Análisis comparativo de tareas de pre procesamiento de textos sobre contenido extraído de redes sociales

Leonardo Esnaola[1 4], Juan Pablo Tessore[2 4], Hugo Ramón[3 4], Claudia Russo[3 4]

Instituto de Investigación y Transferencia en Tecnología (ITT)
Comisión de Investigaciones Científicas (CIC)
Escuela de Tecnología (ET)
Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA)

Sarmiento y Newbery, 236-4636945/44

leonardo.esnaola@itt.unnoba.edu.ar / juanpablo.tessore@itt.unnoba.edu.ar / hugo.ramon@itt.unnoba.edu.ar / claudia.russo@itt.unnoba.edu.ar

Resumen

El texto que surge de la interacción entre usuarios en redes sociales suele ser más disperso que el contenido tradicional. Es decir, contiene errores ortográficos, uso informal del lenguaje, emoticones, urls y otras construcciones que no suelen estar presentes en el lenguaje formal. Dicha dispersión puede afectar el desempeño de los clasificadores de texto basados en aprendizaje automático.

El presente trabajo propone medir el desempeño de diferentes tareas de pre-procesamiento, aplicadas primero de manera aislada y luego combinadas, sobre contenido extraído de redes sociales. Se busca determinar cuán aptas resultan ser estas tareas para corregir errores en textos de este tipo. Para ello, en primer lugar, se determinará en qué magnitud se reduce el porcentaje de palabras "incorrectas" y, en segundo lugar, cómo impactan en la precisión final alcanzada por clasificadores basados en aprendizaje automático.

Este trabajo, se enmarca en una línea de investigación más amplia que propone la construcción de un clasificador automático de opiniones utilizando algoritmos de aprendizaje automático, el cual fuera presentado previamente en otra edición de este Workshop [1], y que permitirá realizar análisis automáticos de bajo costo para determinar las emociones manifestadas por consumidores o usuarios acerca de productos o servicios, a partir del análisis de sus opiniones escritas. Este clasificador será entrenado a partir de los comentarios en lenguaje informal presente en redes sociales.

***Palabras clave*:** Minería de textos, Pre-procesamiento, Inteligencia artificial, Redes sociales.

Contexto

Esta línea de investigación forma parte del proyecto "Tecnología y aplicaciones de Sistemas de Software: Innovación en procesos, productos y servicios" presentado en el marco de la convocatoria a Subsidios de

---

[1] Docente Investigador ITT / Doctorando UNLP
[2] Docente Investigador ITT / Becario Doctoral CIC
[3] Docente Investigador ITT - Investigador Asociado Adjunto sin director CIC
[4] ITT - Centro Asociado CIC

Investigación Bianuales (SIB2019) de la Secretaría de Investigación, Desarrollo y Transferencia de la UNNOBA. A su vez se enmarca en el contexto de un plan de trabajo aprobado por la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires y por la Secretaría de Investigación de la UNNOBA, en el marco de la convocatoria "Becas de Estudio Cofinanciadas 2015 CIC Universidades del interior bonaerense".

El proyecto se desarrolla en el Instituto de Investigación y Transferencia en Tecnología (ITT) dependiente de la mencionada Secretaría, y se trabaja en conjunto con la Escuela de Tecnología de la UNNOBA.

El equipo está constituido por docentes e investigadores pertenecientes al ITT y a otros Institutos de Investigación, así como también, estudiantes de las carreras de Informática de la Escuela de Tecnología de la UNNOBA.

Introducción

Con el auge de las redes sociales, la gente está cada vez más involucrada en muchos aspectos en los en los cuales antes solo eran consumidores pasivos [2]. Dichas redes permiten a las personas expresarse de manera libre y rápida acerca de una gran variedad de temas, y están siendo aprovechadas por los sitios comerciales para influenciar a los usuarios con campañas de marketing dirigidas [3]. Según [4], las reacciones de Facebook proveen una oportunidad para medir el compromiso emocional de los consumidores. Sin embargo, procesar cientos de textos para entender que emoción reflejan no es una tarea sencilla y requiere mucho trabajo manual.

Según [5], la minería de textos se encarga del descubrimiento automático o semiautomático de información nueva, previamente desconocida y de alta calidad a partir de un gran número de textos no estructurados. La información que se quiere inferir de los textos se suele especificar manualmente de antemano. En [6] los autores definen que la minería de textos combina técnicas de tres campos específicos:

- Recuperación de información: implica la recopilación de información relevante para realizar una tarea determinada.
- Procesamiento de lenguaje natural: combina una variedad de técnicas para analizar y representar textos que ocurren naturalmente en uno o más niveles de análisis lingüístico con el propósito de lograr un procesamiento de lenguaje similar al humano para una variedad de tareas o aplicaciones [7].
- Minería de datos: descubre patrones en la información estructurada que se ha construido a partir de los textos.

El desarrollo de las redes sociales ha aumentado la disponibilidad de contenido en forma de texto, creando la materia prima necesaria para aplicar allí técnicas de minería de texto y extraer información significativa.

