ATICA 2020

Aplicación de Tecnologías de la Información y Comunicaciones Avanzadas y Accesibilidad

Luis Bengochea
Gerardo Contreras Vega
(Editores)

OBRAS COLECTIVAS
TECNOLOGÍA 32

UAH

Desarrollo de un recurso léxico de palabras informales en español de Argentina para el análisis de sentimientos en Twitter

Víctor Rojo [1, 2], María Florencia Pollo-Cattaneo[1, 2], Paola Britos[3]

[1] Programa de Maestría en Ingeniería en Sistemas de Información. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional. Argentina.

[2] Grupo de Estudio en Metodologías de Ingeniería de Software (GEMIS). Facultad Regional Buenos Aires. Universidad Tecnológica Nacional. Argentina.

[3] Universidad Nacional de Río Negro. Laboratorio de Informática Aplicada. Río Negro, Argentina.
{vmrojo, flo.pollo}@gmail.com, pbritos@unrn.edu.ar

Resumen. La red social Twitter se destaca como uno de los repositorios de opiniones más grandes en línea y mantiene el interés de investigadores y organizaciones que buscan explotar los datos generados en la plataforma. El análisis de sentimientos como método para el procesamiento automático de textos subjetivos tiene como meta la clasificación de los mensajes en categorías de polaridad, por lo que las soluciones que se implementan a menudo hacen uso de recursos léxicos para auxiliar a los clasificadores. Si bien estos recursos resultan fundamentales para la tarea, los léxicos disponibles son limitados, por lo que en trabajos pasados se ha identificado la necesidad de desarrollar recursos exclusivos para el tratamiento de textos en español. En este artículo se describen los pasos seguidos en el desarrollo de un recurso léxico a disponibilizar a futuro que incorpora modismos, lunfardismos y otras palabras informales utilizadas en Argentina.

Palabras clave: Recurso Léxico. Análisis de Sentimientos. Minería de Opiniones. Twitter.

1. Introducción

En tiempos recientes, la red social Twitter se mantiene como una de las plataformas más populares para la discusión de temas y el intercambio de opiniones sobre personas, marcas y eventos recientes. Los millones de mensajes que se generan en la aplicación todos días a raíz de las interacciones entre sus usuarios resultan de interés para organizaciones ya que de los datos agregados es posible generar perfiles, relevar los sentimientos de consumidores o votantes y realizar predicciones que ayuden en la toma de decisiones estratégicas.

El análisis de sentimientos ha surgido como un método para el tratamiento computacional de las opiniones y sentimientos en textos subjetivos. El objetivo de este tipo de análisis comúnmente consiste en la clasificación automática de los mensajes en categorías de polaridad como "positivo", "negativo" o "neutral". Los enfoques que se observan con mayor frecuencia en los trabajos relacionados al análisis de sentimientos en Twitter se pueden separar en 3 grupos: soluciones basadas en léxicos, en aprendizaje automático o un abordaje híbrido [1,2].

Aquellas soluciones que siguen técnicas basadas en léxico hacen uso de recursos léxicos, también llamados léxicos de opiniones, los cuales pueden ser generales o derivados de un corpus [3]. Un recurso léxico es una lista de palabras o frases ("bueno", "malo" o "me gusta", por ejemplo) que han sido previamente anotadas con etiquetas de polaridad. Si bien otros casos como negaciones o el uso de sarcasmo deben ser considerados, la utilización del recurso léxico permite relacionar la presencia de una palabra o frase en una oración a un sentimiento.

En el enfoque basado en aprendizaje automático, la tarea es planteada como un problema de clasificación [4]. Generalmente, en estos trabajos se emplean modelos supervisados o no supervisados que clasifican los textos en una de las etiquetas de polaridad denominadas. Estos métodos requieren de grandes cantidades de datos, en ocasiones etiquetados, para el entrenamiento de los modelos lo cual resulta costoso; algunos sistemas tratan de mitigar el problema por medio de métodos semi-supervisados [5].

Por último, para aprovechar los puntos fuertes o compensar las desventajas de las técnicas anteriores, algunos trabajos [6] se inclinan por realizar una combinación de los enfoques mencionados.

Ya sea para implementar una solución basada en léxico o para la extracción de características para los clasificadores, los investigadores utilizan recursos auxiliares que resultan fundamentales para la clasificación de los textos [7,8]. En trabajos pasados se ha explorado la necesidad de generar recursos, incluyendo recursos léxicos, exclusivos para el análisis de sentimientos en español [9,10]. En este trabajo se continúa por esta línea de investigación para producir una lista de palabras informales, lunfardismos y otras voces utilizadas en el habla popular junto con las anotaciones de polaridad correspondientes necesarias para el análisis de sentimientos.

