

Análisis de emociones en español para el dominio médico

Emotion analysis in Spanish for the medical domain

Flor M. Plaza del Arco
Sinai Group
Universidad de Jaén
Campus Las Lagunillas s/n. E-23071
fmplaza@ujaen.es

Resumen: El Análisis de Sentimientos (AS) es una tarea del Procesamiento del Lenguaje Natural (PLN) que ha captado la atención de varias áreas no solo relacionadas con la inteligencia artificial sino también con la política, economía o psicología. El reconocimiento y análisis de emociones en documentos textuales se considera una tarea avanzada dentro del AS que consiste en determinar la categoría emocional (tristeza, alegría, sorpresa, entre otras). Hasta ahora, la mayor parte de la investigación se ha desarrollado en inglés, sin embargo, la adaptación de herramientas de AS a otros idiomas diferentes al inglés se hace cada vez más necesaria, ya que la información en la web social crece de manera exponencial. En este trabajo nos centramos en el análisis de emociones en español. Por otra parte, las aplicaciones del AS hasta hoy están centradas en áreas como el análisis de mercado, la política, la detección de Spam, los sistemas de recomendación, etc. Sin embargo, otras áreas como el dominio médico no han sido exploradas todavía o no suficientemente. Por ello, proponemos la aplicación de la tecnología desarrollada en el ámbito biomédico.

Palabras clave: Reconocimiento de emociones, Análisis de sentimientos, Dominio médico, Procesamiento de Lenguaje Natural

Abstract: Sentiment Analysis is a well-known task of Natural Language Processing that is attracting the attention of several areas not only related to artificial intelligence but also to politics, economics or psychology. The recognition and analysis of emotions in textual documents is considered an advanced task within the AS that consists in determining the emotional category (sadness, joy, surprise, among others). To date, most of the research has been developed in English. However, increasingly, adapting AS tools to languages other than English is becoming more and more important. In this work we focus on the analysis of emotions in Spanish. On the other hand, the applications of the AS until now are centered on areas like market analysis, politics, Spam detection, recommendation systems, etc. Nevertheless, other areas such as the medical domain have not sufficiently been explored yet. Therefore, we propose the application of technology developed in the biomedical domain.

Keywords: Emotion recognition, Sentiment Analysis, Medical domain, Natural Language Processing

1 Justificación de la investigación propuesta

El Análisis de Sentimientos (AS) es una tarea del Procesamiento de Lenguaje Natural (PLN) que trata la subjetividad en documentos (Liu, 2012). Aunque se trata de una tarea relativamente reciente, ya cuenta con una gran cantidad de investigación en el área no solo del PLN sino también de otras disciplinas de la

Inteligencia Artificial así como de la Economía, Política, Sociología, etc. Sin embargo, la mayoría de la investigación relacionada se ha realizado sobre documentos en inglés aunque cada vez con mayor interés se empiezan a estudiar otros idiomas como por ejemplo el español (Martín-Valdivia et al., 2013).

Por otra parte, una de las áreas más complejas y menos estudiadas en el AS es el análisis de la emoción (Cambria, 2016). En

el análisis de emociones se han de identificar diferentes categorías emocionales en el texto. Al igual que ocurre en el resto de tareas del AS, la mayor parte de la investigación se ha realizado en inglés, encontrando muy pocos trabajos para otros idiomas y concretamente para español.

En lingüística computacional, la detección automática de las emociones en los textos es cada vez más importante desde el punto de vista práctico. Consideramos, por ejemplo, las aplicaciones basadas en minería de opiniones, análisis de mercado, computación afectiva o interfaces de lenguaje natural como entornos de e-learning. Los posibles efectos beneficiosos de las emociones sobre la memoria y la atención de los usuarios son bien conocidos en el campo de la psicología (Strapparava and Mihalcea, 2007).

Mientras que la identificación de la polaridad de la opinión (positiva, negativa) es un campo activo en el AS, creemos que la identificación de la emoción en los textos aumentaría la eficacia de las aplicaciones.

