

Análisis Automatizado de Sentimiento en Textos Breves de la Plataforma Twitter

Fernando Balbachan

Facultad de Filosofía y Letras, Universidad de Buenos Aires (UBA)
Buenos Aires, Argentina
Socialmetrix - NLP Research
fbalbachan@socialmetrix.com

Diego Dell'Era

Facultad de Filosofía y Letras, Universidad de Buenos Aires (UBA)
Buenos Aires, Argentina
Socialmetrix - NLP Research
ddellera@socialmetrix.com

Abstract

In Computational Linguistics or Natural Language Processing (NLP), Sentiment Analysis in opinionated text is one of the most challenging standard tasks. Although some unsupervised learning approaches based on machine learning use statistical techniques such as Bayesian classifiers, semantic-oriented bigrams, etc. [1], the most successful works in the field lean on lexical-grammar resources under the form of ontologies with sentiment values and several basic syntax rules.

Our solution is based on such symbolic paradigm of lexical-grammar resources and local and global syntax rules. We detect early hints such as emoticons in order to come up with sentiment verdicts. Then, we pre-process the text by standardizing it and segment it into significant units smaller than sentences. Those small units are processed by a lemmatizer (Freeling) with POS-tagging optimized for Spanish. Thus, we get accurate lemmata for each unit of analysis. Those lemmata are valued by our own ontology -similar to SentiWordNet [2]- with more than 4,000 hand-annotated lemmata. Finally, we apply rules for modal changes in negated structures or modal subjunctive and polarity shifters for phrases such as *sin respeto*, *imposible de enojarse*, etc.

Keywords: Sentiment Analysis, Opinion mining, Ontologies, Freeling, SentiWordNet.

Resumen

En el área de la Lingüística Computacional o Procesamiento de Lenguaje Natural (PLN), una de las tareas estándares más desafiantes es el análisis de sentimiento (*sentiment analysis*) en texto *opinionado*. Aunque algunos enfoques de aprendizaje no supervisado (*machine learning*) hacen uso de técnicas estadísticas como clasificadores bayesianos, bigramas de orientación semántica, etc. [1], los trabajos más exitosos en el campo recurren a recursos léxico-gramaticales bajo la forma de una ontología con valoración de sentimiento y diversas reglas de sintaxis básica.

Nuestra solución sigue tal enfoque simbólico de recursos léxico-gramaticales y reglas de sintaxis

local y global. En forma temprana detectamos indicios como emoticones para dar veredictos de sentimiento. Luego, pasamos a la etapa de pre-procesamiento de texto: estandarización del texto y segmentación en unidades significativas menores a la oración, las cuales pasan a ser procesadas por un lematizador optimizado para el español (Freeling) con anotación morfosintáctica (*POS-tagging*). De esta manera, obtenemos lemas muy confiables por cada unidad de análisis, los cuales son valorados por nuestra propia ontología -similar a SentiWordNet [2]- con más de 4.000 lemas anotados a mano. Finalmente, aplicamos reglas de cambios de modalidad para estructuras negadas y subjuntivo con modalidad irreal y reglas de polaridad para frases como *sin respeto, imposible enojarse*, etc.

Palabras claves: Análisis de sentimiento, Minería de texto *opinionado*, Ontologías, Freeling, SentiWordNet.

1. INTRODUCCION

En el área de la Lingüística Computacional o Procesamiento de Lenguaje Natural (PLN), una de las tareas estándares es el análisis de sentimiento (*sentiment analysis*) en texto *opinionado*. La dificultad de la tarea se percibe en que la medida de efectividad de estos mecanismos computacionales no llega, en general, al 85%. Incluso para los propios hablantes resulta a veces problemático dar un veredicto de análisis de sentimiento sobre un texto *opinionado*, ya que el soporte lingüístico de nuestras opiniones involucra frecuentemente fenómenos simultáneos pertenecientes a la estructura sintáctica, semántica, pragmática y discursiva del lenguaje, y aun de nuestro conocimiento de mundo [1].

Otro aspecto crucial para encarar esta tarea es la falta de disponibilidad de recursos. Por un lado, antes de la masiva irrupción de Internet y de las plataformas de redes sociales (Facebook, Twitter), el interés en un análisis computacional de opiniones era escaso [3]. Sólo recientemente diversas compañías y organizaciones académicas han incursionado en este tipo de proyectos de investigación como punto de partida para lo que se conoce en marketing como *Social Media Monitoring*, tomando en cuenta el inmenso volumen de información disponible bajo la forma de opiniones de clientes, prospectos de clientes, competidores, etc. Esto explica la notoria falta de recursos léxicos para lenguas que no sean inglés, e incluso la escasez de los mismos para esta lengua.

Finalmente, en lo que hace a este trabajo particular de análisis de sentimiento aplicado a textos breves de Twitter (tweets) en español, debemos destacar el desafío que significa trabajar con texto espontáneo, plagado de errores ortográficos y cierto desapego a las normas gramaticales (especialmente en lo que atañe a puntuación).

2. ANTECEDENTES

2.1. En los inicios de la tarea de análisis de sentimiento

Los primeros trabajos en el campo de análisis de sentimiento fueron desarrollados a partir de 2000.

Los enfoques en esta tarea apelan tanto al paradigma simbólico de recursos léxicos y reglas de manipulación de símbolos al paradigma estadístico de aprendizaje automático.

Entre los primeros enfoques simbólicos, muchos investigadores demostraron un gran ingenio para lidiar con la escasez de recursos anotados que sirvieran de guía inicial para la valoración de

opiniones. Por ejemplo, Kamps y Marx (2002) utilizan distancia semántica o *Minimal Path Length* MPL (distancia entre nodos) en WordNet¹ (una ontología computacional sin valoración de sentimiento disponible para varios idiomas) entre una palabra blanco, cuya orientación semántica se quiere conocer, y los adjetivos *good* (*bueno*) y *bad* (*malo*). Por ejemplo, la MPL entre *good* y *proper* es 2.

