



Modelamiento analítico de un sistema que permita
la anticipación de peligros geoestratégicos para
Colombia, a través del análisis de patrones y
tendencias en Redes Sociales

Juan Andrés Suárez Acuña

Trabajo de grado para optar al título profesional:
Maestría en Ciberseguridad y Ciberdefensa

Escuela Superior de Guerra "General Rafael Reyes Prieto"
Bogotá D.C., Colombia

2018

711CIBER 2018

011

G.2

**MINISTERIO DE DEFENSA NACIONAL
COMANDO GENERAL FUERZAS MILITARES
ESCUELA SUPERIOR DE GUERRA**



Modelamiento analítico de un sistema que permita la anticipación de peligros geoestratégicos para Colombia, a través del análisis de patrones y tendencias en Redes Sociales.

106282

**Teniente de Fragata
JUAN ANDRES SUAREZ ACUÑA
COMANDO CONJUNTO CIBERNÉTICO**

**Maestría en Ciberseguridad y Ciberdefensa
Trabajo de grado
Bogotá – Colombia
2018**

**MINISTERIO DE DEFENSA NACIONAL
COMANDO GENERAL FUERZAS MILITARES
ESCUELA SUPERIOR DE GUERRA**



Modelamiento analítico de un sistema que permita la anticipación de peligros geoestratégicos para Colombia, a través del análisis de patrones y tendencias en Redes Sociales.

**Teniente de Fragata
JUAN ANDRES SUAREZ ACUÑA**

DIRECTOR

MsC SANDRA MILENA SUAREZ

MAESTRIA EN CIBERSEGURIDAD Y CIBERDEFENSA

**TRABAJO DE GRADO PARA OPTAR EL TITULO DE MAGISTER EN
CIBERSEGURIDAD Y CIBERDEFENSA**

BOGOTA – COLOMBIA

2018

Abstract

Resumen Ejecutivo

The objective of this research is to develop an analytical modeling that allows the

El objetivo del presente trabajo de investigación consiste en desarrollar un modelamiento analítico que permita la anticipación de situaciones anómalas que comprometan los intereses nacionales, la seguridad y defensa de Colombia, por medio del análisis de patrones y tendencias en el campo de la geopolítica y geoestrategia. Este modelamiento contará con el apoyo de metodologías ya reconocidas como lo es SEMMA, la cual permitirá escoger las herramientas que se acoplen a la necesidad de analizar los datos presentes de la extracción de conocimiento no implícito, derivado del análisis de los patrones y tendencias que a simple vista son imposibles de detectar, cuyo resultado servirá para la toma de decisiones estratégicas.

Keywords: Cyber intelligence, Geostrategy, Data Mining, Patterns and Trends.

Palabras claves: Ciberinteligencia, Geoestrategia, Minería de datos, patrones y tendencias.

Abstract

The objective of this research is to develop an analytical modeling that allows the anticipation of anomalous situations related to the national interests, security and defense of Colombia, through the analysis of patterns and trends in the field of geopolitics and geostrategy. This modeling will be supported on well-known methodologies like SEMMA, which allows choosing the tools that fit in the need for analyzing the data obtained from non-implicit knowledge extractions, such as the analysis of patterns and trends that are impossible to detect by human intuition, resulting useful this way, for strategic decision-making.

Keywords: Cyber intelligence, Geostrategic, Data Mining, Patterns and Trends.

Lista de abreviaturas

DATA MINING	:	Minería de Datos
SEMMA	:	Acrónimo (Sample, Explore, Modify, Model, Assess).
KDD	:	Knowledge Discovery in databases.
ICCN	:	Infraestructura Critica Cibernética Nacional
URL	:	Localizador Uniforme de Recursos.
ELN	:	Ejército de Liberación Nacional.
FARC	:	Fuerzas Armadas Revolucionarias de Colombia.
GAO	:	Grupos Armado Organizados.
BACRIM	:	Bandas Criminales.
SSM	:	Social media marketing
S/N	:	Relación Señal a Ruido
UAV	:	Unmanned Aerial Vehicle- Vehículo aéreo no tripulado.
VPN	:	Virtual Private Network. Red Privada Virtual.

	Contenido	44
LISTA DE ABREVIATURAS		V
1. METODOLOGÍA		11
2. INTRODUCCIÓN		14
3. DESARROLLO		16
3.1. CAPITULO I. MARCO TEÓRICO		16
3.1.1 <i>Adquisición de Información a través de Minería de Datos</i>		16
3.1.2 <i>Herramientas de Código Abierto para el procesamiento y visualización de Minería de Datos en Fuentes Abiertas</i>		18
3.1.3 <i>Metodologías de Análisis de captura de datos</i>		20
3.1.4 <i>Métodos de autenticación de las API</i>		21
3.1.5 <i>Abstracción de datos [Minería de datos]-[MUESTREO]</i>		24
3.1.6 <i>Exploración de la Información</i>		29
3.1.7 <i>Manipulación y Modelado de los datos</i>		29
3.1.8 <i>Algoritmos de Procesamiento Mediante el Agrupamiento</i>		33
3.1.9 <i>Valoración de la Información</i>		37
3.2. CAPÍTULO II. EFECTIVIDAD DE ABSTRACCIÓN DE DATOS (MINERÍA DE DATOS) EN LA RED SOCIAL TWITTER		37
3.2.1 <i>Características de Shiny para R</i>		38
3.2.2 <i>Instalación</i>		39
3.2.3 <i>Acceso a la API de Twitter</i>		39
3.2.4 <i>Preparación del código para realizar la aplicación</i>		42
3.2.5 <i>Twitter: Twitter basado en R</i>		42

