La opinión en textos con un enfoque interdisciplinar: propuesta de contexto y léxico ad hoc

Silvana Grazia Temesio Vizoso*

Artículo recibido:
25 de octubre de 2017
Artículo aceptado:
27 de agosto de 2018

Artículo de revisión

Resumen

El artículo aborda la cuestión de cómo determinar la polaridad u opinión en textos, particularmente en Twitter. El análisis se hace desde múltiples perspectivas: ética de la información, comunicación, ciencia de la información e informática, enfatizando en dos análisis que se complementan: el procesamiento de lenguaje natural y la organización del conocimiento.
Se desarrolla un prototipo que propone anotar las palabras cargadas de opinión como los adjetivos y los adverbios contextualizadas por su núcleo: sustantivos y verbos respectivamente. Estos núcleos sirven también para modelar la terminología del dominio y se propone la creación de un lexicón *ad hoc*.

* Facultad de Información y Comunicación, Universidad de la República
Montevideo, Uruguay silvana.temesio@fic.edu.uy

Palabras clave: Twitter; Análisis de Sentimientos; PLN; Ciencia de la Información

A context and ad hoc lexicon proposal to an interdisciplinary approach to opinion analysis

Silvana Grazia-Temesio Vizoso

Abstract

This paper addresses the question of how to determine polarity of opinion among texts and, more specifically, those published on Twitter. The analysis is performed from multiple perspectives, including information science, and communication and information ethics with emphasis on the analysis of natural language processing and knowledge organization.

A prototype is developed that proposes annotation of opinion conveying words as adjectives and adverbs contextualized by grammatical core, i.e., nouns and verbs, respectively. These nuclei also serve to model the terminology of the domain, and the development an ad hoc lexicon is proposed.

Keywords: Twitter; Sentiment Analysis; PLN; Information Science

Introducción

El problema de la caracterización de los textos de acuerdo a los tópicos o asuntos es de importancia para lidiar con la gran cantidad de textos disponibles en la web. Una cuestión que se deriva de esto es la caracterización de esos textos en cuanto a la orientación de opinión.

La opinión es importante para medir la popularidad de un producto, una política, o cualquier otro asunto susceptible de ser analizado en forma subjetiva. Mientras que los tópicos se identifican por palabras, los sentimientos son más complejos de identificar, requieren mayor contexto, mayor comprensión del uso del lenguaje y detección de la ironía, entre otros aspectos. Según Liu (2010), la información textual puede ser categorizada en dos tipos principales: hechos y opiniones. Los hechos son expresiones objetivas acerca de entidades,

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

eventos y sus propiedades. Las opiniones son expresiones usualmente subjetivas que describen los sentimientos y valoraciones hacia las entidades o eventos y sus propiedades.

Una opinión se define por un conjunto de cinco elementos:

- *e*: es la entidad objetivo, que es el objeto sobre el que se emite la opinión; puede ser un producto, una persona, evento, tópico, etc.
- *a*: aspecto o característica de la entidad analizada. Existe una jerarquía de componentes o partes, subcomponentes, etc. *a* es un conjunto de atributos del objeto. El objeto puede ser representado por un árbol donde la raíz es el objeto y cada nodo un componente o subcomponente del objeto en una relación "parte de". Cada nodo se asocia a un conjunto de atributos. Por ejemplo, si hablamos de una entidad *e*=teléfono móvil, *a* puede ser la batería, el teclado, el sistema operativo, etc., es decir, una serie de características de la entidad.
- *so*: orientación del sentimiento (valencia).
- *h*: persona quien expresa la opinión (*holder*).
- *t*: momento en que se expresó la opinión.

Las redes sociales y en particular Twitter constituyen una fuente de información actualizada sobre los tópicos más diversos proporcionando la posibilidad de conocer la opinión acerca de éstos. El análisis de sentimientos o minería de opinión consiste en la clasificación del sentimiento de un texto respecto a su polaridad: positiva o negativa. Cuando el texto es la explicitación de un hecho, no tiene una carga de opinión, se considera neutro. Cuando hay una opinión se puede además realizar una gradación de intensidad de la polaridad: positivo, muy positivo, etc.

En este trabajo se abordan los desafíos del análisis de opinión en Twitter desde distintas ópticas: ética de la información, aspectos de comunicación, ciencia de la información e informática, con énfasis en el procesamiento de lenguaje natural (PLN) y la perspectiva de la organización del conocimiento.

La metodología empleada es la revisión y análisis del estado del arte desde una perspectiva multidimensional, reflexionando en torno a brindar mayor contexto a las palabras que expresan sentimientos. Esto se plasma en un prototipo donde se anotan las palabras cargadas de opinión como los adjetivos y los adverbios tomando en consideración sus núcleos: sustantivos y verbos respectivamente. Los núcleos modelan también la terminología del dominio por lo que son una herramienta para la creación de un lexicón *ad hoc* que capture el lenguaje usado.

Caso de estudio: Twitter

Twitter es una red social en la cual se puede escribir contenido con un límite de hasta 140 caracteres, aunque se ha considerado aumentar dicha extensión. Cada uno de estos textos cortos o tuits aparece en el perfil del autor con el criterio del más reciente primero. Este texto se puede reenviar (un retuit), con lo cual aparece en el usuario reenviante como un reenvío. Esta característica de reenvío permite evaluar la popularidad del tuit.

Los usuarios de Twitter se pueden suscribir al contenido que genera otro usuario, esto es una práctica muy común para seguir noticias suscribiéndose al canal de Twitter de un diario. La popularidad de los usuarios de Twitter se mide por la cantidad de "seguidores". Los tuits, además de los 140 caracteres, tienen una serie de elementos que se denominan entidades: URL, imágenes, menciones a un usuario (@usuario) y # (*hashtag*) o etiqueta que se utiliza para agrupar tuits.

El volumen de información generado en Twitter es muy grande, por lo cual no es factible un tratamiento manual y por tanto existen diversas líneas de investigación en el tratamiento automático del sentimiento. Muchos de los estudios de sentimiento se realizaron en textos largos donde se puede expresar una idea con un contexto suficiente. Respecto al estudio en textos muy cortos, como es el caso de Twitter, establecer el contexto es un desafío. Además, el vocabulario usado en Twitter tiene características especiales en cuanto a que no se restringe a un dominio y las formas agramaticales son frecuentes.

La restricción en la extensión produce transformaciones en la modalidad expresiva y muchas veces se genera una confusión en los conceptos, se distorsiona la ortografía llegando a crearse un nuevo estilo ortográfico, la gramática se diluye y se hace un uso importante de neologismos. Por otra parte, también se propicia una capacidad de síntesis necesaria para este tipo de comunicación. Más allá de toda opinión, es una práctica comunicativa instaurada y desde el punto de vista del tratamiento del lenguaje natural subyace la cuestión de si civilizar el discurso para analizarlo o generar una nueva forma de tratamiento, porque el uso de esta comunicación constituye una tendencia ya afincada.

Algunos aspectos característicos de los aspectos sintácticos de los tuits son:

- Signos de puntuación o caracteres especiales

Los signos de puntuación suelen faltar o ser usados incorrectamente. En general se preprocesan los tuits y con diferentes criterios se homogeneizan los textos. Este tema merece un estudio por sí mismo, porque si bien hay un uso incorrecto desde el punto de vista gramatical es una situación de uso y forma parte de una comunicación hipervinculada donde el texto en Twitter

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

corresponde a un contexto que suele estar en otra parte y es un comentario, un aviso o un llamador que constituye un enunciado no oracional o un titular.

