación, lo que motivó a eliminar las palabras “energía” y “nuclear” y otros verbos muy genéricos como “hacer”, “decir”, “estar” o “poder” ya que iban a ser sobreentendidas en la mayor parte de los tweets, no aportando ninguna información novedosa.

Tokenización: se refiere al proceso de segmentar un texto o varios en tokens ya sea por un espacio en blanco o por signos de puntuación (Jo, 2019). Como estos últimos ya fueron removidos previamente, las palabras serán divididas a través de los espacios en blanco existentes. Cada tuit se transformará entonces en una lista de tokens. A modo de corolario de este proceso de limpieza y transformación se muestra un ejemplo de uno de los textos de los tweets procesados.

Figura 6:
Ejemplo del resultado del procesamiento realizado hasta la obtención de tokens



Gráfica de elaboración propia

2.3. Análisis exploratorio sobre los datos obtenidos de la red social

Una vez recopilados los datos de la red social y preparados reconociendo su naturaleza de texto, es necesario conocer la conformación de la base obtenida ya que permite anticipar y retroalimentar el proceso de limpieza y transformación antes de iniciar el apartado siguiente con los procesos de aprendizaje automático que han sido propuestos. El recorrido que se realizará parte de conocer la estructura de los textos de los tweets, los usuarios recurrentes y los que tienen cierta popularidad, por último, abordaremos la primera aproximación a los contenidos analizando *ngrams* y nubes de palabras.

La primera noción que se requiere de los tweets tiene que ver con su extensión luego de la limpieza realizada. El promedio actual es de 96 caracteres, con un máximo de 271 y un mínimo de 4. Como se observa en la siguiente figura, la distribución esta sesgada a la izquierda. Respecto a la cantidad de tokens el mínimo es de 1 y el máximo de 29, con un promedio de 11 palabras. Tweets con solo 1 o 2 tokens no permitirían interrelacionar las palabras a un tema de

forma confiable, para que la calidad del análisis de tópicos sea aceptable. Previo a realizar los análisis del apartado siguiente, se eliminarán estos tweets cortos, quedando la base de análisis con 3014 mensajes argentinos.

Otro dato de interés es conocer cuáles son las palabras más relevantes en el corpus total de tweets, para lo cual, nos valdremos gráficamente de una nube de palabras, lo cual nos permite identificar fácilmente los conceptos más repetidos ya que pondera este factor en el tamaño mostrado en cada una. Resalta a priori el uso bélico de la tecnología nuclear, representado por las palabras "bomba", y acompañado por otras como "guerra", "submarino", y "arma". Esto ya es un resultado en sí: la primera idea al utilizar el concepto nuclear no se relaciona a sus usos pacíficos, sino a su accionar destructivo y a los efectos de la explosión.

Figura 7:
Nube de Palabras a partir del corpus de Tweets

Gráfica de elaboración propia utilizando la librería WordCloud (Mueller et al, 2018)

Para comprender mejor la utilización de parte de los usuarios del lema 'Bomba' tomó una selección de ejemplos para corroborar que se relación a temas de relaciones internacionales, armamentos de países, política nacional, autocríticas nacionales, fútbol y hasta cocina. Esta distribución temática variada, y la alta frecuencia que muestra el lema muy superior a todas las demás, hace pensar que quizá ensucie las relaciones entre palabras al efectuar el modelado de tópicos. Sin embargo, tendrá que ser evaluado cuando alcancemos esa etapa en el siguiente apartado. Además, se denota en estos tweets la importancia de haber lematizado los textos, se

ha unificado de esta forma las palabras “bombardero”, “bombardear” y “bombas” en el único lema “Bomba”. El segundo tema que se destaca ya tiene que ver con la “Central” y otras palabras con tamaño preponderantes son “planta” (lo cual podría considerarse un sinónimo de la anterior, en algunos contextos), China (con quien se gestiona una nueva central). “Reactor”, “tecnología”, “mundo”, “país” también se destacan en la nube de palabras elaborada.

Las palabras por sí solas ya brindan cierta información, pero es posible analizar sus asociaciones en n.*grams*, conjuntos de n miembros de palabras que son asociadas en los textos como un concepto único. Cabe recordar que la palabra “nuclear” que podría asociarse rápidamente con otras como “medicina nuclear” o “bomba nuclear” fue eliminada durante el preprocesamiento explicado en el sub-apartado anterior, por lo que no es de esperarse este tipo de asociaciones obvias. En este caso, se analizaron casos de bigramas y trigramas (Pedregosa *et al.*, 2011.). Los mismos no arrojaron frecuencias altas, sin embargo, se destaca que respecto a los bigramas la mención de Corea del Norte, un centro de medicina, los riesgos que afronta el mundo (en varias asociaciones) y otro tema medico: la resonancia magnética. Por su parte en los trigramas hay tres conceptos quedaron resaltados; Un alto nivel de riesgo en 4 décadas, el cambio climático, y el nuevo servicio de resonancia magnética (que se instaló en el municipio de Tigre). Los resultados de bigramas y trigramas pueden verse en la siguiente tabla.

Tabla 1:
Frecuencias de Bigrams y Trigrams, relacionados a riesgos y medicina nuclear)

frequency	bigram	frequency	trigram
49	corea norte	32	mundo afronta mayor
33	centro medicina	30	riesgo cuatro décadas
32	mundo afronta	28	mayor nivel riesgo
32	afronta mayor	28	afronta mayor nivel
31	resonancia magnética	26	nivel riesgo cuatro
31	cuatro décadas	24	servicio resonancia magnética
30	riesgo cuatro	24	resonancia magnética campo
28	nivel riesgo	24	nuevo servicio resonancia
28	mayor nivel	17	advirtió mundo afronta
27	cambio climático	16	punto capacidad país

Cuadros de elaboración propia utilizando la librería Scikit-learn – CountVectorizer (Pedregosa *et al.*, 2011.)

