Usando la minería de emociones para la detección de problemas reales

Using emotion mining to detect real-life problems

Flor Miriam Plaza-del-Arco
Departamento de Informática, Escuela Politécnica Superior de Jaén
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
fmplaza@ujaen.es

Resumen: Las emociones juegan un papel importante en la inteligencia y el comportamiento humano y son un vehículo esencial para la comunicación. La minería de emociones es una tarea reciente que tiene como objetivo la identificación de diferentes categorías emocionales en un texto. Debido a su complejidad y a la escasa disponibilidad de recursos léxicos anotados, se encuentra en una primera etapa de investigación. Además, la mayoría de los estudios y recursos existentes se han realizado para el inglés, pero la presencia en Internet de otras lenguas, como el español, es cada vez mayor. Por esta razón, en este trabajo, se describe un proyecto de tesis cuyo objetivo es el desarrollo de sistemas orientados al reconocimiento de emociones en textos en español. Además, se pretende utilizar dichos sistemas para resolver otras tareas de gran relevancia hoy en día, como por ejemplo, la incitación al odio en las redes sociales o la detección de trastornos mentales.
Palabras clave: Minería de emociones, natural language processing, recursos léxicos afectivos, incitación al odio

Abstract: Emotions play an important role in human intelligence and behaviour and are a major vehicle for communication. Emotion mining is a relatively recent task that attempts to identify different emotional categories in text. However, due to its complexity and the limited availability of annotated lexical resources, it is still in the early stages of research. In addition, most of the work and resources have been focus on English texts, but the presence of other languages, such as Spanish, is growing on the Web. Therefore, in this work, we describe a thesis project that will focus on the development of emotion recognition systems in Spanish texts. In addition, we aim to use these systems to solve other relevant tasks, such as, hate speech identification on social media or mental disorders detection.
Keywords: Emotion mining, natural language processing, affective lexicons, hate speech

1 Justificación de la investigación propuesta

Las emociones juegan un papel adaptativo, social y motivacional en nuestro día a día ya que representan diferentes características indicativas del comportamiento humano, como el estado emocional, el nivel de interés o el estado de alerta.

El objetivo de la computación afectiva es permitir que los ordenadores puedan reconocer, interpretar y procesar emociones humanas. Por lo tanto, esta rama es un elemento clave para el progreso de la Inteligencia Artificial. La minería de emociones se enmarca dentro del análisis de sentimientos y de la computación afectiva y trata de identificar diferentes categorías emocionales en un texto, tales como la tristeza, la alegría, el enfado o el miedo. En los últimos años, ha surgido un creciente interés en la detección automática de las emociones en un texto dando lugar a trabajos muy prometedores en el área (Strapparava, 2016).

Por otra parte, actualmente cada vez son más los usuarios que utilizan las redes sociales, blogs o foros para comunicarse, por lo que el texto es una fuente de datos particularmente importante con contenido emocional en la Web. El tratamiento de estos datos requiere la identificación y el análisis automatizado de las emociones expresadas por los usuarios en el texto (Hasan, Rundensteiner, y Agu,

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Sepúlveda-Torres, R. (eds.) *Proceedings of the Doctoral Symposium of the XXXV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*, p. 7–12 Bilbao, Spain, September 25th 2019.

2019).

La minería de emociones tiene el potencial de humanizar las interacciones digitales y ofrecer beneficios en una gama casi ilimitada de aplicaciones. Por ejemplo, en el campo de la psicología, puede ayudar a los profesionales a comprender rápidamente el estado de ánimo de un paciente o buscar signos de que el usuario sufre alguna enfermedad mental como la depresión o la anorexia. En las redes sociales, se podría identificar usuarios que están sufriendo ciberbullying o incluso usuarios que piensan en suicidarse (Hinduja y Patchin, 2010). Al igual, es posible identificar en estas plataformas mensajes que incitan al odio provocando efectos psicológicos negativos a otros usuarios.

La incitación al odio se define comúnmente como el lenguaje hostil, malicioso y motivado por prejuicios dirigido a una persona en específico o a un grupo de personas en base a alguna característica, como puede ser la raza, la sexualidad, el color, la etnia, la apariencia física, la religión o la discapacidad (Cohen-Almagor, 2011; Erjavec y Kovačič, 2012). Inicialmente, este tipo de contenido se difundía a través de medios tradicionales, como la televisión, la radio o los periódicos. Actualmente, con el continuo crecimiento de las redes sociales, desafortunadamente encontramos una gran variedad de contenido malintencionado en la Web. Este hecho preocupa a la sociedad, a los gobiernos y a las plataformas de redes sociales. Según un informe sobre la evolución de los incidentes relacionados con los delitos motivados por el odio en España realizado por el Ministerio del Interior en 2017[1], Internet y las redes sociales aparecen como los medios más utilizados para difundir la incitación al odio, con un 36,5 % y un 17,9 %, respectivamente.