El presente trabajo busca medir la efectividad de diferentes técnicas de pre procesamiento sobre textos "ruidosos" provenientes de las redes sociales. A diferencia de otros estudios [8], el proceso se lleva a cabo sobre un *dataset* de comentarios en español compilados desde Facebook. Teniendo en cuenta los estudios relevados, no existe una investigación que informe la efectividad de estas técnicas sobre este tipo de conjunto de datos.

Esta investigación es parte de una investigación más amplia que se centra en la

construcción de un detector automático de emociones a partir de texto donde, a diferencia de otros estudios [9], las etiquetas que denotan la emoción reflejada en cada comentario se obtienen automáticamente de las reacciones de Facebook, en lugar de clasificarlos manualmente. Esta aplicación podría ofrecer una amplia gama de aplicaciones potenciales, como detectar la emoción que surge de la opinión de grandes grupos de personas sobre ciertos productos, servicios [10, 11, 12] o incluso políticas públicas. También podría utilizarse para identificar demandas o quejas no cumplidas de ciudadanos; En seguridad, para la detección automática de factores de riesgo en redes sociales como amenazas o ciberacoso [13].

Para desarrollar esta aplicación, basándose en un enfoque de aprendizaje automático, es necesario numerizar el texto de entrada. Esto se logra utilizando algunas métricas de texto. Una de las métricas más ampliamente adoptadas es "*term frecuency - inverse document frecuency*" (tf-idf). Es bien sabido que los datos provenientes de redes sociales suelen ser muy dispersos [14], y esta dispersión puede afectar el rendimiento de las aplicaciones que se basan en estadísticas de frecuencia de palabras [15], como tf-idf.

Debido a esto, es necesario aplicar algunas tareas de pre procesamiento a los datos de entrada para mejorar el rendimiento del clasificador de emociones a implementar.

Reparar el texto de los comentarios, realizar, por ejemplo, una corrección ortográfica, eliminar signos de puntuación incorrectos, enlaces y otras intervenciones puede eliminar características innecesarias de la entrada, haciendo que el análisis posterior sea más rápido y más preciso.

Diversos trabajos que realizan pre procesamiento de textos provenientes de las redes sociales [14, 15, 16] hacen, en primer lugar, una clasificación de las palabras del *dataset* analizado en dos categorías. Estas son palabras "*in vocabulary*" (IV) y palabras "*out of vocabulary*" (OOV), siendo estas últimas las palabras incorrectas según el lenguaje formal. Para clasificar la entrada de dicha manera, se emplean un diccionario en el idioma correspondiente y una herramienta de corrección ortográfica. Entre estas últimas, las más utilizadas [17], son *Hunspell* [18] y *GNU Aspell* [19].

Una vez identificadas las OOV, se aplican una serie de filtros al texto para reducir el porcentaje de OOV sobre el número total de palabras del conjunto de datos.

Líneas de Investigación, Desarrollo e Innovación

La presente investigación se encuadra dentro del eje "Gestión de la Innovación" del mencionado proyecto SIB. Dicho eje incluye la investigación de procesos metodológicos para abordar Innovación y su implicación medible en procesos productivos.

En este sentido, es importante transformar al usuario potencial en un partícipe activo del proceso de producción de un determinado producto, o en la concepción y desarrollo de un determinado servicio. Internet permite estar en contacto con un número inmenso de potenciales usuarios, pero poder procesar todas las interacciones buscando reconocer si estos usuarios están satisfechos o no (y en qué medida) sobre alguna característica de un producto o servicio, requiere analizar estas múltiples interacciones lo cual podría consumir mucho tiempo de trabajo.

Un clasificador automático de emociones a partir del análisis de texto podría ser útil para este propósito. Así, se presentan en la Figura 1 las diferentes etapas identificadas para su desarrollo y puede apreciarse en cuál de ellas se enmarcan las tareas de pre-procesamiento presentadas en este trabajo:

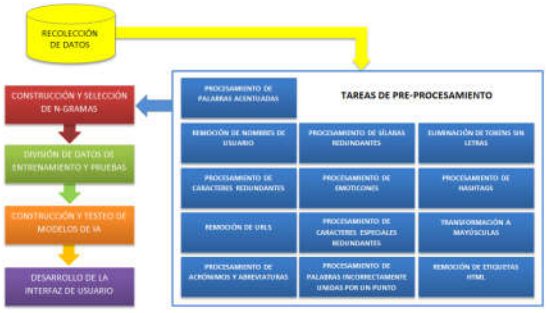

Figura 1: Etapas del proceso de construcción de la herramienta, incluyendo las tareas de pre procesamiento.

Resultados y Objetivos

Se espera que la presente línea de I/D posibilite una mejora en la exactitud de un clasificador automático de textos, a partir de la reparación de los comentarios utilizados para su entrenamiento. Esta reparación se presume conducirá a reducir la variabilidad de *tokens* a utilizar, haciendo que, en primer lugar, el clasificador construido se pueda entrenar más rápidamente y, en segundo lugar, que los *tokens* que podrían generar "ruido" sean suprimidos.

