MASAA

Modelo de análisis de sentimientos con algoritmos de aprendizajes para detectar actitudes peligrosas o violentas de los usuarios en redes sociales

Autores

Juan Carlos Calloni
jcalloni@gmail.com
Federico Degiovanni
federicoadegiovanni@gmail.com
Sergio Paez
sergio_paez@hotmail.com
Eduardo Scarello
eduardo.scarello@gmail.com
Lucia Scharff
luciascharff@gmail.com
Andrés Bianciotti
andresbianciotti@gmail.com
Leandro Banchio
lbanchio@gmail.com
Micaela Mulassano
mulassano.micaela@gmail.com
Federico Francia
federicomatiasfrancia@gmail.com
Javier Saldarini
saldarinijavier@gmail.com
Juan Carlos Cuevas
juancarloscue@gmail.com

Resumen

Es de amplio conocimiento la utilización masiva de las distintas redes sociales. Éstas han cambiado los hábitos y características de la comunicación, tal como la facilidad de intercambio de información, la existencia de receptores globales y la accesibilidad de éstas tecnologías a todos los sectores sociales. Las actitudes violentas y peligrosas en redes sociales constituyen un campo de estudio objeto de varias disciplinas. Detectar este tipo de actitudes de manera temprana colaboran a la prevención de los efectos que éstas podrían causar.

Este proyecto plantea un modelo conceptual para una herramienta de detección de mensajes de actitudes violentas y peligrosas en redes sociales, mediante algoritmos de inteligencia de artificial, extrayendo información estática para demostrar, con métodos de aprendizaje de análisis de sentimientos y minería de opiniones, qué mensaje se aproximan a ser peligrosos.

Palabras Clave

Redes sociales, análisis de sentimiento, inteligencia artificial, actitudes de usuarios, seguridad.

Introducción

En primera instancia se aborda el fenómeno de las actitudes desde una perspectiva psicosocial, históricamente, el concepto de actitud ha sido y es uno de los temas centrales de la Psicología social. La afirmación de G. W. Allport (1935) según la cual "el concepto de actitud es probablemente el más distintivo e indispensable de la Psicología social[1].

La mayoría de los estudiosos del tema estarían de acuerdo en definir las actitudes como: "Evaluaciones globales y relativamente estables que las personas hacen sobre otras personas, ideas o cosas que, técnicamente, reciben la denominación de objetos de actitud." De una manera más concreta, al hablar de actitudes se hace referencia al grado positivo o negativo con que las personas tienden a juzgar cualquier aspecto de la realidad, convencionalmente denominado objeto de actitud[1].

Las actitudes tienen que ver con los juicios evaluativos que realizan las personas en las dimensiones de bueno-malo, o positivo-negativo. La evaluación de los estímulos del entorno nos permite reconocerlos y saber cómo comportarnos en relación con ellos. Aunque existen diferencias individuales y culturales, todo el mundo tiende a juzgar los estímulos del entorno en dimensiones evaluativas[1].

Hay numerosos procedimientos diseñados para medir actitudes. Estos procedimientos o test se presentan organizados en dos grandes categorías: los procedimientos directos consisten en preguntar directamente y explícitamente a las personas por las opiniones y evaluaciones que sustentan en relación a un determinado objeto de actitud y los procedimientos indirectos, que tratan de conocer las evaluaciones de las personas sobre el objeto de actitud sin preguntar directamente por él[1].

Las redes sociales logran que las personas que las utilizan estén un tiempo, cada vez mayor, conectadas entre sí[2]. La facilidad para expresarse con mensajes y el anonimato que brindan las redes sociales permite a los usuarios expresarse de manera peligrosa o violenta sin tener una percepción real de las consecuencias[3].

Entonces cuando se habla de actitudes peligrosas en redes sociales, podemos mencionar como uno de los dominios de estudios el Bullyng. El bullying (acoso escolar) es un tipo de comportamiento invisible, agresivo y permanente, que se observa cuando interactúan los estudiantes entre sí, donde uno es la víctima y otro u otros los acosadores. Se tiene claro que dicho fenómeno no es monocausal, pero se pueden destacar las más trascendentes como, la influencia de los medios masivos de comunicación y el uso masivo que hacen los niños y

jóvenes de las Tecnologías de la Información y la Comunicación cuando se relacionan sin supervisión ni control y que lleva a la interacción virtual que posibilita la extensión del acoso que quizá inició frente a frente en la escuela[4].