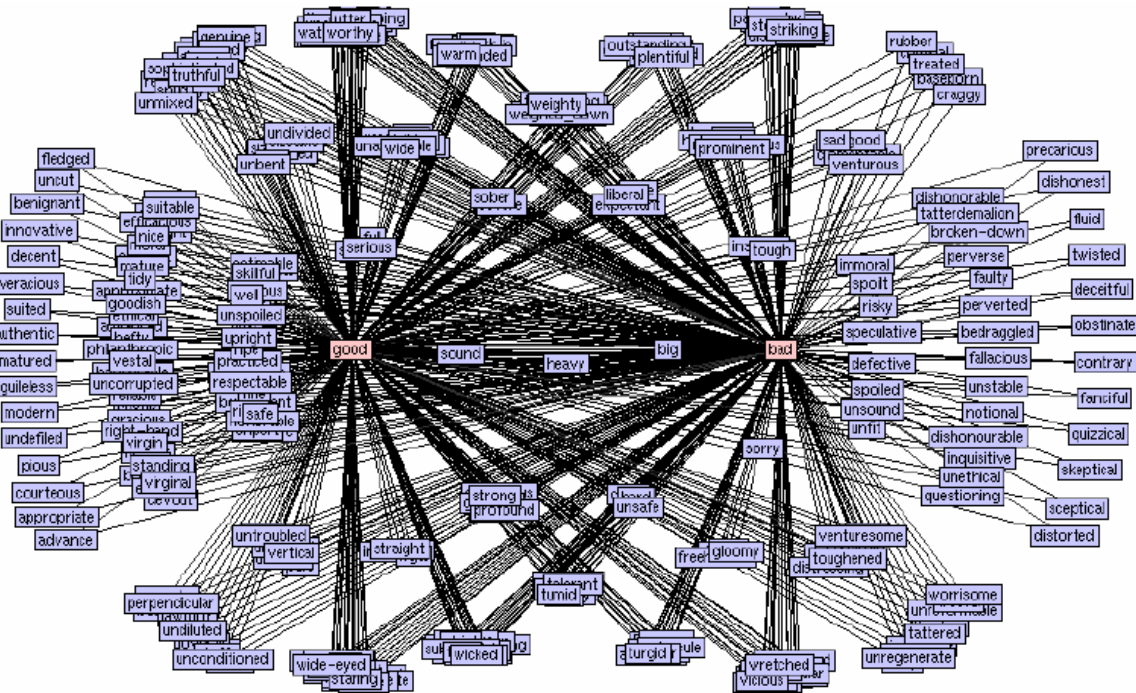


Figura 1: *Minimal Path Lengths* MPLs para algunos adjetivos del inglés, según Kamps y Marx (2002).

Si bien el trabajo de Kamps y Marx (2002) fue pionero en mostrar cómo computar valoraciones subjetivas de texto *opinionado*, la efectividad y el alcance de este enfoque son poco significativos, como los propios autores reconocen [4]. En primer lugar, llega un punto de separación de adjetivos en el que la distancia entre nodos ya no es índice confiable de valoración axiológica entre *bueno* y *malo* (la MPL de *noble* para *good* y para *bad* es 4). En segundo lugar, la ocurrencia de adjetivos axiológicos puede verse afectada por el alcance de un operador modal (negación, cambio de polaridad, etc.). Finalmente, considerar que el soporte lingüístico para opiniones está únicamente basado en adjetivos axiológicos es una concepción demasiado reduccionista: tal como sostiene la teoría de los subjetivemas [5], las palabras que actúan de soporte de la valoración subjetiva del hablante respecto del enunciado pueden ser verbos, sustantivos, adverbios y adjetivos.

Los seguidores del paradigma estadístico también aportaron trabajos iniciáticos para esta tarea. Turney (2002) [6] describe un clasificador no supervisado basado en la orientación semántica de bigramas extraídos mediante simples patrones morfosintácticos en un texto. La orientación semántica es calculada mediante el algoritmo PMI-IR (*Pointwise Mutual Information*). Este enfoque explota la poderosa noción de información mutua (probabilidad de co-aparición de dos términos) en función de dos conjuntos de palabras “semillas” positivas y negativas. Turney [6] reporta una medida $F = 84\%$. No obstante, como mencionan Cruz *et al.* [1] esta aparentemente alta efectividad puede deberse a una sobreadaptación al corpus de evaluación, compuesto de reseñas de automóviles en inglés. El género de las reseñas como textos breves presenta un cierto sesgo hacia una opinión polarizada, sea ésta positiva o negativa.

¹ <http://wordnet.princeton.edu/>

2.2. Trabajos recientes y análisis de sentimiento en español

Más recientemente, la tarea de análisis de sentimiento atrajo un mayor interés por soluciones más integradoras y abarcativas. Los trabajos más exitosos recurrieron a recursos léxico-gramaticales bajo la forma de una ontología con valoración de sentimiento y diversas reglas de sintaxis básica.

Enmarcado en el paradigma simbólico, Taboada, Brooke *et al.* (2010) [7] desarrollaron el *Semantic Orientation Calculador* (SO-CAL), un conjunto de recursos léxicos con valoración de sentimiento que se ven afectados por intensificadores, índices de modalidad (negación y contrafáctica) y reglas de sintaxis básica. Esta solución integral para el inglés ha sido evaluada en distintos *corpora* de reseñas de dominio específico (hotelería, tecnología, películas, etc.) reportando medidas F que varían entre 74% y 90%, resultados auspiciosos por demás.

