Revista Española de Lingüística

PROCESAMIENTO DE LENGUAJE NATURAL APLICADO A DATOS MASIVOS GENERADOS EN MEDIOS SOCIALES

Jordi Porta Zamorano[1] y José Luis Sancho Sánchez[2]
Centro de Estudios de la Real Academia Española

Resumen

La aparición y auge de la comunicación canalizada digitalmente, especialmente de las llamadas redes sociales, reclama capacidades analíticas automatizadas para extraer información y patrones a partir de datos masivos baja o pobremente estructurados con el objetivo de predecir tendencias, acciones y eventos futuros. Este ámbito concita el interés de investigadores y empresas, con implicaciones para la lingüística, la informática, la psicología, las ciencias sociales o la estadística, entre otras áreas.

Palabras clave: Procesamiento de Lenguaje Natural; datos masivos; medios sociales; análisis de sentimiento.

NATURAL LANGUAGE PROCESSING APPLIED TO BIG DATA IN SOCIAL NETWORKS

Abstract

The advent and rise of computer mediated communication, mainly of so-called social networks, ask for automated analytical capabilities to extract information and patterns from massive poorly structured data in order to anticipate future trends, events and actions. This area attracts both researchers and industries, with Linguistics, Computer Science, Psychology, Social Sciences or Statistics involved, among others.

Keywords: Natural Language Processing; big data; social media; sentiment analysis.

Recibido: 18/02/2019 Aprobado: 28/10/2019

1. porta@rae.es; ORCID ID: https://orcid.org/0000-0001-5620-4916
2. sancho@rae.es; ORCID ID: https://orcid.org/0000-0002-2319-8641

DOI: https://doi.org/10.31810/RSEL.51.2.7
RSEL 51/2 · 2021 · pp. 111-124 · ISSN 0210-1874 · e-ISSN 2254-8769

1. INTRODUCCIÓN

La fórmula «redes» o «medios sociales» hace referencia a un amplio rango de plataformas accesibles a través de la web mediante las que sus usuarios pueden crear y compartir contenidos e interactuar entre ellos. Pueden clasificarse según diferentes criterios: las destinadas a compartir contenido multimodal (Facebook, Twitter, Youtube, MySpace, LinkedIn, Vimeo, etc.); foros (StackOverflow, CNET, Apple Support, etc.), generalmente acotados temáticamente, en los que los usuarios publican información especializada, debaten asuntos o formulan preguntas y reciben respuestas; blogs (Gizmodo, Boing Boing o Microsiervos en español) orientados a contenidos de interés particular o microblogs (Twitter, Sina Weibo, Tumblr), limitados a textos cortos para compartir información y opiniones. Este modelo de comunicación canalizada digitalmente se extiende también a la relación entre consumidores y proveedores de productos, bienes o servicios (plataformas de atención al cliente o portales de reseñas de productos).

El número de usuarios de estos medios sociales ha aumentado en varios miles de millones en los últimos ocho años. Se estima que el 45 % de la población mundial está en línea, o que el 76 % de la población en EE. UU. participa en redes. Facebook, Twitter, Pinterest e Instagram son actualmente las plataformas que acogen más usuarios, que de manera creciente acceden a ellas mediante dispositivos móviles (se estima que el 52 % de los accesos a la web proviene de estos dispositivos, con picos de hasta un 95 % en el caso de Facebook)[3]. En estas plataformas los usuarios generan una gran cantidad de contenido, incluido el lingüístico, en una rica variedad de escenarios sociolingüísticos (relacionados con la formalidad, multilingüismo, identidad o filiación, etc.) que proporcionan datos que exhiben las llamadas tres «uves» de los macrodatos (*BigData*): volumen, velocidad y variedad[4].

La relación entre los datos masivos y la lingüística se establece en dos direcciones. En una de ellas, el Procesamiento del Lenguaje Natural (PLN, en español; Natural Language Processing, NLP, en inglés) contribuye al procesamiento y análisis de los datos como se verá más adelante. En la otra, que no se abordará aquí, estos datos proporcionan nuevos medios de indagación y averiguación sobre la facultad del lenguaje. Se han publicado aportaciones en fonología (Eisenstein, 2013), dialectología (Donoso y Sánchez, 2017), sociolingüística y variación (resumidas en Nguyen

3. *https://www.statista.com/topics/1164/social-networks/*

4. Características definitorias (originalmente) de los macrodatos: gran cantidad, heterogeneidad e inmediatez.

y otros, 2016), lexicografía computacional (Wang y otros, 2012b) o pragmática conversacional (Danescu-Niculescu-Mizil y otros, 2011), entre otros ámbitos. Obviamente, las relaciones son bidireccionales: el trabajo de Wang y otros (2012b) hizo evolucionar una ontología del léxico emocional mediante *hashtags* (etiquetas descriptivas) de Twitter que, a su vez, mejoró un clasificador de emociones; los hallazgos en pragmática conversacional, por su parte, permiten adecuar al interlocutor los agentes conversacionales (*chatbots* en inglés: sistemas informáticos que «mantienen» conversaciones con humanos) o los sistemas de recomendación.

