Facultad de Información y Comunicación
UNIVERSIDAD DE LA REPÚBLICA URUGUAY
San Salvador 1944
11200-Montevideo, Uruguay
Tel: [+598] 2413 80 20 / 21 / 22
fic@fic.edu.uy
www.fic.edu.uy

Análisis de sentimiento en twitter

Sentiment analysis on twitter

Grupo de trabajo: GT3- Humanidades digitales

Silvana Temesio

Inscripción Institucional: FIC – Instituto de Información

silvana.temesio@fic.edu.uy

Resumen

El análisis de un texto puede realizarse desde varios enfoques y se realizan aproximaciones desde el procesamiento de lenguaje natural (PLN) y la ciencia de la información.

Utilizando PLN se buscan las palabras que se cargan de polaridad como los adverbios y los adjetivos para confeccionar lexicones de sentimientos. De acuerdo a la gramática HPSG el adverbio modifica al verbo y el adjetivo al sustantivo. A partir de esto se formula la idea de que los modificadores de los núcleos (sustantivos y verbos) aportan a la definición temática de lo que se predica y de este modo brindan contexto en los exigüos textos de twitter y pueden mejorar de ese modo el análisis de polaridad. Desde la ciencia de la información se plantea la construcción de una terminología de dominio que ayude a contextualizar el análisis de polaridad en un medio con un jerga muy cambiante y una construcción gramatical que resulta difícil de analizar con metodologías que se aplican a textos de tipo tradicional con una gramática y un lenguaje más pulido.

Finalmente se presenta un prototipo para concretar esta propuesta.

Palabras clave: Análisis de sentimiento, twitter, PLN, ciencia de la información

Abstract

Text analysis can be done from several approaches, here approximations are made from natural language processing (PLN) and from information science.

Through PLN we use the search of words that are loaded with polarity such as adverbs and adjectives to make sentiment lexicons. According to the HPSG grammar the adverb modifies the verb and the adjective the noun. From this point is formulated the idea that the modifiers of the nuclei (nouns and verbs) contribute to the thematic definition of what is preached and in this way provide context in the required texts of twitter and also can improve the analysis the polarity analysis. From the point of view of information science is stated that the construction of domain terminology can help to contextualize the polarity analysis. Twitter is a medium with a very changing jargon and a grammatical construction that is difficult to analyze with methodologies that apply to traditional texts that use a more polished grammar and language.

Finally, a prototype is presented to make this proposal concrete.

Keywords: Sentiment analysis, twitter, PLN, information science

San Salvador 1944
11200-Montevideo, Uruguay
Tel: [+598] 2413 80 20 / 21 / 22
fic@fic.edu.uy
www.fic.edu.uy

Introducción

La web es un enorme repositorio de texto desestructurado e impone la necesidad de recuperar esos documentos digitales para lo cual es necesario caracterizarlos. A partir de esta necesidad la minería de textos es una disciplina que se ocupa de la extracción de información de textos no estructurados e involucra entre otras a la recuperación de la información, la estadística y el procesamiento del lenguaje natural.

El reconocimiento de entidades (NER named entity recognition) es la tarea de extracción de elementos de un texto que corresponden a categorías como nombres de personas, organizaciones, lugares geográficos u otros elementos que se denominan en forma genérica entidades. El reconocimiento automático de estas entidades es una tarea del área del procesamiento del lenguaje natural (PLN) y es ofrecido por distintos servicios con una progresiva mejora en su identificación. Hay una fuerte dependencia del reconocimiento respecto al idioma y al dominio porque en general se utilizan métodos de aprendizaje que involucran la anotación de corpus.

Existe cierta analogía de estas entidades con las entidades de FRBR del grupo 3 (Tillet, 2005), en cuyo caso las entidades representan objetos claves que interesan a los usuarios de los datos bibliográficos a los efectos de la recuperación de la información en un catálogo. En FRBR hay tres grupos de entidades, el primer grupo corresponde a las obras, el segundo grupo a los productores de las obras, ya sea personas o entidades corporativas y el tercer grupo refiere a las entidades que se utilizan como materias de las obras:

- Concepto - una idea o una noción abstracta
- Objeto - algo material
- Acontecimiento – una acción o suceso
- Lugar – una localización

En cierto sentido se puede ver la catalogación como una extracción de entidades del recurso a describir por lo que hay una cierta analogía con el reconocimiento de entidades en PLN.

