Universidad de Buenos Aires
Facultad de Ciencias Económicas
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 de graduados en la gestión universitaria. Un análisis de las expresiones en la red social Twitter.

Aplicación de algoritmos de minería de texto

Autor: Laura Ester Marmolejo

Mentora: Mg. Natalia Salaberry

Diciembre 2022

Resumen

Actualmente "la universidad[1]" gestiona sus procesos estratégicos valiéndose de información que surge de datos estructurados. No obstante, los alternativos representan una porción significativa, siendo un recurso desaprovechado en la gestión estratégica. Una variada cantidad de ellos provienen de redes sociales, pero actualmente no hay procesamiento de éstos. La organización simplemente utiliza las cuentas de *Twitter* como canal de comunicación de eventos y hechos relevantes. Como consecuencia de lo expuesto, pierde un recurso fundamental para ponerla a la vanguardia de los retos que plantea la era digital.

El presente trabajo tiene por objeto detectar tópicos que expongan la percepción social sobre la formación de graduados universitarios de una organización educativa de nivel superior situada en la provincia de Mendoza. Para ello se efectuará un relevamiento de *Tweets* y se aplicarán algoritmos de minería de textos para generar conocimiento. De esta manera se identificarán los principales tópicos de interés tratados en mensajes publicados por usuarios que arroban a cuentas oficiales de la organización. Finalmente se evaluará el aporte de la información obtenida a través de este proceso reflexionando sobre los desafíos que plantea el uso de los datos como activo estratégico.

Palabras clave: Opinión de graduados, Gestión universitaria, *Twitter*, Algoritmos, Minería de texto.

[1] Se reserva en nombre de la institución sobre la que se realizará el presente trabajo, por no contar con autorización expresa para mencionarla. En adelante para referirse a la institución objeto de análisis se utilizará el término "la universidad" en letras minúsculas y utilizando comillas.

Índice

Introducción **4**

1. El Big Data y las instituciones universitarias **7**

1.1. Datos no estructurados, su caracterización y tecnologías involucradas. 8

1.2. La gestión de datos no estructurados en organizaciones. 10

1.3. El Text Mining como herramienta de valor para organizaciones universitarias. 14

2. Los algoritmos de *Text Mining* para la detección de tópicos **17**

2.1. Obtención y procesamiento de tweets 18

2.2. Análisis descriptivo de los datos 20

2.3. Algoritmos de *Text Mining* para la detección de tópicos 21

3. Detección eficiente de tópicos en vinculación con la formación recibida en la universidad mendocina **25**

3.1. Aplicación de métodos para la detección de tópicos 26

3.2. Evaluación y análisis de resultados 31

3.3. Relevancia del análisis de *tweets* para la gestión universitaria 33

Conclusión **35**

Referencias bibliográfica **39**

Introducción

La tecnología ha impactado a las organizaciones afectando fuertemente su competitividad y productividad. Esto ha generado procesos de cambios organizacionales que posibiliten responder a las actuales necesidades de un mercado cambiante y globalizado. Estas modificaciones incluyen las diferentes capas de la arquitectura empresarial e involucra la redefinición de sus modelos de negocio, la gestión de sus datos, la implementación de aplicaciones y la infraestructura tecnológica necesaria para apoyar los procesos y servicios prestados (Josey et al., 2013).

Las organizaciones de educación superior están transitando estos cambios y resulta interesante identificar los desafíos a los que se enfrentan. Entre ellos se encuentra analizar como utilizan los aportes que brinda la tecnología como herramienta para mejorar la gestión organizacional. Es necesario reflexionar el impacto de la aplicación de tecnologías de información y comunicación (TIC) en los métodos de enseñanza aprendizaje, o en la generación de información oportuna para modificar planes de estudio, programas de espacios curriculares o generar nuevos espacios de formación que responda a las necesidades del mercado. Este proceso no es sencillo, y debe considerar como elemento prioritario la cultura organizacional, que permita hacer viable un camino de transición, lo que implicará involucrar a los diferentes agentes que conforman el sistema universitario.

Actualmente una universidad de la provincia de Mendoza usa sus redes sociales y medios digitales como canal de comunicación con la sociedad y difusión de diferentes actividades realizadas. Pero éstos no son aprovechados como herramienta para generar información que permita de manera sistemática identificar situaciones problemáticas como factor clave de gestión. Las respuestas de los usuarios capturadas de manera oportuna surgen como potencial valor para reorientar ejes de acción estratégicos a ser desarrollados por la organización, como la revisión de las competencias desarrolladas por sus egresados.

Si bien “la universidad” actualmente procesa grandes volúmenes de datos estructurados, no utiliza la potencialidad de datos alternativos para acompañar sus procesos cotidianos. Por ello, la aplicación de un conjunto de métodos para capturar, procesar y obtener información mediante el análisis de *tweets* puede resultar oportuno. En este contexto se plantea el siguiente interrogante: ¿Qué tópicos de interés son planteados por la sociedad mendocina a través de la

red social *Twitter* respecto a la formación de sus egresados para ser tratados por la universidad desde la vinculación, extensión e investigación universitaria?

El presente trabajo tiene por objetivo detectar tópicos referidos a la formación recibida en "la universidad" mendocina a partir de los comentarios realizados en la red social *Twitter* y determinar la vinculación entre sus planes de estudios y las capacidades demandadas en el mercado laboral. Para alcanzarlo se desarrollarán aspectos conceptuales de la gestión de datos no estructurados para generar valor en la organización universitaria y las tecnologías necesarias para su implementación. Por otro lado, se aplicarán algoritmos de *Tex Mining* que permitan identificar tópicos de interés social en vinculación con la formación recibida en "la universidad". A partir de este análisis metodológico, se evaluará su contribución a la redefinición de los planes de formación en la organización universitaria en función de las necesidades profesionales demandadas.

Para poder alcanzar el objetivo planteado, el trabajo se organiza en tres capítulos. El primero de ellos tiene por objetivo describir el fenómeno del *Big Data* en una institución universitaria constituyendo un instrumento para generar valor en el marco de su gestión. El recorrido comenzará con el análisis conceptual de los datos alternativos y la identificación de las tecnologías necesarias para su gestión. Luego, se tratarán los elementos claves a considerar para su gobernanza considerando la gestión del ciclo de vida, la gestión de calidad del dato, la gestión de la seguridad y privacidad. Finalmente se analizará como los algoritmos de minería de textos pueden considerarse como una herramienta que genera valor para la toma de decisiones.

El segundo capítulo tiene por objetivo presentar y describir los datos a utilizar así como los métodos con los cuales se trabajará. En primer lugar, se describirá el proceso de obtención de *tweets* y el tratamiento necesario del texto obtenido mediante el procesamiento del lenguaje natural (*NLP*). Luego se efectuará un análisis exploratorio y descriptivo de los datos recolectados. Finalmente, se especificarán, los algoritmos de Words Cloud, Sentyments Analysis, y *Latent Dirichlet Allocation* (LDA) a utilizar.

En el tercer capítulo se establecerá cómo los resultados obtenidos de la aplicación de los algoritmos seleccionados constituyen una técnica para detectar tópicos vinculados a la formación recibida en "la universidad". Se comenzará con su proceso de aplicación. Luego, se realizará la evaluación de los resultados obtenidos como aporte a la solución del problema planteado en el presente trabajo. Finalmente se especificarán los aspectos que la institución

analizada debiera considerar para generar información a través de datos que surgen de la red social Twitter, para que esta se transforme en una herramienta que contribuya a su gestión estratégica.

El análisis realizado permitirá realizar una aproximación que permita evidenciar cómo la aplicación de técnicas de *Text Mining* constituye una herramienta para identificar áreas de vacancia en una universidad de la provincia de Mendoza. La gestión de los comentarios de los usuarios de la red social resulta necesaria para generar información de valor en la organización universidad. El análisis de textos mediante diferentes algoritmos permite identificar tópicos relevantes referidos a las actuales capacidades del egresado universitario y la percepción social sobre su formación. De esta manera, la organización contará con información relevante para redefinir los actuales planes de estudio que doten al egresado de las habilidades que requiere actualmente el mercado laboral.

1. El Big Data y las instituciones universitarias

La sustentabilidad de las organizaciones requiere actualmente de un proceso de gestión de datos obtenidos de fuentes externas e internas al efecto de aplicar herramientas que colaboren en la toma de decisiones. La analítica convencional contempla el uso de datos estructurados, que constituyen grupos estáticos, utilizando métodos de análisis basado en hipótesis y cuyo propósito es brindar soporte a las decisiones internas aplicando la estadística y matemática. (Davenport, 2014). No obstante, este enfoque deja fuera las potencialidades que brinda el *Big Data*, que incluye el análisis de grandes volúmenes de datos de diferentes tipos, que son tratados como flujos y utiliza métodos de aprendizaje automático para generar productos basados en datos. (Davenport, 2014)

El *Big Data* es un fenómeno asociado a la volumetría de la cantidad de bytes de información generados en espacios digitales, aunque no es un factor determinante de este concepto, por cuanto lo que hoy se considera grande, en poco tiempo dejará de serlo. Lo que importa en realidad, es la masividad de los datos generados como consecuencia de los avances tecnológicos al efecto de su caracterización. (Davenport, 2014). Por otro lado, el fenómeno está impactado por el tipo de datos, en los que los no estructurados toman relevancia con la aparición de internet. Éstos pueden provenir de diferentes fuentes, tales como las redes sociales y diversos dispositivos conectados (internet de las cosas) como teléfonos, televisores, sistemas de seguridad, termostatos, sensores para animar objetos o seres humanos y se caracterizan por su rápido movimiento (Davenport, 2014). Otra situación para considerar es el flujo en el que se mueven los datos dentro de una organización. Muchos son captados y procesados en procesos de *batch*, mientras que otros, por su velocidad requieren ser procesados en tiempo real (*real-time*) (Santiago y Zakhmen, 2022)

La digitalización ha impactado en las organizaciones, generando cambios profundos que han transformado organizaciones consolidadas o han generado un crecimiento exponencial en otras. Entre estas se encuentran las educativas de nivel superior, cuyo objetivo es la generación del conocimiento y formación de futuros profesionales. Actualmente se puede observar la existencia de un ecosistema de generación y distribución del conocimiento por fuera de las universidades, basado en internet y conformado por diversas plataformas de contenidos y de colaboración (Chinkes & Julien, 2019). Por otro lado, el mundo cuenta con una sociedad globalizada en la que los procesos de internacionalización han sido potenciados como

consecuencia de las tecnologías de la información y comunicación. Estos factores generan un nuevo desafío para las universidades, que debe replantear sus modelos de gestión considerando sus tres funciones claves: la docencia, la extensión y la investigación. (Chinkes & Julien, 2019)

Por ello, la analítica del *Big Data* es un recurso de vital importancia para la adaptación de las Universidades a los requerimientos de la sociedad actual. Esta debe ser incorporada en sus procesos estratégicos para mejorar la gestión educativa, el desarrollo de métodos para la enseñanza aprendizaje, la creación de nuevas carreras y perfiles profesionales entre otros aspectos (Argonza, 2016). Su implementación requiere un sistema de gobernanza de datos que contemple modificar su actual arquitectura, para adaptarla y brindar una respuesta adecuada a los procesos de ingesta, almacenamiento, procesamiento. Esto implica, trabajar sobre un cambio en la cultura organizacional, capacitando a su personal docente y no docente de manera que se integren a los actuales desafíos (Argonza, 2016)

En los siguientes subapartados se desarrollará conceptos necesarios para comprender la gestión de datos no estructurados en la generación de valor en una organización universitaria y determinar las tecnologías necesarias para su implementación. Para poder llevarlo adelante se analizará el impacto de los datos no estructurados en la gestión de instituciones de educación superior, caracterizando los mismos y las tecnologías involucradas para su tratamiento. Asimismo, se analizarán los elementos que caracterizan la gobernanza de los datos alternativos. Finalmente se tratará la importancia de utilizar algoritmos de minería de textos como activo de valor en las organizaciones universitarias.