Este artículo pretende ofrecer algunos ejemplos, sin ánimo de exhaustividad, en los que el PLN contribuye a la resolución de alguna tarea de tratamiento o análisis de datos masivos. La existencia, por una parte, de un programa científico, sumamente excitante y cuyas implicaciones probablemente trasciendan lo tecnológico hacia lo social, que puede llegar a «humanizar» la interacción con dispositivos y sistemas informáticos (asistentes virtuales, interfaces de voz, traducción automática, interrogación en lenguaje natural a bases de datos, etc.) y, por otra, la acumulación apabullante de datos que propicia lo digital, necesitados de tratamiento, en parte lingüístico, para ser de alguna utilidad, otorgan al PLN y a la lingüística plena vigencia y justifican la oportunidad de este resumen.

El artículo arranca con un somero análisis de algunas características del lenguaje canalizado digitalmente; continúa con una introducción al PLN para adentrarse después en algunas de sus aplicaciones, antes de cerrarse con una conclusión y la bibliografía citada.

2. EL LENGUAJE DE LAS REDES SOCIALES

En los medios de comunicación tradicionales, como los periódicos, la televisión, la radio o las revistas científicas, la comunicación es unidireccional. Las noticias periodísticas o los artículos científicos han sido el material de investigación en PLN durante los últimos veinticinco años. El objetivo principal del PLN ha sido la extracción de la intención comunicativa usando el conocimiento generado por la informática, la inteligencia artificial y la lingüística. Los textos contenidos en medios sociales, cuyos autores no son escritores profesionales, tienen a menudo una naturaleza informal, cercana al habla real, y muestran un registro en el que abundan recursos expresivos como abreviaturas, alargamientos, reduplicaciones o contracciones, léxico jergal y vulgar o señaladores de implicación emocional, como intensificadores y emoticonos. Con frecuencia se mezclan o alternan lenguas o variedades, y no suelen observarse las convenciones tipográficas, ortográficas

o gramaticales (entre otras razones porque algunos dispositivos de escritura no favorecen su observancia). Incluso discursivamente pueden observarse idiosincrasias estructurales, como el «mensaje», la «publicación» (*post*) o el «hilo» (*thread*), cuya relación con las categorías tradicionales de enunciado o turno no es evidente (Androutsopoulos, 2013). En cuanto a la carga informativa, mientras que los medios de comunicación escritos tradicionales se esfuerzan en presentar una información factual, neutra y objetiva, los medios sociales se caracterizan por una información subjetiva con una carga importante de opinión. La deriva temática es mucho más prominente en medios sociales por la naturaleza conversacional de algunas plataformas y el flujo continuo de información generado. Además, en tanto que herramientas sociales, estas plataformas posibilitan la construcción y visibilización de identidades y comunidades, con un comportamiento lingüístico específico: se han observado (Nguyen y otros, 2016) correlaciones entre diferencias en textos según su función (felicitaciones, informes, juego de palabras, alertas, etc.) y las variables sociolingüísticas clásicas (sexo, edad, localización, grupo social o cultural).

Todas estas características acercan los contenidos de los medios sociales a la expresión de una corriente de consciencia más que a un pensamiento cuidadoso meticulosamente editado, como el esperado en los medios de comunicación escritos tradicionales. Algunas de ellas pueden parecer rupturistas y generan alarma entre los puristas, preocupados por una eventual degradación del lenguaje. Sin embargo, algunos de los mecanismos expresivos definitorios de los medios sociales (logogramas, pictogramas, acrónimos, inicialismos, abreviaturas y acortamientos, alargamientos expresivos, omisión de letras, uso de variantes ortográficas heterodoxas) no lo son en exclusiva. Tal como afirma Crystal (2008), ninguno de los rasgos de la escritura en medios sociales parece especialmente novedoso ni incomprensible, aunque es posible encontrar algunas veces combinaciones poco frecuentes fuera de los medios sociales: *ijc2sailuvu* → *I just call to say I love you* → forma plena + dos inicialismos + logograma + acortamiento con equivalencia fonética + forma plena + ortografía heterodoxa + logograma. Con todo, el uso de etiquetas (*hashtags*) e hipervínculos (URL) y la multimodalidad caracterizan intrínsecamente los contenidos sociales.

