Revista Argentina de Ciencias del Comportamiento

ISSN 1852-4206

Julio de 2017, Vol. 9, N°2, 65-76

revistas.unc.edu.ar/index.php/racc

Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología

Mariñelarena-Dondena, Luciana*, [a; b] ; Errecalde, Marcelo Luis [a]; Castro Solano, Alejandro [b; c]

Artículo Metodológico

Resumen

La *extracción de conocimiento en bases de datos* es un proceso complejo que en última instancia busca darle sentido a los datos. La minería de datos sólo constituye una etapa de este proceso cuyo objetivo consiste en la obtención de patrones y modelos aplicando métodos estadísticos y técnicas de aprendizaje automático. El presente artículo de revisión examina cómo pueden aplicarse las *técnicas de minería de textos* en el campo de la psicología. En este contexto, se describen los dos grandes propósitos de las *técnicas de minería de textos*: la descripción y la predicción. Finalmente, se destaca que la aplicación de *técnicas de minería de textos* en nuestra disciplina hace posible la medición o evaluación de distintos constructos psicológicos, a diferencia de la utilización de los tradicionales cuestionarios o encuestas.

Palabras clave:
Técnicas de Minería de Textos, Ciencias de la Computación, Evaluación, Psicología.

Abstract

Knowledge discovery applying text mining techniques in Psychology. The knowledge discovery in databases (KDD) is concerned with the non-trivial process of making sense of data. Data mining is only a step in the KDD process that consists in pattern recognition using statistics and machine learning techniques. This literature review focuses on how text mining techniques can be applied in Psychology. In this context, the two main purposes of text mining techniques will be introduced: description and prediction. Finally, this paper highlights the use of text mining techniques as a psychological assessment tool, which differs from the use of standard questionnaires or scales.

Keywords:
Text Mining Techniques, Computer Sciences, Assessment, Psychology

Tabla de Contenido

Recibido el 6 de noviembre de 2015; Aceptado el 15 de marzo de 2017
Editaron este artículo: Mariana Bentosela, Carlos Sabena, María Micaela Marín, Daniela Alonso y Estefanía Caicedo

1. Introducción

En los últimos años se ha difundido ampliamente el concepto de *Big Data*, pero no siempre se utiliza este término de manera correcta. Ya en el año 2001, Laney remarcó los

[a] Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (Universidad Nacional de San Luis).
[b] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
[c] Universidad de Palermo, Buenos Aires, Argentina
*Enviar correspondencia a: Mariñelarena-Dondena, L. E-mail: lucianamd.psico@gmail.com

Citar este artículo como: Mariñelarena-Dondena, L; Errecalde, M.L. & Castro Solano, A. (2017). Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. *Revista Argentina de Ciencias del Comportamiento, 9*(2), 65-76

tres principales desafíos que implicaba el análisis de grandes datos y definió "las tres V" que han caracterizado a *Big Data*: Volumen, Variedad y Velocidad. En primer lugar, *volumen* se refiere a la magnitud de los datos, al hablar de grandes datos nos referimos por ejemplo a millones de publicaciones en Facebook o al análisis de billones de críticas de películas. En segundo lugar, *variedad* alude a la heterogeneidad de los datos que deben analizarse: textos, imágenes, audios, videos, etc. Por último, *velocidad* implica que grandes flujos de información deben ser analizados en tiempo real, por ejemplo los datos provenientes de los *smartphones*. Recientemente algunos autores han destacado otras características de *Big Data*: veracidad, variabilidad (o complejidad) y valor (Gandomi & Haider, 2015).

En este contexto, las ciencias sociales han ingresado en la era de la ciencia de los datos ya que es posible analizar el material disponible a través de los medios sociales: "*Language data available through social media provide opportunities to study people at an unprecedented scale*" (Kern et al., 2016). Esta oportunidad conlleva el desafío de realizar investigaciones interdisciplinarias, por ejemplo entre la informática y la psicología (Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013).

James Pennebaker sostiene que el lenguaje natural que las personas utilizan cotidianamente refleja su personalidad, su situación social y las relaciones interpersonales que entablan con los demás. Incluso afirma que el lenguaje puede servir para evaluar la salud física y mental de los sujetos (Pennebaker, 2002).