2. Elaboración del Recurso Léxico

2.1. Fuentes y recolección de términos

Para compilar la lista de palabras y frases que eventualmente pasaron a conformar el léxico de términos se consideraron distintas fuentes. Existen recursos especializados como resultado del estudio de otras disciplinas que recopilan vocablos de usanza popular en Argentina como el "Diccionario del habla de los argentinos" (DiHA) en su edición del 2017 o su sucesor publicado en el 2019, el "Diccionario de la lengua de la Argentina" (DiLA), ambos producidos por la Academia Argentina de Letras (AAL).

Si bien estas fuentes concentran un gran listado de regionalismos y frases comunes, no fue posible al momento de la investigación encontrar una versión digital que facilitara la adopción y procesamiento de sus contenidos. Por este motivo, se le dio

preferencia a tres diccionarios disponibles en línea. Los diccionarios que se consideraron para la tarea se presentan a continuación.

2.1.1. Diccionario Argentino

El Diccionario Argentino es un recurso gratuito en línea que presenta una recopilación de palabras y modismos argentinos tradicionales y modernos con el fin de compartir información sobre su uso y promover su entendimiento [11]. El listado de palabras se conforma en su totalidad de contribuciones de sus propios usuarios; como una forma de control sobre la calidad de los aportes, el sitio proporciona mecanismos para calificar positiva o negativamente las definiciones que se agregan a los términos. Las características de esta fuente son muy similares a las del diccionario en inglés Urban Dictionary, el cual se ha utilizado en varias ocasiones en la construcción de otros recursos léxicos y como auxiliar en las tareas relacionadas al análisis de sentimientos en Twitter [12,13,14].

Al momento de consultarse, se obtuvieron 1868 términos. El sitio se encuentra disponible en https://www.diccionarioargentino.com.

2.1.2. Diccionario de Argentinismos

El Diccionario de Argentinismos es una pequeña colección de palabras empleadas por el autor en lo cotidiano, recolectadas con la intención de resaltar diferencias entre el idioma que se habla en España y el de Argentina [15]. La lista, según su autor, no pretende ser exhaustiva y maneja un tono informal en sus definiciones. De esta fuente se recuperaron 566 palabras con sus definiciones.

La página podía ser visitada a través de http://argentinismos.tripod.com aunque dejó de estar disponible a principios del 2020; aún es posible consultar su contenido por medio de sitios de archivo.

2.1.3. Lunfa2000

La página Lunfa2000 es un repositorio dedicado en su mayoría a recopilar y documentar palabras del lunfardo. Entre las múltiples listas de palabras y artículos que se publicaron en el sitio, destacan dos listas formadas por palabras que, en su momento, no se habían incluido en las ediciones más recientes del Diccionario de la Real Academia Española (DRAE) o el DiHa [16,17]. Siguiendo el ejemplo de estas fuentes, cada término es anotado con etiquetas como “coloq.”, “lunf.” y otras abreviaciones comunes que indican cualidades propias de la palabra y su uso.

Entre las dos colecciones se totalizan 2002 palabras que fueron analizadas para determinar su inclusión en el nuevo recurso. Las listas se encuentran disponibles en http://geocities.ws/lunfa2000/aal.htm y http://geocities.ws/lunfa2000/aal2.htm.

Todos los términos de las distintas fuentes fueron recolectados utilizando rastreadores web desarrollados en Python y diseñados específicamente para cada sitio. Dentro de las particularidades de cada fuente se tuvieron que considerar los siguientes escenarios para poder generar listas más completas:

- La expansión de algunas abreviaciones para indicar género o plurales. Ej. “desprolijo, ja” a “desprolijo” y “desprolija”.
- La expansión de términos si se especifican múltiples usos. Ej. “rabona (hacerse la)” a “rabona” y “hacerse la rabona”.
- Ignorar la inclusión de palabras con una mala puntuación en el sitio (menor a cero).
- Ignorar términos formados únicamente por palabras vacías (*stop words*) o en listas negras.

Si bien de las distintas fuentes se recolectaron más de 4400 registros en esta primera etapa, como se puede ver en la Tabla 1, este número no representa el total definitivo ya que durante una segunda revisión se terminaron de separar varios términos agrupados, lo que resultó en poco más de 1000 nuevas palabras que se agregaron. El refinamiento que se realiza en la siguiente fase parte de una base que alcanza las 5488 palabras.

Tabla 1. Número de registros por diccionario y totales.