- Uso importante de locuciones verbales o neologismos

El lenguaje es coloquial y se usan términos que no aparecen en el diccionario y locuciones verbales que no tienen significado literal (por ejemplo, "estar en la chiquita", "darle gas"). Estas locuciones tienen una componente de localidad muy acentuado.

- Referencia a otro usuario de Twitter a través del símbolo @.

La referencia puede estar dentro de un contexto gramatical:

> Habla Tania Da Rosa de @CAinfouy sobre la necesidad de empoderar a la unidad de acceso a la información pública **#UAIP2016**

o puede ser un elemento informativo vinculado fuera de la estructura gramatical:

> Hay documentos que no deberían conservarse El mundo va hacia el Archivo Único Electrónico #uaip2016 @**agesic** @DestinoUruguay @ mvdbureau

En este último caso –como información complementaria– suele ponerse al final o al principio del tuit. En cualquier caso el @ como etiqueta Part of Speech (PoS[1]) constituye un nombre.

- # Hashtag que se usa para agrupar tuits

Puede ser una palabra normal que se usa con el signo # para agrupar mensajes o ser una concatenación de palabras que pueden o no tener significado. Por ejemplo:

> Conoce la versión preliminar del Paquete de #**DatosAbiertos** contra la #corrupción

#corrupción no tiene un significado adicional, aunque también alude a la agrupación, pero desde el punto de vista sintáctico el # se puede eliminar o ignorar.

1 PoS (Part of Speech). Se trata de las categorías de palabras que tienen propiedades gramaticales similares: nombres, verbos, adjetivos, adverbios, preposiciones, entre otras.

> Llega a su final el Seminario de Acceso a la Información Pública #UAIP2016 ¡Los esperamos el próximo año!

En este caso #UAIP2016 designa un evento y agrupa todos los comentarios del evento, pero no participa de la oración; desde el punto de vista sintáctico se podría sacar sin que la oración pierda sentido.

Una estrategia bastante usada es eliminar el # si está al inicio o al final del tuit presuponiendo que se quiere etiquetarlo y que no hay en este caso un aporte sintáctico o semántico.

- Palabras cambiadas para dar énfasis: jiiiijiiii, muuuchas gracias, Goooooooooolazo de Peñarol Junior Arias aparece el goleador
- Emoticones. Existen tablas de clasificación de emoticones con los valores positivo, negativo o neutro.
- URL. En algunos casos se utilizan para contextualizar, usando el título de la URL
- Titulares o texto difícilmente procesable:

  #Fútbol #Especial #FENvsDEF Final Primer Tiempo @CAFenix_ 0 - 1 @DefensorSpClub #DaleDefe

Fronteras compartidas

El análisis de un tuit desde el punto de vista de opinión puede hacerse desde los aspectos informáticos, desde el procesamiento de lenguaje natural (PLN), desde la implicancia comunicativa de las opiniones en los medios y desde la organización del conocimiento entre otros enfoques.

Las redes sociales y el *microblogging* constituyen una fuente de información que permite monitorear la opinión sin necesidad de estudios más complejos, como las encuestas. Por otra parte el volumen de información que produce Twitter es muy alto, así como la cantidad de usuarios en todo el mundo, que incluyen personajes políticos, celebridades, deportistas, científicos, académicos, instituciones privadas y de gobierno. Los periodistas y políticos lo utilizan como un elemento de publicidad y marca personal y como en 140 caracteres no se puede profundizar muchas veces se usa como un titular y la información se amplía en una URL o se complementa con elementos gráficos.

Desde los aspectos de comunicación, puede analizarse la red como una topología donde es posible reconocer los aspectos de alcance y difusión a

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

través del análisis de grafos utilizando herramientas de análisis y visualización de redes como Gephi,[2] que desentrañan los papeles de los nodos en la red.

Aspectos éticos del análisis de textos

Los mensajes en Twitter constituyen una situación de fragmentación de la información en la que al descontextualizarla, aislarla de la coyuntura histórica y geográfica en la que se emite, se abren múltiples perspectivas de interpretación del fragmento, algunas de las cuales pueden no corresponderse con la intención del emisor.

En otros casos de recuperación de información, como en un catálogo bibliográfico o un fondo archivístico, se recupera desde una interrogación que es en sí misma un fragmento pero que está contextualizada tanto para el buscador como para el objeto informativo con la utilización de un elemento común que es el control terminológico bajo la forma, por ejemplo, de un tesauro. El objeto recuperado es generalmente una porción de información sin fragmentar y con un contexto embebido. Capurro (1987) comenta:

> Podemos considerar al proceso de almacenamiento y recuperación de la información bajo un punto de vista hermenéutico como la articulación de la relación entre la apertura existencial al mundo del interrogador, sus distintos horizontes de pre-comprensión abiertos y compartidos socialmente y el horizonte prefijado del sistema. El proceso de búsqueda de información es básicamente un proceso de interpretación que tiene que ver con el contexto vital y el trasfondo del interrogador y el de aquellos que almacenan diferentes tipos de expresiones lingüísticas que tienen un significado dentro de contextos de comprensión fijos como son un tesauro, palabras claves y esquemas de clasificación.

El análisis de sentimientos utiliza fragmentos (palabras de búsqueda) para recuperar otros fragmentos (mensajes de Twitter) y luego interpretar la inclinación emocional de estos últimos. Ese espacio de precomprensión que se utiliza para buscar los mensajes de Twitter a través de términos ya constituye un espacio de interpretación en sí mismo porque no funciona como ese puente preconstruido del catálogo bibliográfico.

En la segunda etapa, la interpretación del sentimiento del mensaje requiere a su vez ese contexto señalado para poder extraer la polaridad del emisor, porque en este caso no es una libre interpretación sino que se busca desentrañar el sentimiento del emisor.

2 https://gephi.org/ es el sitio de la herramienta de software libre Gephi que realiza exploración y visualización para grafos y redes.

Para fijar este grado de incertidumbre se diseña una estrategia diferente a la de la recuperación de la información en un catálogo bibliográfico. No se establece un contexto predefinido, ese espacio de precomprensión señalado como es el caso de un tesauro, sino que se aborda el contexto dentro del propio fragmento. Lo que se predica con una polaridad se predica sobre un objeto interno al mensaje, el adjetivo se aplica a un sustantivo, el adverbio se aplica a un verbo. Ese espacio de precomprensión se construye al vuelo en el momento de analizar, porque ese espacio de precomprensión es volátil y local. Se construye en un contexto histórico y geográfico específico y nada asegura que persista.

Por otra parte se abre un ámbito de discusión respecto a la utilización de este escudriñamiento de polaridad en un espacio que es público pero abordable con una intención comercial, política o de una índole que no fue prevista por el emisor pero que puede recursivamente incidir en el propio emisor para beneficio del que analiza la polaridad. Este análisis está inmerso en un panorama más vasto que corresponde a la utilización de la información como un instrumento de manipulación, control social y político tergiversando la propia información.