Este comportamiento agresivo que tienen los estudiantes, se observa en el ámbito escolar y, más tarde, se extiende al ámbito virtual, que podemos denominar cyber-bullying (acoso virtual) hasta convertirse en un serio problema que pone en riesgo la calidad del sistema educativo así el bullying ya no se limita al contacto cara a cara en el ámbito escolar, sino que se puede ejercer mediante el uso de tecnologías.

El término cyber-bullying lo empezó a estudiar el psicólogo Peter K. Smith[5]. Acuñó el término en el año 1999 describiendo el fenómeno y en 2006 el uso de la tecnología informática como herramienta para el acoso. El cyber-bullying se ha incrementado a partir del el fácil acceso a las Tecnologías de la Comunicación y la Información, este fenómeno se ha ampliado y se vuelve más agresivo entre los estudiantes[4].

Este proyecto tiene como finalidad brindar un modelo que, al ser utilizado como herramienta, haga un análisis de la información extraída de los mensajes de redes sociales de acceso público. Los métodos de aprendizaje de máquina supervisados y no supervisados permiten clasificar mensajes en función de su contenido y realizar análisis de sentimientos.

La minería de opiniones (MO), también conocida como análisis de sentimientos (AS), es una disciplina que se centra en detectar la información subjetiva de un texto y clasificarla. Existen muchos trabajos centrados en la MO, pero la mayor parte de las investigaciones se han realizado sobre opiniones escritas en inglés.

Sin embargo, cada vez es mayor la presencia de otros idiomas en Internet, entre los que se encuentra el español. En las revisiones del estado del arte del AS se muestran como desafíos el tratamiento de la negación, de la ironía y del sarcasmo, la adaptación al dominio, el análisis a nivel de aspecto y la detección de opiniones spam[6].

Existe una gran variedad de métodos de aprendizaje en dónde podemos distinguir a agentes inteligentes que aprenden y los utilizan. La distinción más importante de los agentes es el elemento de aprendizaje y el elemento de actuación es que el primero está responsabilizado de hacer mejoras y el segundo se responsabiliza de la selección de acciones externas. El elemento de actuación es lo que anteriormente se había considerado como el agente completo: recibe estímulos y determina las acciones a realizar. El elemento de aprendizaje se realimenta con las críticas sobre la actuación del agente y determina cómo se debe modificar el elemento de actuación para proporcionar mejores resultados en el futuro.

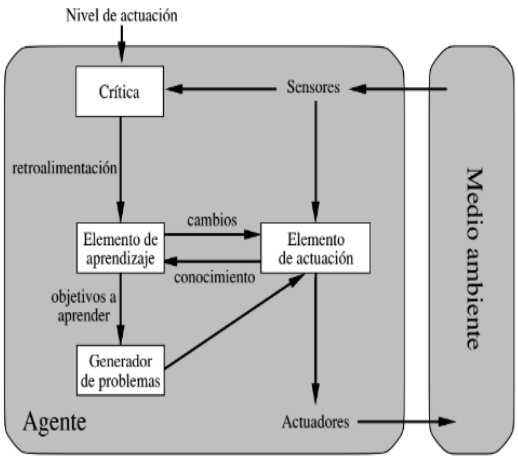

Figura 1: Modelo general para agentes que aprenden. [Russell.]

El algoritmo de aprendizaje, Support Vector Machines[7] (SVM), pertenece a un conjunto de algoritmos de aprendizaje supervisado que están propiamente relacionados con problemas de clasificación y regresión a partir de un conjunto de ejemplos de entrenamiento (de muestras). Es posible etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Por otro lado, K-means[8], es un método de agrupamiento y pertenece a un conjunto de algoritmos de aprendizaje no supervisado que tiene como objetivo la partición de un conjunto de 'n' observaciones en 'k' grupos en el que cada observación pertenece al grupo más cercano a la media. También las redes neuronales artificiales (RNA) son una familia de modelos de aprendizaje estadísticos inspirados en las redes neuronales biológicas y pueden ser utilizadas como algoritmos de aprendizaje supervisados y no supervisados[9].

Existen herramientas no automatizadas que integran los algoritmos mencionados, como Weka[10] que proveen interfaces de conexión y permiten analizar un conjunto de datos. Diversas redes sociales proveen Application Programming Interface (API's) con mecanismos para filtrar mensajes de acuerdo determinados criterios.

La práctica de la minería de datos en redes sociales necesita mucha capacidad de almacenamiento. En la actualidad existen bases de datos NoSQL que permiten brindar soporte a las necesidades de este proyecto. Entre las ventajas destacan[11]:

- Tratamiento de grandes volúmenes de datos
- No utilizan un modelo de datos relacional.
- Buen desempeño en clusters
- Existe un gran número de soluciones de código abierto.
- Son independientes de la definición de un esquema.