3.2.6. *Librería para la creación de mapas*..... 44

3.3. CAPITULO III. ESTABLECIENDO LOS CRITERIOS DE SELECTIVIDAD Y BÚSQUEDA DE INFORMACIÓN EN EL CIBERESPACIO PARA PARAMETRIZAR UN PELIGRO GEOESTRATÉGICO PARA COLOMBIA..... 48

3.3.1. *Conflictos armados*..... 50

3.3.2. *Terrorismo*..... 52

3.3.3. *Crimen Organizado*..... 53

3.3.4. *Espionaje y Ciberespionaje*..... 55

3.3.5. *Amenazas en el Ciberespacio*..... 56

3.3.6. *Soberanía Territorial*..... 57

3.4. CAPITULO IV. ELABORAR LA METODOLOGÍA DE ANÁLISIS DE LOS DATOS RECOPIADOS PARA EL PROCESAMIENTO DE RESULTADOS, QUE PERMITAN GENERAR UN COMPORTAMIENTO TENDENCIAL PARA LLEVAR A CABO UNA ALERTA PARA LA TOMA DE DECISIONES..... 69

3.4.1. *Muestreo (Sample) y Exploración (Explore)*..... 70

3.4.2. *Manipulación (Modify)*..... 73

3.4.3. *Modelado (Model)*..... 77

3.4.4. *Valoración (Assess)*..... 80

4. CONCLUSIONES..... **84**

5. REFERENCIAS BIBLIOGRÁFICAS..... **87**

<i>Figura 18. Vulnerabilidades del eSpace para el</i>	57
<i>Figura 19. Sectores Estratégicos</i>	59
Lista de Figuras	
<i>Figura 1. Metodología SEMMA</i>	12
<i>Figura 2. Diagrama de Flujo Metodología SEMMA</i>	13
<i>Figura 3. Creación de la aplicación del API de Twitter</i>	25
<i>Figura 4. Métodos de Autenticación</i>	26
<i>Figura 5. Atributos de una consulta de Twitter</i>	28
<i>Figura 6. La minería de datos como un paso en el proceso de descubrimiento de conocimiento</i>	31
<i>Figura 7. La minería de gráficos en Maltego</i>	32
<i>Figura 8. Un gráfico en 2-D de los datos de la población de Colombia con respecto a las ubicaciones de los habitantes en una ciudad, visualizando tres grupos de datos diferentes</i>	35
<i>Figura 9. Algoritmo de agrupamiento K-means</i>	36
<i>Figura 10. Los límites para los recursos de la API de Twitter se identifican en el documento en línea para cada petición de la API; el recurso API que se visualiza permite 15 solicitudes por "Ventana de límite de velocidad", que actualmente se define en 15 minutos</i>	42
<i>Figura 11. Localización del tweet con el hashtag #Bogotá</i>	45
<i>Figura 12. Línea de código Apariencia visual #Bogotá</i>	46
<i>Figura 13. Línea de código Adquisición de Tweets en el mapa hashtag #Bogotá</i>	46
<i>Figura 14. Espacios comunes globales</i>	49
<i>Figura 15. Amenazas para Seguridad Nacional</i>	50
<i>Figura 16. Ataques Terroristas en Colombia</i>	53
<i>Figura 17. Visualización de los lineamientos de la D.P 0015 del Ministerio de Defensa Nacional</i>	54

Figura 18. Vulnerabilidades del ciberespacio 57

Figura 19. Sectores Estratégicos ICCN...... 59

Figura 20. Diagrama de Flujo Metodología para el modelo analítico de anticipación de peligros geoestratégicos 71

Figura 21. Metadatos de un tweet...... 73

Figura 22. Conversión de mayúsculas a minúsculas 74

Figura 23. Código remoción de la URL 74

Figura 24. Código remoción de espacios. 75

Tabla 7. Diccionario de palabras para las asociaciones semánticas 77

Tabla 8. Asociación de palabras homófonas 77

Tabla 9. Código y Ecuación del algoritmo de Levenshtein 78

Tabla 10. Resultado final Clustering con el algoritmo de Levenshtein 80

Tabla 11. Pruebas de variación de Cluster. Nota: Elaboración propia 81

Tabla 12. Frecuencia repetición de las palabras clave 82

Lista de Tablas

Tabla 1. Limite de páginas por API de Twiiter..... 40

Tabla 2. Persistencia de datos de las peticiones de la API de Twitter. 41

Tabla 3. Mensaje de error..... 41

Tabla 4. Criterios de selectividad de peligros Geoestrategicos 60

Tabla 5. Resultado conversión de mayúsculas 74

Tabla 6. Resultado de la eliminación de los espacion..... 75

Tabla 7. Diccionario de pablaras para la corrección idiomática..... 75

Tabla 8. Remoción de palabras innecesarias 77

Tabla 9. Código y Ejecución del algoritmo de k-means 78

Tabla 10. Resultado final clustering con el algoritmo de k-means..... 80

Tabla 11. Pruebas de variación de Clúster. Nota: Elaboración propia..... 81

Tabla 12. Frecuencia repetición de las palabras clave 82

1. Metodología

La metodología de investigación de este trabajo se basa en el método analítico bajo el enfoque teórico de Jaqueline Hurtado de Barrera descrito en sus libros “Metodología de la investigación” (Barrera, 2010), mediante el desglose de cada uno de los elementos necesarios para orquestar las técnicas de investigación enfocadas a establecer los criterios de selectividad y búsqueda de información en el ciberespacio. Por medio de la minería de datos en redes sociales, se analizan patrones y tendencias de los datos abstraídos, que faciliten el proceso de toma de decisiones estratégicas. Estos procesos en la actualidad existen con gran cantidad de datos almacenados en el ciberespacio, usufructo del constante desarrollo tecnológico de las computadoras y la velocidad con que estas pueden procesar los datos, lo que ha logrado una nueva ventana de posibilidades hacia el procesamiento y análisis de resultados, asociadas bajo el término conocido como “minería de datos”.