> The simultaneous development of high-speed personal computers, the Internet, and elegant new statistical strategies have helped usher in a new age of the psychological study of language. By drawing on massive amounts of text, researchers can begin to link everyday language use with behavioral and self-reported measures of personality, social behavior, and cognitive styles. Beginning in the early 1990s, we stumbled on the remarkable potential of computerized text analysis through the development of our own computer program - Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007). We are now witnessing new generations of text analysis coming from computer sciences and computational linguistics (Tausczik & Pennebaker, 2010, p. 25).

En esa dirección, Schwartz y Ungar (2015) plantean que tradicionalmente los psicólogos y las psicólogas han evaluado los pensamientos, los sentimientos y los rasgos de personalidad mediante cuestionarios administrados a muestras relativamente pequeñas de participantes. En contraposición, ponen de relieve las nuevas alternativas para la evaluación psicológica que brindan el análisis de contenido conducido por los datos *(data-driven content analyses)* o el enfoque del vocabulario abierto *(open vocabulary approach)* si se utilizan los grandes

volúmenes de información disponibles en los medios sociales como Facebook y Twitter.

Al mismo tiempo, otros estudios han demostrado que el análisis del lenguaje disponible en los medios sociales es extraordinariamente útil para realizar estudios epidemiológicos a gran escala o para identificar las características psicológicas que prevalecen en diferentes regiones geográficas, por ejemplo aquellas vinculadas con el bienestar. Este método de evaluación es mucho más rápido y menos costoso que las tradicionales encuestas realizadas por las agencias del gobierno (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al., 2013).

La *extracción de conocimiento en bases de datos (knowledge discovery in databases* - KDD) es "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, pp. 40-41 - la traducción nos pertenece). La minería de datos sólo constituye una etapa de este proceso cuyo objetivo consiste en la obtención de patrones y modelos aplicando métodos estadísticos y técnicas de aprendizaje automático *(machine learning)*. Por último, el proceso de extracción de conocimiento también implica la evaluación e interpretación de los patrones o modelos obtenidos en la etapa de minería de datos (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004) (Figura 1).

Una variante de la minería de datos es aquella que se ocupa específicamente de los datos textuales denominada habitualmente como minería de textos. Con el fin de mantener una terminología uniforme a lo largo del trabajo utilizaremos el término *técnicas de minería textos.*

Figura 1.

Etapas del proceso de extracción de conocimiento en bases de datos (*knowledge discovery in databases* - KDD).

Si bien estos enfoques han despertado un gran interés, todavía se observa en nuestro país un marcado déficit en el desarrollo de investigaciones utilizando *técnicas de minería de textos* para identificar las características del lenguaje natural que las personas utilizan en su vida cotidiana.

El presente artículo de revisión examina el proceso de extracción de conocimiento, en general, y cómo se pueden aplicar las *técnicas de minería de textos* en el campo de la psicología, en particular.

Las *técnicas de minería de textos* persiguen dos grandes propósitos: la predicción y la descripción. Por una parte, las *tareas predictivas* o medición basada en el lenguaje consisten en la construcción de un clasificador automático que estima la variable dependiente, usualmente llamada etiqueta o resultado, en función de determinadas características (variables independientes) extraídas de los documentos. Por la otra, las *tareas descriptivas* buscan obtener patrones que explican o resumen las relaciones subyacentes en los datos. Esto permite, por ejemplo, formular nuevas hipótesis considerando las palabras que utilizan las personas cotidianamente (Fayyad et al., 1996; Schwartz & Ungar, 2015).

En los siguientes apartados se desarrollarán las principales etapas del proceso de extracción de conocimiento como así también las tareas específicas de las *técnicas de minería de textos*.

2. Extracción de conocimiento en bases de datos

KDD es un proceso complejo que involucra distintas etapas, entre ellas las principales son: la preparación de los datos, la minería de datos, la obtención de patrones o modelos y la evaluación e interpretación de los patrones obtenidos previamente. En última instancia, el proceso de KDD busca descubrir conocimiento a partir de los sistemas de información (Fayyad et al., 1996; Hernández Orallo et al., 2004).