Enfoque informático

Sentiwordnet[3] es una lista de palabras con polaridad que establece un *ranking* de valores positivos y negativos que aplica a cada conjunto de sinónimos de Wordnet[4] en inglés. Se puede traducir o mapear al español a través del sitio http://multiwordnet.fbk.eu/english/home.php. Hay estudios (Montejo-Ráez *et al.*, 2014) que utilizan este recurso para tratar la orientación de los tuits.

También cabe mencionar el lexicón de polaridad de licencia libre[5] que fue desarrollado para TASS, que es un taller anual para análisis de sentimiento en español promovido por la Sociedad Española para el procesamiento del lenguaje natural (Urizar y Roncal, 2013).

Entre los métodos de aprendizaje computacional que usan conjuntos de entrenamiento en un aprendizaje supervisado pueden mencionarse SVM (Support Vector Machine), Naive Bayes y Máxima Entropía. Los resultados de estos métodos están en el entorno del 70 al 80 % de precisión (Pang, Lee y Vaithyanathan, 2002).

3 http://sentiwordnet.isti.cnr.it/ presenta un recurso léxico para minería de opinión. A cada conjunto de sinónimos de WordNet le asigna uno de tres valores: positivo, negativo, objetivo.

4 http://wordnet.princeton.edu/ alberga una base de datos léxica para el inglés que comprende nombres, verbos, adjetivos y adverbios. Se agrupan en conjuntos de sinónimos cada uno expresando un concepto diferente. Los conjuntos de sinónimos se vinculan por relaciones semántico-conceptuales y léxicas.

5 http://komunitatea.elhuyar.eus/ig/files/2013/10/ElhPolar_esV1.lex es el sitio del lexicón.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77.57863

Según señalan Vilares, Alonso y Gómez Rodríguez (2015), otros estudios utilizan n-gramas, etiquetas PoS y lexicones. También se ha estudiado la orientación semántica en adjetivos, verbos y adverbios construyéndose una ecuación lineal. Se ha señalado que los métodos de aprendizaje supervisado no tienen buen rendimiento debido a que Twitter utiliza una gran cantidad de palabras fuera de los vocabularios. También se exploraron las gramáticas de dependencias y se propusieron métodos híbridos que combinan el uso de bolsas de palabras, el aprendizaje supervisado y diferentes estrategias para PoS específicos. En todos los casos se busca capturar contexto para aumentar la adecuación del análisis. Específicamente estos autores proponen un tratamiento híbrido que utiliza la gramática de dependencias utilizando triplas (Head, etiqueta arco, dependiente).

Una manera de analizar sentimientos en forma no supervisada es a través de los métodos de lexicones. Los lexicones son diccionarios que incluyen la polaridad semántica de la palabra. Se pueden crear manualmente o en forma automática. Las investigaciones se han centrado principalmente en el uso de adjetivos como indicadores de la orientación semántica. Se extraen los adjetivos de un texto y se anotan con su valor semántico usando el diccionario de sentimientos y luego se adjudica una orientación semántica al texto.

Según Taboada *et al.* (2011) los clasificadores SVM (Support Vector Machine) que entrenan en un conjunto de datos concreto utilizando unigramas o bigramas consiguen una adecuación alta en detectar la polaridad de un texto. No obstante esa adecuación se pierde en gran medida al utilizarlos en otro dominio diferente a aquél en que se entrenaron. Estos autores extraen lo que denominan palabras que involucran sentimiento e incluyen adjetivos, verbos, nombres y adverbios y utilizan estas palabras para calcular la orientación semántica incluyendo la gradación de la orientación. Sostienen que el lexicón creado (SO-CAL) es consistente a través de distintos dominios con base en la tesis de que hay una polaridad primordial que es independiente del contexto, que corresponde a la orientación semántica en la mayoría de los dominios. Si bien crear diccionarios en forma automática o semiautomática tiene algunas ventajas, a causa de la inestabilidad de los mismos estos autores deciden crearlos en forma manual anotando la polaridad de las palabras. Se crean diccionarios separados para nombres, verbos y adverbios.

La experiencia en métodos basados en lexicones o diccionarios de palabras que tienen anotada la orientación semántica para análisis de sentimientos da buenos resultados según Taboada *et al.* (2011). El resultado es estable incluso a través de distintos dominios, aunque cabe aclarar que se aplica para la clasificación de textos más largos (entradas de blogs, etc.) y se utilizan estrategias para sopesar los parágrafos en la clasificación que no son aplicables en textos de las características de los tuits.

Existen empresas que brindan servicios SaaS (Software as a Service) para realizar el análisis de sentimiento de un texto, la extracción de tópicos, categorización de texto, entre otros. Meaning Cloud[6] es uno de ellos y permite utilizar su API a través de una clave de acceso[7] para realizar el análisis de sentimiento (Cervantes, 2016). A través de la API o de un *endpoint* se puede realizar:

- Análisis de sentimiento: se ingresa un texto y como salida se indica su polaridad, sujetividad, ironía o especificación de desacuerdo.
- Extracción de tema: identifica entidades nombradas y conceptos.
- Clasificación de texto de acuerdo a una taxonomía.

Meaning Cloud ofrece API para distintos escenarios. Estas API incluyen diccionarios, taxonomías y otras funcionalidades.

Enfoque PLN

El uso de una aproximación lingüística ofrece varios problemas, entre los cuales el uso de un lenguaje coloquial y muchas veces agramatical así como el escueto tamaño de los textos son los más notorios.

Tratamiento de la negación

Se ha estudiado que existe una tendencia a expresar una opinión negativa no directamente sino a través de la negación de una opinión positiva como una forma de suavizar la opinión negativa. En Vilares, Alonso y Gómez Rodríguez (2015) se hace una clasificación de las negaciones que pretende especificar el alcance de la negación en distintos casos, pero en el lenguaje informal de Twitter no parece aplicable.

En principio el uso del "no" elimina el carácter positivo de una afirmación, como en el caso siguiente:

El servicio es muy bueno vs. El servicio no es muy bueno

En este caso, si "bueno" se puntuara con un sentimiento positivo, al poner "no" el sentimiento de la frase no se neutraliza sino que expresa una opinión negativa (Liu, 2010). Otras partículas negativas son ni, nunca, nadie, ningún, nada, sin. También están los prefijos que tienen una orientación negativa:

6 https://www.meaningcloud.com/es ofrece un servicio para extraer significado de contenido no estructurado: conversaciones sociales, artículos, expedientes.

7 https://www.meaningcloud.com/developer/ es el sitio para desarrolladores de Meaning Cloud.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

- i-: legible, ilegible
- in-: transitable, intransitable
- im-: perfecto, imperfecto
- des-: favorable, desfavorable
- a-: normal, anormal

Habría que analizar a fondo esta cuestión porque puede haber casos en los que el prefijo no tenga una connotación negativa. La partícula "no" también puede tener una connotación que no sea negativa, como en este ejemplo: "El conductor *no* solamente nos acompañó sino que también nos dio indicaciones claras de cómo proceder". Liu (2010) plantea que la negación de una opinión positiva constituye una opinión negativa. También que la negación de una opinión negativa constituye una opinión positiva; sin embargo, también se sostiene que existe la tendencia de suavizar las opiniones negativas por lo cual en muchos casos la negación de una opinión positiva constituye una opinión negativa o a lo sumo neutral como forma educada de exponer las opiniones. Por ejemplo:

El lugar no es tan malo.
La comida no resultó tan incomible.