En este apartado se ha mostrado la obtención de los tweets en base a la utilización periódica de la API con accesos de consulta para realizar investigación académica. Unificadas las múltiples consultas realizadas, se procedió a realizar una aproximación geográfica de los tweets en base a la locación declarada en el perfil de los usuarios y utilizando las API de geo normalización brindadas en forma pública por el estado nacional. El segundo paso consistió en preparar los datos alternativos obtenido a través de normalizar el uso de minúsculas, quitar signos ortográficos, *smileys*, números, emoticones y *stopwords.* También se optó por lematizar las palabras, a fin de unificar en pocas variantes las conjugaciones, pero mantener la legibilidad de los documentos. Al final se tokenizaron los tweets para permitir el inicio del análisis exploratorio que nos anticipó la necesidad de eliminar los registros con pocas palabras (1 o 2), por lo que la base de análisis de tweets argentinos constará de 3014 registros. A través de la visualización de una nube de palabras, la posibilidad que el uso bélico de la tecnología nuclear tiña los resultados de los análisis a realizar en el apartado siguiente. Esto fue consistente con el análisis de bigramas y trigramas, que reflejaron los temores a ciertos riesgos, aunque en contraposición se resalta el uso médico de la tecnología.

Aprendizaje Automático para la comunicación organizacional

Una vez obtenidos y preparados los datos, en este apartado se procura identificar elementos relevantes a ser tenidos en cuenta en la elaboración por parte de una organización de la industria de una campaña eficiente de comunicación sobre el desarrollo de la industria energía el país. En primera instancia se aplican técnicas de modelado de tópicos, para obtener los temas de los Tweets que se desprenden como factores de preocupación entre los usuarios. Este análisis será completado en segunda instancia por la determinación el análisis de sentimiento (polaridad) de los tweets realizado con librerías adaptadas al lenguaje español. Por último, las métricas elaboradas requieren ser interpretadas y relacionadas con la comunicación organizacional, ya que luego permitirán realizar el seguimiento de la eficacia de las acciones emprendidas en el marco del plan de comunicación.

3.1. Modelado de Tópicos sobre los mensajes en Twitter

El modelado de tópicos refiere a una técnica de aprendizaje automático (*machine learning*) que puede realizarse con un enfoque supervisado o no supervisado, de acuerdo a los datos de origen. Este tipo de modelado es un algoritmo de clasificación de textos o documentos

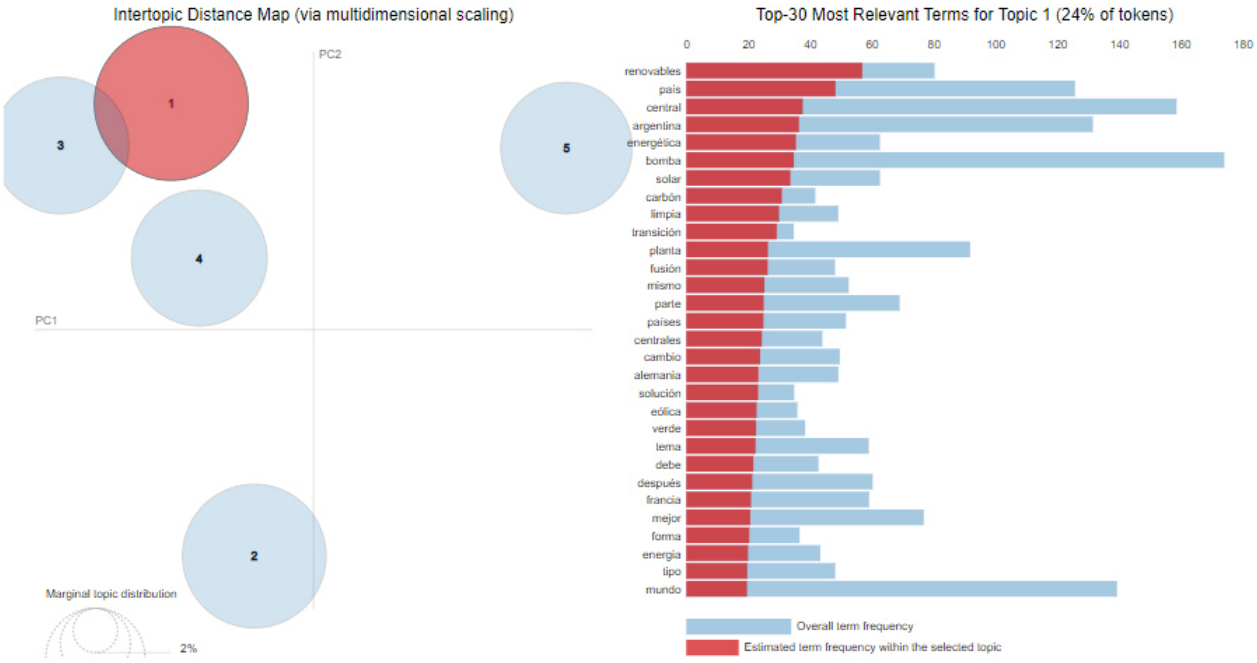
no estructurados, que asignará a cada documento (en nuestro caso, los tweets) la clasificación de a cuál temática pertenece. Un caso de ejemplo de aprendizaje supervisado aplicado a textos, sería si debemos clasificar una serie de documentos en carpetas preexistentes. Las carpetas ya tienen documentos en su interior que puede tomarse de base de entrenamiento para detectar los atributos que permiten predecir a los nuevos documentos a cuál carpeta deberían clasificarse. En este caso la carpeta representa el tópico o tema que actúa como *label* de supervisión de los documentos existentes (Weiss, 2015).