2.1. Construcción de Modelos Predictivos

Dentro de las *técnicas de minería de textos*, las tareas predictivas tienen como objetivo la construcción de un clasificador automático mediante un sistema de aprendizaje supervisado (Lex, 2011; Sebastiani, 2002). Este proceso está compuesto por las etapas que se describen a continuación: etiquetado, extracción de características, entrenamiento, evaluación y uso (Figura 2).

1) *Etiquetado*: consiste en asignar la clase, categoría o valor numérico correcto (etiqueta) a cada documento del conjunto de entrenamiento.

2) *Extracción de características*: a partir de los documentos o textos crudos se genera una representación computacionalmente adecuada para su procesamiento por el módulo de análisis (aprendizaje inductivo).

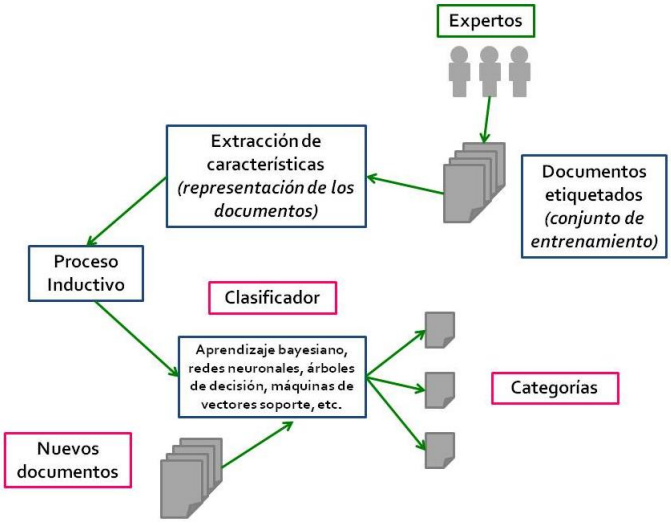

Figura 2.
Etapas involucradas en el proceso de construcción de un clasificador automático mediante un sistema de aprendizaje supervisado *(machine learning).*

Un documento es una unidad de datos textual que puede corresponder a algún

documento del mundo real, por ejemplo: un artículo científico, un escrito personal, un e-mail, los *posts* en los medios sociales como Facebook y Twitter, etc.

A la hora de representar los documentos se pueden utilizar diferentes características. Layton, Watters y Dazeley (2011) establecen la siguiente clasificación: a) *características estáticas*: determinadas a priori, antes del procesamiento de los documentos. Se basan en la frecuencia de caracteres específicos, estadísticas vinculadas a las palabras (por ejemplo, longitud promedio y longitud máxima) y frecuencia de categorías sintácticas particulares (sustantivos, adjetivos, pronombres, etc.), entre otras; y b) *características dinámicas o variables*: se derivan directamente de los términos particulares contenidos en los textos como el modelo bolsa de palabras (*bag of words* - BOW), los n-gramas de palabras y los n-gramas de caracteres.

Lex (2011), por su parte, diferencia las características léxicas de las estilográficas. Los *atributos léxicos* hacen referencia esencialmente a las palabras de contenido. La característica más simple que se puede analizar en un documento es la cantidad de veces que aparece un determinado término. En cambio, los *atributos estilográficos* intentan capturar aquellos aspectos que van más allá del contenido temático del documento reflejando el estilo de escritura del autor, por ejemplo la preferencia por el uso de determinadas palabras de paro o de función (artículos, preposiciones y conjunciones específicas como "el", "de", "con", etc.), categorías de palabras (adverbios, pronombres personales, verbos, etc.), sentencias largas sobre cortas, entre otros.

Para poder categorizar los textos que conforman la colección de documentos por tópicos o temáticas la representación más usual es el *modelo vectorial*. Según este modelo, cada documento se representa por un vector de pesos asociados a los términos que ocurren en la colección de documentos. En otras palabras, a cada uno de los términos que aparecen en la colección de documentos se le asigna un peso específico en función de la cantidad de veces que se registra en cada documento (Figura 3).

