Minería de Opiniones: Análisis de Sentimientos en una Red Social

Alejandra Cardoso [1], Lorena Talame [1], Matias Amor [1], Carlos Neil [1,2]

Grupo de Análisis de Datos /Facultad de Ingeniería e IESIING
[1] Universidad Católica de Salta
Campo Castañares s/n, 4400 Salta, (0387) 426 8536.
{acardoso, mltalame}@ucasal.edu.ar, matiasnicolasamor@gmail.com
[2] Universidad Abierta Interamericana
Carlos.Neil@uai.edu.ar

RESUMEN

El Grupo de Análisis de Datos de la Facultad de Ingeniería y del Instituto de Estudios Interdisciplinares de Ingeniería (IESIING) de la Universidad Católica de Salta, trabaja en proyectos de investigación relacionados a técnicas y aplicaciones de minería de datos y minería de textos. Entre las áreas investigadas se incluyen la categorización de documentos de texto, búsqueda semántica, extracción de entidades con nombre, generación de resúmenes y búsqueda automática de respuestas. Con el proyecto actual se añade un nuevo aspecto en la línea de investigación: el análisis de sentimientos también llamado minería de opiniones. En el tratamiento de opiniones, muchas investigaciones se centraron en el reconocimiento de polaridad en textos. Sin embargo, pocas realizaron una clasificación más exhaustiva y, mucho menos, en lenguaje español. El objetivo del proyecto es clasificar textos cortos (opiniones) en seis sentimientos: miedo, ira, asco, sorpresa, tristeza y felicidad. El proyecto se encuentra en desarrollo. En la primera etapa se recopilaron mensajes de la red social Twitter, muchos de los cuales formarán el corpus de análisis. En la etapa actual se están evaluando diversos enfoques para clasificación y detección de sentimientos en los mismos.

***Palabras clave*:** análisis de sentimientos, minería de textos, Twitter.

CONTEXTO

Este proyecto de investigación continúa la línea de investigación que el Grupo de Análisis de Datos viene desarrollando en minería de textos, financiados por el Consejo de Investigaciones de la Universidad Católica de Salta. El primer trabajo de investigación fue sobre el problema de clasificación o categorización automática de documentos [1]. El segundo versó sobre el reconocimiento de entidades con nombre [2] y la extracción automática de resúmenes de documentos de texto [3]. La plataforma para experimentar con las técnicas descritas fue un buscador semántico en el corpus de más de 8000 documentos conteniendo nueve años de resoluciones rectorales de la Universidad Católica de Salta en distintos formatos. En el tercero, se exploró la búsqueda automática de respuestas a preguntas de usuarios en lenguaje natural en el mismo corpus [4] y en textos extraídos de la web [5]. En esta oportunidad,

el enfoque es el análisis y clasificación de sentimientos de las opiniones de usuarios emitidos en una red social.

1. INTRODUCCION

Debido a la enorme cantidad de documentos disponibles en forma digital y la también creciente necesidad de organizarlos y aprovechar el conocimiento contenido en ellos, en los últimos años, ha aumentado el interés en las técnicas de análisis de información textual. La minería de textos es el proceso de extraer información y conocimiento interesante y no trivial de texto no estructurado. Es un campo relativamente reciente con un gran valor comercial que se nutre de las áreas de recuperación de la información, minería de datos, aprendizaje automático, estadística y procesamiento del lenguaje natural.

La minería de textos incluye una serie de tecnologías: extracción de la información, seguimiento de temas (topic tracking), generación automática de resúmenes de textos (sumarización), categorización, agrupamiento (clustering), vinculación entre conceptos, visualización de la información, respuesta automática de preguntas y análisis de sentimientos. El presente trabajo se centra en la última, también denominada minería de opiniones.

La información puede dividirse en dos grandes grupos, la información objetiva, que representa un hecho, algo comprobable, por ejemplo "*Salta es una provincia del norte argentino*", y la información subjetiva que puede entenderse como la interpretación de una realidad, por ejemplo "*Salta es la provincia más linda del norte argentino*". La minería de opiniones consiste en el análisis de la información subjetiva (opiniones). Un caso particular son las opiniones que generan usuarios de redes sociales, blogs, portales de noticias, foros, etc. que ayudan a revelar información importante sobre un tema específico apoyando, por ejemplo, a campos como la inteligencia de negocios y jugando un papel importante en la toma de decisiones. Una forma de monitorear la opinión de los usuarios sobre determinado producto o tema es proponer encuestas o resaltar comentarios de otros usuarios, por ejemplo otorgando puntaje. Si bien esto puede ser válido, la opinión textual de los usuarios sobre un tema en cuestión se la puede clasificar como positiva o negativa y/o realizar una clasificación más avanzada identificando sentimientos como por ejemplo, tristeza, miedo, sorpresa, alegría, etc. La red social Twitter se ha convertido en una excelente herramienta para conocer en tiempo real las opiniones que los usuarios expresan sobre una gran variedad de temas. Twitter permite identificar estos temas a través de los denominados hashtag o etiquetas, que se caracterizan por comenzar con el carácter # y a continuación una cadena de una o varias palabras concatenadas.

