CMA IADCOM - UBA CENTRO DE INVESTIGACIÓN EN METODOS CUANTITATIVOS APLICADOS A LA ECONOMÍA Y LA GESTIÓN

Universidad de Buenos Aires Facultad de Ciencias Económicas

LATINDEX
http://www.latindex.org/latindex/ficha?folio=2751

http://www.economicas.uba.ar/institutos_y_centros/rimf/

EL APRENDIZAJE AUTOMÁTICO UTILIZADO EN LOS PROCESOS DE INNOVACIÓN DE LAS ORGANIZACIONES[1]

Melisa Elfenbaum

Instituto Nacional de Servicios Sociales para Jubilados y Pensionados (INSSJP). Gerencia de Sistemas. Departamento de Diseño y Desarrollo de Sistemas - Prestaciones. Paraná 468, CABA, C1017AAJ. República Argentina.

melisaelfenbaum@gmail.com

Resumen

Recibido: 03/2019

Aceptado: 09/2019

Palabras clave

Aprendizaje automático.

Big data.

Organizaciones.

En los últimos años, el crecimiento sostenido de acceso a la información está generando nuevos desafíos y necesidades en el ámbito organizacional y en el de las políticas públicas y regulaciones. En este marco, el término *big data* –datos masivos o grandes datos- se utiliza para representar a los activos de información caracterizados por un gran volumen, velocidad y variedad. Frente a estos nuevos retos, y considerando que la cantidad de datos seguirá en aumento, es necesario desarrollar herramientas que permitan realizar la recolección, análisis y predicción de datos con el fin de que las organizaciones puedan crear valor en formas que anteriormente no existían. En tal sentido, surge el aprendizaje automático o *machine learning* como respuesta al fenómeno de *big data*, el cual permite detectar automáticamente patrones utilizando datos de ejemplo o experiencia pasada.

Con el objeto de realizar una aplicación innovadora de los procesos de gestión de la información en la era del *big data*, en el presente trabajo se realiza una implementación de un método de aprendizaje automático con datos obtenidos de la red social Twitter, que les permita a las organizaciones crear valor y obtener una ventaja competitiva en el proceso de desarrollo e innovación de nuevos productos.

ISSN: 2250-687X - ISSN (En línea): 2250-6861

[1] Los conceptos y opiniones contenidos en este trabajo son de exclusiva responsabilidad de la autora.

MACHINE LEARNING USED IN ORGANIZATION INNOVATION PROCESSES

Melisa Elfenbaum

Instituto Nacional de Servicios Sociales para Jubilados y Pensionados (INSSJP). Gerencia de Sistemas. Departamento de Diseño y Desarrollo de Sistemas - Prestaciones. Paraná 468, CABA, C1017AAJ. República Argentina.

melisaelfenbaum@gmail.com

Abstract

KEYWORDS

Machine learning.

Big data.

Organizations.

In the last years, the sustained growth of access to information is generating new challenges and needs in the organizational field and in the public policies and regulations. In this framework, the term big data is used to represent information assets characterized by a large volume, speed and variety. Facing these new challenges, and considering that the amount of data will continue to increase, it is necessary to develop tools that allow the collection, analysis and prediction of data in order that organizations can create value in ways that previously did not exist. In this sense, machine learning emerges as a response to the phenomenon of big data, which allows the automatic detection of patterns using sample data or experience.

In order to make an innovative application of information management processes in the era of big data, in this paper is performed an implementation of a machine learning method with data obtained from the social network Twitter, which allows organizations to create value and obtain a competitive advantage in the process of development and innovation of new products.

ISSN: 2250-687X - ISSN (En línea): 2250-6861

INTRODUCCIÓN

En el entorno competitivo en que se desarrollan las organizaciones actualmente, el *big data* ha asumido un papel central en la transformación de los procesos comerciales clave y la creación de nuevas oportunidades de negocio, a partir de la transparencia, segmentaciones altamente específicas de clientes, creación de nuevos productos y servicios innovadores, entre otros (Manyika et al., 2011).

