Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

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

TRABAJO FINAL DE ESPECIALIZACIÓN

Detección de problemáticas en el uso de la tarjeta SUBE. Un análisis y clasificación de tweets

Implementación de técnicas de Text Mining

AUTOR: NATALIA R. SALABERRY

DICIEMBRE DE 2019

Resumen

La explosión de los medios digitales ha dado lugar a la conformación de un ecosistema de datos, que posee la particularidad de poseer diversidad, de ser a gran escala y sin una estructura definida. Frente a tales características, las organizaciones se enfrentan a un desafío para su tratamiento. Es en este sentido que la gestión y gobierno de datos toma un rol central, dado que se trata de un proceso a partir del cual la organización gestionará la información como un recurso integrado y de calidad, garantizando la confiabilidad sobre estos para la toma de decisiones.

Dentro del marco definido en el párrafo anterior, el presente trabajo persigue el objetivo de establecer el marco necesario de gestión y gobernanza de datos para que los resultados obtenidos a través del uso de técnicas de Data Mining, se constituyan en un verdadero valor agregado en el proceso de toma de decisiones de una organización.

Frente a la necesidad de poder garantizar la integridad y calidad sobre los datos, en primer lugar, se conceptualiza y se analizan las implicancias de establecer un sistema de gestión de datos maestros en una organización. Dada la existencia de tal sistema, luego se lleva a cabo la aplicación de técnicas de Data Mining sobre un conjunto de tweets con el fin de poder detectar problemáticas asociadas al uso de la tarjeta SUBE.

Como resultado, en primer lugar, se logra detectar al menos tres problemáticas que expresan los usuarios en la red social Twitter. Como resultado final, se logra determinar que, del total de tweets analizados, el 50% expresan una problemática asociada. Finalmente, se da cuenta sobre la importancia de establecer un sistema de gobierno de datos cuyo eje central sea conservar la privacidad de datos personales de los usuarios, desde una visión ética y legislativa.

Palabras claves: SUBE, Tweets, Gestión y Gobierno de datos, Aprendizaje automático, Toma de decisiones.

Contenido

Introducción

La explosión de los medios digitales durante el presente siglo ha facilitado la posibilidad de contar con grandes volúmenes de datos. Esto dio lugar a la conformación de un ecosistema de datos, sobre el que las diferentes organizaciones se encuentran cada vez más interesadas, en la medida que les brinda la oportunidad de obtener más información para mejorar su estrategia de toma de decisiones. Pero tales datos cuentan con la particularidad de poseer diversidad, de ser a gran escala y sin una estructura definida, lo que conlleva una dificultad para su tratamiento.

Una de las principales fuentes digitales de datos son las redes sociales. En gran medida esto se debe a que las organizaciones las utilizan como un canal de comunicación, ya sea para brindar informar a sus clientes, así como para recibir diferentes tipos de consultas, reclamos o quejas. De esta manera, se convierten en un espacio generador de datos ya que son los propios individuos los que terminan exponiendo características sobre sí mismos. Es así como, las organizaciones, tienen interés en poder analizar de manera simple y eficiente los comentarios que realizan los usuarios en sus redes sociales.

Es en este sentido que en el presente trabajo se persigue el objetivo de establecer el marco necesario de gestión y gobernanza de datos para que los resultados obtenidos a través del uso de técnicas de Data Mining en el procesamiento y análisis de datos alternativos, se constituyan en un verdadero valor agregado en el proceso de toma de decisiones de una organización. De esta manera, se demuestra la potencialidad que posee la explotación de tal tipo de datos, al mismo tiempo que se pone de relieve el marco bajo el cual debe llevarse a cabo.

Para llevar adelante esta tarea, se abordarán dos visiones simplificadoras. Por un lado, aquella que muestra una posición optimista sobre la utilización de este tipo de datos, mediante la obtención de resultados. Por el otro, aquella que advierte sobre la necesidad de contar con un sistema gobernanza y gestión de datos para garantizar la calidad de estos, reparando en cuestiones sobre la privacidad de datos privados, de forma tal que la información obtenida se constituya en un verdadero valor agregado para la toma de decisiones.

En este contexto, la elección de caso para el desarrollo del presente trabajo consiste en, por un lado, obtener y manipular los datos contenidos en los tweets realizados por los usuarios de la tarjeta SUBE cuando estos arroban a la cuenta oficial de SUBE en Twitter. Por el otro, establecer las condiciones generales necesarias para gestionar y gobernar tales datos para la toma de decisiones. Es en este sentido, entonces, que surge un interrogante principal a resolver: ¿Cuál es la intención del contenido de los tweets realizados por los usuarios de la Tarjeta SUBE cuando arroban a la cuenta oficial de SUBE en Twitter?

Para poder responder al interrogante planteado, en primer lugar, se evalúa la relevancia de la utilización de datos alternativos para la toma de decisiones en las organizaciones, así como también, se analizan las implicancias del procesamiento involucrado. Mediante este objetivo se pretende demostrar que el procesamiento de datos alternativos en un contexto de grandes volúmenes de datos permite generar valor para la toma de decisiones en una organización. De esta manera, se expone la necesidad de contar con un sistema de gestión de datos que permita establecer un repositorio de datos maestros con el fin de lograr la integridad y calidad de los datos obtenidos de fuentes alternativas.

En segundo lugar, relevar datos no estructurados y analizar su impacto para la toma de decisiones en un contexto organizacional. Mediante el procesamiento de datos obtenidos de tweets realizados por usuarios de la tarjeta SUBE, se mostrará el tratamiento necesario que debe darse a los datos mediante técnicas de Data Mining para que los mismos puedan estructurarse y de ese modo evaluar, a partir de la aplicación de algoritmos de Text Mining y la obtención de resultados, su potencial impacto para la toma de decisiones. La hipótesis que busca resolver este objetivo es que al menos el 50% de los usuarios de la tarjeta SUBE que arroban a la cuenta SUBE Oficial en Twitter es para realizar algún tipo de reclamo.

Para poder cumplir con los objetivos mencionados, el presente trabajo se estructura del siguiente modo. En el capítulo uno se realizará un análisis textual de diferente material bibliográfico teórico, exponiendo diferentes puntos de vista que permitan dar un abordaje amplio sobre la relevancia de la utilización de datos alternativos en grandes volúmenes para la toma de decisiones en un contexto organizacional. A su vez, se complementará con diferentes estudios de casos realizados de forma tal de lograr el entendimiento e identificación de aquellos puntos relevantes en el proceso involucrado.

En el capítulo dos, se abordará el relevamiento de datos y su posterior análisis. En cuanto a la recolección de datos se obtendrán datos de la red social Twitter, específicamente tweets realizados por usuarios de la tarjeta SUBE que arroban a la cuenta oficial de SUBE en Twitter. Esto se realizará a través de una API de Twitter de uso público y gratuito, con conexión a través del software R. Mediante la implementación de técnicas de Data Mining, se buscará obtener el procesamiento de los datos obtenidos para, en primera instancia, poder estructurarlos de forma tal que facilite su interpretación y posterior análisis.

Para la construcción de un análisis de impacto en la toma de decisiones en un contexto organizacional, se implementarán algoritmos de Text Mining de forma tal de arribar a resultados concretos, que se constituirán en medidas cuantitativas de la generación de valor para la toma de decisiones. Los algoritmos propuestos a utilizar son Nube de Palabras (o Word Cloud), Análisis de Sentimiento (o Sentiment Analysis) y LDA (Latent Dirichlet Allocation).

Finalmente, para una gestión responsable de datos en un contexto de grandes volúmenes de datos y, teniendo en cuenta la realización del primero y segundo objetivo, en el capítulo tres se propone evaluar la responsabilidad involucrada en los procesos implementados de forma tal de proponer algunos lineamientos sobre calidad y privacidad que hacen a un sistema de gobierno de datos. Para ello se hará especial foco en la privacidad de los datos. Se abordará desde una visión regulatoria, con énfasis en la ética individual como de investigación, y legislativa, en el marco de la existencia de ley 25.326 sobre protección de datos personales en Argentina.

Capítulo 1: Gestión de datos en contextos organizacionales

Desde fines del siglo pasado, la evolución tecnológica dio lugar a la obtención de grandes volúmenes de datos (Big Data). En una primera etapa, la problemática asociada se centró en la obtención de un almacenamiento eficiente de los datos. Superada la misma y con el correr de los años, el concepto evoluciona hacia un sistema de mayor complejidad, abarcando un conjunto de tecnologías que fueron transformando de manera disruptiva a la sociedad en su conjunto. En este sentido es posible hablar de un ecosistema de Big Data (Kolanovic y Krishnamachari, 2017) que ha tenido un alto impacto generando desde nuevas formas comunicacionales hasta cambios de hábitos en los individuos, causando un cambio socioeconómico alrededor del mundo entero.

Con la llegada del presente siglo, el mercado no quedo exento de tales consecuencias y son las organizaciones las que comienzan un proceso transformador a partir de la producción y disponibilidad de grandes volúmenes de datos. De aquí que la importancia de los datos en las organizaciones cobra una relevancia transformadora. Podría decirse que, se comienza a transitar una revolución empresarial impulsada por los datos (Schmarzo, 2013).

Dentro de ese marco, las organizaciones ya no son su único proveedor de datos. También comienzan a surgir nuevas fuentes de datos. Entre las más disruptivas de este último tiempo se encuentran las redes sociales. Dichas redes presentan la característica principal de ser de libre acceso para todos los individuos, dado el alcance masivo del uso de internet, acompañado por la evolución de los diferentes dispositivos tecnológicos.

Su expansión dio origen a concentrar a los individuos en nichos posibilitando la segmentación de los usuarios en función de miles de características que ellos mismos exponen en las mencionadas redes. Entonces, su potencial radica en el hecho de ser una fuente de datos que permite obtener atributos sobre los diferentes individuos (Preot¸iuc-Pietro, Lampos y Apetras, 2015). Es así como surge el interés por parte de las organizaciones para poder extraer y explotar los datos que las diferentes redes sociales ponen a disposición, con el fin de poder sumar información nueva o completaría a sus procesos productivo, permitiéndoles generar mayor o nuevo valor agregado.

Al mismo tiempo, no resulta de menor importancia las cuestiones que giran en torno a calidad y gobernanza de datos en un contexto organizacional. En este sentido un eficiente sistema de gobierno de datos permite que una organización pueda apalancar la información extraída de los datos como un recurso con valor agregado para la toma de decisiones (Martínez, 2012). Ahora bien, el resultado final de las decisiones tomadas dependerá de la calidad de los datos sobre las cuales se les dio sustento, resultando un punto de vital importancia para ser tenido en cuenta en un contexto de Big Data.

Es entonces como la gestión de datos toma un rol central, dado que se trata de un proceso a partir del cual la organización gestionará la información como un recurso integrado (Smith y McKeen, 2007), frente a la necesidad de establecer reglas y políticas que permitan garantizar la confiabilidad en los datos por parte de los usuarios finales. De esta manera la claridad en las definiciones de conceptos que son transversales a toda la organización se constituye en el objetivo central de un plan de gestión de datos.

El capítulo se estructura de la siguiente manera: en el siguiente aparatado, se explica el concepto de Big Data de manera sencilla, a partir de tres características esenciales. Tales características son el volumen, la velocidad y la variedad de los grandes volúmenes de datos. En el siguiente apartado, se realiza un abordaje de las implicancias que conlleva la determinación de un sistema de gestión de grandes volúmenes de datos, basado en las características definitorias del Big Data. Finalmente, en el tercer apartado, se aborda la gestión de datos alternativos, fundamentado en la existencia de una gran diversidad en los datos. Así, este capítulo se constituye en el punto de partida del presente trabajo, exponiendo los conceptos iniciales que dará sustento al desarrollo de los siguientes capítulos.