Todas estos conceptos y tecnologías permiten realizar un modelo de aplicación para un sistema clasificador de mensajes de redes sociales.

Metodología

Inicialmente fue planificado trabajar con modelos no supervisados, y construir un nuevo modelo basado en técnicas de aprendizaje supervisado, lo cual requiere una extensa tarea de etiquetado y validación pero que, en contrapartida, trae aparejados mejores resultados para el modelo.

En el caso de los modelos no supervisados, se trabaja con el algoritmo K-means. Y una vez etiquetados se transforman en ejemplos de entrenamiento y se aplican los clasificadores basados en regresión logística y RNA.

El Idioma elegido es el español y se utilizan algoritmos de aprendizaje específicos para determinar el modelo que mejor resuelve la clasificación.

Luego de un análisis de requerimientos sobre potenciales demandas fue modelado cada componente del sistema teniendo en cuenta su aplicación sobre situaciones reales. Para ello se realizaron pruebas utilizando contenedores Docker [12] para simular el comportamiento de las capas físicas en una infraestructura actual.

El modelo de componentes obtenido representa el modelo conceptual aplicado en este trabajo:

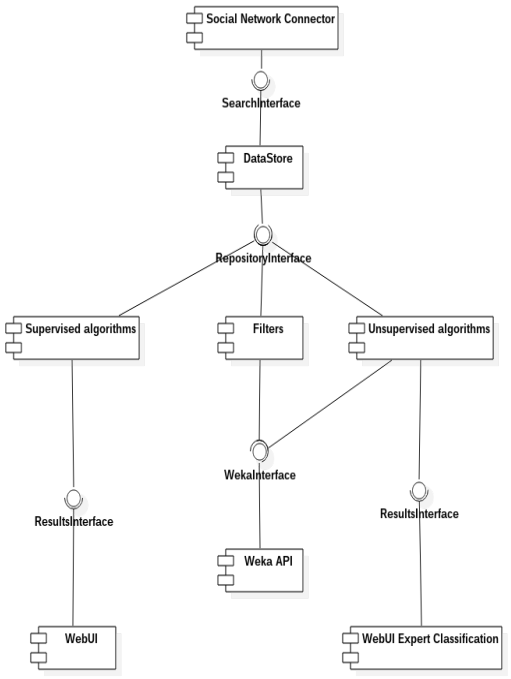

Figura 2: Diagrama de componentes del modelo

Recopilación de datos

Para realizar el análisis sobre los mensajes, primero es necesario contar con un dataset apropiado. Este conjunto es posible obtenerlo mediante la utilización de diversos mecanismos y formas de acceso a los mensajes. Para este proyecto es utilizada la API Rest que provee la red social Twitter. La elección de ésta fuente se basa en la disponibilidad pública de los datos, la cantidad de mensajes obtenidos (quota y rate limit de la API) y capacidad de filtrado.

La representación de cada tweet utilizando la API es una estructura de datos (Javascript Object Notation) JSON [13] que es almacenada en una base de datos NoSQL documental. Para las pruebas es utilizada la base datos MongoDB.

Selección y limpieza de datos

En esta parte el dataset inicial es depurado eliminando aquellos tweets repetidos, citas y respuestas. De esta forma queda sólo el mensaje original. Éste es sometido a una corrección ortográfica automática utilizando el dataset de Google N-Grams[14] montado sobre una instancia dockerizada con MySQL. Por último se aplican técnicas de normalización de palabras utilizando reemplazos mediante el uso de expresiones regulares. El resto de los datos disponibles en la metadata es guardado como atributos del mensaje para un posterior análisis de correlación de variables.

Análisis e interpretación de datos

El análisis de los datos obtenidos se corresponde con un paradigma de aprendizaje y procesamiento automático. Se parte de un conjunto de datos de entrada suficientemente significativo con el objetivo de conseguir que el algoritmo aprenda automáticamente las propiedades deseadas. Cuando el entrenamiento es el adecuado, una vez concluido, el sistema puede recibir mensajes no clasificados y obtener su clasificación con un buen grado de seguridad. Luego, en función del score queda a criterio de los especialistas del dominio establecer la validez del mensaje.

Resultados

El resultado del proyecto es un modelo conceptual aplicable para la creación de una herramienta para el análisis de mensajes de actitudes peligrosas o violentas en la red social Twitter que respondan a los criterios planteados. Es posible la utilización del modelo en soluciones específicas (como la utilizada en este trabajo) o bien para la creación de servicios online de procesamiento y detección bajo demanda mediante el esquema SaaS, PaaS o IaaS[15].

