# ANÁLISIS DE SENTIMIENTOS EN UN CORPUS DE REDES SOCIALES

GUADALUPE AGUADO-DE-CEA [1]
M. AUXILIADORA BARRIOS [2]
M. SOCORRO BERNARDOS [1]
INÉS CAMPANELLA [3]
ELENA MONTIEL-PONSODA [1]
ÓSCAR MUÑOZ-GARCÍA [3]
VÍCTOR RODRÍGUEZ [1]
*Universidad Politécnica de Madrid [1]*
*Universidad Complutense de Madrid [2]*
*Havas Media [3]*

RESUMEN

*El análisis de sentimientos de textos en las redes sociales se ha convertido en un área de investigación cada vez más relevante debido a la influencia que las opiniones expresadas tienen en potenciales usuarios. De acuerdo con una clasificación conceptual de sentimientos y basándonos en un corpus de diversos dominios comerciales, hemos trabajado en la confección de reglas que permitan la clasificación de dichos textos según el sentimiento expresado con respecto a una marca, empresa o producto. Con la ayuda de una base de datos de colocaciones (Badele3000) y un gestor de corpus (Calíope) se han creado 200 reglas en español que han puesto de manifiesto algunas consideraciones a tener en cuenta en la siguiente fase del trabajo.*

Palabras clave: análisis de sentimientos, análisis de emociones, *opinion mining*, redes sociales

ABSTRACT

*Sentiment analysis of user-generated content on the Web has become a relevant research area because of the influence that user opinions have on*

*potential users. Basing on a classification of sentiments and a corpus of different business domains, we have worked on the creation of linguistic rules to classify texts according to the sentiment classification mentioned with respect to a brand, company or product. For the creation of the 200 linguistic rules we have relied on a collocations database (Badele3000) and a corpus management tool (Calíope). The specific linguistic characteristics of the corpus have evidenced some problems in the creation of the rules that need to be approached in the next stages of the work*

Keywords: sentiment analysis, opinion mining linguistic rules, social media

# 1. INTRODUCCIÓN

En el mundo empresarial, cada día es más importante conocer el sentimiento que despiertan las marcas y los productos que las empresas lanzan al mercado. Según un informe reciente de Nielsen (2012), un 70% de usuarios de medios sociales prestan atención a la experiencia de consumo de otros usuarios, un 65% declara buscar información sobre marcas, productos y servicios, un 53% expresa comentarios positivos sobre marcas, y otro 50% expresa quejas y reclamaciones al menos una vez al mes. Dada la ingente cantidad de comentarios que se generan diariamente a través de los distintos canales, como son los blogs, microblogs (Twitter, o Tumblr), fórums, redes sociales como Facebook, LinkedIn, etc., se hace prioritaria la necesidad de un sistema que pueda extraer de forma automática el sentimiento global asociado a una marca, a una empresa, o a alguno de sus productos, por la relevancia que puede tener a la hora de planificar las estrategias de mercado.

Los estudios sobre sentimientos han estado presentes en diferentes campos y con fines distintos. Desde los primeros trabajos cognitivistas de Arnold (1960), muchas han sido las líneas de investigación seguidas para el análisis de las emociones y los sentimientos, casi tantas como los términos adoptados dentro de los diversos campos para referirse a la manifestación de las emociones, opiniones, gustos y valoraciones. En el campo de la psicología, los continuadores de los principios cognitivistas de Arnold, han tratado de plasmar las diferencias entre el proceso de valoración, y los principios de la valoración como contenido y como resultado (Lazarus 1991, Roseman y Smith 2001, entre otros). En lingüística, Martin y White (2005), dentro de la corriente funcionalista de la lengua, han planteado la teoría de la valoración o *appraisal theory* centrándose en determinados subsistemas como la actitud, el compromiso y la gradación, y han dirigido su investigación hacia la modalidad epistémica, la evidencialidad, y la intensificación, es decir, su objetivo era saber cómo se expresa lingüísticamente la valoración, el aplauso, la crítica o la opinión en general. Ortony et al. (1988) trataban de buscar las bases para encontrar un “computationally tractable model of emotion” que pudiera utilizarse en Inteligencia Artificial. Sin