La metodología sustentada en el proceso analítico - particularmente en el proceso de minería de datos para extraer información útil de redes sociales a partir de grandes cantidades de datos, requieren de la aplicación de una metodología estructurada para la utilización ordenada y eficiente de las técnicas y herramientas disponibles. El estado del arte y trabajos de investigación como la “Adaptación de Estándares de Dirección de Proyectos Particularizados para la minería de datos” (Sánchez, 2016) ha vislumbrado que una de las metodologías utilizadas por los analistas para la realización de proyectos de minería de datos, es SEMMA.

La metodología SEMMA (Sample, Explore, Modify, Model, Assess) (Montequín et al., 2015), se define como el proceso de selección, exploración, manipulación, modelamiento y valoración de grandes cantidades de datos, para descubrir patrones de negocio desconocidos. El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso (Figura 1).

El proceso se inicia con el muestreo y selección de los datos de una red social, para nuestro caso particular, se elige la Red Social Twitter. Una vez determinada la muestra, se procede con la fase de exploración con el fin de simplificar el problema; para llevar a cabo esta fase, se

plantea la utilización de herramientas de visualización o de técnicas estadísticas que faciliten las relaciones entre las variables y de esta forma se determinen cuales participarán como inputs al modelo.

La tercera fase, radica en la manipulación de los datos de acuerdo a la exploración realizada anteriormente, para que se definan y parametricen las variables que serán introducidas en el modelo.

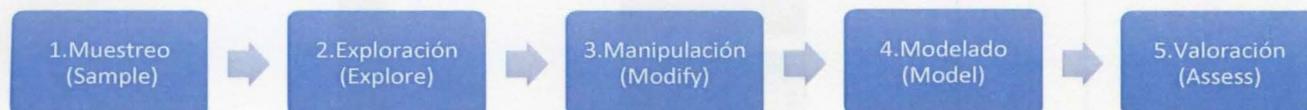


Figura 1. Metodología SEMMA.

Nota: Elaboración propia

Una vez culminada la fase de manipulación, se procede al análisis y modelado de los datos, las técnicas utilizadas para el modelado incluyen métodos estadísticos tradicionales, tales como los métodos de agrupamiento, o análisis de regresión, así como procesos basados en datos como las redes neuronales, técnicas adaptativas, árboles de decisión o reglas de asociación.

En último lugar, la fase de valoración de los resultados busca evaluar los resultados obtenidos producto del análisis y modelado de datos, comprobando que la información obtenida sea congruente con los criterios de búsqueda establecidos por el usuario y en caso de requerir una nueva evaluación de los resultados, adaptar mecanismos para corregir dichas incongruencias. (Figura 2. Diagrama de Flujo Metodología SEMMA).