Sin embargo, en el caso de aprendizaje 'No Supervisado', no se trabaja con una base de entrenamiento que ya tuviese etiquetados los tópicos existentes a los cuales forzar la clasificación, sino que por el contrario se trata de un procedimiento de descubrimiento de tópicos desconocidos (latentes). Con los Tweets se desconoce a priori las temáticas que abarcan (más allá que todos sean de temática "Nuclear"). En estos modelos de clasificación no supervisada esta subyacente la idea que los diferentes temas utilizarán palabras afines en alguna proporción (Blei, 2012). Formalmente, este autor, define "tópico" como la distribución de un vocabulario específico: una lista de palabras fijas que tienen mayor o menor preponderancia en la definición del tópico. Cada texto o documento (o Tuit, en este caso), se contrasta con estos tópicos hallados y se obtiene una probabilidad de que el texto pertenezca a esos tópicos, seleccionando al final para ese texto cual es el tópico más probable al que pertenezca (momento de clasificación). Se han desarrollado diferentes métodos para lograr implementar estos algoritmos, y aplicaremos para nuestra selección de tweets los modelos llamados BITERM y LDA cuyos resultados se comentan a continuación.

Mas allá de la alta informalidad de los textos de los tweets, el lenguaje diverso y la fuerte temporalidad al contexto que tienen y con lo que ya hemos lidiado en el apartado anterior, el principal inconveniente a la hora de encontrar patrones es su extensión: tan solo 280 caracteres de texto que no brindan muchas oportunidades de obtener *features* (atributos), principalmente porque la correlación entre palabras se ve empobrecida por la limitación. La principal idea que desarrolla el modelo BITERM (o BTM) es que las palabras pueden enriquecerse conformando términos de a pares, es decir, alimentando el modelo a partir de las asociaciones desordenadas de dos palabras (Yan *et al*, 2013). La idea es que esta asociación de pares es más probable que pertenezcan a un mismo tópico. Al agregar estos bigramas al corpus, demostraron mejor estabilidad y confiabilidad que el análisis por palabra simple para la revelación de los tópicos latentes que otros algoritmos aplicados a textos cortos (Cheng et al, 2014)

La aplicación de este modelo, solicitando 5 tópicos a los tweets, mostró como resultados una clasificación donde el término "bomba" se presenta en todos los tópicos. Esta posibilidad se había advertido en el apartado anterior, cuando se analizó la nube de palabras. Asimismo, se ha planteado en el primer apartado que contemplamos como posibilidad que, a partir de los resultados de cada etapa, se deba volver uno o varios pasos atrás a la luz de los resultados que se obtengan. En el primer tópico (Figura 15) aparecen relevantes las siguientes palabras: Renovables, País, Central, Argentina, Energética, Bomba, Solar, Carbón, limpia transición, relacionadas con las energías no emisoras de carbono, al igual que la energía nuclear, aunque también aparece la palabra transición y Carbón. El tópico abarca el 24% de los tweets, por lo que es la temática más relevante.

Figura 8:
Modelado BITERM de tópicos. Se visualiza el 1er Tópico refiere a Energías Renovables.



Elaboración propia utilizando la librería PyLDAvis (Sievert & Shirley, 2014).

En el segundo tópico con 20,9% de los tokens, involucra conceptos bélicos y relaciona diversos países, como ser: "china", "submarino", "acuerdo", "eeuu", "misil", "bomba", "arsenal", "argentina", "guerra", "mundo". El tercer tópico (18,9%) es de difícil conceptualización ya que involucra las palabras tales como: "bomba", "central", "reactor", "guerra", "medicina", "argentina", "mejor", "centro", "cuenta", "mundo", "tecnología", "tema", "agua", "planta", "país", "cnea_arg", "ingeniero", "lugar", "física". La diversidad de términos contemplados permite inferir que habla de las aplicaciones tecnológicas en general:

generación de energía, medicina y el uso bélico (nuevamente) y los asocia con las ciencias detrás (ingeniero, física, CNEA) de las mismas.

Por su parte, el cuarto tópico con un 18,7% de los tokens, habla de los riesgos de esta tecnología, utilizando las palabras "central", "mundo", "mayor", "nivel", "riesgo", "tecnología", "décadas", "bomba", "proyecto", "país", "nacional", "Atucha", "planta", "afronta". Esto esta directamente relacionado con un comunicado de la ONU a fines de septiembre advirtiendo sobre el mayor riesgo nuclear en 4 décadas, asociado a la capacidad armamentística de los países, que fue replicado de diferentes formas tanto por cuentas asociadas a medios de comunicación como por particulares. Por tanto, también tiene relación con el concepto bélico, si bien no se descarta que incluya los temores a otros riesgos de la tecnología. El último tópico (17,5%) se centra en la aplicación en medicina. Se destacan los términos: "nuevo", "china", "bomba", "campo", "resonancia", "magnética", "servicio", "mundo", "misil", "medicina", "parte", "tigre".

En cambio, al quitar el término "bomba" y evitar así que se conjugue en multitud de bitérminos ganando mayor importancia, se aplicó el algoritmo solicitando 8 tópicos. No surgieron temas principales nuevos, sino algunas aperturas de los temas ya identificados. Por ejemplo, en cuanto al tópico energías Renovables, se divide en dos: Separa un tópico que integra a Alemania y Francia, dos países muy comparados al renunciar el primero a la energía nuclear manteniendo las plantas de carbón, y mantiene un tópico genérico de renovables con ciertos términos asociados a lanzamientos espaciales.