En ciencias de la Información se analizan los recursos (documentos, libros, artículos) recabando los elementos que los definen a través de la catalogación y buscando los puntos de acceso o los términos por los cuales el recurso puede ser recuperado temáticamente a través de la indización. La indización aporta a la recuperación de la información cuando el usuario busca por los términos indizados. En este caso la indización es un proceso manual que realiza el profesional de la información y el objetivo es tender un puente entre los términos de búsqueda del usuario y los términos con los cuales se ha categorizado el texto. La indización suele realizarse ajustándose a un “patrón estandarizado” como una lista de encabezamientos o un tesauro.

Los conceptos pueden tener varios términos que los caracterizan que son sinónimos o quasi sinónimos. El tesauro funciona realizando una partición en una clase de equivalencia en donde se elige el

representante de clase como el término tope y los otros términos de la clase remiten a éste desde un "VÉASE". De esta forma al buscar por el término tope no es necesario buscar por todos los sinónimos ya que el documento estará indizado por el representante de la clase de equivalencia y se recuperarán todos los documentos que correspondan.

El problema suele ser que el usuario que busca debe conocer este patrón estandarizado que se usa. Si el usuario es especializado de a poco irá dominando esta práctica y encontrará rápidamente lo que busca, en caso contrario será guiado por un bibliotecólogo.

En el caso de la categorización de textos por técnicas del lenguaje natural el desafío es que se pueda indizar el documento utilizando distintas técnicas que tomarán en cuenta las palabras del propio documento. Las técnicas usarán frecuencias y también contexto.

Cuando el texto es muy corto como un mensaje de Twitter el contexto es casi inexistente. Por otra parte el texto en twitter tiene características muy especiales en cuanto al propio lenguaje usado y la sintaxis, lo cual constituye un desafío en sí mismo.

Pese a estas dificultades analizar la caracterización de los textos en twitter resulta de gran interés porque permite conocer lo que se "murmura" en la web sobre distintos asuntos de una manera muy rápida. El volumen de información en el micro blogging no solo es muy grande sino que es una cantera que se actualiza constantemente. Analizar la información de twitter por ejemplo puede ser usado para campañas políticas y publicidad sin necesidad de realizar encuestas y otro tipo de estudios más complejos y de mayor demanda de tiempo.

Existen otros estudios además de la caracterización de tópicos que es el análisis de sentimientos. No solo se desea saber cuál es el tema del texto sino la opinión que se tiene sobre el tema: positiva, negativa o neutra en los casos en que no se emita una opinión.

La polaridad de opinión es más compleja de analizar que los tópicos, porque es necesario detectar la ironía entre otros elementos que pueden dar indicaciones incorrectas.

Según (Liu, 2010):

La información textual puede ser categorizada en dos tipos principales: hechos y opiniones. Los hechos son expresiones objetivas acerca de entidades, 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,a, so,h, t) donde:

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

En este trabajo se pretende explorar la situación que afecta el análisis de opinión en tweeter explorando algunos problemas que se presentan y realizando una aproximación práctica.

El ensayo práctico consta de las siguientes etapas:

- Elegir un tema o un usuario para descargar tweets
- Descargar los tweets y recuperar datos sobre las entidades que son de interés:
  - Usuario productor
  - Fecha
  - Contenido del tweet
  - otros
- Recuperar las palabras del tweet y sus categorías sintácticas - PoS (Part of Speech)- . Se ha estudiado que algunas categorías detentan subjetividad y se utilizan para analizar el sentimiento del texto, específicamente adjetivos y adverbios. En nuestro caso estas palabras cargadas de subjetividad se analizarán junto con otros PoS que les aportan contexto, nombres y verbos. Los nombres aportan contexto al adjetivo y los verbos al adverbio.
- Realizar una anotación manual de estas palabras que expresan subjetividad enmarcadas en su contexto para confeccionar un lexicón de polaridad con contextualidad.
- Confeccionar otros productos como diccionarios de hashtags y terminología del dominio que incluya nombres y verbos.
- Realizar el análisis de polaridad de los tweets con las herramientas generadas
- Comparar el resultado con un servicio de análisis de sentimiento