Recientemente, con el vertiginoso desarrollo de la Web 2.0, gran cantidad de opiniones son asignadas por los usuarios con categorías emocionales, tales como la felicidad, la tristeza y la sorpresa. Dichas emociones se pueden tratar para realizar categorización de documentos, y por tanto ayudar a los usuarios conectados a seleccionar documentos basándose en sus preferencias emocionales.

Este trabajo se centra en el análisis de la emoción en español en el dominio médico. Este dominio es una de las áreas más importantes para el ser humano debido a que la salud es uno de los temas que más preocupa al mismo. Comprender la experiencia de los pacientes es un pilar fundamental para la calidad de la asistencia sanitaria.

2 Origen y trabajo relacionado

Desde principios del año 2000, el Análisis de la Opinión ha crecido hasta convertirse en una de las principales áreas de investigación del PLN. Esto es debido a la aparición de gran cantidad de información subjetiva disponible en forma digital en la Web 2.0.

Las tareas iniciales en AS se centraron en la clasificación de polaridad de las opiniones determinando simplemente si dichas opiniones eran positivas o negativas (Pang, Lee, and others, 2008). Sin embargo, actualmente

el AS engloba el tratamiento computacional de la opinión, emoción y la subjetividad en cualquier tipo de documento (Cambria, Livingstone, and Hussain, 2012). Podemos encontrar una visión general del campo de la computación afectiva (Affective Computing) en (Picard and Picard, 1997), que pone el foco más en el tratamiento de las emociones.

Precisamente, una de las tareas más complejas del AS y que aún no ha sido estudiada en profundidad es la identificación y análisis de la emoción en los textos. Liu, Lieberman, and Selker (2003) realizan un estudio sobre la inferencia textual de la afinidad emocional a nivel de oración. Los autores adoptan la noción de emociones básicas y utilizan seis categorías emocionales básicas (Ekman et al., 1987): ira, disgusto, temor, felicidad, tristeza y sorpresa. El trabajo no resulta exitoso ya que trabajan a nivel de oración.

Uno de los pilares fundamentales en la investigación relacionada con AS se centra en los recursos lingüísticos disponibles. Los recursos léxicos son indispensables y existen varios disponibles para el idioma inglés, como SentiWordNet (Esuli and Sebastiani, 2006) o General Inquirer (Stone, Dunphy, and Smith, 1966). Si nos centramos en el reconocimiento de emociones encontramos menos recursos, si bien podemos destacar por ejemplo ANEW (Affective Norms for English Words) (Redondo et al., 2007) y WordnetAffect (Strapparava, Valitutti, and others, 2004). Concretamente, para el desarrollo de WordnetAffect, se consideró como punto de partida WordNet Domains (Magnini and Cavaglia, 2000) una extensión multilingüe de WordNet, desarrollada en ITC-irst. WordnetAffect incluye un subconjunto de synsets adecuados para representar conceptos afectivos.

Con respecto a la disponibilidad de recursos de AS en un idioma distinto al inglés, nos encontramos con que el número es bastante más limitado. Concretamente para español podemos citar el recurso CRiSOL (Molina-González, Martínez-Cámara, and Martín Valdivia, 2015) que es el resultado de la combinación de dos recursos lingüísticos destinados al Análisis de Opiniones. Uno de esos recursos es la lista de palabras de opinión en español iSOL (Molina-González et al., 2013), y el otro es el lexicon de opinión en inglés SentiWordNet. (Ortiz, Pozo, and

Sánchez, 2010)

En lo referente al análisis de emociones, encontramos el corpus EmotiBlog (Boldrini et al., 2009) anotado con emociones que incluye los idiomas español, inglés e italiano.

El único recurso marcado con emociones encontrado en español es el desarrollado por (Díaz Rangel, Sidorov, and Suárez Guerra, 2014), que presentan un método para la creación de diccionarios marcados con un valor específico para su uso en varias tareas de PLN realizadas por computadoras. Dicho método se diferencia en varios aspectos de la adaptación al español del ANEW; primero por el hecho de que cada palabra proporciona información acerca de su categoría emocional, tomando como referencia las seis emociones básicas de Ekman; segundo, las palabras incluidas indican la frecuencia con que se usan con un sentido emocional, considerando diferentes contextos de aplicación; tercero, se aplica una adaptación del método de concordancia ponderada en las evaluaciones realizadas, mientras que ANEW no lo hace.