Se analizan algunos casos para el tratamiento de la negación:

- Hay un elemento de polaridad positiva que al tener una negación cambia su polaridad a negativa: Es muy bueno/no es muy bueno.
- Hay un no que refuerza lo negativo: No se puede ir. Es inaguantable.
- Hay un no que no produce un efecto negativo: ¿Quién dijo que no se puede comer un helado en invierno?
- La gradación de la negatividad (Vilares, Alonso y Gómez Rodríguez, 2015): negativo, muy negativo.

Como se puede ver hay una casuística variada por lo que un tratamiento general del "no" no daría resultados correctos en todos los casos. En el estudio realizado lo que se hace –adhiriéndose a Liu (2010)– es invertir la polaridad del tuit si aparece la partícula "no". Si el tuit es positivo (sin tomar en cuenta la palabra no), si tiene un "no" se invierte la polaridad y se considera negativo. Si el tuit es negativo (sin tomar en cuenta la palabra no), si tiene un "no", se considera positivo.

Subjuntivo

En español el uso del subjuntivo indica que lo que se predica no es un hecho, puede ser una afirmación hipotética o un deseo y está caracterizado por el rasgo irrealis. Las oraciones que hacen uso del subjuntivo tienen grandes probabilidades de expresar un sentimiento.

Locuciones

Las locuciones son ampliamente usadas en el lenguaje coloquial y especialmente en Twitter. Las locuciones que pueden tener una carga mayor de opinión son las locuciones adjetivales, adverbiales y verbales. Al parecer, el caso menos explorado es el uso de las locuciones verbales, que como parte de la expresión coloquial aparecen detentando polaridad como en el caso de "meter la pata", "hacerse el oso", "comérselo en dos panes", "estar de fiesta", "costar un ojo de la cara", "cortarle la cara", etc.

Resulta de interés incorporar a los lexicones de polaridad algunos verbos que denotan sentimiento y locuciones verbales que también tienen una carga de polaridad semántica importante y que son muy locales y difícilmente se comprenden fuera del ámbito de uso.

Cláusulas subordinadas

Liu (2010) sostiene que el "pero" en inglés cambia la orientación de la cláusula anterior y que en español funciona igual, planteando una situación que se opone o que parcialmente contradice lo anterior, como en los casos siguientes:

La comida es buena pero hay que esperar mucho.
Hay que esperar mucho pero la comida es buena.

El "aunque" funciona de manera análoga:

La casa está en un buen punto, aunque es un poco chica.

En los casos de ejemplo hay dos opiniones, una en la oración independiente y otra en la subordinada, aunque es difícil extraer una regla sobre la opinión general porque incluso manualmente no está clara la opinión global que dependerá del peso que se otorgue en forma particularizada a cada una de las predicaciones en cierto sentido antagónicas que plantea la oración.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

En lo que sigue se utiliza Freeling,[8] aplicación de licencia libre (Affero GPL) mantenida por la Universidad Politécnica de Cataluña para el análisis de lenguaje que incluye distintos análisis: morfológico, anotación de PoS, *parsers* de gramáticas libres de contexto y dependencias, entre otros, para una variedad de idiomas y en particular el español.

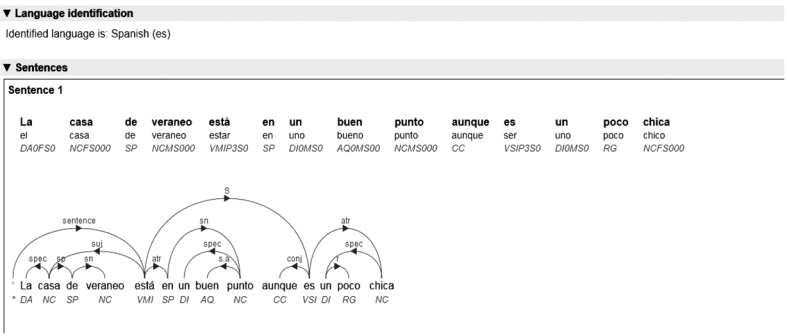

Figura 1. Freeling usando *dependency parser*

En la *Figura 1* se muestra el análisis de Freeling para la oración "La casa de veraneo está en un buen punto aunque es un poco chica" utilizando el *parser* de dependencias, que funciona correctamente indicando las dos oraciones. En la primera el adjetivo "buen" (denotado por Freeling como adjetivo calificativo) se aplica a "punto" de tipo nombre. Si se anotara manualmente un lexicón dando contexto se pediría la polaridad de "buen" aplicada a "punto" lo que se presume como positivo. De esta forma la primera oración sería positiva. En la segunda oración la palabra que presumiblemente tiene la carga semántica es el adverbio (poco) utilizada con el verbo (es) como contexto, podría anotarse manualmente como negativo, con lo cual la segunda oración subordinada sería negativa. Si se realizara la suma de las dos polaridades para la oración completa el resultado sería neutro, una parte positiva y una parte negativa.

Una línea para profundizar sería el análisis de que en casos como el de esta oración la carga semántica está en la etiqueta *atr* (*attribute*) y el contexto está en el sujeto de la primera oración. Entonces para la oración subordinada la anotación de polaridad con el contexto sería para *chica* con el contexto *casa*, que igualmente sería negativo pero mucho más ajustado. Otra cuestión

8 http://nlp.lsi.upc.edu/freeling/node/1 Sitio de un conjunto de herramientas de acceso abierto para análisis de lenguaje, desarrolladas y mantenidas por la Universidad Abierta de Cataluña.

interesante respecto a la coordinación la plantea Hatzivassiloglou y McKeown (1997), quienes realizaron un estudio cuando hay dos adjetivos unidos por una conjunción coordinante. Sostienen que cuando dos palabras aparecen vinculadas por "y" siendo ambas adjetivos existe una alta probabilidad de que tengan la misma polaridad, en cambio cuando la vinculación es con "pero" la polaridad tiende a ser opuesta. Por ejemplo, "bonito y barato", "complicado pero interesante".

Contexto a través del uso de las características sintácticas de las palabras a través de distintos enfoques

El desafío de mejorar la clasificación automática de la polaridad del tuit pasa por establecer contexto para determinar la polaridad de las palabras. Si bien se adhiere a que existe una polaridad primordial para algunas palabras y que esta polaridad se aplica a la mayoría de los dominios, en muchos casos las palabras utilizadas tienen una polaridad que se relativiza de acuerdo al contexto en que se usan.

Las palabras con polaridad consideradas son los adjetivos y los adverbios, y aunque también los verbos y las locuciones verbales se presumen de carga de polaridad no se han considerado para este estudio. La cuestión es analizar, dado un adjetivo, a qué se aplica, a qué adjetiva; lo mismo con el adverbio.

Rojo (1975) hace una distinción entre adjetivos restrictivos y no restrictivos y analiza su ubicación en la frase nominal estableciendo que en general el adjetivo restrictivo va luego del sustantivo y el no restrictivo antes. En el ejemplo que sigue el adjetivo restrictivo modifica al sustantivo (político) y el no restrictivo a la agrupación de sustantivo y adjetivo (personaje político):

conocido **personaje político**
terrible **accidente aéreo**
evidente **crisis económica**

En estos casos la agrupación sustantivo-adjetivo restrictivo funciona en forma similar a una locución adjetival, lo que se califica es el "accidente aéreo" y la carga de polaridad está entonces en el adjetivo no restrictivo (terrible).