Debido a la importancia que tiene el *big data* en la actualidad -definido como datos grandes, diversos, complejos y longitudinales generados por instrumentos, sensores, transacciones en internet, e-mails, videos y todos las fuentes digitales disponibles en la actualidad y en el futuro- (Collmann, FitzGerald, Wu, Kupersmith, & Matei, 2016), las organizaciones requieren nuevas arquitecturas, algoritmos y técnicas de análisis para gestionar los datos, extraer el conocimiento y generar valor a través de la toma decisiones más inteligentes y oportunas no sólo a partir de su propia información, sino también de las crecientes fuentes externas a la organización, como son los datos meteorológicos, datos demográficos de los clientes, datos de crédito al consumidor de las agencias de crédito y datos de sitios de medios sociales (Casonato, Lapkin, Beyer, Genovese, & Friedman, 2011; Schmarzo, 2013).

En tal sentido, si bien el avance de la tecnología puede permitir que la economía mundial almacene y procese cantidades cada vez mayores de datos, existen límites para la capacidad humana para procesar los grandes volúmenes de datos (Manyika et al., 2011). Es por ello que surge el aprendizaje automático –también conocido comúnmente por su denominación en inglés, *machine learning*-, que permite detectar automáticamente patrones para predecir datos futuros o realizar toma de decisiones bajo incertidumbre, y se ha convertido en parte central en el funcionamiento de las organizaciones en los últimos años ya que se puede aplicar en cualquier campo que necesite interpretar datos (Murphy, 2012).

En este marco, el trabajo presenta un caso práctico aplicado en el sector de la telefonía celular, en el cual se implementa un método de aprendizaje automático no supervisado en combinación con técnicas de *text mining* sobre los datos obtenidos de la red social Twitter con el objetivo de extraer las características que deberían tener los nuevos teléfonos celulares de acuerdo a las opiniones de personas de todo el mundo.

1. EL APRENDIZAJE AUTOMÁTICO EN LA ACTUALIDAD

El aprendizaje automático es definido como un conjunto de métodos que permiten detectar automáticamente patrones tan generales como sea posible sin intervención o asistencia humana utilizando datos de ejemplo o experiencia pasada, que deben ser significativos ya que proporcionan información, permiten predecir datos futuros o realizar toma de decisiones rápida y precisa bajo incertidumbre que conducen a alguna ventaja (Murphy, 2012; Witten, Frank, Hall, & Pal, 2016).

Entre las aplicaciones más utilizadas del aprendizaje automático se destacan la clasificación de páginas web, el filtrado de correo electrónico no deseado, la clasificación de imágenes, el reconocimiento de escritura a mano y de rostros, la detección de fraudes, la traducción automática de documentos, la segmentación de clientes y el análisis de canasta de mercado, el procesamiento del lenguaje natural, problemas de regresión como la predicción de valores de acciones o de temperatura, los sistemas de recomendación –como por ejemplo el de Netflix o Amazon-, entre otros (Murphy, 2012).

Más allá de la diversidad de algoritmos desarrollados actualmente, cualquier problema de *machine learning* se puede clasificar un uno de aprendizaje supervisado o no supervisado. Mientras que en el aprendizaje supervisado o predictivo se cuenta con un conjunto de ejemplos cuya respuesta es conocida, y lo que se necesita realizar es una regla que permita aproximar la respuesta para todos los datos que se presenten, en el aprendizaje no supervisado o descriptivo no se cuenta con un conjunto de ejemplos cuya respuesta es conocida, por lo tanto, los problemas se abordan con poca o ninguna idea de cómo deben ser los resultados. En otras palabras, en el aprendizaje supervisado para cada observación $x^{(i)}, i = 1, \dots, n$ existe una medida de respuesta asociada $y^{(i)}$, por lo que se desea ajustar un modelo que relacione la respuesta a los predictores. Por el contrario, el aprendizaje no supervisado describe una situación en la que para cada observación $i = 1, \dots, n$, se obtiene un vector de medidas $x^{(i)}$ pero ninguna respuesta asociada $y^{(i)}$ (James, Witten, Hastie, & Tibshirani, 2013).

Los métodos de aprendizaje supervisado principales son los de regresión -cuando la variable que se intenta predecir puede tomar valores continuos- y los de clasificación - si la respuesta puede tomar solamente valores discretos-. Mientras que los algoritmos principales de aprendizaje no supervisado son los de agrupamiento –en los que se agrupan los datos de acuerdo a las relaciones entre las variables- y el descubrimiento de anomalías o de estructuras en un ambiente caótico.