Las graves consecuencias de este problema, combinadas con la gran cantidad de datos que los usuarios publican diariamente en la Web, requieren el desarrollo de algoritmos capaces de detectar automáticamente comentarios inapropiados.

Recientemente, un gran número de investigadores han comenzado a trabajar en la tarea de detección automática del odio en redes sociales (Fortuna y Nunes, 2018; Fersini, Rosso, y Anzovino, 2018). Se considera una tarea muy compleja para estas plataformas tanto que, para resolver este problema, a menudo dependen de su comunidad para reportar el contenido malicioso.

Este proyecto de tesis se centra en el análisis de las emociones y en la aplicación de dicho análisis a problemas reales, como puede ser la detección del odio o de trastornos mentales en las redes sociales. Además, a diferencia de la mayoría de los trabajos existentes hasta el momento, la tesis principalmente se centrará en el tratamiento de textos en español, ya que su presencia en Internet es cada vez mayor, por lo que surge la necesidad de desarrollar sistemas aplicados a dicho idioma.

El resto del artículo está organizado de la siguiente forma: en primer lugar, en la Sección 2 se mencionará el origen y trabajo relacionado con el proyecto de tesis. En la Sección 3 se describe la investigación propuesta. La Sección 4 expone la metodología y los experimentos que se van a desarrollar y por último, se presentan los elementos de investifación propuestos para su discusión en la Sección 5.

2 Origen y trabajo relacionado

Uno de los primeros estudios relacionados con la computación afectiva es el de Picard (Picard, 1997). Ella propuso la idea de entrenar a los sistemas para identificar las emociones humanas. La construcción de sistemas afectivos requiere un procesamiento multimodal, ya que un ser humano puede expresar emociones a partir de una amplia gama de señales de comportamiento. Los investigadores realizan el análisis a través de diferentes fuentes de información, como los gestos, el habla, los movimientos, la expresión facial o las señales fisiológicas. El reconocimiento de la emoción en un texto es considerado una de las ramas más recientes de la computación afectiva. De hecho, las redes sociales representan una fuente enorme de expresividad emocional textual mayor que cualquier otra. Esta es una de las razones por las que muchos investigadores de áreas como el Procesamiento del Lenguaje Natural (PNL), la Inteligencia Artificial (IA) o la psicología están interesados en este campo.

Los estudios científicos sobre la clasificación de las emociones humanas datan de la década de 1960. Muchos teóricos han propuesto conjuntos de emociones que tienden a

[1] http://www.interior.gob.es/documents/10180/7146983/ESTUDIO+INCIDENTES+DELITOS+DE+ODIO+2017+v3.pdf/5d9f1996-87ee-4e30-bff4-e2c68fade874

ser básicos con características innatas y universales (Tomkins, 1962; Izard, 1992). Si bien los psicólogos no están de acuerdo sobre qué modelo describe con mayor precisión el conjunto de emociones básicas, el más utilizado en la investigación informática es el propuesto por Ekman (1992) con 6 emociones (enfado, repulsión, miedo, alegría, tristeza y sorpresa) (Gholipour Shahraki, 2015).

Uno de los pilares fundamentales en la investigación relacionada con la minería de emociones se centra en los recursos lingüísticos disponibles. Los recursos léxicos son indispensables y existen varios disponibles para el idioma inglés, como WordnetAffect (Strapparava y Valitutti, 2004), Emolex (Mohammad y Turney, 2013) NRC Affect Intensity Lexicon (Mohammad y Kiritchenko, 2018) y LIWC (Pennebaker, Francis, y Booth, 2001). Sin embargo, con respecto a la disponibilidad de recursos para otros idiomas, nos encontramos con que el número es bastante más reducido (Yadollahi, Shahraki, y Zaiane, 2017). Concretamente, para el español podemos citar el recurso Spanish Emotion Lexicon (SEL) de Díaz Rangel, Sidorov, y Suárez Guerra (2014).

Los algoritmos basados en el reconocimiento de emociones en el texto pueden clasificarse en dos categorías: enfoques basados en el léxico y enfoques basados en aprendizaje automático (Cambria, 2016). El primero trata sobre el uso de recursos léxicos u ontologías clasificados por emoción (Mohammad, 2012). El segundo aplica algoritmos estadísticos sobre características lingüísticas, los cuales pueden ser supervisados o no supervisados (Chaffar y Inkpen, 2011).