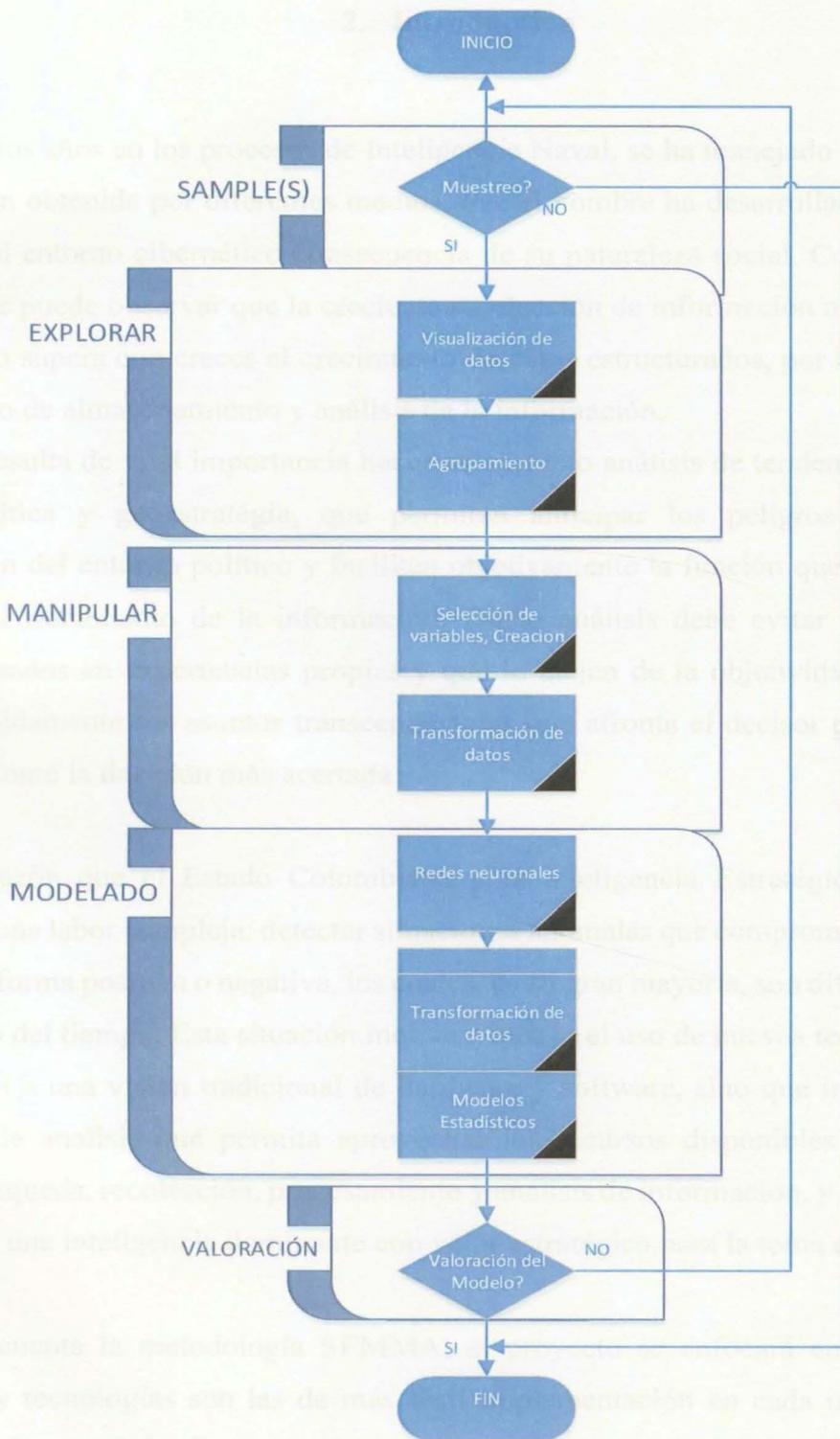


Figura 2. Diagrama de Flujo Metodología SEMMA.

Nota: Elaboración propia

2. Introducción

A lo largo de los años en los procesos de Inteligencia Naval, se ha manejado un gran volumen de información obtenida por diferentes medios, que el hombre ha desarrollado a través de su inclusión en el entorno cibernético consecuencia de su naturaleza social. Como resultado de este proceso se puede observar que la creciente recolección de información no estructurada en el ciberespacio supera con creces el crecimiento de datos estructurados, por lo que plantea un reto en término de almacenamiento y análisis de la información.

Es así como resulta de vital importancia hacer un correcto análisis de tendencias en el campo de la geopolítica y geoestrategia, que permitan anticipar los peligros que acarrea la desinformación del entorno político y faciliten objetivamente la función que desempeñan los analistas en el tratamiento de la información. Dicho análisis debe evitar caer en criterios personales basados en experiencias propias y que lo alejen de la objetividad, permitiéndole identificar rápidamente los asuntos trascendentales que afronta el decisor político nacional, para que este tome la decisión más acertada

Es por esta razón que el Estado Colombiano y la Inteligencia Estratégica Naval se ven enfrentados a una labor compleja: detectar situaciones anómalas que comprometan los intereses nacionales en forma positiva o negativa, los cuales, en su gran mayoría, son difíciles de detectar en el transcurso del tiempo. Esta situación motiva a buscar el uso de nuevas tecnologías que no solo se limiten a una visión tradicional de hardware y software, sino que incluyan toda una metodología de analisis que permita aprovechar los recursos disponibles para ejercer un proceso de búsqueda, recolección, procesamiento y análisis de información, y así obtener como producto final una inteligencia dominante con valor estratégico para la toma de decisiones.

Teniendo en cuenta la metodología SEMMA, el proyecto se enfocará en vislumbrar que herramientas y tecnologías son las de más fácil implementación en cada una de las etapas propuestas por la metodología y permitirán realizar un análisis más adecuado a la información recopilada por las redes sociales o fuentes abiertas.

De acuerdo a lo anterior, este proyecto verificará el estado del arte que ha tenido la minería de datos en las redes sociales, y en aquellas técnicas y métodos que permitan el análisis de dichas

fuentes, con el fin de identificar los patrones, tendencias y relaciones que no son evidentes a simple vista y que coadyuvan a la toma de decisiones.

Procesos como el KDD (knowledge Discovery in databases) han servido de apoyo a metodologías como SEMMA (Sánchez, 2016), ya que estos se vuelven una herramienta útil para el análisis de datos y extracción de conocimiento no implícito, producto de diferentes procesos inherentes a una actividad en el ciberespacio, que al ser procesada correctamente a partir de grandes volúmenes de datos se extrae conocimiento útil que ayuda a la toma de decisiones en los ámbitos en donde fueron extraídos, por esta razón se comenzará a ahondar en esta dinámica cognitiva, para la cual encontramos varias fuentes de información sobre su base teórica, metodología de implementación y recomendaciones para su éxito.

De la situación planteada, se desprende el siguiente interrogante: ¿Cómo analizar los patrones y tendencias de los trinos de Twitter, con el fin de aportar una variable significativa para anticipar un peligro geoestratégico para Colombia?