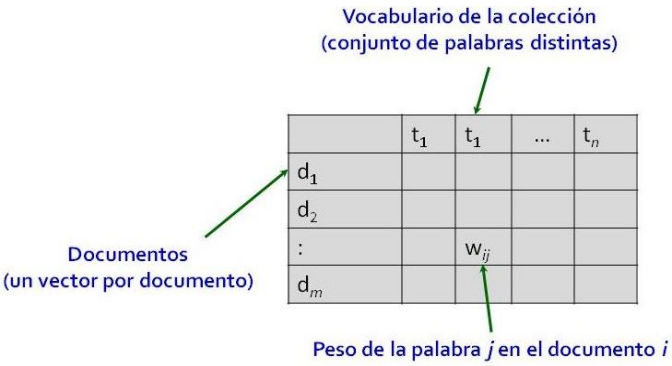

Figura 3.
Representación vectorial de los documentos

De lo expuesto anteriormente se deduce que la importancia o peso de un término se incrementa proporcionalmente al número de veces que aparece en un documento, ya que permite describir el contenido temático del mismo. Mientras que la importancia general de cualquier término decrece proporcionalmente al total de sus ocurrencias en la colección de documentos, puesto que los términos muy frecuentes no facilitan la discriminación de los

textos en distintas clases o categorías.

En los estudios psicológicos, las *técnicas de minería de textos* utilizadas para la extracción de características de los documentos se dividen en dos grandes grupos: 1) las conducidas manualmente o enfoque del vocabulario cerrado (*closed-vocabulary*), centradas en el empleo de diccionarios de palabras especificadas a priori, y 2) las conducidas por los datos también conocidas como enfoque del vocabulario abierto (*open-vocabulary approach*) en las cuales los diccionarios, los tópicos, las palabras y las frases se determinan a partir del conjunto de datos objeto de estudio. Siguiendo la terminología propuesta por Layton et al. (2011), los enfoques de vocabulario cerrado que veremos a continuación, como los diccionarios manuales, entrarían dentro de la categoría de *características estáticas*, mientras que los enfoques de vocabulario abierto como los diccionarios derivados de los textos, el estudio de los tópicos o temas, las palabras y las frases de los documentos corresponden a la categoría de *características dinámicas o variables*.

Tal como lo plantean Schwartz y Ungar (2015), estos dos grandes enfoques presentan distintas variantes que difieren precisamente en el grado de participación manual o de conducción por los datos en el proceso de extracción de características para la representación de los documentos:

a) *Diccionarios manuales*: son listas de términos asociadas a categorías previamente determinadas, se basan en el conteo de palabras. Dado un documento el programa registra cuántas veces aparece cada uno de los términos de todas las categorías. Entre sus principales ventajas se destacan, por un lado, que es una técnica accesible y fácil de aplicar en muestras de datos pequeñas; por el otro, cabe remarcar que fueron los expertos en un área del conocimiento quienes elaboraron las categorías y seleccionaron los términos objeto de estudio.

b) *Diccionarios generados en forma masiva*: compuestos por listas de miles de términos en cuya elaboración participaron cientos de personas buscando incluir todas las palabras más comunes dentro de cada categoría. A diferencia de los anteriores, no estarían sujetos a los posibles sesgos de un pequeño grupo de expertos.

c) *Diccionarios derivados de los textos*: se basan en un proceso de aprendizaje automático donde en primer lugar se etiqueta una vasta colección de documentos, tomando en consideración las características de las personas que los escribieron o los atributos de los textos propiamente dichos. Luego se identifican las palabras y las frases que presentan las correlaciones más altas con un determinado resultado.

d) *Tópicos*: también es posible extraer automáticamente diccionarios de palabras prescindiendo de categorías determinadas previamente. Vale decir, en base al conjunto variable de palabras obtenidas de la colección de documentos se pueden extraer tópicos o grupos de términos semánticamente coherentes.