2. EL MODELO LATENT DIRICHLET ALLOCATION

El método *Latent Dirichlet Allocation* (LDA) es un algoritmo de aprendizaje automático no supervisado de agrupamiento –particularmente de modelado de temas-, el cual permite organizar y resumir documentos digitales a una escala que sería imposible a través de la inspección humana, a partir del descubrimiento de temas que se encuentran ocultos en una gran colección de documentos no estructurados, y asimismo, tiene la ventaja de no requerir ningún etiquetado previo, sino que los temas surgen del análisis de los textos (Blei, 2012).

El modelo LDA se utiliza principalmente para extraer temas de una colección de documentos, con el cual se trata cada documento como una mezcla aleatoria de temas, y cada tema como una mezcla de palabras, lo que permite que los documentos se superpongan entre sí en términos de contenido, en lugar de separarse en grupos discretos, y a diferencia de los métodos de agrupamiento en donde se asigna un tema a cada documento, con este método se asigna un conjunto de etiquetas –o temas- ponderadas a cada uno de ellos (Blei, Ng, & Jordan, 2003).
El modelo se denomina "latente" debido a que solo se observan documentos y palabras, mientras que los temas no se pueden observar directamente, sino que son parte de la estructura oculta o latente de los documentos. El objetivo del método LDA es inferir la estructura del tema latente dadas las palabras y el documento, a partir de la recreación de los documentos en el corpus ajustando la importancia relativa de los temas en los documentos y las palabras en los temas de forma iterativa.

Formalmente, si cada palabra w se define como un elemento de un vocabulario, un documento es una secuencia de N palabras denotado por $\mathrm{W} = (w_1 + w_2 + \cdots + w_N)$ y el corpus es la colección de M documentos denotado por $D = (W_1 + W_2 + \cdots + W_M)$, se define el proceso generativo para cada documento W en un corpus D en la ecuación 1, luego de definir la cantidad de temas K -de forma arbitraria, a partir de resultados anteriores o realizando la elección de acuerdo a los resultados obtenidos luego de ejecutar el algoritmo para distintos valores de K- (Blei et al., 2003).

1. $Para\ cada\ documento: \theta \sim Dirichlet(\alpha)$

2. *Repetir i veces* { *Para cada una de las N palabras* w_n *de cada documento*:

$$a.\,Se\ elige\ un\ tema\ z_n \sim Multinomial(z_n|\theta)$$
$$b.\,Se\ elige\ una\ palabra\ w_n \sim Multinomial(w_n|z_n,\beta)\ \} \quad (1)$$

En el paso 1 se asigna aleatoriamente cada palabra de cada documento a uno de los K temas y de esta forma el modelo proporciona tanto las representaciones de temas de todos los documentos como las distribuciones de palabras de todos los temas, es decir que se elige aleatoriamente una distribución sobre temas, por lo que cada documento exhibe los temas en diferentes proporciones (Blei, 2012). Luego, en el paso 2 se realiza un proceso iterativo para cada una de las palabras de cada documento para actualizar las asignaciones de los temas, de acuerdo a la frecuencia de cada palabra en todos los temas y la frecuencia de los temas en el documento. En este paso en primer lugar se calcula para cada tema la proporción de palabras en cada documento que están actualmente asignadas al tema z_n – es decir que se elige aleatoriamente un tema de la distribución sobre los temas en el paso 1-, y en segundo lugar la proporción de asignaciones al tema z_n sobre todos los documentos que provienen de la palabra w_n, es decir que cada palabra en cada documento se extrae de uno de los temas, asumiendo que todas las asignaciones de temas excepto el de la palabra actual son correctas, y luego se actualiza la asignación de la palabra que se está evaluando (Blei, 2012).

La variable aleatoria θ -que representa los temas disponibles por documentos- sigue una distribución de Dirichlet, la variable z_n -tema- una distribución multinomial y w_n -palabra- una multinomial condicionada al tema z_n. La función de densidad de probabilidad de θ se determina en la ecuación 2, donde $\Gamma(x)$ es la función Gamma.

