Observatorio Vial Nacional Inteligente (OVNI)

Análisis de las Redes Sociales con Técnicas de Minería de Datos para Aumentar la Seguridad Vial en Venezuela

Juan Vicente Cisneros Arocha
Equipo de investigación socio-tecnológico “Irvin Cuervo”
Colectivo Teletriunfador
Los Teques, Venezuela
juanv.cisneros@gmail.com

Elías Oswaldo Cisneros Arocha
Equipo de investigación socio-tecnológico “Irvin Cuervo”
Colectivo Teletriunfador
Los Teques, Venezuela
cisnerose@cantv.net

***Resumen* — Según la Organización Mundial de la Salud, todos los años, más de 1.2 millones de personas mueren como consecuencia de accidentes de tránsito. En Venezuela los accidentes de tránsito representan la sexta causa de muerte, ante esta situación se crea el Observatorio Vial Nacional Inteligente (OVNI), el cual es un proyecto de investigación desarrollado con tecnologías libres, que pretende brindar un servicio de información vial actualizada, oportuna y transparente, con el objetivo de contribuir en el aumento de la seguridad vial en el país. Con el análisis de los datos viales obtenidos de la red social *Twitter* se podrá, a través de técnicas de procesamiento de lenguaje natural y minería de datos, realizar clasificación de información vial en formato de texto para luego hacer análisis descriptivo y predictivo de los escenarios de mayor riesgo vial en las principales vías del país.**

Palabras claves — seguridad vial; minería de datos; red social Twitter; procesamiento de lenguaje natural; software libre.

I. Introducción

Según el más reciente informe de la Organización Mundial de la Salud (OMS) [1], todos los años, más de 1,2 millones de personas mueren como consecuencia de accidentes en las vías de tránsito y 50 millones sufren traumatismos. Según cifras oficiales, en Venezuela, las muertes por accidentes de tránsito de vehículos automotor representa la sexta causa de muerte en el país [2]. Según [1] se prevé que para el año 2030 las muertes por accidentes de tránsito sean la quinta causa de muerte en todo el mundo, esto debido en parte, al incremento del parque automotor y a falta de políticas públicas en materias de seguridad vial, que aborden el tema desde un punto de vista integral, como lo es la salud, el transporte y la cuerpos de seguridad. Estas cifras son más alarmantes cuando se observa que los traumatismos causados por los accidentes de tránsito, constituyen la principal causa de muerte entre los jóvenes, en edades comprendidas entre los 15 y los 29 años.

De acuerdo a la OMS, Venezuela ocupa el noveno lugar en el mundo con más accidentes de tránsito, alrededor de 7.000 personas fallecen anualmente por este tipo de accidentes. La ocurrencia de accidentes de tránsito se ven considerablemente disminuidos, cuando se implementan leyes que disminuyan los factores de riesgo, tales como: exceso de velocidad, conducción bajo los efectos del alcohol, el uso obligatorio del casco en motorizados, el uso del cinturón de seguridad y de sistemas de retención de niños [1].

Uno de los aspectos a considerar con respecto a la medición de indicadores de seguridad vial, tiene que ver con los datos oficiales de accidentes de tránsito reportados por los cuerpos de seguridad y registrados por los organismos del estado y organizaciones no gubernamentales (ONG), los cuales periódicamente recogen información y generan informes públicos que permiten diseñar políticas públicas en materia vial. Según un estudio publicado por [3], una de las debilidades en materia de seguridad vial en Venezuela, es que se desconocen los datos oficiales de heridos por accidentes de tráfico de vehículos y motos, así como su distribución por edad y sexo. Dado que la publicación de datos oficiales no son realizados frecuentemente, es necesario buscar la forma de obtenerlos a través de otras fuentes, para poder analizarlos con el objeto de hacer propuestas que deriven en políticas públicas que permitan mejorar las condiciones de seguridad vial en el país.