Estos grupos de palabras relacionadas obtenidos a partir de técnicas de agrupamiento *(clustering)* se basan en un aprendizaje de tipo no supervisado.

e) *Palabras y Frases*: tradicionalmente el análisis lingüístico comienza con la división de la secuencia de caracteres que componen un mensaje en palabras. No obstante, muchas veces las expresiones compuestas por más de una palabra (n-gramas de palabras) constituyen la mejor unidad de análisis.

Entre las principales ventajas del denominado *open-vocabulary approach*, dentro del cual se ubican las últimas tres técnicas, deben mencionarse las siguientes: permite el análisis de grandes volúmenes de datos y refleja de una manera más transparente las características de la propia colección de documentos, posibilitando así la captura de conexiones y asociaciones entre clases de palabras que no se habían considerado inicialmente produciendo resultados inesperados.

3) *Entrenamiento (aprendizaje automático)*: En las tareas predictivas, dada una colección de documentos representada con alguno de los enfoques expuestos más arriba, el siguiente paso será asignarle a cada documento una etiqueta o rótulo que representa una clase, categoría o valor numérico particular. Por ejemplo, si queremos identificar quién es el autor de un documento, las categorías corresponderán a los nombres de los distintos autores de los textos. En cambio, en la predicción numérica se le asignan valores numéricos a los textos y se denomina a la tarea de regresión; en este caso si nuestro objetivo consiste en evaluar el nivel de bienestar de los sujetos mediante los mensajes en Twitter el resultado será una puntuación que nos indicará su grado de satisfacción con la vida.

La construcción de un *clasificador automático* que pueda realizar este tipo de tarea, se basa en un proceso inductivo de aprendizaje automático que busca reproducir un proceso correcto o ideal, vale decir, que para cada *input* o documento a clasificar siempre se genere el mismo *output* o asigne dicho documento a la misma clase. Como vimos previamente, los ejemplos de este comportamiento ideal se especifican en los datos de entrenamiento. Sin embargo, este sistema de aprendizaje inductivo debe ser capaz de extraer las características distintivas de los documentos del conjunto de entrenamiento para luego poder analizar otros textos no observados previamente lográndose así la capacidad de generalización del clasificador que se suele evaluar sobre otro conjunto de prueba separado. Este proceso en matemática se lo conoce como aproximación de una función.

Luego de haber etiquetado el corpus de entrenamiento - haberle asignado a cada representación del documento su clase, categoría o valor numérico correspondiente - se puede entrenar un clasificador utilizando distintos enfoques o algoritmos (Mitchell, 1996; Russell & Norvig, 2009): aprendizaje bayesiano (McCallum & Nigam, 1998), de redes neuronales,

de árboles de decisión, máquinas de vectores soporte (Joachims, 1998, 1999), etc.

Este proceso de aprendizaje busca en un espacio de hipótesis, una hipótesis (modelo/clasificador) que sea consistente con los datos de entrenamiento pero que pueda además clasificar correctamente otros datos (nuevos documentos) no presentes en ese conjunto. La siguiente etapa es la encargada de verificar la capacidad de generalización del clasificador obtenido.

4) *Evaluación y uso*. Si un clasificador automático sólo fuera evaluado sobre los datos de entrenamiento con que fue generado se correría el serio riesgo de obtener modelos que han "memorizado" dichos datos pero que tienen bajo desempeño sobre nuevos documentos; en este caso, se dice que la capacidad de generalización del clasificador es pobre y se denomina a este fenómeno "sobreajuste".

Por lo tanto, se evalúa la utilidad de las representaciones de los documentos y del modelo obtenido sobre un conjunto de prueba separado o utilizando esquemas más complejos como el denominado validación de k-pliegues (*k-fold validation*). En todos estos casos, se mantiene separado el conjunto de entrenamiento del de prueba y se evalúa la precisión del clasificador midiendo la exactitud (porcentaje de documentos clasificados correctamente) en el caso de categorización o midiendo el error cuadrático medio (diferencia entre el valor numérico predicho y el real) en el caso de la regresión. Como resultado de este proceso, se puede generar un *ciclo de ajuste* que involucre la re-ejecución de los experimentos con atributos y/o algoritmos de aprendizaje diferentes. Una vez obtenido un clasificador con un desempeño "aceptable" de acuerdo al dominio de aplicación, éste es puesto en funcionamiento y sus resultados (predicciones) comienzan a ser aplicadas sobre los nuevos datos que vayan ingresando al sistema.