Fuente	**Registros**
Diccionario Argentino	1868
Diccionario de Argentinismos	566
Lunfa2000 I	1001
Lunfa2000 II	1001
Total (recolección inicial)	4436
Total (luego de revisión)	**5488**

2.2. Refinamiento y Anotación de palabras

Luego de normalizar las listas de términos aplicando operaciones para pasar las palabras a minúsculas y remover tildes y otros símbolos, el primer refinamiento que se empleó fue para sacar registros repetidos entre los distintos diccionarios. En total se removieron del listado original 779 palabras al terminar el proceso de eliminación de duplicados.

Durante el proceso de anotación (ver más adelante), se realizó una nueva revisión de las palabras a medida que se procesaban y se identificaron nuevos criterios para excluir registros del léxico final. Las razones incluyen:

- *Común*. La palabra es de uso común, y si bien puede existir alguna diferencia con lo que se usa en otros países, es difícil encontrar su uso en el sentido informal. Esta categoría también incluye referencias a nombres de personas (Diego), nacionalidades o lugares (turco), marcas (Gancia Sprite) y personajes (Pikachu).
- *Derivada*. La palabra es derivada de otra palabra en el listado que ya fue considerada o eliminada (tirarle una onda).
- *Duplicada*. Similar a la categoría anterior y parte del primer refinamiento. La palabra coincide exactamente con otra palabra extraída de alguno de los otros diccionarios.

- *No aporta*. La palabra no contribuye de forma significativa en lo relacionado a polaridad. Si bien se incluyen palabras en el léxico que tienen la etiqueta NEU (neutral), se procuró que estas sean únicas de la región (rioba, ñoba) y no solo variaciones menores de algún término (aceto balsámico, azúcar impalpable).
- *No es un término*. Como el nombre explica, el registro en la lista no es una palabra o frase. La categoría incluye emoticons (:v), abreviaciones (idk) exclamaciones (eaaaa) o lo que se puede considerar como un meme (f to pay respects).
- *Poco uso*. Como resultado de un análisis que se llevó a cabo sobre una gran colección de tweets recolectados a lo largo de varios meses, se eliminaron aquellas palabras sin presencia en los textos o con menciones esporádicas. Esto también incluye búsquedas puntuales posteriores sobre algunos términos que regresaron resultados nulos o insignificantes.
- *Sin definición*. Durante el proceso de recolección de palabras, no fue posible extraer la definición de la palabra o frase. En la mayoría de los casos, esto lo ocasionaba la presencia de algún carácter especial o una página malformada.
- *Misceláneo*. Casos que no encajan en las categorías anteriores y ocurren con poca frecuencia. Incluye, por ejemplo, palabras que no pertenecen al lenguaje informal o al lunfardo (aburrido), palabras en otros idiomas (beef, moishe) o palabras irrelevantes y que escaparon otros filtros (breadwhatwhat).

En la Tabla 2 se presenta el total de términos eliminados y su detalle.

Tabla 2. Número de registros excluidos por motivo y totales

Motivo	**Registros**
Común	317
Derivada	28
Duplicada	779
No aporta	995
No es un término	32
Poco uso	465
Sin definición	155
Misceláneo	137
Total	**2908**

Al descontar estas palabras, el listado final resulta en 2580 términos a etiquetar en el recurso léxico.

Los métodos que se han utilizado en el pasado para producir las anotaciones varían según su automatización: se pueden realizar anotaciones manuales llevadas a cabo por una o varias personas o aplicar traducciones automáticas de léxicos ya calificados [9]. En vista de que la mayoría de las palabras no cuentan con una traducción por ser propias del español de Argentina, se tuvo que recurrir a la asignación manual de las polaridades. Debido a la dificultad que representa encuestar a personas en un número tan extenso de palabras, se optó por un enfoque híbrido en el que primero se realiza

una clasificación por el autor, similar a la que se hizo para el léxico AFINN [12], y posteriormente se encuesta a un grupo de voluntarios en las palabras que no cuentan con una polaridad clara o que tratan conceptos más complejos. Las palabras fueron categorizadas en 3 grupos que reflejan un grado de polaridad: NEG (negativo), NEU (neutral), y POS (positivo).

Las encuestas se realizaron por medio de QuestionPro, una plataforma que permite definir cuestionarios de forma similar a herramientas como Survey Monkey que han sido utilizadas en la generación de otros recursos [18]. Las instrucciones de la encuesta consistían en pedirle al voluntario que evaluara cada una de las palabras o frases según su uso hipotético en una oración informal. Las opciones a elegir correspondían a las categorías mencionadas anteriormente ("Negativo", "Neutral" y "Positivo"), además de una cuarta opción en caso de que se desconozca la palabra ("No la conozco"). Se limitó el número de palabras a menos de 20 términos por página y no más de 3 páginas por encuesta para disminuir la tasa de abandono.