embargo, la implementación de cualquier modelo solo sería posible gracias al desarrollo de las técnicas de procesamiento de lenguaje natural y a la posibilidad de integración en diferentes sistemas que fueron apareciendo posteriormente. Según Pang y Lee (2008), la aplicación de los análisis sobre las emociones, opiniones y valoraciones en relación con la empresa comienzan a tomar impulso a partir del año 2001, y adoptan diferentes nombres en inglés: *sentiment analysis, opinion mining, brand monitoring, buzz monitoring, online anthropology, market influence analytics, conversation mining, online consumer intelligence, user generated content.* Estas divergencias terminológicas muestran diferencias en las connotaciones que cada grupo quiere proyectar en su trabajo, así como en los usos que se dan en diferentes comunidades epistemológicas.

En este trabajo de carácter interdisciplinar hemos adoptado un término que satisface tanto la proyección lingüística como la computacional: "análisis de sentimientos", donde se conceptualiza el sentimiento como reacción humana detectable, esto es, rastreable e identificable y con una valencia concreta (Clore at al. 1987). Se excluyen así estados cognitivos que no tienen un signo (positivo o negativo) específico como la sorpresa, aburrimiento, etc.

El trabajo descrito en este artículo forma parte de un proyecto de más alcance, cuyo objetivo es extraer el sentimiento asociado a una entidad dada, es decir, a una marca o producto. Aquí se recoge lo realizado en una primera fase, que se centra en (1) identificar las expresiones que indican sentimientos en un corpus con la ayuda de dos recursos: Badele3000 (Bernardos y Barrios, 2008) y Calíope (Aguado y Bernardos, 2007), y (2) crear reglas que formalicen el conocimiento lingüístico y sirvan para clasificar los textos de acuerdo a una categorización de sentimientos que se describirá a continuación.

Tras esta introducción, en la sección 2 se presenta el corpus y la metodología, y se describen los pasos seguidos con los recursos mencionados. El apartado 3 recoge unas conclusiones iniciales y aspectos que están en estudio.

## 2. MÉTODO SEGUIDO Y RECURSOS EMPLEADOS

El **corpus** analizado está formado por textos en lengua española, procedentes de diferentes canales (blogs, microblogs, foros…) correspondientes a diversos dominios (automoción, deporte, banca, telecomunicaciones, alimentación, cosmética, seguros, etc.), cada uno con más de 2000 frases.

En una primera fase, se ha pre-procesado el corpus, es decir, se ha normalizado eliminando aquellos símbolos (como @, #, …) y abreviaciones muy coloquiales (como "ktal", "x", y otras del mismo estilo) que son habituales en estos tipos de textos y que causan "ruido". Así el corpus queda preparado para pasar el etiquetador POS utilizado, Freeling 3.0 (Padró y Stanilovsky, 2012). Tras esta fase de pre-procesamiento, se ha anotado manualmente el corpus con la finalidad de que sirva para la posterior evaluación. Para llevar a cabo esta anotación se parte de una **clasificación conceptual de los sentimientos** basada en Ekman (1982), Richins (1997) y Shaver et al. (1987). Dicha clasificación contempla las siguientes categorías de sentimientos, no disjuntas entre sí y cada una con sus correspondientes polaridades: satisfacción-insatisfacción, confianza-temor, amor-odio, felicidad-tristeza (véase la Tabla 1). La anotación consiste en asignar a cada mención (frase o texto) alguno de esos ocho sentimientos y, en caso de no corresponder a ninguno de los sentimientos, clasificarla como neutra (no clasificada).

| **Categoría** | **Polaridad** | |
|---|---|---|
| | + | - |
| SD | satisfacción | insatisfacción |
| TF | confianza | temor |
| HS | felicidad | tristeza |
| LH | amor | odio |

Tabla 1. Categorías para la clasificación de sentimientos organizadas según su polaridad