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{\mathrm{i=1}}^{\mathrm{k}}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\dots\theta_k^{\alpha_k-1} \quad (2)$$

Dados los parámetros α y β, la distribución conjunta de una mezcla de temas θ, un conjunto de N temas z, y un conjunto de N palabras w viene dada por la siguiente ecuación:

$$p(\theta,\mathrm{z},\mathrm{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta) \quad (3)$$

Integrando sobre θ y sumando sobre z, se obtiene la distribución marginal de cada documento en la ecuación:

$$p(\mathrm{z}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta \quad (4)$$

El problema clave que se debe resolver en el modelo es inferir la distribución posterior de las variables ocultas dado un documento –representada en la ecuación 5-, donde el numerador es la distribución conjunta de todas las variables aleatorias –definida en la ecuación 3- y el denominador es la probabilidad marginal de las observaciones que en teoría se puede calcular sumando la distribución conjunta en todas las instancias posibles de la estructura de temas oculta, sin embargo, el número de estructuras de temas posibles es exponencialmente grande, por lo que distribución posterior es difícil de computar (Blei, 2012; Blei et al., 2003). Por lo que se utilizan métodos para realizar una aproximación, entre los que se destacan los algoritmos basados en muestreo -que intentan obtener muestras de la parte posterior para aproximarla con una distribución empírica- y los algoritmos variacionales –que colocan una familia parametrizada de distribuciones sobre la

estructura oculta y luego encuentran el miembro de esa familia que está más cerca de la posterior, es decir que el problema de inferencia se transforma en un problema de optimización- (Blei, 2012).

$$p(\theta, \mathrm{z}|\mathrm{w}, \alpha, \beta) = \frac{p(\theta, \mathrm{z}, \mathrm{w}|\alpha, \beta)}{p(\mathrm{w}|\alpha, \beta)} \quad (5)$$

Dentro de los algoritmos basados en muestreo, el más utilizado para el modelado de temas es el muestreo de Gibbs[2], que es una forma de cadena de Markov Monte Carlo fácil de implementar y eficiente para extraer un conjunto de temas de un gran corpus (Blei, 2012; Steyvers & Griffiths, 2007). Entre los métodos de variaciones, se encuentran el algoritmo de inferencia variacional de ascenso coordinado[3] -que utiliza la desigualdad de Jensen para obtener un límite inferior ajustable en la probabilidad- y el modelo variacional online de Bayes[4] -el cual se basa en la optimización estocástica en línea con un paso de gradiente natural- (Blei, 2012; Blei et al., 2003; Hoffman, Bach, & Blei, 2010).

En cuanto a los parámetros α y β –determinados de forma arbitraria o a partir de una estimación- pueden tomar valores entre 0 y 1 donde cuanto mayor es el valor de α, más probable es que cada documento contenga una mezcla de la mayoría de los temas -en lugar de uno solo-, mientras que un valor alto de β denota que cada tema es probable que contenga una mezcla de la mayoría de las palabras y no una palabra específica.

3. EL MODELO LDA APLICADO PARA OPTIMIZAR LA INNOVACIÓN EN LAS ORGANIZACIONES

En la era de *big data,* las organizaciones pueden tener una ventaja competitiva a partir del nuevo conocimiento y nuevos modelos innovadores de negocio para lograr mejorar productos y servicios considerando las expectativas de los clientes – que están cada vez más en el epicentro de la economía-, ya que la creciente transparencia, el compromiso y los nuevos patrones de comportamiento de los consumidores obligan a las organizaciones a adaptar la forma en que diseñan, comercializan y entregan los productos y servicios, lo que genera una mejor correspondencia entre los productos y las necesidades del consumidor (Manyika et al., 2011; Schwab, 2017).

En tal sentido, debido a que el estudio de las necesidades y preferencias de los clientes por parte de las organizaciones es fundamental en el diseño de propuestas innovadoras de nuevos productos, el *microblogging* –siendo Twitter la plataforma más utilizada, con rápido crecimiento y adopción significativa- se ha convertido en una fuente de información relevante acerca de las preferencias y opiniones de las personas. Si bien una gran parte de los mensajes compartidos en Twitter -denominados *tweets*- pueden ser considerados como ruido e inútiles para extraer conocimiento, debido a que los mensajes son cortos y se generan constantemente, se considera una fuente de datos potencialmente valiosa que se puede utilizar para profundizar en los pensamientos y opiniones de millones de personas de cualquier cosa en tiempo casi real, y de esta forma descubrir lo que está sucediendo en cualquier momento, en cualquier parte del mundo (Bifet & Frank, 2010; Brown, 2012).