Estos enfoques basados en recursos léxicos guiaron nuestra investigación inicial hacia un mecanismo de análisis de sentimiento para el español. Desafortunadamente, nos topamos con el obstáculo de la ausencia total de recursos léxicos para este idioma. Desde 2007, se encuentra disponible un poderoso recurso léxico para el inglés: SentiWordNet², una ontología con valoración de sentimiento en 3 dimensiones (positivo, negativo, neutral) y una amplia cobertura de acepciones. Definitivamente necesitábamos de algo similar para el español. Diversos trabajos en lenguas que no sean el inglés han apelado a enfoques que incluían etapas de traducción automática, pero sin lograr un éxito mayor.

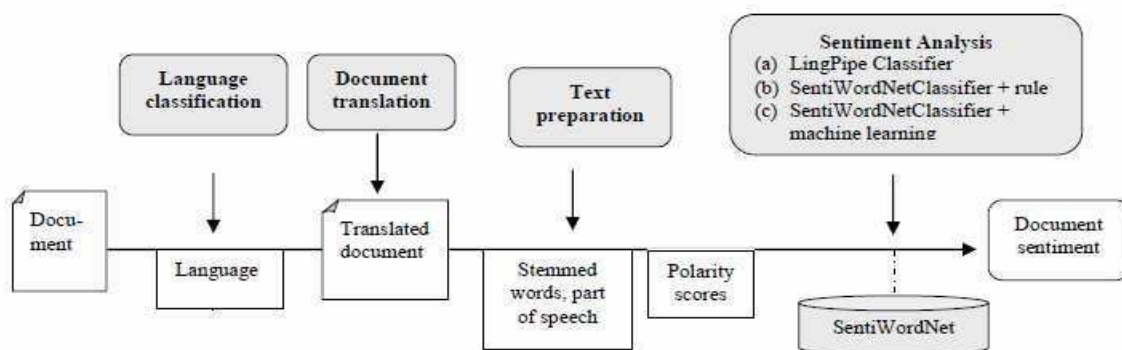


Figura 2: Procesamiento para análisis de sentimiento en idiomas que no sean el inglés (Denecke 2008).

3. NUESTRA SOLUCIÓN

3.1. Arquitectura general de la solución

Nuestra solución integral, desarrollada para Socialmetrix³, sigue el enfoque simbólico de recursos léxico-gramaticales y reglas de sintaxis local y global. En forma temprana detectamos indicios como emoticones para dar veredictos de sentimiento. Luego, pasamos a la etapa de pre-procesamiento de texto: estandarización del texto y segmentación en unidades significativas menores a la oración, las cuales pasan a ser procesadas por un lematizador optimizado para el español (una de las poderosas herramientas de análisis de la suite *open source* y multilinguaje Freeling⁴) con anotación morfosintáctica (*POS-tagging*). De esta manera, obtenemos lemas muy confiables por cada unidad de análisis, los cuales son valorados por nuestra propia ontología de

² <http://sentiwordnet.isti.cnr.it>

³ <http://socialmetrix.com>

⁴ <http://nlp.lsi.upc.edu/freeling/> y <http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf>

valoración de sentimiento en español -similar a SentiWordNet [2]- con más de 4000 lemas anotados a mano. Finalmente, aplicamos reglas de cambios de modalidad para estructuras negadas y subjuntivo con modalidad irreal y reglas de polaridad para frases como *sin respeto*, *imposible enojarse*, etc.

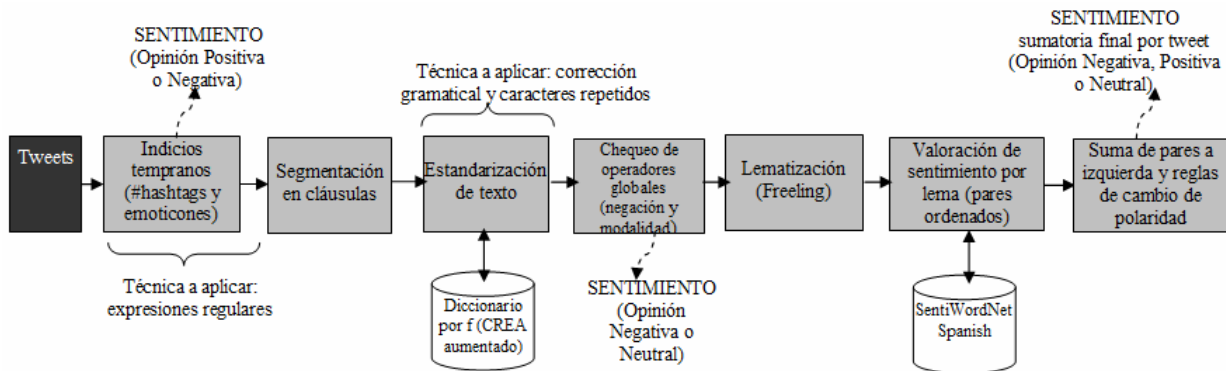


Figura 3: Arquitectura general de la solución de análisis de sentimiento en español para Twitter

3.2. Descripción detallada de cada etapa de procesamiento

3.2.1. Indicios tempranos (hashtags y emoticones)

En el uso del lenguaje en Internet [8] es una práctica común que los usuarios marquen su propia valoración subjetiva respecto de lo expresado a través de emoticones, como por ejemplo :) o :(, o hashtags (#WTF, #FTW, etc.). Este recurso supralingüístico a menudo no ha sido muy tomado en cuenta por la academia para el procesamiento de lenguaje natural en modelizaciones más abarcativas. No obstante, diversas implementaciones de la industria se basan exclusivamente en este tipo de indicios (por ejemplo, Tweetfeel⁵). En nuestro caso, al estar trabajando con opiniones expresadas en textos breves (tweets) de la plataforma Twitter, estadísticamente el 5% de los datos a analizar en cualquier set de datos hacían uso de este recurso.