3. PROCESAMIENTO DE LENGUAJE NATURAL

Se puede pensar en el PLN como un conjunto de técnicas computacionales para representar y procesar automáticamente el lenguaje humano. El PLN dista todavía de alcanzar el nivel de «maestría» que tienen los humanos, ya que el manejo del

lenguaje requiere capacidades simbólicas de alto nivel, entre las que se incluyen la manipulación de estructuras de constituyentes recursivas; la creación y propagación de vínculos dinámicos; la adquisición y acceso a memorias léxicas, semánticas y episódicas; la representación de conceptos abstractos; la manipulación de las convenciones semióticas (creatividad, prevaricación), etc.

Las primeras aproximaciones al procesamiento del lenguaje fueron simbólicas (basadas en reglas, lógica y ontologías), pero el pobre rendimiento de los sistemas pioneros, especialmente orientados a la traducción automática, llevó a la introducción de un enfoque algorítmico estadístico en los ochenta al que se le sumaron en los noventa los algoritmos de aprendizaje automático y más recientemente el aprendizaje profundo (*deep learning*, un conjunto de técnicas computacionales de análisis de datos que pretenden dotar a los sistemas informáticos de capacidad de generalización), aunque no han faltado intentos de hibridación (Klavans y Resnik, 1996).

El PLN ha progresado considerablemente en las últimas décadas. Se han desarrollado recursos como corpus anotados, lexicones y gramáticas electrónicas para un número importante de lenguas, y algoritmos para procesar grandes cantidades de texto y producir resultados prácticos. Hay toda una familia de técnicas usuales en PLN, algunas con notable madurez, que se aplican al tratamiento de los datos sociales en función de la tarea y el objetivo: segmentación en unidades de análisis, lematización o reducción morfológica y categorización, tratamiento de palabras ausentes de los repertorios léxicos, eliminación de términos irrelevantes (palabras de baja frecuencia, palabras vacías o *stopwords*, términos que se descartan durante el procesamiento), representación léxico-semántica del contenido (TF-IDF, Term Frequency – Inverse Document Frequency, la medida de la relevancia de un término en un documento), representación vectorial de la distribución (representación de los documentos mediante matrices de frecuencias de aparición de sus términos), análisis semántico latente (LSA, técnica estadística que permite analizar las relaciones entre documentos y términos contenidos en los documentos, produciendo un conjunto de conceptos relacionados, etc.), identificación de lengua o dialecto, detección de entidades nombradas, sumarización o traducción automática.

Las herramientas tradicionales de PLN presuponen homogeneidad y adecuación a la norma en el material que procesan, de modo que los rasgos del contenido de los medios sociales vistos en la sección anterior constituyen ruido que impacta negativamente en su rendimiento: la ausencia o inconsistencia en el uso de la puntuación o las mayúsculas dificultan la detección de los límites oracionales, incluso a los humanos; los emoticonos, la laxitud o la creatividad ortográfica, el

uso intensivo de abreviaciones, las desviaciones morfofonológicas (reduplicaciones, alargamientos o acortamientos expresivos, sufijación heterodoxa, etc.), o el multilingüismo complican la segmentación de palabras y el análisis morfosintáctico. En las plataformas que limitan la extensión de los mensajes (Twitter, por ejemplo), la brevedad, con la consecuente limitación del contexto, supone una dificultad añadida. Inicialmente, la estrategia para procesar los contenidos sociales fue adaptarlos a lo que esperaban las herramientas, generalmente normalizándolos. En la actualidad, es más habitual modificar el diseño de las herramientas, no ya para que tengan la robustez necesaria para manejar la especificidad de estos contenidos, sino para que las exploten en favor de su cometido: p. ej., Gimpel y otros (2011) declaran una mejora del 5 % en la tasa de acierto de un etiquetador morfosintáctico adaptado a los tuits.