1.1. Datos no estructurados, su caracterización y tecnologías involucradas.

Los datos están presentes desde hace tiempo en la vida organizacional y se presentan bajo diferentes formas. Los estructurados son aquellos de mayor facilidad para acceder. Los mismos poseen una estructura especificada que incluye una colección finita de elementos en formatos definidos del mismo tipo. Son homogéneos y ordenados por un índice (Camargo Vega, 2015). Por otro lado, se observan los semiestructurados en los que el esquema de datos existe implícitamente en la instancia del dato, pero podría evolucionar y existir un esquema a posteriori a la existencia de éste (Santiago y Zakhmen, 2022). Los datos no estructurados son aquellos que no tienen tipos definidos ni están organizados bajo algún patrón. Tampoco son almacenados de manera relacional o con base jerárquica de datos, debido a que no son un tipo de dato predefinido. Éstos se pueden observar a diario en correos electrónicos, archivos de

texto, un documento de algún procesador de palabra, hojas electrónicas, una imagen, un objeto, archivos de audio, blogs, mensajes de correo de voz, mensajes instantáneos, contenidos *Web* y archivos de video, entre otros (Camargo Vega, 2015).

Las instituciones de nivel superior y en particular "la universidad"[i] sobre la cual se realizará el presente trabajo cuenta con datos generados por los individuos, datos generados o producidos por la organización, como aquellos generados por sensores que reúnen estas tipologías. Los primeros están representados en buena parte por texto y se distribuyen a través de múltiples plataformas tales como las redes sociales, páginas webs, aplicaciones móviles, correos electrónicos entre otros (Kolanovi M.; Krishnamachari R., 2017). Los datos recopilados por la propia organización surgen de registros operativos relacionados, por ejemplo, con la evolución académica de los estudiantes en sistemas como SIU Guaraní o datos presupuestarios contenidos en el sistema de gestión presupuestaria. Éstos tienen naturaleza estructurada. También cuenta con los generados a través de sensores integrados en varios dispositivos conectados a computadoras o tecnologías inalámbricas que se caracterizan por ser no estructurados. (Kolanovi y Krishnamachari, 2017). Entre éstos se encuentran los que surgen de dispositivos y aplicaciones que conectan a docentes con estudiantes ubicados en distintos puntos geográficos que viabiliza el dictado de clases híbridas, que son grabadas y subidas a la plataforma educativa. Los registros de ingreso y salida de personal o videos capturados por cámaras de seguridad ubicadas en el predio universitario son otro ejemplo de estos datos.

El tratamiento de datos alternativos dentro de la organización requiere un replanteo de la arquitectura[2] tradicional, de manera de lograr un diseño centralizado, alineada con los procesos comerciales o de gestión, que se adapta al crecimiento del negocio, y que evolucione con los avances tecnológicos (Santiago y Zakhmen, 2022). Además, resulta necesario el rediseño de la infraestructura de almacenamiento, acceso y análisis de volúmenes masivos de transacciones, integrando datos semi estructurados y no estructurados que agregue nuevas dimensiones, atributos dimensionales, métricas, informes y tableros de *Business Inteligent.* Ello con el fin de

[2] La arquitectura de datos son modelos, políticas, reglas y estándares que indican de qué manera se tienen que almacenar, organizar e integrar los datos que recoge una organización con el objetivo de que sean aprovechables y útiles (Santiago y Zakhmen, 2022)

acceder y procesar datos en tiempo real aplicando análisis predictivo de pronóstico y recomendaciones que se puedan integrar en los sistemas operativos de la organización (Schmarzo, 2013).

Entre las tecnologías de almacenamiento de datos no estructurados se encuentra el *Apache Hadoop*. Es un software de código abierto que admite aplicaciones nativas paralelas, distribuidas y con uso intensivo de datos. Admite la ejecución de aplicaciones en grandes clústeres de hardware básico mediante una arquitectura de escalamiento horizontal. *Hadoop* implementa un paradigma computacional llamado *MapReduce* donde la aplicación se divide en pequeños fragmentos de trabajo, cada uno de los cuales se puede ejecutar en cualquier nodo del *cluster*. Además, proporciona un sistema de archivo distribuido (llamado HDFS) que almacena datos en los nodos de cómputo y proporciona un ancho de banda agregado muy alto en todo el *cluster*. Tanto *MapReduce* como *HDFS*, están diseñados para que las fallas de los nodos se manejen automáticamente (Schmarzo, 2013).

Resumiendo lo expuesto, las organizaciones en general y "la universidad" cuenta con grandes volúmenes de datos como consecuencia del crecimiento de los no estructurados, cuyo tratamiento requiere un replanteo de la arquitectura tradicional. Esto es posible dado la potencialidad que brindan los actuales desarrollos tecnológico y que permiten un rediseño de su captura, almacenamiento y procesamiento que contemple la velocidad, variedad y volumen de los datos generados. En el próximo subapartado de desarrollaran qué elementos claves se deben considerar para realizar una buena gobernanza de datos, identificando aspectos relevantes vinculados al tratamiento de datos no estructurados.

1.2. La gestión de datos no estructurados en organizaciones.

Actualmente, el 80% de los datos con que cuentan las organizaciones son alternativos. Se encuentran disponibles a través de bases de datos no relacionales y requieren un proceso de gestión y gobernanza debido a sus particularidades que los distinguen de los estructurados. En el presente subapartado se caracterizarán las áreas claves de buen gobierno de datos que permita un tratamiento adecuado como activo estratégico.

Toda organización que desee realizar una gestión eficaz de los datos debe contar con un sistema de gobernanza de datos. Este es un proceso que compete a la dirección organizacional, que conlleva ejercer la toma de decisiones consensuada y comunicada, la autoridad y el control

sobre la gestión de los activos de datos (Serrano, 2022). Este sistema de gobernanza surge de la combinación de elementos (principios, políticas, equipos, instalaciones, programas informáticos, documentación técnica, servicios, personas) interrelacionados y organizados que permita la gestión estratégica del dato, para obtener una ventaja competitiva (Serrano, 2022). Su principal objetivo es la creación de nuevo valor a partir de los datos en función de los objetivos del negocio (Delloitte, 2011). El sistema de gobernanza está vinculado con tres áreas claves: la gestión del ciclo de vida, la gestión de calidad del dato, la gestión de la seguridad y privacidad e incluye determinar que puestos o roles están involucrados en el proceso de decisiones y cómo se relacionan entre ellos (Serrano, 2022).

Para materializar el valor de los datos, es necesario considerar los cambios que ocurren durante su ciclo de vida. La primera fase es la captura, que incluye la ingestión o integración de datos existentes en sistemas de la organización, la creación de datos específicos o la adquisición del dato a proveedores. La segunda se refiere a su almacenamiento de manera de hacerlos disponibles, estableciendo política de acceso y requerimientos de seguridad. La fase de procesamiento y mantenimiento incluye tareas de integración, limpieza y enriquecimiento, aplicando un determinado procesamiento analítico para agregarle valor al dato. En la fase de uso de los datos, los usuarios pueden recuperar datos, procesarlos, e integrarlos a otras aplicaciones para tomar decisiones. La fase final es el archivado, en la que está prevista la instancia de destrucción. La eliminación de los datos representa una operación importante en organizaciones reguladas y sobre todo en aquellas que mantienen información personal identificatoria *(Santiago y Zakhmen, 2022)*.

Las potencialidades del Big Data y las oportunidades vinculadas al tratamiento de los datos no estructurados han alterado algunos aspectos de las fases de captura y almacenamiento. El enfoque tradicional *ETL (Extract, Transform, Load)* se refiere a la extracción del dato de la fuente original para luego transformarlos mediante procesos de normalización, limpieza, agregación, para finalmente realizar su carga en los almacenes de datos. Pero el orden de este proceso se ve alternado cuando se tratan datos alternativos, que se producen a gran velocidad y sufren modificaciones durante su vida. Esto provoca que en muchos casos, el proceso de ingestión y procesamiento deba realizarse en tiempo real, para dar una respuesta adecuada que satisfaga los objetivos organizacionales. Por ello, es necesario aplicar un sistema de lectura y carga de los datos sin procesar (*ELT*), es decir sin definir un determinado esquema, agilizando una rápida ingestión de datos masivos. De esta manera se los podrá incorporar a procesos de

aprendizaje automático que permite crear nuevas métricas, dimensiones y atributos y mejorar la performance de los modelos de predicción y el valor de la información generada. (Schmarzo, 2013)

En relación a los modelos de almacenamiento, los datos no estructurados cuentan con distintos tipos de almacenes tales como los de datos de documentos; estos representan un conjunto de campos de cadena con nombre y valores de datos de un objeto en una entidad y son guardados en forma de documentos JSON. También se encuentran los de columnas que se caracterizan por un enfoque desnormalizado, que contiene datos dispersos en formato tabular con filas y columnas, pero las últimas se dividen en familias que contiene un conjunto de elementos que están relacionadas de forma lógica y que se recuperan o se manipulan como una unidad. El almacén de datos clave/valor es básicamente una tabla *hash*[3] que asocia cada valor de datos con una clave única y están optimizados para aplicaciones que realizan búsquedas simples mediante el valor de la clave, o por un intervalo de éstas. Un almacén de datos de grafos administra dos tipos de información: nodos que representan entidades y bordes que especifican las relaciones entre estas entidades facilitando la realización de consultas como "Buscar todos los empleados que dependen directa o indirectamente de Sarah" o "¿Quién trabaja en el mismo departamento que John?". Los almacenes de datos de series temporales son un conjunto de valores organizados por tiempo, admiten un número elevado de operaciones de escritura y recopila grandes cantidades de datos en tiempo real de diferentes orígenes. El almacenamiento de objetos está optimizados para guardar y recuperar objetos binarios grandes o blobs como imágenes, archivos, transmisiones de vídeo y audio, objetos de datos de aplicación de gran tamaño, documentos e imágenes de disco de máquina virtual y se compone de metadatos y un identificador único para acceder a él. Finalmente, los almacenes de datos de índice externo proporcionan la capacidad de buscar información que se encuentra en otros almacenes de datos y servicios y actúa como un índice secundario que puede utilizarse para indexar grandes volúmenes de datos y proporcionar acceso a ellos casi en tiempo real (Microsoft, 2022).

[3] Una tabla hash es una estructura de datos que asocia llaves o claves con valores. La operación principal que soporta de manera eficiente es la búsqueda: permite el acceso a los elementos (teléfono y dirección, por ejemplo) almacenados a partir de una clave generada usando el nombre, número de cuenta o id.

El ecosistema de datos actual caracterizados por su volumen, variedad y velocidad impone a las organizaciones la necesidad de tomar decisiones respecto de los datos que se encuentran almacenados en diferentes fuentes, para lo cual deberán definir el real aporte de mantener estos en los almacenes respectivos (Salaberry, 2019). El crecimiento exponencial de los datos, fundamentalmente de aquellos no estructurados, imponen la necesidad de revisar aspectos tales como su duplicidad en diferentes repositorios, su seguridad y el periodo de almacenamiento.