La mayoría de los enfoques actuales se basan principalmente en dos tipos de tareas: detección de la polaridad de la opinión y análisis de sentimientos basado en características.

1) Detección de la polaridad: consiste en determinar si una opinión es positiva o negativa. Algunos trabajos extienden la polaridad básica a un rango de valoración, por ejemplo, en [6] realizaron experimentos con un corpus en español de críticas de cine obtenidas de un sitio web sobre novedades, cartelera y opiniones sobre cine. Las críticas fueron puntuadas en un rango de 1 a 5,

significando el 1 una película muy mala, y el 5 una película muy buena. Las películas puntuadas con 3 se catalogaron como neutras, o que el crítico (usuario de la web) no las consideraba ni malas ni buenas.

Para obtener la polaridad, existen dos métodos más populares: el aprendizaje computacional y el basado en diccionarios léxicos.

- El aprendizaje computacional analiza la información automáticamente de forma supervisada, basándose en conjuntos de entrenamiento que son utilizados para catalogar al resto de las opiniones, realizando pruebas y luego validándolas. Un ejemplo de aplicación de algoritmos de aprendizaje supervisado se encuentra en [7].

- El método de diccionarios léxicos se basa en una lista de palabras con un determinado peso y/o categoría emocional. Estos diccionarios presentan principalmente adjetivos, que son los que aportan mayor información al momento de analizar los sentimientos, aunque también incluye verbos, adverbios y sustantivos. En [8] se utilizó un diccionario de palabras positivas y negativas con puntaje para determinar la polaridad de mensajes de Twitter durante un acto electoral.

2) Análisis del sentimiento basado en características: consiste en determinar las distintas características o entidades del producto tratadas en la opinión escrita por el usuario, y para cada una de esas características mencionadas en la opinión, ser capaces de extraer una polaridad. En [9] se recopilaron opiniones de un sitio gastronómico y se definieron las entidades comida, ambiente y servicio, así de cada comentario se identificó la opinión (positiva, negativa, neutra) sobre cada entidad.

Algunas investigaciones se centraron en analizar opiniones y detectar algún sentimiento en particular. En [10] se clasificó una serie de palabras según seis sentimientos: alegría, enojo, miedo, tristeza, sorpresa y repulsión formando un diccionario de emociones. El trabajo de [11] propone un modelo conceptual para la detección de mensajes violentos o peligrosos. El análisis realizado por [12] se enfocó en diferenciar distintos sentimientos (me encanta, sublime, patético, lamentable, etc.) manifestados por los usuarios.

La propuesta del presente proyecto es llegar un poco más allá de la extracción de polaridad en mensajes de textos, de tal forma de detectar el sentimiento que se expresa en los mismos. Para ello se están analizando las distintas alternativas mencionadas para la clasificación. Así, servirá como puntapié inicial para otro tipo de tareas que posibilite, por ejemplo, eliminar textos agresivos, tan frecuentes en una red social.

2. LINEA DE INVESTIGACION Y DESARROLLO

Este proyecto de investigación propone detectar y clasificar sentimientos expresados en textos, en particular, opiniones textuales emitidas en una red social. El proyecto se desarrolla con las siguientes tareas:

- Revisión de la literatura relevante al problema de minería de opiniones y sentimientos.
- Exploración y evaluación de las diferentes técnicas de recopilación de mensajes
- Recopilación de mensajes y creación de un corpus de opiniones
- Exploración y evaluación de las técnicas de clasificación de textos.

Hasta el momento se logró capturar más de 60000 tweets utilizando la API[1] de Twitter, los cuales fueron recopilados, diariamente durante tres meses, a partir de una serie de hashtags sobre temas de actualidad de nuestro país. Los mensajes se almacenaron en una base de datos NoSQL. Muchos de estos tweets serán descartados por contener solo imágenes, íconos o textos con poca información para el análisis.