El AS se ha aplicado a varios dominios tales como el cine, la política, el análisis de mercado, entre otros. Sin embargo, la investigación en el dominio médico es muy escasa. Quizás uno de los primeros enfoques es el propuesto por (Niu et al., 2005). Los autores aplican aprendizaje automático (SVM) para clasificar la polaridad de las oraciones extraídas de la base de datos de la bibliografía médica MEDLINE. (Greaves et al., 2013) aplicaron técnicas de aprendizaje automático para clasificar las opiniones de los pacientes relacionadas con su experiencia en un hospital en el Servicio Nacional de Salud Inglés. El objetivo principal era predecir automáticamente la información textual en el comentario si el paciente recomendaba un hospital, si el hospital estaba limpio y si él/ella fue tratado/a con dignidad. El objetivo principal era estudiar la aplicabilidad de los métodos de análisis de opiniones típicas en los textos clínicos. Un enfoque similar al trabajo de (Greaves et al., 2013) es el expuesto en (Plaza del Arco et al., 2016) donde se recopila el corpus COPOS con opiniones de pacientes sobre médicos y clínicas en español extraídas de la página web *masquemedicos*. Aplican dos enfoques (supervisado y no supervisado) para clasificar la polaridad (positiva, negativa) de las opiniones escritas por los pacientes.

Con respecto a la aplicación de técnicas

de emociones sobre el dominio médico, hasta donde sabemos, no conocemos ningún trabajo.

3 Descripción de la investigación propuesta

En este trabajo proponemos un sistema automático de reconocimiento de emociones en español.

En primer lugar, haremos una investigación a fondo de los trabajos sobre reconocimiento e identificación de emociones que se han desarrollado para inglés. A continuación, se verán los posibles estudios que haya en otros idiomas, centrándonos en los que se enfoquen en español.

Se intentarán implementar los sistemas ya existentes en inglés adaptándolos al español mediante la generación de diferentes recursos, tanto corpus como lexicones.

Elegimos realizar el trabajo en español ya que son escasos los recursos disponibles actualmente en nuestra lengua y cada vez el español tiene más presencia en la red.

Por último, los sistemas se desarrollarán en un principio de manera general, pero en una segunda fase se adaptarán a un dominio específico como es el dominio médico, ya que consideramos que es posible extraer de redes sociales diversos corpus interesantes que expresen emociones. Podremos utilizar estos recursos para entrenar nuestros sistemas así como para validar el modelo propuesto.

Otro de los motivos para centrarnos en el ámbito biomédico es que el crecimiento de los documentos médicos en Internet en la última década se ha incrementado exponencialmente. En España, un estudio realizado por Doctoralia en 2015 muestra que el 62% de la población española consulta internet para buscar información acerca de algún tema relacionado con la salud. Por tanto, se requiere el desarrollo de sistemas más eficientes para tratar este tipo de información.

Los objetivos concretos que se pretenden alcanzar con este proyecto son los siguientes:

- Extraer información subjetiva de las diferentes plataformas (blogs, redes sociales, foros, etc) que dispongan de emociones.
- Generar distintos recursos para reconocimiento de emociones en español, tanto corpus como lexicones.
- Procesar dicha información para desarrollar sistemas que sean capaces

de identificar las diferentes categorías emocionales.

- Adaptar los sistemas desarrollados al dominio médico.

4 Metodología y experimentos propuestos

La metodología que se propone para la consecución de esta tesis se presenta a continuación:

1. Estudio y revisión del estado del arte. Se comenzará con el estudio y análisis de la bibliografía existente sobre el AS, en concreto, sobre la tarea de análisis de emociones.
2. Adaptación de recursos existentes para poder realizar un análisis de los métodos propuestos. Se intentará adaptar ciertos recursos tales como WornetAffect y SentiWordnet al español.
3. Desarrollo de un prototipo. Se tratará de implementar un sistema de detección de emociones en español.
 - Diseño de una arquitectura modular que permita integrar nuevas funcionalidades a medida que se vaya avanzando en la investigación.
 - Construcción de la arquitectura modular diseñada.
 - Prueba del correcto funcionamiento del prototipo.
4. Experimentación y evaluación. Se utilizarán los recursos generados para llevar a cabo la experimentación y posteriormente se procederá a la evaluación del prototipo, llevando a cabo una comparación de los resultados obtenidos con los ya existentes. Los resultados obtenidos se pondrán a disposición de la comunidad científica.