3. Los resultados

Al momento de la redacción de este trabajo, se encuentra en curso una última encuesta de la que se espera obtener alrededor de 50 nuevas etiquetas una vez que concluya. Actualmente, la lista anotada con la que se cuenta se distribuye según los valores de la Tabla 3.

Tabla 3. Número de términos por polaridad en el recurso léxico

Polaridad	Términos
NEG	1170
NEU	1106
POS	251
Total	**2527**

4. Futuros trabajos

En futuros trabajos se planea disponibilizar el listado de palabras y frases informales junto con las polaridades asignadas que surgen como resultado de este desarrollo. Además se contempla aplicar el nuevo recurso léxico en un clasificador de tweets para poner a prueba su efectividad.

5. Conclusiones

Los recursos léxicos juegan un papel fundamental en una gran cantidad de sistemas que implementan el análisis se sentimientos en Twitter. Como se ha visto en trabajos anteriores, el desarrollo de recursos exclusivos para el español es una necesidad que ha sido reconocida por los investigadores dedicados al estudio de estos problemas.

Se espera que el recurso generado sirva como base para otros recursos en el lenguaje o como parte del tratamiento de textos subjetivos que hagan uso de modismos, regionalismos, lunfardismos y otras figuras de la comunicación informal en Argentina.

6. Referencias

1. Wehrmann J, Becker W, Cagnini HEL, Barros RC. A Character-Based Convolutional Neural Network for Language-Agnostic Twitter Sentiment Analysis. In Neural Networks (IJCNN), 2017 International Joint Conference; 2017; Anchorage, AK, USA: IEEE. p. 2384-2391.
2. Giachanou A, Crestani F. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. ACM Computing Surveys (CSUR). 2016 Noviembre; 49(2).
3. Kharde VA, Sonawane SS. Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications. 2016; 139.
4. Hurtado Oliver LF, Pla F, Buscaldi D. ELiRF-UPV en TASS 2015: Análisis de Sentimientos en Twitter. In TASS workshop at SEPLN 2015; 2015. p. 75-79.
5. Nakov P. Semantic Sentiment Analysis of Twitter Data. In Encyclopedia on Social Network Analysis and Mining (ESNAM).; 2017.
6. Kolchyna O, Souza TTP, Treleaven PC, Aste T. Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. Handbook of Sentiment Analysis in Finance. 2015.
7. Pang B, Lee L. Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval. 2008; 2(1-2): p. 1-135.
8. Liu B. Sentiment Analysis and Subjectivity. Handbook of natural language processing. 2010; 2: p. 627-666.
9. Rojo V, Britos P, Pollo-Cattaneo MF. Revisión de enfoques y comparación de recursos para el análisis de sentimientos en español en Twitter. In Desarrollo e Innovación en Ingeniería – Cuarta Edición.: Editorial Instituto Antioqueño de Investigación; 2019. p. 5-16.
10. Rojo V, Pollo-Cattaneo MF, Britos P. Análisis de Sentimientos en Twitter: Desarrollo de Recursos en el Español Rioplatense de Argentina. In XXII Workshop de Investigadores en Ciencias de la Computación; 2020; El Calafate, Santa Cruz.
11. Diccionario Argentino. [Online]. [cited 2019 09 17. Available from: https://www.diccionarioargentino.com/.
12. Nielsen FÅ. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages; 2011. p. 93-98.
13. Tang D, Wei F, Qin B, Zhou M, Liu T. Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. In Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers; 2014. p. 172-182.
14. Wu L, Morstatter F, Liu H. SlangSD: Building and Using a Sentiment Dictionary of Slang Words for Short-Text Sentiment Classification. [Online].; 2016. Available from: https://arxiv.org/abs/1608.05129.

15. Diccionario de Argentinismos. [Online]. [cited 2019 09 17. Available from: http://argentinismos.tripod.com.
16. López N. 1001 Palabras que se usan en la Argentina y no están en el Diccionario del Habla de los Argentinos. [Online].; 2004. Available from: http://geocities.ws/lunfa2000/aal.htm.
17. López N. 1001 Palabras que se usan en la Argentina y no están en el Diccionario del Habla de los Argentinos (II). [Online].; 2005. Available from: http://geocities.ws/lunfa2000/aal2.htm.
18. Hinojosa JA, Martínez-García N, Villalba-García C, Fernández-Folgueiras U, Sánchez-Carmona A, Pozo MA, et al. Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. Behavior research methods. 2016; 48(1): p. 272-284.
19. Karami A, Bennett LS, He X. Mining Public Opinion about Economic Issues: Twitter and the U.S. Presidential Election. International Journal of Strategic Decision Sciences. 2018 Enero; 9(1): p. 18-28.