Por un lado, la incidencia de estos indicios tempranos de emotividad sobre el set de datos totales no es estadísticamente despreciable. Por otro lado, disponemos de un método trivial para reconocer o identificar estas marcas a través de expresiones regulares, sin hacer uso de ningún recurso léxico o regla adicional. Así pues, logramos un porcentaje de efectividad muy alto (entre 85% y 90%) en la tarea de determinar si el sentimiento del tweet es positivo o negativo, porque en general los indicios positivos y negativos no se superponen ni se contradicen.

En esta etapa, no obstante, no podemos dar cuenta del estado NEUTRAL para los tweets, ya que no existen indicios tempranos de tal marcación subjetiva. La gran masa de todos los tweets que no incluyen estos indicios (aproximadamente 95%) debe continuar su camino en el procesamiento de cascada, tal como se describe en la figura 3.

3.2.2. Segmentación

La segmentación del texto del tweet involucra decisiones de diseño fundamentales. Las propias convenciones de marcación en Twitter para retweets (RT), menciones, recomendaciones (#followfriday) vínculos (links) y vínculos acortados obligan a tomar ciertas precauciones a la hora de aislar el soporte lingüístico que podría conllevar texto opinado.

⁵ <http://www.tweetfeel.com>

En el caso de nuestro algoritmo, decidimos segmentar la unidad de análisis en base a una lista de signos de puntuación y conjunciones coordinantes y subordinantes. Así pues, la unidad de análisis resulta típicamente menor que una oración. Como se verá más adelante, uno de los inconvenientes más difíciles de subsanar a la hora de analizar sentimiento en función de la ocurrencia de subjetivas es el alcance de negaciones y operadores de modalidad sobre los mismos (véase sección 3.2.6 *Operadores globales y operadores locales de cambios de polaridad*).

- (1) [*Te felicito*] *porque* [*has mejorado*]. (valoración positiva y positiva, respectivamente)
- (2) [*No me alegra*] *que* [*hayas mejorado*]. (valoración neutra o negativa y positiva, respectivamente)
- (3) [*Me alegra*] *que* [*no hayas mejorado*]. (valoración positiva y neutra o negativa, respectivamente)
- (4) [*Mejoraste*] *pero* [*todavía me das asco*]. (valoración positiva y negativa, respectivamente)
- (5) [*No mejoraste*]. (valoración neutra o negativa)

Esta problemática también se da en español con índices de modalidad irreal (subjuntivo irreal) o del enunciado o volitiva de la enunciación (modo imperativo) [5].

- (6) [*Dudo*] *que* [*haya mejorado*]. (valoración neutra y neutra -por subjuntivo irreal-, respectivamente)
- (7) [*Mejorá*]. (valoración negativa o neutra)

Existe bastante bibliografía especializada en el problema específico de detectar el alcance de las negaciones, desde enfoques simbólicos con expresiones regulares y diccionarios de frases negativas hasta métodos estadísticos de *machine learning* como *Support Vector Machine* (SVM) o clasificadores bayesianos ingenuos [9]. El trabajo de Goryachev *et al.* (2008) [9] demuestra que los enfoques simbólicos tienen nuevamente una efectividad mayor que la de los métodos estadísticos para determinar el alcance de las negaciones.

En este sentido, y ante la falta de un mecanismo que entendiera la estructura sintáctica de la oración a alto nivel (*chunking* o *shallow parsing*), nos vimos obligados a minimizar la extensión de las cláusulas de modo tal de garantizar que el alcance de operadores de modalidad no excediera el ámbito de la localidad inmediata de ocurrencia de la partícula negativa hacia la derecha (cambio de polaridad para expresiones privativas como las encabezadas por *sin*) o el ámbito de la globalidad de la cláusula (operadores globales de negación o de modalidad). Para mayor información, véase la sección 3.2.6 *Operadores globales y operadores locales de cambios de polaridad*.

3.2.3. Estandarización de texto (*distancia mínima de edición y caracteres repetidos*)

El uso del lenguaje en plataformas de redes sociales en Internet representa un interesante desafío para la modelización en lingüística computacional. Al tratarse de escritura espontánea y creativa con marcados rasgos de oralidad, “la ortografía resulta sumamente variable y [...] ha sido despojada de su función normalizadora y puesta al servicio de la expresividad de los usuarios” [8]. En este sentido, nuestro proceso de estandarización del texto debe abarcar no sólo una tarea de corrección ortográfica, sino también una tarea de análisis de caracteres repetidos (recurso muy tipificado como intensificador de carga emotiva o expresión enfática).

Con respecto al recurso de intensificación emotiva a partir de la repetición de caracteres, especialmente en el caso de vocales (por ejemplo, “*te amooooo*”), implementamos un control basado en expresiones regulares con *back reference* que reemplaza la ocurrencia de dos o más de estos caracteres por uno solo, a excepción de los grupos *cc,ll,rr*.

Para la tarea de corrección ortográfica, recurrimos al algoritmo de Levenshtein [10] y su noción de “distancia”. Se define como distancia a la medida en que dos términos difieren en cuanto los caracteres que los componen. Así, una distancia de valor 0 indica que dos términos son idénticos,

mientras que valores mayores de distancia indican una creciente disimilitud. La máxima distancia está acotada por la longitud del término más extenso. El algoritmo calcula la distancia entre dos términos como la mínima cantidad de operaciones de edición (inserción, borrado y sustitución) necesarias para convertir a un término en el otro.

En la tabla 1 los términos se representan como secuencias de caracteres alineados, y las operaciones de edición como una matriz con los “costos” acumulados. Si bien es posible asignar un costo diferencial a cada una de estas operaciones, para penalizar ciertos cambios considerados menos frecuentes en el ámbito del que provengan los términos, pruebas preliminares demostraron que se podían obtener iguales resultados con un valor uniforme de 1 para todas las operaciones.

