Una exploración de la semántica del concepto de ‘Derechos’ en Twitter

Luz Marina Barreto Guerra, Haydemar Núñez
(Universidad Central de Venezuela)

UCV ESCUELA DE FILOSOFÍA

apuntes filosóficos

Vol. 30 No. 58

Una exploración de la semántica del concepto de "Derechos" en Twitter

Exploring the semantic of the concept of "Rights" in Twitter

Luz Marina Barreto Guerra, Haydemar Núñez[*]
(Universidad Central de Venezuela)

Resumen: Este artículo estudia tres técnicas de modelado de tópicos, o reducción de la dimensionalidad, para la recuperación de la semántica en matrices dispersas de bolsas de palabras y las aplica sobre tres conjuntos de datos de texto tomados de la red social Twitter. Se trata de técnicas destinadas al modelado de tópicos o conceptos latentes en conjuntos desestructurados de datos de texto, también conocidas como métodos para modelar la *semántica latente* de emisiones comunicativas. Las tres técnicas son: el análisis semántico latente o LSA, el análisis semántico latente de índole probabilística o PLSA y la atribución o adjudicación latente de Dirichlet o LDA. El artículo estudia los fundamentos teóricos que subyacen al desarrollo de sus algoritmos y aplica sus implementaciones, en el lenguaje de programación Python, a un corpus de documentos tomados de la red social Twitter. El corpus consta de tres conjuntos de datos de texto en los cuales se busca reconstruir la semántica latente del concepto "derechos humanos", tal y como se expresa en emisiones de usuarios de Twitter provenientes del entorno Iberoamericano. Al analizar los resultados obtenidos en las aplicaciones se pudo comprobar que el algoritmo de la LDA ofrece una semántica más general y profunda del concepto estudiado, al atravesar transversalmente los documentos, que la que arroja el algoritmo del PLSA, cuyos resultados dan mejor cuenta de la semántica *ad intra* de los documentos. Al mismo tiempo, fue posible constatar que los algoritmos que implementan modelos de inferencia bayesiana son más eficientes para la tarea de modelado de tópicos que los algoritmos que calculan valores singulares en matrices factorizadas. También se pudo comprobar un manejo competente de la semántica de la noción de "derechos humanos" por parte de los usuarios de esa red social, el cual evidencia familiaridad con el significado teórico e institucional de dicho concepto. No obstante, ese manejo se mantiene siempre en un nivel elevado de convencionalidad.

Palabras clave: Modelado de Tópicos, Reducción de la Dimensionalidad, Semántica Latente, Factorización de Matrices, Modelos de Inferencia Bayesiana, Algoritmos LSA, PLSA y LDA.

Abstract: The following paper focuses upon three techniques of topic modeling, or dimensionality reduction, aimed at reconstructing the semantics of sparse data text, and presents their applications to three data sets retrieved from Twitter. These techniques usually model latent topics or concepts on unstructured data sets, also

[*] (Dr. Phil. FU Berlin), Profesora Titular de la Escuela de Filosofía de la Universidad Central de Venezuela. Este artículo está basado en mi Tesis de Maestría para optar al Grado de Magister Scientiarium en Ciencias de la Computación, el cual fue realizado bajo la tutoría de la Profesora Dra. Haydemar Núñez (Facultad de Ciencias, UCV).

known as methods intended to model the *latent semantics* of communicative emissions. The three techniques studied are: the Latent Semantic Analysis or LSA, the Probabilistic Latent Semantic Analysis or PLSA and the Latent Dirichlet Allocation or LDA. The paper deals with the theoretical foundations that underlie their algorithms and applies their implementations, in Python, on a corpus of documents taken from the social network Twitter. The corpus consists of three sets of data text whose analysis attempts to reconstruct the latent semantic of the concept "Rights", as it is revealed in the emissions by users of Twitter from the Ibero-American region. It was possible to confirm that the LDA algorithm offers a broader and a more general semantic of the concept "Rights" than that offered by the PLSA algorithm, as the LDA results retrieved a more transversal semantic throughout all the documents. Conversely, the PLSA algorithm gives a better perspective of the topics related to the studied term ad intra the documents of the corpus. It was also established that the models that implement Bayesian inference are better for the task of topic modeling than those algorithms, such as the LSA, that calculate singular value decompositions in factorized matrices. It was observed a competent understanding of the semantic of the concept of "Rights" among the users of Twitter, which evidences some familiarity with the institutional and theoretical meaning of the term. This kind of semantic competence, however, never surpassed a conventional understanding of the notion of "Rights".

Keywords: Topic Modeling, Dimensionality Reduction, Latent Semantics, Matrix Factorization, Bayensian Inference Models, Algorithms LSA, PLSA and LDA.

1. Introducción

Este artículo aplica tres técnicas para la reconstrucción automatizada o el reconocimiento automatizado (*machine learning*) de la semántica en conjuntos de datos de texto, también conocidos como *modelado de tópicos o conceptos*, sobre tres conjuntos de datos de texto tomados de la red social Twitter. Su propósito general es explorar la "semántica latente", como prometen dichos métodos, del concepto de *derechos humanos*, tal y como dicha semántica se encuentra expresada implícitamente en las emisiones de usuarios de la mencionada red de microblogging.

La conjetura que anima este ensayo supone que dicha aplicación pudiera poner de relieve efectivamente la comprensión semántica, o conceptual, que usuarios de Twitter del entorno Iberoamericano pudieran tener del concepto de *derechos*, un término que típicamente integra oraciones que expresan juicios valorativos o juicios que expresan presunciones de validez sobre el concepto de *derechos humanos*, así como sus extensiones conceptuales referidas a derechos de

seres sintientes o no humanos. Este tipo de estudio, que habitualmente es realizado por un filósofo o pensador solitario llevado por su intuición o capacidad para el análisis conceptual, ¿cómo puede hacerse de manera automatizada? En otras palabras: ¿puede la filosofía analítica extender sus alcances a través del reconocimiento automatizado (o *machine learning*) de las implicaciones semánticas de un concepto básico usado por una comunidad amplia de hablantes?

Las tres técnicas de reconocimiento de la semántica latente estudiadas y aplicadas en el presente estudio son:

1. El análisis semántico latente o LSA (por sus siglas en inglés: *latent semantic analysis*)

2. El análisis semántico latente de índole probabilística o PLSA (por sus siglas en inglés *probabilistic latent semantic analysis*) y

3. La atribución o adjudicación latente de Dirichlet o LDA (*latent Dirichlet allocation*).

Estas tres técnicas pueden enmarcarse, a su vez, en dos paradigmas más amplios que son objeto de mucha discusión en las ciencias de la computación contemporáneas, a saber:

1. Los modelos de estadística "frecuentista" y

2. Los modelos de inferencia probabilística o modelos bayesianos.

La distinción entre la estadística frecuentista y la inferencia probabilística es también una distinción relativamente reciente. Los filósofos que estudian la teoría de la elección racional, en particular los que estudian ética y filosofía política, están familiarizados con ella gracias al trabajo clásico de Duncan Luce y Howard Raiffa, *Games and Decisions*[1], quienes extendieron al campo de la ciencia social aplicada y al de los modelos conductistas de actores sociales los desarrollos pioneros de los matemáticos John Von Neumann y Oskar Morgenstern, en *Theory of Games and Economic Behavior*[2], para comprender de manera matemáticamente precisa los juegos de n-personas en donde se producen conflictos de interés y en los cuales hay que tomar decisiones sobre estrategias cuyas probabilidades de éxito se estiman como distribuciones de probabilidad.

[1] LUCE, R. Duncan y RAIFFA, Howard, *Games and Decisions*, New York, John Wiley & Sons, 1957.
[2] VON NEUMANN, John y MORGENSTERN, Oskar, *Theory of Games and Economic Behavior*, Princeton, Princeton University Press, 1944.

En Luce y Raiffa la teoría de las decisiones racionales formalizadas, teoría que, entre tanto, se constituyó en uno de los pilares de la inteligencia artificial (y, por ende, de las técnicas que estudiamos aquí), se divide conforme a los siguientes criterios:

1. Si la decisión es hecha por un individuo o un grupo (dos o más individuos) y

2. Si la decisión se realiza bajo condiciones de: *certidumbre*, *riesgo* e *incertidumbre*.

En el momento en el que Luce y Raiffa escriben su trabajo, la inferencia estadística combina en un mismo ámbito de indagación decisiones en condiciones de *riesgo* (cuando se conocen los parámetros de un modelo estadístico) y decisiones en condiciones de *incertidumbre* (cuando no se conocen los parámetros, por ejemplo, cuando no se sabe qué tipo de estrategia elegirá un contendor en un juego de suma no nula). No obstante, con el paso de los años, y con el crecimiento de nuestra capacidad de computación, ambos tipos de decisión se han ido distanciando una de la otra y es más fácil ahora encarar modelos no-paramétricos con métodos bayesianos.