1.1 Acerca de grandes volúmenes de datos

Con la explosión del Big Data, toma mayor fuerza la idea de que no es posible manejar aquello que no se puede medir (McAfee, Brynjolfsson, Davenport y Barton, 2012). La disponibilidad de datos en formato digital comienza a tomar mayor relevancia en el proceso de toma de decisiones, desde el momento que surge la posibilidad de obtener mejores resultados que llevan a un mayor conocimiento del propio negocio. A su vez,

permite obtener ventajas competitivas en la medida que permite conocer no solo cuales son las preferencias y características de los clientes o consumidores sino, también, ofrece la oportunidad de conocer cuál es el accionar de sus competidores. En particular, la disponibilidad online de datos conlleva una dinámica de volumen, variedad y velocidad de disponibilidad de información que puede ser utilizada para adoptar cambios estratégicos con mayor inmediatez.

La definición de Big Data no solo refiere a la idea de disponer de grandes volúmenes de datos, que pueden ser de tipo estructurados y no estructurados, sino también refiere a un conjunto de tecnologías que abarcan desde el almacenamiento, procesamiento y transformación de los datos. En primer lugar, se establecen tres características definitorias del Big Data: volumen, velocidad y variedad.

Por volumen se hace referencia a la tecnología necesaria para recolectar y almacenar grandes volúmenes de datos con el fin de procesarlos para transformarlos en información de utilidad. Dada la explosión de datos digitales en la última década, el volumen de datos comenzó a crecer de manera exponencial, con la particularidad de ser mayoritariamente datos de tipo no estructurado agregando complejidad para su almacenamiento y procesamiento. Este tipo de datos son los obtenidos de redes sociales, sensores, imágenes entre muchas otras fuentes. La magnitud del volumen de tal tipo de datos ha llevado al desarrollo de nuevas formas de negocio o de transformación de los existentes, en la medida que se ha encontrado una forma clara de explotarlos (Eberendu, 2016).

En cuanto a la velocidad, resulta ser en muchos casos más importante que el volumen (McAfee, Brynjolfsson, Davenport y Barton, 2012) en la medida que una organización cuente con la agilidad suficiente para captar los datos en tiempo real frente a su competidor. En este sentido, los datos ya dejan ser un stock para ser un flujo constante (Eberendu, 2016), lo que lleva a la necesidad de procesamiento diario y hasta por hora en ocasiones. De aquí que, el procesamiento requiere de cierta inteligencia superior para lograr captar la mayor cantidad de tipos de datos provenientes de diferentes fuentes y con la mayor velocidad posible, de forma tal de convertirlos en valor agregado para la toma de decisiones.

La variedad de datos resulta de la existencia de una amplia diversidad a partir de la disponibilidad digital de los mismos. Desde datos obtenidos a partir de emails, transacciones online hasta los derivados de videos, audios entre otros. Dada esta diversidad, el procesamiento de datos para la toma de decisiones requiere de procesos que sinteticen y simplifiquen la información que pueda obtenerse de ellos. A su vez, presentan la posibilidad de contar con información sobre cualquier tema de interés para el negocio (McAfee, Brynjolfsson, Davenport y Barton, 2012), llegando a conformar entre el 70 y 80 por ciento de los datos utilizados por las organizaciones (Eberendu, 2016).

De esta manera, la conjunción de volumen, velocidad y variedad de datos que ofrece un sistema de Big Data, lo convierten en un ecosistema de datos que ha implicado una transformación disruptiva en las organizaciones a la hora de delinear estrategias de negocio. Y es en este sentido que surge la necesidad de contar con un maestro de datos que sea transversal a la organización con el fin de brindar consistencia y calidad sobre estos para que realmente sean convertidos en valor agregado.

1.2 Gestión de grandes volúmenes de datos

En un contexto de grandes volúmenes de datos, comienza a surgir la necesidad de estructurar un proceso que se convierta en el maestro de datos centrales de toda una organización, dada la diversidad de orígenes de datos de la que se dispone. De aquí que el maestro de datos (Master Data o sus siglas MD) se define como las tecnologías necesarias para desarrollar un conjunto de atributos consistentes que describen los conceptos centrales de los datos para toda una organización (Smith y McKeen, 2008). Conceptos que permitirán integrar, analizar y extraer el valor que contienen los datos independientemente de la fuente origen de estos (Cleven y Wortmann, 2010).

Las organizaciones al enfrentarse con un ecosistema de datos que surge en un contexto de Big Data requieren contar con un repositorio de metadatos. Este se convertirá en el objetivo central de definiciones que será transversal a toda la organización, permitiendo que convivan diferentes sistemas y procesos para la construcción de información a partir de datos diversos. Tal necesidad surge como consecuencia de inconsistencia en los datos por encontrarse almacenados en diferentes bases, sin reglas transversales que permitan su fácil entendimiento y utilidad para la organización en su conjunto. A su vez, contribuirá

a facilitar la construcción de la arquitectura de almacenamiento en la medida que permitirá captar las necesidades de información de los diferentes usuarios dentro de la organización.

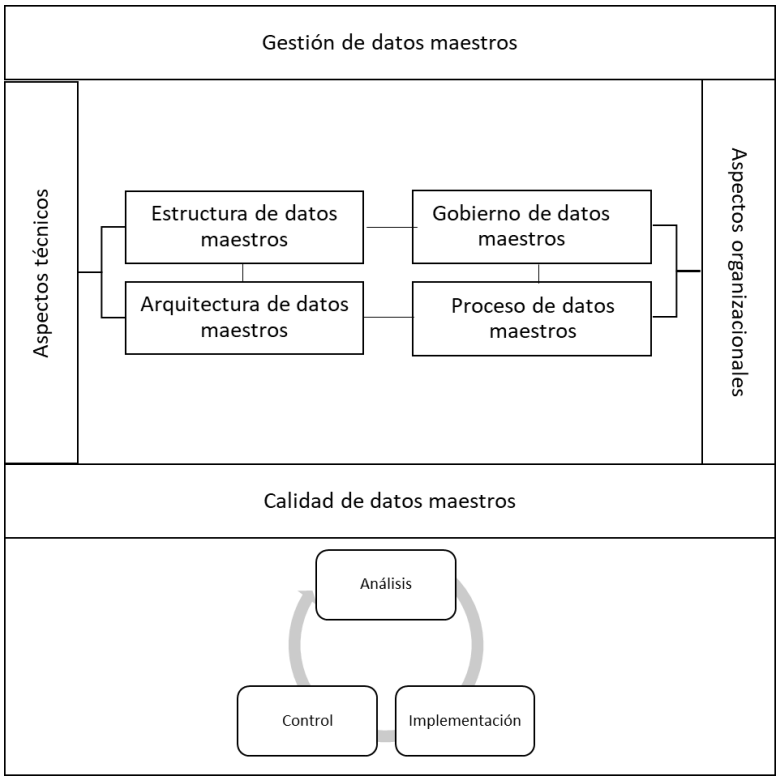
En este sentido, los datos maestros se refieren a entidades centrales del negocio de una organización que se utilizan repetidamente en muchos procesos y sistemas (Cleven y Wortmann, 2010). A diferencia de datos transaccionales, los datos maestros no cambian durante su ciclo de vida, dado que tienen existencia propia y presentan la característica de ser de un volumen relativamente constante. De esta manera, se convierten en un activo clave dentro una organización. Por consiguiente, la falta de una adecuada gestión de estos puede derivar en el surgimiento de diferentes problemas, como fallas en el funcionamiento de procesos hasta una mala toma de decisiones por generación de datos erróneos. Es entonces que surge la necesidad de contar con un responsable de datos maestros, ya que será quien determine el significado de las entidades maestras además de sus reglas de uso.

En función del riesgo asociado a la falta de datos maestros se pone de relevancia la problemática a la cual deben enfrentarse las organizaciones para gestionar grandes volúmenes de datos. En este sentido surgen cinco aspectos principales para tener en cuenta a la hora de diseñar un sistema de datos maestros: la estructura, la arquitectura, el gobierno, el proceso y la calidad de datos maestros (Cleven y Wortmann, 2010). En la figura 1 se resumen los cinco aspectos principales de un sistema de gestión de datos. La estructura de datos maestros implica establecer un acuerdo sobre cada definición para cada entidad al mismo tiempo que requiere definir la estructura relacional entre las diferentes entidades. Luego, la arquitectura de datos maestros implica la necesidad de establecer el diseño de sistemas adecuados que den soporte al ciclo de vida de los datos maestros.

En cuanto a la gobernanza de datos, se requiere establecer una articulación entre todas las áreas de la organización para delinear adecuadamente una estructura de datos sólida. Para ello se deberán definir los roles y responsabilidades sobre los datos en cada área. En línea con este punto, el proceso de datos maestros refiere a la necesidad de establecer un proceso que determine la manera en que los datos maestros deben ser creados, usados,

mantenidos y almacenados. Finalmente, la calidad de datos maestros resultará como consecuencia de la implementación correcta de las cuatro fases anteriores.

Figura 1: Estructura de Gestión de datos maestros



Elaboración propia en base a (Cleven y Wortmann, 2010)

Para garantizar un resultado exitoso en la calidad de datos maestros se requiere de un proceso interactivo continuo (Cleven y Wortmann, 2010). El mismo consta de tres etapas: análisis, implementación y control. En una etapa inicial de análisis se identificará los datos claves que representan las entidades maestras de la organización. Luego en una etapa de implementación se buscará definir los conceptos que serán transversales a toda la organización. Y, finalmente, se requerirá de una etapa de control que permita garantizar la sincronización de los datos maestros. De esta manera, se logrará un sistema de datos maestros consolidado.

Podría resultar mas sencillo de implementar un sistema de gestión de datos como el descripto si los datos que maneja una organización fueran únicamente estructurados. Pero a la hora de enfrentarse con datos de tipos no estructurados, requerirá de algunas transformaciones en el proceso estándar de incorporación de datos, para que los mismos puedan integrarse de manera exitosa al sistema de datos de la organización.

1.3 Gestión de datos alternativos

En una era de datos digitales, los datos resultan de una gran diversidad y con enorme escalabilidad, siendo en su mayoría datos de tipo no estructurado. La incorporación de estos requiere de una transformación en el diseño del proceso de almacenamiento de datos. Tradicionalmente, dado un tipo de dato estructurado, el proceso de almacenamiento de los datos consistía en la extracción, transformación y carga (ETL sus siglas en inglés). Frente a la necesidad de incorporar un nuevo tipo de dato que esencialmente resultan ser de tipo no estructurado, se plantea pensar de manera diferente aquel proceso. En este sentido, se propone un enfoque de extracción, carga y transformación (Schmarzo, 2013).

Una característica particular de este tipo de datos es que son datos automáticos. La particularidad que poseen es que se obtienen de manera casi inmediata, a través de diferentes dispositivos como por ejemplo sensores, GPS entre muchos otros. Frente a ello, las organizaciones se encuentran interesadas en poder captarlos con mayor velocidad a diferencia de un procesamiento ETL. Tal interés surge con el objetivo de poder obtener una ventaja de negocio frente a sus competidores hasta poder mejorar la toma de decisiones en forma ágil. Para este tipo de casos, existen diferentes alternativas de almacenamiento de datos que están orientadas a un procesamiento en paralelo de forma tal de garantizar un acceso performante a los datos y a alta velocidad. En este sentido, podría decirse que, desde una visión técnica, la tecnología necesaria para procesar tales datos se encuentra resuelta.