Una de las fuentes de datos que han surgido en los últimos años, con el auge de los sistemas de *microblogging*, ha sido la red social *Twitter*, con más de 350 millones de usuarios activos y un promedio diario de 100 millones de tuits o mensajes, se ha convertido en un flujo de información constante en tiempo real usado con diferentes propósitos, desde el entretenimiento, las comunicaciones de instituciones públicas y privadas hasta la investigación en diversos campos. Para [4] las redes sociales son estructuras sociales que se pueden representar en uno o varios grafos en los cuales los nodos representan individuos y las aristas representan las relaciones entre ellos. Las relaciones pueden ser de distintos tipos: familiares, amistad, intereses, deportes, etc. Las redes sociales en la web, han tomado este concepto general, y han sido informatizadas para crear virtualmente estos espacios de intercambio e interrelaciones. De acuerdo con [5], una red social en la web es un servicio que permite a los individuos construir un perfil público o semipúblico dentro de un sistema delimitado, articular una lista de otros usuarios con los que comparten una conexión, y ver y recorrer su lista de las conexiones realizadas por otros dentro del sistema. En Venezuela la red social *Twitter* es usada para el reporte de tránsito, donde usuarios, organismos del estado y empresas privadas participan para colaborar en la disminución de la congestión vehicular, así como en reportar accidentes de

tránsito para su atención temprana por parte de los cuerpos de seguridad.

El análisis de redes sociales ha intervenido en el desarrollo de muchas ciencias sociales en los últimos años, como una nueva herramienta de análisis de realidad social. Dado el gran potencial que tienen las redes sociales en la web, es posible a través de diversas técnicas informáticas, extraer información relevante y de interés sobre un área en estudio. En función a esta realidad, se plantea a partir del análisis de los datos de información vial obtenidos desde *Twitter*, aplicar técnicas de descubrimiento de conocimiento que permitan extraer patrones e información de interés que aporten en la disminución de accidentes viales.

Para [6] el Descubrimiento de Conocimiento en Bases de Datos o *Knowledge Discovery in Database* (KDD) es un proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos. Entre las metodologías existentes para el KDD, resaltan SEMMA y CRISP-DM, siendo la última la más difundida, al ser independiente de las plataformas y herramientas [7]. CRISP-DM propone un marco de trabajo para el descubrimiento de conocimiento, que incluye las siguientes actividades: conocimiento del negocio, conocimiento de los datos, preparación de los datos, elaboración del modelo, evaluación de los resultados (minería de datos) y despliegue de la información. Es por esto, que la minería de datos en redes sociales, es una oportunidad de investigación en diferentes áreas, ya que según [4] al menos el 50% de los usuarios de Internet usan redes sociales de forma habitual para comunicarse y mantenerse informados.

En función del escenario planteado, se crea el Observatorio Vial Nacional Inteligente (OVNI), el cual es un proyecto socio-tecnológico de investigación, desarrollado con herramientas de software libre. El OVNI en su versión *beta*, extrae datos viales de la red social *Twitter* de la carretera Panamericana ubicada en Venezuela, tramo Distrito Capital - Altos Mirandinos y a su vez informa sobre estos incidentes viales a los suscriptores registrados en OVNI a través del correo electrónico.

Los usuarios de OVNI podrán desde la web, explorar los datos viales obtenidos a través de *Twitter* y realizar diferentes búsquedas personalizadas, así como visualizar estos incidentes en el mapa interactivo. De esta manera es posible realizar análisis sobre la realidad vial por parte de los propios usuarios. En un futuro se espera que OVNI pueda a través de técnicas de minería de datos brindar información vial descriptiva y predictiva de los escenarios de mayor riesgo en las principales vías del país. El proyecto puede ser accedido desde Internet en el sitio web Ciencia con Conciencia a través de la siguiente ruta: http://observatoriovial.cienciaconciencia.org.ve.

A continuación en la figura 1 se muestra la interfaz de usuario del sistema, en la parte superior se encuentra el menú de opciones, así como una sección resumen con estadísticas generales sobre: cantidad de tuits analizados, cantidad de tuits clasificados y cantidad total de usuarios clasificados.

Figura 1. Interfaz de usuario OVNI

En la figura 2 se muestran los filtros de búsqueda por fecha que proporciona OVNI para realizar búsquedas personalizadas.

Figura 2. Filtros de búsqueda de OVNI

En la figura 3 se observa el formato de correo electrónico enviado a los suscriptores de OVNI cuando ocurre un incidente vial detectado en la red social *Twitter*.

Figura 3. Correo electrónico enviado por OVNI

II. Procesamiento de lenguaje natural en redes sociales

Dada las características de las redes sociales de *microblogging* como *Twitter*, es necesario aplicar algoritmos de procesamiento de lenguaje natural a la información que ahí se presenta, con el propósito de poder clasificar cada uno de los tuit. Para [8] la clasificación automática de texto se requiere realizar un análisis léxico, el cual ayudará a extraer características que detallarán a cada tuit, así como, sus clases o categorías. El análisis léxico de los tuits se centró en las siguientes actividades: estudio de la terminología usada en *Twitter* para reportar incidentes viales, estudio de cuentas de *Twitter* oficiales y particulares para el reporte vial, estudio de los resultados obtenidos en la clasificación en tiempo real. Para el análisis de cada tuits, se realiza una limpieza del texto para poder realizar una mejor clasificación de cada tuit, para esto se realizaron las siguientes tareas: eliminación de acentos, conversión a minúsculas y eliminación de caracteres especiales como # , @ e hipervínculos.