Cuando se usa el analizador sintáctico de Freeling, éste no realiza distinción entre los adjetivos. En el caso de la oración:

El terrible accidente aéreo ocurrió cerca de la ciudad.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77.57863

El *parsing* PoS en formato CONLL (https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/sequences/CoNLLDocumentReaderAndWriter.html):

1 El el DA0MS0 DA pos=determiner|type=article|gen=masculine|num=-singular
2 terrible terrible AQ0CS00 AQ pos=adjective|type=qualificative|gen=common|num=singular
3 accidente accidente NCMS000 NC pos=noun|type=common|gen=masculine|num=singular
4 aéreo aéreo AQ0MS00 AQ pos=adjective|type=qualificative|gen=masculine|num=singular
5 ocurrió ocurrir VMIS3S0 VMI pos=verb|type=main|mood=indicative|tense=past|person=3|num=singular
6 cerca cerca RG RG pos=adverb|type=general
7 de de SP SP pos=adposition|type=preposition
8 la el DA0FS0 DA pos=determiner|type=article|gen=feminine|num=singular
9 ciudad ciudad NCFS000 NC pos=noun|type=common|gen=feminine|num=-singular

El *parsing* de dependencias en formato CONLL tampoco distingue la etiqueta:

1 El el DA0MS0 DA pos=determiner|type=article|gen=masculine|num=-singular - - - 3 spec - - -
2 terrible terrible AQ0CS00 AQ pos=adjective|type=qualificative|gen=common|num=singular - - - 3 s.a - - -
3 accidente accidente NCMS000 NC pos=noun|type=common|gen=masculine|num=singular - - - 5 suj - - A1
4 aéreo aéreo AQ0MS00 AQ pos=adjective|type=qualificative|gen=masculine|num=singular - - - 3 s.a - - -
5 ocurrió ocurrir VMIS3S0 VMI pos=verb|type=main|mood=indicative|tense=past|person=3|num=singular - - - 0 sentence - ocurrir.00 -
6 cerca cerca RG RG pos=adverb|type=general - - - 5 cc - - AM-TMP
7 de de SP SP pos=adposition|type=preposition - - - 6 sp - - -
8 la el DA0FS0 DA pos=determiner|type=article|gen=feminine|num=singular - - - 9 spec - - -
9 ciudad ciudad NCFS000 NC pos=noun|type=common|gen=feminine|num=-singular - - - 7 sn - - -
10 . . Fp Fp pos=punctuation|type=period

Freeling no distingue entre terrible y aéreo para el caso y los considera adjetivos calificativos. Si se pudiera establecer esta distinción se podría especificar el contexto (accidente aéreo) para el adjetivo que tiene la carga de polaridad (terrible).

El posicionamiento de los adjetivos en torno de los sustantivos adquiere distintos formatos por lo cual el posicionamiento no sería una indicación para la determinación de los tipos de adjetivos. Pueden darse posicionamientos diversos como en los siguientes casos:

Sustantivo + adjetivo restrictivo + adjetivo restrictivo Novelas policíaca inglesas
Sustantivo + adjetivo restrictivo + adjetivo no restrictivo Crisis económica manifiesta

Si utilizamos gramáticas independientes de contexto probabilísticas (GLCP) para analizar esta situación y definimos un grupo adjetival restrictivo (GAr) que instancie aquellos casos en que los adjetivos sean de tipo restrictivos para las oraciones 1) y 2) tenemos:

1) Las *novelas policíacas inglesas* están en el estante superior.
2) La *crisis económica* manifiesta afecta a los ciudadanos.
O -> GN GV
GN -> det nombre adjetivo adjetivo | *det GAr* | *det GAr* adjetivo |det nombre adjetivo
GAr -> nombre adjetivo | nombre adjetivo adjetivo
GV -> verbo GP
GP -> prep GN

En el caso 1) se puede aplicar GN-> det GAr (GAr -> nombre adjetivo adjetivo), pero en el caso 2) GN -> det GAr adjetivo (GAr-> nombre adjetivo)

Si se midiera la probabilidad de las reglas en un *corpus* podría estimarse cuál es la que tiene mayor probabilidad. La cuestión es que en ningún caso se toman en cuenta los aspectos léxicos que en este caso son los más significativos para determinar la característica de restrictivo de un adjetivo y por tanto las gramáticas libres de contexto probabilísticas no darían una buena solución.

En ese sentido la idea de que el núcleo en un grupo selecciona los complementos en función de sus propiedades léxicas es un punto a tomar en cuenta y parecería que un tratamiento con gramáticas HPSG (*head-driven*

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

phrase structure grammar) fuera más apropiado para capturar el contexto de los adjetivos y los adverbios. En HPSG el adjetivo, núcleo de la frase adjetival, tiene la característica de que modifica a un elemento que en este caso es un nombre. Puede verse como ejemplo en "acto programado":

```
programado := word &
[ ORTH "programado",
 HEAD adj_mas_sg,
 SPR < >,
 COMPS < >,
 MOD < [HEAD noun_mas_sg]>].
```

En el caso del adverbio, que es el núcleo de la frase adverbial, modifica a un elemento que es el verbo. Puede verse como ejemplo en "llegar tarde":

```
tarde := word &
[ ORTH "tarde",
 HEAD adv,
 MOD <[head verb]> ].
```

Puede plantearse entonces que el contexto para un adverbio es el verbo y para un adjetivo es el nombre, y de este modo realizar una anotación manual no ya sólo de adjetivos, sino de adjetivos en su contexto (nombre) y de adverbios en su contexto (verbos).

Enfoque Ciencia de la Información: temática

Según Barité *et al.* (2013) la organización del conocimiento es el

> Área del conocimiento de formación reciente, que estudia las leyes, los principios y los procedimientos por los cuales se estructura el conocimiento especializado en cualquier disciplina, con la finalidad de representar temáticamente y recuperar la información contenida en documentos de cualquier índole, por medios eficientes que den respuesta rápida a las necesidades de los usuarios.
>
> La Organización del Conocimiento se nutre de los aportes recibidos de la Informática, la Lingüística, la Terminología y la Ciencia de la Información.
>
> […]
>
> El objeto de estudio de la Organización del Conocimiento es el conocimiento

> socializado o registrado, y en lo que hace a Bibliotecología y Documentación, da cuenta del desarrollo teórico-práctico para la construcción, la gestión, el uso y la evaluación de clasificaciones, taxonomías, nomenclaturas, ontologías temáticas y lenguajes documentales. Asimismo, ampara el conjunto de conocimientos vinculados al análisis de información en general, considerando aspectos semánticos, cognitivos e informáticos.

La organización del conocimiento genera productos para categorizar los recursos a describir con distintos grado de estructuración y riqueza como las taxonomías, las clasificaciones, los tesauros y las ontologías. Estos productos son utilizados para categorizar los recursos informacionales de modo que el usuario pueda encontrar los recursos adecuados a su consulta.