No obstante, para lograr la calidad e integridad de los datos se requiere aplicar un proceso de datos maestros como el mencionado en el apartado anterior. La necesidad de hacer convivir datos de diferente tipo provenientes de diferentes fuentes solo puede garantizarse su éxito a través de la existencia de un maestro de datos. Este proceso permite integrar datos de calidad con el fin de obtener información enriquecida que permita generar valor agregado. Para ello la etapa de control se convierte en vital, dado que permite sincronizar, de ser necesario en tiempo real, los datos maestros. Esto es, durante la extracción y carga de datos podría detectarse nuevos valores que requieran una modificación de los datos maestros para que los nuevos datos puedan ser integrados. Tal

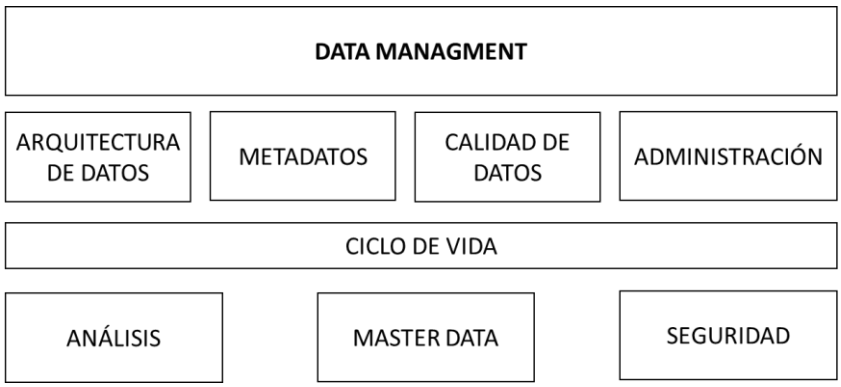
detección puede obtenerse de manera automática o como resultado de un análisis realizado por alguna persona a cargo.

Bajo esta nueva metodología de integración de datos alternativos, podrían surgir metadatos que sean temporales y otros permanentes. Estos últimos requieren de almacenamiento en el maestro de datos mientras que los primeros pueden ser útiles durante el tiempo de ejecución de un proceso (Russom, 2015). A diferencia de los datos estructurados y su procesamiento en formato ETL, los datos no estructurados podrían presentar un ciclo de vida corto que no requerirán de un almacenamiento permanente. Es en este sentido que el tratamiento de datos alternativos marca una diferencia a ser tenida en cuenta de cara al futuro.

Finalmente, a pesar de que la integración de datos alternativos requiere de adaptaciones o modificaciones de los procesos tradicionales de gestión de datos, representan un claro potencial para la obtención de ventajas competitivas. Bajo un esquema adecuado de gestión, puede lograrse la integridad de estos, y garantizar su calidad. De esta manera, se logrará un proceso de toma de decisiones más robusto en la medida que se sustenta en mayor conocimiento a partir de obtener más información.

Por lo tanto, el ciclo de vida del dato debe ser administrado por un proceso de gestión de datos (Data Managment) de forma tal de garantizar la integridad y calidad de los datos. Para llevar a cabo una gestión eficiente se propone el siguiente esquema:

Figura 2: Estructura de Data Managment.



Elaboración propia.

donde la arquitectura de datos se concentra principalmente en establecer el almacenamiento, organización e integridad de los datos. En la etapa de elaboración de los

metadatos, se busca establecer la documentación y posterior publicación sobre la información contenida en los datos maestros para toda la organización. La calidad de los datos busca garantizar la calidad de estos, así como su integridad y su enriquecimiento. La administración refiere a la administración de la base de datos, la cual debe ser de forma diaria para brindar mayor performance. En cuanto a la seguridad, refiere a la definición y administración del acceso a los datos. El administrador de datos maestros (Master Data Manager) se encarga de gestionar y dar coherencia a los datos maestros. Y, finalmente, una etapa de análisis que facilite la generación de informes otorgando un valor agregado para la toma de decisiones.

A modo de conclusión, en este primer capítulo se pone de relieve las características definitorias del Big Data, donde el volumen, velocidad y variedad conllevan a la necesidad de contar con un sistema de gestión de datos maestros que permitan lograr la integridad y calidad de los datos. Datos que pueden ser de diferentes tipos y provenir de diferentes fuentes origen. En particular, los datos alternativos desafían a las organizaciones a considerar adaptaciones de estructuras tradicionales en el tratamiento de los datos frente a la oportunidad de obtener ventajas competitivas, generando un valor agregado en la producción de información para la toma de decisiones. Para lograr con éxito esta tarea, se debe tener en cuenta que el proceso de obtención de datos de calidad debe estar sustentado en un proceso de análisis, implementación y control continúo. En este sentido, la automatización de procesos puede contribuir de manera eficaz para el ejercicio de detección de nuevos valores en los datos, otorgando la posibilidad de adoptar cambios estratégicos de manera inmediata.

En el siguiente capítulo se llevará a cabo una implementación con técnicas de Data Mining que pondrán de manifiesto el tipo procesamiento de datos alternativos. Asumiendo que los mismos pueden integrarse a un sistema de gestión datos existente, este procesamiento permitirá estructurar los datos de forma tal que facilite la implementación de técnicas de Text Mining con el objetivo de arribar a resultados concretos. Con tales resultados se buscará cumplimentar con el segundo objetivo específico del presente trabajo: el 50% de los usuarios de la Tarjeta SUBE que arroban a la cuenta SUBE Oficial en Twitter es para realizar algún tipo de reclamo.

Capítulo 2: Procesamiento y análisis de datos alternativo

Los datos pueden ser principalmente de dos tipos: estructurados y no estructurados. Por datos estructurados se refiere a aquellos que se encuentran ordenados en formatos de columnas y filas perfectamente legibles y listos para ser procesados. Por el contrario, los datos no estructurados son aquellos que no poseen una estructura interna, razón por la cual requieren de un tratamiento inicial para poder hacer uso y comprensión de estos.

En particular, los datos no estructurados, también llamados datos alternativos, pueden presentar alguna ventaja en la información que proporcionan. Tal ventaja puede consistir en descubrir nueva información no contenida en fuentes tradicionales, o descubrir una misma información, pero anticipadamente (Kolanovic y Krishnamachari, 2017). Entre este tipo de datos se encuentran los producidos y publicados por los individuos como por ejemplo las publicaciones que realizan en redes sociales, los generados a partir de transacciones online como los provenientes de comercio electrónico entre otros, o aquellos generados por sensores como por ejemplo las imágenes satelitales.

Dada las características de los datos no estructurados, las técnicas de Data Mining ofrecen la posibilidad de aplicar algoritmos que permiten procesar datos hasta en tiempo real y de gran volumen. Una metodología frecuente en la que se sustentan para el procesamiento de texto como es el caso del texto contenido en las publicaciones en redes sociales, es el procesamiento de lenguaje natural (o Natural Language Processing).

El procesamiento de lenguaje natural consiste en un análisis de datos de texto, utilizando métodos computacionales, cuyo objetivo es la construcción de una representación sobre el contenido del texto que agregue una estructura al lenguaje natural no estructurado contenido en el cuerpo de aquel (Verspoor y Khoen, 2013). 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.

Tal proceso consiste en determinar las 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 los sufijos y la pluralidad (Stemming) con el fin de obtener un procesamiento más veloz (Fiore, Almodovar, Assoumou, Dutta, y Cotoranu, 2017). A partir de dicha estructura base de procesamiento es que surgen diferentes técnicas de análisis para este tipo de datos que se engloban en el conjunto de algoritmos de Text Mining.

El capítulo se estructura de la siguiente manera: en el siguiente aparatado, se explica cómo se obtuvieron los datos que conforman el set de datos a ser procesado y analizado. A partir de ello se desarrolla el procesamiento de datos llevado a cabo. En el siguiente apartado, se implementan técnicas de Text Mining sobre los datos estructurados obtenidos del procesamiento implementado, realizando un análisis sobre los resultados obtenidos. Finalmente, en el tercer apartado, se especifican los resultados alcanzados. En función de estos, se pone de relevancia el potencial que tienen dichas técnicas como medio de generación de valor agregado para la toma de decisiones.

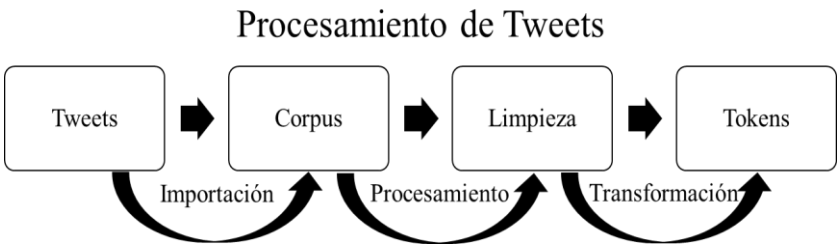
2.1 Obtención y procesamiento de tweets

Frente al objetivo de relevar datos no estructurados y analizar su impacto para la toma de decisiones en un contexto organizacional, en el presente apartado se desarrollará el procesamiento de datos realizado mediante técnicas de machine learning. Los datos están conformados por tweets realizados por usuarios de la tarjeta SUBE que arroban a la cuenta oficial de SUBE en Twitter. Tales datos se obtuvieron a través de una API de Twitter de uso público y gratuito, con conexión a través del software RStudio. Dado que existe una limitación en cuanto a la cantidad de tweets que es posible extraer para un limitado período de 7 días, se realizaron diferentes extracciones en diferentes periodos de tiempo, entre febrero y septiembre de 2019. De esta manera, se logró conformar un set de datos con 5.851 tweets, el cual se almacenó en formato CSV (Comma-separated values).

Conformado el set de datos, se procedió a implementar técnicas de Data Mining para realizar la limpieza de estos y luego estructurarlos de forma tal que facilite su interpretación y posterior análisis. En la figura 2 se especifican las etapas de procesamiento. En primer lugar, se importa el archivo que contiene los datos. Luego, se construyó un corpus a partir del texto contenido en cada tweet, siendo un corpus una colección de documentos, donde cada documento es un texto. En tercer lugar, se lleva a

cabo un proceso de limpieza del texto contenido en cada documento. Esto, con el fin de eliminar todo tipo de simbología y direcciones web que no constituyen una palabra que pueda aportar significado. Adicionalmente, se eliminan aquellas palabras que no aportan significado al sentido del texto (Stop Word), como son las preposiciones entre otras. También se realiza Stemming con el fin de reducir las palabras a su nivel raíz removiendo los sufijos y la pluralidad. Finalmente, se obtienen Tokens, siendo unidades de textos conformadas por palabras simples.

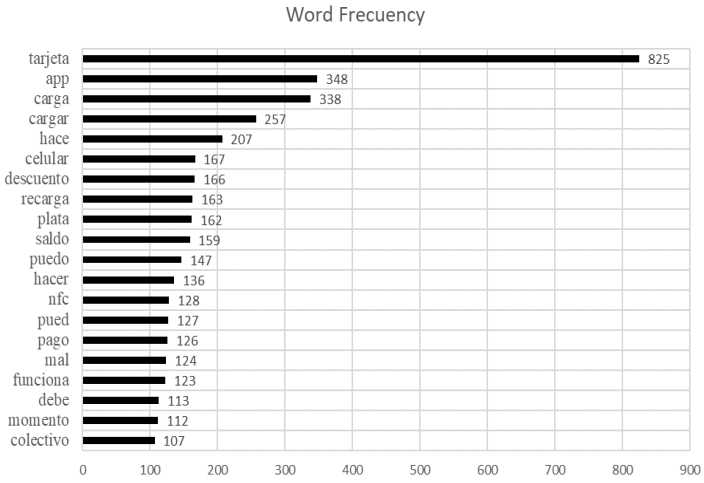
Figura 3: Etapas de procesamiento.



Elaboración propia.

De esta manera, se estructuraron los tweets iniciales en un corpus que contiene 5.851 documentos, donde cada documento contiene tokens. La totalidad de los tokens alcanzan las 49.837 unidades de textos. Realizando un top de palabras en función de su frecuencia de aparición a lo largo del corpus, se puede observar en la figura 4, que la palabra “tarjeta” se encuentra en el primer lugar. Este es un resultado esperado dado que se trata de detectar problemáticas asociadas al uso de la tarjeta SUBE. Resulta interesante que en segundo, tercer y cuarto lugar aparecen las palabras “app”, “carga” y “cargar”. Es de suponer que pueda asociarse alguna problemática a dichos términos.

Figura 4: Top 20 de frecuencia de palabras.



Elaboración propia con RStudio.

Dada esta suposición, se construyen bigrams, siendo una conjunción de dos palabras. Tal conjunción permite asociar una palabra a la palabra tarjeta con el objetivo de determinar un sentido concreto para definir una posible problemática inicial.