Para brindar una respuesta a las problemáticas planteadas en el párrafo anterior, es de fundamental importancia definir un sistema de datos maestros[4] que permita su adecuada gestión. Esto evitará problemas como fallas operativas, decisiones inadecuadas y hará posible que la totalidad de los datos centrales de una empresa sean únicos, consistentes, confiables y rastreables. Para su gestión es necesario definir una estructura datos maestros que se ocupe del diseño de sistemas que respalde cada etapa del ciclo de vida. También el establecimiento de un gobierno de datos maestros que defina su misión, las estructuras organizativas referidas a funciones y responsabilidades, actividades y área de decisión. Por otro lado, los procesos de datos maestros deben corresponderse con la arquitectura definida, de manera tal que la captura, uso, mantenimiento y archivo responda a ésta. Todo esto permite lograr calidad de los datos maestros respondiendo a estándares, políticas, métricas e indicadores de desempeño previamente definidos. (Clevent y Wortmann, 2010).

Otro pilar a considerar en la gobernanza de los datos es la seguridad, privacidad y riesgo de los datos. Algunas consideraciones para ello consisten en identificar los datos sensibles y establecer clasificaciones de estos según los niveles de seguridad definidos por la organización, teniendo en cuenta los requisitos propios del negocio, regulaciones, normativas y legislación que le son aplicables. También establecer políticas de autenticación y autorización que definan qué se puede hacer con los datos y quiénes pueden hacerlo. Otro aspecto es definir políticas que permitan la protección de la información efectuando *backup* y recuperación de datos. Las políticas de seguridad deben considerar la distribución de los datos a través de los diferentes

[4] Datos maestros son las entidades centrales del negocio u organización que se usan repetidamente en muchos procesos y sistemas. A diferencia de los datos transacciones, lo maestros no cambian durante el ciclo de vida, son independiente de otros objetos y resultan de un volumen relativamente constante. (Clevent A & Wortmann F, 2010)

sistemas de integración vertical y horizontal, incorporando estándares vinculados a las nuevas plataformas (Serrano, 2022).

Para resumir, el tratamiento de datos no estructurados es un desafío organizacional que impacta en el sistema de gobierno de los datos. Las características de volumen y velocidad imponen a las organizaciones modelos de ingesta y procesamiento en tiempo real que modifica el enfoque tradicional ETL, por uno que brinde reducción en periodos de latencia para ser procesados utilizando técnicas de aprendizaje automático. La incorporación masiva de estos datos para aplicarlas permitirá mejorar su performance, generando información que aporta valor a la organización. Por otro lado, se observa que el almacenamiento de los datos alternativo no responde a los incorporados en las bases de datos estructuradas y usan modelos optimizados para los requisitos específicos del tipo de datos que se almacena. Por ejemplo, los datos se pueden almacenar como pares clave/valor simple, como documentos *JSON* o como un grafo que consta de bordes y vértices. Otro aspecto para considerar es la calidad de los datos, que requiere implementar una gestión eficaz de datos maestros. Por último, se deben atender a políticas de seguridad, privacidad y riesgo de los datos para garantizar el cumplimiento de las normas legales existentes y normativa interna, asegurando una conducta responsable y ética que resguarde la imagen corporativa. En el siguiente subapartado se introducirá conceptualmente el tema de minería de textos y se comentarán algunos algoritmos de *Tex Mining* y su importancia como herramienta de gestión de las universidades.

1.3. El Text Mining como herramienta de valor para organizaciones universitarias.

Las universidades poseen gran cantidad de información disponible en diversos formatos estructurados y no estructurados disponible en diferentes fuentes. Muchos de ellos se encuentran en formato de texto incluido en acervos bibliográficos, normativas internas, planes de estudio, contratos o convenios suscriptos, información incluida en la plataforma educativa, páginas webs y redes sociales. Esta situación lleva a plantear la incorporación de la minería de textos como una herramienta para generar valor en sus decisiones. En presente subapartado se realizará un análisis conceptual de *Tex Mining,* describiendo algunos algoritmos que permiten construir conocimiento a partir del análisis de datos textuales.

Previo a incluir el concepto de *Text Mining* resulta oportuno comenzar el análisis revisando el concepto de *KDD.* Esta sigla representa las palabras en inglés *Knowledge Discovery in*

Databases y consiste en el descubrimiento de conocimiento en base de datos por medio de un proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia comprensible en los datos (Fayyad et al., 1996). Éste incluye varias etapas que requieren una aplicación iterativa y deben aplicarse al conjunto de datos de interés. Estas son: a) entendimiento del negocio, b) entendimiento de los datos, c) preparación de los datos, d) modelado, e) evaluación y d) despliegue (Hotho et al. 2005).

La minería de datos es un término que muchas veces se utiliza como sinónimo de KDD, pero hay autores que la consideran como parte de este proceso incluida en la etapa de modelado (Hotho et al., 2005). El *Data Mininig al igual que la* minería de datos permite obtener nuevo conocimiento, pero éste se genera a partir del análisis de texto. Esta metodología se encuentra vinculada con diversas áreas de investigación entre las que se encuentran las bases de datos, el *Machine Learning* o aprendizaje automático y las estadísticas. La estadística es un método científico que proporciona instrumentos para la toma de decisiones cuando éstas se adoptan en ambiente de incertidumbre, siempre que esta incertidumbre puede ser medida en términos de probabilidad (Martín, 2007). El *Machine Learn*ing es un área de inteligencia artificial relacionada con el desarrollo de técnicas que permiten a las computadoras "aprender" mediante el análisis de conjunto de datos (Hotho et al, 2005).

La minería de texto apunta a la extracción de información significativa presente en textos escritos en lenguaje natural. Su aplicación se efectúa a través de métodos que analizan automáticamente datos textuales con el fin de obtener, a partir de fuentes no estructuradas, el conocimiento utilizado (Villaverde Medina, 2017). Las áreas de investigación involucradas en *Tex Mining* son la recuperación de la información, el procesamiento del lenguaje natural (*PNL*) y la extracción de la información. La recuperación de la información se ocupa de toda la gama de procesamiento de información, desde la recuperación de datos hasta recuperación de conocimientos para obtener una descripción general. El objetivo del *PNL* es lograr una mejor comprensión del lenguaje mediante el uso de computadoras. La extracción de información tiene por objeto extraer información específica de los documentos de textos que son almacenados y puestos a disposición para su posterior uso. (Hotho et al., 2005)

El procesamiento del lenguaje natural tiene por objeto construir una representación sobre el contenido del texto que agregue una estructura al lenguaje natural contenido en discurso del documento. Dicha estructura puede ser de naturaleza sintáctica, capturando las relaciones

gramaticales entre los componentes del texto, o semántica, obteniendo el significado que está transmitiendo el contenido textual (Salaberry, 2019). Este proceso consiste en determinar reglas en cada oración, eliminar aquellas palabras que no aportan significado al sentido del texto (*Stop Word*) y reducir las palabras a su nivel raíz removiendo sufijos y pluralidad con el fin de obtener un procesamiento más veloz. A partir de dicha estructura base se aplican diferentes técnicas de análisis que se engloban en el conjunto de algoritmos de *Tex Mining* (Salaberry, 2019).

Los algoritmos para el análisis de textos permiten la clasificación del texto, la categorización y agrupación en *clusters.* Su objetivo es descubrir relaciones, tendencias y patrones ocultos que son una base sólida para la toma de decisiones empresariales. Entre ellos se encuentran los algoritmos de agrupamiento de *K-Means*, el clasificador Bayes Naive, K-Vecinos más cercanos *(KNN*), Máquina de soporte vectorial, el método de Kernel, Árbol de decisión, Modelos lineales generalizados, Redes Neuronales, Normas de asociación, *Latent Dirichlet Allocation (LDA*), *Word Cloud*, *Sentiment Analysis* o Análisis de sentimientos.

En el presente trabajo se aplicarán los algoritmos de *Word Cloud* o nube de palabras, análisis de sentimiento y *Latent Dirichlet Allocation (LDA*). Word Cloud es una herramienta que se utiliza en la etapa inicial con el fin de detectar en un texto las palabras más frecuentes utilizadas, mediante una visualización rápida y sencilla, permitiendo una visión rápida del contenido del texto, que se logra determinando la frecuencia de las palabras. (Kabir et al., 2018). *LDA* es un modelo matemático probabilístico que pretende encontrar la mezcla de palabras que esté asociada con un tema o tópico, al mismo tiempo que determina el conjunto de temas que describe a cada documento. La distribución de tópicos incluidos en éste viene dada por una distribución de probabilidad de Dirichlet (Alvarez et al, 2018). El análisis de sentimiento es una técnica que trata de identificar los sentimientos contenidos en un texto utilizando una lista de características léxicas, que permiten etiquetar a las palabras según su orientación semántica como positivas o negativas o con una orientación semántica neutral.

En resumen, la minería de textos es una potente herramienta cuya aplicación requiere el análisis de datos no estructurados, la utilización de conceptos estadísticos y la implementación de modelos de *Machine Learning*. Para trabajar con texto es necesario aplicar un procesamiento del lenguaje natural para lograr estructurar los datos contenidos y luego generar información por medio de la aplicación de distintos algoritmos que permiten hacer una lectura de información contenida en grandes volúmenes y extraer conocimiento. Entre estos se encuentran

la nube de palabras, el análisis de sentimientos y el *Latent Dirichlet Allocation.* En el próximo capítulo se desarrollará la metodología aplicada a "la univesidad" mediante el análisis de *tweets* al efecto de detectar tópicos de interés que aporte información de valor a su gestión.

2. Los algoritmos de *Text Mining* para la detección de tópicos

La minería de textos es una herramienta que permite extraer información de documentos heterogéneos que provienen de diversas fuentes. Su utilidad radica en identificar patrones repetitivos, tendencias en el uso de palabras y reglas que expliquen el comportamiento de los datos incluidos en el texto (Chang, 2018). En organizaciones educativas de nivel superior su aplicación ofrece ventajas que apoyan el trabajo de investigadores, docentes, directivos y personas que intervienen en servicios de apoyo, facilitando la interpretación de material bibliográfico, normativas, resoluciones, planes de estudio, programas de los espacios curriculares, filtrado de currículo, análisis de opiniones obtenidas en encuestas y redes sociales entre otros aspectos.

Su implementación requiere considerar cuatro etapas principales que incluyen la determinación del objetivo del análisis, el preprocesamiento de datos, la determinación del modelo y la interpretación de los resultados obtenidos (NETEC, 2019). El preprocesamiento es una instancia de transformación del texto para obtener una representación estructurada o semiestructurada, que permitirá el descubrimiento de patrones facilitando su posterior análisis. Entre las técnicas utilizadas se encuentra el procesamiento de lenguaje natural u otras que permiten la adquisición de patrones léxico sintáctico, extracción automática de términos, localización de trozos específicos de texto o indexación. Una vez que el texto ha sido tratado, se aplican algoritmos de minería de texto, que incluyen métodos descriptivos como la visualización de documentos, el *Clustering*, reglas de asociación y análisis estadístico. También métodos predictivos de aprendizaje supervisado que permite la clasificación y categorización del documento (Montes Gomez, (s.f.))

De lo antes expuesto, se puede afirmar que *TexMining* es una herramienta aplicable a diversas áreas. Las organizaciones pueden utilizarla para facilitar el trabajo y descubrir conocimiento oculto en los documentos, posibilitando entre otras cosas la identificación de tópicos de interés incluidos en un texto. Otro aporte es la minería de opiniones, cuyo objetivo es identificar información subjetiva a partir de publicaciones realizadas por las personas, detectando la

polaridad del texto y analizando el sentimiento expresado a través de las palabras (Polo Aumada, 2022).