4. APLICACIONES

La aparición y auge creciente de los medios sociales ha propiciado la creación de muchas aplicaciones sobre sus contenidos: las necesidades analíticas en los sectores de la industria, medios de comunicación y periodismo, salud, política, seguridad, etc. requieren capacidades de procesamiento de grandes cantidades de datos heterogéneos. La industria está interesada en mejorar su inteligencia de negocio, imagen corporativa, compromiso del cliente, mejora de los servicios al cliente, *marketing* en línea, predicción del mercado de valores, recomendaciones de productos, o reputación. Los medios de comunicación tienen en las redes sociales sus mayores fuentes de información (25 %), donde recopilar opiniones y analizar el sentimiento social. En política, la minería de opinión o percepción permite obtener una mejor idea de la realidad sobre determinadas cuestiones y mejorar posiciones. En el ámbito de la salud puede haber interés en las discusiones sobre las recomendaciones de proveedores y consumidores, dolencias y enfermedades. Los foros sobre medicina surgen de la necesidad de los pacientes de discutir y expresar sus sentimientos y experiencias. También son interesantes las aplicaciones para la monitorización de enfermedades y su propagación. En seguridad, el análisis de las redes sociales permite comprender situaciones y analizar la opinión de grupos de individuos con intereses compartidos y generar alertas sobre potenciales amenazas a la seguridad pública, gestionar emergencias y catástrofes o asuntos más mundanos como monitorizar el tráfico urbano. El comportamiento social agregado puede dar información a nivel nacional.

Colectivamente, este conjunto de técnicas aplicadas a tareas concretas recibe el nombre de analítica predictiva, y su objetivo es extraer información y patrones de los datos para predecir tendencias, acciones y eventos futuros sobre la base de los datos históricos. Las predicciones se basan en percepciones señalizadas por el sentimiento, la emoción, el volumen o cantidad de datos y su tema. A continuación, se exponen aisladamente algunas aplicaciones orientadas a detectar unitariamente algunas de esas señales, pero debe tenerse en cuenta que con frecuencia se combinan para resolver tareas de mayor complejidad (p. ej., Daniulaityte y otros (2015) clasifican tuits sobre drogas –identificación de contenido– en función del tipo de usuario –perfil de autor–) y que los rasgos relevantes en diferentes tareas son codependientes (Barbieri (2008) determinó que la implicación emocional y el posicionamiento covarían con la edad del autor).

4.1. *Análisis de sentimiento*

Desde una perspectiva de las ciencias sociales, las interacciones informales dan pistas acerca de la opinión pública sobre temas variados: canciones, blogs, discursos presidenciales, etc. Las opiniones guían la toma de decisiones. El análisis de sentimiento u opinión tiene como objetivo identificar la polaridad de un documento, bien sea de forma discreta (positivo, negativo o neutro) o continua (p. ej., en una escala de 1–5 en reseñas de películas). Ni siquiera los humanos coinciden siempre, por lo que la tarea es difícil para algunos algoritmos, especialmente cuando el texto es corto. En una variante de esta tarea, se intenta conectar la opinión con la entidad que la suscita o alguna de sus características (peso de un móvil, limpieza de un hotel, etc.). Los microposts como los de Twitter plantean mayor dificultad, ya que a las características genéricas de los contenidos sociales señaladas en el apartado 2, suman el hecho de que buena parte de la carga nocional recaiga en *hashtags* y, a menudo, se recurra a una importante dosis de ironía y sarcasmo, cuyo tratamiento automático es todavía pobre (véase Sección 4.3).

4.1. *Análisis de emoción*

Las emociones no son entidades lingüísticas, pero se expresan mediante el lenguaje e influyen en la manera de comunicarnos. La interpretación de la emoción de un texto la multimodalidad se explota pobremente (Calvo y otros, 2010) puede ser subjetiva, y la relación simbólica entre lenguaje y emoción puede estar modulada por la ideología u otros aspectos culturales que se pueden inferir mejor a partir del

análisis cuantitativo del uso que de los índices léxicos. Igualmente, el género textual, el momento temporal o las características sociolingüísticas del autor intervienen en el análisis de la emoción o los estados de ánimo (*mood*), más transitorios, cuya clasificación también se ha abordado.

Teóricamente, las emociones se han clasificado en conjuntos discretos o espacios dimensionales, aunque ningún modelo parece superior a otro y la elección del soporte teórico suele hacerse por razones estratégicas. Las técnicas de detección de emociones pueden estar basadas en recursos construidos para este fin (WordNetAffect o ANEW (Affective Norms for English Words), LIWC (Linguistic Inquiry and Word Count) u ontologías como EmotiNet (Balahur y otros, 2011) o EmojiNet (Wijeratne y otros, 2017) o en aprendizaje automático a partir de ejemplos.