La incitación al odio y el análisis de emociones están estrechamente relacionados, ya que generalmente las emociones negativas aparecen en comentarios maliciosos. En los últimos años, el interés por desarrollar sistemas para combatir contenido malicioso en las redes sociales apoyándose en técnicas basadas en el reconocimiento de emociones ha incrementado, tanto que, cada vez son más los trabajos que proliferan en el ámbito del PLN. Algunos trabajos realizados hasta el momento siguen un enfoque en el que se aplica en primer lugar un clasificador para detectar comentarios negativos antes de que el clasificador final verifique específicamente si hay evidencia de odio (Dinakar et al., 2012; Sood, Churchill, y Antin, 2012; Gitari et al., 2015).

3 Descripción de la investigación propuesta

Este proyecto de tesis se propone con la finalidad de desarrollar un sistema automático de reconocimiento de emociones en español con el objetivo de aplicarlo a tareas reales, como por ejemplo, la detección del odio en las redes sociales.

En primer lugar, se están estudiando en detalle los trabajos que tratan el reconocimiento de emociones en inglés y en español. Este estudio es fundamental para obtener conocimiento de los enfoques más utilizados en inglés y reproducirlos con objeto de conocer y comparar su funcionamiento en español.

Elegimos realizar el trabajo en español, ya que son escasos los recursos disponibles actualmente en nuestra lengua, a pesar de ser la segunda lengua más hablada en el mundo y la tercera lengua más usada en la Web.

A continuación, se mencionarán los trabajos que se han realizado hasta el momento para la persecución del objetivo de la tesis.

En 2018, uno de nuestros primeros trabajos fue el desarrollo de tres sistemas multilingües con motivo de nuestra participación en tres subtareas (EI-oc, EI-reg, E-c) de la Tarea 1 de SemEval: Affect in Tweets (Mohammad et al., 2018). Son tareas relacionadas con la identificación de la intensidad de la emoción y con la clasificación de emociones en tweets. Nuestra principal contribución fue la implementación de un sistema para adaptar WordNet-Affect al español (Plaza-del-Arco et al., 2018a) utilizando diferentes recursos como BabelNet (Navigli y Ponzetto, 2012) o Babelfy (Moro, Raganato, y Navigli, 2014). Otro de los trabajos relacionado con los recursos léxicos, fue la adaptación del lexicón NRC Affect Intensity construyendo un nuevo lexicón para el español probado sobre el conjunto de datos liberado en la tarea 1 de la competición de SemEval 2018. (Plaza-del-Arco et al., 2018d). Por otra parte, se realizó un sistema automático para categorizar emocionalmente artículos de noticias para la tarea 4 de la competición TASS 2018 (Plaza-del-Arco et al., 2018c). Además, participamos en WASSA 2018 Implicit Emotion desarrollando un sistema basado en redes neuronales para predecir la emoción que expresa una palabra excluida en el texto (Plaza-del-Arco et al., 2018a). Por último, se desarrolló un sistema para el reconocimiento de emociones en el dominio político (Plaza-del-Arco et al.,

2018b).

En 2019, no solo nos centramos en el reconocimiento de emociones en sí, si no que también optamos por realizar sistemas orientados a diferentes aplicaciones dentro del ámbito de la minería de la emoción, como por ejemplo, la identificación del lenguaje ofensivo o la detección de trastornos mentales (anorexia, depresión) en las redes sociales. Por ello, hemos participado en diferentes competiciones que tratan dichas tareas. En primer lugar, participamos una vez más en SemEval, en concreto, en las tareas EmoContext, HatEval y OffensEval. En la primera de ellas, implementamos un sistema automático orientado al reconocimiento de cuatro emociones (enfado, tristeza y alegría y otras) en un diálogo textual entre dos personas incorporando características derivadas de diferentes lexicones afectivos. En la segunda tarea, implementamos un sistema multilingüe para la detección del odio en redes sociales dirigido a dos objetivos específicos: inmigrantes y mujeres. Para la tercera tarea, desarrollamos un sistema con el objetivo de identificar el lenguaje ofensivo en las redes sociales. Por otra parte, este año hemos participamos por primera vez en CLEF eRisk 2019: Early risk prediction on the Internet (Losada, Crestani, y Parapar, 2019) en la tarea 1: Detección temprana de signos de anorexia.

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

- Extraer información subjetiva de las diferentes plataformas (blogs, redes sociales, foros, etc) que dispongan de emociones.
- Generar y adaptar distintos recursos para el reconocimiento de emociones en español, tanto corpus como lexicones.
- Procesar dicha información para desarrollar sistemas que sean capaces de identificar las diferentes categorías emocionales.
- Aplicar los sistemas desarrollados a aplicaciones reales para solucionar problemas actuales.