Tabla 1: Top 10 Bigrams.

Bigrams	Frecuency
carga tarjeta	20
registrar tarjeta	19
nueva tarjeta	18
numero tarjeta	18
baja tarjeta	15
cargar tarjeta	13
saldo tarjeta	13
perdí tarjeta	11
problema tarjeta	8

Elaboración propia con RStudio.

Frente a la exploración inicial realizada, es posible notar que un procesamiento como el realizado permite comenzar a ver posibles problemáticas asociada al uso de la tarjeta SUBE que los usuarios expresan a través de la red social Twitter. En el siguiente aparatado se propone implementar diferentes técnicas de análisis que permitirán arribas a resultados más concretos acerca de la detección de problemáticas.

2.2 Análisis de datos no estructurados mediante técnicas de Text Mining

La minería de texto (o Text Mining), puede definirse como el proceso de descubrimiento y extracción de conocimiento a partir de un texto no estructurado (Kao y Poteet, 2007). En este sentido se constituye en un conjunto de herramientas que brinda la posibilidad de examinar grandes volúmenes de texto, con el objetivo de generar información para su posterior análisis de forma tal de arribar a un resultado concreto.

Existen diversos algoritmos de Text Mining, que permiten obtener nuevo conocimiento a partir de un texto. Entre los más frecuentemente utilizados se encuentra Nube de Palabras (o Word Cloud). Este suele utilizarse en una etapa inicial de análisis con el fin de detectar en un texto las palabras más frecuentemente utilizadas, mediante una visualización rápida y sencilla. Su potencial, para el caso de analizar tweets, radica en el hecho de que permite obtener una visión de lo que está pensando el usuario, basada

en la frecuencia de utilización de determinadas palabras (Kabir, Karim, Newaz y Hossain, 2018).

Para poder construir una nube de palabras primero debe crearse un Document Term Matrix (DTM) que consiste en una matriz cuyas filas son los documentos creados y cuyas columnas son cada uno de los términos contenidos en los documentos. Luego, el contenido de cada celda será la frecuencia en que ocurrió cada término en cada documento. La ventaja que presenta esta representación consiste en que se pasa de un formato de texto a un formato numérico, permitiendo realizar operaciones en memoria de manera optimizada. Llevando a cabo esta implementación se obtuvo:

Figura 5: Nube de palabras.

Elaboración propia con RStudio.

En la figura 5, se puede observar fácilmente que la palabra “tarjeta” resulta ser la más utilizada en los comentarios realizados por los usuarios de la red, lo cual resulta esperable, como ya se mencionó anteriormente. Adicionalmente, es de destacar que surgen diferentes palabras con importancia alta a su alrededor. “Descuento, cargar, carga, app, nfc” son las que presentan mayor importancia alrededor de la palabra tarjeta. Llegado a este punto, puede establecerse que se cuenta con un indicio de posibles problemáticas a partir de la identificación de palabras claves.

Tales problemáticas surgen frente a la imposibilidad de realizar una recarga de crédito en la tarjeta ya que implica la imposibilidad de poder realizar un viaje en transporte

público. A su vez, que surjan palabras como “app” y “nfc”, también cobra sentido en la medida que la organización SUBE lanzó la posibilidad de realizar acreditación de recargas de crédito a través de una App (abreviatura de Application) en celular. El requisito técnico, además de contar con sistema Android, es que el modelo de celular cuente con tecnología NFC (Near Field Communication). Sin esta, no es posible realizar la aplicación de recargas, es decir, el aplicativo no funciona.

Por último, también surge la palabra “descuento” como otra posible problemática que expresan los usuarios. Por lo que se pudo analizar, los tweets en relación con este concepto son por reclamos sobre la no obtención de descuentos a pesar de poseer algún tipo de beneficio. Entre los principales, se encuentran los reclamos asociados al Sistema Red Sube. Este sistema, implica la aplicación de un descuento del 50% en la tarifa del segundo viaje realizado, siempre que el mismo este dentro un periodo de tiempo no mayor a dos horas. Y de un 75% de descuento a partir del tercer viaje si el mismo también es realizado dentro del período horario especificado anteriormente. De esta manera, se obtiene una clara primera aproximación a establecer algunas problemáticas aparentes.

Una primera aproximación para establecer asociaciones entre palabras claves se puede obtener utilizando un algoritmo de clúster aglomerativo, como resulta ser el clúster jerárquico a través del método Ward. Este método se basa en calcular la distancia al cuadrado entre la frecuencia de las observaciones respecto de un centro, siendo en general la media de los valores, buscando establecer uniones de puntos de mínima varianza. Esto es, dado dos clústeres A y B:

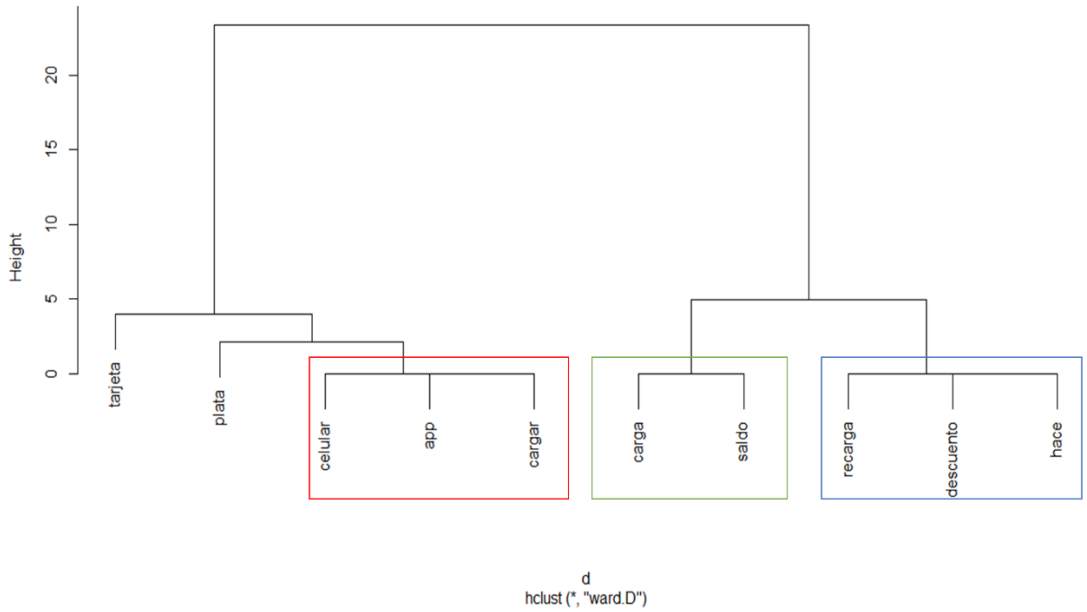
$$\Delta(\boldsymbol{A}, \boldsymbol{B}) = \frac{\boldsymbol{n_A n_B}}{\boldsymbol{n_A + n_B}} \|\overrightarrow{\boldsymbol{m_A}} - \overrightarrow{\boldsymbol{m_B}}\|^2$$

donde $\overrightarrow{\boldsymbol{m_j}}$ es el centro del clúster j, y $\boldsymbol{n_j}$ es el número de puntos contenido en el clúster. Luego Δ es el costo marginal de combinar el clúster A y B. De esta manera el algoritmo, consiste en los siguientes pasos:

1- Inicia evaluando cada punto para asignarle un grupo
2- Selecciona los pares de puntos más cercanos y los junta en un grupo
3- Finalmente, devuelve el árbol de grupos

Partiendo de la matriz de documentos creada, se lleva a cabo esta implementación y se obtuvo:

Figura 6: Clúster entre palabras.



Elaboración propia con RStudio.

De la figura 6 se observa que surgen tres grupos principales en torno a las palabras “carga y saldo”, “cargar, celular y App” y “recarga y descuento”. Con esta primera aproximación se puede observar una asociación entre las palabras que fueron detalladas anteriormente como indicadoras de posibles problemáticas asociadas al uso de la tarjeta SUBE, basada en una medida robusta de cálculo.

Otro método frecuentemente utilizado es LDA (Latent Dirichlet Allocation). Se trata de un método que busca asignar una probabilidad de pertenencia a categorías determinadas. Tales categorías (Política, Salud, Economía entre muchas otras) son previamente fijadas por el analista en función del interés que posee. De esta manera, la probabilidad de pertenencia se asigna en función del patrón de palabras contenidas en el texto, asociadas a temas particulares (Munzert, Rubba, Meißner, y Nyhuis, 2014).

La noción básica detrás de LDA es que los documentos, es decir, cada tweet, se representan como mezclas aleatorias de temas latentes, siendo cada tema latente caracterizado por una distribución determinada a partir de las palabras. Secuencialmente, el proceso implica (Blei, Ng y Jordan, 2003):

- Selección de N palabras que siguen una distribución de Poisson

- Selección de θ tópicos que siguen una distribución de Dirichlet
- Para cada palabra contenida en una secuencia de palabras:
    - asigna un tópico a partir de una distribución multinomial sobre θ
    - le asigna a cada palabra una probabilidad multinomial de pertenencia, condicionada al tópico.

De esta manera, la función de densidad de probabilidad Dirichlet se define como:

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

donde el parámetro α es un vector k aleatorio de θ, cuyos componentes son mayores que 0, y $\boldsymbol{\Gamma(x)}$ 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 $\mathbb{z}$ de tópicos θ, viene dada por:

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

En definitiva, 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 documento viene dada por una distribución de probabilidad de Dirichlet, siendo una generalización de la distribución Beta (Alvares, Armero, y Forte, 2018), lo que implica que un documento sea parte de varios de tópicos donde cada uno posee un peso diferente.

Para llevar a cabo la implementación de un modelo LDA, se requiere que los datos de texto se encuentren estructurados en un formato de matriz, cuyas filas son los documentos creados (tweets) y cuyas columnas son cada uno de los términos contenidos en los documentos. Dado que ya se realizó este procesamiento previamente, a continuación, se llevó a cabo la implementación del algoritmo a partir del cual se obtuvo el siguiente resultado:

Figura 7: Tópicos obtenidos a partir de LDA.

Elaboración propia con RStudio

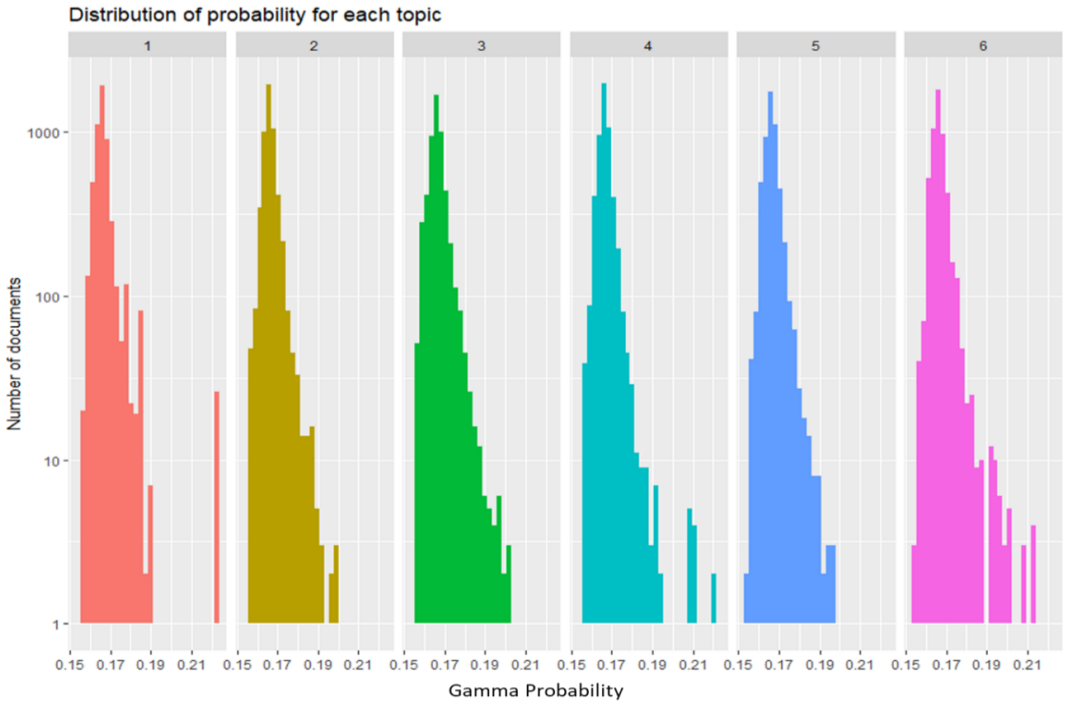
La visualización de la figura 7 permite comprender los tópicos que se extrajeron de los tweets bajo análisis. El eje horizontal de cada gráfico expresa la probabilidad beta de pertenencia de una palabra al tópico. De esta manera, las palabras más comunes en el tópico 1 resultan ser "tarjeta", "carga", "cargar" y "recarga", lo que sugiere una problemática con la carga de crédito en la tarjeta SUBE. Similar situación se presenta para el tópico 2 y 4. En cambio en el tópico 3, surge la palabra descuento como la de mayor pertenencia lo que sugiere una problemática asociada con la obtención de descuentos a la hora de utilizar la tarjeta SUBE, siendo que el individuo cuenta con algún beneficio. En el caso del tópico 5 y 6, surgen las palabras “tarjeta” y “app” como más comunes en el tópico, lo que sugiere una problemática asociada al uso de la App de SUBE donde se aplican las recargas de crédito realizadas a la tarjeta SUBE a través del celular.