Aproximaciones desde distintas disciplinas

La opinión en un tweet puede analizarse como un fenómeno comunicativo. Este fenómeno se puede estudiar para monitorear la opinión política o un producto que se va a lanzar al mercado. Sin duda existen aspectos éticos a considerar y no son menores. Los emisores de una opinión pueden no querer que se utilice su opinión con estos fines aunque el tweet sea público. Cuando se llena una encuesta se sabe que la misma va a ser utilizada con los fines de difusión de la opinión que se emite, pero en el caso del tweet si bien no es un mensaje privado, no se tiene tan clara esta perspectiva y existe una falta de límites entre lo que constituye el ámbito público y el ámbito privado. Muchas veces no se tiene una proyección de los posibles alcances de una opinión emitida sobretodo las nuevas generaciones para las cuales es una práctica instaurada y no se analizan las repercusiones.

Otro aspecto es la repercusión de un tweet dentro de la red. Si se estudia la red de difusión como una topología se pueden observar distintos comportamientos como la centralidad o la intermedialidad y

entonces el mensaje o la opinión difundida tendrán repercusiones y alcances diferentes. La importancia del emisor puede medirse también por la cantidad de seguidores y la cantidad de retweets que se realizan (seguidores de seguidores). Esto se puede estudiar a través del análisis de grafos.

Implementación de enfoques de Procesamiento de Lenguaje Natural por métodos informáticos

Gran parte 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 tweeter establecer el contexto es un desafío.

Para la determinación de la polaridad lo que se hace es buscar dentro del texto a estudiar la concordancia con una lista de palabras (lexicones) previamente elaborada que tiene establecida la polaridad.

Sentiwordnet [1] 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 [2] en inglés. Se puede traducir o mapear al español a través de http://multiwordnet.fbk.eu/english/home.php. Hay estudios (Montejo-Ráez, 2014) que utilizan este recurso para tratar la orientación de los tweets.