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 realizará un estudio de la bibliografía existen sobre la minería de emociones en inglés y en español.
2. Creación, adaptación e integración de recursos existentes para poder realizar un análisis de los métodos propuestos. Se intentará crear recursos lingüísticos además de adaptar ciertos recursos ya disponibles en inglés.
3. Desarrollo de un prototipo. Se tratará de implementar un sistema de detección de emociones para el español y se aplicará a determinadas tareas tales como la detección de la incitación al odio.
    - 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 Elementos de investigación específicos propuestos para discusión

Las principales cuestiones de investigación a las que se pretende responder con este proyecto de tesis son las siguientes:

- Estudios psicológicos muestran que las emociones del ser humano van ligadas a su cultura e idioma. Por tanto, ¿es necesario crear recursos emocionales teniendo en cuenta el idioma? o ¿una simple traducción entre recursos es suficiente?
- ¿Qué características se deben tener en cuenta en el proceso del análisis de emociones? ¿Cómo se pueden utilizar estas características para mejorar los sistemas de reconocimiento de emociones?

- ¿Qué algoritmos son los que nos proporcionan una mayor exactitud para reconocer las diferentes categorías emocionales en un texto?
- Dado que los usuarios en las redes sociales es donde más suelen expresar sus emociones, ¿es útil esta información para la creación de recursos léxicos?
- ¿Es útil incorporar conocimiento afectivo para detectar problemas como el odio o los trastornos mentales?. En caso afirmativo, ¿qué características afectivas aportan más valor?

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por el Fondo Europeo de Desarrollo Regional (FEDER) y el proyecto REDES (TIN2015-65136-C2-1-R) del Gobierno de España.

Bibliografía

Cambria, E. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Chaffar, S. y D. Inkpen. 2011. Using a heterogeneous dataset for emotion analysis in text. En *Canadian Conference on Artificial Intelligence*, páginas 62–67. Springer.

Cohen-Almagor, R. 2011. Fighting hate and bigotry on the internet. *Policy & Internet*, 3(3):1–26.

Díaz Rangel, I., G. Sidorov, y S. Suárez Guerra. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein*, 1(29).

Dinakar, K., B. Jones, C. Havasi, H. Lieberman, y R. Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Ekman, P. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Erjavec, K. y M. P. Kovačič. 2012. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6):899–920.

Fersini, E., P. Rosso, y M. Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018.

Fortuna, P. y S. Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Gholipour Shahraki, A. 2015. Emotion mining from text.

Gitari, N. D., Z. Zuping, H. Damien, y J. Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Hasan, M., E. Rundensteiner, y E. Agu. 2019. Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, 7(1):35–51.

Hinduja, S. y J. W. Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221.

Izard, C. E. 1992. Basic emotions, relations among emotions, and emotion-cognition relations.

Losada, D. E., F. Crestani, y J. Parapar. 2019. Overview of eRisk 2019: Early Risk Prediction on the Internet. En *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019*, Lugano, Switzerland. Springer International Publishing.

Mohammad, S. M. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741.

Mohammad, S. M., F. Bravo-Marquez, M. Salameh, y S. Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. En *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Mohammad, S. M. y S. Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. En *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference, Miyazaki, Japan.*

Mohammad, S. M. y P. D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Moro, A., A. Raganato, y R. Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Navigli, R. y S. P. Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Pennebaker, J. W., M. E. Francis, y R. J. Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Picard, R. W. 1997. Affective computing. 1997.

Plaza-del-Arco, F. M., S. M. Jiménez-Zafra, M. Martin, y L. A. Ureña-Lopez. 2018a. Sinai at semeval-2018 task 1: Emotion recognition in tweets. En *Proceedings of the 12th International Workshop on Semantic Evaluation*, páginas 128–132.

Plaza-del-Arco, F., S. M. Jiménez-Zafra, M.-T. Martın-Valdivia, y L. A. Urena-López. 2018b. Using facebook reactions to recognize emotion in political domain.

Plaza-del-Arco, F. M., E. Martínez-Cámara, M. T. M. Valdivia, y L. A. U. López. 2018c. SINAI en TASS 2018: Inserción de conocimiento emocional externo a un clasificador lineal de emociones (SINAI at TASS 2018: Lineal classification system with emotional external knowledge). En *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, páginas 125–130.

Plaza-del-Arco, F. M., M. D. Molina-González, S. M. Jiménez-Zafra, y M. T. Martín-Valdivia. 2018d. Lexicon adaptation for spanish emotion mining. *Procesamiento del Lenguaje Natural*, 61:117–124.

Sood, S. O., E. F. Churchill, y J. Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2):270–285.

Strapparava, C. 2016. Emotions and nlp: Future directions. En *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, página 180.

Strapparava, C. y A. Valitutti. 2004. Wordnet affect: an affective extension of wordnet. En *Language Resources and Evaluation Conference (LREC)*, volumen 4, páginas 1083–1086.

Tomkins, S. 1962. *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.

Yadollahi, A., A. G. Shahraki, y O. R. Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Comput. Surv.*, 50(2):25:1–25:33, Mayo.