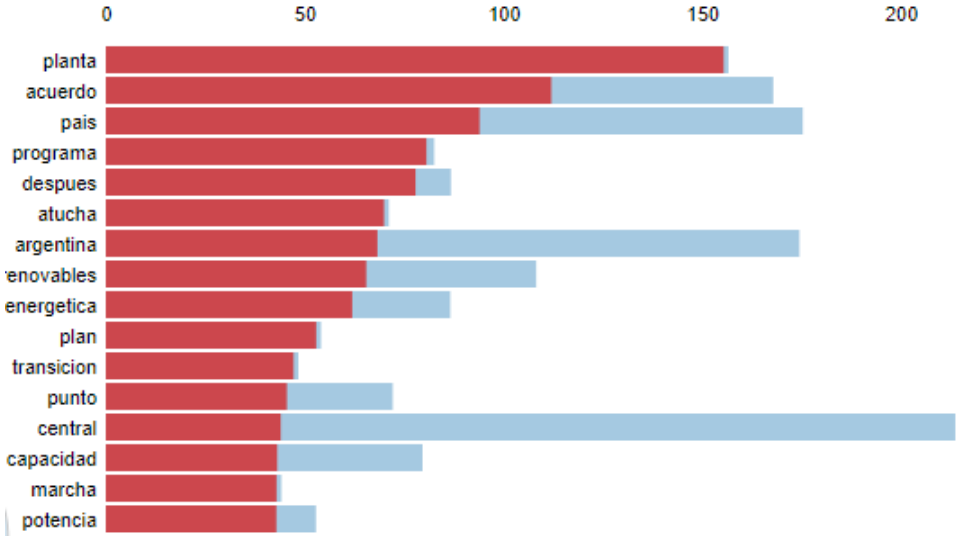
Por su parte, el modelo de Asignación de Dirichlet Latente (LDA, por sus siglas en inglés) es ampliamente utilizado según demuestra la abundante literatura para la detección de tópicos. Blei & Jordan (2013) describen este proceso como un modelo bayesiano de probabilidad en 3 niveles. Determina a las palabras como la unidad discreta, los documentos como el conjunto de palabras y el corpus como el conjunto de documentos. Cada documento puede tratar varios tópicos y será la probabilidad la que defina a cuál de ellos definitivamente se asigne como principal. En este caso se requirió la detección de 6 tópicos, que se detallan a continuación.

El primer tópico bajo esta modalidad vuelve a referirse a energías renovables, sin embargo, incluye directamente la idea de un plan/programa, una transición y menciona "acuerdos" y a "Atucha", por lo que puede inferirse que hable de determinados proyectos energéticos. A modo de ejemplo se muestra la gráfica obtenida de los principales términos asociados a este primer concepto (18,3% de los tokens). El segundo tópico, en cambio, no tiene un concepto directamente asociado si bien agrupa al 18,1% de los tokens, pero se puede extraer

conceptos de tecnologías nuevas (fusión) y accidentes famosos como Fukushima y Chernobyl. Las palabras principales que lo componen son “central” , “eeuu” , “ingeniero” , “tecnologia” , “fisica” , “fusion” , “china” , “chernobyl” , “informe” , “global” , “generacion” , “agua” , “argentina” , “centrales” , “pais” , “tiempo” , “falta” , “fukushima” , “accidente”.

El resto de los tópicos son conceptualmente mixtos, lo que reafirma que, para textos cortos, la utilización del método BITERM brinda mejor asociación de ideas. El tercer (17,4%) y cuarto (17,1%) de los tópicos detectados a través del LDA vuelven a los conceptos bélicos uno relacionado con ideas de desarrollo tecnológico y el otro más a la seguridad de los países. Un quinto tópico (14,7%) mezclas la idea de energías renovables con palabras como “china”, “corea del norte” y “misil”. El sexto tópico pedido a esta metodología, en cambio vuelve a la idea del riesgo asumido por el mundo combinados con términos de medicina.

Figura 9:
Composición del Tópico 1 bajo LDA. Menciona proyectos energéticos y energías renovables



Elaboración propia utilizando la librería PyLDAvis (Sievert & Shirley, 2014). Nota: Las franjas celestes simbolizan la frecuencia del término en todo el corpus, las rojas dentro del tópico seleccionado.

3.2. Análisis de Sentimiento de la opinión pública

Ya obtenidos los temas principalmente tratados y continuando con las tareas de minería de opiniones a través de las expresiones en la red social Twitter, se centra la tarea en el análisis de sentimiento. Este tipo de análisis se centra en la detección automatizada de una clasificación polarizada: usualmente se segregan los comentarios en positivos, neutrales y negativos. Otros enfoques realizan clasificaciones afectivas sobre la felicidad o el estado de ánimo (Cambria &

Grassi, 2012). Principalmente, señalan estos autores, los esfuerzos computacionales se centran en la identificación de palabras que expresan estas clasificaciones. Así palabras positivas (bueno, lindo, excelente, mejor) y palabras negativas (malo, pobre, desafortunado, peor) serán determinantes para calificar la polaridad del texto.

Se utilizó la librería Pysentimiento que está entrenada en textos en idioma español, con la intención de facilitar los análisis de minería de opiniones (Pérez, Giudici & Luque, 2021). En su versión actual (lanzada en octubre 2021), permite realizar tareas de análisis de sentimiento, análisis de emociones y análisis de discurso de odio. El análisis de sentimiento brinda tres etiquetas: Neutral (1669 Tweets), Negativo (1117) o Positivo (228). A priori estos resultados confirman la minoría de mensajes positivos. La provincia con mejor porcentaje de positivos fue Formosa, con un 20%, pero sólo cuenta con 20 tweets, por lo que el resultado no resulta concluyente para realizar conclusiones. Considerando aquellas provincias con más de 50 tweets, Neuquén resultó la mejor posicionada en cuanto a la cantidad de mensajes calificados positivos mientas que Santa Fe resultó la más adversa, estos resultados pueden observarse en la siguiente tabla.

Tabla 2:
Análisis de sentimiento Porcentaje de tweets según polaridad por provincia con más de 50 mensajes.

Provincia	Cantidad	NEG	NEU	POS
BUENOS AIRES	1523	35,5%	55,4%	9,1%
CÓRDOBA	233	44,6%	50,2%	5,2%
SANTA FE	224	50,0%	45,1%	4,9%
MENDOZA	162	32,7%	61,1%	6,2%
SIN DATO	158	32,3%	65,2%	2,5%
RÍO NEGRO	108	36,1%	58,3%	5,6%
CABA	73	35,6%	56,2%	8,2%
CHUBUT	57	36,8%	56,1%	7,0%
NEUQUÉN	55	32,7%	56,4%	10,9%
CORRIENTES	51	35,3%	58,8%	5,9%
Nación	**3014**	**37,1%**	**55,4%**	**7,6%**

Cuadro de elaboración propia en base a los resultados obtenidos.