3. Desarrollo

3.1. Capítulo I. Marco Teórico

Este capítulo menciona los conceptos y la documentación relevante de la bibliografía disponible electrónicamente, como insumo para la cimentación del presente trabajo de investigación, la cual permite tener un punto partida para posteriores necesidades de investigación que se requieran en el ámbito de la minería de datos en redes sociales a través de la Metodología de investigación “SEMMA”.

3.1.1 Adquisición de Información a través de Minería de Datos

La revolución digital ha hecho posible que la información digitalizada sea fácil de capturar, procesar, almacenar, distribuir, y transmitir (Achayra, 2003). Con el constante progreso en informática y las tecnologías relacionadas a estos procesos, la minería de datos cumple un papel fundamental en diferentes aspectos de la vida como en el campo empresarial, permitiendo integrar y almacenar en bases de datos gran cantidad de información que realmente es relevante para la toma de decisiones estratégicas, que con el debido análisis de los datos podría descubrir patrones y tendencias que a simple vista son imposibles de detectar. (Riquelme, Ruiz, & Gilbert, 2006).

Gracias a estos avances en el tema computacional y a la necesidad de mejorar la rentabilidad de mercadeo en el ciberespacio, todas las empresas han contribuido a que cada vez sea más creciente el interés en la minería de datos¹ en el sector público (mejora de programas, acceso a información privilegiada, etc.) así como también en el sector privado (banca, seguros, ventas, entre otros). Esto se puede observar en una sociedad hiperconectada, anidado al crecimiento de las redes móviles, las cuales son utilizadas para conectarse a bases de datos de donde constantemente se quiere tener acceso.

Por esta razón, la minería de datos en el ciberespacio se acopla rápidamente a las necesidades del gremio y es una herramienta vital para comprenderla y darle un valor estratégico a los datos

¹ La Minería de Datos es el proceso de descubrir patrones interesantes y novedosos, así como modelos descriptivos, comprensibles y predictivos de datos a gran escala. (Zaki, Meria, & Meira, 2013)

que obtenemos de ella (Chakrabarti, 2002). En la actualidad y de acuerdo a la búsqueda realizada se distinguen tres tipos de técnicas de minería Web (Ricardo Baeza-Yates, 2009):

- A. Minería de contenido: texto, imágenes, etiquetas (tags), metadatos, etc.
- B. Minería de estructura: enlaces y sus relaciones;
- C. Minería de uso: interacción de las personas con la Web.

De acuerdo al autor Ricardo Baeza, los dos primeros tipos, se obtienen recolectando todo el contenido de los sitios Web usando un software especial llamado recolector (crawler). Un recolector se inicializa con un conjunto de sitios iniciales y continua con todos los enlaces que encuentra en las páginas según un cierto conjunto de reglas predeterminadas (por ejemplo, qué dominios o tipos de ficheros recorrer). Los datos de uso provienen de los registros (logs) de los servidores web y de aplicaciones específicas, como de las de búsqueda (Ricardo Baeza-Yates, 2009).

Y la tercera técnica, la que más popularidad ha obtenido últimamente es la que han denominado los grandes programadores “aprendizaje automático”. La cual consiste en aprender como predecir variables en función de otras variables a través de subconjuntos de datos completos, para posteriormente evaluar cuán buena es la predicción en otro subconjunto de datos. El algoritmo resultante se usa en los datos reales con la suposición de que su desempeño será similar. Este proceso se repite en el tiempo para ir mejorando la herramienta con casos difíciles. Para esto se pueden utilizar árboles de decisión, máquinas de soporte vectorial o redes neuronales, entre otros. (Ricardo Baeza-Yates, 2009).

Es así como el Ingeniero Cristóbal Romero Morales, en el marco de la minería de datos, visualizó varias técnicas de minado, por el momento las más utilizadas aplicada a los sistemas de e-Learning o de aprendizaje automático son: clasificación y agrupamiento, descubrimiento de reglas de asociación, y análisis de secuencias. A continuación, se describirán los principales trabajos de investigación agrupados dentro de estos tres tipos de técnicas, aunque algunos de los investigadores no sólo utilizan una única técnica sino varias. (Morales, Soto, & Hervás Martínez, 2005).

A. Clasificación y Agrupación:

Conocidos bajo el nombre de Clustering² esta técnica consiste en ordenar o dividir fenómenos complejos (descritos por conjuntos de objetos agrupados por datos altamente dimensionales) en pequeños y comprensibles unidades o clases que permiten un mejor control o comprensión de la información. (Morales, Soto, & Hervás Martínez, 2005).

B. Reglas de Asociación

Las reglas de asociación (Agrawal, Imielinski, & Swami, 1993) descubren las relaciones entre los atributos de un conjunto de datos que superan unos determinados umbrales. Su aplicación más típica ha sido en los sistemas de comercio electrónico para informar sobre las preferencias de compra de los clientes. (Morales, Soto, & Hervás Martínez, 2005).

C. Análisis de Secuencias

El análisis de secuencias o secuencia de patrones (Agrawal & Srikant, 1995) es una técnica de minería de datos que descubre las secuencias dentro de un conjunto de datos. Al igual que las reglas de asociación, también se han aplicado en sistemas de comercio electrónico para descubrir secuencias de acciones de los clientes. Su aplicación a sistemas de aprendizaje automático permite analizar secuencias de páginas visitadas durante una sesión o en distintas sesiones de un mismo usuario. (Morales, Soto, & Hervás Martínez, 2005).