En los apartados que conforman el presente capítulo se describirán el proceso de obtención de los datos objeto de estudio mediante recolección de *tweets*. Luego se describirá el tratamiento del texto mediante el procesamiento del lenguaje natural (*NLP*). Finalmente se efectuará un análisis exploratorio descriptivo de los datos recolectados y se especificarán, los algoritmos de Words Cloud, Sentyments Analysis, y *Latent Dirichlet Allocation* (LDA) a utilizar para obtener información de interés.

2.1. Obtención y procesamiento de tweets

Al efecto de conformar el conjunto de datos que será analizado aplicando minería de texto, se efectúa un relevamiento de *tweets* generados por usuarios que arrobaron a cuentas oficiales de "la universidad", de sus unidades académicas y de la Secretaría Académica y de Investigación y Posgrado. Para acceder a esta información se solicitó acceso a la *API (Application Programming Interface)* de *Twitter*. Ésta es una aplicación de uso público y gratuito, no obstante, su utilización requiere actualmente autorización por parte de la compañía, quien otorga un conjunto de credenciales necesarias para autenticar las solicitudes efectuadas. El relevamiento se desarrolla en tres etapas, realizando la captación de *tweets* los días 07 de noviembre, 14 de noviembre y 01 de diciembre del 2022. En cada etapa de recolección, la conexión con la *API* se realizó utilizando la herramienta *Google Colaboratory*[5] a través del lenguaje de programación *Python*[6]. Cada instancia permitió obtener los *tweets* publicados durante los últimos siete días, excluyendo los *retweets*, obteniendo un total de 622 mensajes. Los datos obtenidos se almacenan en un archivo CSV (*Comma-Separated Values*) para su posterior tratamiento.

Una vez completada la tarea descripta en párrafo anterior, se realizó el procesamiento del texto aplicando las técnicas de Procesamiento del Lenguaje Natural (PNL). El lenguaje, desde el punto de vista lingüístico, se define como una función que expresa pensamientos y comunicaciones entre las personas y puede ser utilizado para analizar situaciones complejas y

[5] Acerca de Google Colaboratory: https://colab.research.google.com/

[6] Acerca de Python: https://www.python.org/

razonar sutilmente (Vazquez et al., 2009). La sintaxis del lenguaje natural puede ser modelada por un lenguaje formal, es decir aquel desarrollado por el hombre para expresar situaciones que se dan en diferentes áreas del conocimiento científico como la matemática, la lógica o la computación

El *PNL* consiste en la utilización del lenguaje natural para establecer comunicación con las computadoras, mediante la utilización de programas que ayudan a comprender los mecanismos humanos relacionados con el lenguaje (Vazquez et al., 2009). La arquitectura el *PNL* muestra como la computadora interpreta y analiza las oraciones que le son proporcionadas en sentidos morfológico y sintáctico, verificando si las frases contienen palabras compuestas por morfemas (componentes léxicos definidos a priori) y si la estructura de la oración responde a un orden gramatical entre sus elementos, identificando cómo las palabras pueden unirse para formar oraciones. Luego se analiza las oraciones semánticamente verificando cómo se unen las palabras en una oración para otorgar un significado. Finalmente se realiza el análisis pragmático, es decir analizan un conjunto de oraciones en diferentes contextos, identificando cómo el significado de una oración se ve afectado por las inmediatas anteriores.

Para ayudar a las computadoras a interpretar y manipular el lenguje humano es necesario un conjunto de técnicas de *PNL,* que permitirá posteriormente minar el documento para extraer conclusiones. En primer lugar y para obtener un correcto cálculo de la frecuencia de cada palabra, se normaliza el texto obtenido conviertiendo a minúscula las palabras, de manera de representar con una codificación común aquellas escritas con letra mayúscula en el texto original. Luego se efectua una tokeninzación del texto, dividiendo al mismo en unidades pequeñas llamadas *tokens*. Este proceso se efectua utilizando el módulo de expresiones regulares *NLTK.Tokenizer en Python*, que permite establecer una secuencia de caracteres que ayuda a encontrar otras cadenas o conjunto de cadenas utilizando esa secuencia original como patrón (Datapeaker, (s.f.)). Finalmente, se excluye del texto las *stopwords* o palabras vacias, que incluyen pronombres, advervios y preposiciones, por cuanto no aportan significado alguno. También se amplía el diccionario con símbolos y palabras que no son útiles tales como “http”, “#”, “@”, “rt” al efecto de que el programa no las considere en el análisis.

Otra técnica que permite mejorar la lectura por parte de las máquinas es la lematización del texto. Lematizar es la acción de encontrar el lema de una palabra, es decir establecer su raiz, forma que por convenio se acepta como representante de las formas flexionadas (palabras en

plural, femenino o conjugadas) (Castillo Fadic, 2020). Por ejemplo, "decir" es el lema de "dije", "diré", "dijéramos". Para el presente trabajo se utiliza el método que ofrece *NTLK*, descargando *Open Multilingual Wordnet* (*omw*) y *Wordnet*, permitiendo encontrar en el texto objeto de análisis, lemas en idiomas distintos al inglés.

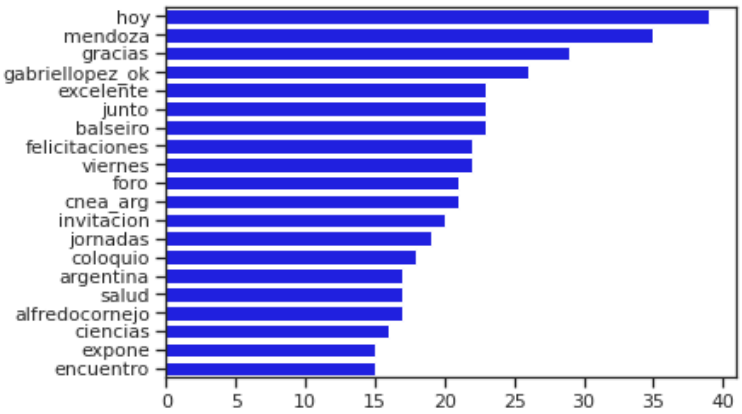
Para sintetizar, en el presente apartado ~~se~~ describe el proceso desarrollado para obtener el conjunto de datos que será analizado. Para recolectar esta información, se utilizó la API de *Twitter*, permitiendo acceder al conjunto opiniones expuestas por usuarios que arroban a la cuenta de "la universidad" mendocina. También se desarrolla aspectos conceptuales del procesamiento del lenguaje natural, describiendo las técnicas utilizadas para la estructuración de los datos. A continuación, se realizará un análisis descriptivo sobre el conjunto de datos finalmente obtenido.

2.2. Análisis descriptivo de los datos

El PNL que se describió en el apartado anterior, es un proceso que contribuye a que el procesamiento algorítmico para efectuar minería de textos sea más eficiente. No obstante, para aprender del comportamiento pasado, es importante efectuar un análisis descriptivo que permita comprender qué es lo que pasó y efectuar un diagnóstico inicial. Éste puede efectuarse a través del cálculo de indicadores estadísticos y la ayuda de gráficos que permitan visualizar el comportamiento de los datos.

Para comenzar el análisis, se estructuraron los tweets iniciales en un corpus que contiene 622 documentos, donde cada documento contiene tokens. La totalidad de los tokens alcanzan las 5.468 unidades de textos. Luego se identificó la frecuencia de cada palabra determinando la cantidad de veces que aparece cada término dentro del texto, ordenando las mismos de mayor a menor. Del total de tokens que componen el texto transformado, se efectúo un top de veinte palabras en función de su aparición a lo largo del corpus, que se pueden observar en figura 1.

Figura 1: Top 20 de frecuencia de palabras



Elaboración propia con Python

De la figura anterior se observa que las palabras con mayor frecuencia son: "hoy" y "Mendoza" por lo que los comentarios que surgen de los *Tweets* pueden asociarse a temas actuales que vinculan a la provincia con "la universidad". Complementando este análisis preliminar, se observa la aparición de nombres de instituciones (Instituto Balseiro, Comisión Nacional de Energía Nuclear) y de personas vinculadas actualmente a la actividad política. También aparecen términos vinculados a encuentros de formación e intercambio de ideas tales como foros, jornadas, coloquio, expone, encuentro, y palabras referidas a ejes de trabajo como la ciencia, salud y otras vinculadas a emociones tales como gracias, felicitaciones.

Del análisis descriptivo realizado, se comienza a observar elementos que contribuyen a la definición tópicos que actualmente son tratados por usuarios que arroban a las cuentas oficiales de "la universidad". Esto debe ser profundizado por medio de diferentes algoritmos que permitan detectar con mayor claridad los temas principales. En el próximo apartado se describirá conceptualmente los algoritmos de nube de palabras, *Latent Dirichlet Allocation* y análisis de sentimientos que serán aplicados al texto objeto de análisis.

2.3. Algoritmos de *Text Mining* para la detección de tópicos

La minería de textos es un área que apunta a la extracción de información significativa presente en textos escritos en lenguaje natural, para lo cual se aplican métodos que los analizan automáticamente, con el fin de obtener conocimiento utilizable a partir de datos no estructurados (Fachin, (s.f)). Para ello es necesario desarrollar un proceso de recuperación y construcción de un corpus de documentos a partir de la selección de textos disponible en

diversas fuentes. Luego éstos son preprocesados con el objeto de efectuar una limpieza de los datos y generar una representación estructurada, para incorporarlos a la fase de *Data Mining* con el objeto de descubrir conocimiento oculto en el corpus original (Villaverde Medina, 2017). Este apartado tiene por objeto efectuar una descripción conceptual de esta última fase, describiendo los algoritmos de Nube de palabras, Análisis de sentimiento y *Latent Dirichlet Allocation (LDA).*

La nube de palabras es un algoritmo que se utiliza en la etapa inicial de análisis. Es una herramienta visual que brinda un diagrama claro del contenido del texto y muestra las palabras que se usan regularmente dentro de él. El algoritmo muestra una representación gráfica útil que permite una aproximación de los temas tratados, por medio de un esquema estadístico en el que hace corresponder la frecuencia de las palabras con el tamaño de la fuente (Kabir et al., 2018).

Latent Dirichlet Allocation (LDA) es un modelo predictivo generativo donde se asumen que las palabras observadas en varios documentos de un corpus (para el presente trabajo cada *Tweets es un documento*) son generadas por temas latentes, por lo que cada documento es una mezcla de temas. Por otro lado, un tema se determina en función de una distribución sobre términos de un vocabulario fijo diseñado y generalmente se representa con aquellos que posee la probabilidad más alta (Hamoel, 2018).