En este contexto, el análisis de los mensajes de Twitter se enmarca dentro de los problemas de aprendizaje automático no supervisado ya que no se conoce de qué temas pueden tratar los *tweets*

[2]Para una explicación detallada del muestro de Gibbs se pueden consultar: Griffiths (2002); Griffiths & Steyvers (2004); Steyvers & Griffiths (2007).
[3]Se puede ver el método de inferencia variacional de ascenso coordinado aplicado para el modelo LDA en Blei et al. (2003)
[4]En Hoffman, Bach, & Blei (2010) se desarrolla el modelo variacional online de Bayes.

debido a que no se tiene un conjunto de mensajes de los cuales se sepan los temas que tratan que sirvan como set de entrenamiento. Dentro de los algoritmos de aprendizaje automático no supervisado, es conveniente utilizar un algoritmo de modelado de tópicos para poder obtener los diferentes temas contenidos en cada *tweet* en vez de uno solo por cada mensaje, con el objetivo de aprovechar al máximo el análisis de las opiniones de las personas acerca de un producto específico. Entre los algoritmos de modelado de temas, se elige el método *Latent Dirichlet Allocation* – el cual se construye sobre la base del modelo *Latent Semantic Analysis* (LSA) y soluciona un problema que surge en el método *Probabilistic Latent Semantic Indexing* (pLSI)-, ya que es el más simple y utilizado en la actualidad, y asimismo puede utilizarse fácilmente como punto de partida en métodos más complejos que se requieran para cumplir objetivos más complicados –como el modelado de temas correlacionados, modelado de temas esféricos, modelado de temas supervisados y modelado de temas dispersos- (Blei, 2012).

3.1 Recolección y almacenamiento

Con el objetivo de obtener las opiniones de los usuarios de Twitter acerca de los mejores celulares que existen en el mercado, se desarrolla una aplicación en el lenguaje de programación PHP[5] que se conecta con Twitter a través de una API[6] de búsqueda gratuita proporcionada por la red social que permite obtener una colección de mensajes que coinciden con una consulta especificada. Para ello se realizan las búsquedas en el idioma inglés por palabra clave en relación a las tres organizaciones líderes en el sector de telefonía celular: -#s9 para Samsung, #iphonex para Apple y #HuaweiP20Pro para Huawei para identificar los mensajes que mencionan los celulares más vendidos de cada compañía, sin tener en cuenta los *retweets*, respuestas, imágenes y videos.

La API de Twitter en cada petición del programa PHP devuelve resultados de acuerdo a los parámetros de búsqueda que incluyen la información del mensaje –id, fecha y hora de publicación, el texto, el lenguaje, entre otros-, los parámetros que se utilizaron en esa búsqueda y el usuario – como el id, descripción, nombre, localización, seguidores-, que se insertan en la base de datos. Si bien se encuentran disponibles los datos de los usuarios que escriben los *tweets* y esos datos son públicos, solo se almacena la información relacionada a los mensajes –el texto del mensajes, id, fecha de publicación del tweet y el idioma- ya que al proteger la privacidad de los individuos y la seguridad de sus datos personales, el trabajo de esta forma se enmarca dentro de la investigación responsable e innovación, conocido comúnmente por el acrónimo RRI -*Responsible research and innovation*- (Stahl, 2013).

Luego de definir los parámetros utilizados en cada llamada a la API de Twitter y cuáles son los datos que se almacenan, se ejecuta el programa PHP una vez por semana durante aproximadamente tres meses del año 2018 y se insertan a través de la ejecución de procedimientos almacenados incluídos en un paquete en Oracle. Adicionalmente, con el objetivo de realizar una primera limpieza de datos –debido a que diversos *tweets* pueden tratarse de publicidades o mensajes repetidos-, se realiza un procedimiento para descartar los similares a otros en un porcentaje mayor al 90%, a partir del algoritmo Jaro-Winkler –con la función "utl_match.jaro_winkler" que es nativa de Oracle -, que calcula la distancia de edición entre dos cadenas. Luego de ejecutar el proceso, la cantidad de mensajes de 72398 disminuye a 51151 -43491 de Iphone, 2651 para Samsung, y 5009 de Huawei-, que se analizan con un programa desarrollado en el lenguaje R, el cual se detalla a continuación.