En relación a las clases o categorías definidas para clasificar los tuits, se realizó una revisión de los textos almacenados para identificar los incidentes viales que surgían con mas frecuencia en los reportes viales, de esta manera en la tabla 1 se propone la siguiente clasificación de los textos de cada tuit:

Tabla 1. Clasificación de los tuits por incidentes viales

Categorías por incidentes viales	
Palabras de interés	***Clasificación***
Choque	Siniestro vehículo
Colisión	Siniestro vehículo
Accidente	Siniestro vehículo
Incendiado	Siniestro vehículo
Motorizado	Siniestro vehículo
Volcamiento	Siniestro vehículo
Volcado	Siniestro vehículo
Accidentado	Falla vehículo
Herido	Persona lesionada
Lesionado	Persona lesionada
Arrollado	Persona lesionada
Fallecido	Persona fallecida

En la tabla 2, se propone la siguiente clasificación de los tuits en función de la ubicación geográfica que ha sido notificada por los usuarios en cada uno de los reportes viales, de esta manera se pudo identificar las zonas viales de mayor riesgo vial. En la tabla que se presenta a continuación solo se señala una muestra de todos los lugares de interés analizado por OVNI.

Tabla 2. Clasificación de los tuits por zonas viales

Categorías por zonas viales	
Lugares de interés	***Clasificación***
Hipódromo, entrada la vega, Km 0 , Km 1, Km 2, Km 3, Km 4, Km 5.	Kilómetro 0 al 5
IUT, Hotel La Orquídea, Hotel Bosque Dorado, Km 6, Km 7, Km 8, Km 9, Km 10	Kilómetro 6 al 10
San Antonio, IVIC, recta de la minas, Km 11, Km 12, Km 13, Km 14, Km 15	Kilómetro 11 al 15
C.C La Casona, Lomas de Urquia, Km 16, Km 17, Km 18, Km 19	Kilómetro 16 al 19
C.C La Cascada, Montaña Alta, Makro, Km 20, Km 21, Km 22, Km 23, Km 24, Km 25, Km 26	Kilómetro 20 al 26
Club Cumbre Azul, IUTA, Km 27, Km 28, Km 29, Km 30, Km 31, Km 32	Kilómetro 27 al 32

III. Minería de datos en redes sociales

Cada vez son más comunes las investigaciones que muestran el potencial de las redes sociales para el análisis de la realidad social, como por ejemplo, investigaciones en el área de mercadeo, turístico y salud. Recientemente la universidad norteamericana de Northwestern ha desarrollado un sistema para la detección de brotes de gripe [9], a partir del análisis de minería de datos sobre los *tuits* reportados por personas que han consultado *Twitter* para informarse sobre síntomas y enfermedades. A partir de la localización de estas consultas en *Twitter* fue posible construir un mapa y gráficas sobre la propagación de esta enfermedad, incluso detectar zonas de posibles nuevos brotes de la misma.

Según [10], la minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. La minería de datos en redes sociales se basa en extraer información útil y comprensible para un área en estudio, tomando como origen de datos las diversas plataformas de redes sociales en la web. Para el proceso de extracción de los datos viales, se usa el *API* (*Application Programming Interface*) para desarrolladores que proporciona *Twitter* para consultar los datos de la línea de tiempo, tendencias, etiquetas y usuarios. Según la metodología CRISP-DM, se requiere realizar un conjunto de tareas para extraer y publicar la información de interés, en el caso de la información vial se procedió en el siguiente orden:

- Estudiar las zonas viales y los datos reportados por los usuarios en *Twitter*, para ello se realizó un seguimiento diario de la etiqueta #PNM para extraer la información vial de la carretera Panamericana. De esta manera fue posible identificar un modelo sencillo sobre las palabras de interés más usadas y horarios de reporte de los usuarios.
- Diseñar una rutina automática que extrajera de Twitter a través de su API, alrededor de 300 *tph (*tuit por hora*)* donde se indiquen palabras relacionadas a

incidentes viales tales como: colisión vehicular, arrollamientos, volcamientos, heridos y fallecidos por accidentes de tránsito.