La idea básica de LDA es que los documentos representan mezclas aleatorias sobre temas latentes, donde cada tema se caracteriza por una distribución de palabras. El algoritmo asume un proceso generativo para cada documento en un corpus D que contempla la siguiente secuencia: a) la selección de una cantidad N de palabras que siguen una distribución de Poisson, b) la selección de θ tópicos que siguen una distribución de Dirichlet. c) para cada palabra contenida en una secuencia de palabras, se asigna un tópico a partir de una distribución multinomial sobre θ. d) a cada palabra contenida se le asigna una probabilidad multinomial de pertenencia condicionada al tópico. Por otro lado, la función de densidad de probabilidad Dirichlet se define como:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}$$

donde el parámetro α es un vector k aleatorio de θ, cuyos componentes son mayores que 0, y $\boldsymbol{\Gamma}(\boldsymbol{x})$ es la función de distribución Gamma. Dados los parámetros α y β, la distribución conjunta de θ, para un set ω de palabras N y set ƶ de tópicos θ, viene dada por (Salaberry, 2019):

$$p(\theta, \mathrm{z}, \omega|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(\mathrm{z}_n|\theta) p(\omega_n|\mathrm{z}_n, \beta)$$

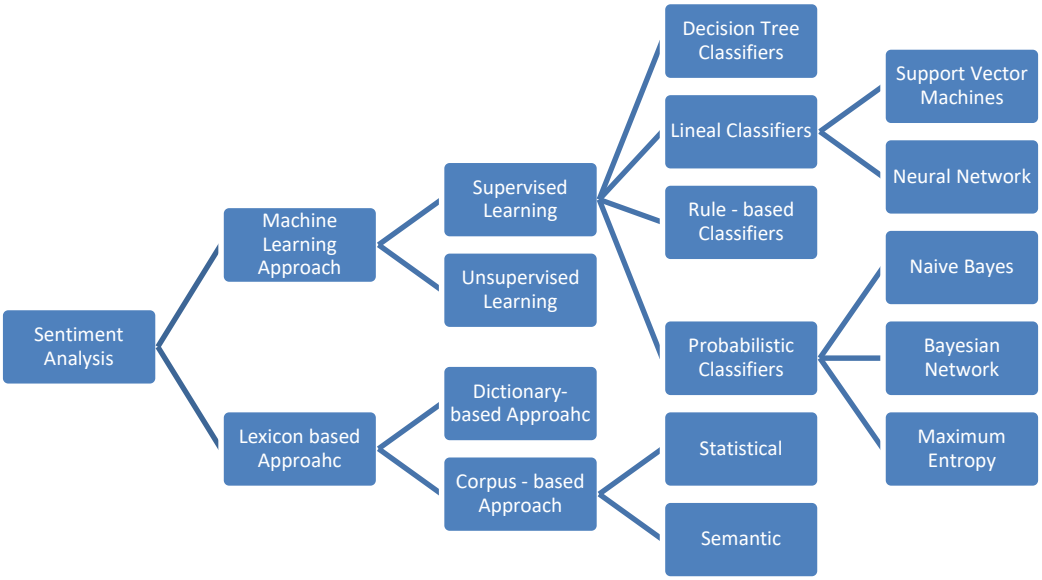
En definitiva, siguiendo lo expuesto por Natalia Salaberry:

> *"LDA es un modelo matemático probabilístico que pretende encontrar la mezcla de palabras que está asociada con cada tema o tópico, al mismo tiempo que determina el conjunto de temas que describe a cada documento. La distribución de tópicos que describen a un corpus viene dada por una distribución de probabilidad de Dirichlet, siendo una generalización de la distribución Beta, lo que implica que un documento sea parte de varios de tópicos donde cada uno posee un peso diferente." (Salaberry, 2019, pag. 24)*

El análisis de sentimientos aplicados a mensajes publicados en Twitter constituyen un material de gran interés para detectar tendencias de opinión entre los usuarios, que hacen públicas sus ideas y debates en la red social. En principio, el algoritmo permite asignar a cada mensaje publicado un valor relacionado con la carga emocional que éste transmite, identificando por un lado la polaridad, intensidad y emoción del mensaje. La polaridad indica si el mensaje tiene un sentimiento positivo, negativo o neutro. La intensidad proporciona un valor numéridco en relación a la intensidad del sentimiento. La emoción contenida puede referirse a alegría, tristeza o ira contenido en el mensaje (Baviera, 2016)

El sentimiento entendido de modo general puede ser evaluado a nivel de documento, oración o entidad determinada como podría ser una organización, un partido político o un candidato. Por esto el análisis debe clarificar cuál es el nivel al que se aplica (Baviera, 2016). En el caso desarrollado en el presente trabajo se refiere al sentimiento expresado en cada documento (o *Tweets*). Por otro lado hay que considerar las técnicas de análisis de sentimiento. Estas se dividen en dos grupos, las que se basan en aprendizaje automático (*Machine Learning Approach*) y las que se basan en diccionario (*Lexicon-Based Approach*). Un resumen de ellas pueden observarse en Figura 2 (Methat, et al, 2014)

Figura 2: Tipologías de las técnicas de análisis de sentimientos



Fuente: Tomado de Methat, et al, 2014

Las técnicas de análisis de sentimiento basadas en *Machine Learning* pueden ser supervisadas y no supervisadas. Los algoritmos no supervisados realizan el procesamiento en base a datos de entrada y tiene la capacidad de configurarse a medida que procesa las observaciones, efectuando un análisis multidimensional de grandes volúmenes de textos. Por otro lado, los algoritmos de aprendizaje automático supervisado cuentan con un corpus manualmente clasificado, llevando a cabo los procesos de entrenamiento del modelo para encontrar los mejores parámetros y evaluación del nivel de fiabilidad del algoritmo, para luego pasar a la fase de clasificación en la que se efectúa la predicción de sentimientos (Baviera, 2016).

El camino alternativo para el análisis de sentimiento se apoya en el uso de diccionarios, siendo éste un listado de términos (palabras o multipalabras) que tienen asociada una determinada orientación de sentimientos. De acuerdo con la distinción inicial, es posible determinar la polaridad, intensidad y tipo de emoción. El algoritmo no detecta patrones sintácticos o aprende a partir de un determinado corpus, sino que detecta coincidencias con los diccionarios y articula un modo de evaluación del sentimiento en base al número de concurrencias encontradas (Baviera, 2016). Esta es la técnica que se aplicará en el presente trabajo.

Para sintetizar, se puede afirmar que la Minería de Textos aporta un conjunto de algoritmos que permiten el análisis de datos no estructurado entre los que se encuentra *Words Clouds, Sentyments Análysis y Latent Dirichechlet Allocation*. La nube de palabra es un algoritmo útil

en la etapa preliminar para orientar al analista sobre las palabras que representan mayor frecuencia en el texto mediante un esquema gráfico que facilita su lectura. Por otro lado, LDA es un modelo que permite especificar tópicos incluidos en un corpus. Cada palabra del texto se asigna aleatoriamente a un tópico, recibiendo una puntuación basada en la probabilidad de que esa palabra pertenezca a él, en el conjunto de documentos. Ésta puede ser asignada a otros temas y se calcula la misma puntuación. Después de este proceso iterativo, se obtiene una lista de términos en cada uno de los temas con probabilidades asociadas. Así, es posible identificar las palabras con mayor probabilidad de pertenecer a un tópico en particular y, la combinación de estas permite describir el tema. Finalmente, el análisis de sentimiento es una herramienta útil para analizar las emociones que contiene el texto, que aporta un valor significativo para el análisis de mensajes publicados en redes sociales. En el próximo capítulo se realizará una aplicación de los algoritmos antes enunciado al conjunto de datos recolectados, analizando cómo estos pueden aportar valor a la organización analizada.

3. Detección eficiente de tópicos en vinculación con la formación recibida en la universidad mendocina

La minería de datos es una herramienta que aporta un conjunto de recursos para el análisis e interpretación de grandes volúmenes de datos incluidos en un corpus textual, conformando un potente recurso para mejorar la gestión organizacional. Ésta se relaciones con disciplinas como la lingüística, la computación, la estadística y la inteligencia artificial, no obstante, su desarrollo no es actual, y su relevancia se ha potenciado por los avances tecnológicos, que han impulsado entre otras cosas, la gestión de los datos como canal para generar diferenciación en la organización

Actualmente las universidades se encuentran cada vez más impactadas por la intersección entre las tecnologías digitales y la sostenibilidad, constituyendo ambos aspectos, ejes transversales que deben priorizarse al definir su estrategia. Esto requiere un proceso de innovación que implique una revisión de sus modelos de gestión, poniendo foco en los datos masivos no estructurados, como recurso para el desarrollo estratégico. Estos cambios requieren trabajar con diversos mecanismos que impacten en la cultura organizacional, diseñando una arquitectura que soporte los desafíos actuales.

En el presente capítulo se describirá el proceso metodológico de aplicación de algoritmos de *TexMining* al conjunto de datos relevados de “la universidad”. Luego se realizará un análisis de

resultados, evaluando los mismos y su aporte al problema planteado en el trabajo. Finalmente se analizará si la información obtenida permite efectuar aportes a la gestión universitaria.

3.1. Aplicación de métodos para la detección de tópicos

Las técnicas de minería de textos persiguen dos grandes propósitos: la predicción y la descripción. Las tareas predictivas o medición basadas en el lenguaje consiste en la construcción de un clasificador automático que estima la variable dependiente, usualmente llamada etiqueta, en función de determinadas características extraídas de los documentos, con el objeto de obtener patrones que explican las relaciones subyacentes en los textos. Por otro lado, las tareas descriptivas buscan obtener patrones que expliquen o resuman las relaciones subyacentes en los datos (Mariñeranela Dondena, et. al, 2017).

Actualmente "la universidad" cuenta con grandes volúmenes de datos que provienen de textos, algunos de los cuales están vinculados estrechamente con las opiniones y sentimientos de las personas que interactúan con la organización. Para capturar su valor, se ha obtenido un conjunto de datos alternativos que provienen de la red social *Twitter*, aplicándose algoritmos que permiten identificar tópicos de interés expresados a través de cuentas oficiales de la organización analizada y sentimientos ocultos en éstos. En el presente apartado se efectuará la descripción metodológica de la aplicación de *Words Clouds, Latent Dirichlet Allocation y Sentyments Analysis*.

El algoritmo de nube de palabras es una técnica visual que manifiesta la importancia del conjunto de palabras incluidas en ~~u~~ corpus. Para su aplicación, se transformó el conjunto de tweets en una variable que contiene *tokens* conformando un vector. Para mostrar el mapa de palabras se usan el módulo *Matplotlib.pyplot y WordCloud de Python* que permite el diseño de la gráfica, obteniéndose la figura 3.

Figura 3: Nube de palabras

Fuente elaboración propia con Python

De la figura anterior se observa como palabras importantes Mendoza, hoy, felicitaciones, gracias, equipos, junto, Comisión Nacional de Energía Atómica, equipo, programa. En menor tamaño palabras vinculadas a personas tales como Alfredo Cornejo (exgobernador y senador nacional por Mendoza), Gabriel Lopez (concejal de municipalidad de Maipú) y Mario Sebastiani (Doctor en medicina). También aparece el nombre de organizaciones tales como Conicet Mendoza, UBAonline, Instituto Balseiro y actividades realizadas en "la universidad" tales como coloquio, jornada, encuentro, foro. Conforme los términos que aparecen de manera representativa en el mapa, los *tweets* estarían a priori, vinculado con situaciones actuales que vincula a "la universidad" con la provincia de Mendoza, pudiéndose referir a actividades realizadas en la provincia o a temas que marcan la agenda de ésta. También aparece como relevantes las palabras Comisión de Energía Atómica, equipo, programa, que podrían indicar aspectos vinculados al área académica o de investigación y el término felicitaciones y gracias que podría asociarse a la entrega de premios o reconocimientos por parte de la institución.

Con el fin de focalizar el análisis e identificar el conjunto de tópicos abordados en los mensajes se aplicó el algoritmo *LDA*, que permite agrupar palabras de un conjunto de documentos automáticamente sin una lista predefinida de etiquetas. El algoritmo requiere para el modelado el diseño de una matriz de palabras insertas en el corpus y un diccionario como entradas principales. En ésta, las filas se corresponden con los documentos (*tweets*) y las columnas con las palabras. Este objeto describe la frecuencia de los términos que aparecen en la colección de documentos. También se crea un diccionario con los *tokens* únicos presentes en el documento y finalmente se genera un modelo base predeterminando la cantidad de tópicos. De esta manera

se obtiene un conjunto de temas integrados por palabras claves que poseen ponderaciones conforme su importancia dentro del tópico, lo que posibilita etiquetarlos. Para el caso bajo análisis, el algoritmo se definió con seis grupos, siendo su resultado el expuesto en la figura 4.