Hjørland (2009) discute la "teoría del concepto", el "átomo" en la descripción del conocimiento que utilizan estos productos y concluye que los conceptos deben ser comprendidos como significados negociados socialmente identificados estudiando los discursos más que como principios *a priori*. Afirma:

> Cuando las palabras son colocadas en bases de datos, el contexto que proporciona cierta parte de su significado se pierde en parte. Desde la perspectiva de los paradigmas social e históricamente integrados, la futura mejora de la tecnología de recuperación de información está conectada a las posibilidades de reestablecer los contextos perdidos que determinan el significado de las palabras y los conceptos. (1530, traducción propia)

Realizar una clasificación de tuits por temática o por dominio es una cuestión previa al análisis de opinión porque el lenguaje y las entidades que aparecen están en relación a la temática. Si no se determina la temática existe mucho ruido en la recuperación. Este problema se menciona como trabajo futuro en Selva Castelló (2015) y en Vilares Calvo (2014).

Cuando la búsqueda es por términos específicos, por ejemplo, un modelo de celular, los resultados serán bastante adecuados, pero si se trata de un tópico más general empieza a producirse mucho ruido. Una manera de resolver la cobertura es utilizar sinónimos para la búsqueda.

La precisión está afectada por las palabras que tienen más de un significado y una manera de resolverlo es asociar la palabra al dominio en la que el significado tenderá a ser más estable. Determinar el dominio sobre el que se descarga la muestra de tuits afectará tanto la cobertura como la precisión. En muchos estudios de sentimiento en Twitter se trabaja sobre dominios específicos como cine, etc., partiendo ya de un conjunto de datos clasificados.

Vilares Calvo (2014) recopila algunas iniciativas de clasificación de tópi-

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

cos desde algunas clasificaciones en 12 tópicos (política, altruísmo, eventos, tecnología, juegos, idiomas, música, personalidad, películas, celebridades, estilo de vida y deportes) hasta 50 a través de la utilización de distintas aproximaciones. En ese panorama el *corpus* TASS ya mencionado provee en español su caracterización propia.

La definición de un tópico en lingüística corresponde a un tema o un asunto principal del que se habla, se explica, se predica o se comunica algo, en una frase o en un discurso y sobre el que se va aportando nueva información.

El ejemplo de un tuit como "arriba defensor…" constituye un caso claro de la necesidad de contexto, de un tema. En Uruguay una persona rápidamente asociaría que la palabra no alude a alguien que defiende sino al equipo uruguayo de futbol de primera división Defensor. La extracción de tópicos permitiría poner Defensor bajo el tópico "Fútbol" y no bajo "Justicia".

En el artículo de Gattani *et al.* (2013) se construye una base de conocimientos que tiene una serie de conceptos, subconceptos, un conjunto de instancias y un conjunto de relaciones entre los conceptos. Un planteo muy ingenioso es el uso de Wikipedia para mapear la base de conocimiento (conceptos e instancias). Se señala la importancia que en el caso de Twitter reviste el contar con una base de conocimientos en tiempo real ya que la evolución de los temas, los eventos y las instancias sobre las que versa el tuit es muy dinámica. La propuesta es mapear lo que se denominan "entidades" –que coinciden en gran parte con los sustantivos– a la base de conocimiento para encontrar el tópico.

Esta operación es conocida como *Named Entity Recognition* (NER, https://nlp.stanford.edu/software/CRF-NER.shtml) y es el etiquetado de secuencias de palabras en un texto que constituyen nombres de cosas tales como personas, nombres de empresas o proteínas, entre otros. Stanford proporciona un modelo para el inglés con tres categorías (personas, organizaciones, localizaciones) y en el 2014 se generó un modelo en español.

Esta necesidad de categorizar las entidades es cercana a la iniciativa de establecer características de los objetos. Ambas son necesidades de ubicar los objetos, los conceptos o las instancias en un sistema jerárquico, una taxonomía, un tesauro o una base de conocimientos que de alguna forma organiza los elementos en categorías y modela las relaciones entre ellos, fundamentalmente las relaciones jerárquicas.

El *Aspect Based Sentiment Analysis* (ABSA) plantea un análisis en el sentido de lo que Liu (2010) considera una opinión sobre un objeto y sus componentes en una estructura de árbol. Los aspectos (características o componentes) se definen como una combinación de una entidad (por ejemplo, restaurante) y un atributo de esa entidad (por ejemplo, precio) y hay una dife-

rencia entre la opinión sobre una entidad o sobre un aspecto de la entidad o un atributo de la entidad. Puedo tener una opinión positiva del restaurante, pero negativa de un aspecto de éste –el precio–. Por supuesto que la opinión del atributo participa en la opinión de la entidad, pero esa participación no es tan trivial, porque otros atributos también participan y componen la opinión sobre el objeto.

Esta estructuración de los componentes de los objetos en forma jerárquica a través de una relación "se compone de" o la granularización de los conceptos a través de la relación "es una" constituye un primer paso en la categorización, que va en el sentido de delinear un dominio, y la jerarquización, que permite ajustar el enfoque al nivel de detalle que se desee; ambos son pasos incipientes para disminuir el ruido y encaminarse a subir desde el nivel léxico al nivel conceptual en el sentido de un tesauro.

Un tesauro es un vocabulario controlado y estructurado formalmente, constituido por términos que guardan entre sí relaciones semánticas y genéricas de equivalencia, jerárquicas y asociativas. Se trata de un instrumento de control terminológico que permite convertir el lenguaje natural de los documentos en un lenguaje controlado, ya que representa, de manera unívoca, el contenido de estos, con el fin de servir tanto para la indización como para la recuperación de los documentos (Lapuente, 2007).

En el tesauro también se consideran términos compuestos, la estructura sintáctica según Lapuente (2007) es:

- sustantivo + adjetivo: documentos digitales
- sustantivo + sintagma preposicional: documentos de archivo
- sustantivo + sintagma aposicional: documentos RDF/XML

Hay una analogía entre las bases de conocimiento y los tesauros aunque el nivel expresivo de estos últimos es menor. En el tesauro la relación jerárquica impone una taxonomía de conceptos, clases y subclases que se corresponden a términos generales y términos específicos, esta taxonomía es un árbol donde los nodos son los conceptos y los arcos la relación "es una" que mapea a una subclase. Las relaciones de términos modelan asimismo una relación asociativa.

La primera etapa de la elaboración de un tesauro es la recopilación de los términos del dominio. Luego, esos términos empezarán a organizarse de manera que sea posible establecer una taxonomía, relaciones de clase y subclase (término tope y términos específicos), relaciones de véase además (asociativas), y dentro de los términos sinónimos elegir un término tope (el representante de la clase de equivalencia) y los otros términos sinónimos que

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

remitirán a este.

Las folksonomías parten de un punto distinto: para identificar un texto por su contenido se aportan etiquetas representativas de los temas que son términos sin ningún tipo de normalización y sin establecer relaciones. La riqueza de estas etiquetas o términos descriptivos reside en el hecho de que son aportados en las redes sociales y se reutilizan y comparten. Existen gestores de estos marcadores sociales en los que esta interacción se lleva a cabo,[9] la cual, junto con la reutilización, puede relacionarse con la idea de la memética y la hipótesis de la replicación de las ideas en otros huéspedes en forma análoga a como sucede en la genética.