4.1. *Detección de ironía y sarcasmo*

La ironía y el sarcasmo son dos figuras retóricas que se asientan sobre una característica al parecer exclusiva del lenguaje humano: la prevaricación. La diferencia entre ironía y sarcasmo es sutil y se acostumbra a considerar el sarcasmo un tipo de ironía expresada generalmente de forma agresiva y ofensiva. Los sistemas que detectan estas figuras usan principalmente técnicas de aprendizaje automático y se entrenan con ejemplos anotados para este fin. Como estas figuras no tienen ninguna estructura lingüística particular que permita identificarlas fácilmente, se usan características superficiales, como la presencia de signos de puntuación, emoticonos y mayúsculas; la polaridad de las palabras; presencia de intensificadores e inversores de polaridad; y la presencia de determinados marcadores textuales añadidos por los usuarios para comunicar su intención en ausencia de otros signos paralingüísticos. Algunos sistemas hacen también uso de información contextual, como el foro en el que se ha publicado el texto. La identificación de la ironía y el sarcasmo incide en la detección de sentimiento, ya que actúan como inversores de polaridad, o en la detección de abuso: un intercambio sarcástico entre usuarios conocidos no es sospechoso, a diferencia del que se produce entre desconocidos (Xu y otros, 2012).

4.5. *Detección de personalidad*

La personalidad es una combinación de patrones estables de comportamiento, emoción, motivación y pensamiento, que puede ser descrita por tipos o rasgos. Los textos pueden reflejar varios rasgos de la personalidad de su autor y su detección automática tiene varias aplicaciones importantes. Algunos sistemas pueden

hacer recomendaciones de productos y servicios evaluados positivamente por otros usuarios con el mismo tipo de personalidad. En salud mental, ciertos diagnósticos están correlacionados con rasgos de personalidad. En recursos humanos, la personalidad de un candidato es importante para determinar su adecuación a un trabajo. El modelado de personalidad dominante se basa en los conocidos como cinco grandes, los siguientes rasgos binarios: extroversión, neuroticismo, cordialidad (*agreeableness*), responsabilidad (*conscientiousness*) y apertura a la experiencia. Un experimento reciente (Majumder y otros, 2017), utilizando aprendizaje automático y teniendo en cuenta rasgos lingüísticos (como recuentos de palabras, longitud media de la frase, representaciones distribucionales del léxico, etc.), obtuvo tasas de acierto de entre un 56% y un 62 % en cada uno de los cinco rasgos de personalidad mencionados, superiores en los cinco casos al estado del arte.

4.4. *Detección de posicionamiento*

La detección de posicionamiento tiene por objetivo determinar si un texto es favorable, desfavorable, neutro o no relacionado con respecto a una entidad o a un tema de discusión, normalmente controvertido y que no siempre se menciona explícitamente en el texto que se analiza (un texto que apoye los derechos del feto estará en contra del aborto, aunque no lo mencione). La detección de posicionamiento (*stance detection*) ofrece información complementaria al análisis de sentimiento, ya que se ocupa de la perspectiva evaluativa del autor, con independencia de si su texto es positivo, negativo o neutro. Los primeros resultados compartidos sobre esta tarea fueron presentados por Mohammad y otros (2016) sobre un conjunto de tuits sobre ateísmo, cambio climático, movimiento feminista, Hillary Clinton y la legalización del aborto, anotados manualmente con una tasa de acuerdo entre anotadores del 73,1 %. El mejor de los sistemas obtuvo una puntuación media de 67,83 %. La detección de posicionamiento ha sido aplicada a temas políticos debatidos en medios sociales, como la independencia de Cataluña (Bosco y otros, 2016; Taulé y otros, 2017), el apoyo a Hillary Clinton y a Donald Trump durante las elecciones presidenciales estadounidenses de 2016 (Lai y otros, 2017) o el Brexit (Grčar y otros, 2017).