		T	E	C	O	R
		1	2	3	4	5
T	1	0	1	2	3	4
E	2	1	0	1	2	3
M	3	2	1	1	2	3
O	4	3	2	2	1	2
R	5	4	3	3	2	1

Tabla 1: Distancia mínima de edición entre la no-palabra TECOR y la palabra TEMOR con un valor acumulado de 1

Con excepción de los nombres propios, reconocidos por el uso de mayúsculas, una no-palabra es un token alfabético que no ha sido validado contra nuestro diccionario, el cual se compone de la lista completa de formas del *Corpus de Referencia del Español Actual* (CREA)⁶ de la Real Academia Española (RAE), ordenada por frecuencia de uso y aumentada con las formas conjugadas del español rioplatense (voseo), totalizando aproximadamente 190.000 formas. La no-palabra genera una serie de posibles candidatos a palabras en base a las sustituciones, inserciones y borrados sucesivos de cada carácter que la componen (distancia de edición igual a 1). Nuestro algoritmo reemplaza la no-palabra con la palabra que surja de esta lista de candidatos a distancia de edición igual a un carácter y que esté en nuestro diccionario con frecuencia más elevada. Obviamente, este método de corrección ortográfica resulta más confiable a medida que la no-palabra es más extensa y, por tanto, existen menos candidatos a distancia de un carácter que puedan resultar palabra.

3.2.4. Lematización (Freeling)

La lematización es el proceso de obtención de una forma base (*lema*) a partir de formas flexionadas y derivadas morfológicamente [11]. En lenguas como el español, donde el paradigma verbal y flexivo es muy extenso, este proceso de modelización adquiere vital importancia a la hora de minimizar esfuerzos de anotación de recursos léxicos y maximizar la cobertura de los mismos.

Numerosos trabajos en lingüística computacional recurren a herramientas especializadas como Freeling para este tipo de tareas [1][12]. En nuestro caso aprovechamos la salida morfosintácticamente etiquetada de Freeling como un lematizador optimizado. Freeling resuelve muy bien el escollo de la ambigüedad morfosintáctica del español, mediante un *POS-tagger* (etiquetador morfosintáctico) a partir de reglas simbólicas y cálculo de probabilidades. No obstante, la variedad dialectal del español rioplatense representa un problema, porque las formas verbales pertenecientes al pronombre personal *vos*, cuyo uso se encuentra extendido en toda la población, no son reconocidas por el español de Freeling. Por lo tanto, configuramos el archivo *afixos.dat* de la herramienta para obtener el verbo en infinitivo a partir de las formas conjugadas de voseo.

⁶ <http://corpus.rae.es/lfrecuencias.html>

Otra modificación al proceso de lematización de Freeling consistió en derivar por afijación los adverbios terminados en *-mente*, ya que nuestra ontología de valoración de sentimiento (*SentiWordNetSpanish*) incluía los primeros y no los últimos. Cabe recordar que el uso de adverbios de modo terminados en *-mente* (*afortunadamente*, *desgraciadamente*, etc.) está muy difundido como índice de modalidad del enunciado [5] y, por lo tanto, resulta un excelente vehículo para expresar la opinión del hablante.

En el diseño de nuestra etapa de lematización nos topamos con otro obstáculo. El módulo de reconocimiento de entidades nombradas (*Named-Entity Recognition* NER) de Freeling considera como una única unidad varios tokens seguidos que comiencen con mayúscula o estén unidos por partículas unitivas (por ejemplo, *Banco de Bilbao y Vizcaya*). Esta característica interfiere con otra práctica difundida en Internet: la escritura TOTALMENTE EN MAYÚSCULAS como recurso enfático. Decidimos, por lo tanto, deshabilitar en Freeling el módulo NER para multpalabras.

3.2.5. Valoración de carga subjetiva (*SentiWordNetSpanish*)

A la hora de planificar la actividad de anotación manual para generar recursos léxicos es importante contar con criterios que optimicen los esfuerzos. Nuestra inspiración original provino del importante recurso léxico *SentiWordNet*⁷.

Dado el procesamiento en cascada descrito en la figura 3, sabíamos de antemano que sólo las palabras que estuvieran en nuestro diccionario de estandarización y en el diccionario que usa Freeling para su configuración (archivo *dicc.src*) serían tomadas en cuenta para la asignación de sentimiento. Comenzamos por obtener los lemas de Freeling y ordenarlos en función de la frecuencia de uso de nuestro diccionario, basado en el CREA. De ese modo, nos asegurábamos comenzar por los lemas más frecuentes. En la lista resultante de 40.000 palabras desechamos las primeras 300 palabras por ser consideradas palabras funcionales con altísima frecuencia y sin valoración de sentimiento alguno.

En cuanto a los criterios de asignación de puntaje positivo y negativo (en una escala de 0 a 100), recurrimos a nuestros juicios como hablantes nativos y nuestro conocimiento de la teoría de los subjetivemas [5]. De ese modo, obtuvimos 4.000 lemas muy confiables con valoración (un 10% aproximadamente de todos los lemas), los cuales representan una cobertura de aproximadamente 40.000 formas no lematizadas del español, en virtud del extenso paradigma verbal y de la rica morfología para lemas nominales del español, todas ellas conllevando valoración de sentimiento.

