Técnicas de Análisis de Sentimientos Aplicadas a la Extracción de Opiniones en el Lenguaje Español

Rosenbrock Germán, Trossero Sebastián, Goette Pablo,
Llorente María Emilia, Pascal Andrés, Cían Damián

Laboratorio de Análisis, Procesamiento, Almacenamiento y Control de Datos
Facultad de Ciencia y Tecnología
Universidad Autónoma de Entre Ríos

german009@gmail.com, sebastian@lambdasi.com.ar, pgoette@hexacta.com,
maria.fcyt@gmail.com, andrespascal2003@yahoo.com.ar, damiancian@gmail.com

RESUMEN

Actualmente existe una gran cantidad de datos textuales disponibles, principalmente en Internet, que crece día a día. El texto es el tipo de dato más utilizado en la web, ya que es fácil de publicar y generar. Lo complejo es obtener información a partir de los mismos en forma automática, y manualmente es sumamente costoso.

La información textual puede dividirse en dos tipos principales: hechos y opiniones. Mientras que los hechos son objetivos, las opiniones representan los sentimientos de cada autor. La Minería de Opinión o Análisis de Sentimientos estudia la extracción de información a partir de datos subjetivos y es relativamente reciente.

Desde hace ya varios años existen sitios web donde los usuarios pueden expresar sus opiniones respecto a diversos temas, por ejemplo, nuevos productos o servicios, imagen empresarial, propuestas de leyes, etc.

Este proyecto propone analizar distintas técnicas de Análisis de Sentimiento aplicadas a opiniones expresadas en el lenguaje Español, evaluar sus resultados para distintos casos reales, y realizar mejoras a las mismas.

CONTEXTO

Este trabajo se encuentra en el marco del Proyecto de Investigación y Desarrollo de Inserción (PIDIN) denominado "Evaluación y Mejora de Técnicas de Minería de Opinión / Análisis de Sentimientos", de la Facultad de Ciencia y Tecnología de Oro Verde de la Universidad Autónoma de Entre Ríos.

Su ejecución forma parte de las actividades del Laboratorio de Análisis, Procesamiento, Almacenamiento y Control de Datos (LAPACDA) perteneciente a la mencionada facultad.

1. INTRODUCCION

En un proceso de toma de decisiones, es fundamental contar con la información oportuna, confiable e íntegra que permita un análisis real de la situación en estudio. En ciertos casos, los datos de origen son opiniones personales. En forma previa a la Web 2.0, su importancia no era alta debido a la falta de grandes cantidades de textos que registren opiniones. Con la disponibilidad masiva de este tipo de información, surgen nuevas oportunidades y desafíos en la búsqueda, comprensión e interpretación de la misma. Sin embargo, la búsqueda en

estos sitios y la posterior valoración de las opiniones en forma manual es un trabajo intenso y costoso, por lo tanto, es necesario contar con sistemas que automaticen este proceso.

El Análisis de Sentimientos o Minería de Opiniones estudia la interpretación automática de opiniones y sentimientos expresados mediante el lenguaje natural.

Es utilizada por ejemplo, por algunas organizaciones para el análisis de su imagen. Además, la literatura muestra varios otros tipos de aplicaciones, incluyendo: valoración de películas [1], opiniones sobre deportes [2], turismo [3, 4], política [5], educación [6], salud [7], finanzas [8] y automóviles [9].

Este trabajo propone aplicar las técnicas Máquinas de Vectores de Soporte (SVM) [2, 10, 11], Clasificador Bayesiano (Naive-Bayes) [12, 13, 14], Redes Neuronales Profundas (DNN) [15, 16], Máxima Entropía [17] y K-Vecinos más Cercanos [18], a conjuntos de opiniones expresadas en el lenguaje español, para luego evaluar sus comportamientos y proponer mejoras.

Los casos de estudio son tres:

- www.cinesargentinos.com.ar (revisión de películas)
- www.booking.com/hotel/ar (comentarios sobre hoteles, versión español)
- www.mercadolibre.com.ar/ (opiniones sobre productos)

La selección de estos sitios se realizó teniendo como criterios la disponibilidad de los datos, la cantidad de opiniones, el nivel de informalidad en el uso del lenguaje, la disponibilidad de una valoración ya registrada para cada opinión, y la existencia de distintos aspectos a evaluar por cada opinión.

2. LINEAS DE INVESTIGACION Y DESARROLLO

La línea principal de investigación es la Minería de Opiniones, para la cual se utilizan técnicas de Inteligencia Artificial; en particular, Procesamiento del Lenguaje Natural (PLN).

3. RESULTADOS OBTENIDOS / ESPERADOS

El objetivo general de este trabajo consiste en evaluar la aplicación de análisis de sentimientos para la obtención de calificaciones cuantitativas a partir de valoraciones textuales cualitativas de reseñas generadas por usuarios de sitios web en el idioma español/castellano, ya que la mayoría de los estudios actuales están hechos sobre el lenguaje Inglés.