2.2. Análisis exploratorio o Descriptivo

El otro gran objetivo de las *técnicas de minería de textos* es la descripción. Estas tareas buscan comprender las características psicológicas y comportamentales de una población a partir de los patrones del lenguaje; a su vez, estos hallazgos pueden servir para el desarrollo de futuras investigaciones, entre ellas la construcción de modelos predictivos. Un ejemplo de este tipo de análisis exploratorio basado en el enfoque del vocabulario abierto *(open vocabulary approach)* es el método del Análisis Diferencial del Lenguaje *(Differential Language Analysis - DLA)* desarrollado por Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al. (2013).

Este método está compuesto por tres grandes etapas: 1) la extracción de las características del lenguaje, 2) el análisis correlacional, y 3) la visualización. Su objetivo principal consiste en hallar y discriminar aquellas características del lenguaje que mejor representen los atributos psicológicos y

demográficos de una determinada región geográfica o comunidad; vale decir, se establecen correlaciones entre las características del lenguaje extraídas de los propios documentos y las variables de salud o psicológicas objeto de estudio. Para resumir y representar gráficamente los resultados obtenidos los autores utilizan nubes de palabras, ya que las mismas les permiten agrupar los términos según los tópicos o temáticas (Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013).

3. Discusión

El lenguaje de las personas, sobre qué hablan y cómo lo hacen, refleja información que permite diagnosticar su estado de salud física y examinar sus características de personalidad. Aquí la aplicación de *técnicas de minería de textos* en el campo de la psicología hace posible la medición o evaluación de distintos constructos, a diferencia de la utilización de los tradicionales cuestionarios o encuestas.

Se ha comprobado que el análisis de las características del lenguaje empleado en los medios sociales como Facebook y Twitter permite identificar los rasgos de personalidad, el género y el sexo de los participantes (Mairesse, Walker, Mehl, & Moore, 2007; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013) como así también predecir el nivel bienestar de las personas que viven en distintas regiones geográficas de los Estados Unidos de Norteamérica (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al., 2013). Asimismo, se han realizado estudios epidemiológicos a gran escala identificando en el lenguaje de los medios sociales aquellas características psicológicas presentes en la comunidad asociadas con el riesgo de mortalidad por arterosclerosis (Eichstaedt et al., 2015).

También a través de la información disponible en Twitter se han analizado los episodios de *bullying*. Los mensajes o *posts* permiten explorar quiénes estuvieron involucrados, cuál fue el tipo de agresión y quiénes reportan estos hechos. Al mismo tiempo puede registrarse de qué lugar provienen los *posts* y cuándo (qué día y a qué hora) fueron realizados (Bellmore, Calvin, Xu, & Zhu, 2015).

Investigaciones recientes sugieren que las *técnicas de minería de textos* podrían usarse incluso para la detección temprana y la prevención del suicidio. Con tal fin Desmet y Hoste (2013) analizaron el contenido y, más precisamente, las emociones de notas suicidas.

Por otro lado, se ha usado la información política comunicada en los medios sociales para seguir o monitorear las opiniones de los usuarios en tiempo real. En ese sentido, a partir del estudio de los *posts* de 1000 usuarios de Twitter de Estados Unidos de Norteamérica se comprobó que es posible identificar la alineación política - de izquierda o de derecha - de los individuos mediante el análisis de semántica latente (Conover, Gonçalves, Ratkiewicz, Flammini, & Menczer, 2011).

Entre los desafíos y las posibles líneas de

investigación a futuro debemos mencionar las siguientes. En primer lugar, la necesidad de realizar estudios en distintas comunidades y contextos culturales buscando así superar el problema de las falacias ecológicas. En segundo término, analizar no sólo el contenido textual disponible en los medios sociales sino también las imágenes, las grabaciones de audio, los videoclips, etc. Esta información multimodal puede reflejar otros aspectos de los pensamientos, los sentimientos y las preocupaciones de las personas que no llegan a ser capturados sólo mediante el análisis de los textos (Schwartz & Ungar, 2015).