De esta manera, el algoritmo implementado posibilitó la obtención de las palabras que están asociadas, con una determinada probabilidad, a cada tema o tópico. En esta

instancia, es posible afirmar que existen diferentes problemáticas asociadas a la tarjeta SUBE con tres principales tópicos: recarga de crédito (tópicos 1, 2 y 4), obtención de descuento (tópico 3) y aplicación de recargas en la tarjeta a través de la App de SUBE (tópicos 5 y 6).

Al mismo tiempo es posible determinar el conjunto de tópicos que describe a cada documento (tweet) en función de determinar la distribución de probabilidad asociada a tal conjunto de tópicos. La distribución de probabilidad utilizada en este caso resulta ser Gamma. Cada uno de los valores obtenidos será una proporción estimada de las palabras contenidas en los documentos que se generan a partir de cada tópico. Llevando a cabo la implementación se obtiene:

Figura 8: Distribución de probabilidad (Gamma) de cada tópico.



Elaboración propia con RStudio

De la figura 8 se puede determinar que alrededor del 17% de las palabras en más de 1000 tweets son generadas por el tópico 1 definido previamente como “recarga de crédito”, con similar interpretación para los tópicos 2 y 4. Del mismo modo, alrededor del 17% de las palabras en más de 1000 tweets son generadas por los tópicos 5 y 6 definido previamente como “aplicación de recargas de crédito a través de la App de SUBE”.

En consecuencia, a través de la aplicación de la metodología LDA se pudo detectar de manera concisa al menos tres problemáticas asociadas a la tarjeta SUBE: inconvenientes con recargas de crédito, obtención de descuentos y aplicación de recargas de crédito a través de la App de SUBE en celulares. En una etapa inicial, a partir de la realización de una nube de palabras que permitió una visualización sencilla de frecuencia de términos se establecieron las principales palabras que podían estar asociadas a una problemática en el uso de la tarjeta SUBE. En esta instancia, se pudo determinar que alrededor de tales palabras frecuentes efectivamente se encuentran problemáticas asociadas.

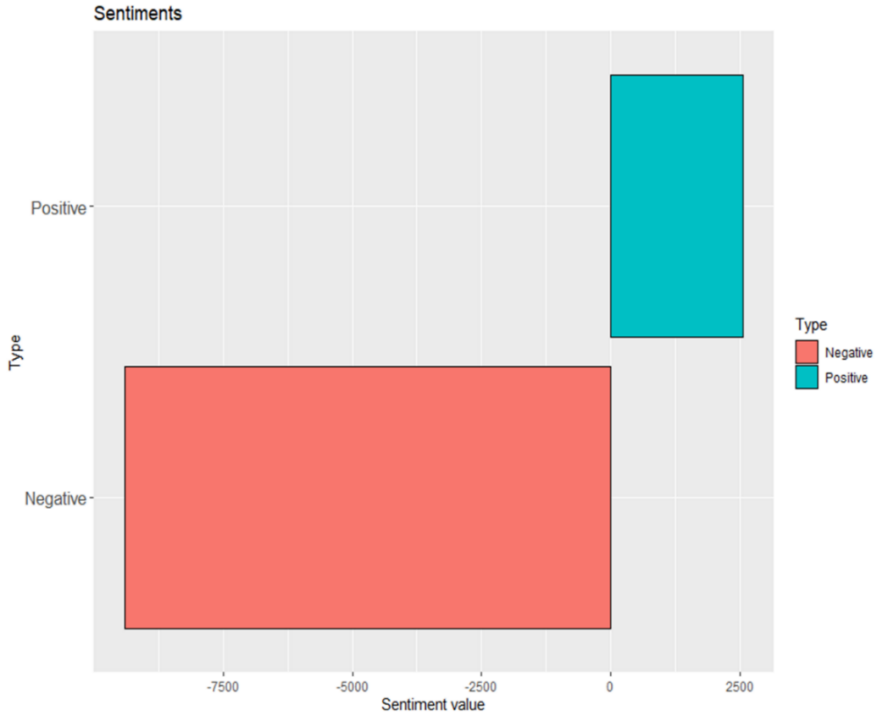
Dado que los usuarios de redes sociales son seres humanos, frecuentemente expresan emociones o sentimientos a través de su escritura, dejando una huella acerca de su pensamiento. Es por ello por lo que resulta interesante poder captar tales sentimientos a la hora de analizar tweets. Una metodología frecuentemente utilizada en este sentido es el Análisis de Sentimiento (o Sentiment Analysis). Este método, trata de buscar el sentimiento contenido en un texto, por ejemplo, en un tweet, en términos de actitudes, emociones y opiniones, con el fin de proporcionar una respuesta adecuada al usuario (Chen y Franks, 2016).

El método de análisis de sentimiento depende en gran medida de un léxico de sentimiento (u opinión) subyacente. Un léxico de sentimiento es una lista de características léxicas (por ejemplo, palabras) que generalmente se etiquetan según su orientación semántica como positiva o negativa (Hutto y Gilbert, 2014). Tal lista es validada previamente por el analista, de forma tal que pueda adecuarla al idioma con el cual esta expresado el texto bajo análisis. Luego, para aquellos sentimientos que no pueden ser definidos, se le asigna una semántica neutral.

A su vez, dentro de la clasificación en negativo o positivo, se asigna un puntaje entre 1 y 6 para determinar la intensidad del sentimiento conformando un rango de valores positivos o negativos para cada caso. La determinación del puntaje es resultado de diferentes análisis psicológicos y grafológicos sobre el impacto sentimental que le causa a un individuo determinada palabra, teniendo en cuenta la simbología utilizada, el uso o no de mayúsculas, entre otros.

Para llevar a cabo la implementación de un modelo de análisis de sentimiento, se requiere tokenizar los documentos contenidos en un corpus, es decir, generar un set de datos conformado por cada una de las palabras en formato de lista. Una vez llevada a cabo esta tarea, se realiza una unión con la lista de léxicos con el fin de clasificar a cada palabra en negativa, neutral o positiva. Finalmente, se toma el puntaje asignado a cada palabra en cada documento y se sumariza el mismo. De esta manera, se obtiene una clasificación de tweets en positivo o negativo. Llevando a cabo tal método, se obtuvo el siguiente resultado:

Figura 9: Sentimientos asociados a cada tweet.



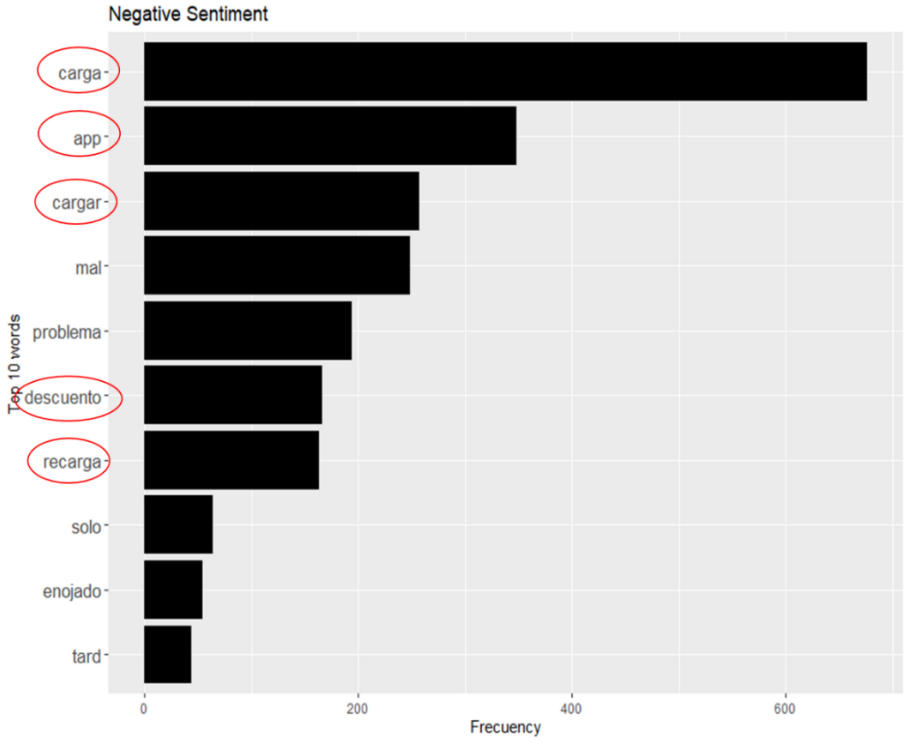
Elaboración propia con RStudio.

En la figura 9 se puede observar que el puntaje obtenido a partir de las palabras utilizadas en cada tweet resulta ser mayoritariamente con sentimiento negativo. Esto da lugar a pensar que el contenido de cada tweet está expresando una disconformidad por parte del usuario. Tal disconformidad puede ser interpretada en términos de reclamos o quejas con algún servicio relacionado a la utilización de la tarjeta SUBE. Finalmente, el promedio de puntaje obtenido es:

Sentiment Type	Average Sentiment
Negative	-9.400
Positive	2.548

De esta manera se obtuvo que en promedio el 79% del contenido de los tweets contiene un sentimiento negativo mientras que el 21% resulta ser positivo, sin considerar aquellas palabras que reciben un puntaje neutral, es decir, cero. A continuación, se realiza un top 10 de palabras asociadas a un sentimiento negativo:

Figura 10: Sentimientos negativos asociados a cada palabra.



Elaboración propia con RStudio.

De la figura 10 se puede observar que las principales palabras que surgen expresando un sentimiento negativo son aquellas que previamente se habían asociado a diferentes problemáticas. Así, la palabra “carga” se encuentra en el top de las palabras con sentimiento negativo. Y en segundo lugar surge la palabra “app”. Luego las palabras “cargas”, “recarga” y “descuento” también se encuentran como las de mayor frecuencia de utilización asociadas a un sentimiento negativo.

