ANÁLISIS DE SENTIMIENTOS EN TWITTER: DESARROLLO DE RECURSOS EN EL ESPAÑOL RIOPLATENSE DE ARGENTINA

Rojo, V.[1,2], Pollo-Cattaneo, Ma. F[1,2]; Britos, P.[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

Twitter se ha posicionado como una de las redes sociales más importantes para el intercambio y concentración de opiniones, convirtiéndose en un ambiente idóneo para el procesamiento automático de estos textos subjetivos a través del análisis de sentimientos. El desafío de la clasificación de tweets se trata de afrontar con ayuda de recursos especializados, tales como corpora, léxicos y herramientas de análisis, los cuales suelen tener un impacto significativo en los resultados de los clasificadores. Este, junto con futuros trabajos, busca colaborar a equiparar las condiciones del análisis de sentimientos en Twitter en español a sus homólogos en otros idiomas por medio del desarrollo de un nuevo recurso léxico y corpus enfocados en el lenguaje informal de Argentina.

Palabras clave: Análisis de Sentimientos en Twitter, Minería de Opiniones, Procesamiento del Lenguaje Natural, Léxicos de Sentimientos

CONTEXTO

En el marco de las actividades conjuntas que realizan el Grupo de Estudio en Metodologías de Ingeniería de Software (GEMIS) perteneciente a la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional, y el Grupo de Estudio en Ciencias de Datos (GECS) perteneciente al Laboratorio de Informática Aplicada de la Universidad Nacional de Río Negro, se comienza una nueva línea de trabajo que se articula dentro de los objetivos de ambos grupos en el campo de la Informática, vinculando la Inteligencia Artificial (que está asociada al Procesamiento del Lenguaje Natural y específicamente, al análisis de sentimientos) y la aplicación de sistemas de información (buscando la automatización en el procesamiento de textos subjetivos). En este contexto, se prevé generar nuevo conocimiento en el área de la Ingeniería de Software con la aplicación de tecnologías no convencionales provenientes del Aprendizaje Automático (o Machine Learning en inglés), por lo que sus actividades se desarrollan dentro del ámbito del PID con incentivos UTN UTI5103TC, y UNRN PI 40-C-542.

Por un lado, el Grupo GEMIS busca la sistematización de cuerpos de conocimientos y promoción sobre el campo de la Ingeniería en Sistemas de Información y la Ingeniería en Software, sus aplicaciones y abordajes metodológicos en todo tipo de escenarios (convencionales y no convencionales).

Por otro lado, el Grupo GECS se encuentra conformado por un equipo de docentes y alumnos dentro del ámbito de la Universidad Nacional de Río Negro. Este grupo busca la sistematización de cuerpos de conocimientos y promoción sobre el campo de la Ciencia de

Datos e Inteligencia Artificial, sus aplicaciones y abordajes metodológicos en todo tipo de escenarios (convencionales y no convencionales).

1. INTRODUCCIÓN

1.1. Análisis de sentimientos

Una de las redes sociales más populares para el intercambio de opiniones en torno a una gran variedad de temas es el servicio de *microblogging* Twitter[1]. Actualmente, según cifras oficiales, la plataforma cuenta con cerca de 335 millones de usuarios activos al mes, de los cuales 68 millones de ellos pertenecen a los Estados Unidos y los 267 millones restantes a la comunidad internacional [1]. Recolectar y analizar datos de la plataforma, a diferencia de métodos más tradicionales, es una alternativa que permite encuestar un amplio número de participantes con un menor número de recursos [2]. Es por este motivo que áreas relacionadas a las ciencias políticas, económicas, sociales y de investigación de mercado estudian la red social activamente con un interés especial en las estadísticas agregadas que surgen de los millones de mensajes (o tweets) producidos por sus usuarios todos los días [3]. El conjunto de las opiniones extraídas durante la ejecución de las técnicas sirve para delinear perfiles, conocer sentimientos e ideas de futuros consumidores o votantes, relevar expectativas y realizar predicciones. Esta información resulta de gran valor para empresas, gobiernos y demás organizaciones por sus posibles aplicaciones, las cuales, a menudo, buscan auxiliar en la toma de decisiones estratégicas que se alineen con los objetivos de la organización.