Asimismo, se espera poder comparar las diferencias entre los *tokens* generados por esta implementación con aquellos generados por tokenizadores conocidos, como Gensim y NLTK. Esta comparación podrá realizarse no sólo sobre los textos de la fuente seleccionada, sino que podría realizarse sobre textos de otras fuentes, como por ejemplo Twitter, dado que hay estudios [20] que afirman que los resultados serían similares.

Así mismo, se buscan generar informes técnicos en base al trabajo realizado, en donde se registren los avances, el grado de implementación y los resultados obtenidos. Como así también difundir y transferir los resultados y logros alcanzados mediante la presentación y participación en diferentes congresos, jornadas y workshops de carácter nacional e internacional vinculados a la temática de estudio.

Formación de Recursos Humanos

En esta línea de I/D se han obtenido y se encuentran desarrollando actualmente dos becas de iniciación a la investigación. Asimismo, se esperan desarrollar una tesina de grado y una PPS, todas ellas dirigidas por miembros de este proyecto.

Bibliografía

[1] Tessore J. Esnaola L. Russo C. Ramón H. and Pompei S. 2018. Análisis automático de grandes volúmenes de datos en redes sociales mediante minería de textos combinado con algoritmos inteligentes. XX Workshop de Investigadores en Ciencias de la Computación. Universidad Nacional del Nordeste. Corrientes, Argentina.

[2] Xia Hu and Huan Liu. 2012. Mining Text Data. Text Analytics in Social Media. Springer, Boston, MA. DOI: https://doi.org/10.1007/978-1-4614-3223-4_12

[3] Charu C. Aggarwal and ChengXiang Zhai. 2012. Mining Text Data. An Introduction to Text Mining. Springer, Boston, MA. DOI: https://doi.org/10.1007/978-1-4614-3223-4_1

[4] Sarah Turnbull and Simon Jenkins.

2016. Why Facebook Reactions are good news for evaluating social media campaigns. Direct, Data and Digital Marketing Practice. (Feb. 2017), 17-156. DOI: https://doi.org/10.1057/dddmp.2015.56

[5] Henning Wachsmuth. 2015. Text Analysis Pipelines: Towards Ad-hoc Large-Scale Text Mining. Springer International Publishing. DOI: 10.1007/978-3-319-25741-9_2.

[6] Sophia Ananiadou and John Mcnaught. 2005. Text Mining for Biology And Biomedicine. Artech House Inc., Norwood, MA.

[7] Elizabeth Liddy. 2001. Encyclopedia of Library and Information Science, 2nd Ed. Natural Language Processing. Marcel Decker Inc., New York, NY.

[8] Eric S. Tellez, Sabino Miranda-Jimnez, Mario Graff, Daniela Moctezuma, Oscar S. Siordia, and Elio A. Villaseor. 2017. A case study of Spanish text transformations for twitter sentiment analysis. Expert Syst. Appl. 81, C (September 2017), 457-471. DOI: https://doi.org/10.1016/j.eswa.2017.03.071

[9] Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis. Know.-Based Syst. 89, C (November 2015), 14-46. DOI: https://doi.org/10.1016/j.knosys.2015.06.015

[10] Johan Bollen and Huina Mao. 2011. Twitter Mood as a Stock Market Predictor. Computer 44, 10 (October 2011), 91-94. DOI: http://dx.doi.org/10.1109/MC.2011.323

[11] Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (K-CAP '03). ACM, New York, NY, USA, 70-77. DOI: http://dx.doi.org/10.1145/945645.945658

[12] Alvaro Ortigosa, José M. Martín, and Rosa M. Carro. 2014. Sentiment analysis in Facebook and its application to e-learning. Comput. Hum. Behav. 31 (February 2014), 527-541. DOI: http://dx.doi.org/10.1016/j.chb.2013.05.024

[13] Mifta Sintaha, Satter, S. Bin, Niamat Zawad, Chaity Swarnaker and Ahanaf Hassan. 2016. Cyberbullying detection using sentiment analysis in social media. Brac University, Dhaka, Bangladesh.

[14] Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2015. TweetNorm: a benchmark for lexical normalization of Spanish tweets. Lang. Resour. Eval. 49, 4 (December 2015), 883-905. DOI: http://dx.doi.org/10.1007/s10579-015-9315-6

[15] Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. ACM Trans. Intell. Syst. Technol. 4, 1, Article 5 (February 2013), 27 pages. DOI: http://dx.doi.org/10.1145/2414425.2414430

[16] Naradhipa, Aqsath Rasyid, and Ayu Purwarianti. 2012. Sentiment classification for Indonesian message in social media. International Conference on Cloud Computing and Social Networking. IEEE, 2012.

[17] Eleanor Clark and Kenji Araki. 2011. Text Normalization in Social them Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. Procedia - Social and Behavioral Sciences, vol. 27. 2-11.

[18] Hunspell.github.io. 2019. Hunspell: About. Retrieved from: http://hunspell.github.io/.

[19] GNU Aspell. 2019. GNU Aspell. Retrieved from: http://aspell.net/.

[20] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay and Li Wang. 2013. How noisy social media text, how diffrnt social media sources?. Proceedings of the 6th International Joint Conference on Natural Language Processing (January 2013).