Para construir una taxonomía es necesario determinar la categoría de los términos extraídos. Si queremos subir un nivel en la taxonomía es necesario mapear los conceptos extraídos en un sistema jerárquico que vaya delineando un dominio con clases y subclases. Wikipedia es especialmente adecuada porque tiene un sistema en el que cada página –que corresponde a un concepto o a una instancia– tiene etiquetada la categoría a la que pertenece en un sistema jerárquico. Cada categoría en Wikipedia corresponde a su vez a una categoría más amplia. Una primera división es Ciencia, Arte, Naturaleza y Sociedad; luego, por ejemplo, la categoría Ciencia incluye las siguientes 13 subcategorías:

- Científicos‟ (12 cat, 1 pág.)
- Ciencia por año‟ (359 cat)
- Ciencia y tecnología por continente‟ (7 cat)
- Ciencia y tecnología por país‟ (178 cat)
- Wikiproyectos de ciencias‟ (3 cat)
- Anexos: Ciencias‟ (8 cat, 2 págs.)
- Ciencia y sociedad‟ (10 cat, 10 págs.)
- Filosofía de la ciencia‟ (15 cat, 122 págs.)
- Historia de la ciencia‟ (19 cat, 37 págs.)
- Investigación científica‟ (7 cat, 39 págs.)
- Problemas científicos‟ (3 cat, 2 págs.)
- Protociencia‟ (4 cat, 7 págs.)
- Ramas de la ciencia‟ (5 cat, 2 págs.)

Otra herramienta que se puede usar para construir el árbol jerárquico del objeto o del concepto es Wordnet porque incorpora las relaciones semánticas (lemas, sinónimos, hiperónimos, hipónimos). En el caso de un discurso con las características del tuit el uso de un tesauro formal difícilmente se ajustará a esos conceptos socialmente entendidos pero gramatical y sintácticamente transgresores, por lo que se busca desarrollar una tendencia en delinear una

9 https://del.icio.us/ Sitio que ofrece servicio de marcadores web.

jerarquía que se ajuste con gran flexibilidad.

Prototipo

El prototipo realizado busca determinar la polaridad global de un tuit. No considera la gradación de intensidad. La idea es clasificar un tuit como positivo, negativo o neutro.

Se crea un lexicón que registra la polaridad semántica de algunas palabras (adjetivos y adverbios) a través de una anotación manual utilizando el contexto de los términos sobre los que se expresa el sentimiento (nombre y verbo respectivamente). El lexicón se crea *ad hoc* porque de esta forma se estará tomando en cuenta la localidad y la temporalidad del lenguaje usado en el tema en particular.

El lexicón tendrá como entradas los adjetivos y los adverbios como sujetos pasibles de detentar la polaridad. Los nombres referirán el tema de la opinión y su granularidad con lo cual es posible establecer un mapeo al vector de Liu del objeto de opinión y sus componentes. Algunos verbos tienen también una carga de polaridad y se pueden incluir en el lexicón de sentimientos. En el caso de los verbos en el lexicón se considera el lema ya que en español las variantes de persona y tiempo son muchas.

El lexicón no solamente incluye las palabras que expresan polaridad y su polaridad sino que además registra la palabra a la que están modificando, con el fin de establecer un contexto de aplicación de la misma. Se obtiene entonces un archivo para los adjetivos de la forma Adjetivo(nombre): polaridad, donde polaridad toma los valores 1 (positivo), 0 (neutro), -1 (negativo). Se obtiene un archivo para los adverbios de la forma Adverbio(verbo): polaridad, donde polaridad toma los valores 1 (positivo), 0 (neutro), -1 (negativo).

Los nombres además son utilizados para construir una terminología *ad hoc* o social, que recopila las entidades y sus componentes. Estos nombres son recopilados a ese efecto en un archivo que contiene los nombres que son utilizados en los tuits. Sin llegar a la construcción de un tesauro o una taxonomía se realiza la extracción de términos que ayuden a delimitar el dominio y generar una terminología (no normalizada) pero con la ventaja de tener un conjunto de términos usados en tiempo real. En cierto sentido lo que se modela es una terminología de uso, una terminología folksonómica podríamos decir.

La idea no es crear un tesauro o una base de conocimiento del tópico, sino establecer una recopilación de términos a modo de nube de palabras que sirvan para guiar la búsqueda de tuits y establecer el contexto para la

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

polaridad de los adjetivos y adverbios. No obstante, como una línea futura de trabajo puede plantearse la construcción no de una terminología *ad hoc*, sino de una taxonomía *ad hoc* mapeando Wikipedia o Wordnet.

La idea que se plantea parte de la hipótesis de que las palabras que se utilizan para expresar opinión son distintas en distintos dominios y en distintos lugares porque el lenguaje y los giros son elementos culturales muy localizados geográficamente y sensibles a los espacios temporales e incluso a lo que podemos denominar "tribus urbanas". Por lo antedicho la tarea de generar un lexicón que contenga palabras de aplicación global es muy dificultosa y de aplicación muy restringida.

La nube de palabras que se genera *ad hoc* involucra conceptos, instancias y verbos que sirven para dos cosas: la primera, delinear la terminología del tópico; la segunda, contextualizar el aporte de la polaridad de las palabras.

El problema de la ambigüedad en el procesamiento del lenguaje natural es un problema general y en la expresión de sentimiento también se manifiesta. Si un lexicón de polaridad puede advertir sobre el objeto a que se califica tendrá mayor probabilidad de adecuarse, como en el siguiente ejemplo:

desestructurante (jugador) positivo
desestructurante (masaje) neutro
desestructurante (relación) negativo

El prototipo se generó como software libre y está disponible en *Sourceforge* (https://sourceforge.net/projects/pln-polaridad-en-twitter/). Debe señalarse que lo que se diseña es un prototipo pero no resulta apropiado para realizar un análisis porque hay un problema metodológico que proviene de extraer la polaridad de los términos en un conjunto de tuits y sobre el mismo conjunto asignar la polaridad con los términos ya relevados. Para poder extraer conclusiones se debería realizar la extracción de términos en un conjunto y realizar el análisis de opinión en otro conjunto. Para lograr un resultado plausible se debería trabajar con un volumen apropiado para verificar una cobertura adecuada. Otra posibilidad es incorporar el lexicón como diccionario complementario en los servicios como *Meaning Cloud*.

Conclusiones

El lenguaje constituye un rasgo importante de la identidad cultural de las personas, e incluso dentro de una misma zona geográfica varía de acuerdo al

grupo social que lo utiliza, el cual tendrá su propia jerga, sus modismos, etc.

El uso del lenguaje en las redes sociales es coloquial, en formato muy informal, sin cuidado de la ortografía, la sintaxis o la gramática pero constituye un formato expresivo de amplio uso. La expresión en Twitter es particularmente fragmentada, constituye una expresión transmedia, hipervinculada, signo de un discurso que se fragmenta: empieza en un medio, continúa en otro, alude a un contexto que está no sólo en otro lugar, en otro medio, sino que además está expresado en formatos diversos, texto, imagen, video, etc., utilizando signos que se continúan en otros lugares y que dan pistas de un camino expresivo laberíntico.

La ciencia de la información tiene un amplio desarrollo en la categorización temática de los documentos y la categorización por sentimientos aparece como una línea de trabajo emergente que es de interés. En este sentido cabe señalar la existencia de un área de interdisciplina entre el procesamiento automático de lenguaje natural y la anotación manual de la polaridad semántica o la extracción de entidades por los profesionales de la información. La anotación o revisión manual aparece como una necesidad ya sea para la anotación de un corpus como para el diseño de distintas estrategias para contextualizar el sentimiento y es un área que puede desarrollarse en conjunto. Modelar la terminología de un dominio, establecer una base de conocimiento de un dominio, elaborar una taxonomía o un tesauro o incluso delinear los atributos de los objetos son operaciones muy cercanas y que también están en esa zona común.