Al procurar combinar los resultados de tópicos con el análisis de polaridad realizado, los resultados no brindan diferenciación relevante que amerite tratar con diferencias un tema que otro. Prima siempre con más de la mitad de los tweets la opinión neutral, seguida de cerca por la opinión negativa. Con lo cual, no sugiere que los temas de energía sostenible (aun teniendo el mayor porcentaje de negativos), medicina nuclear (con el mayor porcentaje de positivos) o

los proyectos y otras aplicaciones sean tratados diferentes a los temas bélicos o temores por los riesgos que implica la tecnología.

Tabla 3:
Cantidad porcentual de los mensajes por tópico (biterm) y Polaridad detectado.

Tópico	Cantidad	%tweets	NEG	NEU	POS
Uso Belico y Armamentístico	677	32,5%	32%	62%	5%
Energías renovables	674	22,4%	45%	50%	5%
Aplicaciones Tecnológicas	578	19,2%	39%	53%	8%
Medicina Nuclear	543	18,0%	34%	55%	11%
Riesgos	542	18,0%	35%	56%	9%
Todos los temas	**3014**		**37%**	**55%**	**8%**

Cuadro de elaboración propia en base a los resultados obtenidos.

De esta forma se determina que del análisis de sentimiento predominan en el país los comentarios neutros, esta situación no tiene diferencias al considerar el análisis por tópicos, pero, al realizar un análisis por provincia pueden verse algunas con mayor preponderancia de negativos.

El análisis de discurso, en cambio, brinda porcentajes para dar indicios si los tweets expresan odio (400), son misóginos o racistas (14) o agresivos (71). En este dato es de consideración que haya un 13% de discurso de odio, y un 2% de mensajes agresivos. Respecto al discurso de odio en general, los tópicos donde mayor incidencia tuvo fueron los de Riesgos (15,6%), uso bélico (14,2%) y el de menor incidencia fue el de las aplicaciones tecnológicas en general (11,1%).

Por su parte del resultado de la clasificación por emociones, la librería no pudo hacer una determinación muy variada, asignando en todo el país la mayoría a la clasificación “Otros” que es la categoría utilizada cuando no puede determinarse si el tweet tiene expresiones definidas como “ira” (355), “alegría” (132), “sorpresa” (40), “tristeza” (36), “disgusto” (5) y “miedo” (3). La clasificación “Otros” con 2443 tweets representa más de un 81% de los tweets, por lo que no pueden realizarse conclusiones sobre este dato en particular, sino que podría servir de indicador configurado como un registro temporal que vaya midiendo las variaciones.

3.3. Consideraciones para el plan de comunicación organizacional

La comunicación nuclear es necesaria para mejorar la percepción pública sobre energía nuclear. La distancia entre la organización y la opinión pública se disminuye con la transmisión de información precisa, comprensible y creíble de divulgación científica. Además, el plan debe ser proactivo, es decir, tomando la iniciativa de transparencia. (Cobos Urbina, 2018). Este autor, referenciando a especialistas en el tema, indica además que parte de la estrategia comunicacional es incorporarse activamente a las redes sociales, ya que el público se encuentra inmerso en la multicanalidad; donde lo importante no es sólo hablar sino también escuchar, ya que la reputación se construye a partir de la interacción. Esta reputación puede volverse una ventaja competitiva, señala Cobos en su desarrollo.

El primer hallazgo encontrado durante el análisis exploratorio es la baja cantidad de Tweets recuperados referidos a la temática nuclear que pudieron ser identificados al país en 4 meses de recabar información. Este podría ser un indicador de desinformación y desinterés, sobre todo considerando que en el período abarcado se celebró la COP26, y la iniciativa Stand Up For Nuclear. El primero es la Conferencia de las Naciones Unidas sobre el Cambio Climático de 2021, durante la cual la industria nuclear internacional reflejó los beneficios de su tecnología y la segunda es un evento descentralizado que procura la difusión de la tecnología bajo el lema “átomos por el clima”.

Lo segundo es que la organización tiene bajo ratio de menciones (el usuario más mencionado llegaba al 1% de los tweets y no es un referente argentino). Por lo tanto, el plan comunicacional debe establecer metas de las interacciones con los otros usuarios, para adoptar esta actitud proactiva y de escucha al público. Esto no se remite únicamente a Twitter, el concepto está vigente en cualquier red social. Además, considerando la sinergia y cercanía de otras organizaciones del sector, podría considerarse mayor interacción conjunta en redes a fin de promocionarse mutuamente. La organización podría organizar acciones de difusión cruzada entre las diferentes organizaciones.

En tercera instancia, hay que considerar que no parece existir un referente nacional con el mismo alcance que el usuario mencionado en el párrafo anterior. Por tanto, se abre la posibilidad de facilitar la información necesaria a *influencers* que realicen divulgación científica para la elaboración de sus posteos o hilos. Esto posiblemente requiera una revisión de las políticas, procedimientos o instructivos internos para asegurarse que no haya inhibición de los posibles colaboradores con conocimiento necesario para volverse referentes en las redes, a la vez de preservar la rigurosidad técnica.

Del análisis de tópicos realizado, se detecta que la comunicación nuclear debe incluir temáticas relacionadas a dar en conocimiento las aplicaciones pacíficas y científicas de la energía nuclear, y a procurar poner en conocimiento de la dimensión y mitigación de los riesgos de esta tecnología, así como la desmitificación. También se destacó la relación de la tecnología nuclear con energías renovables, tanto por el conocimiento de la experiencia internacional como nacional.

Thill & Bovee (2021) sugieren que se puede aplicar un enfoque centrado en la audiencia que implica comprender y respetar al público de la red social, haciendo todo lo posible para transmitir los mensajes de una manera que sea significativa para ellos. Para este enfoque es fundamental conocer que temas le interesan y allí es donde las conclusiones del modelado de tópicos efectuado se hacen relevantes. Indican, además que relacionarse con las necesidades de los demás es una parte clave de la inteligencia emocional, la capacidad leer las emociones de otras personas de manera productiva a la comunicación, para lo cual hemos realizado el análisis de sentimiento en varios aspectos. La métrica de polaridad y de emociones, puede ser utilizada como referencia para el seguimiento general

Consecuentemente, a raíz del trabajo realizado se reconoce a los indicadores necesarios para el seguimiento del éxito del plan de comunicación que elabore la organización, como ser: la medición y determinación de la cantidad de los tweets y menciones, los resultados del modelado de tópicos (en detalle) y de sentimientos (incluyendo emociones), tanto con la información a nivel nacional como la segregada por provincia.