4.1. *Rumores y veracidad (posverdad)*

Un rumor es una historia en circulación de veracidad cuestionable, aparentemente creíble pero difícil de verificar, que produce suficiente escepticismo o preocupación

para motivar el descubrimiento de la verdad (Zubiaga y otros, 2015). Los rumores abundan en los medios sociales y las afirmaciones falsas afectan a la percepción que las personas tienen sobre un determinado tema y a su comportamiento, a veces de manera nociva, contribuyendo a la desinformación y la manipulación de la opinión pública. La creciente dependencia de los medios sociales como fuentes de información y de noticias hace que el impacto de los rumores sea alto e influyente (Hermida, 2012). Analizar y determinar la veracidad de un contenido es una tarea de gran interés y dificultad en la que se aplican técnicas de PLN (Dale, 2017). Las técnicas para el análisis de rumores se fundamentan habitualmente en el posicionamiento que se adopta sobre el rumor: afirmación, negación, o cuestionamiento. Uno de los estudios más influyentes en análisis de rumores en Twitter es el de Mendoza y otros (2010), en el que se analizaron los rumores sobre la erupción de un volcán y las alertas de tsunami surgidos durante el terremoto de Chile de 2010. Demostró que los falsos rumores son negados en un 50 % de los casos, mientras que los rumores ciertos solo se negaron el 0,3 %. Los porcentajes obtenidos por agregación en dichos posicionamientos permiten detectar la veracidad de los rumores. Varol y otros (2017) estudiaron características estadísticas de las redes de usuarios como predictores de material patrocinado o promovido deliberadamente o emitido por robots.

4.7. *Perfil de autor*

A diferencia de la detección de autoría, que trata de determinar, de entre un conjunto de autores candidatos, cuál es el autor de un texto anónimo dado, el perfilado de autor aspira a discriminar entre clases de autores, en función de algunas características como su edad, sexo, lengua nativa o tipo de personalidad, usando para ello las características estilísticas y de contenido de un texto (Argamon y otros, 2009). Esta tarea es cada vez más importante en aplicaciones muy diversas como el análisis lingüístico forense, el análisis de tendencias, la identificación de relaciones de poder, la seguridad o el *marketing*, especialmente en medios sociales donde el número de autores anónimos es muy alto. Desde el punto de vista forense, p. ej., la posibilidad de determinar el perfil lingüístico del autor de un texto malicioso únicamente mediante el análisis del texto permite contrastar sospechosos. Igualmente, desde la perspectiva del *marketing*, las empresas pueden tener interés en conocer qué conjuntos de personas aprueban o reprueban sus productos mediante el análisis de blogs o reseñas. Por poner un ejemplo, Schler y otros (2006) alcanzaron

una tasa de acierto del 80 % para la identificación del sexo y del 75 % para la de la edad en autores de blogs.

Menos atención ha recibido el problema complementario de difuminar el estilo de un autor, esto es, de generalizar un texto mediante la ocultación de los rasgos particulares de su autor, aprendidos a partir del análisis de un conjunto de sus textos. La dificultad no estriba solo en la dilución de los rasgos de autor, sino que el texto resultante debe ser gramaticalmente correcto y nocionalmente equivalente.

4.8. *Detección de eventos y temas (topic)*

Detectar el tema de un texto, es decir, determinar de qué trata, es otra de las tareas habituales de clasificación textual. Se utiliza para estructurar y organizar mensajes, frecuentemente en entornos corporativos: agrupar la comunicación con los clientes en función de su asunto u organizar artículos periodísticos por su temática, por ejemplo. En el ámbito de los medios sociales, se utiliza para comprender qué subtemas asociados con un evento o circunstancia y qué aspectos de estos atraen mayoritariamente la atención de los usuarios. Ha sido combinado con éxito con otras tareas, como la detección de opinión (Chen y otros (2012) asignan la polaridad en función del argumento: *predecible* es *malo* para una obra literaria o cinematográfica, pero *bueno* para el mercado de valores), clasificación textual (Hong y otros (2010) obtuvieron mejoras en la predicción de popularidad y en la clasificación de tuits en categorías temáticas) o la predicción de delitos (Wang y otros, 2012a).

4.9. *Identificación y clasificación de lenguaje ofensivo y expresiones de odio*

Muchos usuarios de medios sociales aprovechan el anonimato para usar un lenguaje ofensivo. Las expresiones de odio conllevan denostar a una persona o un grupo a propósito de alguna característica racial, de género, de orientación sexual, o nacionalidad, entre otras. Pero existe además un trasfondo cultural que hace que la percepción de la ofensa sea subjetiva. Las comunidades en línea y las plataformas sociales están interesadas en prevenir comportamientos abusivos en sus medios y en evitar las expresiones de odio. También los gobiernos pueden monitorizar comunidades para intentar anticiparse a posibles incidentes de violencia racial, ataques terroristas u otros crímenes. Desde el punto de vista del análisis de sentimiento, las expresiones de odio son negativas. Los experimentos que han abordado esta tarea han obtenido mejores resultados combinando varias de las tareas detalladas hasta ahora (análisis lingüístico del contenido, detección de emoción o intención) con

el análisis de otras características de la comunicación digital, como los metadatos disponibles sobre los usuarios, el contexto comunicativo, la topología de la red de usuarios o la dinámica evolutiva de la comunidad que los acoge (Xu y otros, 2012).