Figura 4: Tópicos identificados aplicando LDA

Items	Tokens ponderados
0	0.008*mendoza + 0.006*felicitaciones + 0.004*gracias + 0,004*conicetmendoza + 0,004* mate + 0.004*equipo + 0.004*argentina + 0.003*políticas + 0.003*cna_arg +0.003*premio
1	0.017*hoy + 0.009*balseiro + 0.007*salud + 0,007*foro + 0,006* jornadas + 0.006*compartimos + 0.006*viii + 0.006*derecho + 0.005*vida +0.005*nacionales
2	0.005*msebastiani + 0.005*ubaonline + 0.005*alfredocornejo + 0,005*junto + 0,004* mendoza + 0.004*gracias + 0.004*nota + 0.004*trabajo + 0.004*viernes +0.004*hoy
3	0.006*felicitaciones + 0.005*mendoza + 0.005*beatles + 0,005*colplay + 0,005* sinfonico + 0.005*rolando + 0.005*egresado + 0.005*somma + 0.005*expone +0.005*invitacion
4	0.006*acto + 0.005*jornada + 0.005*realizo + 0,005*encuentro + 0,005* sociales + 0.005*foro + 0.004*evento + 0.004*mendoza + 0.003*dia +0.003*seguridad
5	0.011*gabriellopez_ok + 0.010*excelente + 0.007*junto + 0,007*capacitacion + 0,006* gracias + 0.005*facultad + 0.005*alfredocornejo + 0.004*hoy + 0.004*oeruuncuyo +0.003*ubaonline

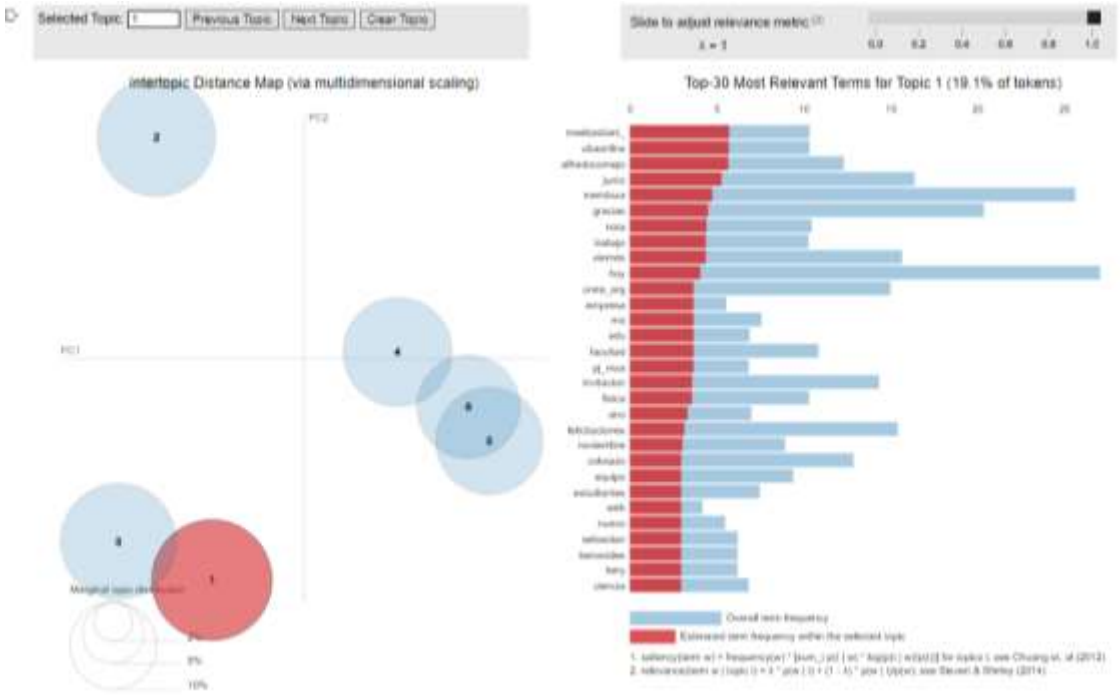
Fuente: elaboración propia con Python

En esta figura se observa seis grupos que están compuestos por palabras ponderadas que permiten describir los temas latentes. Los ítems cero y uno incluyen palabras como Mendoza, felicitaciones, gracias, Conicet Mendoza, mate, equipo, argentina, políticas, cnea_arg, hoy, Balseiro, salud, foro, jornadas, compartimos, derecho, vida que está vinculadas a actividades de investigación científica generadas en el ámbito de “la universidad”. En el ítem dos aparece msebastianelli, ubaonline, alfredocornejo, junto, mendoza que se relación con vinculaciones de la casa de estudios con reconocidos referentes del ámbito político y científico. El ítem tres está integrado por felicitaciones, Mendoza, Beatles, *Colplay*, sinfónica, relacionado con actividades culturales. Por otro lado, el ítem cuatro está conformado por los términos acto, jornadas, encuentro, sociales, foro, evento, seguridad se relaciona con acciones de formación académica. Finalmente, el ítem cinco incluye gabriellopez_ok, excelente, junto, capacitación, facultad, ubaonline por lo que podría asociarse con actividades de vinculación.

Para mejorar la interpretación de este algoritmo, se complementa el análisis con la herramienta visual que propone el módulo *pyDavis en Python*, que permite el diseño de un gráfico

interactivo. Del lado izquierdo de la imagen, se muestra el mapa de distancia entre temas representados por burbujas. Cuanto más grandes son éstas, mayor es el número de documentos en el corpus que pertenecen a cada tema. Por otro lado, cuanto mayor es la distancia entre las burbujas, más diferentes son los temas tratados. En el lado derecho, se muestran un gráfico de barras, en el que en eje "y" se exponen las palabras de mayor relevancia por tema, y en el eje "x", por medio de barras celestes, su frecuencia. Al seleccionar cada tópico, una porción de éstas toma color rojo representando la frecuencia de la palabra en relación con el tema seleccionado. En el caso bajo análisis, se puede observar una imagen del gráfico interactivo en figura 5.

Figura 5: Mapa interactivo de tópicos identificados aplicando LDA



Elaboración propia con Python

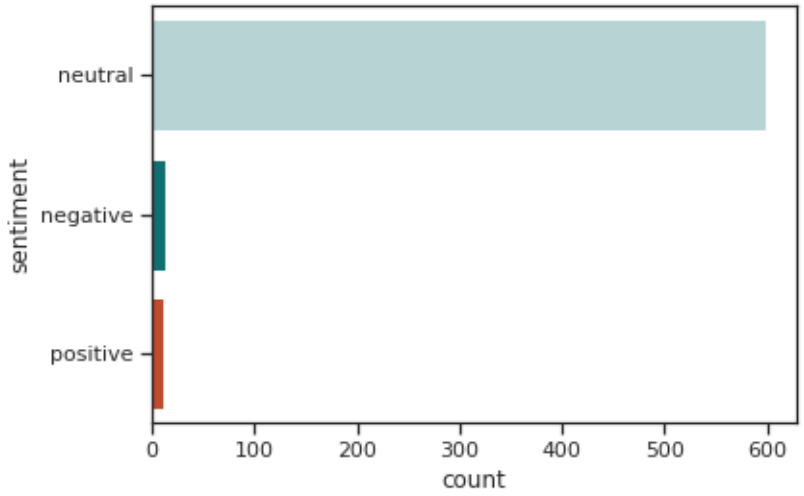
Esta representación gráfica aporta varios elementos que enriquecen la interpretación. Por un lado, las burbujas que guardan mayor distancia entre si se refieren a temas diferentes, como son los tópicos uno y tres ubicados en el tercer cuadrante (ítems 2 y 5 párrafo anterior) y el dos en el cuarto cuadrante (ítems 4 párrafo anterior). Los primeros están referidos a actividades que vincula a "la universidad" con el medio a través de representantes de la ciencia y la política, mientras que tópico cuatro se refiere a diversas actividades de formación académicas

desarrolladas por las unidades académicas. De manera superpuestas y compartiendo parte del cuadrante uno y dos están los tópicos cuatro, cinco y seis que se relacionan con actividades científicas, culturales y de investigación (ítems 1, 0 y 3 apartado anterior).

Del lado derecho del gráfico se puede complementar la información de cada burbuja. Si se ubica el cursor sobre cada tópico, el gráfico de barras muestra un conjunto de 30 palabras que participan para la identificación de un tema, mostrando la frecuencia de cada una de ellas e indicando en términos porcentuales la participación relativa del conjunto de tokens que representan cada burbuja. Para ejemplificar, la burbuja uno incluye el 19.10% de los tokens incluidos en el diccionario para conforman un tema y la tres posee una participación relativa inferior del 17.60%, que puede visualizarse por la reducción en el tamaño del círculo.

El tercer algoritmo aplicado es el análisis de sentimientos realizando el análisis a nivel de documento (o *Tweets)*. Su implementación se efectúa aplicando el módulo *SentimentIntensityAnalyzer* de *NLTK.sentiment*, cargando el diccionario "vader_lexicon". Éste permite medir los sentimientos, y las emociones de un mensaje a partir del tratamiento computacional de la subjetividad del texto. VADER (V*alence Aware Dictionary For Sentiment Reasoning*) asigna características léxicas a intensidades de emoción, a través de un puntaje de sentimiento que se obtiene sumando la intensidad que aporta cada palabra del texto (Aditya, 2020). Para el caso analizado se toma la cadena de caracteres compuesta por lemas y se aplica la función *Analyzer.plolarity_scores,* devolviendo un diccionario de puntajes que está asociado a la categoría negativo, neutral o positivo. El resultado de este proceso se observa graficado en figura 6.

Figura 6: Análisis de sentimientos



Fuente: elaboración propia con Python

Como se observa en el gráfico el conjunto de sentimientos expresados en los tweets refieren una actitud de neutralidad. Esto está vinculado a que los mismos expresan un conjunto de actividades concretas realizadas por "la universidad" y sus unidades académicas, que utiliza la red social como medio de comunicación. La difusión de contenidos informativos publicados por la casa de estudio no genera en el lector una posición polarizada. Por otro lado, debe destacarse que la cantidad de interacciones son reducidas, lo que demuestra que "la universidad" no ha logrado fidelizar al lector por medio de *Twitter*.

Para sintetizar, en el presente capítulo se aplicaron tres algoritmos que permitieron detectar que la organización bajo análisis utiliza las redes sociales para publicar el conjunto de actividades que realiza cotidianamente. De estas se desprenden actividades vinculadas a la investigación, la cultura, la actividad académica y la vinculación con la sociedad. Finalmente se observa que la respuesta de los usuarios ante estas publicaciones es baja, no observándose polarización de sentimientos en el contenido de los mensajes. En el próximo apartado se efectuará una evaluación que describa cómo los resultados obtenidos pueden aportar al problema planteado en el presente trabajo.

3.2. Evaluación y análisis de resultados

Twitter es una plataforma de *microblogging* que permite el contacto de personas a través de mensajes cortos en formato de texto, fotos y videos. La masividad en el uso de esta plataforma ha generado que las organizaciones la utilicen como canal de comercialización y análisis de la

conducta de los usuarios para personalizar la experiencia del usuario. También se utiliza para observar la acción de los competidores o dar a conocer eventos y hechos importantes (Fachin, (s.f)). "La universidad" no es ajena a esta situación y realiza periódicamente publicaciones. Un conjunto de ellas han sido objeto de estudio en el presente trabajo. El resultado del análisis se sintetiza en el presente apartado.