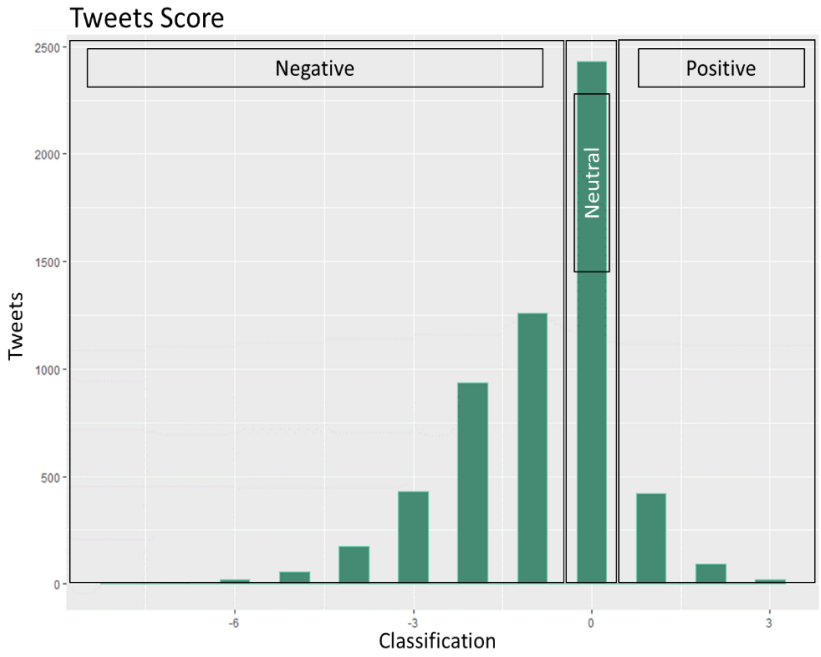
Por lo tanto, a través de la aplicación de la metodología de análisis de sentimiento se pudo determinar que la mayor parte del contenido de los tweets que realizan los usuarios de la tarjeta SUBE expresan al menos una disconformidad a la hora de realizar una recarga de crédito en general o, a través de la App de SUBE por medio de un celular. También, se detecta una asociación negativa con respecto a la obtención de descuentos. Teniendo

en cuenta que inicialmente se llevó a cabo una nube de palabras lo que permitió una visualización sencilla de frecuencia de términos donde también surgían las mismas palabras como las de mayor frecuencia, es posible determinar que efectivamente se encuentran problemáticas asociadas en torno a tales términos.

2.3 Resultados

Ahora bien, el objetivo final de análisis consiste en poder determinar una clasificación sobre el sentimiento asociado a cada tweet. Para ello, y teniendo en cuenta la metodología de análisis de sentimiento, se llevó a cabo la realización de un score de tweets que permitió ya no una clasificación individual de sentimientos por palabras sino del tweet completo. Para ello, se sumariza el puntaje obtenido por cada palabra en cada tweet. Si este resulta menor que cero, se considera que el tweet puede ser clasificado de manera negativa, mientras que si resulta mayor a cero se considera clasificado de manera positiva. En caso de sumar cero, se considera que su contenido es neutral, es decir no pudo determinarse en forma polarizada cual es el sentido que expresa el mismo. Llevando a cabo tal implementación se obtuvo:

Figura 11: Score de Tweets en función del sentimiento asociado.



Elaboración propia con RStudio.

De la figura 11 se puede observar que la cantidad de tweets clasificados como negativos es muy superior a los clasificados como positivos. A su vez, aproximadamente la mitad de la totalidad de los tweets no pudo ser clasificada de manera polarizada resultando en neutral la expresión del sentimiento de su contenido. Específicamente se pudo establecer que el 50% de los tweets analizados resultan ser negativos, entiendo que la negatividad se encuentra asociada a reclamos que realizan los usuarios:

Classification Tweets		Percentage
Negative	2.904	50%
Neutral	2.420	41%
Positive	527	9%
	5.851	100%

De esta manera, en función de los resultados obtenidos a partir de la implementación de las diferentes técnicas de Text Mining, fue posible determinar que el 50% de los tweets que realizaron los usuarios implica una disconformidad o un reclamo puntual sobre algún servicio o funcionalidad asociada a la tarjeta SUBE.

A modo de conclusión, en este segundo capítulo se llevó a cabo una implementación de Data Mining para poder estructurar datos provenientes de un contenido textual, con el objetivo de poder aplicar diferentes metodologías de Text Mining que permitan arribar a un resultado concreto. Como resultado final se pudo clasificar los tweets en función de su sentimiento contenido, detectando que el 50% de los mismos expresan alguna disconformidad o queja. Dado este escenario, se identificaron diferentes problemáticas asociadas al uso de la tarjeta SUBE. Tales problemáticas identificadas se relacionan con dificultades en la realización de recargas de créditos, utilización de la App de SUBE y obtención de descuentos a la hora de utilizar la tarjeta SUBE, siendo que el individuo cuenta con algún beneficio. De esta manera se logró demostrar que la utilización de técnicas de Text Mining permiten la generación de valor agregado a partir de obtener información sobre datos alternativos, para la toma de decisiones.

En el siguiente capítulo se llevará a cabo un abordaje sobre la gobernanza de datos. Esto permitirá comprender la necesidad de coordinación y gestión para poder generar una sinergia con el área de negocio, frente a la incorporación de datos tan diversos. En particular se abordará el tratamiento de la privacidad de datos, considerando que, al

utilizar datos provenientes de redes sociales, podría incurrirse en violar normas y regulaciones existentes sobre la privacidad de datos de los individuos.

Capítulo 3: Gobierno responsable de datos

En un contexto de Big Data, donde la variedad de tipo de datos está a la orden día, no resulta de menor importancia las cuestiones que giran en torno a calidad y gobernanza de datos en las organizaciones. Esto se debe a que la implementación de un modelo adecuado de gobierno de datos permite la generación de valor a partir de los datos, así como la coordinación y gestión de una sinergia con el área de negocio. Adicionalmente, cuando se incorporan datos de tipo personal, debe existir una estrategia de protección de la información para evitar futuros problemas. En este sentido, la manipulación de datos personales conlleva asociado un riesgo que, mediante el diseño de un adecuado plan de gobierno de datos, puede ser controlado.

De esta manera, la noción de calidad de datos se encuadra en la necesidad de contar con un plan de gobierno de datos, bajo el cual la gestión continua de los mismos por parte del negocio evite que aquellos se degraden perdiendo valor para la toma de decisiones. Es por ello por lo que debe mantenerse una estructura lógica sobre los datos centrales para que los mismos sean consistentes, persistentes y útiles (Jones, 2018). A su vez, la falta de calidad en los datos conlleva un riesgo asociado para la toma de decisiones.

Frente a la falta de calidad en los datos, se puede incurrir en una mala toma de decisiones, a partir de haber realizado un análisis sobre datos erróneos. Es así como surge la necesidad de garantizar un nivel de confiabilidad sobre aquellos fijada previamente (Kim, y Cho, 2018). Adicionalmente, no resulta de menor importancia, el hecho de que los datos provenientes de internet pueden contener datos distorsionados, lo que suma una complejidad adicional a la hora de establecer un criterio de calidad sobre los mismos.

Otro riesgo que surge a la hora de trabajar con datos, en particular con datos personales, es acerca de la privacidad. Esta puede ser abordada desde dos enfoques, un enfoque ético y un enfoque normativo. Desde una perspectiva ética, refiere a que se trata de evaluar los métodos de manipulación de datos en un contexto de Big Data en la medida que pueden afectar la privacidad de los individuos. Tal evaluación se debe fundamentar

en los valores humanos sobre el buen uso, la justicia, la autonomía y la confianza (Steinmann, Matei y Collmann, 2016).

Desde un enfoque normativo, se hace referencia al hecho de que deberá tenerse en cuenta la normativa vigente en cada país a la hora de diseñar un sistema de gobierno de datos, independientemente de la normativa interna que pueda tener cada organización en particular. En el caso particular de Argentina existe la ley 25.326 de Protección de Datos la cual establece, entre otros, que "Los datos objeto de tratamiento no pueden ser utilizados para finalidades distintas o incompatibles con aquellas que motivaron su obtención.", estableciendo un límite preciso que deberá ser tenido en cuenta.

El capítulo se estructura de la siguiente manera: en el siguiente apartado, se explica cómo un sistema de gobierno de datos contribuye a la generación de valor en línea con las necesidades del negocio. A su vez, como este sistema contribuye a la reducción del riesgo asociado que conlleva el uso de datos personales. En el siguiente apartado, se realiza un abordaje ético acerca del tratamiento de la privacidad de datos personales. Finalmente, en el tercer apartado, se expondrá un enfoque normativo general sobre datos personales, teniendo en cuenta la legislación existente en Argentina sobre este punto. Así, con este capítulo, y en función de lo desarrollado en los dos capítulos previos, se podrá completar una visión integral del tratamiento de datos.

3.1 Gobernanza de datos

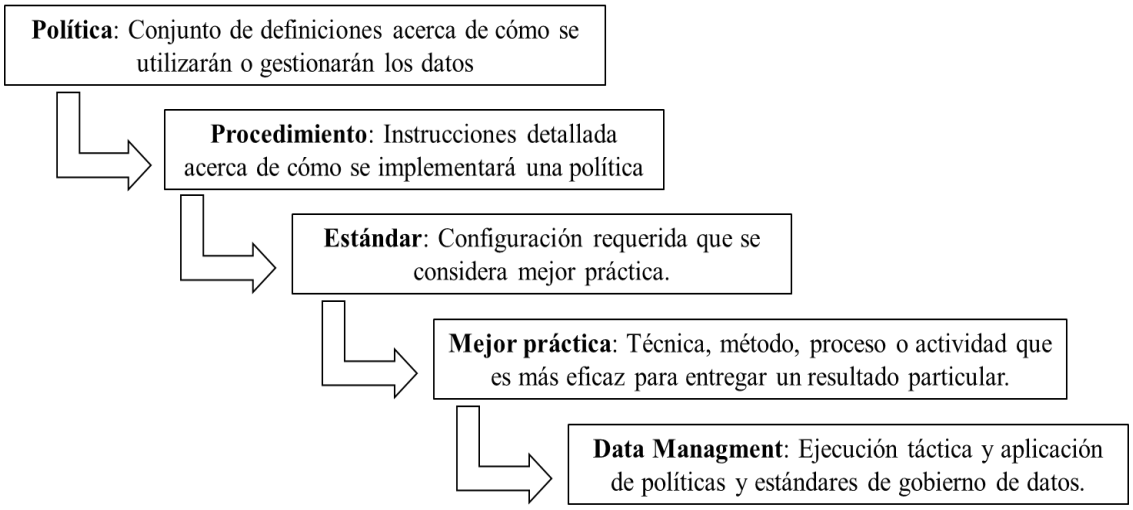
Dado el gran volumen de datos disponibles, es cada vez es más evidente que los problemas en el manejo de la información afectan la toma de decisiones en las organizaciones, frente a la inexistencia de procesos y políticas que permitan garantizar la confiabilidad en los datos. Es entonces como el tratamiento de grandes volúmenes de datos a puesto de manifiesto que en un plan de gobierno de datos ya no es suficiente solo garantizar la calidad de los datos, sino que, resulta necesario que los datos estén disponibles en tiempo adecuado, sean confiables, significativos y suficientes (Kim y Cho, 2018).

Un sistema de gobierno de datos operativo permite establecer un marco para la transparencia y usabilidad de los datos, a partir de determinar las estrategias de

recolección de datos, procesos de integración de datos y, finalmente, la gestión de la información obtenida de los datos. Permite llevar adelante una organización sistemática de pensamientos y comunicaciones sobre conceptos complejos y ambiguos dentro de una organización (Martínez, 2012). La necesidad de tal marco surge entonces como consecuencia de que los datos a menudo no son verificados, son redundantes, incompletos y están peligrosamente desactualizados. En respuesta a esta problemática una puesta en práctica de un gobierno de datos surge como necesaria para controlar los datos disponibles.