La distinción entre *riesgo* y verdadera *incertidumbre* es muy importante en las ciencias de la computación actuales porque muchos investigadores están preocupados por diseñar mejores y más eficientes algoritmos destinados a modelar conjuntos de datos que crecen exponencialmente, en particular datos de texto emanados de redes sociales. De este modo, la creciente popularidad que comienzan a gozar los modelos computacionales bayesianos descansaría en su capacidad para permitir a las máquinas de reconocimiento la toma de decisiones en condiciones de cada vez mayor incertidumbre, es decir, en donde el cálculo de los parámetros probabilísticos y de la probabilidad marginal de una variable se vuelven intratables computacionalmente o se van perdiendo en cada nueva iteración de un algoritmo. Por esta razón, en contraste con el escenario que Luce y Raiffa tenían ante sí, hoy en día el cálculo de la incertidumbre está en manos de modelos computaciones que se apoyan en la capacidad de la regla de Bayes para encarar el cálculo de la probabilidad posterior sobre familias exponenciales profundas.

Como veremos, los algoritmos LSA y el PLSA, aplicados al modelado de la semántica latente del concepto de “derechos” usado en emisiones de Twitter, son modelos típicamente frecuentistas, en donde el PLSA, como su nombre lo indica, calcularía las probabilidades que se

desprenden de las frecuencias detectadas o parámetros disponibles en un conjunto de datos[3]. Por contraste, la LDA aspira a dar un salto cualitativo para el cálculo de la probabilidad posterior, -una cantidad desconocida, es decir, no detectada como frecuencia en un conjunto de datos-, en el modelado de tópicos de índole probabilístico y por eso los bayesianos lo consideran un enfoque orientado ya al cálculo de la incertidumbre (por su enfoque orientado a optimizar el cálculo de la máxima probabilidad posterior o MAP, *maximum a posteriori*, y su capacidad de conjugar la probabilidad previa para sucesivas iteraciones de los algoritmos).

Las tres técnicas de modelado de tópicos que son objeto de estudio tienen interés para el científico social y para el filósofo porque atañen a un problema central del análisis del lenguaje ordinario: la reconstrucción de la semántica latente en conjuntos de datos de texto.

En efecto, la capacidad de *entender* el sentido que emerge de un conjunto de palabras, que es también la capacidad de *clasificar* algo como algo, es, sin duda, el problema más antiguo del conocimiento y, por lo tanto, uno de los problemas más importantes de la filosofía. Se trata de comprender cómo se pueden formar enunciados y juicios que son considerados verdaderos o válidos por hablantes competentes, en la medida en que se apoyan en un significado compartido en común. Por esta razón, el problema de la comprensión del sentido o de la comprensión de la semántica de una oración está ligado al establecimiento de la verdad, de lo que *sea el caso*. Se trata de un problema que define la metafísica antigua, especialmente la aristotélica, pero es en el siglo XX que la teoría de la filosofía define el problema como uno que atañe a los métodos de análisis de conceptos básicos o fundamentales[4].

Aristóteles definió, en su tratado *Los analíticos posteriores,* lo que en los siglos subsiguientes seguía siendo considerado el método de la fundamentación del conocimiento: un conjunto de proposiciones ligadas por relaciones de fundamentación es verdadero si es posible hacer un recuento exhaustivo de todos los axiomas que se encuentran en su base. La tarea de todas las ciencias era, de acuerdo con el Estagirita, encontrar sus axiomas básicos, de los cuales

[3]El LSA también puede decirse que calcula una probabilidad a partir de una frecuencia detectada. En efecto, el coeficiente de *x* en una función de regresión puede definirse como la *probabilidad* de que un punto en el conjunto de datos se acerque a la *media* de la función de regresión y, por lo tanto, pueda ser modelado por ella. Son los modelos bayesianos los únicos que calculan una probabilidad sobre cantidades desconocidas. Por lo tanto, a diferencia de los otros dos, lidian con la incertidumbre. De este modo, si bien tanto el PLSA como el LDA son modelos que invocan probabilidades de modo explícito, el único que lidia con la incertidumbre o con una probabilidad desconocida es el LDA. Veremos esto con detalle más adelante.

[4] STRAWSON, P.F., *Analysis and Methaphysics*, Oxford, Oxford University Press, 1992.

dependería la validez de sus teoremas, reglas de inferencia y la verdad de sus enunciados. La comprensión racional o semántica no sería sino un derivado natural de este proceso.

Sin embargo, a medida que el proyecto de encontrar los axiomas últimos, (o verdades autoevidentes), de la ciencia va perdiendo vigencia a lo largo del siglo XX, en parte gracias al impacto que el teorema de incompletitud de Kurt Gödel tiene sobre ese proyecto, el problema de explicar cómo es que podemos decir que algo es verdadero en un sentido racional o universalizable, es decir, cómo podemos entendernos sobre ello, comienza a exigir otras explicaciones.

El gran filósofo austríaco del siglo XX, Ludwig Wittgenstein, en línea con el verificacionismo de Ernst Mach[5], sugiere en su *Tractatus Logico-Philosophicus* que nos entendemos sobre algo si es posible señalar aquel "estado de cosas" que hace que un enunciado sea comprensible a un interlocutor y, por lo tanto, verdadero en algún tipo de contexto de verificación. Con ello, se deja de lado la necesidad de encontrar axiomas últimos que sean susceptibles de un acuerdo convencional entre científicos. Sólo aquel que "tiene un mundo", bien sea porque vive en él con otros hablantes o porque comparte un lenguaje común con otros, puede decir que sabe lo que alguien quiere decir cuando afirma algo.

Las cosas son mucho más complejas de lo que se puede decir en este rápido esbozo, pero, en definitiva, el énfasis sobre la semántica del lenguaje es fundamental para nuestra comprensión actual de la filosofía como una forma de análisis conceptual profundo o sobre categorías básicas.

2. Materiales y métodos

2.1. Captura de los *tuits*

En primer lugar, se creó una aplicación en Twitter. Seguidamente, los días 24/9/2017, 29/9/2017 y 7/10/2017 se capturaron en Caracas, Venezuela, en una conexión en tiempo real que se extendió alrededor de 45 minutos en las tres ocasiones, tres conjuntos de documentos o tuits que constaban de: 1.148 tuits o documentos para 13.565 atributos o dimensiones (conjunto o muestra A); 3.415 tuits para 32.407 atributos o dimensiones (conjunto o muestra B) y 2.020 documentos o tuits para 20.881 atributos o dimensiones (conjunto o muestra C).

[5] MOULINES, Carlos Ulises, *El desarrollo moderno de la filosofía de la ciencia (1890-2000)*, México D.F., UNAM Instituto de Investigaciones Filosóficas, 2011.

2.2. Pre-procesamiento de los documentos

2.2.1. Análisis sintáctico (*parsing*)

Los tweets suelen ser típicamente un corpus de documentos particularmente desetructurado y "ruidoso". Por esta razón, el siguiente paso fue procesar los tweets para hacerlos aptos para el análisis de su semántica latente. Para ello, se utilizó un script que pasa el documento inicial por tres filtros sucesivos: uno para las "stopwords" o palabras comunes de poco peso semántico (tales como artículos determinados e indeterminados, preposiciones, etc.), otro filtro para puntuaciones y caracteres extraños (tales como la arroba o los *slash*) y otro para los caracteres numéricos.

2.2.2. Ponderacion TF·IDF de matrices de términos documentos

A este tipo pre-procesamiento le sigue otro que es crucial para las tareas de análisis semántico de conjuntos desestructurados de datos de texto. En todas las tareas de análisis de conjuntos de datos de texto, los datos se representan como matrices de datos multidimensionales que registran la frecuencia con las que ocurren en un documento dado, una técnica que se conoce como "*bag of words*", bolsas de palabras o BOW por su siglas en inglés. Las palabras también se conocen como "términos".

En estas matrices, se registra simplemente la frecuencia de un término o una dimensión en un documento o, en representaciones categoriales, solo si está presente o no.