Como se comentó en el apartado 2.2, el relevamiento de *tweets* se efectuó en tres oportunidades, que incluyeron la captura de mensajes generados durante 21 días a través de usuarios que arrobaron a 12 cuentas oficiales, obteniendo un set de 622 *tweets* excluyendo los *retweets*. A este conjunto de datos, se aplicaron algoritmos de *TexMining* al efecto de detectar tópicos referidos a la formación recibida en "la universidad" mendocina a partir de los comentarios realizados *y* determinar la vinculación entre sus planes de estudios y capacidades demandadas en el mercado laboral.

Del análisis de los algoritmos *Cloud Words, Latent Dirichlet Allocation* y *Sentyments Analysis* se observó que los temas tratados por los canales oficiales se refieren a actividades de investigación, culturales o académicas realizadas por la casa de estudios y expresa la vinculación que ésta mantiene con representantes de la política, de la ciencia y la cultura. No obstante, los datos no mostraron tópicos que representen la percepción social referida a la formación del graduado universitario. Por otro lado, el sentimiento que surge del análisis del texto no muestra polaridad positiva o negativa. Esto manifiesta que quienes arrobaron a las cuentas oficiales, en general no abren debates sobre los temas planteados, ni plantean problemáticas u opiniones que generen participación masiva. Simplemente acompañan el sistema comunicacional que abre "la universidad" ante la publicación de eventos o actividad de interés.

Lo antes planteado manifiesta que actualmente "la universidad" no está aprovechando la potencialidad que provee la red social para transformarla en una valiosa herramienta de gestión. Entre los posibles usos, la casa de estudio podría establecer una estrategia de alcance, que incremente la cantidad de usuarios que interactúan con las cuentas oficiales, generando contenidos que habiliten debates para capturar su opinión sobre aspectos de interés. Actualmente, es de suma importancia conocer las competencias que requiere actualmente un graduado universitario, porque apunta a objetivos clave de la organización educativa, pudiendo obtener datos dinámicos que manifiesten las tendencias en tiempo real. Esta información

representa un genuino aporte, posibilitando cambios fundados en planes de estudio vigentes, programas y cursos de actualización. En el próximo apartado se analizará, cómo la información obtenida por medio de *tweets* puede impactar en gestión de "la universidad" objeto de estudio.

3.3. Relevancia del análisis de *tweets* para la gestión universitaria

Actualmente la demanda creciente para acceder a la educación superior exige que las universidades diversifiquen su oferta académica, y las modalidades de enseñanza–aprendizaje actualizando sus estrategias formativas (CRES, 2018). Esto pone de manifiesto que la oferta de servicios debe adecuarse oportunamente para responder a las actuales necesidades del mercado, capturando en tiempo real el conjunto de requerimientos demandados. Para brindar respuestas adecuadas, el *Big Data* presenta una oportunidad que "la universidad" debe poner en el centro del debate. En este apartado, se reflexionará sobre los cambios que la organización debe plantear, para responder adecuadamente a los desafíos mundo actual.

Como medio para construir y mantener una ventaja competitiva, los instrumentos tradicionales de la planificación estratégica centran su análisis en el contexto interno y externo identificando fortalezas, debilidades, oportunidades y amenazas. El conjunto de herramientas desarrolladas actualmente proporciona una imagen estática del momento en que se recopilan los datos, mediante el cálculo de indicadores que suponen una representación simplificada del entorno empresarial. No obstante, las organizaciones, para brindar una ventaja competitiva, poseen recursos heterogéneos que provienen de datos alternativos y posibilitan generar múltiples capacidades. Pero para lograrlas, es necesario gestionar procesos que permitan su utilización para volverse relevantes en el desarrollo de la estrategia (Constantiou yKalliniko, 2015).

Actualmente los procesos de captura de datos están insertos en ecosistemas digitales cuyo almacenamiento está vinculado a propósitos genéricos. Éstos poseen esquemas estructurados, y son procesados y almacenados a partir de sistemas de clasificación y agregación previamente definidos. Por otro lado, se encuentran los datos que provienen de redes sociales, que se generan a partir de situaciones diarias y triviales. Sus repositorios de almacenamiento requieren considerar formatos diferentes que no conviven con los sistemas alfanuméricos que pueblan las organizaciones actuales. Esto genera una barrera de acceso, por cuanto los sistemas tradicionales de captura y procesamiento no son directamente transferibles para el tratamiento de los alternativos (Constantiou yKalliniko, 2015)

Por otro lado, los datos no estructurados han generado un cambio de paradigma en la definición de la estrategia. La nueva forma de generar información como consecuencia de la naturaleza de los datos alternativos implica cambios en las representaciones mentales de muchos decisores (Constantiou yKalliniko, 2015). Los líderes deben ser capacer de interpretar las tendencias para modelar la estrategia, capturando los cambios que se generan constantemente, estableciendo estructuras flexibles, adaptables, capaces de aprender rápidamente y brindar una respuesta ajustada a las necesidades el cliente. Esto incluye procesos ágiles, que consisten en abordar decisiones estratégicas como procesos de descubrimiento y aprendizaje en condicones competitivas, que cambián rápidamente (Espósito at el., (s.f)). Los miembros de la organización deben acompañar este proceso, integrando al conjunto de actores que participan en la organización y haciéndolos partícipe en la definición de estos cambios.

Los datos no estrucucturados aportan a “la universidad” un activo que debería ser explotado, de manera que la misma replantee su actual modelo de gestión, generando un proceso gradual que modifique el actual. Esta transformación requiere una revisión de la arquitectura actual , planteando un modelo ideal y un proceso de transición que acompañe los cambios. Para ello, se deben considerar la capa del negocio, revisando la visión, ejes estratégicos, modelo de gobernanza y procesos clave. Luego intervenir en la arquitectura de aplicacaciones, diseñando un plano para cada uno de los sistemas que se decida implementar, y desarrollar una estructura de datos físicos y lógicos y una arquitectura tecnológica que soporte los cambios previstos (Piorum, 2019).

Conclusión

El principal resultado logrado a través del desarrollo del presente trabajo es la identificación de tópicos a partir de *tweets* sobre la vinculación de "la universidad" con las actividades de formación de sus estudiantes y graduados. De este modo, se logra poner en valor el potencial que brinda el tratamiento de grandes volúmenes de datos no estructurado en la obtención de información para la toma de decisiones. A su vez, esto requiere de aplicar un modelo de gobierno de datos para lograr una adecuada gestión. De esta manera, surge que la organización debería replantear el modelo actual de arquitectura de datos para incorporar a los alternativos como un activo.

El objetivo general que guio la elaboración de este trabajo fue detectar tópicos que expongan la percepción de la formación recibida en "la universidad" mendocina a partir de los comentarios realizados en la red social *Twitter* y determinar la vinculación entre sus planes de estudios y las capacidades demandadas en el mercado laboral. Para lograrlo se aplicaron diferentes algoritmos de *Text Mining* a través del análisis de *Tweets* que se obtuvieron de la red social *Twitter*, relevando las cuentas oficinales de "la universidad" objeto de estudio y de sus unidades académicas. Para ello se conformó un conjunto de datos con el objeto de analizar los comentarios de usuarios que arrobaban a la "Universidad" y sus Facultades. Por otro lado, se realizó un análisis del gobierno de datos y sus tres ejes claves: el ciclo de vida de los datos, la gestión de calidad del dato y la gestión de su seguridad y privacidad. En base a los resultados obtenidos se logró establecer los tópicos principales que surgen de los comentarios de los usuarios que arroban a las cuentas oficiales de "la universidad", identificando los temas tratados.

El primer capítulo tuvo por objetivo desarrollar conceptos necesarios para comprender la gestión de datos no estructurados en la generación de valor en una organización universitaria y determinar las tecnologías necesarias para su implementación. Mediante el desarrollo realizado se establecieron los aspectos que caracterizan los datos no estructurados tales como la velocidad con que estos se generan, su variedad y masividad. Esto requiere necesariamente un proceso de gobernanza que tenga en cuenta sus particularidades, para transformarlos en un activo en la organización. La captación de datos alternativos requiere un proceso de ingesta en tiempo real que permita su tratamiento aplicando modelos de aprendizaje automático para anticipar o predecir diferentes situaciones. A su vez ello implica considerar el tratamiento de datos

maestros como instrumento que permita asegurar su calidad. Finalmente se concluyó la importancia de trabajar sobre la seguridad y privacidad de los datos como elemento para que asegure una conducta ética y responsable que debe caracterizar a la organización educativa. De este modo, el aporte realizado en este capítulo es demostrar el impacto de los datos no estructurados como fuente para generar información que aporta valor a organización universitaria para la toma de decisiones.

El segundo capítulo tuvo como objetivo preparar el corpus textual sobre el que se aplicaron algoritmos de *Tex Mining* para identificar tópicos de interés social en vinculación con la formación recibida en "la universidad". Para lograrlo se efectuó un relevamiento de *tweets* publicados por usuarios que arrobaron a cuentas oficiales de "la universidad" conformando 622 documentos. Sobre este conjunto de datos se aplicaron técnicas de procesamiento de lenguaje natural para estructurar el texto y se efectuó un análisis descriptivo de los datos obtenidos. Finalmente se describieron conceptualmente los algoritmos de Nube de Palabras, Análisis de Sentimientos y *Latent Dirichlet Allocation.* De este modo, el aporte realizado en este capítulo consistió es construir un conjunto de datos normalizados aplicando PNL y efectuar su diagnóstico, describiendo algoritmos de *TexMining* que se aplicarán en el apartado tres.

El tercer capítulo tuvo como objetivo evaluar, cómo el análisis de datos alternativos aplicando técnicas de minería de textos, puede contribuir a la redefinición de los planes de formación en la organización universitaria en función de las necesidades profesionales demandadas. Para ello se aplicaron técnicas de *TexMining* identificando que los principales tópicos están referidos a actividades de vinculación, formación, eventos científicos y culturales realizados por "la universidad", pero no se observó de manera concreta aspectos vinculados a las capacidades de sus estudiantes para responder al mercado. Como consecuencia del análisis se puede concluir que el tratamiento de datos no estructurados a través de algoritmos de minería de texto constituye una herramienta útil para obtener información de valor. Esta podría potenciarse si la organización moviliza la discusión de diferentes temas, que permitan conocer la opinión de los agentes sociales que conforman el sistema universitario, para identificar tendencias vinculado a la formación de sus graduados y el desarrollo de competencias. De este modo, el aporte realizado por este capítulo consistió en aplicar las técnicas de minería de textos e identificar cómo éstas permiten a la organización objeto de estudio obtener información de valor que aportan a su gestión.

Tras el desarrollo de cada uno de los capítulos, el trabajo final de especialización realiza un aporte de especial interés para "la universidad". Se pudo mostrar que la organización está en una etapa inicial en la que simplemente utiliza la red social para comunicar, pero no está gestionando los datos que surgen de la misma como herramienta para generar valor. Esta situación manifiesta la necesidad de realizar un proceso de cambio que incluya a los datos alternativos como un activo organizacional. Esto implica generar un cambio de la arquitectura organizacional que permita capturar la variedad de datos que provienen del entorno externo e interno e identificar la magnitud y dimensión de diferentes tendencias sociales para reformular la estrategia y el modelo de gestión.

El trabajo presenta diversas posibilidades para continuar ampliando su alcance. Para complementar el análisis realizado en este trabajo sería de utilidad incorporar algoritmos complementarios de minería de textos, para revisar diferentes redes sociales en las que participan los estudiantes al efecto de detectar problemáticas por las que estos transitan durante su proceso de formación. Esto permitirá a la institución aplicar recursos de manera eficiente, considerando información obtenida oportunamente, para acompañar su trayectoria académica.