La Gramática sintagmática nuclear (HPSG-Head-driven phrase structure grammar), plantea que la oración tiene grupos nucleares y que el núcleo en un grupo selecciona los complementos en función de sus propiedades léxicas. Esto es una manera 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 Linguistic Knowledge Building (lkb) [3]

```
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:

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

Entonces puede plantearse 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 solo de adjetivos, sino de adjetivos en su contexto (nombre) y de adverbios en su contexto (verbos).

Ciencias de la información

Realizar una clasificación de tweets por temática o por dominio es una cuestión previa al análisis 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 v.g. un modelo de celular, los resultados serán bastante adecuados, pero en cuanto a un tópico más general empieza a producirse mucho ruido.

Una manera de resolver la cobertura es utilizar sinónimos para la búsqueda. Sin embargo 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 tweets afectará tanto la cobertura como la precisión.

En muchos estudios de sentimiento en tweetter se trabaja sobre dominios específicos como cine, etc., partiendo ya de un conjunto de datos clasificados.

(Vilares, 2014) recopila algunas iniciativas de clasificación de tópicos 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 [4] provee en español su caracterización.

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 tweet como "*arriba Defensor …*" constituye un caso claro de la necesidad de contexto, de un tema. En Uruguay un ser humano rápidamente asociaría que Defensor no es alguien que defiende sino el cuadro de fútbol "Defensor[5]". La extracción de tópicos permitiría poner Defensor bajo el tópico "Fútbol" y no bajo "Superhéroes".

En el artículo de (Gattani, 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 tweeter 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 tweet 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 de extracción de entidades, Named Entity recognition (NER) (Finkel, 2005), 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 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 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 (v.g. restaurant) y un atributo de esa entidad (v.g. precio) y hay una diferencia 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 restaurant, 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 de alguna manera componen la opinión sobre el objeto.

Un tesauro es un vocabulario controlado y estructurado formalmente, formado 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).

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 relacionados 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 de 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 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 es 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[6] . Esta interacción y reutilización se relaciona con la idea de la memética y la hipótesis de la replicación de las ideas en otros huéspedes en forma análoga a como se da 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 la 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 abarcativa. Una primera división es: Ciencia, Arte, Naturaleza y Sociedad.

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)

Estudio

El estudio a realizar buscará determinar la polaridad global de un tweet, no se considerará gradación de intensidad. La idea es clasificar un tweet como positivo, negativo o neutro.

Se creará un lexicón que registre la polaridad semántica de algunas palabras (adjetivos y adverbios) a través de un aprendizaje 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 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 lexicon de sentimientos. En el caso de los verbos en el lexicon 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 en este caso va a registrar la entidad a la que se aplica estableciendo un contexto de aplicación de la misma.

Los nombres son utilizados además para construir una terminología adhoc o social, que recopila las entidades y sus componentes.

Sin llegar a la construcción de un tesauro o una taxonomía se plantea 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 del trabajo 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 tweets y establecer el contexto para la polaridad de los adjetivos, adverbios y verbos. No obstante como una línea futura de trabajo puede plantearse la construcción no de una terminología adhoc, sino una taxonomía adhoc mapeando a Wikipedia o a 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 genere adhoc involucra, conceptos, instancias y verbos que sirven para dos cosas, la primera delinear la terminología del tópico y 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. Por ejemplo el adjetivo terrible es usado en lenguaje coloquial en ciertos ámbitos con una connotación positiva como en el caso de ”..terrible jugador” , que expresa que se trata de un jugador muy bueno. Entonces el problema es determinar la polaridad de este adjetivo, o es positivo o es negativo, lo que claramente no es neutro y dependerá del contexto en que se usa.

Si un lexicon de polaridad puede advertir sobre el objeto a que se califica tendrá mayor probabilidad de correctitud:

Terrible (jugador) positivo

Terrible (inundación) negativo

Metodología

Se extraen los tweets a través de una API. Tweeter usa un protocolo abierto OAuth (Open Authorization) con el cual se puede acceder a los tweets a través de una aplicación. Al registrar una aplicación en tweeter se obtienen las claves: consumer key y consumer secret que son las que vinculan a la aplicación. Con las access token y access token secret nos conectamos con la API y podemos descargar tweets.

Los tweets los podemos descargar por nombre de usuario o por tópico.

Dentro de las principales funciones de la API está la búsqueda de tweets con un determinado filtro (GET search/tweets), el que incluye idioma, geolocalización, fecha de inicio y fin de la búsqueda,

palabra que incluye el tweet, mensajes de un usuario específico, etc. Esta función está restringida a 7 días en el pasado y a 1500 mensajes, lo cual representa una limitante importante.

Otras funciones usadas para obtener un mayor número de tweets son las usadas vía *streaming*, donde se obtienen los mensajes en tiempo real, la cual, al igual que la anterior, puede incluir filtros tales como palabras, usuarios específicos, etc. Estas funciones no tienen restricción de número de mensajes, lo cual lo hace ideal para juntar una gran cantidad de información sobre un tema específico

(Liu, 2010) plantea una función para obtener la opinión de las distintas características o facetas de un objeto sobre el que se emite la opinión tomando para una oración s un conjunto de características f {f1, …fn} y un conjunto de palabras o frases que denotan opinión o {op1, ….opn} con su ranking de polaridad. La orientación de la opinión para cada característica fi en la oración s se forma con una función que realiza la sumatoria de las palabras de opinión por su ranking de polaridad divido la distancia entre la característica y la palabra de opinión.

En nuestro caso buscamos solo dar una indicación de la polaridad, no su gradación, por tanto las palabras de opinión se tomarán como positivas (1), negativas (-1) o neutras (0) y se realizará la sumatoria en la oración s.

Tampoco se tomarán en cuenta las características, sino el objeto con sus características como un todo y la opinión es una opinión global.

Luego de extraer la opinión de la oración se tomarán en cuenta algunas situaciones particulares como el caso de la negación que aplicada a una oración tiene implicancia sobre su polaridad. El criterio que se toma es que dada la polaridad de la oración sin el “no”, la aplicación del “no” invierte la polaridad.

Las tareas que desarrolla el ensayo luego de determinar los términos para la búsqueda son:

1. Descargar los tweets (Anexo NLTK)
   Se utiliza además el programa captura.py proveniente del curso Análisis Semántico de redes sociales (Cervantes, 2016)
2. Preprocesar ortográficamente
   a. sacar signos
3. Extraer hashtags si están al principio o al final para elaborar un diccionario de hashtags.
4. Confección manual de un lexicon de hashtags que consigne el hashtag y su polaridad (positivo, negativo o neutro)