Conclusión

Como resultado del trabajo realizado se han identificado los tópicos relevantes sobre el desarrollo nuclear en las expresiones de los usuarios de Twitter realizadas en Argentina y determinado la polaridad de estas expresiones. Esta información tiene utilidad como insumo para el diseño de una campaña comunicacional de la organización a fin de incrementar la adhesión positiva a los proyectos en curso, por tanto, se considera que ha cumplido el propósito general del trabajo.

En el primer apartado se han establecido las necesidades de información de la organización, que son costosas de obtener por métodos tradicionales ya que serían necesarias encuestas con un alcance nacional. A cambio, se ha determinado que una opción operativa más económica es la opción de utilizar los datos obtenidos a través de la red social Twitter. Como estos datos alternativos, en formato semi-estructurado, contienen un campo de texto que, en su estado original, no aporta valor a la compañía se plantearon los procesos necesarios para su transformación en información.

Estos procesos fueron detallados en el segundo apartado, incluyendo la ingesta de datos, la limpieza y transformación como enriquecimiento, normalización, y otros, según los objetivos planteados. Se han mostrado los resultados de recopilar los tweets relacionados a la energía nuclear a partir de mediados de agosto hasta el 22 de noviembre 2021 conformando una base de 3014 tweets localizados en argentina luego de las operaciones de limpieza. El manejo de estos datos requirió normalización como utilización de minúsculas, eliminación de palabras de uso común y sin significado (*stopwords*), lematización, descarte de palabras cortas, entre otros procesos de lenguaje natural empleados. En atención a las necesidades informativas de la organización, se aplicó una lógica heurística a fin de aproximar geográficamente el origen de los tweets; para lo cual, se hizo uso entre otras técnicas, de la API del Servicio de Normalización de Datos Geográficos (https://datosgobar.github.io/georef-ar-api/) a partir de los textos de ubicación proporcionados por los mismos usuarios de la red social. Luego de estos procesos de carga, transformación y limpieza de los datos, se realizó el análisis exploratorio para determinar una nube de palabras más utilizadas, los usuarios más involucrados con los temas, y otros datos estadísticos que permitieron conocer mejor la base de datos de tweets y el corpus léxico que han sido preparado.

Por último, en el tercer apartado, se determinó el modelado de tópicos para conocer la asociación entre las temáticas de las opiniones de los usuarios, y luego aproximar las polaridades de los mensajes obtenidos, estableciéndose métricas necesarias para la

conformación de un plan de comunicación y su seguimiento. Los aportes dados se resumen en llegar al público con la utilización pacífica de la tecnología nuclear, su impacto favorable en la descarbonización en el medio ambiente, el soporte que hace como energía de base a las tecnologías de generación renovable (solar, eólica), la divulgación científica sobre los temas de seguridad de las instalaciones. Luego, el principal sentimiento obtenido en el estudio es la neutralidad, con diferencias interprovinciales en cuanto a la relación de positividad/negatividad. Se determinaron como métricas relevantes para el seguimiento continuo de la percepción pública: la cantidad de tuits recopilados, la cantidad de respuestas/menciones, los tópicos a nivel detallado, la polaridad y emoción determinadas por provincia.

Los resultados son de aplicación inmediata en la organización, además de los organismos públicos y otras empresas del sector nuclear, cuya acción en medios sociales les permite un alcance comunicacional de amplio alcance, pero requiere dejar de lado los esfuerzos aislados y una planificación de acciones que brinde continuidad a la reputación organizacional, a la divulgación científica, y al conocimiento desmitificado de los alcances de la tecnología nuclear.

Las líneas de trabajo futuras, pueden dividirse según el aspecto a mejorar. Con un criterio técnico, es viable incluir la mejora en los diccionarios de *stopwords* y lematización en español, el entrenamiento de modelos de sentimientos entrenados a partir de una base propia a la jerga argentina. Con un criterio organizacional, los objetivos pueden llevar a buscar mejor cobertura de determinadas áreas geográficas, y la recopilación de datos de otras redes sociales, como por ejemplo la inclusión de comentarios en Instagram, blogs, plataformas de video como Youtube, entre otras.

Referencias bibliográficas

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) pp. 30-38.

Bello, H (2021) sentiment_analysis_spanish. Github. https://github.com/sentiment-analysis-spanish/sentiment-spanish

Blei, D. M. (2012). *Probabilistic topic models. Communications of the ACM, 55(4), 77.* doi:10.1145/2133806.2133826

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.iculo.oa?id=515751524008

Bird, Steven, Edward Loper and Ewan Klein (2009) *Natural Language Processing with Python*. O'Reilly Media Inc.

Bovée & Thill (2021). Business Communication Today. Global Edition Pearson.

Buneman, P. (1997). *Semistructured data.* Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems - *PODS '97*. doi:10.1145/263661.263675

Burscher, B., Vliegenthart, R., & Vreese, C. H. de. (2016). Frames Beyond Words. Social Science Computer Review, 34(5), 530–545. doi:10.1177/0894439315596385

Cambria, E., Grassi, M., Hussain, A., & Havasi, C. (2012). Sentic computing for social media marketing. Multimedia tools and applications, 59(2), 557-577.

Chahab, M. (2016) Imágenes y símbolos en la opinión pública argentina sobre la energía nuclear y el medio ambiente: 'La necesidad de una nueva estrategia comunicacional'. México.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 26(12), 2928-2941.

Comisión Nacional de Energía Atómica (2017). Proyecto CAREM. "CAREM escala comercial". Revista de la CNEA. Año XVII, Número 67-68, julio/diciembre

Google Research (2017). Te damos la Bienvenida a Colaboratory. https://colab.research.google.com/notebooks/welcome.ipynb?hl=es-419

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning 42(1-2), 177-196. Machine Learning. 42. 177-196. 10.1023/A:1007617005950.