Este proceso de anotación ha sido paralelo al del análisis lingüístico, en el que se han identificado las expresiones que corresponden a esos sentimientos. Para ello, se ha partido de un conjunto de sustantivos que reflejan los sentimientos mencionados en la Tabla 1 y se ha ido ampliando progresivamente, al identificar otras unidades paradigmáticas que encajan con esos sentimientos, como los englobados en la Tabla 2 de sentimientos secundarios, basada en una reformulación de Richins (1997) y Shaver et al. (1987). Tras este recorrido inicial por los sustantivos, se ha seguido con los adjetivos, verbos y adverbios que contribuyen a indicar el tipo de sentimiento, la polaridad y algún tipo de gradación (+2 muy positivo, +1 positivo, -1 negativo, -2 muy negativo). Una vez extraídas y analizadas las expresiones relevantes se han creado las reglas que formalizan ese conocimiento lingüístico y permiten al sistema determinar el sentimiento global de un conjunto de textos dado.

| Sentimientos primarios | Sentimientos secundarios |
|---|---|
| **Confianza** | • optimismo, esperanza, seguridad |
| **Satisfacción** | • gratificación, contento, conformidad |
| **Felicidad** | • alegría, agrado, disfrute, placer, ilusión, entretenimiento,<br>• jovialidad, entusiasmo, júbilo<br>• orgullo, triunfo |
| **Amor** | • pasión, excitación, euforia, éxtasis, |
| **Temor** | • nerviosismo, alarma ansiedad, tensión, aprehensión, preocupación,<br>• shock, miedo, terror, pánico, histeria, mortificación,<br>• agonía, derrota |
| **Insatisfacción** | • disgusto, rechazo, inconformidad, repulsión, asco<br>• irritación, agravio, exasperación, frustración, molestia |
| **Tristeza** | • depresión, derrota, infelicidad, angustia, pena<br>• melancolía, nostalgia<br>• desilusión, decepción. desesperanza, derrota, abatimiento<br>• vergüenza, culpa, arrepentimiento, remordimiento<br>• alienación, aislamiento, soledad, inseguridad, humillación |
| **Odio** | • rabia, furia, ira, hostilidad, ferocidad<br>• amargura, resentimiento, rencor, desprecio, revanchismo<br>• envidia, celos |

Tabla 2. Sentimientos primarios y secundarios

Para el análisis de las expresiones se han utilizado dos recursos desarrollados dentro del grupo de trabajo, Badele3000 (Bernardos y Barrios, 2008) y Calíope (Aguado y Bernardos 2007), así como el etiquetador Freeling 3.0[1]. Para una mejor comprensión del proceso seguido, a continuación se explica brevemente cada uno de estos **tres recursos** y cómo se han creado las reglas.

*2.1 BADELE3000*

Badele3000 (Bernardos y Barrios, 2008, Barrios 2010) es una base de datos (BD) léxica que recoge los 3300 sustantivos más frecuentes del español, así como 2800 verbos de uso habitual que, conjuntamente, dan lugar a más de 20.000 colocaciones agrupadas por sentidos generales. Esta base de datos sigue los principios de la Teoría Sentido-Texto (TST) (Mel'čuk, 1996) referentes al concepto de Funciones Léxicas (FFLL). En la TST, una FL relaciona dos unidades léxicas: la 'palabra clave', o base de la colocación, y el colocativo, también denominado 'valor' de la función léxica.

La organización del contenido de Badele3000 ayuda a identificar y clasificar las expresiones del corpus en donde aparecen los sentimientos. Esto se debe a que sus unidades léxicas están clasificadas según una jerarquía de etiquetas semánticas (que suelen corresponder con el hiperónimo o genérico inmediato que aparece en su definición) y al hecho de que al emplear FFLL permite contar con información semántica que aportan conjuntamente el colocativo y la base. Así, gracias a esta BD se conocen los sustantivos que se corresponden con sentimientos y se obtienen los verbos que son colocativos de los sentimientos en general y de cada sentimiento en particular.