De esta manera, un segundo paso importante en el preprocesamiento de conjuntos dispersos de datos de texto es la normalización de la frecuencia de los términos a través de su representación *tf·idf*, que pondera la frecuencia con la que aparece un término en *un* documento con la frecuencia del mismo término en la *colección* de los documentos que forman parte del corpus.

2.3. Aplicación de los algoritmos

2.3.1. El análisis semántico latente o LSA.

El primer método de modelado de tópicos que se aplicó sobre los tres conjuntos de datos de texto es una forma especial de la *descomposición en valores singulares* (SVD, *singular value decomposition*): la descomposición en valores singulares parcial *truncada*. Esta técnica proyecta

tanto términos como documentos en un subespacio de baja dimensionalidad. Se trata de un algoritmo iterativo que computa autovalores y autovectores de matrices grandes y dispersas, matrices como las que consideramos aquí, usando la multiplicación de matrices y vectores. La descomposición en valores singulares se encuentra estrechamente relacionada con el análisis de componentes principales dado que apela al mismo tipo de estrategia, a saber, el análisis de la forma espectral de una matriz.

En el álgebra lineal, el proceso de factorización, que consiste en la descomposición de una matriz en otras matrices que, al ser multiplicadas de nuevo entre sí, ofrecerían como producto la matriz original, busca ayudar a descubrir atributos latentes que revelan las relaciones implícitas entre las entidades que conforman la matriz que se descompone. La primera técnica de modelado o descubrimiento de tópicos que se estudia utiliza un algoritmo de descomposición en valores singulares o SVD y, por lo tanto, es una forma de SVD aplicada a los conjuntos de datos de texto.

La descomposición en valores singulares SVD, y también el análisis de componentes principales o PCA (por sus siglas en inglés: *Principal Components Analysis*), descubren relaciones implícitas entre términos al interior de una matriz de términos-documentos grande y dispersa. Estas relaciones implícitas con frecuencia son agrupamientos o *clusters*. La factorización de matrices es una técnica de reducción de la dimensionalidad porque permite agrupar las entidades de una matriz bajo atributos o tópicos (o conceptos) de mayor envergadura o capacidad englobadora, de forma que entidades aisladas que pudieran haber sido consideradas como su propia clase o categoría, por decirlo así, pueden ahora ser agrupadas bajo categorías analíticas más abarcadoras o de mayor profundidad ontológica.

La descomposición en valores singulares, SVD, y el análisis de componentes principales PCA, de matrices de datos multidimensionales supone un procedimiento análogo a la operación de centrado en la media que ajusta los datos a una función de regresión en modelos lineales. Esta operación de centrado consiste, como en todas las medidas de error en modelos lineales, en sustraer la media del conjunto de datos con respecto a cada punto del mismo; su resultado es un conjunto de datos "centrado" ya en el origen.

Sin embargo, en conjuntos de datos dispersos, y, por lo tanto, no susceptibles de ser representados directamente en modelos lineales, un centrado en la media es especialmente difícil.

Por esta razón, las técnicas de SVD y PCA proponen un análogo de este tipo de centrado. El "centrado en la media" que propone el SVD y el PCA no es sino un método ingenioso para encontrar el valor esperado medio que definiría una línea de regresión en datos cuya elevada dimensionalidad impide un cálculo más sencillo, como la suma del cuadrado de los errores por ejemplo, para encontrar correlaciones potenciales entre datos.

Desde el punto de vista de la representación geométrica de su estrategia algorítmica, el objetivo del SVD y PCA es poder rotar los datos a un sistema de ejes en donde la mayor cantidad de varianza de los datos sea capturada con un número mínimo de dimensiones. Esto se hace, como ya hemos señalado, con un análisis de la forma espectral de la matriz BOW.

En nuestra aplicación hemos usado la implementación del SVD truncado que ofrece la biblioteca de métodos y códigos para el aprendizaje automático en Python "Scikit-learn"[6].

2.3.2. El análisis semántico de índole probabilística o PLSA

Las técnicas de factorización de matrices que extraen valores singulares de las matrices de bolsas de palabras dejan de funcionar bien con conjuntos masivos de datos de textos, en donde los *outliers*, que en presencia de una matriz de frecuencias acotada o finita se pueden reconocer y dejar de lado, pudieran pasar a tener un peso inesperado en la computación. En este caso, cuando todos los datos en un plano, por el crecimiento exponencial de los datos, se vuelven cada vez más difíciles de describir estadísticamente con una función de regresión o con un análogo de un centrado en la media, -como con los métodos de factorización de matrices que apelan al cálculo de la forma espectral de una matriz-, una buena solución es calcular probabilidades de la distribución de tópicos en un corpus de documentos.

La decisión automatizada de una aplicación de modelado de tópicos basada en una técnica de PLSA es, pues: ¿Cuál es la probabilidad de que una palabra observada en un conjunto de datos de textos apunte a un tópico en vez de otro, es decir, pueda ser subsumida o proyectada en un espacio de baja dimensionalidad?

El PLSA es, ante todo, una continuación del LSA, en el sentido de que atribuye palabras a tópicos o conceptos latentes con base en la frecuencia ponderada de unos términos en algunos

[6] PEDRAGOSA *et al., Scikit-learn: Machine Learning in Python*, 12, 2011, JMLR, pp. 2825-2830.

documentos en vez de en otros, aunque interpreta o representa esas frecuencias como una distribución de probabilidad. Por esta razón, el PLSA sigue siendo una técnica de estadística descriptiva de tipo frecuentista, en la medida en que la probabilidad de que un término forme parte de *clusters* de términos pertenecientes a un tópico o concepto dado depende de parámetros arrojados por un conteo de frecuencias dadas en una matriz de bolsas de palabras, frecuencias que posibilitan un cálculo de probabilidad multinomial.

El objetivo del PLSA es estimar la distribución de probabilidad multinomial de algunas palabras en un tópico en contraste con el otro, de modo que, por ejemplo, sea posible encontrar una proporción o distribución de probabilidad de tópicos en un documento cualquiera, dado un conjunto de palabras observadas en el mismo, que permita luego clasificarlos como documentos que aluden a uno de los tópicos presentes en el corpus.

En los sistemas de PLSA, las probabilidades de que una palabra de un documento sea generada por un tópico se actualizan a través de un cálculo de la probabilidad conjunta, en donde las probabilidades de que una palabra probablemente fuera generada por un tópico θ_k (la probabilidad condicional o *likelihood*) se multiplican por la distribución de probabilidad previa de la palabra en el tópico θ_k

Esta probabilidad condicional se computa con el algoritmo EM (Esperanza- Maximización o *Expectation-Maximization*, por sus siglas en inglés), en donde se trata de definir un conjunto de variables escondidas o latentes *z*, cada una de las cuales sería un índice de tópicos.

Ahora bien, de acuerdo con Crain y Zhou[7], maximizar la probabilidad condicional en sistemas de PLSA es lo mismo que minimizar la divergencia de Kullback-Leibler, un método de evaluación de la distancia o de la entropía relativa entre dos distribuciones de probabilidad, una aproximada y una "real", que ofrece una medida de ganancia o pérdida de información. La divergencia KL entre dos funciones de densidad de la probabilidad evalúa la distancia entre la distribución medida empíricamente o aproximada, y la distribución "verdadera". Por esta razón, el PLSA utiliza básicamente un método para la minimización de la divergencia KL.

[7] CRAIN, S, KE ZHOU et al., "Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond" en AGGARWAL, Charu y ZHAI, ChengXiang (Eds.), *Mining Text Data*, New York, Springer, 2012.

En la aplicación, siguiendo la implementación del PLSA en la biblioteca de métodos de Python Scikit-Learn ya citada, el análisis semántico latente de índole probabilístico se concibe como una factorización de una matriz no negativa X (la matriz BOW de tweets/documentos-términos que se usa) con la función objetivo definida por la divergencia Kullback-Leibler generalizada[8].

2.3.3. La atribución latente de Dirichlet o LDA.

La tercera y última técnica de modelado de tópicos aplicada a los tres conjuntos de datos de texto es la atribución o adjudicación latente de Dirichlet o LDA, *latent Dirichlet allocation*.