Actualmente, el grupo de investigación se encuentra en la etapa de limpieza y preparación de los tweets, y evaluando distintos algoritmos de clasificación de textos.

3. RESULTADOS OBTENIDOS/ESPERADOS

El proyecto tiene como objetivo general, analizar mensajes de textos generados en la red social Twitter e identificar los sentimientos que se expresen en ellos.

Una de las primeras etapas consistió en la captura de tweets, para lo cual se evaluó distintas formas de recopilación.

Los próximos objetivos a alcanzar son:

- Realizar la limpieza y preparación de los tweets descartando aquellos que no posean suficiente información textual para el análisis.
- Evaluar las formas de etiquetado de los mensajes que formarán el conjunto de entrenamiento para los algoritmos.
- Evaluar y comparar algoritmos de clasificación de textos
- Seleccionar los algoritmos que mejor clasifiquen las opiniones

Se espera que esta línea de investigación amplíe los conocimientos sobre las diferentes técnicas de minería de textos y procesamiento de lenguaje natural.

Por otro lado, se espera que este proyecto anime el interés por la investigación y por esta temática a los alumnos de nuestra Facultad.

4. FORMACION DE RECURSOS HUMANOS

El equipo de trabajo está integrado por tres docentes de la carrera de Ingeniería en Informática y dos alumnos de la carrera, una de ellas se encuentra realizando su proyecto de grado en la temática de esta investigación.

REFERENCIAS

[1] A. Pérez Abelleira y A. Cardoso, «Categorización automática de documentos,» de *Simposio Argentino de Inteligencia Artificial, 40 Jornadas Argentinas de Informática (JAIIO)*, Córdoba, 2011.

[2] A. Pérez Abelleira y A. Cardoso, «Técnicas de extracción de entidades con nombre,» de *Simposio Argentino de Inteligencia Artificial, 42 Jornadas Argentinas de Informática (JAIIO)*, Córdoba, 2013.

[3] A. Cardoso y A. Pérez Abelleira, «Generación automática de resúmenes,» de *Congreso Nacional de Ingeniería Informática/Sistemas de Información (ConNaIISI)*, Córdoba, 2013.

[4] A. Cardoso, A. Bini y A. Pérez Abelleira, «Una Arquitectura de un Sistema de Búsqueda de Respuestas,» de *2° Congreso Nacional de Ingeniería Informática/ Sistemas de Información, CoNaIISI*, San Luis, 2014.

[5] A. Cardoso, A. Pérez Abelleira y E. Notario, «Búsqueda de respuestas como aplicación del problema de extracción de relaciones,» de *4º Congreso Nacional de Ingeniería Informática/Sistemas de Información (ConNaIISI)*, Salta, 2016.

[6] E. Cámara, M. Valdivia, J. Ortega y A. Ureña Lopez, «Técnicas de clasificación de opiniones aplicadas a un corpus en español,» *Procesamiento del Lenguaje Natural,* nº 47, pp. 163-170, 2011.

[7] T. Baviera, «Técnicas para el análisis del

[1] https://developer.twitter.com/

sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength,» *Digitos. Revista de comunicación digital,* vol. 1, nº 3, pp. 33-50, 2017.

[8] L. Montesinos García, «Análisis de sentimientos y predicción de eventos en Twitter,» Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago de Chile, 2014.

[9] L. Dubiau, *Procesamiento de Lenguaje Natural en Sistemas de Análisis de Sentimientos,* Buenos Aires: Facultad de Ingenieria, Universidad de Buenos Aires, 2013.

[10] I. Diaz Rangel, G. Sidorov y S. Suarez Guerra, «Creación y evaluación de un diccionario marcado con emociones y ponderado para el español,» *Onomazein. Revista semestral de lingüística, filología y traducción,* pp. 31-46, 2014.

[11] J. C. Calloni, A. Bianciotti, S. Páez, E. Scarello, L. Banchio, M. Mulassano, J. Saldarini, F. Francia, F. Degiovanni, L. Scharff y J. C. Cuevas, «Modelo de análisis de sentimientos con algoritmos de aprendizajes para detectar actitudes peligrosas o violentas de los usuarios en redes sociales,» de *IV Congreso Nacional de Ingeniería Informática y Sistemas de Información (CONAIISI)*, Salta, 2016.

[12] SM Reputation Metrics, «SM Reputation Metrics,» 2015. [En línea]. Available: https://smreputationmetrics.wordpress.com/2015/10/19/el-debate-albertvspablo-analisis-de-sentimiento-twitter/.