- Diseñar una rutina automática de clasificación de tuits en función de su contenido, extrayendo información tal como: tipo de incidente, usuario que reporta el incidente, ubicación geográfica del incidente, entre otros.
- Diseñar una rutina automática de georeferenciación de cada tuit, a partir de la ubicación reportada por los usuarios, para que luego sea publicada en el servicio de mapas *OpenLayers*, el cual se encuentra integrado a los servicios de OVNI.
- Diseñar una rutina automática de notificación a usuarios a partir de la detección de un nuevo incidente vial detectado por OVNI. Actualmente las notificaciones son realizadas a través de un correo electrónico a los usuarios suscritos al servicio

Luego de la clasificación de cada *tuit*, se requiere el análisis de los datos obtenidos a través del uso de técnicas de minería de datos, para con ello, poder describir o predecir según sea el caso, la información de seguridad vial reportada por los usuarios. Dentro de los modelos de minería de datos existen diferentes tipos de técnicas, no existe una técnica universal que pueda ser aplicada para la resolución de cualquier tipo de problema [7]. Cada técnica presenta sus ventajas e inconvenientes y la elección de la misma dependerá del objetivo perseguido por el analista; siendo, según [11], las más utilizadas en el campo de la seguridad vial las Redes Neuronales Artificiales, las Redes Bayesianas, las Reglas de Asociación y los Árboles de Decisión .

Particularmente, los árboles de decisión son una técnica de minería de datos muy apropiada para el estudio de los accidentes de tránsito, ya que constituyen uno de los modelos más utilizados en aprendizaje supervisado y en aplicaciones de minería de datos [12]. Según [11], entre las ventajas de los árboles de decisión de cara a su utilización para el estudio de los accidentes se puede destacar que permiten la extracción de reglas de decisión del tipo “SI- ENTONCES”. Estas reglas son fácilmente comprensibles para las autoridades de seguridad vial y pueden ser usadas para descubrir determinados patrones de comportamiento que ocurren dentro de un conjunto de datos. Estos patrones pueden ayudar a la comprensión del suceso de un accidente, a la identificación de las principales variables que determinan su gravedad, así como al establecimiento de actuaciones concretas por parte de los organismos competentes con el fin de mejorar la seguridad vial de las carreteras analizadas.

IV. Conclusiones parciales

A través de la creación del Observatorio Vial Nacional Inteligente (OVNI), se pretende brindar a los ciudadanos y al poder público, un servicio socio-tecnológico que suministre información actualizada, oportuna y transparente que pueda incidir en la disminución de accidentes de tránsito, brindando información descriptiva y predictiva de los escenarios de mayor riesgo en las principales vías del país.

Desde el mes enero hasta septiembre de 2015 se han analizado alrededor de 100.000 *tuits* con la etiqueta de *Twitter* #PNM, logrando clasificar alrededor de 5.000 tuits de interés e identificando alrededor de 1.500 usuarios. En la figura 4 se detallan las categoría de incidentes viales obtenidos a través de los reportes viales de los usuarios y analizados por OVNI.

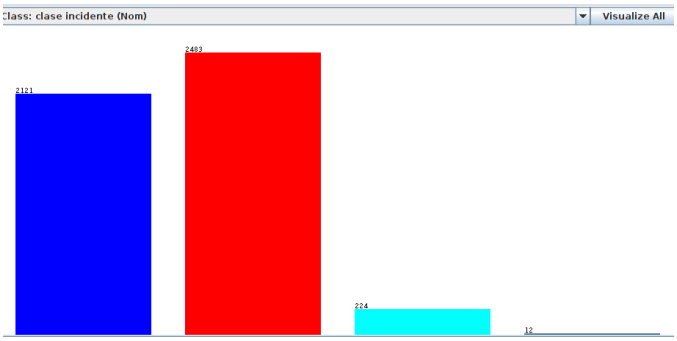

Figura 4. Cantidad de incidentes viales por categorías

En la tabla 3 se detallan los datos de la figura 4.

Tabla 3. Cantidad de incidentes viales por categorías

Cantidad de incidentes viales por categorías	
Clasificación	***Cantidad***
Siniestro vehículo (rojo)	2.483
Falla vehículo (azul oscuro)	2.121
Persona lesionada (azul claro)	224
Persona fallecida (gris)	12

En la figura 5 se detallan la cantidad de incidentes viales discriminados por zonas viales, obtenidos a través de los reportes viales de los usuarios y analizados por OVNI.