3.1.2. Herramientas de Código Abierto para el procesamiento y visualización de

Minería de Datos en Fuentes Abiertas

Teniendo en cuenta la importancia de llevar a cabo un proceso de minería de datos, la segunda fase que se debe tener en cuenta y en la que la gran mayoría de autores convergen en sus investigaciones, es en la recolección y análisis del texto de las fuentes abiertas, ya que el texto es uno de los principales insumos de datos que deben ser recolectados, por tal motivo es de vital importancia verificar que herramientas de código abierto servirán para el análisis y

² Los algoritmos de clustering o agrupamiento tienen como objeto descubrir una estructura oculta utilizando similitudes o diferencias (como distancias) entre puntos individuales del conjunto de datos que está bajo consideración, es decir, se trata de un procedimiento de agrupación de acuerdo a un criterio. (Kaufman & Rousseeuw, 2009)

visualización de los datos textuales para ser analizados y posteriormente difundirlos oportunamente.

A nivel de implementación y desarrollo de programación de acuerdo a los criterios de Marco Bonzanini (2016) podemos desglosar nuestro proyecto en dos grandes sistemas que pueden ser utilizados:

- Tipos de herramientas: Lenguajes de programación.
- Entornos de edición visual de workflow.

A nivel de lenguajes de programación, R y Python (Bonzanini, 2016) son los de más uso generalizado y permiten a los expertos programar código altamente versátil para sofisticadas tareas analíticas. (Hofmann, Chisholm, Chisholm, & Berthold, 2016)

En el otro extremo del espectro, están las de herramientas de trabajo visual que permiten incluso a los no expertos utilizar plantillas predefinidas (o planos) y modificar los análisis a los cuales se quieren llegar.

El uso de un workflow visual tiene el beneficio adicional de poseer una documentación intuitiva y la guía a través del proceso que se crea implícitamente. RapidMiner (versión 5.3, que sigue siendo de código abierto) y KNIME son ejemplos de este tipo de herramientas. Vale la pena señalar que especialmente este último está sobre los hombros de los gigantes: KNIME no sólo integra R y Python sino también varias bibliotecas. Estos permiten el uso de métodos de última generación a través de un editor gráfico fácil de usar. (Hofmann et al., 2016)

Independientemente del lenguaje de programación o de la metodología a realizar, es necesario comprender que las interacciones y relaciones sociales nos unen y facilitan el flujo de información, ideas, percepciones, etc., en un sistema de comunicación, global y local, en red.

El término red o network es muy gráfico, ya que evoca imágenes de nodos (individuos) conectados por medio de líneas (relaciones). Una red social consiste en “un conjunto finito de actores y la relación definida entre esos actores” (Wasserman Faust, 1994) considerados como unidades interdependientes y que pueden ser representadas por medio de un grafo. (Del-Fresno-García, 2014).

La teoría y aplicación del Análisis de Redes Sociales brinda una serie de métodos para el análisis de las interacciones sociales que permiten de forma específica investigar las estructuras relacionales y su representación como redes. Algunos científicos y matemáticos han contribuido con conceptos y métricas aplicables al estudio sistemático de procesos sociales impensables hace unos años para las ciencias sociales. Estos estudios proporcionan acceso a nuevos conocimientos como la representación de las estructuras relacionales y como éstas pueden ser consecuencia de la acción individual y colectiva. Para Linton C. Freeman el análisis tiene cuatro características singulares:

- A. Obtiene una intuición estructural de las relaciones sociales.
- B. La obtención de datos empíricos ha de efectuarse de manera sistemática.
- C. Los modelos matemáticos son una parte fundamental del análisis con la ayuda de ordenadores como herramientas para su explotación y visualización.
- D. Analizar y difundir las visualizaciones de las relaciones, los patrones de las interacciones, permite la generación de ideas estructurales significativos y su comunicación a otros, de alguna manera pueden afectar la toma de decisiones estratégicas.

Su aplicación al contexto de mercadeo social o web 2.0 aún está en fase de desarrollo embrionario (Xu; Lin, 2013; Del Fresno-García; Daly, 2014) y es de esperar que se desarrollen más investigaciones empíricas en breve.

3.1.3. Metodologías de Análisis de captura de datos

La captura y extracción de datos puede llevarse a cabo desde cualquier Red Social (Twitter, Facebook, YouTube, Google+) interactuando directamente con su respectiva API (application programming interface), para así definir un objeto de estudio por medio de una palabra, frase clave, hashtag, etc.

De acuerdo a los estudios realizados, esta es la manera más acertada para hacer uso de la obtención de los datos en las redes sociales, la cual no nos proporciona mucho valor si poseemos todos los datos en bruto. Pero si nos general el interrogante: *¿Qué podemos hacer con los datos para llegar a un producto de inteligencia y que concluya en una toma de decisión acertada?*

Dadas estas condiciones se formulará una hipótesis a un problema que afecte los intereses de la nación, y de acuerdo a sus características se responderán unos interrogantes que nos permitan dirimir la incertidumbre por medio de una visualización gráfica y estadística, captando estas variables en el entorno de las redes sociales.

Ya una vez que entendemos lo que podemos hacer con los datos, la última pregunta que surge es: ¿Cómo realizaríamos el procesamiento y difusión de la información? La herramienta que hasta el momento ha tenido más auge y acogida es Python y R, por ser uno de los principales lenguajes para la ciencia de los datos, su sintaxis y semántica son fáciles de entender, junto con su rico ecosistema para la computación científica.