[1] http://twitter.com/

Una de las formas utilizadas para analizar el contenido de los datos de la red social es por medio del análisis de sentimientos, aunque los términos empleados en los trabajos que estudian y exploran el tratamiento computacional de las opiniones, sentimientos y subjetividad en textos incluyen expresiones como minería de opiniones (*opinion mining*), análisis de subjetividad (*subjectivity analysis*) y minería de reseñas (*review mining*), entre otras [4]. En este sentido, los enfoques de las soluciones actuales al problema del análisis de la información subjetiva abarcan una variedad de técnicas pertenecientes a distintas ciencias, campos y subcampos interdisciplinarios como la Lingüística Computacional, el Procesamiento del Lenguaje Natural (PLN), la Inteligencia Artificial (IA) y la ciencia de los datos, o *Data Science*.

El objetivo principal del análisis de sentimientos consiste en la clasificación automática de textos subjetivos en categorías de polaridad previamente determinadas. Normalmente, el contenido puede ser etiquetado como positivo, negativo o neutral [5], aunque estas agrupaciones pueden variar dependiendo de la granularidad que busque el estudio.

Las soluciones encontradas en la bibliografía perteneciente a dicho tema se pueden categorizar, en líneas generales, en tres grupos según las técnicas empleadas para afrontar el desafío del reconocimiento de polaridad en opiniones: enfoque basado en léxico, en aprendizaje automático y el abordaje híbrido. Si bien gran parte de los trabajos pueden ser representados por una de estas categorías, las clasificaciones nombradas no son exhaustivas y en la literatura es posible encontrar soluciones que no se adaptan por completo a ninguno de los tres paradigmas mencionados [6].

1.2. Recursos para el análisis de sentimientos en Twitter

En el análisis de sentimientos en Twitter se tratan de complementar las tareas de clasificación de tweets por medio del uso de distintos recursos.

Muchas de las soluciones hacen uso de léxicos, los cuales se conforman con listas de palabras o frases clasificadas en categorías que comúnmente denotan una polaridad y que han sido calificadas en una escala para describir la intensidad de la misma. Los métodos de elaboración de este tipo de recursos pueden alcanzar distintos grados de automatización, desde lo manual (cuando las etiquetas de polaridad son colocadas por humanos), a lo mecánico (al aplicar motores de traducción automática a léxicos en otros idiomas).

En aquellos trabajos que hacen uso de técnicas de aprendizaje automático, se vuelve indispensable el uso de conjuntos de datos compuestos de tweets recolectados de la plataforma, también llamados corpora, para el entrenamiento y validación de modelos estadísticos. Si la solución lo demanda, estos sets de datos son enriquecidos con anotaciones, normalmente de polaridad, por medio de una variedad de técnicas. Al igual que en el caso de los léxicos, esta pueden involucrar distintos grados de intervención humana.

Por último, existen otras herramientas que ofrecen un amplio espectro de funcionalidades como transformaciones de texto, implementaciones de modelos y algoritmos utilizados para el procesamiento del lenguaje natural y extracción de características comunes para el análisis de sentimientos.

Estos recursos son empleados en las soluciones para el diseño, implementación, entrenamiento y validación de características y clasificadores, por lo que tienen un impacto en los resultados obtenidos.

1.3.Relevancia del problema

Según estimaciones del año 2019, la cantidad de usuarios de Internet que se comunican en español es de alrededor de 344 millones, representando cerca de un 8% de la población total de usuarios, posicionándose en el tercer puesto de los lenguajes más utilizados [7].