Lo que distingue los modelos probabilísticos que se apoyan en probabilidades basadas en un conteo de frecuencias de las aplicaciones para modelado de tópicos que utilizan la adjudicación latente de Dirichlet es la capacidad que exhibe esta última para inferir, no simplemente *la probabilidad condicional de las palabras dados los tópicos en un corpus de documentos (*conocida como su *Maximum Likelihood Estimation), sino distribuciones de probabilidad basadas en cantidades desconocidas, las cuales resultan más apropiadas para calcular probabilidades en conjuntos de datos de texto que, o bien son muy grandes, o bien crecen exponencialmente*. Por esta razón, los modelos de LDA no solo calculan una máxima distribución condicional de la probabilidad de palabras dados tópicos en una colección de documentos que se conoce, y cuya frecuencias de palabras se aspira a registrar (el *MLE*), sino también el así llamado *maximum a posteriori* (o *MAP*), es decir, el cálculo de las distribuciones de la probabilidad posterior (futura o desconocida), en conjuntos de datos que crecen o pudieran crecer exponencialmente.

La LDA supone, pues, una implementación completa de la regla de Bayes en sistemas de modelado probabilístico de tópicos y, por esta razón, posibilitan modelos de decisión sobre tópicos caracterizados por la *incertidumbre*. Con la LDA, el modelado de tópicos se separa realmente del enfoque frecuentista y asume plenamente su carácter de *inferencia probabilística*.

A medida que crece el corpus, el número total de sus palabras, tópicos ocultos y documentos, la computación de la inferencia bayesiana se convierte en un problema de cálculo de

[8] *Scikit-learn*. Documentos y ejemplos. [En línea] http://scikit-learn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda.html#sphx-glr-auto-examples-applications-plot-topics-extraction-with-nmf-lda-py.

optimización de funciones, en donde la probabilidad se calcula con una función de densidad para definir el área del gráfico que abarca los datos.

Que el cálculo de la posterior exija, primero, elevar una presunción de probabilidad para futuras iteraciones del algoritmo de modelado probabilístico de datos con base en probabilidades condicionales calculadas para conjuntos de datos anteriores y, segundo, que la posterior exija, a su vez, un cálculo de optimización sobre máximos que solamente pueden ser locales, es lo que hace que los modelos de LDA sean sumamente difíciles desde el punto de vista operacional. Comportan un componente subjetivo que puede resultar polémico, dado no sólo por la distribución inicial de la previa, que se apoya en una opinión facilitada por expertos sobre una posible distribución de tópicos en documentos, sino también porque el cálculo de la probabilidad posterior, que ofrece presunciones racionales basadas en el teorema de Bayes sobre distribuciones futuras, lo hace iterando la regla de Bayes con base en probabilidades posteriores que pasan a definir la previa de la siguiente iteración. Como es evidente, sucesivas iteraciones de las distribuciones de la probabilidad posterior, en tanto que previas de la siguiente iteración del algoritmo, pudieran sesgar el análisis de los datos, por pérdida de los parámetros, alejándolos de una evidencia que ya no es completa. El LDA ofrece, así, una solución ingeniosa al problema de calcular la probabilidad de tópicos en corpora de documentos cuya probabilidad marginal se desconoce.

En el enfoque de la inferencia bayesiana, que calcula una distribución de probabilidad posterior, una distribución de probabilidad previa se multiplica por la distribución condicional, que se apoya en una muestra de los datos observados, y luego este resultado se normaliza por la distribución de probabilidad marginal, es decir, la sumatoria igual a 1 (recordemos que se llama "marginal" porque antiguamente se sumaba "en el margen" de la matriz de datos) de todas las variables aleatorias de la distribución de probabilidad que abarca la evidencia completa. Para una distribución posterior (predictiva):

$$p(T \mid P) = \frac{p(P|T)p(T)}{p(P)}$$

En donde *T* denota los tópicos de un corpus y *P* todas las palabras de ese corpus.

La inferencia bayesiana predictiva ofrece, a la par de una probabilidad *condicional* sobre la evidencia disponible $p(P|T)$, un cálculo de la probabilidad *posterior* (como se expresa en la ecuación anterior: $p(T|P)$). El teorema de Bayes se usa, entonces, para estimar la probabilidad posterior de una hipótesis a medida que más información o más evidencia se vuelve disponible y poder utilizarla como previa $p(T)$ para la siguiente aplicación en el conjunto de datos.

Ello vuelve necesario apelar a métodos de *inferencia variacional* para resolver el problema del cálculo de la posterior en distribuciones que forman parte de familias exponenciales y que, por ello, no pueden dominarse totalmente con métodos frecuentistas[9]. Los métodos de inferencia variacional generalizan, para un conjunto de datos que no se puede dominar en su totalidad, los métodos de inferencia de la posterior, o *MAP*, un método de optimización que sustituye la necesidad de computar toda la evidencia (una tarea imposible) y conocer todos los parámetros para los puntos del espacio de datos[10][11].

La LDA o adjudicación latente de Dirichlet ofrece una alternativa al problema planteado por distintas iteraciones de la previa que se alejan de la evidencia ofrecida por el conjunto de datos, proponiendo una técnica matemática, la distribución de Dirichlet, que calcula *distribuciones de probabilidad sobre distribuciones de probabilidad posibles*. La LDA aplica este tipo de técnica matemática a la búsqueda de tópicos latentes.

La propiedad matemática que define a la LDA como modelo que computa estimados de probabilidad posterior y los utiliza como previas en la siguiente iteración es su capacidad para *conjugar* esas posteriores como previas. Cuando nos vamos alejando de la evidencia disponible y ya no conocemos cómo están distribuidas las probabilidades de los tópicos en los documentos, la capacidad de la LDA para conjugar las previas o distribuciones de probabilidad en familias exponenciales permite a este tipo de modelo proponer parámetros posibles, lo cual es importante si se quiere evitar que se modele como una distribución gaussiana lo que es una distribución multinomial, por ejemplo, una mezcla de gaussianas. La previa conjugada permite que la posterior de la cual deriva sea del *mismo tipo* de distribución de probabilidad que esta última.

[9] BLEI, David. *Variational Inference*. [En línea] 2011. https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf.
[10] BLEI, David, KUCUKELBIR, Alp y MCAULIFFE, Jon. *Variational Inference: A Review for Statisticians*. arXiv.org. [En línea] 11 de Mayo de 2018. https://arxiv.org/pdf/1601.00670.pdf. 1601.00670v4 [stat.CO].
[11] BLEI, David, *Variational Inference: Foundations and Innovations*. You Tube. [En línea] 2017. https://www.youtube.com/watch?v=Dv86zdWjJKQ&list=PLjc-tqOYn1C73S5-g1fE4EksnMrzsaIYb&index=9.

De este modo, la LDA es un modelo de inferencia bayesiana que no se limita a la optimización de la probabilidad condicional o *MLE* de los parámetros del modelo, como sucede con el PLSA, sino que incorpora también un cálculo de la posterior (*MAP*), la probabilidad predictiva que funcionará como probabilidad previa en iteraciones sucesivas, *para distintas distribuciones de probabilidad de tópicos por documentos*. Para ello, incorpora al cálculo de la probabilidad, además de los parámetros que vimos en el modelo anterior, un nuevo parámetro α, también llamado un "hiperparámetro", el cual agrega una dimensión adicional a la tarea de cálculo de la distribución de probabilidad de tópicos por corpus al calcular *una distribución de probabilidad sobre distribuciones posibles de probabilidad* de tópicos sobre documentos dados en el corpus. El hiperparámetro α ofrece la posibilidad de calcular distribuciones de probabilidad sobre previas conjugadas como distribuciones multinomiales, el tipo de cálculo de probabilidad que caracteriza a los conjuntos de datos de texto multidimensionales.

3. Resultados

Los resultados obtenidos revelan las concepciones y definiciones semánticas implícitas que subyacen al uso que hizo cada uno de los usuarios cuyos tweets fueron recuperados para estos conjuntos de datos de texto tomados de Twitter. En un sentido importante, los tweets son una suerte de "encuesta de opinión" que revela lo que este grupo de usuarios de Twitter en particular piensa del concepto "derechos": cuáles son las asociaciones, implicaciones, sentidos sugeridos y semántica implícita en el uso de este concepto particular.

Desde un punto de vista general, las aplicaciones de los algoritmos en los tres conjuntos de datos arrojaron dos tipos de resultados claramente discernibles: 1) por una parte, ofrecen una visión reducida de los tipos de problemas, preocupaciones o dimensiones que los usuarios de Twitter tienen en el momento de emitir sus tweets. 2) Por la otra, ofrecen una visión general de la semántica que define el uso del término "derechos", y que es transversal a las preocupaciones o dimensiones que definen los tópicos o dimensiones reducidas, en los usuarios que emiten sus tweets en el momento de la captura.