[5] PHP es un acrónimo que significa *Hypertext Pre-processor*. Para obtener más información sobre el lenguaje, se puede acceder a http://php.net/

[6] Se denomina API al conjunto de llamadas a ciertas bibliotecas que ofrecen acceso a servicios determinados. Son las siglas de su denominación en inglés: *Application programming interface*. Se puede acceder a la versión actual de la API de Twitter con la siguiente URL: https://api.twitter.com/1.1/

3.2 Análisis de datos

Se realiza un programa en R -que es un lenguaje gratuito especializado en al análisis estadístico ampliamente utilizado en la actualidad- con el objetivo de extraer información valiosa para las organizaciones a partir de los 51151 mensajes de Twitter que fueron recopilados y almacenados en una etapa previa. Debido a que los *tweets* incluyen sentimientos, opiniones, y escepticismo de las personas previamente a la implementación del modelo de *machine learning* se aplican técnicas de *text mining* para realizar análisis de sentimiento –con el objetivo de filtrar solo los mensajes con sentimiento positivo- y una limpieza de datos – a través de la cual se remueven los signos de puntuación, los números y las palabras que se consideran generales y sin sentido a través del método "*Stopword*".

Una vez que se realiza la limpieza de datos y se filtran los mensajes con opiniones positivas, se procede a aplicar el modelo LDA con la librería "textmineR". Por un lado, se elige una cantidad de 50 temas[7] para los *tweets* de cada una de las tres compañías de teléfonos celulares ya que un valor superior podría arrojar buenos resultados para Iphone que tiene mayor cantidad de mensajes, sin embargo no se cuenta con suficiente información para Samsung y Huawei. Por otro lado, se utilizan los valores de los parámetros α y β establecidos por defecto en la librería –de 0.1 y 0.05 respectivamente-, lo que implica que es muy poco probable que cada mensaje contenga una mezcla de la mayoría de los temas y aún menos probable que cada tema contenga una mezcla de la mayoría de las palabras, lo que tiene sentido para los *tweets* que pueden tratar de diversos temas ya que son escritos por personas en todo el mundo, de diferentes edades que no están relacionados entre sí.

En cuanto a los supuestos que realiza el modelo LDA, se destacan tres –adicionalmente a las distribuciones de los parámetros-: el orden de las palabras en cada documento no importa, el orden de los documentos es indistinto, y la cantidad de temas es fija y conocida (Blei, 2012). Si bien es poco realista que el orden de las palabras en cada mensaje de Twitter sea indiferente, es razonable ya que el objetivo es descubrir la estructura de temas y previamente se realizó un análisis de sentimiento. En cuanto al orden de los documentos, no es realista al analizar colecciones de larga duración que abarcan años ya que los temas cambian con el tiempo, sin embargo en el presente trabajo se analiza un periodo de tiempo de tres meses, por lo que el orden de los *tweets* es indistinto para el modelo. Por último, el supuesto de la cantidad de temas fijo y conocido se puede resolver realizando una estimación o pruebas de diferentes valores de números de temas.

En este marco, se ejecuta el modelo LDA, en el cual se utiliza el muestreo de Gibbs para aproximar la distribución posterior y se realizan 10.000 iteraciones sobre los mensajes pre-procesados. Para cada uno de los teléfonos celulares que se estudian en el presente trabajo el algoritmo descubre 50 temas, de los cuales se seleccionan los 10 que tienen mayor frecuencia dentro del conjunto de *tweets* para cada una de las compañías para realizar el análisis, cuyos resultados se muestran a continuación.

[7]Se realizaron pruebas de ejecución del modelo con diferentes valores de temas -30, 80 y 100- arrojando resultados similares.

Tabla 1: Resultados del modelo LDA

Huawei	**Samsung**	**Iphone**
price	cool	camera_quality
camera_good	love	worth_money
formats_supported	awesome_mobile	finally_upgraded
triple_camera	sunrise_gold	wireless_charging
luck	camera_good	beautiful_day
mate_design	screen_protector	happy_birthday
real_time	supplies_restored	face_id
valuable_global	week_bargain	big_shopping
night_mode	deal_buy	photography
love	intelligent_scan	love

Fuente: Elaboración propia.

Luego de aplicar el modelo LDA a los mensajes con sentimiento positivo que mencionan al celular Samsung 9, Iphone X o Huawei P20 Pro publicados en Twitter en un lapso de aproximadamente tres meses, se obtienen las características que los clientes destacan de cada uno de los teléfonos, que deberían considerarse en el proceso de desarrollo de nuevos celulares. En primer lugar, se destaca la calidad de la cámara de fotos de los tres celulares, sin embargo el Huawei P20 Pro tiene una ventaja sobre los otros dos ya que los clientes priorizan el modo para sacar fotos de noche y asimismo tiene la particularidad de tener tres cámaras. En segundo lugar, para los usuarios es importante el reconocimiento facial que realiza el celular, como el "Face ID" del Iphone X y el "Intelligent Scan" del Samsung 9. Asimismo se debe tener en cuenta que los clientes consideran el cargador inalámbrico de Iphone X como uno de sus principales atributos, y con respecto al diseño de los teléfonos, se menciona la edición dorada "Sunrise Gold" del Samsung 9. Por último, en cuanto al precio, se menciona como una ventaja en los mensajes del celular Huawei P20 Pro, se destacan las ofertas de Samsung 9, sin embargo, por otro lado, los clientes del Iphone X consideran que vale la pena pagar un precio mayor por las prestaciones que brinda.