La exploración del AST en lenguajes diferentes al inglés es una necesidad que ha sido reconocida por investigadores dentro de la comunidad hispanohablante, así como aquellos que no forman parte de ella [8]. Debido a que la mayoría de los trabajos en el dominio se enfocan en solucionar el problema considerando únicamente un idioma, los recursos creados varían en calidad y cantidad al pasar de un lenguaje a otro. Al contar con el mayor número de herramientas en inglés, la solución de algunos investigadores ha sido, en ocasiones, optar por traducciones automáticas de recursos léxicos para adaptarlos al análisis de tweets en español [9], [10]; sin embargo, como han señalado algunos autores, los resultados pueden variar dependiendo de la calidad de la traducción automática [9]. Es por esta razón que los investigadores con foco en el AST en español han puesto en evidencia en varios estudios [11], [12] la falta de recursos dedicados a este lenguaje.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Luego de la revisión de trabajos presentandos a la edición 2017 de la competencia organizada por la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), el Taller de Análisis Semántico en la SEPLN (TASS), surge del trabajo una herramienta comparativa enfocada en los recursos utilizados para el análisis de sentimientos en Twitter en español [13]. Por medio de este análisis se respalda la necesidad de direccionar esfuerzo hacia la creación de nuevos recursos en español, y específicamente se pone en evidencia la falta de léxicos en los que se incluya lenguaje informal, común en las redes sociales, para el idioma.

3. RESULTADOS Y OBJETIVOS

La presente línea de trabajo posee los siguientes objetivos.

3.1. Objetivo general

El objetivo general es proponer un recurso léxico compuesto por palabras informales focalizado en el español rioplatense de Argentina para su aplicación en el análisis de sentimientos en español en Twitter.

3.2.Objetivos específicos

Asociados al presente objetivo general se definen los siguientes objetivos específicos:

- Relevar los desafíos y el estado del arte del análisis de sentimientos en Twitter en español y los recursos asociados.
- Identificar y documentar las características que formarán parte del recurso léxico.
- Analizar las características identificadas para determinar su aplicabilidad.
- Desarrollar un recurso léxico que haga uso de las características evaluadas.
- Desarrollar un conjunto de textos subjetivos anotado para el entrenamiento y validación de un clasificador utilizado para la validación.
- Aplicación y validación del léxico creado en un proceso de análisis de sentimientos en español.
- Reportar los resultados y conclusiones.

4. FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo se encuentra formado por 2 investigadores formados, 3 investigadores en formación, 5 alumnos avanzados de carreras de grado, 4 estudiantes avanzados de carreras de postgrado, 1 becario CIN. En su marco se desarrollan 2 Tesis de Maestría, 2 Trabajos Finales de Especialidad y 3 trabajos de Fin de Carrera de grado.De esta manera se espera generar un verdadero espacio integrado de investigación en carreras de grado y posgrado.

5. REFERENCIAS

1. Twitter Investor Relations. Twitter Q2 2018 Earnings Report. [Online].; 2018 [cited 2018 Octubre 03. Available from: https://twitter.com/i/moments/1022804623717875712.
2. 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.
3. Rosenthal S, Farra N, Nakov P. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In 11th International Workshop on Semantic Evaluation (SemEval-2017); 2017. p. 502-518.

4. Pang B, Lee L. Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval. 2008; 2(1-2): p. 1-135.
5. Escortell Pérez MA, Giménez Fayos M, Rosso P. El Impacto de las Emociones en el Analisis de la Polaridad en Textos con Lenguaje Figurado en Twitter. Procesamiento del Lenguaje Natural. 2017 Marzo;(58): p. 85-92.
6. Giachanou A, Crestani F. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. ACM Computing Surveys (CSUR). 2016 Noviembre; 49(2).
7. Miniwatts Marketing Group. Top Ten Internet Languages in the World - Internet Statistics. [Online].; 2019 [cited 2020 02. Available from: https://www.internetworldstats.com/stats7.htm.
8. Nakov P. Semantic Sentiment Analysis of Twitter Data. In Encyclopedia on Social Network Analysis and Mining (ESNAM).; 2017.
9. Vilares D, Alonso MA, Gómez-Rodríguez C. Supervised Sentiment Analysis in Multilingual Environments. Information Processing & Management. 2017 Mayo; 53(3): p. 595-607.
10. 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.
11. Jiménez-Zafra SM, Martín-Valdivia MT, Martínez-Cámara E, Ureña-López LA. Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter. IEEE Transactions on Affective Computing. 2017.
12. Sidorov G, Galicia Haro SN, Camacho Vázquez VA. Construcción de un corpus marcado con emociones para el análisis de sentimientos en Twitter en español. Revista Escritos BUAP. 2016; 1(1).
13. 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.