Honnibal M., Montani I., Van Landeghem S, Boyd A. (2020) *spaCy: Industrial-strength Natural Language Processing in Python*. DOI: 10.5281/zenodo.1212303

Jefatura de Jefe de Ministros (2021) Informe 129 Honorable Cámara de Senadores de la Nación. https://www.argentina.gob.ar/sites/default/files/informe_129_-_hsn.pdf

Jefatura de Jefe de Ministros (2021) Informe 130 Honorable Cámara de Diputados de la Nación. https://www.argentina.gob.ar/sites/default/files/informe_130-_hdn.pdf

Jo T. (2019). *Text Mining. Concepts, Implementation, And Big Data Challenge,* (Vol. 45). Springer. https://doi.org/10.1007/978-3-319-91815-0

Jungherr A. (2016) Twitter use in election campaigns: A systematic literature review, Journal of Information Technology & Politics, 13:1, 7291, DOI: 10.1080 / 19331681.2015.1132401

Kulkarni A. & A. Shivananda A.(2021), *Natural Language Processing Recipes* (2nd Edition). Apress.. https://doi.org/10.1007/978-1-4842-7351-7

Kwartler, T (2017) *Text Mining in Practice with R.* John Wiley & Sons Ltd

Liu, Z., & Na, J.-C. (2018). Aspect-Based Sentiment Analysis of Nuclear Energy Tweets with Attentive Deep Neural Network. Maturity and Innovation in Digital Libraries, 99–111. doi:10.1007/978-3-030-04257-8_9

Mueller A., Fillion-Robin J.C., Boidol R., et all (2018). amueller/word_cloud: WordCloud 1.5.0 (1.5.0). Zenodo. https://doi.org/10.5281/zenodo.1322068

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, No. 1, pp. 41-48).

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. *arXiv preprint arXiv:2106.09462* https://github.com/pysentimiento/pysentimiento

Polino, C. y Fazio, M. E. (2009). Energía nuclear en Argentina: opinión pública y riesgo percibido. En Moreno Castro, C. (Ed.) Comunicar los riesgos. Ciencia y tecnología en la sociedad de la información (65-84). Madrid: Biblioteca Nueva

Ranco G, Aleksovski D, Caldarelli G, Grčar M, Mozetič I (2015) *The Effects of Twitter Sentiment on Stock Price Returns*. Plos One 10(9): e0138441. https://doi.org/10.1371/journal.pone.0138441

Roesslein, J. (2020). *Tweepy: Twitter for Python!* https://Github.Com/Tweepy/Tweepy

Sampietro N. (2018) Datos Geográficos: un servicio de normalización. *Datos Argentina (en Medioum*, https://medium.com/datos-argentina/datos-geogr%C3%A1ficos-un-servicio-de-normalizaci%C3%B3n-eecd32fe4d8d

Salaberry N. (2019). *Detección de problema ticas en el uso de la tarjeta SUBE. Un análisis y clasificación de tweets.* Trabajo Final de Maestría -UBA FCE.

Sammy A. (2018) *Auditando modelos analíticos*. Internal Auditor Magazine https://iaonline.theiia.org/2018/Pages/Auditing-Analytic-Models.aspx

Sarasa A. (2016) *Introducción a las bases de datos NoSQL usando MongoDB*. Editorial UOC

Secretaría de Ciencia, Tecnología e Innovación Productiva (2006) La percepción de los argentinos sobre la investigación científica en el país. Segunda Encuesta Nacional.

Segarra, S. M. (2020). *Big data: la revolución de los datos y su impacto en la comunicación corporativa*. Comunicación y Hombre, (16), 115-132.

Statista Research Department (Sep 7, 2021) *Leading countries based on number of Twitter users as of July 2021.* https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. *In Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

Paul, M. J., & Dredze, M. (2011, July). *You are what you tweet: Analyzing twitter for public health*. In Fifth international AAAI conference on weblogs and social media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2011) *Scikit-learn: Machine Learning in Python*. JMLR 12, pp. 2825-2830.

Tajima, K. (2013). Schemaless semistructured data revisited. *In Search of Elegance in the Theory and Practice of Computation*, 466-482.

The IIA (2020) *El Modelo De Las Tres Líneas Del IIA 2020.* https://na.theiia.org/translations/PublicDocuments/Three-Lines-Model-Updated-Spanish.pdf

Twitter (2021) *More about restricted uses of the Twitter APIs.* https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual. Scotts Valley,* CA: CreateSpace.

Warner T. (2020, Diciembre 17). Understanding XML: The Human's Guide to Machine-Readable Data. *Safe Software Blog.* https://www.safe.com/blog/2016/07/understanding-xml-humans-guide-machine-readable-data/

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 60(6). https://doi.org/10.21105/joss.03021