5. Procesar sintácticamente a través de Freeling [7], extraer el PoS de las palabras del tweet que se utilizarán de la siguiente forma:
   a. los nombres y verbos para extraer la terminología del dominio (folksonomía)
   b. Los adjetivos y adverbios (palabras expresivas de polaridad) para entrenar el sentimiento con el contexto del nombre y el verbo.

San Salvador 1944
11200-Montevideo, Uruguay
Tel: [+598] 2413 80 20 / 21 / 22
fic@fic.edu.uy
www.fic.edu.uy

6. Asignación manual de polaridad a las palabras expresivas de polaridad (positivo, negativo, neutro) utilizando como contexto el nombre y el verbo
7. Procesar cada tweet utilizando el lexicon de polaridad generado y ajustar la polaridad tomando en cuenta características especiales
    a. Tratamiento de partículas negativas:

    Chequear la existencia de la palabra "no" y en caso de hallarla invertir la polaridad
8. Resultados a obtener:
    a. Folksonomía para modelar terminología del dominio
    b. Lexicon de palabras con sentimientos entrenado
    c. Tweets con resultado de sentimiento: positivo, negativo o neutro
    d. Diccionario de hashtags
9. Comparación con Meaning Cloud del resultado de la polaridad del tweet
10. Los resultados que pueden reutilizarse son:
    a. Nube de términos para categoría
    b. Lexicón de sentimientos entrenado

El conjunto de rutinas y los datos de prueba del prototipo se encuentran disponibles en abierto en: https://sourceforge.net/projects/pln-polaridad-en-twitter/files/polaridad/
El resultado es un prototipo que incluye rutinas y la prueba con un conjunto de datos. Este prototipo se realizó como un ejercicio práctico de propuesta de estrategia, utilizando distintas herramientas ya disponibles. Los resultados no son concluyentes a nivel estadístico y tienen la falencia metodológica que calculan la polaridad sobre el mismo conjunto entrenado previamente.
Las herramientas utilizadas se detallan en el anexo 1 y 2.

Conclusiones

El lenguaje es una identidad cultural y tiene una localidad muy importante incluso dentro de una misma lengua. El lenguaje usado incluso dentro de una misma zona geográfica es diferente de acuerdo al grupo social que lo utiliza que 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 tweeter 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 solo 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.

Las ciencias de la información tienen 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 y está vinculado con el contexto de las palabras usadas. 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) 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.

El estudio usando la metodología descripta arrojó un resultado que se equiparó al logrado con el uso de Meaning Cloud como se puede observar en el juego de prueba (incluído con las rutinas). Este resultado es meramente indicativo ya que se señalaron las falencias metodológicas, pero aún así abre la posibilidad de un estudio más profundo.

Otro trabajo a futuro sería abordar las locuciones verbales que tienen una alta carga de subjetividad.

La propuesta de contextualizar se funda en la hipótesis de que un lexicon de polaridad para redes sociales no puede ser genérico, el lexicon de polaridad está vinculado no solo 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 adhoc de lexicones para cada situación en particular, lexicones que se entrenen con una herramienta como la que se desarrolló y se usen ajustando, entallando para lograr una adecuación a la situación concreta.

San Salvador 1944
11200-Montevideo, Uruguay
Tel: [+598] 2413 80 20 / 21 / 22
fic@fic.edu.uy
www.fic.edu.uy

Esta propuesta es practicable puesto que un servicio como Meaning Cloud permite y alienta la combinación con 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 micro escala. Este modelado constituye 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 tweets así como su polaridad constituyen indicadores significativos que pueden ser tomados en cuenta para el desarrollo y mejora de productos, políticas o acciones aunque esto se constituye dentro de un necesario debate ético.

El enfoque abre una problematización multidisciplinar de la polaridad en micro blogging que puede ser de interés para el avance en su estudio y también realiza una herramienta para su instanciación en la práctica.

Bibliografía

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

Cervantes, Ofelia (2016). Curso Análisis semántico de redes sociales Disponible en: https://eva.fing.edu.uy/course/view.php?id=939

Finkel, Jenny, Grenager, Trond, Manning, Christopher (2005) Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

Hu, X., Tang, J., Gao, H., & Liu, H. (2013, May). Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on World Wide Web (pp. 607-618). ACM.

Kouloumpis, E., Wilson, T., & Moore, J. D. (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 Ch. (2010) NLP Handbook, University of Illinois, Chicago.

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

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

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

Sag, I. A., Wasow, T., Bender, M. E. (2003) Syntactic Theory. A Formal Intoduction. 2ª edición. CSLI Publications. Stanford.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter.

Selva Castelló, Javier (2015). Desarrollo de un sistema de análisis de sentimiento sobre Twitter (Doctoral dissertation).

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

Tillett, B. (2005). ¿Qué es FRBR? : *Un modelo conceptual del universo bibliográfico. https://www.loc.gov/catdir/cpso/Que-es-FRBR.pdf*.

Urizar, X. S., & Roncal, I. S. V. (2013). Elhuyar at TASS 2013. In Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS 2013) (pp. 143-150).

Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (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. Universidad da Coruña, Facultad de Informática, España.

Anexo 1. Herramientas utilizadas

NLTK

Se utilizó Python y la librería NLTK que contiene prestaciones para tokenizar, lematizar, anotación POS (Part of Speech), parsear, etc.

Se utilizó el paquete twitter de NLTK[8]. Este paquete incorpora una serie de prestaciones que son muy prácticas y se detallan en el Anexo NLTK. NLTK utiliza la modalidad de recolección de datos de tweeter a través de la API que está disponible[9]. El procedimiento es registrarse y obtener las cuatro claves que luego pueden ser utilizadas para recolectar tweets. Los elementos más destacados se incluyen en el anexo 2 NLTK.

Freeling

Se utiliza Freeling a través de su demo en la web[10], levantando los resultados obtenidos para cada tweet configurando las opciones de: number recognition, date/time recognition, quantities/ratios and percentages, named entity detection, named entity classification, multiword detection. Se utiliza la salida para etiquetado PoS en formato CoNLL.

Algunos comportamientos detectados en Freeling a señalar son, el de considerar algunas expresiones compuestas como “mientras que “ que son tratadas como una solo término (mientras_que), pero es posible que haya algunas expresiones que no estén consideradas.

En el caso del hashtag pegado a una palabra, Freeling separa el # de la palabra y considera el # como un signo –sin darle tratamiento especial- y la palabra la trata normalmente, lo cual resulta apropiado en el caso de la palabra a la interna de la oración.

Freeling no reconoce las abreviaturas que son muy utilizadas en tweeter.

MeaningCloud

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, etc. El caso de Meaning Cloud [11] es una de ellas y permite utilizar su API a través de una clave de acceso [12] para realizar el análisis de sentimiento (Cervantes, 2016). A través de la API o de un endpoint se puede realizar:

- análisis de sentimiento: Se ingesta un texto y como salida se indica su polaridad, subjetividad, ironía o especificación de desacuerdo
- Extracción de tema: identifica entidades nombradas y conceptos

- Clasificación de texto: Clasifica un texto de acuerdo a una taxonomía

Meaning Cloud ofrece APIs para distintos escenarios. Estas APIs incluyen diccionarios, taxonomías y otras funcionalidades. Para utilizar MeaningCloud hay que registrarse como desarrollador y obtener una clave de acceso al API de MeaningCloud y luego realizar una petición con la clave obtenida.

Se utilizó esta prestación para analizar los tweets:

```
def sentimiento(tweet):
 url = "http://api.meaningcloud.com/sentiment-2.1"
 headers = {'content-type': 'application/x-www-form-urlencoded'}
 payload = "key=….&lang=es&txt=" + str(tweet) + "&model="
 response = requests.request("POST", url, data=payload, headers=headers)
 return response
```

El resultado del análisis muestra lo siguiente:

- status: Contiene información acerca del proceso de extracción.
  - code: Número que indica el estado del proceso.
  - msg: Una cadena de texto que describe el estado del proceso.
  - credits: Número que indica los créditos consumidos en la petición.
  - remaining_credits: Número de créditos restantes.
  - model: modelo utilizado para la evaluación.
- score_tag: Indica la polaridad encontrada en el texto.
  - P+: strong positive
  - P: positive
  - NEU: neutral
  - N: negative
  - N+: strong negative
  - NONE: without sentiment

La API puede configurarse para detectar si el texto es objetivo o subjetivo y si contiene marcas de ironía dando información sobre la confiabilidad de la polaridad obtenida del análisis de sentimientos.

Los elementos que aparecen a configurar son:

- **agreement:** Este campo indica el acuerdo entre los sentimientos detectados en el

texto.

- AGREEMENT: Los elementos tienen la misma polaridad.
- DISAGREEMENT: Existe desacuerdo entre la polaridad de los elementos.

❍ **subjectivity**: Indica la subjetividad en el texto.

- OBJECTIVE: El texto no tiene marcas de subjetividad.
- SUBJECTIVE: El texto tiene marcas subjetivas.

❍ **confidence:** Representa la confianza asociada con el análisis de sentimientos realizada al texto.

- Su valor es un número entero en el rango de 0-100.

❍ **irony:** Indica la ironía en el texto.

- NONIRONIC: El texto no tiene marcas de ironía.
- IRONIC: El texto tiene marcas de ironía

Los errores que se pueden dar en la respuesta están establecidos en la página[13] Y son los siguientes:

101: License expired

102: Credits per subscription exceeded

103: Request too large

104: Request rate limit exceeded

200: Missing required parameter(s) - [name of the parameter]

201: Model not supported

202: Engine internal error

203: Cannot connect to service

204: Model not suitable for the identified text language

205: Language not supported

206: Number of IDs does not match number of texts

207: Mode not supported

Meaning Cloud permite usar recursos propios en el análisis de sentimiento de entidades y conceptos creando diccionarios propios y también se puede personalizar el modelo de análisis de sentimiento. El usuario puede definir sus propias entidades y conceptos y aplicarlos a cualquier tipo de escenario. El

sitio establece sus propósitos académicos y ofrece accesos mayores de los servicios para lo cual solicita una descripción del uso de la entidad universitaria.

Sostienen que el uso de diccionarios de usuario es favorable porque la API que ofrecen se limita al uso de recursos genéricos con lo cual la precisión puede verse afectada por el uso en dominios específicos, lo cual es justamente la hipótesis que planteamos en el trabajo.

Anexo 2. Uso de NLTK

La librería Tweepy incluye funciones adecuadas para la conexión, descarga y tratamiento de los twits

Las claves para trabajar se obtienen de la api https://apps.twitter.com

app_key= CONSUMER KEY
app_secret= CONSUMER SECRET
oauth_token= ACCESS TOKEN
oauth_token_secret= ACCESS TOKEN SECRET
.
Para extraer tweets con determinadas palabras – palabra1, palabra2- hasta un límite de 10:

```
from nltk import Twitter
tw = Twitter()
tw.tweets(keywords='palabra1, palabra2', limit=10 )
```

Si se quiere extraer tweets de una cuenta en especial a través de su id. numérico:

```
tw = Twitter()
tw.tweets(follow=['759251', '612473'], limit=10 )
```

Se puede recolectar a través de la modalidad OAuthHandler y se obtienen un archivo.json.
Se instaló twython que es un paquete de terceros que se apoya en este paquete:
Lo que se encontró más útil es el paquete json2csv[14] que permite pasar un archivo json a csv.
Ejemplo:

```
from nltk.twitter.common import json2csv
json2csv('descarga.json', 'tweets_ejemplo.csv',
['created_at', 'favorite_count', 'id', 'in_reply_to_status_id' ,
'in_reply_to_user_id', 'retweet_count', 'retweeted' ,
'text', 'truncated', 'user.id'])
```

Este ejemplo levanta de un archivo descarga.json que contiene los tweets y genera un archivo tweets_ejemplo.csv con los campos:

```
'created_at', 'favorite_count', 'id', 'in_reply_to_status_id', 'in_reply_to_user_id',
'retweet_count', 'retweeted','text', 'truncated', 'user.id'
```

Tweeter recurpera elementos que denomina entidades (entities) y lugares.

12. "entities" Entities provide metadata and additional contextual information about content posted on Twitter. Entities are never
divorced from the content they describe. In API v1.1, entities are returned wherever Tweets are found in the API. Entities are
instrumental in resolving URLs.
13. "entities"" user_mentions" (Array of Object) Represents other Twitter users mentioned in the text of the Tweet.

San Salvador 1944
11200-Montevideo, Uruguay
Tel: [+598] 2413 80 20 / 21 / 22
fic@fic.edu.uy
www.fic.edu.uy

"entities" "user_mentions"" id" (Int)ID of the mentioned user, as an integer.
14. "entities""user_mentions""indices" (Array of integer) An array of integers representing the offsets within the Tweet text
where the user reference begins and ends. The first integer represents the location of the '@' character of the user mention.
The second integer represents the location of the first nonscreenname
character following the user mention.
15. "entities""user_mentions""id_str" (Int) Id of the mentioned user, as a Int.
16."entities""user_mentions"" of the referenced user.
17. "entities""user_mentions""name" (String) Display name of the referenced user.
18. "entities""symbols"
19. "entities""hashtags"(String) Represents hashtags which have been parsed out of the Tweet text.
20. "entities""urls"(String) Represents URLs included in the text of a Tweet or within textual fields of a user object.

Para recuperar las entidades:

```
from nltk.twitter.common import json2csv_entitie s
json2csv_entities('descarga.json', 'tweetsConHashtags.csv' ,
['id', 'text'], 'hashtags', ['text'])
json2csv_entities('descarga.json', 'tweets_user_mentions.csv' ,
['id', 'text'], 'user_mentions', ['id', 'screen_name'] )
json2csv_entities('descarga.json', 'tweets_media.csv' ,
['id'], 'media', ['media_url', 'url'])
json2csv_entities('descarga.json', 'tweets_urls.csv' ,
['id'], 'urls', ['url', 'expanded_url'])
json2csv_entities('descarga.json', 'tweets_place.csv' ,
['id', 'text'], 'place', ['name', 'country'] )
json2csv_entities('descarga.json', 'tweets_place_bounding_box.csv' ,
['id', 'name'], 'place.bounding_box', ['coordinates'] )
```

Cuando un tweet es un retweet, el tweet original puede ser incorporado al mismo archivo haciendo lo siguiente:

```
json2csv_entities('descarga.json', 'tweets_original_tweets.csv' ,
['id'], 'retweeted_status', ['created_at', 'favorite_count' ,
'id', 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweet_count' ,
'text', 'truncated', 'user.id'])
```

El primer id corresponde al tweet retweeteado y el segundo al tweet original

Una cuestión a considerar es la elección de un número manejable de tweets, que den una cobertura apropiada, para lo cual una propuesta es ponderar algunos elementos que se detallan:

- el autor del tweet, considerando si es un influencer, es decir si tiene un número importante de seguidores, porque si él hace un comentario éste les llegará a sus seguidores. Se utiliza el atributo user_followerscount.
- si el tweet ha sido retweeteado, lo cual seguramente indica que hay un afiliación activa al comentario (podría ser que se retweetee como un exponente de algo en contrario, pero en cualquier caso el comentario mereció ser señalado). Se utiliza el atributo retweet_count

Se extraen los sigs. elementos:

- 3. "text" (String) Tweet content
- 6. "id" (Int64) The integer representation of the unique identifier for this Tweet. This number is greater than 53 bits and some programming languages may have difficulty/silent defects in interpreting it. Using a signed 64 bit integer for storing this identifier is safe.

● 10. "coordinates"(floats) The longitude and latitude of the Tweet's location, as an collection in the form of [longitude, latitude].
● 23. "retweet_count" (Int) Number of times this Tweet has been retweeted. This field is no longer capped at 99 and will not turn into a String for "100+"
● 35. "user"" followers_count" (Int) The number of followers this account currently has. Under certain conditions of duress, this field will temporarily indicate "0."

NOTAS

[1] http://sentiwordnet.isti.cnr.it/
[2] http://wordnet.princeton.edu/
[3] http://www.swmath.org/software/20948
[4] http://www.sepln.org/workshops/tass/2013/papers/tass2013-submission3-Elhuyar.pdf
[5] https://es.wikipedia.org/wiki/Defensor_Sporting_Club
[6] https://del.icio.us/
[7] http://nlp.lsi.upc.edu/freeling/node/1
[8] http://www.nltk.org/howto/twitter.html
[9] https://apps.twitter.com
[10] http://nlp.lsi.upc.edu/freeling/demo/demo.php
[11] https://www.meaningcloud.com/es
[12] https://www.meaningcloud.com/developer/
[13] https://www.meaningcloud.com/developer/documentation/error-codes
[14] jsoncsv está en common no util como indica la documentación de nltk