Respecto a este último punto, se han evaluado las relaciones existentes entre la cantidad de fotos que suben los usuarios de Facebook y las interacciones de los mismos con sus amigos (por ejemplo, a través de la cantidad de "me gusta" y comentarios recibidos) con sus rasgos de personalidad (Eftekhar, Fullwood, & Morris, 2014). Asimismo se ha examinado si la foto de perfil que eligen los usuarios refleja su personalidad, ya que dicha fotografía determina en gran parte la identidad *online* del sujeto (Jim Wu, Chang, & Yuan, 2015).

Las investigaciones interdisciplinarias de las ciencias de la computación, sociales y de la salud constituyen sin lugar a dudas un campo promisorio. En nuestra disciplina en particular, estos enfoques abren la puerta para la medición o evaluación de los constructos psicológicos mediante la aplicación de *técnicas de minería de textos*. Esta perspectiva podría convertirse a futuro en un nuevo método de evaluación psicológica a nivel individual y para la realización de estudios epidemiológicos a gran escala.

Normas Éticas

En la realización del presente artículo de revisión *(literature review)* se respetaron las normas éticas internacionales establecidas por la *American Psychological Association* (http://www.apa.org/ethics/code/index.aspx) y la *Declaración de Helsinki* (http://www.wma.net/es/30publications/10policies/b3/).

Referencias

Bellmore, A., Calvin, A. J., Xu, J. M., & Zhu, X. (2015). The five W's of ''bullying'' on Twitter: Who, What, Why, Where, and When. *Computers in Human Behavior, 44*, 305–314.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the Political Alignment of Twitter Users. *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*. Boston, USA, 192-199. doi: 10.1109/PASSAT/SocialCom.2011.34

Desmet, B., & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications, 40*(16), 6351–6358.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... Seligman, M. E. P. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science, 26*(2), 159-169. doi: 10.1177/0956797614557867

Eftekhar, A., Fullwood, C., & Morris, N. (2014).

Capturing personality from Facebook photos and photo-related activities: How much exposure do you need? *Computers in Human Behavior, 37*, 162–170.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37-54.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos*. Madrid: Pearson Prentice Hall.

Jim Wu, Y. C., Chang, W. H., & Yuan, C. H. (2015). Do Facebook profile pictures reflect user's personality? *Computers in Human Behavior, 51*(B), 880-889.

Joachims, T. (1998) Text categorization with Support Vector Machines: Learning with many relevant features. En C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)* (pp. 137-142). Heidelberg: Springer.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. En I. Bratko & S. Dzeroski (Eds.), *Proceedings of ICML-99, 16th International Conference on Machine Learning* (pp. 200–209). San Francisco: Morgan Kaufmann Publishers.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining Insights From Social Media Language: Methodologies and Challenges. *Psychological Methods, 21*(4), 507-525.

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. *Application Delivery Strategies,* META Group Inc. Recuperado de http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Layton, R., Watters, P., & Dazeley, R. (2011). Recentred Local Profiles for Authorship Attribution. *Natural Language Engineering, 18*(3), 293-312.

Lex, E. (2011). *Content Facets for Individual Information Needs in Media.* (Tesis Doctoral). Graz University of Technology, Styria, Austria.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research, 30*, 457-500.

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *Learning for Text Categorization: Papers from the 1998 AAAI Workshop, 752*, 41-48.

Mitchell, T. M. (1996). *Machine Learning*. New York: McGraw Hill.

Pennebaker, J. W. (2002). What our words can say about us: Toward a broader language psychology. *Psychological Science Agenda, 15*(1), 8-9.

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). New Jersey: Prentice Hall.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., … Ungar, L. H. (2013). Characterizing Geographic Variation in Well-Being using Tweets. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (ICWSM), Boston, USA, 583-591

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE, 8*(9), e73791.

Schwartz, H. A., & Ungar, L. H. (2015). Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 78-94.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys, 34*(1), 1-47.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology, 29*(1), 24-54.