Estos dos tipos de resultados no se ofrecen con independencia el uno del otro, sino que ambos constituyen el rendimiento esperado de la aplicación de un algoritmo de modelado de tópicos sobre un conjunto determinado de datos de texto. La distinción es, por tanto, más bien

analítica. Ambos permiten al investigador dos tipos de tareas diferentes, dependiendo de su foco de interés.

Puede, por ejemplo, en primer lugar, reconstruir los tópicos que caracterizaron una determinada transmisión de tweets. En los tres conjuntos de datos se pudo observar que el tema de Cataluña es prominente para los tres conjuntos de datos, incluso estando separados por varios días. Igualmente importantes en los tres conjuntos de datos son las referencias a la situación política de Venezuela para aquel momento, la situación política de Argentina para el mismo período y distintas referencias a crímenes, derechos de minorías y situaciones sucedidas a activistas políticos en la región. No obstante, la preeminencia del tema de Cataluña resulta notable, habida cuenta de que la aspiración a la independencia de una nación no es, desde el punto de vista filosófico, un problema relacionado con la defensa de los derechos humanos.

Por esta razón, es importante observar que las aplicaciones de modelado de tópicos rinden un segundo resultado analíticamente distinto al que se acaba de mencionar. Se trata de la reconstrucción de la semántica de un término, en este caso, "derechos". La presente investigación ha arrojado un contexto semántico enormemente rico para la comprensión de la palabra "derechos" en el ámbito Iberoamericano.

En efecto, alrededor de la palabra "derechos" se tejen sentidos de una gran diversidad. Se asocia, como en su significado convencional, a los derechos humanos individuales. Pero también al derecho de autodeterminación de los ciudadanos de una nación (Cataluña y España), a los derechos económicos y sociales (una extensión de los derechos que se desarrolla en los años 60 del siglo XX como una prologanción de la Declaración de los Derechos Universales de 1948, y que se vincula con la filosofía política de la izquierda latinoamericana). Pudimos observar que este último sentido es muy importante para los tuiteros del ámbito argentino. La palabra "derechos" está muy vinculada también al derecho a la vida y a la seguridad en nuestra región. La semántica de la palabra se encuentra también asociada a la idea de "gozar de libertades", en particular en relación con actores como las mujeres, la comunidad LGBT (comunidad lesbiana, gay, bisexual y transexual) y los defensores de los derechos de los animales, un sentido muy reciente de la comprensión del derecho que se ha incorporado cómodamente al ámbito comunicativo de la región. La idea del derecho a la libertad también es prominente en los tweets de usuarios vinculados con Venezuela. También resulta muy importante su vinculación con el

"derecho a la vida", un sentido sofisticado desde el punto de vista filosófico, que fue central para los autores de la Carta de los Derechos de la ONU en 1948 y cuya presencia en los tópicos obtenidos evidencia cómo los conceptos e ideario filosófico de la mencionada Carta se han permeado al mundo comunicativo cotidiano o a la cultura política de la nación.

También en un sentido general pudo observarse la presencia activa de usuarios de Twitter preocupados por el tema de los "derechos" en distintos países de América Latina, llamando la atención la relativa ausencia, para el momento de la captura, de tweets significativos de países como Paraguay, Uruguay y Ecuador. De Centro América el país más activo es México.

3.1. Interpretación del primer conjuto de datos de texto (muestra A):

La primera muestra fue tomada la mañana del 24 de septiembre de 2017, una semana antes del pautado Referendo Independentista de Cataluña. Pudo observarse en los conceptos arrojados por el LSA (Tabla 1) que uno de los usuarios más activos durante el período de tiempo en el que se recuperaron los tweets (de 8:27 am a 8:57 am, hora local), fue Albert Rivera, (o su "community manager"), el líder del partido de centro derecha español Ciudadanos, quien repudió en distintos tweets, que eran retuiteados repetidamente, un ataque a una sede de diputados anti-independentistas. El repudio a ese ataque, pues, queda asociado en el conjunto de datos a la palabra "derechos", cuya semántica se redefine en ese momento como defensa de libertades políticas.

Aspectos de la salida del algoritmo del LSA. Concepto "derechos"	
Países/Nombres	albert_rivera, kikomonasterio, Avila
Instituciones	amnistía, defensor, guardia, policía, presidente, procurador, sindicato
Verbos	acosa, atacado, apoyo
Sustantivos	acuerdo, amenaza, antiespecie, defensa, democracia, diputados, golpistas, humanos, libertades, sedes, veganismo
Adjetivos/Adverbios	no, todo, pero

Tabla 1: Semántica LSA de la muestra A.

Similarmente, se obtuvieron los siguientes resultados para el algoritmo de PLSA:

Aspectos de la salida del algoritmo del PLSA. Concepto "derechos"	
Países/Nombres	albert rivera, Barcelona, Bogotá, Cataluña, El Salvador, España, Guatemala, Miranda, Pablo Iglesias, Soria, Venezuela
Instituciones	amnistía, defensor, guardia, policía, presidente, procurador, sindicato
Verbos	abstenerse, acosar, apoyar, atacar, callar, ceder, deben, defender, destruir, disfrutar, excederse, expresar, expresar, garantizar, mirando, permitir, queremos, reclamar, reprimir, respetar, reventar, robar, roben, rodeaban, romper, sentir, ver, violar, votar, vulnerar (total: 30)
Sustantivos	abstención, abusadores, acuerdo, amenaza, asamblea, cambio, chicas, ciudadano, claro, compañeros, constituyente, cristiano, defensa, democracia, derechos, diputados, fiscal, foto, fundamentales, futuro, gente, golpistas, guerra, humanos, impuestos, independencia, joven, laborales, ley, libertades, lucha, manifestación, mujer, mujeres, noche, oposición, pacto, patriotas, personas, políticos, promulgar, pueblos, régimen, renuncia, sedes, sentido común, soviets, violencia, voto, vulnerar (total: 51)
Adjetivos/Adverbios	demoledor, entendido, estable, hoy, importar, necesario, nuevo, orgulloso, plenamente, popular, prioridad, todo

Tabla 2: Semántica PLSA de la muestra A.

La Tabla 2, que expresa la salida del algoritmo de análisis semántico de índole probabilística, PLSA, arroja resultados mucho más diversificados y ricos que los obtenidos durante la aplicación del método de factorización de matrices LSA. Pudieron identificarse 7 menciones a países diferentes, dos españoles (España y Cataluña) y el resto latinoamericano, entre ellos Venezuela.

También se mencionaron 7 instituciones públicas relevantes en la semántica del concepto de "derechos" y 30 verbos y 51 sustantivos asociados a dicha semántica.

Mucho más elocuentes son los resultados de la Tabla 3, tabla que recoge los resultados del algoritmo LDA, entre los cuales podemos reconocer 35 verbos y 68 sustantivos asociados a la

semántica del concepto de "derechos" en el mismo conjunto de datos. Como observaremos también en el segundo conjunto de datos, los resultados del algoritmo LDA son mucho más precisos en el sentido de que son más diversos y "elocuentes" en la recuperación de la semántica de la palabra que estamos examinando.

Para el filósofo o científico social que trabaja sobre la semántica del concepto de "derechos" los resultados no revelan mayores sorpresas, aunque sí llama la atención que la comprensión del concepto en el ámbito del lenguaje ordinario, en el "mundo de la vida" (para utilizar la expresión clásica de la fenomenología que se refiere a las comunicaciones cotidianas), incorpora rápidamente los desarrollos académicos más sofisticados.

Por ejemplo, la mención a los "derechos de los animales", un desarrollo teórico de la ética aplicada relativamente reciente en el mundo hispanohablante[12] se ha incorporado rápidamente al lenguaje del "planeta Twitter". Aquí es importante tener presente que el usuario de esa red social no necesita más que un texto breve de pocos caracteres para explicar el sentido de la expresión "derecho de los animales", con lo que está seguro de que su frase no despertará extrañeza o perplejidad: un claro indicativo que el sentido de la expresión "derechos de los animales" ya ha "bajado" al lenguaje ordinario en el mundo de la vida.

También es interesante para el teórico social observar que la semántica asociada a la Declaración de los Derechos Económicos y Sociales de 1966 (bastante posterior a la Declaración de la ONU sobre Derechos Humanos de 1948), también se ha incorporado a la semántica de los "derechos", con términos como "laborales", "industria", "impuesto", "trabajo", todos alejados de la semántica tradicional en torno a los Derechos Humanos Universales.