*2.2. CALÍOPE*

Calíope (Aguado y Bernardos 2007) es una aplicación web diseñada inicialmente con una doble finalidad: por un lado, proporcionar a estudiantes universitarios de Ingeniería informática la forma de aprender los términos propios de la especialidad dentro de su contexto y, por otro lado, mostrar las relaciones sintácticas y léxico

semánticas que se establecen entre ellos. Para ello, esta aplicación gestiona dos recursos: un corpus de textos correspondientes a distintos géneros profesionales y académicos del ámbito informático, tanto en inglés como en español, y un glosario de términos en ambas lenguas. Las operaciones de Calíope relevantes para este trabajo son las relacionadas con el manejo del corpus, principalmente la posibilidad de filtrar los textos con los que se quiere trabajar, analizar la frecuencia de las palabras, y buscar tanto las concordancias de un término como las coocurrencias de varios términos, que pueden aparecer juntos o no.

### *2.3. FREELING 3.0*

Freeling 3.0 (Padró y Stanislovsky, 2012) consta de un conjunto de herramientas de análisis de lenguaje natural, con código abierto, y disponible para varias lenguas. La facilidad de uso, robustez de la herramienta y su fiabilidad para el español son las principales razones que han contribuido a que la incorporemos para realizar el análisis POS. La anotación morfosintáctica empleada sigue las recomendaciones de EAGLES[2], como se detalla a continuación.

### *2.4 Las reglas*

Las reglas siguen el siguiente formato: *antecedente* → *consecuente*. El antecedente (a la izquierda de la flecha) representa una expresión en lenguaje natural y está constituido por las unidades léxicas que lo conforman y la parte de anotación POS de Freeling que sea relevante. El consecuente (a la derecha de la fecha) indica la categoría de sentimiento correspondiente a esa expresión, su polaridad y el grado de esa polaridad. Comenzando desde el inicio de un texto, el clasificador automático busca si algún antecedente de las reglas corresponde a la parte del texto que está examinando. Si lo encuentra, realizará la asignación que aparece en el consecuente y pasará a la parte del texto que sigue a la expresión identificada para volver a buscar una regla, y así sucesivamente hasta el final. La clasificación global será el resultado del cálculo del grado de polaridad de cada categoría de sentimientos de todos los consecuentes aplicados.

En i, ii, iii y iv se muestran ejemplos de algunas reglas creadas a partir de expresiones extraídas para “odio” tras analizar sus concordancias, obtenidas con Calíope (véase Tabla 3), y sus colocaciones, con Badele (véase Tabla 4):

i. mi/este **odio** a/por entidad:

```
[D] ODIO#NC [SP] _ENTITY_ -> LH - 1
```

ii. Siento **odio** a/por entidad:

```
SENTIR#V ODIO#NC [SP] _ENTITY_ -> LH - 1
```

iii. (cómo/cada día) **odio** (más) a (el/la/esta/…) entidad:

```
ODIAR#V A#SP /1/ _ENTITY_ -> LH – 1
ODIAR#V MÁS#RG A#SP /1/ _ENTITY_ -> LH – 2
CÓMO ODIAR#V A#SP /1/ _ENTITY_ -> LH – 2
```

iv. entidad es (muy/tan/…) **odiosa**

```
_ENTITY_ SER#V ODIOSO#A -> LH - 1
_ENTITY_ SER#V MUY#RG ODIOSO#A -> LH - 2
```

**TEXTO:** Havas-Bebidas. CÓDIGO: 1008. **OCURRENCIAS ENCONTRADAS:** 8

| | | |
|---|---|---|
| España sopetero 6 Mahou 0 | Odio | la mahou desde pixiquito viva |
| en buena compañía RT mayteea2012 | Odio | los bares que intentan engañarme |
| pero esas me gustan mucho | Odio | la sol y la estrella |
| t co RUkgI5pT RT JaimeALastra | Odio | los anuncios de Estrella damm |
| Heineken bien frias no te | odio | pero ojalá panza y él |
| y picando alguito buenas noches | Odio | a Quilmes Stella Artois antes |
| Estrella Damm están creando mucho | odio | en este país Empiezan las |
| ocurrían más ejemplos todohalloween net | Odio | los sueños que parecen reales |