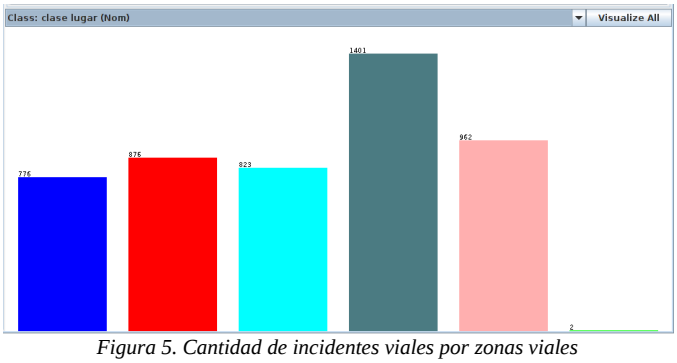

Figura 5. Cantidad de incidentes viales por zonas viales

En la tabla 4 se detallan los datos de la figura 5.

Tabla 4. Cantidad de incidentes viales por por zonas viales

Cantidad de incidentes viales por zonas viales	
Clasificación	***Cantidad***
Kilómetro 0 al 5 (verde oscuro)	1.401
Kilómetro 6 al 10 (rosado)	962
Kilómetro 11 al 15 (rojo)	876
Kilómetro 16 al 19 (azul oscuro)	776
Kilómetro 20 al 26 (azul claro)	823
Kilómetro 27 al 32 (verde claro)	2

A continuación se presentan datos de interés, que fueron obtenidos del análisis de los tuits registrados por OVNI.

- El incidente vial mas reportado es de la categoría “Siniestro Vehículo”.
- La zona vial donde ocurren mas incidentes es la categoría “Kilómetro 0 al 5”.

V. Próximos pasos

Entre las próximas acciones a realizar en el desarrollo del Observatorio Vial Nacional Inteligente (OVNI) destacan:

- Incorporar técnicas de inteligencia artificial, como sistemas multiagentes al proceso de extracción, clasificación y notificación de incidentes viales.
- Incorporar técnicas procesamiento de lenguaje natural que permita la clasificación no supervisaba de los tuits.
- Aplicar algoritmos de minería de datos, que permitan realizar descripciones y predicciones de incidentes viales.

Referencias

[1] Organización Mundial de la Salud (OMS), “Informe sobre la situación mundial de la seguridad vial ”. Año 2013.

[2] Ministerio del Poder Popular para la Salud, República Bolivariana de Venezuela , “Anuario de mortalidad”. Año 2012

[3] Asociación Civil Observatorio Vial. “II Informe sobre la situación de seguridad vial en Venezuela”. Año 2013.

[4] P. García, C. Azaustre (2012). “Minería de Datos aplicada a las Redes Sociales”.

[5] N. B. Ellison and D. Boyd. (2013). Sociality through Social Network Sites. In Dutton, W. H. (Ed.), The Oxford Handbook of Internet Studies. Oxford: Oxford University Press, pp. 151-172.

[6] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. “Knowledge discovery and data mining: towards a unifying framework”. In Simoudis, E., Han, J., and Fayyad, U., editors, *Proceedings of KDD'96, Second International Conference on Knowledge Discovery & Data Mining*, pages 82-88. AAAI Press, Menlo Park, CA. 1996.

[7] J. H. Orallo, J. Ramirez, C. Ferri. “Introducción a la minería de datos”. Pearson – Prentice Hall. Madrid, Año 2004.

[8] A. P. Bográn, J. L. Alonso Berrocal y L. C. García de Figuerola Paniagua. (2013) . “Análisis Léxico sobre los Tweets de Twitter”.

[9] N. Roales, “Detección de tendencias en Twitter utilizando minería de datos adaptativa”. Tesis Universidad de Granada. Año 2014.

[10] I. H. Witten, E. Frank, and M. A. Hall. (2011). “Data Mining: Practical Machine Learning Tools and Techniques”. Morgan Kaufmann, Burlington, MA, 3 edition.

[11] G. López. “Análisis de la severidad de los accidentes de tráfico utilizando técnicas de minería de datos”. Tesis PhD, Universidad de Granada, Año 2013.

[12] J. Gehrke, V. Ganti, R. Ramakrishnan, and W.-Y. Loh. “BOAT-Optimistic Decision Tree Construction”. SIGMOD Conference, pp. 169-180. Año 1999