En definitiva, el principal objetivo de un sistema de gobierno de datos es la creación de nuevo valor a partir de los datos en línea con los objetivos del negocio. Para lograr este objetivo, el sistema debe contar con cuatro aspectos principales en su diseño: estructura, responsabilidad de vigilancia, talento y cultura e infraestructura (Deloitte, 2011). Por estructura se hace referencia al establecimiento de reglamentos claros que permitan establecer un diseño de gobierno adecuado al modelo de negocio. En cuanto la funcionalidad de vigilancia refiere a delinear las políticas que determina la junta directiva para especificar las funcionalidades administrativas que hacen a las prácticas del negocio de la organización. En cuanto al talento y la cultura, busca alinear los principios del negocio con las creencias centrales para establecer una cultura de trabajo. Y, por último, en cuanto a la infraestructura, se hace referencia al establecimiento manuales de políticas y procedimientos que permitan alinear la tecnología con el sistema de gobierno de datos.

Figura 12: Esquema conceptual de un Gobierno de datos.



Elaboración propia.

En este sentido, un sistema de gobierno de datos podría decirse que posee tres pilares fundamentales: las personas, los procesos y la tecnología. A partir de ello, se establece un ciclo funcional que comienza por la identificación de desafíos, con el fin de establecer el impacto sobre el negocio que permita definir prioridades. En una segunda etapa, se desarrollan políticas y procedimientos alineados con las necesidades y objetivos del negocio. Luego se procede con la ejecución de tales políticas con el fin de lograr los objetivos. Y, finalmente, se realiza monitoreo y medición de resultados. Este ciclo, es de naturaleza continua en función de la dinámica del ciclo de vida de los datos para la toma de decisiones.

Es así como, bajo un sistema de gobierno de datos, el dato se convierte en un activo, el cual debe ser gestionado formalmente en toda la organización. Ello implica que las personas asuman responsabilidad frente a cualquier situación adversa en la calidad de los datos, al mismo tiempo que brinda confiabilidad sobre los mismos. En este sentido, la tecnología colabora en el proceso de forma tal que ayuda a manejar la información para que pueda ser utilizada por toda la organización. De esta manera, busca garantizar la integridad y calidad de los datos con el fin de generar nuevo conocimiento al mismo tiempo que resulta un desafío para toda la organización.

A su vez, busca prevenir efectos secundarios como ser la fuga de información privada o la violación en la privacidad de la información (Kim y Cho, 2018). En particular, se entiende por datos personales o privados a cualquier elemento que pueda llevar a la identificación de un sujeto determinado, lo cual no se limita solo a nombres y apellidos. De este modo, cuando se incorporan datos de tipo personal, debe existir una estrategia de protección de la información para evitar futuros problemas. En este sentido, la manipulación de datos personales conlleva asociado un riesgo que, mediante el diseño de un adecuado plan de gobierno de datos, puede ser controlado.

3.2 Enfoque ético sobre privacidad de datos

En ocasiones, mediante la incorporación de datos personales, se puede crear información sin el consentimiento de las personas, aunque la intención de su uso no sea mal intencionado. En tales circunstancias, los riesgos asociados pueden derivarse de cuatro acciones específicas: reciclar, reutilizar, recombinar y reanalizar (Steinmann,

Matei y Collmann, 2016). La conjunción de estas cuatro acciones puede llevar a la generación de nueva información que excede al objetivo primario para el cual el individuo cedió sus datos. Por esta razón, la gobernanza de datos requiere de un marco estratégico para la transparencia y el uso general de los datos. En dicho marco, deben definirse los accesos, controles y responsabilidad sobre las personas que accedan a tales datos.

A partir de ello, y desde un punto de vista ético, la manipulación de datos personales puede entrar en conflicto con la libertad de expresión, reunión y manifestación o a la presunción de inocencia; en valores como la confianza y la cohesión social; y en procesos humanos importantes como el desarrollo de la identidad (Buenadicha, Galdon, Hermosilla, Loewe, y Pombo, 2019). En este sentido, un correcto sistema de gobierno de datos puede surtir un efecto positivo en la medida que establezca claras condiciones sobre la seguridad de la información y su almacenamiento. Prácticas como implementación de antivirus, el cifrado de datos, la anonimización y la codificación de datos suelen ser modalidades de protección de los datos personales que resultan efectivas.

Por esta razón, el marco estratégico de un gobierno responsable de datos debe establecer que se debe hacer un buen uso de los datos para lograr el objetivo deseado. Esto es, no debe hacerse uso abusivo en función de un interés particular a partir de utilizar los datos privados de los individuos. En segundo lugar, debe hacerse un uso justo de los datos, evitando cometer cualquier acto discriminatorio ya que puede tener un impacto social. En tercer lugar, se debe respetar la autonomía de los individuos. En este sentido, si el uso de los datos tuviera un fin diferente al original, debería consultarle al usuario si está dispuesto a facilitar sus datos a tal fin. Por último, debe establecer un principio de confianza entre el usuario y la organización a la hora de utilizar sus datos para que el mismos pueda tener autonomía de decisión y protección.

Un aspecto adicional para tener en cuenta en un contexto de implementación de técnicas de Data Mining, aborda la noción de discriminación algorítmica. La discriminación es un tratamiento perjudicial que se le da a una persona a través de categorizaciones arbitrarias. La implementación automática puede proliferar aquella a través de los sistemas informáticos. Dado que una implementación de este tipo no hace mas que reproducir la discriminación que sucede en la realidad, es entonces que debe tenerse en cuenta a la hora de llevar a cabo tal metodología. Es por ello por lo que, un

buen sistema de gobierno de datos debe establecer un marco de responsabilidad acerca de los criterios a implementarse por parte de aquellos que tiene la responsabilidad de llevar a adelante cualquier implementación de este tipo.

A su vez, lo anterior conlleva un riesgo asociado en torno a la falta de transparencia de la información. En general, los sistemas informáticos son percibidos por la mayoría de los individuos como cajas negras (Buenadicha, Galdon, Hermosilla, Loewe, y Pombo, 2019), es decir, como mecanismos incomprensibles. Dado que no se puede pretender que todos los individuos tengan el suficiente conocimiento respecto de la funcionalidad de las diferentes metodologías, se propone establecer un sistema que este fundamentado en la transparencia de la información. En este sentido, un buen sistema de gobierno de datos debe tener entre sus fundamentos que la información personal debe ser protegida a la vez que accesible para quien solicite la misma. Así mismo debe garantizar a través de sus políticas que los datos solicitados deben dejar en claro la utilización que se hará de ellos, con el fin de obtener un consentimiento claro por parte de quien los otorga. De esta manera, se logrará evitar una crisis de confianza entre el usuario o cliente y quien manipula su información.

3.3 Enfoque normativo sobre privacidad de datos

Dado todos los riesgos asociados a la manipulación de información personal, en función de tener en cuenta los aspectos éticos mencionados en el apartado anterior, es que ha surgido diversa regulación en los diferentes países. Un estándar de referencia global es el Reglamento General de Protección de Datos (RGPD) de la Unión Europea, con entrada en vigor en el año 2018 (Buenadicha, Galdon, Hermosilla, Loewe, y Pombo, 2019). Este reglamento establece seis principios básicos sobre el manejo de datos personales:

(i) deben ser tratados de forma lícita, leal y transparente;

(ii) se deben recolectar con fines determinados explícitos y legítimos;

(iii) deben ser adecuados, pertinentes y limitados a lo necesario dependiendo del uso;

(iv) deben ser exactos y estar siempre actualizados;

(v) deben mantenerse de forma tal que se permita la identificación de los interesados durante no más tiempo del necesario para los fines del tratamiento; y

(vi) deben ser tratados de tal manera que se garantice su seguridad.

Sin entrar en mayores detalles sobre el RGDP, uno de los elementos esenciales que se establece en este reglamento es que el consentimiento es la base de la gestión de datos personales. Esto implica que, además de las cuestiones entorno a la seguridad, quienes recolecten y gestionen los datos deberán asegurarse siempre de haber informado a sus propietarios (los individuos) y obtener su consentimiento tantas veces como sea necesario si la finalidad del uso que se le dará a sus datos cambia. Además de respetar los derechos de los individuos, esta noción se constituye en un eje fundamental del contrato social que permite generar confianza entre quienes proporcionan los datos y quienes los manejan.

En esta misma línea, la Organización para la Cooperación y el Desarrollo Económicos (OCDE), de la cual forma parte Argentina, establece unos principios generales, vigentes desde el año 2013, sobre la protección de datos personales, que constituyen un marco general para delinear unas políticas claras sobre el resguardo de los datos personales. En este sentido postula que se debe establecer límites claros para la obtención de los datos, así como determinar la relevancia de los datos para el uso previsto. En cuanto a la recolección, establece definir con claridad el uso que se dará a los datos antes de solicitarlos. Luego, propone abstenerse de utilizar los datos para usos distintos al determinado originalmente sin el consentimiento de las personas afectadas. En cuanto a la protección de datos, establece asegurarse de proteger los datos contra el acceso ilícito o piratería. Respecto de la transparencia de la información se propone asegurar que los avances, prácticas y políticas sobre el uso de los datos sean abiertos y transparentes. Finalmente, establece garantizar que las personas cuyos datos se han recolectado tengan acceso a los mismos y puedan solicitar modificaciones o su eliminación definitiva.

También deberá tenerse en cuenta la normativa vigente en cada país a la hora de diseñar un sistema de gobierno de datos. En el caso particular de Argentina existe la ley 25.326 de Protección de Datos Personales en cuyo objeto establece “la protección integral de los datos personales asentados en archivos, registros, bancos de datos, u otros medios técnicos de tratamiento de datos, sean éstos públicos, o privados destinados a dar informes, para garantizar el derecho al honor y a la intimidad de las personas, así como también el acceso a la información que sobre las mismas se registre, de conformidad a lo establecido en el artículo 43, párrafo tercero de la Constitución Nacional.” (Art. 1°).

Particularmente, por tratamiento de datos refiere a “Operaciones y procedimientos sistemáticos, electrónicos o no, que permitan la recolección, conservación, ordenación, almacenamiento, modificación, relacionamiento, evaluación, bloqueo, destrucción, y en general el procesamiento de datos personales, así como también su cesión a terceros a través de comunicaciones, consultas, interconexiones o transferencias.” Y en particular, para el caso de datos electrónicos se los define como “datos personales sometidos al tratamiento o procesamiento electrónico o automatizado”. (Art. 2°). Por último, se considera de vital importancia lo establecido en el artículo 4° “Los datos objeto de tratamiento no pueden ser utilizados para finalidades distintas o incompatibles con aquellas que motivaron su obtención.”

A modo de conclusión, en este tercer capítulo se pone de relieve la necesidad de establecer un plan de gobierno de datos cuyo principal objetivo es garantizar la creación de nuevo valor a partir de los datos, siempre en línea con los objetivos del negocio. El mismo deberá preservar el nivel de calidad de los datos y su integridad, así como la definición de responsabilidad sobre estos. Para ello deberá considerarse la protección de la información personal, como estrategia clave del sistema ya que, un mal diseño estratégico puede derivar en la pérdida de confianza en la organización por parte de los clientes y, en consecuencia, en la pérdida de estos. De esta manera, la gestión en torno de la privacidad de datos se convierte en un punto central que deberá ser diseñada bajo consideraciones éticas y dentro del contexto de la normativa vigente.

Conclusiones

A lo largo del presente trabajo se ha demostrado cómo llevar a cabo el procesamiento de un volumen considerable de tweets, y la implementación de técnicas de Data Mining, que permitieron la detección de tópicos o temas asociándolos a diferentes problemáticas a las que hacen referencia los usuarios de la tarjeta SUBE en la red social Twitter. Al mismo tiempo, se logra dar cuenta de que la incorporación y tratamiento de este tipo de datos, en un contexto de grandes volúmenes de datos, requiere de la implementación de un modelo de gobierno y gestión de datos para que finalmente los resultados obtenidos se conviertan en un verdadero valor agregado para la toma de decisiones.