Freeling dispone de un diccionario de aproximadamente 5.000 giros lingüísticos del español (archivo *locutions_extended.dat*), el cual puede ser ampliado manualmente. Ésta es una interesante característica que permite identificar frases por sobre la ocurrencia de las palabras que las componen (*multiword recognition*). Así pues, podemos refinar los criterios de anotación, tal como muestra la tabla 2:

Lema o frase	Valoración positiva	Valoración negativa
<i>bueno</i>	50	0
<i>buenos_días</i>	0	0
<i>bueno_para_nada</i>	0	40

Tabla 2: Anotación diferenciada para lemas o frases

⁷ <http://sentiwordnet.isti.cnr.it>

3.2.6. Operadores globales y operadores locales de cambios de polaridad

A esta altura estamos en condiciones de entender el procesamiento algorítmico de valoración de sentimiento: simplemente nos hacemos pasar cada token de cada unidad de análisis (cláusula) a través de nuestra SentoWordNetSpanish, reemplazando cada token por pares ordenados con las dimensiones (positivo,negativo) que correspondan. Si un token no es encontrado en la SentiWordNetSpanish, se le asigna el par (0,0). Los pares ordenados de una cláusula se van reduciendo (sumando) a izquierda. No obstante, existen dos tipos de ocurrencias de palabras que generan sus propias reglas de procesamiento: operadores globales para negación y modalidades y operadores locales de cambio de polaridad (por ejemplo, expresiones privativas como *sin* o *falta de*).

Los operadores globales son una lista de expresiones negativas (*no, ni, nadie, negar*, etc.) cuya ocurrencia provoca que la cláusula sea considerada negativa -asignando valor (0,60) a la cláusula-, independientemente de cualquier subtotal de arrastre de pares ordenados, constituyendo un fuerte sesgo hacia las opiniones negativas. Evaluamos distintas versiones de sesgo de los operadores globales, por ejemplo hacia la neutralidad -asignando valor (0,0) a la cláusula-; pero las pruebas sugirieron la conveniencia del sesgo hacia opiniones negativas ante la ocurrencia de dichos operadores globales (véase tabla 5). Los ejemplos (8) a (10) ilustran esta disyuntiva entre las opiniones neutrales y negativas para la ocurrencia de expresiones de negación:

- (8) *Mejoraste.* (valoración positiva)
- (9) *No mejoraste.* (valoración neutra o negativa)
- (10) *Nadie me ayudó.* (valoración negativa)

Los operadores locales producen cambios en la polaridad de los subtotales de arrastre de pares ordenados a derecha. Se trata de listas de palabras y expresiones como *sin, falta de, difícil de, deber*, etc. cuyo funcionamiento queda ilustrado en los ejemplos (11) y (12):

- (11) *Se manejó con respeto.* (valoración positiva)
- (12) *Se manejó sin respeto por las normas.* (valoración negativa)

En nuestro algoritmo, simplemente invertimos la polaridad del subtotal de pares ordenados a derecha de la ocurrencia de un operador local dentro de una cláusula.

4. EVALUACIÓN

Nuestra solución fue inicialmente evaluada en un corpus de 800 tweets de la industria de la telefonía celular, llegando a tener una efectividad de 76%, respecto del *gold standard* anotado a mano. Como grupo de control evaluamos el desempeño de los propios humanos en esta tarea. Por ejemplo, en los papers en donde se comparan clasificadores humanos entre sí, la tasa de *interconfiabilidad* (*interreliability ratio*) no supera el 70%, incluso para desacuerdos no fatales (cualquier discrepancia que no sea positivo por negativo y viceversa) [13].

Para nuestro set de datos, los clasificadores humanos cometieron errores más o menos significativos de inconsistencias en el 17% de los casos en promedio, por lo que la comparación final entre humanos y el algoritmo podría ser enunciada como un mecanismo muy costoso y muy lento (clasificación manual con humanos) con un éxito del 83% *versus* un mecanismo algorítmico sin costo alguno, instantáneo y escalable con un éxito del 76%.