Apéndices

Apendice 1 – Formato JSON

Un ejemplo completo de tuit en formato JSON extraído con la librería `tweepy` de *Python*.


```
{
 "created_at": "Sun Oct 17 23:03:07 +0000 2021",
 "id": 1449873639130779653,
 "id_str": "1449873639130779653",
 "full_text": "CAMBIO CLIM\u00c1TICO: \u00bfLA ENERGIA NUCLEAR, ALIADA CONTRA
     EL CAMBI... https://t.co/Afo2zxnTB5",
 "truncated": false,
 "display_text_range": [
 0,
 96
 ],
 "entities": {
 "hashtags": [],
 "symbols": [],
 "user_mentions": [],
 "urls": [
 {
 "url": "https://t.co/Afo2zxnTB5",
 "expanded_url": "https://crisisambiental-
     cambioclimatico.blogspot.com/2021/10/la-energia-nuclear-aliada-contra-
     el.html?spref=tw",
 "display_url": "\u2026mbiental-cambioclimatico.blogspot.com/2021/10/la-
     ene\u2026",
 "indices": [
 73,
 96
 ]
 }
 ]
 },
 "metadata": {
 "iso_language_code": "es",
 "result_type": "recent"
 },
 "source": "<a href=\"https://mobile.twitter.com\" rel=\"nofollow\">Twitter
     Web App</a>",
 "in_reply_to_status_id": null,
 "in_reply_to_status_id_str": null,
 "in_reply_to_user_id": null,
 "in_reply_to_user_id_str": null,
 "in_reply_to_screen_name": null,
 "user": {
 "id": 570927159,
 "id_str": "570927159",
 "name": "Juan Jos\u00e9 Olivieri",
 "screen_name": "COOL_JJO",
 "location": "Buenos Aires - Argentina",
 "description": "Soy Ingeniero Qu\u00edmico, con una Maestr\u00eda en
     Sociolog\u00eda, y un posgrado en Econom\u00eda Ambiental. Estoy
     investigando el Cambio Clim\u00e1tico, y tengo un grupo en Linkedin",
 "url": "http://t.co/bLxB19AV1j",
 "entities": {
 "url": {
 "urls": [
 {
```

```
"url": "http://t.co/bLxB19AV1j",
"expanded_url": "http://www.crisisambiental-
    cambioclimatico.blogspot.com.ar/",
"display_url": "\u2026ental-cambioclimatico.blogspot.com.ar",
"indices": [
0,
22
]
}
]
},
"description": {
"urls": []
}
},
"protected": false,
"followers_count": 58,
"friends_count": 89,
"listed_count": 5,
"created_at": "Fri May 04 13:47:46 +0000 2012",
"favourites_count": 1,
"utc_offset": null,
"time_zone": null,
"geo_enabled": false,
"verified": false,
"statuses_count": 2897,
"lang": null,
"contributors_enabled": false,
"is_translator": false,
"is_translation_enabled": false,
"profile_background_color": "C0DEED",
"profile_background_image_url":
    "http://abs.twimg.com/images/themes/theme1/bg.png",
"profile_background_image_url_https":
    "https://abs.twimg.com/images/themes/theme1/bg.png",
"profile_background_tile": false,
"profile_image_url":
    "http://pbs.twimg.com/profile_images/2296936771/lwkwsj2jo6muswtqao2v_norm
    al.png",
"profile_image_url_https":
    "https://pbs.twimg.com/profile_images/2296936771/lwkwsj2jo6muswtqao2v_nor
    mal.png",
"profile_link_color": "1DA1F2",
"profile_sidebar_border_color": "C0DEED",
"profile_sidebar_fill_color": "DDEEF6",
"profile_text_color": "333333",
"profile_use_background_image": true,
"has_extended_profile": false,
"default_profile": true,
"default_profile_image": false,
"following": false,
"follow_request_sent": false,
"notifications": false,
"translator_type": "none",
"withheld_in_countries": []
},
"geo": null,
"coordinates": null,
"place": null,
"contributors": null,
"is_quote_status": false,
"retweet_count": 0,
"favorite_count": 0,
"favorited": false,
"retweeted": false,
"possibly_sensitive": false,
"lang": "es"
}
```

Apéndice 2 – Índice de Figuras

Trabajo Final de Especialización de Adriana Cruz:

"IMPACTO DE LA OPINIÓN PÚBLICA SOBRE EL DESARROLLO DE LA ENERGÍA NUCLEAR.
Un análisis de tópicos y sentimientos de las expresiones de los usuarios de la red social Twitter en Argentina."

Mentora: Natalia Salaberry

Fecha: diciembre 2021

La alumna Adriana Cruz en su trabajo final de especialización logra desarrollar de manera concisa y acabada su proyecto inicial de trabajo. Plantea de manera clara la existencia de una necesidad clara en una organización dedicada a la explotación de energía nuclear que luego resuelve con las herramientas aprendidas en el campo diciplinar de la especialización en Métodos Cuantitativos para la Gestión y Análisis de Datos en Organizaciones. En este sentido, el problema propuesto a resolver se vincula a la falta de información acerca de la opinión ciudadana siendo clave para la estrategia comunicacional que derive en la aprobación de proyectos de explotación de energía nuclear en Argentina. A partir de esto, se propone incorporar la utilización de datos no estructurados constituyendo un alto desafío de análisis.

Para cumplimentar con el objetivo planteado, en un primer apartado desarrolla de manera clara y acabada la concepción de los datos no estructurados advirtiendo sobre el desafío que implica su gestión y posterior utilización en una organización. La claridad de la exposición permite comprender fácilmente los conceptos involucrados que son clave en este desafío. Dada la complejidad de operar con estos, selecciona una red social para la obtención de los datos base y realiza una selección de técnicas a implementar tanto para el procesamiento como el posterior análisis de los datos no estructurados.

En un segundo apartado, desarrolla con una clara exposición el complejo procesamiento de obtención y posterior procesamiento de datos no estructurados, demostrando haber adquirido un conocimiento acabo en la temática tratada. Esto le permite finalmente obtener los datos en una forma estructurada para realizar un análisis descriptivo de los mismos. Desde el aspecto técnico de implementación, realiza un trabajo perfecto, respetando todas las etapas necesarias de extracción, transformación y análisis descriptivo.

Finalmente, en un tercer apartado avanza en la aplicación de las técnicas seleccionadas para la obtención de resultados. A partir de una correcta implementación técnica y posterior interpretación de estos logra cumplir con su objetivo principal exponiendo como la incorporación de estos datos le permitirá a la organización gestionar eficientemente la comunicación institucional. Al mismo tiempo, advierte sobre la necesidad de ampliar las fuentes origen de datos implicando un desafío aún mayor para la organización en términos de una gestión eficiente.

De esta manera, la alumna logra cumplimentar con el objetivo principal y los específicos de forma ordenada y clara, realizando un trabajo completamente estructurado en el marco de las pautas establecidas. Existe una coherencia entre la problemática planteada, el título y las palabras claves lo que le permitió poder elaborar un planteo y desarrollo adecuado del trabajo. Un aspecto a mejorar en futuros trabajos que se sugiere es ahondar en las

problemáticas asociadas a incorporar este tipo de datos en una organización con el fin de llevar a cabo una gestión responsable de datos.