La categorización del sentimiento reviste mayor complejidad que la categorización por tópicos y en particular con los métodos automáticos, ya que el sentimiento involucra elementos como el contexto, la ambigüedad, la ironía, que son fácilmente detectables por los humanos pero no sucede lo mismo en forma automática.

El mayor desafío es la adecuación en la determinación del sentimiento de las palabras y está vinculado con el contexto de las mismas. El contexto incluye los giros idiomáticos, la ironía, el sarcasmo y el lenguaje particular que se usa que puede tener distintos significados dependiendo del tema, el sitio o el emisor.

El aporte de contexto asignando el aspecto calificado (nombre) al calificador (adjetivo) –o adverbio (verbo)– es una iniciativa de mejoramiento y sería posible explorar en trabajos futuros si la polaridad varía en función del objeto al que se aplica o están involucrados otros aspectos no considerados. Otro trabajo a futuro sería abordar la polaridad de los verbos y sobre todo las locuciones verbales que tienen una alta carga de polaridad.

La propuesta de contextualizar se funda en la hipótesis de que un lexicón de polaridad para redes sociales no puede ser genérico, pues está vinculado

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863

no sólo al tema, al lugar, al tiempo, al idioma, sino que varía rápidamente porque constituye una expresión de un idioma cambiante que es un organismo vivo en continua transformación.

Si bien es posible que exista una polaridad general o primordial como sostienen algunos autores, ésta es seguramente sobreseída por las particularidades de la instancia concreta en un alto porcentaje. Lo que se propone como estrategia es la utilización de lexicones generales en combinación con la generación *ad hoc* de lexicones para cada situación en particular, lexicones que se entrenen con una herramienta como el prototipo que se desarrolló y se usen ajustando, entallando para lograr una adecuación a la situación concreta. Esta propuesta es practicable puesto que un servicio como *Meaning Cloud* permite combinar diccionarios propios.

Existe por otra parte la necesidad de modelar el objeto y sus atributos para ajustar los distintos elementos de la opinión sobre un objeto en concreto estableciendo el árbol de componentes de los objetos para realizar un modelado de los aspectos (ABSA), lo cual desde otro punto de vista constituye una taxonomía o un tesauro a microescala. Este modelado es un análisis que recoge tanto conceptualizaciones como terminología de uso en las redes sociales y permite un análisis más relevante de los objetos.

El análisis de sentimientos se puede complementar con otros análisis como los de la ubicación del emisor en la red de difusión. Identificar el emisor en un grafo permite interpretar la importancia de las opiniones, ya que la opinión de un personaje influyente –con muchos seguidores – llega a muchos otros nodos y resulta más significativa.

La extracción de entidades o terminología de los tuits así como su polaridad constituyen indicadores significativos que pueden ser tomados en cuenta para el desarrollo y mejora de productos, políticas o acciones.

Referencias

Agarwal, A., B. Xie, I. Vovsha, O. Rambow y R. Passonneau. 2011. "Sentiment analysis of twitter data". *Proceedings of the workshop on languages in social media* (Association for Computational Linguistics): 30-38.

Barité, M. *et al*. 2013. *Diccionario de organización del conocimiento: clasificación, indización, terminología*, 5a. ed. Montevideo: PRODIC. http://archivos.liccom.edu.uy/diccionario/Diccionario%20Definitivo%20%5B3-11-13%5D.html

Capurro, Rafael. 1987. "La hermenéutica y el fenómeno de la información". *Cuaderno de psicoanálisis freudiano* 8: 107-120. http://www.capurro.de/herminf.html

Cervantes, Ofelia. 2016. Curso *Análisis semántico de redes sociales*. https://eva.fing.

edu.uy/course/view.php?id=939

Gattani, A., D. S. Lamba, N. Garera, M. Tiwari, X. Chai, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan y A. Doan. 2013. "Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach". *Proceedings of the VLDB Endowment* 6 (11): 1126-1137. http://www.vldb.org/pvldb/vol6/p1126-gattani.pdf

Hatzivassiloglou, V. y K. R. McKeown. 1997. "Predicting the semantic orientation of adjectives". *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 174-181. Association for Computational Linguistics.

Hjørland, Birger. 2009. "Concept theory". *Journal of the American Society for Information Science and Technology* 60: 1519-1536.

Hu, X., J. Tang, H. Gao y H. Liu. 2013. "Unsupervised sentiment analysis with emotional signals". *Proceedings of the 22nd international conference on World Wide Web*, 607-618. ACM.

Kouloumpis, E., T. Wilson y J. D. Moore. 2011. "Twitter sentiment analysis: The good the bad and the omg!". *Icwsm* 11: 538-541.

Lapuente, M. J. L. 2007. Hipertexto: *El nuevo concepto de documento en la cultura de la imagen*. http://www.hipertexto.info/documentos/web_tecnolog.htm

Liu, C. 2010. *NLP Handbook*. Chicago: University of Illinois.

Montejo Ráez, A., E. Martínez Cámara, M. T. Martín Valdivia y L. A. Ureña López. 2014. "Ranked wordnet graph for sentiment polarity classification in Twitter". *Computer Speech & Language* 28 (1): 93-107.

Pang, B., L. Lee y S. Vaithyanathan. 2002. "Thumbs up?: sentiment classification using machine learning techniques". *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* 10: 79-86. Association for Computational Linguistics.

Pang, B. y L. Lee. 2008. "Opinion mining and sentiment analysis". *Foundations and trends in information retrieval* 2 (1-2): 1-135.

Rojo, G. 1975. "Sobre la coordinación de adjetivos en la frase nominal y cuestiones conexas". *Verba* 2: 193-224.

Saif, H., M. Fernandez, Y. He y H. Alani. 2014. "On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter". *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. https://pdfs.semanticscholar.org/68ff/581ba7162f5c2e809c72823bf418a58d4ff9.pdf

Selva Castelló, Javier. 2015. "Desarrollo de un sistema de análisis de sentimiento sobre Twitter". Tesis de doctorado.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll y M. Stede. 2011. "Lexicon-based methods for sentiment analysis". *Computational linguistics* 37 (2): 267-307.

Urizar, X. S. e I. S. V. Roncal. 2013. "Elhuyar at TASS 2013". *Proceedings of the Workshop on Sentiment Analysis* at SEPLN (TASS 2013), 143-150. http://www.sepln.org/workshops/tass/2013/papers/tass2013-submission3-Elhuyar.pdf

Vilares, D., M. A. Alonso y C. Gómez-Rodríguez. 2015. "On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages". *Journal of the Association for Information Science and Technology* 66 (9): 1799-1816.

Vilares Calvo, David. 2014. *Análisis de contenidos en Twitter: clasificación de mensajes e identificación de la tendencia política de los usuarios*. España: Universidad da Coruña, Facultad de Informática.

Para citar este texto:
Temesio-Vizoso, Silvana Grazia. 2018. "La opinión en textos con un enfoque interdisciplinar: propuesta de contexto y léxico *ad hoc*". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 32 (77): 73-99.
http://dx.doi.org/10.22201/iibi.24488321xe.2018.77.57863

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.77. 57863