4.10. *Detección del propósito (intent)*

La clasificación textual se usa también en entornos corporativos para determinar automáticamente el propósito de las comunicaciones con los clientes, en ocasiones con el objetivo de analizar los productos o servicios o para automatizar operativas. Especial interés tiene identificar qué clientes van a dejar una compañía, ya que los costes de captación de nuevos clientes exceden los costes de retención, por lo que la identificación del riesgo de deserción o baja suele ser la primera fase de las campañas de retención de clientes (Huang y otros, 2012). Este problema ha sido estudiado sobre los registros de llamadas (*call detail record* o CDR) en el contexto de redes sociales de compañías de telecomunicaciones (Verbeke y otros, 2014), en juegos en línea (Kawale y otros, 2009) y comunidades en chats, foros y microblogs (Amiri y otros, 2015).

5. CONCLUSIÓN

Este artículo ha enumerado algunas de las características de la comunicación canalizada digitalmente y cómo el PLN aborda e incluso explota esas características para contribuir a la resolución de algunas tareas de tratamiento y análisis de datos masivos producidos en la red. Igualmente, la pertinencia del PLN y la lingüística en este ámbito ha quedado patente, así como las nuevas posibilidades de investigación sobre la facultad del lenguaje que auspician la cantidad y naturaleza de los datos masivos.

REFERENCIAS BIBLIOGRÁFICAS

Amiri, H. y Daume III, H. (2015): «Target-dependent churn classification in microblogs», en *Proceedings of the 29th AAAI Conference on AI*, pp. 2361-2367.

Androutsopoulos, J. (2013): «Online data collection», en Mallinson, C., Childs, B. y Van Herk, G. (eds.) *Data Collection in Sociolinguistics: Methods and Applications*, Routledge, pp. 236-249.

Argamon, S., Koppel, M., Pennebaker, J. W. y Schler, J. (2009): «Automatically profiling the author of an anonymous text», *Communications of the ACM* 52, pp. 119-123.

Balahur, A., Hermida, J. M., Montoyo, A. y Muñoz, R. (2011): «EmotiNet: a knowledge base for emotion detection in text built on the appraisal theories», en *Proceedings of*

the 16th international conference on NLP and information systems, Alicante, pp. 27-39.
Barbieri, F. (2008): «Patterns of age-based linguistic variation in American English», *Journal of Sociolinguistics* 12, pp. 58-88.
Bosco C., Lai M., Patti V., Rangel F. y Rosso P. (2016): «Tweeting in the debate about Catalan elections», en *Proceedings of the LREC Emotion and Sentiment Analysis Workshop*, Portorož, pp. 67-70.
Calvo, R. A. y D'Mello, S. (2010): «Affect detection: An interdisciplinary review of models, methods, and their applications», en *IEEE Transactions on Affective Computing* 1, pp. 18-37.
Chen, L., Org, C., Wang, W., Org, W., Nagarajan, M., Wang, S., Sheth, A. P. y Org, A. (2012): «Extracting diverse sentiment expressions with target-dependent polarity from Twitter», en *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, Dublín, pp. 50-57.
Crystal, D. (2008): *Txtng: The Gr8 Db8*, Oxford, Oxford University Press.
Dale, R. (2017): «NLP in a post-truth world», *Natural Language Engineering* 23, pp. 319-324.
Danescu-Niculescu-Mizil, C., Gamon, M. y Dumais, S. (2011): «Mark my words!: Linguistic style accommodation in social media», en *Proceedings of the 20th International Conference on WWW*, Hyderabad, pp. 745-754.
Daniulaityte, R., Nahhas, R. W., Wijeratne, S., Carlson, R. G., Lamy, F. R., Martins, S. S., Boyer, E. W., Smith, G. A. y Sheth, A. (2015): «'Time for dabs': Analyzing Twitter data on marijuana concentrates across the U.S. HHS Public Access», *Drug Alcohol Depend* 155, pp. 307-311.
Donoso, G. y Sanchez, D. (2017): «Dialectometric analysis of language variation in Twitter», en *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pp. 16-25.
Eisenstein, J. (2013): «Phonological factors in social media writing», en *Proceedings of the Workshop on Language Analysis in Social Media*, Atlanta, pp. 11-19.
Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. y Smith, N. A. (2011): «Part-of-speech tagging for Twitter: Annotation, features, and experiments», en *Proceedings of the ACL-2011*, Stroudsburg, pp. 42-47.
Grčar, M., Cherepnalkoski, D., Mozetič, I. y Novak, P. K. (2017): «Stance and influence of Twitter users regarding the Brexit referendum», *Computational Social Networks* 4.
Hermida, A. (2012): «Tweets and truth: Journalism as a discipline of collaborative verification», *Journalism Practice* 6, pp. 659-668.
Hong, L. y Davison, B. D. (2010): «Empirical study of topic modeling in Twitter», en *Proceedings of the 1st Workshop on Social Media Analytics*, Washington D.C., pp. 80-88.
Huang, B., Kechadi, M. T. y Buckley, B. (2012): «Customer churn prediction in telecommunications», *Expert Systems with Applications* 39, pp. 1414-1425.