El modelo cuenta con un conector para redes sociales que permite realizar las consultas mediante la interfaz provista por ésta (o alguna desarrollada para tal motivo) y luego almacenar el resultado usando la funcionalidad provista por el subsistema de almacenamiento.

Los mecanismos para acceder a los datos almacenados por los colectores constituyen un repositorio que implementa funcionalidad común para el manejo de los datos. Ésta funcionalidad tiene como ventaja abstraer los métodos de almacenamiento para permitir el uso del sistema de bases de datos que mejor se adapte a las condiciones del dominio.

Cada uno de los algoritmos clasificadores accede a los datos mediante el repositorio. De acuerdo a las

necesidades particulares, puede ser necesario un filtro. Éste, inspirado en el patrón proxy[Gang of four], permite agregar la lógica requerida para que el mensaje reúna las condiciones para ser procesado. Una de los ventajas principales del modelo es la capacidad de intercambiar algoritmos de detección y análisis sin necesitar aplicar retrabajo sobre aquellas operaciones ajenas a la responsabilidad de los mismos.

El proyecto plantea dar conocimientos sobre AI, Machine Learning, NLP, NoSQL Databases, así como también aplicar conceptos de arquitecturas en la nube para una escalabilidad horizontal y un buen desempeño.

Discusión

Si bien en el modelo fue utilizado sólo un conector de redes sociales, es posible añadir tantos como sean necesarios. Esto podría ocasionar un cuello de botella en el subsistema de almacenamiento. Aunque constituye un problema, se puede solucionar con futuros trabajos aplicando técnicas de almacenamiento distribuido.

La idea inicial contempla la posibilidad de aplicación mediante esquemas SaaS, sin embargo queda pendiente analizar en profundidad diversos esquemas de implementación de un potencial producto.

La comunicación actual en redes sociales utilizan distintos recursos para intercambiar información. Si bien el modelo no analiza otros formatos adicionales al texto debido a la alta complejidad computacional que esto representa, éstas implementaciones quedan pendientes para futuras líneas de investigación.

Conclusión

Como resultado de la investigación se obtuvo un modelo con la posibilidad de adaptarse a distintos conectores de redes sociales, a una arquitectura de cloud computing capaz de escalar vertical y horizontalmente y una estrategia de integración para otros algoritmos supervisados y no supervisados que mejoren los resultados sobre el análisis del conjunto de datos.

Se espera que los resultados aporten a la sociedad herramientas para mitigar los efectos de las actitudes peligrosas o violentas en usuarios de redes sociales.

Referencias

[1] Morales, J. and Arias Orduña, A. (2007). Psicología social. Madrid: McGraw-Hill Interamericana de España. p.457, p.459, p.489.

[2] Anderson, Janna; RAINIE, Lee. Millennials will benefit and suffer due to their hyperconnected lives. Washington DC, Pew Research Center, 2012.

[3] Hernández Prados, María Ángeles and Solano Fernández, Isabel María, 2007, Ciberbullying, un problema de acoso escolar. Revista Iberoamericana de Educación a Distancia. 2007.

[4] Cervantes Benavides, L. (2013). Una propuesta para identificar, clasificar y tipificar el Bullying (Acoso Escolar). Revista Iberoamericana para la Investigación y el Desarrollo Educativo.

[5] Smith, P. (1999). The nature of school bullying. London: Routledge.

[6] Zafra, Salud MaJiménez, et al. Desafios del Análisis de Sentimientos. V Jornadas TIMM, p. 15.

[7] Cortes, Corinna; Vapnik, Vladimir. Support-vector networks. Machine learning, 1995, vol. 20, no 3, p. 273-297.

[8] Kanungo, Tapas, et al. An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 2002, vol. 24, no 7, p. 881-892.

[9] Russell, Stuart J and Norvig, Peter, 2004, Artificial intelligence. Englewood Cliffs, N.J. : Prentice Hall.

[10] Weka 3 - Data Mining with Open Source Machine Learning Software in Java, 2016. Cs.waikato.ac.nz [online].

[11] Sadalage, Pramod J.; Fowler, Martin. NoSQL distilled: a brief guide to the emerging world of polyglot persistence. Pearson Education, 2012.

[12] "Docker". Docker. N.p., 2016. Web. 13 Julio 2016.

[13] Rfc-editor.org. N.p., 2016. Web. 13 Julio 2016.

[14] Google Ngram Viewer, 2016. http://storage.googleapis.com/books/ngrams/books/datasetsv2.html [online].

[15] Kavis, Michael J. Architecting The Cloud: Design Decisions For Cloud Computing Service Models. John Wiley & Sons, 2014. Print.