Los objetivos específicos son:

- Evaluación de técnicas de análisis de sentimientos aplicadas a opiniones escritas en español/castellano. Las técnicas principales a ser evaluadas son: Naive Bayes, SVM, Entropía Máxima y combinaciones de las anteriores con el uso de diccionarios/léxicos. Además se agregará al menos una más, a seleccionar entre Random Forest, Redes Neuronales Profundas y K-NN, teniendo como criterio su grado de complementación con las anteriores (en el leguaje español), para definir enfoques híbridos.
- Establecer cuáles obtienen mejores resultados para los distintos casos propuestos.
- Realizar mejoras a dichas técnicas.

El aporte principal de este trabajo es el diseño de una técnica híbrida que permita obtener mejores resultados que las técnicas nombradas anteriormente, particularmente en su aplicación a opiniones escritas en español/castellano.

Actualmente hemos extraído datos del sitio www.cinesargentinos.com.ar y conformado una base de datos con 52.307

opiniones sobre diversas películas. Luego utilizamos técnicas de PLN (análisis léxico, eliminación de stopwords, stemming) para extraer las palabras de mayor relevancia para el análisis. Posteriormente entrenamos un modelo usando Naive-Bayes y estamos realizando ajustes y validaciones para asegurar la calidad de los resultados.

4. FORMACION DE RECURSOS HUMANOS

El equipo de trabajo está formado en su totalidad por miembros del Laboratorio de Análisis, Procesamiento, Almacenamiento y Control de Datos (LAPACDA) de la Facultad de Ciencia y Tecnología de la UADER de la sede de Concepción del Uruguay y de Oro Verde. Participan docentes de la carrera Lic. en Sistemas de Información de la UADER (dos de ellos realizando una Maestría en Minería de Datos), y alumnos avanzados de la carrera de grado.

5. BIBLIOGRAFIA

[1] Kuat Yessenov. Sentiment Analysis of Movie Review Comments. 2009.

[2] N. LI and D. D. W. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems, vol. 48, nº 2, pp. 354 - 368, 2010.

[3] L. C. Fiol, J. S. García, M. M. T. Miguel and S. F. Coll, «La importancia de las comunidades virtuales para el análisis del valor de marca. El caso de TripAdvisor en Hong Kong y París,» Papers de turisme, nº 52, pp. 89-115, 2012.

[4] C. Henriquez, J. Guzmán and D. Salcedo. Minería de Opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles. Procesamiento del Lenguaje Natural, vol. 56, pp. 25-32., 2016.

[5] S. Rill, D. Reinel, J. Scheidt and R. V. Zicari. PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. Knowledge-Based Systems, vol. 69, pp. 24-33, 2014.

[6] A. Ortigosa, J. M. Martín and R. M. Carro. Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior, vol. 31, pp. 527-541, 2014.

[7] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi and L. Donaldson. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. Journal of medical Internet research, vol. 15, nº 11, 2013.

[8] X. Dong, Q. Zou and Y. Guan. Set-Similarity joins based semi-supervised sentiment analysis. Neural Information Processing. Springer Berlin Heidelberg, 2012., from Neural Information Processing, Springer Berlin Heidelberg, 2012, pp. 176-183.

[9] P. D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics, Stroudsburg, PA, USA, 2002.

[10] A. Abbasi, H. Chen and A. Salem. Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Transactions on Information Systems (TOIS), vol. 26, nº 3, p. 12, 2008.

[11] F. Pla and L.-F. Hurtado. Sentiment Analysis in Twitter for Spanish. Natural Language Processing and Information Systems, pp. 208-213, 2014.

[12] A. Lazaridou, I. Titov and C. Sporleder. A Bayesian Model for Joint Unsupervised Induction of Sentiment, Aspect and Discourse Representations.. ACL, vol. 1, pp. 1630 -1639, 2013.

[13] A. Pak and P. Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC, vol. 10, pp. 1320 -1326, 2010.

[14] J. Ortigosa-Hernández, J. D. Rodríguez, L. Alzate, M. Lucania, I. Inza and J. A. Lozano. Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers. Neurocomputing, vol. 92, pp. 98-115, 2012.

[15] M. Ghiassi, J. Skinner and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. vol. 40, nº 16, pp. 6266-6282, 2013.

[16] M. Anjaria and R. M. R. Guddeti. A novel sentiment analysis of social networks using supervised learning. Social Network Analysis and Mining, vol. 4, nº 1, pp. 1-15, 2014.

[17] Pang, Bo & Lee, Lillian & Vaithyanathan, Shivakumar. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. EMNLP. 10. . 10.3115/1118693.1118704.

[18] Jędrzejewski, K. & Zamorski, M. (2013). Performance of K-Nearest Neighbors Algorithm in Opinion Classification. Foundations of Computing and Decision Sciences, 38(2), pp. 97-110. 2017.