1	TEXT (para ver analisis cuantitativo y confusion matrix ir al final de la tabla) STEP 2...SENTI DETECTED BY HINTS	ASSIGNED	ERROR POS	ERROR NEG	ERROR NEUTR
2					
3	ya se estan vendiendo las entradas para #RedHotChiliPeppers en Chile por TicketMaster y 20% Desc @Entel http://bit.ly/9qDQGm.D	POSITIVE			
4	RT @entel: ¡Ya tenemos el Sony Ericsson Xperia Arc! Ven a conocerlo :) http://on.fb.me/q5qKwm	POSITIVE			
5	@entel Tres meses con un iphone y se echa a perder?! Pfff! Cambio de equipo para @clapuinov !!	NEGATIVE			
6	@entel hola! Una pregunta: a mi hermana le robaron hoy su Iphone con plan. Hay como inhabilitarlo o seguirlo? Que pueo hacer? Gracias!	NEGATIVE			1
7	@entel ¿les cuesta mucho avisar cuando se esta acabando la bolsa d navegacion? DEPPREDADORES D MIERDA!	NEGATIVE			
8	Y si el comercial de @entel es cierto ?? Se les lesiona Messi y Nos encontramos con Argentina en 2a Ronda ??... Ahí los quiero ver :D	POSITIVE			1
9	Mierda esta que no me deja conectarme digan algo en entel @entel x favor ... Entel no está ni ahí	NEGATIVE			
10	@hlfx Pero yo ni siquiera soy cliente de @entel y me llaman para ofrecer productos! #FAIL	NEGATIVE			
11	Esto es velocidad mierda!!! 3 y + en cargar en mi super conexión de internet móvil de @entel @entel_ayuda	NEGATIVE	1		
12	RT @entel: ¡Ya tenemos el Sony Ericsson Xperia Arc! Ven a conocerlo :) http://on.fb.me/q5qKwm	POSITIVE			
13	@danielxhuevo @TWITCARTV @SoledadQnetto @entel @coseche Buena huevo, hermosa ella :)	POSITIVE			
14	RT @BancoGalicia: #TipQuiero! ¿Sabías que podés pagar el servicio de cable con puntos Quiero!? Con solo 225 puntos ahorras un 40%!!!	NEGATIVE	1		
15	@BancoGalicia Buenos días! Ayer me contestaron el primer mail con mis datos...y les contesté lo de la sucursal, pero no recibí respuesta :(NEGATIVE			
16	@BancoGalicia les mandé mi mail con los datos :) Cuanto demorarán en contestar?	POSITIVE			1
17	Excelente el servicio de atención de @entel_ayuda... rapido, claro y efectivo :)	POSITIVE			
18	Hasta que me llevo la Galaxy Ace, pero no gracias a @entel_ayuda, nada que decir en todo caso del call center. ellos la llevan :) cc @entel	POSITIVE			
19	@entel_ayuda ok, mensaje enviado :)	POSITIVE			
20	@entel_ayuda enviado :)	POSITIVE			
21	@entel_ayuda no.. mi plan se activa mañana! :(NEGATIVE			
22	gracias chicos de @entel_ayuda por la solución ultra flash! :)	POSITIVE			
23	@entel_ayuda muchas gracias!!!! :)	POSITIVE			
24	@MovistarArg Muchas gracias ~MS, ahora me quedo mas claro :)	POSITIVE			
25	Será un Motorola Spice ??? #motorola #spice #motorolaspice ?? #loquehayenlacaja ?? #OJALAAAH!! LO QUIEROOO! @PersonalAr	NEGATIVE	1		
26	@PersonalAr tengo un iphone4 y ya probé de todo... Tengo 3 amigos c iPhone y les pasa lo mismo... Son Uds. El problema, saludos! =)	POSITIVE		1	
27	@PersonalAr hice todo! ya tengo 4 reclamos hechos! ya nose que hacer... si me compro un equipo nuevo y sigue el problema que pasa?!	NEGATIVE			
28	@PersonalAr Hoy funciona. Qué significa red ocupada? Y por qué en Constitución, plena CABA, JAMÁS (de los jamases) hay señal?!!	NEGATIVE			
29	SUBTOTAL POR OPINION POSITIVA, NEGATIVA, NEUTRO		3	1	5
30					
31			TOTAL	9	
32	Distribucion: 38 POSITIVE 17 NEGATIVE DISTRIBUCION REAL 37 POSITIVE 13NEGATIVE 5 NEUTRAL				
33	Error Positive 3/37 Negative 1/13 Neutro 5/5				
34	Efectividad 92% 92,4% 0%				
35	Efectividad total 100-(9/55) = 83.7%				
36	tráfico de 6.9% de 800 tweets				

Tabla 3: Efectividad total de etapa de indicios de tempranos sobre 55 tweets en un corpus inicial de 800 tweets

1	TEXT SENTI DETECTED BY SENTIWORDNET + BASIC SYNTAX	SENTI ASSIGNED	ERROR POS	ERROR NEG	ERROR NEUTR	SENTI CORRECTED
2						
3	RT @cristianruizc: @zonaEntel @entel y que paso con el 2x1 con nevados de Chillan?	POSITIVE				POSITIVE
4	RT @zonaEntel: Mañana los clientes @entel tienen 2x1 en Valle Nevado. http://awe.sm/5OH6R,	POSITIVE				POSITIVE
5	@entel Muy buena la serie #EICr@ck !,	POSITIVE				POSITIVE
6	@entel como se puede hacer valido el descuento para Red Hot Chili Peppers ??,	NEUTRAL				NEUTRAL
7	Alerta!! molestia de Messi, @entel lo predijo ajajajajaj.	NEGATIVE			1	NEUTRAL
8	en la segunda se les lesiona Messi, NO EXISTEN... que cracks los de @entel ajajajajajaj.	NEUTRAL				NEUTRAL
9	Hasta ahora va a pasar exactamente lo que le dijo el pelado del spot de @entel al argentino, camino a Mendoza.,	NEUTRAL				NEUTRAL
10	Fome el comercial de @entel para el partido contra Uruguay. Se ve que ya se les acabaron las ideas al grupo creativo xd,	POSITIVE		1		NEGATIVE
11	Odio el comercial de @Entel ... Garra Charrúa es lo que van a ver el Viernes!!!,	NEGATIVE				NEGATIVE
12	Los del comercial de @Entel le están achuntando, casi se lesiona Messi xD,	NEUTRAL				NEUTRAL
13	@personalAR seguramente es un mejor servicio 3G para la gente con #Black #loquehayenlacaja,	POSITIVE		1		NEGATIVE
14	@PersonalAR adentro de la caja hay menos redes 3G colapsadas? #loquehayenlacaja,	NEUTRAL		1		NEGATIVE
15	@PersonalAR buen dial estoy llamando hace una hora para hacer un reclamo y no me atienden... como puedo solucionarlo?,	NEGATIVE				NEGATIVE
16	@PersonalAR Ok voy a estar esperando su respuesta,	NEGATIVE			1	NEUTRAL
17	@PersonalAR Por unos minutos mi celular dejo de tener el menu : PERSONAL !,	NEUTRAL				NEUTRAL
18	@PersonalAR me deben un Iphone,	NEUTRAL		1		NEGATIVE
19						
20						
21		muestreo hay#debr	6#48	43#58	50#48	
22	ERROR SUBTOTAL POR OPINION POSITIVA, NEGATIVA, NEUTRO		10	127	73	
23						
24	Distribucion total 192 POSITIVE 238 NEGATIVE 315 NEUTRAL	TOTAL		210		
25	Distribucion corregida a mano extrapolada de muestreo en total 240 POSITIVE 290 NEGATIVE 240 NEUTRAL	TOTAL EXTRAPOLADO = 31*5 = 155				
26	Error extrapolado Positi 10/240 Negative 95/290 Neutro 50/240	(muestreo de 154 sobre 745)				
27						
28	Efectividad total step 3 = 100-(31/154) = 80%					
29	tráfico de 93.1% de 800 tweets					