CONCLUSIONES

En el presente trabajo se implementó el método de aprendizaje automático LDA para desarrollar nuevos celulares a los mensajes de Twitter que publican las personas acerca de los teléfonos líderes del mercado para obtener las características que deberían tener los mismos. Para ello se utilizaron diversas tecnologías –todas gratuitas-: una aplicación en el lenguaje de programación PHP que se conecta a la API de Twitter, una base de datos Oracle Express Edition para almacenar los datos, y un programa en el lenguaje R para implementar los algoritmos de *machine learning* y las técnicas de *text mining*.

En primer lugar, se concluye que a partir del surgimiento del *big data* se pueden explotar nuevas fuentes de información que permiten a las organizaciones crear valor sin implicar altos costos ya que con la información que se obtiene de las redes sociales del tipo *microblogging* como Twitter, las organizaciones pueden no solo explotar la competencia de clientes existentes, sino también identificar nuevos clientes, comprender sus necesidades y evaluar qué recursos son necesarios para dirigirse a ellos, sin que sea necesario realizar encuestas, sondeos de opinión y grupos focales.

En segundo lugar, si bien se encuentran disponibles y son públicos los datos de los usuarios que escriben los mensajes en Twitter, debido a que los datos masivos en general -y los datos publicados en las redes sociales en particular- plantean problemas de privacidad, en este trabajo no se

almacenó la información de los mismos, por lo que por un lado se resuelve uno de los principales inconvenientes que plantea el *big data*, y por otro, la aplicación presentada de la gestión de innovación se enmarca dentro de la investigación responsable e innovación (RRI) ya que es una manera novedosa de gestionar la investigación que permite la inclusión adecuada de los avances tecnológicos de forma ética, responsable y sustentable. Por lo tanto, se demuestra que es posible realizar una implementación de un método de aprendizaje automático de forma responsable con buenas prácticas de protección de la privacidad y seguridad de datos personales que sea de utilidad para las organizaciones.

REFERENCIAS BIBLIOGRÁFICAS

Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. *International conference on discovery science*, 1–15. Springer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

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

Brown, E. D. (2012). Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market. *Proc. of SAIS*.

Casonato, R., Lapkin, A., Beyer, M., Genovese, Y., & Friedman, T. (2011). Information management in the 21st century. *Gartner. Retrieved at March*, *1*, 2016.

Collmann, J., FitzGerald, K. T., Wu, S., Kupersmith, J., & Matei, S. A. (2016). Data Management Plans, Institutional Review Boards, and the Ethical Management of Big Data about Human Subjects. En *Ethical Reasoning in Big Data* (pp. 141–184). Recuperado de http://link.springer.com/chapter/10.1007/978-3-319-28422-4_10

Griffiths, T. (2002). *Gibbs sampling in the generative model of latent dirichlet allocation*.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. *Advances in neural information processing systems*, 856–864.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Recuperado de http://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdf

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Recuperado de http://www.citeulike.org/group/18242/article/9341321

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Schmarzo, B. (2013). *Big Data: Understanding how data powers big business*. Recuperado de https://books.google.com.ar/*books*?hl=es&lr=&id=Tez9AAAAQBAJ&oi=fnd&

pg=PR19&dq=Big+Data:+Understanding+how+data+powers+big+business&ots=wtgWpu7FAj&sig=eaUvxk6EX9PX7CsVm6jdYJL6fwA

Schwab, K. (2017). *The fourth industrial revolution*. Crown Business.

Stahl, B. C. (2013). Responsible research and innovation: The role of privacy in an emerging framework. *Science and Public Policy*, *40*(6), 708–716.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424–440.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*.
Recuperado de
https://books.google.com.ar/books?hl=es&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=Data+Mining:+Practical+machine+learning+tools+and+techniques&ots=8IzKrcjAC8&sig=uD5ipa-nj_Yep6DITdFIZXz4i5g