Tabla 3. Algunas concordancias de “odio” encontradas con Calíope

| Lema | FL | Valor |
|---|---|---|
| odio | FinFunc0 | desaparecer |
| odio | Func1 | emanar |
| odio | Func1 | anidar (en algo/alguien) |
| odio | Func1 | palpitar (en alguien) |
| odio | Func1 | latir (en alguien) |
| odio | Func1 | embargar (a alguien) |
| odio | IncepFunc0 | nacer |
| odio | IncepPredMinus | disminuir |
| odio | IncepPredPlus | aumentar |
| odio | Manif | mostrar |
| odio | Oper1 | sentir |
| odio | Oper1 | tener |
| odio | Real1-M | ocultar |
| odio | Real1-M | disimular |

Tabla 4. Colocaciones de “odio” en BADELE

## 3. DISCUSIÓN Y CONCLUSIONES

En el momento de escribir este artículo se cuenta con aproximadamente 1200 reglas, cuya efectividad sólo se ha podido probar parcialmente puesto que todavía no se tiene la versión definitiva de la parte del sistema que las maneja. No obstante, del trabajo realizado se han extraído algunas consideraciones reseñables que apuntan a nuevas líneas de investigación. Una de las principales es que las reglas son demasiado específicas, por lo que habrá pocas expresiones que casen con su antecedente. Esto significa que la mayor parte de los textos quedaría sin clasificar, aunque, dado el detalle de las reglas, la probabilidad de que aquellos que estuvieran clasificados lo estuvieran correctamente es muy alta.

Para conseguir una cobertura mayor se está trabajando en dos líneas: aumentar el número de reglas y elaborar reglas más generales.

Se puede obtener un mayor número de reglas generando reglas equivalentes a las ya existentes, pero considerando lemas distintos, que no se encuentren en el corpus de análisis. No es necesario que sean sinónimos, pero sí que compartan la misma estructura sintáctica del antecedente y que su clasificación sea la misma. Un candidato idóneo para sustituir un verbo es otro que pueda actuar de colocativo con el mismo sentimiento con respecto a la misma función léxica. Por ejemplo, dado que “sentir” y “tener” son valores de la FL Func1 para “odio”, se podría añadir la regla de v (equivalente a la de i).

v. Posibles reglas para añadir a partir de i

```
TENER#V ODIO#NC [SP] _ENTITY_ -> LH - 1
```

En lo que respecta a la elaboración de reglas más generales, hay que estudiar los efectos de prescindir de alguno de los elementos del antecedente o de hacerlo menos restrictivo. Por ejemplo, la segunda regla de iv no se podría generalizar con una regla que permitiera cualquier adverbio, en lugar de especificar “muy”, puesto que si el adverbio fuera “poco” el grado sería distinto. Por otro lado, teniendo en cuenta que el sistema descarta aquellas oraciones en las que no aparece la entidad, se están analizando los efectos de eliminarla en las reglas en las que aparezca. Por ejemplo, en la regla v, se considerarían únicamente el adverbio y el adjetivo, con el fin de que el clasificador encontrara más encajes.

vi. Posibles generalizaciones de la segunda regla de iv

```
_ENTITY_ SER#V [RG] ODIOSO#A -> LH – 2 (*)
MUY#RG ODIOSO#A -> LH – 2 (?)
```

No se debe olvidar la posibilidad de que en el mundo real no se den las expresiones correspondientes a los antecedentes, puesto que de ser así, la regla correspondiente no contribuiría a aumentar la cobertura. Además, es necesario analizar si las nuevas reglas cambian la precisión del sistema y, en el caso de que disminuya, estudiar si la ganancia de cobertura compensa la pérdida de precisión.