Como dato curioso, la Declaración de los Derechos Humanos del Islam, promulgada en el Cairo en 1990 en un intento por ofrecer una contrapartida no liberal a la filosofía de los DDHH occidentales, se encuentra totalmente ausente de la semántica de los conjuntos de datos, limitándose a la mención de "musulmanes" e "inmigrantes" toda referencia a los contenidos asociados al Islam. Esto pudiera ser un indicativo de un relativo fracaso de los promotores de esa declaración en publicitar sus contenidos más allá de los países con mayoría musulmana, así como

[12] MOSTERÍN, Jesús, *¡Vivan los animales!*, Madrid, Debate, 1998.

la pervivencia para nuestra región de la semántica que ha definido los DDHH en el Occidente moderno a lo largo de los últimos siglos.

Aspectos de la salida del algoritmo del LDA. Concepto "derechos"	
Países/Nombres	Albert Rivera, Argentina, Barcelona, Cataluña, Colombia, El Salvador, España, kirschnerismo, México, Rajoy, sur Bogotá, Venezuela, Zaragoza
Instituciones	autoridades, defensor, estado, estatales, gobierno, govern, instituciones, ministro, Onu, podemos, policia, procurador
Verbos	acostumbra, adoctrinar, advertir, agitar, apoyar, asfixiar, brindar, callar, cambiar, castigar, conocer, convocar, defender, enviar, exceder, expresión, firman, garantizar, imponer, incluir, iniciar, invertir, manifestar manipular, perder, pisotear, proteger, provocar, reclamar recortar, reventar, roben, violar, votar, vulnerar (total:35)
Sustantivos	abstencionista, abusadores, activistas, acuerdo, amenaza, animales, bandera, ciudadano, ciudades, civiles, comentario, constitución, constitucional, consulta, contenidos, convivencia, cristiano, culpa, democracia, derecha, diputados, documentación, ejercicio, experiencia, femenino, golpe, humanos, idea, implicancia, impuestos, independencia, industria, joven, justicia, laboral, laborales legalidad, libertad/es, lucha, lugar, marcha, matriz, mensaje, milicos, muchachos, mujeres, obligaciones, odio, oposición, patriotas, persona, planeta, privilegios,querella, radicales, reforma, régimen, rehenes, represión, retornados, sociales, soviets, trabajo, vicios, violencia, violencia, voto, xenofobia (total: 68)
Adjetivos/Adverbios	cada momento, civismo, criminal, demoledor, destruido, facista, hoy, imposible, mejor, mientras, millones, mismos, nazis, nueva, nunca, orgullosa, público, reaccionaria, realmente, siempre, también, tampoco.

Tabla 3: Semántica LDA de la muestra A.

Para el teórico social también pudiera ser interesante identificar y etiquetar constelaciones de tópicos. Como ejemplo de este posible etiquetado, tomaremos, en el primer conjunto de datos de texto, la salida del algoritmo con los resultados más "elocuentes", el LDA.

A la par de los contenidos convencionalmente asociados al concepto de derecho, tales como libertad, democracia, Constitución, etc., encontramos también las siguientes posibles constelaciones de tópicos. No se presupone un orden de importancia:

1. Independencia de Cataluña. Tópicos: 2, 6, 13, 14, 18, 19, 20, 24, 26, 28, 29.

2. Derechos laborales: 1, 27 (¿Venezuela?)

3. Régimen criminal: 3 (¿Venezuela?)

4. Derechos de las mujeres: 4, 23, 25 (¿Argentina?)

5. Represión, lucha ciudadana: 7 (¿Bogotá?)

6. Derechos de los animales: 10

7. Constitución, congreso mexicano: 12

3.2. Interpretación del segundo conjunto de datos de texto (muestra B):

En la Tabla 4 se puede observar los resultados del LSA para el segundo conjunto de datos de texto, que constaba, como se recordará, de 3415 documentos.

Aspectos de la salida del algoritmo del LSA. Concepto "derechos"	
Países/Nombres	Cataluña, España
Instituciones	Onu
Verbos	debe, violando, violar
Sustantivos	constitución, expertos, fundamentales, gobierno, humanos individuales, medidas, referendum, violaciones
Adjetivos/Adverbios	----

Tabla 4: Semántica LSA de la muestra B.

El algoritmo PLSA arroja 18 verbos y 48 sustantivos alrededor de la semántica del término "derechos".

No obstante, aquí las ventajas y el rendimiento en elocuencia del algoritmo LDA son palpables. Se puede observar que la atribución latente de Dirichlet (Tabla 6) ofrece una visión mucho más amplia y precisa de la semántica implítica en el segundo conjunto de datos de texto,

arrojando un léxico de casi el doble de palabras asociadas al concepto inicial "derechos": 84 vs. 48. También es capaz de ofrecer un contexto más rico para este léxico, al invocar un número mayor de instituciones, adjetivos y adverbios asociadas a las palabras.

Aspectos de la salida del algoritmo del PLSA. Concepto "derechos"	
Países/Nombres	Argentina, Arreaza, Barcelona, Bolivia, Carlos Slim, Cataluña Colombia, Cristina Kirschner, Cuba, Dios, Ecuador, España Estados Unidos, México, Mundo, Rajoy, Venezuela
Instituciones	gobierno, musulmanes, ONU, Televisa, guardia civil
Verbos	abusar, advertir, anteponer, avisar, cambiar, colisionar, defender, denunciar, exigir, garantizar, limitar, merecer parecer, proteger, ratificar, tener, transmitir, violar (total: 18)
Sustantivos	adulto mayor, asamblea, aviso, bandera, cadenas pres., ciudadanos, concepto, constitución, cristiano, decadencia democracia, derecho, derechos animales, desaparición, dreamers, estrategia, fundamentales, grieta, honestidad, humanos, individuales, inmigrantes, izquierda, joven, ley, libertades, marcha animales, mujeres, obligaciones, pago deuda, partidos, personas, prioridad, pueblo, reeleccion indefinida, referendum, regresión, relatores, represión, sexuales, silencio, sociedad, urnas, utilidades, verdad, vida digna, violaciones, vulneración (total: 48)
Adjetivos/Adverbios	ahora, impecable, sagrada, siempre

Tabla 5: Semántica PLSA de la muestra B.

Aspectos de la salida del algoritmo del LDA. Concepto "derechos"	
Países/Nombres	Argentina, Arreaza, Carlos Slim, Cataluña, Colombia, Cuba, EEUU, España, MBachelet, Mundo, Venezuela
Instituciones	Ceofanb, colegio, consejo, gobierno, guardia civil, Hispania ONU, organismos, Partido Popular, Televisa, UniSalamanca
Verbos	anteponer, atentar, avanzar, avisar, colisiona, defender, denunciar, exigir, ganar, ilegalizar, luchar, merecer, piropear, proteger, rechazar, recortar, tomar, usar, violar votar (total: 20)
Sustantivos	adulto mayor, amenazas, animales, autor, bandera, ciudadanos, civil, colectivo, concepto, constitución, contrato, corrupción, cristiano, cuerpos, decadencia democracia, denuncia, derecha, derechos civiles, deshonra, discriminación, diversidad, estudiantes, fallecimiento, fiscal, frente, fundamentales, garantes, gente, grieta, hijos, humanos, igualdad, individuales, individuales, instituciones, izquierda, joven, ley, libertad, luchadora, maniobra, marcha, medidas, memoria, miedo, mujer, nacional, obligaciones, online, pago deuda, partido, patria, personas, piquete, presidencial, profesionales, profesora, propiedad, propuesta, protección, pueblo, racismo, referendum, registro, relatores, represores, respeto, salida, salud, sanciones, selección, servicios, sexual, silencio, solidaridad, trabajo, urnas, verdadera, víctima, victimario, vida digna, violaciones, vulneración (total: 84)
Adjetivos/Adverbios	abierto, absurdo, cientos, contra, frontal, justo, laborales, obligados, prioridad, siempre, total, urgente

Tabla 6: Semántica LDA de la muestra B.

Si, de nuevo, identificamos constelaciones de palabras para discernir tópicos más amplios o generales, encontraremos los siguientes, tomando en cuenta la salida del algoritmo más elocuente, el LDA.