También se considera relevante como objeto de estudio, analizar los cambios que requiere la arquitectura actual de "la universidad", de manera de transformarse en una organización en la que el dato sea un activo de valor para su gestión. Esto permitirá trabajar sobre ejes como la inteligencia colaborativa que permita mejorar los resultados y rendimientos propuestos y sobre sistemas de aprendizaje a través del análisis de datos, que permite tomar decisiones más rápidas y precisas. También agilizar el proceso de toma de decisiones, que ayude a la organización a responder de manera oportuna y cada vez más efectiva, a la incertidumbre manifiesta en el mercado actual. A su vez también podrá brindar nuevas oportunidades, atrayendo talento, innovando y obteniendo el consentimiento social de la actividad desarrollada (Espósito et al., 2020).

Para ello se propone como línea de trabajo indagar sobre la actual arquitectura de "la universidad" y proponer un modelo que incluya a los datos como factor estratégico. Será necesario un desarrollo progresivo de áreas de investigación en las que se trabaje sobre las cuatro capas de la arquitectura, poniendo foco en los datos e infraestructura tecnológica. Esto servirá de apoyo para poner a "la universidad" a la vanguardia en los procesos de formación

requeridos por la sociedad, aportando valor por medio de la investigación y de vinculación con agentes del medio.

Referencias bibliográfica

Aditya, V. (27 de 05 de 2020). *Análisis de sentimientos utilizando Vader*. Obtenido de https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

Alvarez D, Armero C, Forte A. (2018). Whats Does Objective Mean in a Dirichlet-multinomial Process. *Intenational Statistical Review, 86(1),* 106-118.

Argonza, J. S. (2016). Big Data en la Educación. *Revista digital Universitaria*, 1-16 (Vol 17 -Nro 1).

Baviera, T. (2016). Técnicas para el análisis de sentimiento en Twitter: Aprendizaje automático supervisado y SintiStrength. *Dígitos: revista de comunicación digital* , Pags 33-50.

Blei, David, Ng, Andrew y Jordan Michael. (2003). Latent Dirichlet Allocation. *Revista de investigación de aprendizaje automático* , Pags 993-1022.

Camargo Vega, J. C. (2015). Conociendo Big Data. *Revista facultad de Ingeniería*, 63-77.

Castillo Fadic, M. N. (2020). Corpus Básico del Español de Chile ©: metodología de procesamiento y análisis. *Lexis*, Pag 483-523. Vol. XLIV (2).

Chang, j. O. (22 de 02 de 2018). *¿Qué es la minería de textos, cómo funciona y por qué es útil?* Obtenido de https://universoabierto.org/2018/02/22/que-es-la-mineria-de-textos-como-funciona-y-por-que-es-util/

Chang, J;O' Reilly, C; Pontika, N; Owen; G,Haug, K;Oudenhoven (LIBER). (s.f.).

Chinkes, E., & Julien, D. (2019). Las instituciones de educación superior y su rol en la era digital. La transformación digital de la universidad: ¿transformadas o transformadoras? *Ciencia y Educación (Vol 3-Nro 1),* 21-33.

Clevent A & Wortmann F. (2010). Uncovering four strategies to appoach master data management. *Acta de la 43 Conferencia Internacional sobre Ciencias Sociales*, (págs. 1-10). Hawai.

Constantiou, I, Kalliniko, J. (14 de 11 de 2015). *Nuevos juegos, nuevas reglas. Big data y el contexto cambiante de estrategia*. Obtenido de https://journals.sagepub.com/doi/10.1057/jit.2014.17

CRESS. (14 de 06 de 2018). *Declaración de la III Conferencia Regional de Educación Superior en América Latina y el Caribe.* Obtenido de https://www.iesalc.unesco.org/wp-content/uploads/2020/08/Declaracion2018PortFinal.pdf

Datapeaker. ((s.f.)). *¿Qué es la tokenización? Métodos para realizar la tokenización*. Obtenido de https://datapeaker.com/big-data/que-es-la-tokenizacion-metodos-para-realizar-la-tokenizacion/

Davenport, T. H. (2014). *Big Data at work. Dispelling the Myths, Uncovering the Opportunities*. Boston, Massachusetts: Harvard Business Scholl.

Delloitte. (2011). *Desarrollo de un modelo operativo de gobierno que sea efectivo. Una uuía para las juntas y los equipos de administración de servicios financieros.*

Espósito, M, Lanteri A, Tse, T. ((s.f)). Una arquitectura estratégica para la prosperidad corporativa pospandémica. *Harvard Deusto Business Review*, Pag 28 a 39.

Fachin, J. ((s.f)). *¿Qué es Twitter, para qué sirve y cómo funciona esta red de microblogging?* Obtenido de https://josefacchin.com/que-es-twitter-como-funciona/.

Fayyad U. Piatetsky S., Smyth P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Dat. *Communications of the ACM - Vol 39 - Nro 11*.

Hamoel, L. (2018). *Detección de tópicos utilizando el modelo LDA. Trabajo final de especialización.* Buenos Aires. https://ri.itba.edu.ar/server/api/core/bitstreams/162467b1-60d4-4e95-bc21-29f526830652/content.

Hotho A., Nurnberger A., Paass Gerhard. (2005). *Una reserña dela minería de texto.*

Josey A, Harrison R, Homan P, Rouse M, Van Sante T, Turner Mike, Merwe P. (2013). *TOGAP Versión 9.1. Guía de bolsillo.* Reino Unido: Van Haren Publishing, Zaltbommel, ISBN eBook: 978 90 8753 813 2.

Kabir A, Karim R, Newaz S, Istiaque Hossain M. (2018). The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R. *Informática Economica - Vol 22*, 25-38.

Kolanovi M.; Krishnamachari R. (2017). *Estrategia de Big Data e IA. Aprendizaje automático y enfoque de datos altenativos para invertir.* Reino Unido: JP Morgan.

Mariñeranela Dondena, L., Errecalde, M, Castro Solano, A. (2017). Extracción de conocimientos con técnicas de minería de textos aplicadas a la psicología. *Revista Argentina de Ciencias del Comportamiento*, Pags. 65-76.

Martín, P. (2007). *Introducción a la estadística económica y empresarial: teoría y práctica.* Madrid: AC.

Methat, W, Hassan, A y Korashy, H. (2014). Sentiment analysis algorithms and aplicatiions: A survey. *Ain Shams Enfineering Journal*, Pags. 1093-1113, Vol 5, nro 4.

Microsoft. (25 de 10 de 2022). *Datos no relacionales y NoSQL*. Obtenido de https://learn.microsoft.com/es-es/azure/architecture/data-guide/big-data/non-relational-data

Montes Gomez, M. ((s.f.)). *Minería de texto: Un nuevo reto computacional.* https://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf: Centro de Investigación en Computación. Instituto Politécnico Nacional.

NETEC. (27 de Mayo de 2019). *Minería de datos: Qué es, importancia y técnicas de su implementación*. Obtenido de https://www.netec.com/post/mineria-de-datos-que-es-importancia-y-tecnicas-de-su-implementacion

Piorum, D. (05 de 08 de 2019). *Arquitectura empresarial: Desafío Organizacional en la Transformación Digital*. Obtenido de https://degerencia.com/articulo/tag/arquitectura-empresarial/

Polo Aumada, A. M. (03 de Diciembre de 2022). *Mineria de datos, de textos y de sentimientos*. Obtenido de https://www.gestiopolis.com/mineria-de-datos-de-textos-sentimientos/

Roesslein, J. (Copyright 2009-2022). *Tweepy. An easy-to-use Python library for accessing the Twitter API.* Obtenido de https://docs.tweepy.org/en/stable/

Salaberry, N. (2019). Detección de problemáticas en el uso de la tarjeta SUBE a partir del análisis y clasificación de twets. *Trabajo final de Posgrado Especialización en métodos cuantitativos y análisis de datos*. Buenos Aires, Argentina.

Santiago y Zakhmen. (23 de 10 de 2022). *Implementación modelos de aprendizaje automático*. Obtenido de https://e72102.readthedocs.io/es/latest/index.html#

Schmarzo, B. (2013). *Big Data. Undersanding how data powers big business.* Indianápolis: WILEY.

Serrano, J. (2022). *Marco para la construcción de sistema de gobbernanza de datos en entornos de industria 4.0.* Santander: Tesis Doctoral - Universidad de Cantabria.

Vazquez A, Huerta V, Quispe J. (2009). Procesamiento de lenguaje natural. *Revista de Ingeniería de Sistemas de Información*, Págs. 45-54 . Vol 6, Nro 2.

Villaverde Medina, N. (2017). *Nuevas técnicas estadísticas: Text Mining en Web.* España: Facultad de Economía y Empresa de la Universidad da Coruña.

Wikipedia. (17 de 11 de 2022). *Wikipedia: Portal tecnología*. Obtenido de Wikipedia: Portal tecnología: https://es.wikipedia.org/wiki/Portal:Tecnologia

Paquetes y módulos

Biblioteca Gensim. (2009). *Radim Řehůřek.* Obtenido de Gensim topic modelling for human: https://radimrehurek.com/gensim/

Función models.ldamulticore. (2009). *Radim Řehůřek.* Obtenido de Gensim topic modelling for humans: https://radimrehurek.com/gensim/models/ldamulticore.html

Libreria Matplotlib. (2012). *John Hunter, Darren Dale, Eric Firing, Michael Droettboom and the Matplotlib development team.* Obtenido de Matplotlib: Visualization with Python: https://matplotlib.org

Librería Numpy. (2022). *NumFOCUS, Inc.* Obtenido de Numpy. The fundamental package for scientific computing with Python: https://numpy.org

Librería Pandas. (2008). *Wes McKinney.* Obtenido de Pandas: https://pandas.pydata.org/docs/getting_started/install.html

Librería Seaborn. (2012). *Michael Waskom.* Obtenido de Seaborn: statistical data visualization: https://seaborn.pydata.org

Librería Tweepy. (2009). *Roesslein, J.* Obtenido de An easy-to-use Python library for accessing the Twitter API: https://docs.tweepy.org/en/stable/

Módulo NLKT.downloader. (2001). *NLTK team.* Obtenido de Documentation nltk.downloader module: https://www.nltk.org/api/nltk.downloader.html#module-nltk.downloader

Módulo Pickle. (2015). *GitHub, Inc.* Obtenido de Pickle — Python object serialization: https://github.com/python/cpython/commits/3.11/Lib/pickle.py

Módulo Tokenize. (2001). *Python Software Foundation.* Obtenido de tokenize — Conversor a tokens para código Python¶: https://docs.python.org/es/3/library/tokenize.html

Paquete NLKT. (2001). *Tom Aarsen, Joel Nothman, Steven Bird, Alexis Dimitradis, Danny Sepler, Dmitrijs Milajevs, Francis Bond, Ilia Kurenkov.* Obtenido de NLKT documentation: https://www.nltk.org/api/nltk.html

Paquete PyLDAdis. (2015). *Ben Mabey.* Obtenido de Python library for interactive topic model visualization: https://pyldavis.readthedocs.io/en/latest/readme.html

Submódulo nltk.sentiment.sentiment_analyzer module. (2001). *NLKT teams.* Obtenido de Documentation nltk.sentiment.sentiment_analyzer module: https://www.nltk.org/api/nltk.sentiment.sentiment_analyzer.html

Submódulo nltk.sentiment.util. (2001). *NLTK teams*. Obtenido de Documentation nltk.sentiment.util module: https://www.nltk.org/api/nltk.sentiment.util.html

Submódulo nltk.sentiment.vader. (2014). *Hutto, C.J. & Gilbert, E.E.* Obtenido de VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media: https://www.nltk.org/api/nltk.sentiment.vader.html