Otra observación interesante es que se ha comprobado que el número de frases con valor negativo es mayor que el de frases con

valor positivo, ya que es más frecuente escribir para criticar un producto, una marca, o a una empresa que para comentar una experiencia positiva. Por otra parte, más de un 80% del corpus puede considerarse neutro, dado que en muchas ocasiones los textos únicamente constatan un hecho, véase a) y b).

a) Ojo, debido a la actualización de precios tras el cambio de política de subvención de terminales de Vodafone, los mismos aún están sometidos a cierta variación, de hecho, desde la última actualización se confirma una bajada de hasta 90€ en modelos y tarifas seleccionadas
b) Las acciones de Vodafone subieron cerca de un 2% tras la presentación de los resultados

Finalmente, es importante señalar la variedad de problemas encontrados, que incluyen aspectos como la tipología textual (lenguaje de los medios sociales), la incorrección gramatical y estilística (por tratarse de textos extremadamente coloquiales) y la brevedad y concisión de ciertos mensajes (principalmente los procedentes de microblogs) y que dificultan enormemente la elaboración de las reglas.

## NOTAS

[1] http://nlp.lsi.upc.edu/freeling/
[2] http://www.uni-leipzig.de/~burr/Verb/htm/LinkedDocuments/annotate.pdf

## AGRADECIMIENTOS

Ese trabajo ha sido financiado por el Programa CENIT, "Consorcios Estratégicos Nacionales en Investigación Técnica", y el Ministerio de Ciencia e Innovación, en colaboración con el Centro para el Desarrollo Tecnológico Industrial (CDTI) y Social Media -- CEN-20101037

## REFERENCIAS BIBLIOGRÁFICAS

Aguado de Cea, G. y Bernardos, Mª S. 2007. “Calíope: herramienta para gestionar un corpus y un glosario de términos

informáticos". *Proceedings of the 6th Annual Conference of the European Association of Languages for Specific Purposes* (AELFE 2007). Lisboa (Portugal).

Arnold, M.B. 1960. *Emotion and personality*. New York: Columbia University Press

Bernardos, Mª S. y Barrios, M.A. 2008. "Data model for a lexical resource based on lexical functions". *Research in Computing Science*, vol. 27.

Barrios, M. A. 2010. "El dominio de la funciones léxicas en el marco de la Teoría Sentido-Texto". *Estudios de Lingüística del Español* (ELiEs), 30.

Clore, G.L., Ortony, A; y Foss, M.A. 1987. The Psychological Foundations of the Affective Lexicon. *Journal of Personality and Social Psychology*, 53, 751–755.

Ekman, P. 1982. *Emotion in the Human Face*. Cambridge University Press

Lazarus, R.S. 1991. *Emotion and adaptation*. Nueva York: Oxford University Press.

Martin, J.R. y White, P.R.R.. 2005. *The Language of Evaluation, Appraisal in English*, Londres y Nueva York: Palgrave Macmillan.

Mel'čuk I. 1996. "Lexical functions: A tool for the description of lexical relations in a lexicon". *Lexical functions in lexicography and natural language processing*. L. Wanner, (ed.), Amsterdam/ Philadelphia: John Benjamin, 37-102.

Nielsen. 2012. The social media report. [Documento de Internet disponible en http://blog.nielsen.com/nielsenwire/social/2012/]

Ortony, A., Clore, G.L. y Collins, A. 1988. *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Padró, L. y Stanilovsky, E. 2012. "FreeLing 3.0: Towards Wider Multilinguality". *Proceedings of the Language Resources and Evaluation Conference* (LREC 2012) ELRA. Estambul, (Turquía).

Pang, B y Lee, L 2008. "Opinion mining and sentiment analysis". *Foundations and Trends in Information Retrieval*, 2, 1-2, 1-135.

Richins, M. 1997. “Measuring Emotions in the Consumption Experience”. *Journal of Consumer Research* 24, 127-146.

Roseman, I.J. y Smith, C.A. 2001. Appraisal theory: Overview, assumptions, varieties, controversies. K.R. Scherer, A. Schorr y T. Johnstone (eds.), *Appraisal processes in emotion: Theory, methods, research*. New York: Oxford University Press.

Shaver, P., Schwartz, J., Kirson D., y O'Connor C. 1987. “Emotion Knowledge: Further Exploration of a Prototype Approach”. *Journal of Personality and Social Psychology* 52(6), 1061-1086.

Zabin, J. y Jefferies, A. 2008. “Social media monitoring and analysis: Generating consumer insights from online conversation” *Aberdeen Group Benchmark Report*.