En efecto, tomando en cuenta la salida de este algoritmo, para el segundo conjunto de tweets es posible identificar los siguientes tópicos asociados al concepto de "derechos", al lado de los más convencionales tales como libertad, ciudadanía, estudiantes, democracia, etc., sin presuponer un orden de importancia:

1. Cadenas presidenciales: tópico 1 (¿Venezuela?)

2. Derechos laborales: 2, 7, 9 (¿Argentina?)

3. Diversidad sexual: 3, 8, 14 (¿México?)

4. Derechos de la mujer: 4, 12 (¿España?)

5. Derechos humanos violados por sanciones de EEUU: 5, 25 (Venezuela).

6. Derecho a la independencia de Cataluña. De lejos, la semántica más prevalente para este conjunto de datos: 6, 10, 15, 16, 17, 18, 19, 20, 24, 26.

7. Derechos animales: 22 (¿España?)

8. Vida digna, adulto mayor: 27 (Ecuador)

3.3. Interpretación del tercer conjuto de datos de texto (muestra C):

En la Tabla 7 se observan los resultados para el tercer conjunto de datos de texto, que constaba de 2020 documentos y 20881 atributos, de la aplicación del algoritmo para el LSA.

Aspectos de la salida del algoritmo del LSA. Concepto "derechos"	
Países/Nombres	Cataluña, España, Estrasburgo, Murcia
Instituciones	Caritas, Pontifex
Verbos	Luchan
Sustantivos	bandera, constitucional, derechos, gente, humanos, palos, patria, represión
Adjetivos/Adverbios	"Pasando de algo"

Tabla 7: Semántica LSA de la muestra C.

Por su parte, el algoritmo PLSA arroja 44 verbos y 53 sustantivos alrededor de la semántica del término "derechos".

Sin embargo, como se ha observado en los otros conjuntos de datos de texto, aquí las ventajas y el rendimiento en elocuencia del algoritmo LDA son también importantes. Se puede constatar que la atribución latente de Dirichlet (Tabla 9) ofrece un mejor escrutinio de la semántica implítica en el tercer conjunto de datos de texto, arrojando un léxico de más del doble de palabras asociadas al concepto inicial "derechos": 129 vs. 53. También ofrece un contexto más

rico para este léxico por su número mayor de instituciones, adjetivos y adverbios asociadas a las palabras.

Aspectos de la salida del algoritmo del PLSA. Concepto "derechos"	
Países/Nombres	Argentinos, Barcelona, Catalanes, Cataluña, Colombia, España, Estrasburgo, Lenin, LilianTintori, Luis E. Rondón, Macri, Mexicanos, Murcia, Pablo Iglesias, Santander,Tumaco Unes_Táchira
Instituciones	congreso, gobierno, guardia civil, ONU, TSJ
Verbos	amo, arrastrar, asegurar, coaccionar, defender, dejar denunciar, desaparecer, desnucar, devolver, egresar, evitar, fundar, garantiza, habéis, hablar, humillar impedir, independizarnos, jugando, legislar, luchan, mentir, ocuparse, ocurrir, paguen, proteger, querer, quitar, reclamar, reconocer, recordar, recortar, renuncie, restringir, seguir, somos, tenes, vamos, visto, vivir, votar, vulnerar (total: 44)
Sustantivos	animales, asociación, autoritario, bandera, cabeza, comida, consecuentes, constitución, contexto, crisis, democracia, desprotección, dignidad, esperanza, estado, expediente, fascistas, feminismo, futbolistas, futuro, gente, humanitaria, humanos, integridad, justicia, laborales, libertad, libertades, marcha, masacre, mayoría, miedo, muerte, mujeres, mundo, nacionalidad, naturaleza, octubre, palos, personas, profesionales, pueblo, rector tsj, represión respeto, sentencia, sociales, techo, tierra, trabajo, unidad, vida, violaciones (total: 53)
Adjetivos/Adverbios	fracasado, fundamentales, mayores, mejor, nuevos, pocos, siempre, también, todos, vasta

Tabla 8: Semántica PLSA de la muestra C.

Aspectos de la salida del algoritmo del LDA. Concepto "derechos"	
Países/Nombres	Albert Rivera, Argentina, Argentinos, Bachelet, Barcelona, Botero, Catalana, Colombia, Estrasburgo, Lenin, Lilian Tintori, Luis E. Rondón, Macri, Mexicanos, México Murcia, Santos, ThaysPeñalver, Trump, Tumaco Unes_Táchira
Instituciones	CNE, congreso, gobierno, guardia civil, MUD, Netflix ONU, TSJ
Verbos	aprovechar, arrastrada, arrebatar, asegurar, cambiar, coaccionar, comiendo, comprendamos, conocer, cuentan, debemos, defendiendo, dejar, denunciando, desaparecer, desnuquen, devolverle, discuten, egresan, entrenando, esperar, exigen, fortalecen, fundan, garantizar, gobierna, habeis, hablar, hacer, humillar, imagino, impedir, independizarnos, invitar, joder, legislar, lograste, luchando, mentir, ocurre, paguen, perdés, pisoteando, podremos, proteger, puede, queremos, quiero, quitar, realizamos, reclamar, reconoce, recortar, renuncie, respetar, restituir, restringir, seguimos, sentimos, sigamos, somos, tenemos, tenes, vamos, vamos, visto, viviendo, vivir, volar, vota, voy (total: 71)
Sustantivos	aborto, activista, adultos, agenda, animales, arte, asociaciones, campesino, cancelaciones, capitalismo, carro, caza, ciuadadanos, clases, colaborador, comida, comunidad, consecuentes, constitución, construcción, corrupción, crisis, deber, declaración, defensa, defensor, defensores, democracia, desprotección, dictadura, dictaduras, dignidad, docentes, domingo, ejemplo encuentro, escuelas, esperanza, expediente, fascistas, feminismo, fiesta, foro, funcionarios, fútbol, futbolista, futbolistas, futuro, gente, golpe, graduandos, guerra, hijos, huelga, humanos, igualdad, imagen, imperialistas, individuo, injusticia, instituciones, integridad, invencible, jubilaciones, jugadores, laborales, leyes, lgbt, libertad, libertades, lucha, mano, marcha, masacre, mayoría, miedo, muerte, mujer, mujeres, mundo, nacionalidad, naturaleza, nazis, noticia, octubre, ofensiva, oficial, organización, país, partido, paz, periodista, personas, plan, plumazo, poder, policía, presencia, primera, privilegios, profesionales público, pueblo, religión, reputación, respeto, retroceso, rey, ricos, robo, rumbo, sacrificio, salud, sentencia, sicario, sociales, sociedad, soldado, techo, tirano, trabajadores, trabajo, trato, unidad, vía, vicios, violación,

	violaciones, vulneración (total: 129)
Adjetivos/Adverbios	ahora, animalista, autoritario, civil, claro, conmovedora, constitucional, desarmado, fracasada, frecuentemente, gran, hoy, humanitaria, iguales, imposible, injustas, inspiracional, libre, mala, mayores, mejor, muchos, mundial, nada, nuevos, peor, plenamente, pocos, rápido, siempre, social, solo, todavía, todos

Tabla 9: Semántica LDA de la muestra C.

Aquí es posible también identificar, desde el punto de vista del análisis filosófico, palabras alrededor de tópicos posibles. Tomando en cuenta los resultados más elocuentes que se han obtenido, aquellos arrojados por el algoritmo LDA, se puede ahora señalar algunos de ellos:

1. Defensa de derechos laborales, marcha (¿Colombia y México?), comida, techo, ricos, obra de arte inspiradora, Argentina, vivir, aborto, paz, esperanza, derechos sociales, salvaje, joder, pisotear, capitalismo, iguales, asustar, retroceso, docentes, cancelar jubilaciones, Navarro, ¿Venezuela?, lucha, bandera, huelga, guerra, 0, 3, 5, 8, 10, 16, 23.
2. Restitución de derechos, TSJ, Venezuela, ley, golpe, México: fortalecer justicia en instituciones, derechos campesinos (¿Venezuela y Argentina?), Chile: derechos adultos mayores, democracia, sociedad, cambiar el rumbo, mexicanos, libertades, miedo, desaparecer, futbolista, arrebatarnos, quitarnos, humanos, masacre, expediente, respeto, violaciones, derecho, desprotección, luchar, humanos, libertad, hijos, privilegios, agenda, leyes 1, 4, 5, 13, 19, 25, 27.
3. Derechos de las mujeres, feminismo, Cataluña, mujer, libertades, defender, evitar, desnucar, arrastrar, 2, 22.
4. Argentina: contra imperialistas, nazis, fascitas, autoritario, Macri, reconocimientos, mayoría, gobierno, libertades, libertad, dictadura, crisis humanitaria, defender, libertad, Venezuela 5, 11, 21, 26.
5. Cataluña: amenaza a la nacionalidad, constitución, trabajo, coaccionado, humillado, mentir, Murcia, luchando, gente, policía, España, represión, renunciar, poder, imagen, Pablo Iglesias, Barcelona, defiene, constitución 7, 17, 18, 27, 28.
6. Derechos de minorías: Fiesta animalista, invita, Venezuela, Aragua, Ciudadanos lgbt, defensores, salud, integridad, libertades, 14, 20.