Kawale, J., Pal, A. y Srivastava, J. (2009): «Churn prediction in MMORPGs: A social influence based approach», en *Proceedings of International Conference on Computational Science and Engineering*. IEEE Computer Society.

Klavans, L. y Resnik, P. (eds.) (1996): *The balancing act: Combining symbolic and statistical approaches to language*, Cambridge, MIT press.

Lai M., Hernández I., Patti V. y Rosso P. (2017): «Friends and enemies of Clinton and Trump: Using context for detecting stance in political tweets», en *Proceedings of the 15th Mexican ICAI*, pp. 155-168.

Majumder, N., Poria, S., Gelbukh, A. y Cambria, E. (2017): «Deep learning-based document modeling for personality detection from text», *IEEE Intelligent Systems* 32, pp. 74-79.

Mendoza, M., Poblete, B. y Castillo, C. (2010): «Twitter under crisis: Can we trust what we RT?», en *Proceedings of the 1st Workshop on Social Media Analytics*, Nueva York, pp. 71-79.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X. y Cherry, C. (2016): «Semeval-2016 task 6: Detecting stance in tweets», en *Proceedings of SemEval'16,* San Diego, pp. 31-41.

Nguyen D., Doğruöz, A. S., Rosé C. P. y de Jong, F. (2016): «Computational Sociolinguistics: A Survey», *Computational Linguistics* 42, pp. 537-593.

Schler, J., Koppel, M., Argamon, S. y Pennebaker, J. W. (2006): «Effects of age and gender on blogging», en *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, Stanford, pp. 199-205.

Taulé, M., Martí, M.A., Rangel F., Rosso M., Bosco C. y Patti, V. (2017): «Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017», en *Notebook Papers of the 2nd SEPLN IBEREVAL Workshop*, Murcia, pp. 157-177.

Varol, O., Ferrara, E., Menczer, F. y Flammini, A. (2017): «Early detection of promoted campaigns on social media», en *EPJ Data Science*, 6.

Verbeke, W., Martens, D. y Baesens, B. (2014): «Social network analysis for customer churn prediction», en *Applied Soft Computing* 14, pp. 431-446.

Wang, W., Chen, L., Thirunarayan, K. y Sheth, A. P. (2012a): «Harnessing Twitter 'BigData' for Automatic Emotion Identification», en *IEEE International Conference on Social Computing,* Amsterdam, pp. 587-592.

Wang, X., Gerber, M. S. y Brown, D. E. (2012b): «Automatic crime prediction using events extracted from Twitter posts», en *Proceedings of the International conference on social computing, behavioral-cultural modeling, and prediction,* College Park, pp. 231-238.

Wijeratne, S., Balasuriya, L., Sheth, A. y Doran, D. (2017): «EmojiNet: An open service and API for Emoji Sense Discovery», en *Proceedings of the ICWSM-2017*.

Xu, J.-M., Jun, K.-S., Zhu, X. y Bellmore, A. (2012): «Learning from Bullying Traces in Social Media», en *Proceedings of the 2012 Conference of the North American Chapter of the ACL: Human Language Technologies*, Montreal, pp. 656-666.

Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K. y Tolmie, P. (2015): «Towards detecting rumours in social media», en *Proceedings of the AAAI Workshop on AI for Cities,* pp. 35-41.