Tabla 4: Muestra sobre un corpus inicial de 745 tweets (exceptuando los 55 tweets detectados por indicios tempranos)

input total: 745
input total_positive: 120
input total_negative: 333
input total_neutral: 292
output_true_positives: 108
output_true_negative: 280
output_true_neutral: 169
output_false_positive: 62
output_false_negative: 95
output_false_neutral: 31
recall for POSITIVE: 90%
precision for POSITIVE: 63%
recall for NEGATIVE: 84%
precision for NEGATIVE: 74%
recall for NEUTRAL: 57%
precision for NEUTRAL: 84%
F-score for POSITIVE: 76.7647058824 %
F-score for NEGATIVE: 79.3753753754 %
F-score for NEUTRAL: 71.1883561644 %
F-score total (ponderando incidencia de los 3 sentimientos en total): 75.76473076 %

Tabla 5: Métricas totales para efectividad (confussion matrix) sobre 745 tweets en un corpus de 800 tweets (exceptuando los 55 tweets detectados por indicios tempranos, véase tabla 3)

5. CONCLUSIONES

El análisis automatizado de sentimiento es aún un campo fértil y vasto para ser explorado. Si bien el estado del arte para esta tarea ha mostrado recientes avances en los resultados, la misma diversidad del soporte lingüístico de la opinión humana torna complejo el problema. Uno de los aspectos más difíciles de modelizar es los que se conoce como análisis de sentimiento basado en entidades (*entity-based sentiment analysis*) [3]. Es decir, muchas veces las opiniones, incluso en textos breves, conllevan polaridades contrapuestas respecto de diversas entidades, como en el siguiente ejemplo:

(13) *Obama hizo un excelente trabajo con esa maldita pérdida de petróleo en New Orleans.*

Este tipo de opiniones representan un gran escollo para los algoritmos de análisis de sentimiento. La asignación de polaridad positiva y negativa a distintas entidades en la misma cláusula (*Obama y pérdida de petróleo*) hacen que sea imprescindible recurrir a una estrategia de *full parsing* complejo para entender el enunciado y a robustos algoritmos de extracción de entidades (*Named-Entity Recognition NER*). Estos enfoques más avanzados resultan un gran atractivo para investigaciones de marketing en empresas que pueden así refinar las opiniones positivas y negativas de clientes respecto de características en particular de productos.

Nuestra solución basada en reglas y recursos léxicos ha demostrado una razonable eficacia para el análisis de textos breves. No obstante, la evaluación del mismo algoritmo en textos más extensos y con dinámicas textuales complejas ha resultado en una drástica reducción de la efectividad. Esto puede deberse a la complejidad propia de los textos elaborados en base a los recursos retóricos y discursivos que se ponen en juego, como argumentaciones concesivas, polifonía [5], etc.

Aun así, una de las principales fortalezas de nuestro enfoque es que brinda la posibilidad de ampliar la cobertura en función de anotaciones manuales modulares en la valoración de SentiWordNetSpanish, en el agregado de palabras o frases nuevas a nuestro diccionario y a Freeling, etc. Incluso, existe la posibilidad de adaptar los recursos a diferentes dominios específicos o temáticos. Por ejemplo, en la industria de la telefonía móvil, palabras como *interferencia*, *cobertura* o frases como *caída_de_señal* seguramente serán valoradas de distinta manera que en sus

usos coloquiales generales.

En definitiva, estamos convencidos de que el análisis de sentimiento es un muy reciente campo de aplicación práctica de la investigación académica en lingüística computacional con una clara articulación con la industria. Y en esta convergencia de intereses puede hallarse una muy fecunda interacción para nuevos hallazgos y avances en NLP.

Referencias

- [1] Cruz, F., Troyano J., Enríquez, F. y Ortega J. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, (41):73-80, 2008.
- [2] Denecke, K. Using SentiWordNet for multilingual sentiment analysis. En *Proceedings of the IEEE 24th International Conference on Data Engineering Workshop*: 507-512, 2008.
- [3] Liu, B. Sentiment Analysis: a multi-faceted problem. *IEEE Intelligent Systems* 25(3):76-80, 2010.
- [4] Kamps, J. y Marx, M. Words with attitude. En *1st International WordNet Conference*: 332-341, 2002.
- [5] Kerbrat Orecchioni, C. *La enunciación*. Buenos Aires, Edicial, 1993.
- [6] Turney, P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*: 417-424, 2002.
- [7] Taboada, M., Brooke J., Tofiloski, M., Voll, K. y Stede, M. Lexicon-based methods for sentiment analysis. *Computational Linguistics* (1):1-41, 2010.
- [8] Giammatteo, M. y Albano, H. El español en Internet: una mirada a su evolución en los fotologs. *Revista Digital Universitaria UNAM* 10(3):1-17, 2009.
- [9] Goryachev S., Sordo M., Zeng Q.T. y Ngo L. Implementation and evaluation of four different methods of negation detection. DSG reporte técnico, 2008.
- [10] Jurafsky, D. y Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey, Prentice-Hall, 2000.
- [11] Manning, C., Raghavan P. y Schütze, H. *Introduction to Information Retrieval*. Cambridge (Reino Unido), Cambridge University Press, 2008.
- [12] Atserias, J., Casas, B. , Comelles, E., González, M., Padró L. y Padró M. Freeling 1.3: Syntactic and semantic services in an open-source NLP library. En *Proceedings of the 5th International Conference on Language (LREC'06)*:48-55, 2006.
- [13] Gryc, W. y Moilanen, K. Leveraging textual sentiment analysis with social network modeling: Sentiment Analysis of political blogs in the 2008 U.S. presidential election. En *Proceedings of the "From Text to Political Positions" workshop (T2PP 2010)*, Vrije Universiteit, Amsterdam, 2010.