3.1.4. Métodos de autenticación de las API

Gracias a la amplia adopción de plataformas de colaboración y redes sociales, los desarrolladores de aplicaciones tienen la oportunidad de conectar a los usuarios con sus datos en cualquier parte del mundo a través de una plataforma Web o el desarrollo de una API a medida. Por medio de esta integración, se mejora la eficiencia diaria al eliminar la compartimentación de la información, como también permite a los desarrolladores diferenciar sus aplicaciones de la competencia.

OAuth³ proporciona la capacidad para que cualquier aplicación desarrollada por un arquitecto de software logre acceder a los datos de un usuario de forma segura, sin requerir que el usuario tenga que entregar una contraseña de cuenta.

Los tipos de funcionalidad proporcionados por las API autorizadas para OAuth (Boyd, 2012), incluyen lo siguiente:

- Obtener acceso a la red social de un usuario: sus amigos de Facebook, las personas que está siguiendo en Twitter, o sus contactos de Google, etc.
- Compartir información sobre las actividades de un usuario en su sitio mediante la publicación en su Facebook.

³ El estándar de autorización de OAuth permite que una aplicación externa obtenga acceso limitado a un servicio HTTP, ya sea en nombre de un propietario de recursos, organizando una interacción de aprobación entre el propietario del recurso y el servicio HTTP, o permitiendo que la aplicación de un tercero obtenga acceso en su propio nombre. Esta especificación reemplaza el protocolo obsoleto OAuth 1.0 descrito en RFC 5849.

- Acceder a la cuenta de Google Docs. o Dropbox de un usuario para almacenar datos en su sistema de archivos.

Para acceder o actualizar los datos a través de una API, la aplicación requiere la autorización del administrador de los datos, que en nuestro caso son las redes sociales como Facebook, Twitter, etc. Para el año 2018 encontramos alrededor de 19.800 API con soporte OAuth para garantizar el acceso a sus datos. (ProgrammableWeb, 2018)

Tener un protocolo común para manejar la autorización de la API mejora enormemente la experiencia del desarrollador, porque disminuye la curva de aprendizaje requerida para integrarse con una nueva API. Asimismo, un estándar de autorización crea más confianza en la seguridad de las API, ya que estas han sido valoradas por la comunidad de desarrolladores de software.

Para comprender OAuth, es importante entender primero la terminología relevante. En este marco teórico ahondaremos algunos términos clave que permitirán el entendimiento de las sentencias que se describirán en la abstracción de datos.

A. Autenticación

Autenticación es el proceso de verificar la identidad de un usuario, confirmando que el usuario es quien dice ser. La autenticación se realiza típicamente pidiendo al usuario un nombre de usuario y una contraseña.(Boyd, 2012).

B. Autenticación Federada

En la actualidad, algunas aplicaciones tienen su propio sistema de autenticación de cuentas (incluidos los nombres de usuario y contraseñas), otras aplicaciones dependen de servicios externos para verificar la identidad de los usuarios. Esto se llama autenticación federada. En un entorno de TI corporativo, las aplicaciones pueden confiar en un servidor como el Active Directory, el Servidor LDAP⁴ o un proveedor SAML⁵ para autenticar a los usuarios. En la Web,

⁴ Por sus siglas en ingles (Lightweight Directory Access Protocol) cuya traducción al español significa, Protocolo Simplificado de Acceso a Directorios.

⁵ Por sus siglas en ingles (Security Assertion Markup Language) cuya traducción al español significa, Lenguaje de Marcado para Confirmaciones de Seguridad.

las aplicaciones a menudo confían en los proveedores de OpenID⁶ (como Google o Yahoo!) para manejar la autenticación de los usuarios. OpenID es el protocolo web abierto más común para el manejo de la autenticación federada.(Boyd, 2012).

C. Autorización

La autorización es el proceso de verificar que un usuario tiene el derecho a realizar alguna acción, como leer un documento o acceder a una cuenta de correo electrónico. Esto generalmente requiere una identificación válida del usuario (autenticación) para verificar si el usuario es real o está autorizado. En informática, una aplicación web primero verifica su identidad iniciando sesión y luego se asegura de que solo acceda a los datos y servicios que tiene permitido, generalmente verificando en una lista de control de acceso para cada operación.(Boyd, 2012).

D. Autorización delegada

La autorización delegada es otorgar acceso a otra persona o aplicación para realizar ciertas acciones en mi nombre. OAuth funciona de manera similar: un usuario otorga acceso a una aplicación para realizar ciertas acciones en el nombre del usuario y la aplicación se encarga de realizar las acciones autorizadas.(Boyd, 2012).

E. Roles

Hay varios actores clave en los flujos del protocolo OAuth:

a. Servidor de recursos

Es el servidor que aloja los recursos de propiedad del usuario, los cuales están protegidos por OAuth. Esto es, típicamente un proveedor de la API que contiene y protege datos tales como fotos, videos, calendarios, o contactos.(Boyd, 2012).

b. Propietario del recurso

Normalmente, el usuario de una aplicación o el propietario del recurso tienen la capacidad de otorgar acceso a sus propios datos alojados en el servidor de recursos.(Boyd, 2012).

⁶ Estándar de identificación digital descentralizado.

c. Servidor de autorizaciones

El servidor de autorizaciones obtiene el consentimiento del propietario del recurso y los tokens de acceso a los clientes para acceder a recursos alojados por un servidor de recursos.(Boyd, 2012).