En el capítulo uno se pone de relieve que los datos alternativos desafían a las organizaciones a considerar adaptaciones de estructuras tradicionales en el tratamiento de los datos frente a la oportunidad de obtener ventajas competitivas, generando un valor agregado en la producción de información para la toma de decisiones. Para lograr con éxito esta tarea, se debe tener en cuenta que el proceso de obtención de datos de calidad debe estar sustentado en un proceso de análisis, implementación y control continuo. Esto debido a que los datos, en un contexto de grandes volúmenes con gran diversidad, dejan de ser estáticos para ser dinámicos en la medida que la disponibilidad digital de los mismos incrementa el volumen, la velocidad y la variedad con que los mismos se generan. En este sentido, la automatización de procesos puede contribuir de manera eficaz para el ejercicio de detección de nuevos valores en los datos, otorgando la posibilidad de adoptar cambios estratégicos de manera inmediata.

En el capítulo dos mediante técnicas de Data Mining se estructuraron datos provenientes de tweets realizados por usuarios de la Tarjeta SUBE, con el objetivo de poder aplicar diferentes metodologías de Text Mining que permitieron arribar a un resultado. Como resultado final se pudo clasificar los tweets en función de su sentimiento contenido, detectando que el 50% de los mismos expresan alguna disconformidad o queja. Dado este escenario, se identificaron diferentes problemáticas asociadas al uso de la tarjeta SUBE. Tales problemáticas identificadas se relacionan con dificultades en la realización de recargas de créditos, utilización de la App de SUBE y obtención de

descuentos a la hora de utilizar la tarjeta SUBE, siendo que el individuo cuenta con algún beneficio. De esta manera se logró demostrar que la utilización de técnicas de Data Mining permiten la generación de valor agregado a partir de obtener información sobre datos alternativos, para la toma de decisiones.

Finalmente, en el capítulo tres se pone de relieve la necesidad de establecer un plan de gobierno de datos cuyo principal objetivo es garantizar la creación de nuevo valor a partir de los datos, siempre en línea con los objetivos del negocio. El mismo deberá preservar el nivel de calidad de los datos y su integridad, así como la definición de responsabilidad sobre estos. Para ello deberá considerarse la protección de la información personal, como estrategia clave del sistema ya que, un mal diseño estratégico puede derivar en la pérdida de confianza en la organización por parte de los clientes y, en consecuencia, en la pérdida de estos. De esta manera, la gestión en torno de la privacidad de datos se convierte en un punto central que deberá ser diseñada bajo consideraciones éticas y dentro del contexto de la normativa vigente.

De esta manera, se considera que el presente trabajo logra sentar las bases de un aporte para las distintas organizaciones que se encuentren interesadas en llevar adelante la incorporación de datos alternativos con el objetivo de mejorar su proceso de toma de decisiones. En este sentido, los desafíos pueden resultar diversos, implicando desde un cambio cultural en la organización hasta un mayor requerimiento de especialistas en los diferentes eslabones de la cadena de implementación. No obstante, si el interés es poder generar un mayor valor agregado para optimizar las decisiones a tomar, se considera que puede ser un buen comienzo en la búsqueda de tal objetivo.

Pero no solo las organizaciones del sector privado pueden encontrar un interés en el presenta trabajo. Para el caso de las organizaciones públicas, donde el estado debe asumir responsabilidades superiores en el tratamiento de datos, la implementación de un modelo de gobierno de datos basados en normativas vigentes de manera global, pueden resultar un desafío a resolver. Son cada vez más los gobiernos de los diferentes países que están llevando una agenda en torno al tratamiento de la privacidad de datos personales en un contexto de Big Data para garantizar la implementación de modelos para la explotación de información con el fin de llevar adelante mejores políticas públicas.

En cuanto al ámbito académico, se considera que el presente trabajo logra cumplir con los lineamientos generales que fueron expuestos a lo largo de los distintos contenidos de las materias incluidas en la Especialización en Métodos Cuantitativos para la Gestión y Análisis de Datos en Organizaciones. En este sentido, puede ser considerado como un ejemplo de generación de conocimiento, tendiendo un puente entre el sector académico y el sector público o privado.

Por otra parte, resulta de interés llevar adelante una ampliación de este trabajo, que permita realizar una implementación más profunda de otros métodos de Text Mining que sean complementarios o innovadores para ampliar la obtención de resultados, así como del conocimiento. En este sentido, la búsqueda de predicción a partir de llevar a cabo una implementación de Machine Learning, podría ser un objetivo interesante, en la medida que permitiría establecer un modelo automático de detección de reclamos.

También, la incorporación de otras fuentes de datos alternativas, podría ser un desafío para tener en cuenta. Por ejemplo, la incorporación de otras redes sociales, como Facebook o Instagram, podrían aportar nueva información para llegar a nuevos resultados. O así mismos, podría ser de interés el estudio de casos similares de forma tal de contar resultados que lleven a potenciar un nuevo desarrollo de metodologías a implementar.

Referencias

Alpaydin, E. (2009). Introduction to machine learning. MIT press.

Alvares, D., Armero, C., & Forte, A. (2018). What Does Objective Mean in a Dirichlet-multinomial Process? International Statistical Review, 86(1), 106-118.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Buenadicha, C., Galdon, G., Hermosilla, M. P., Loewe, D., & Pombo, C. (2019). La Gestión Ética de los Datos. Por qué importa y cómo hacer un uso justo de los datos en un mundo digital BID, editor.

Chen, H. M., & Franks, P. C. (2016). Exploring Government Uses of Social Media through Twitter Sentiment Analysis. Journal of Digital Information Management, 14(5).

Cleven, A., & Wortmann, F. (2010, January). Uncovering four strategies to approach master data management. In 2010 43rd Hawaii International Conference on System Sciences (pp. 1-10). IEEE.

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

Eberendu, A. C. (2016). Unstructured Data: an overview of the data of Big Data. International Journal of Computer Trends and Technology, 38(1), 46-50.

Fiore, V., Almodovar, K., Assoumou, A., Dutta, D., & Cotoranu, A. (2017). The Correlation between the Topic and Emotion of Tweets through Machine Learning. Seidenberg School of CSIS. Pace University, Pleasantville, New York

Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

Jones, D. T. (2018). Data Governance Framework Implementation Plan. Philadelphia: DBHIDS.

Kabir, A. I., Karim, R., Newaz, S., & Hossain, M. I. (2018). The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R. Informatica Economica, 22(1).

Kao, A., & Poteet, S. R. (Eds.). (2007). Natural language processing and text mining. Springer Science & Business Media.

Kim, H. Y., & Cho, J. S. (2018). Data governance framework for big data implementation with NPS Case Analysis in Korea. Journal of Business and Retail Management Research, 12(3).

Kolanovic, M., & Krishnamachari, R. T. (2017). Big data and AI strategies: Machine learning and alternative data approach to investing. JP Morgan Global Quantitative & Derivatives Strategy Report.

Levy Abitbol, J., Fleury, E., & Karsai, M. (2019). Optimal Proxy Selection for Socioeconomic Status Inference on Twitter. Complexity, 2019.

Martínez, J. (2012). Seis pasos para el Gobierno de Datos. ¿Qué es y cómo se implementa un programa de Gobierno de Datos? IBM, Developer Works.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

McKeen, J. D., & Smith, H. A. (2007). Developments in practice XXIV: information management: the nexus of business and IT. Communications of the Association for Information Systems, 19(1), 3.

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). Automated data collection with R: A practical guide to web scraping and text mining. John Wiley & Sons.

Preoţiuc-Pietro, D., Lampos, V., & Aletras, N. (2015, July). An analysis of the user occupational class through Twitter content. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1754-1764).

Russom, P. (2015). Modernización de la integración de datos para dar cabida a los nuevos requisitos de negocio y Big Data. TDWI, tdwi.org.

Schmarzo, B. (2013). Big Data: Understanding how data powers big business. John Wiley & Sons.

Steinmann, M., Matei, S. A., & Collmann, J. (2016). A theoretical framework for ethical reflection in big data research. In Ethical Reasoning in Big Data (pp. 11-27). Springer, Cham.

Verspoor, Karin & Cohen, Kevin. (2013). Natural Language Processing. 10.1007/978-1-4419-9863-7_158.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. Communication Methods and Measures, 11(4), 245-265.

SOFTWARE: RSTUDIO

R Score Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/,

RStudio (Febrero 2011), AGPL v3, Northern Ave, Boston, https://www.rstudio.com/

PAQUETES UTILIZADOS:

Package ‘broom’ (Abril 2019) “Convert Statistical Analysis Objects into Tidy Tibbles” David Robinson [aut] y otros. http://github.com/tidyverse/broom

Package ‘dplyr’ (Mayo 2008), “A Grammar of Data Manipulation”, MIT + file LICENSE, http://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr

Package ‘factoextra’ (Agosto 2017) “Extract and Visualize the Results of Multivariate Data Analyses” Alboukadel Kassambara [aut, cre], Fabian Mundt [aut] http://www.sthda.com/english/rpkgs/factoextra

Package ‘ggplot2’ (Agosto 2019) “Create Elegant Data Visualisations Using the Grammar of Graphics” r Hadley Wickham [aut, cre], Winston Chang [aut] y otros. http://ggplot2.tidyverse.org, https://github.com/tidyverse/ggplot2

Package ‘gridGraphics’ (Mayo 2019) “Redraw Base Graphics Using 'grid' Graphics” Paul Murrell [cre, aut], Zhijian Wen [aut] https://github.com/pmur002/gridgraphics

Package ‘gridExtra’ (Septiembre 2017) “Miscellaneous Functions for ``Grid'' Graphics”, Baptiste Auguie [aut, cre], Anton Antonov [ctb]

Package ‘lubridate’ (Abril 2018) “Make Dealing with Dates a Little Easier” Vitalie Spinu [aut, cre], Garrett Grolemund [aut], Hadley Wickham [aut], Ian Lyttle [ctb], Imanuel Constigan [ctb], Jason Law [ctb], Doug Mitarotonda [ctb], Joseph Larmarange [ctb], Jonathan Boiser [ctb], Chel Hee Lee [ctb] http://lubridate.tidyverse.org

Package ‘RColorBrewer’(Febrero 2015) “ColorBrewer Palettes” Erich Neuwirth [aut, cre]

Package ‘scales’ (Agosto 2018) “Scale Functions for Visualization” Hadley Wickham [aut, cre], RStudio [cph] https://scales.r-lib.org, https://github.com/r-lib/scales

Package ‘stringr’ (Febrero, 2019), “Simple, Consistent Wrappers for Common String Operations”, Hadley Wickham [aut, cre, cph], RStudio [cph, fnd], http://stringr.tidyverse.org, https://github.com/tidyverse/stringr

Package ‘tidyr’ (Septiembre 2019) “Tidy Messy Data” MIT + file LICENSE URL https://tidyr.tidyverse.org

Package ‘tidytext’ (Octubre 2018), “Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools”, MIT + file LICENSE, http://github.com/juliasilge/tidytext

Package ‘tidyverse’ (Noviembre 2017) “ Easily Install and Load the 'Tidyverse'” Hadley Wickham [aut, cre], RStudio [cph, fnd] http://tidyverse.tidyverse.org

Package ‘tm’ (Diciembre 2018) “Text Mining Package” Ingo Feinerer [aut, cre], Kurt Hornik [aut], Artifex Software, Inc. [ctb, cph] http://tm.r-forge.r-project.org/

Package ‘tokenizers’(Marzo 2018) “Fast, Consistent Tokenization of Natural Language Text” MIT + file LICENSE, https://lincolnmullen.com/software/tokenizers/

Package ‘twitteR’ (Agosto 2016), “Provides an interface to the Twitter web API”, Jeff Gentry, http://lists.hexdump.org/listinfo.cgi/twitter-users-hexdump.org

Package ‘topicmodels’ (Diciembre 2018) “Topic Models” Bettina Grün [aut, cre], Kurt Hornik [aut]

Package ‘widyr’ (Septiembre 2019) “Widen, Process, then Re-Tidy Data” David Robinson [aut, cre], Kanishka Misra [ctb] http://github.com/dgrtwo/widyr

Package ‘wordcloud’ (Agosto 2018), “Word Clouds”, Ian Fellows, http://blog.fellstat.com/?cat=11 http://www.fellstat.com