5 Conclusión

El AS ha sido aplicado a diferentes dominios tales como el cine, la política, el análisis de mercado, etc. La mayoría de los recursos generados están en inglés. Sin embargo, como hemos visto, apenas se han creado recursos en español, al igual que no se han explotado tareas de AS en el dominio médico. Con el crecimiento de la Web 2.0, disponemos de una cantidad ingente de información subjetiva a

la que es necesario aplicar técnicas de PLN. Por tanto y con la finalidad de mejorar el procesamiento del sistema inteligente y el razonamiento, es necesario generar recursos con modelos emocionales.

El propósito de este trabajo es generar recursos en español en el dominio médico para el análisis de emociones, extrayendo información subjetiva de distintas plataformas Web y procesando dicha información para desarrollar sistemas capaces de identificar las diferentes categorías emocionales.

Agradecimientos

Este trabajo está parcialmente subvencionado por el proyecto REDES (TIN2015-65136-C2-1-R) del MICINN del Gobierno de España.

Bibliografía

- Boldrini, E., A. Balahur, P. Martínez-Barco, and A. Montoyo. 2009. Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. In *DMIN*, pages 491–497.
- Cambria, E. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.
- Cambria, E., A. Livingstone, and A. Hussain. 2012. The hourglass of emotions. *Cognitive behavioural systems*, pages 144–157.
- Díaz Rangel, I., G. Sidorov, and S. Suárez Guerra. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein*, 1(29).
- Ekman, P., W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.
- Esuli, A. and F. Sebastiani. 2006. Sentiwordnet: A high-coverage lexical resource for opinion mining. *Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR)*.

- Greaves, F., D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson. 2013. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of medical Internet research*, 15(11).
- Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
- Liu, H., H. Lieberman, and T. Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132. ACM.
- Magnini, B. and G. Cavaglia. 2000. Integrating subject field codes into wordnet. In *LREC*, pages 1413–1418.
- Martín-Valdivia, M.-T., E. Martínez-Cámara, J.-M. Perea-Ortega, and L. A. Ureña-López. 2013. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems With Applications*, 10(40):3934–3942.
- Molina-González, M. D., E. Martínez-Cámara, and M. T. Martín Valdivia. 2015. Crisol: Base de conocimiento de opiniones para el español. *Procesamiento del Lenguaje Natural*, 55:143–150.
- Molina-González, M. D., E. Martínez-Cámara, M.-T. Martín-Valdivia, and J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.
- Niu, Y., X. Zhu, J. Li, and G. Hirst. 2005. Analysis of polarity information in medical text. In *AMIA annual symposium proceedings*, volume 2005, page 570. American Medical Informatics Association.
- Ortiz, A. M., Á. P. Pozo, and S. T. Sánchez. 2010. Sentitext: sistema de análisis de sentimiento para el español. *Procesamiento del Lenguaje Natural*, 45:297–298.
- Pang, B., L. Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.
- Picard, R. W. and R. Picard. 1997. *Affective computing*, volume 252. MIT press Cambridge.
- Plaza del Arco, F. M., M. T. Martín Valdivia, S. M. Jiménez Zafra, M. D. Molina González, and E. Martínez Cámara. 2016. Copos: Corpus of patient opinions in spanish. application of sentiment analysis techniques. *Procesamiento del Lenguaje Natural*, 57:83–90.
- Redondo, J., I. Fraga, I. Padrón, and M. Comesaña. 2007. The spanish adaptation of anew (affective norms for english words). *Behavior research methods*, 39(3):600–605.
- Stone, P. J., D. C. Dunphy, and M. S. Smith. 1966. The general inquirer: A computer approach to content analysis.
- Strapparava, C. and R. Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.
- Strapparava, C., A. Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086.