4. Conclusiones

El presente artículo ha estudiado tres técnicas de modelado de tópicos o reducción de la dimensionalidad en conjuntos de datos de textos que, desde el punto de vista filosófico y teórico, suponen la recuperación de la semántica de matrices dispersas de bolsas de palabras. Estas tres técnicas se desarrollan en el marco de dos paradigmas generales que son muy importantes en la literatura del análisis estadístico de datos: el paradigma frecuentista y el paradigma bayesiano, el cual lidia con la incertidumbre sobre cantidades desconocidas.

Se expuso que, de las tres técnicas de modelado de tópicos que se estudiaron, las dos primeras, el LSA y el PLSA, ofrecen un análisis estadístico del conjunto de datos que se apoya en un conteo preciso de las frecuencias de términos en documentos al interior de conjuntos de datos acotados.

En este sentido, la primera de las técnicas de modelado de tópicos de tipo frecuentista, el LSA, rinde una factorización de la matriz que la descompone en valores singulares, los cuales se convierten en los parámetros para la agrupación de nuevos documentos y términos en los tópicos que ha identificado el sistema. La segunda técnica, el PLSA, por su parte, rinde otro tipo de parámetros: aquellos que definen una distribución de probabilidad para tópicos de los términos y documentos que se incorporan en el sistema.

Finalmente, la tercera de las técnicas, la LDA, se desprende del paradigma frecuentista anteriormente descrito en la medida en que es considerada por la literatura una técnica de inferencia estadística sobre cantidades desconocidas. La LDA debe apelar al cálculo de la distribución de probabilidad posterior en conjuntos de datos cuyos parámetros van "desapareciendo" por pérdida de la posibilidad de un dominio de la distribución de la probabilidad marginal, a causa del crecimiento exponencial de los datos.

Por esta razón, la LDA se aleja del paradigma frecuentista y representa un intento de calcular probabilidades en condiciones de *incertidumbre*. En este sentido, permite un cálculo de probabilidades sobre distribuciones de probabilidad a través de una técnica matemática que conjuga la previa en distribuciones multinomiales, de modo que se conserve a lo largo de todas las iteraciones la familia exponencial a la que pertenece la distribución. Paralelamente, los

modelos de LDA proponen distintos métodos para facilitar la inferencia variacional de los datos de entrada y, con ello, la creación de modelos con una mayor capacidad de generalización.

Al capturar los tres conjuntos de datos de texto, la palabra clave o "filtro" de la aplicación de Twitter ha sido la palabra "derechos". Los documentos de Twitter o tweets se eligieron, entonces, conforme a si mencionaban o no, en el texto, esa palabra. La intuición que guió el calibrado de los distintos programas de código que se usaron para la aplicación de Twitter ha sido que los tweets que mencionaban la palabra "derechos", al ver reducida su dimensionalidad con las tres técnicas que se estudiaron, arrojarían un contexto conceptual o una serie de tópicos característicos que podían iluminar la semántica implícita o latente en el uso de la palabra "derechos" por parte de los usuarios de los tweets capturados en "streaming" o tiempo real, todos provenientes del ámbito Iberoamericano.

Los resultados analizados demuestran que esa intuición, apoyada por la literatura sobre el tema, es correcta y que los documentos estudiados realmente ofrecen interesantes hallazgos respecto al significado del concepto de "derechos" para los usuarios de la red social Twitter en el momento de la captura. En efecto:

1. Un primer hallazgo ha sido la coincidencia del significado semántico de la palabra "derechos" con los resultados ofrecidos en los distintos estudios de este tipo de conceptos que se encuentran en el ámbito de la reflexión filosófica de carácter analítico, el tipo de análisis que caracteriza el desempeño profesional de los filósofos y que persigue una comprensión racional del concepto. Esta coincidencia sugiere que muchos de los aspectos teóricos más importantes que se encuentran en los esfuerzos de comprensión de la noción de "derechos" han sido exitosamente incorporados a la cultura política de la región. En este sentido, la tarea educativa de las Naciones Unidas pudiera calificarse, en un primer nivel, como una tarea lograda: la aguda conciencia, por parte de aquellos que reclaman sus derechos en nuestros países de habla hispana, respecto de lo que involucra realmente apelar a los derechos, constituye, en mi opinión, uno de los hallazgos más importantes del presente trabajo.

El discurso sobre el derecho entendido como derecho a la vida, la libertad, la seguridad, la libertad de expresión, pero también el derecho a una vida digna, seguridad laboral, condiciones de vida razonables (vivienda y bienestar), derecho de los animales, derechos de las mujeres, pero

también de minorías, etc., se ofrece de modo elocuente a partir de los modelos estudiados y abren al profesional de la filosofía y de las ciencias sociales, que quiere fundamentar su análisis con evidencia empírica, un campo de trabajo amplio e importante.

Las herramientas que se han estudiado, pues, iluminan la semántica que define el uso de muchos conceptos filosóficos, y de otras disciplinas cuyos conceptos se expresan en el lenguaje ordinario, y lo hacen de maneras inéditas hasta hace relativamente poco tiempo.

2. El segundo hallazgo ha sido que, al contrario de lo que podría pensarse, el significado semántico de la noción de "derechos", tal y como es usado en contextos de lenguaje ordinario al interior de estos tres conjuntos de documentos de Twitter, no revela ninguna novedad respecto de lo que ordinariamente se entiende por "derechos" en el discurso filosófico especializado. Probablemente una exploración más profunda y con conjuntos mayores pudiera revelar novedades en la semántica del concepto en el futuro. Pero no ha sido así para este estudio. De este modo, aunque la semántica reconstruida por los tres tipos de algoritmos ha sido rica y diversificada, no es particularmente novedosa y se mantiene en un nivel elevado de convencionalidad.[13] Por esta razón, es posible afirmar que el análisis de contenidos semánticos latentes que ofrece el modelado de tópicos automatizado es una suerte de pequeña encuesta de opinión sobre el estado actual de la opinión pública respecto de un concepto.

3. En tercer lugar, fue posible constatar, de acuerdo con Crain y Zhou (6), que la LDA tiende a aprender tópicos más generales o más amplios que el PLSA, lo que los hace tal vez más difusos. Ello tiene las siguientes consecuencias: que los modelos de LDA, según se pudo apreciar también, son muy buenos para analizar transversalmente la semántica de un concepto (lo que era la intención inicial del presente trabajo), pero menos buenos si se trata de identificar cuáles son los tópicos que realmente caracterizan a un conjunto de datos. De este modo, se pudo notar que resulta más fácil identificar esos distintos tópicos en los modelos de PLSA que en los modelos de LDA, mientras que, inversamente, es más nítida la semántica de un término en los modelos de LDA que en los modelos de PLSA.

De este modo, es posible que un periodista, un científico social o un investigador en redes sociales que quiera averiguar de qué se está hablando, de qué temas o tópicos se ocupa la gente

[13] Se usa aquí el concepto de lenguaje moral convencional en el mismo sentido que Lawrence Kohlberg (22).

en un momento dado, obtenga mejores resultados con un modelo de PLSA. Pero, por contraste, un filósofo que quiera averiguar qué entiende la gente por un determinado concepto que se usa de manera transversal en tópicos a lo largo de distintos documentos, tal vez obtenga mejores resultados o resultados más precisos con un modelo de LDA. Por esta razón, aquí se han privilegiado los resultados arrojados por el algoritmo de LDA para el análisis hermenéutico de la semántica encontrada.

4. Un cuarto hallazgo es que, si bien los modelos de LSA arrojan resultados relativamente confiables, los modelos frecuentistas y de inferencia Bayesiana que apelan al cálculo de probabilidades representan una ventaja enorme respecto de los primeros, que apelan al cálculo de valores singulares en matrices factorizadas. Tal vez por esta razón, los modelos probabilísticos son cada vez más populares en la literatura de las ciencias de la computación, quedando como aspecto polémico en la discusión sobre la futura prevalencia de estos modelos el peso de la subjetividad del investigador que distribuye la probabilidad previa, un tipo de subjetividad típico de la inferencia bayesiana.