3.1.5. Abstracción de datos [Minería de datos]-[MUESTREO]

3.1.5.1. API Twitter

El primer paso para obtener el acceso a los datos de Twitter es mediante la utilización de un aplicativo (API), que interactúe con la red social, y permita realizar la recolección de datos con el fin de generar conocimiento que se maleable hacia los fines propuestos.

La red social Twitter cuenta con un desarrollo propio, que, mediante la configuración de una API, lograr tener una comunicación segura para que en una segunda fase sea posible realizar la exploración y manipulación del conocimiento recolectado.

Para ello es necesario realizar una serie de pasos descritos a continuación:

A través de la URL <https://dev.twitter.com/> se puede realizar el inicio de sesión con las credenciales que se tengan para Twitter, para efectuar la configuración de la API:

- Ingresar a la URL <https://apps.twitter.com/app/new>. (Ver Figura 3. Creación de la aplicación del API de Twitter)
- Digitar el nombre de la aplicación en el campo Nombre. Este nombre puede ser cualquiera.
- Del mismo modo, debe ingresar la descripción en el campo “Descripción”.
- El campo del sitio web debe llenarse con una URL válida.
- El siguiente campo de URL puede dejarlo en blanco.

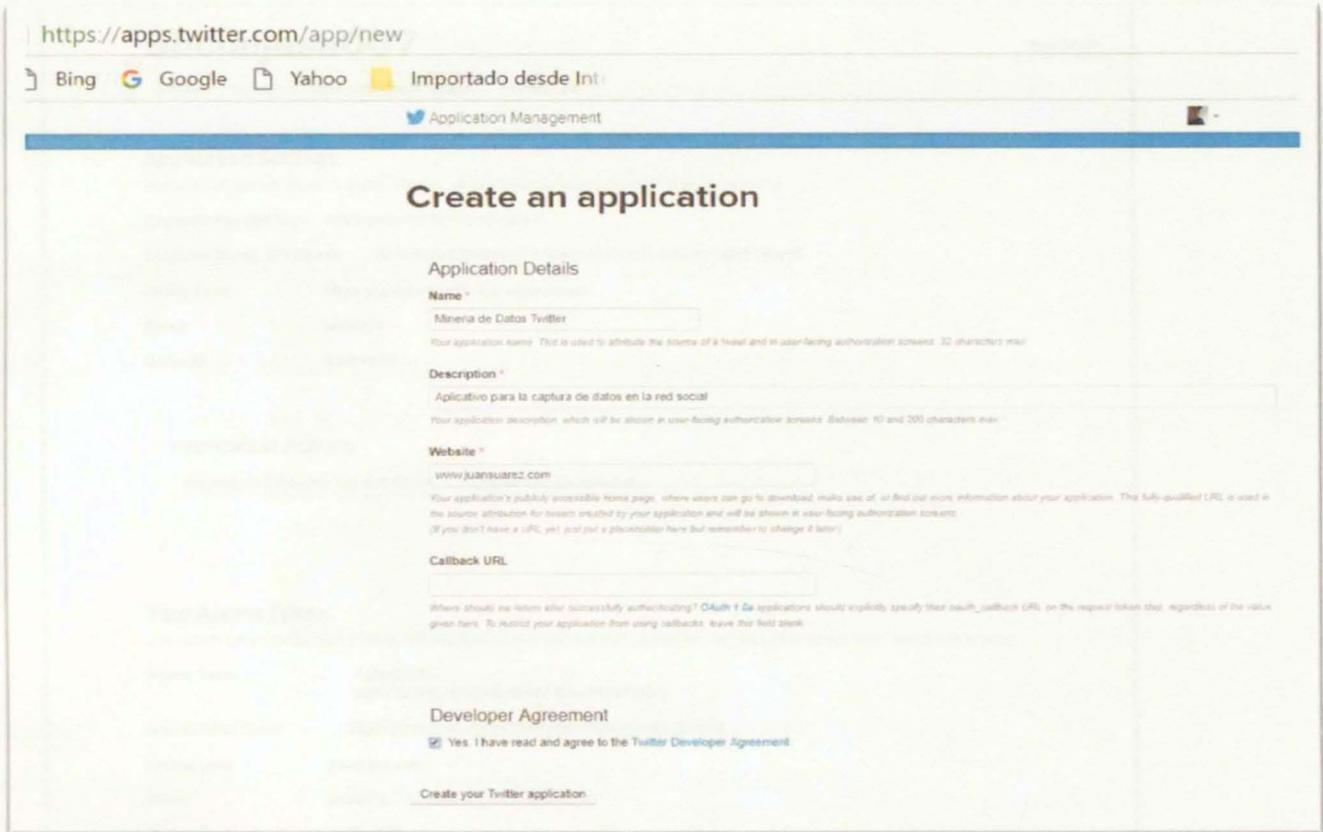


Figura 3. Creación de la aplicación del API de Twitter.

Nota: Elaboración propia.

Después de la creación de esta aplicación, debemos encontrar la “API Key” y el “API Secret”. Como se muestra en la siguiente figura:

Para realizar la comunicación de Twitter con nuestra aplicación, Ver Figura 4. Métodos de Autenticación de la Red Social de Twitter)

- Key
- Secret
- Access token
- Access token secret

Una vez se tiene almacenada la información de las llaves se podrá obtener una aplicación autorizada por el usuario con acceso a los recursos de Twitter. La función OAuth en twitter realiza las interacciones necesarias para la comunicación.

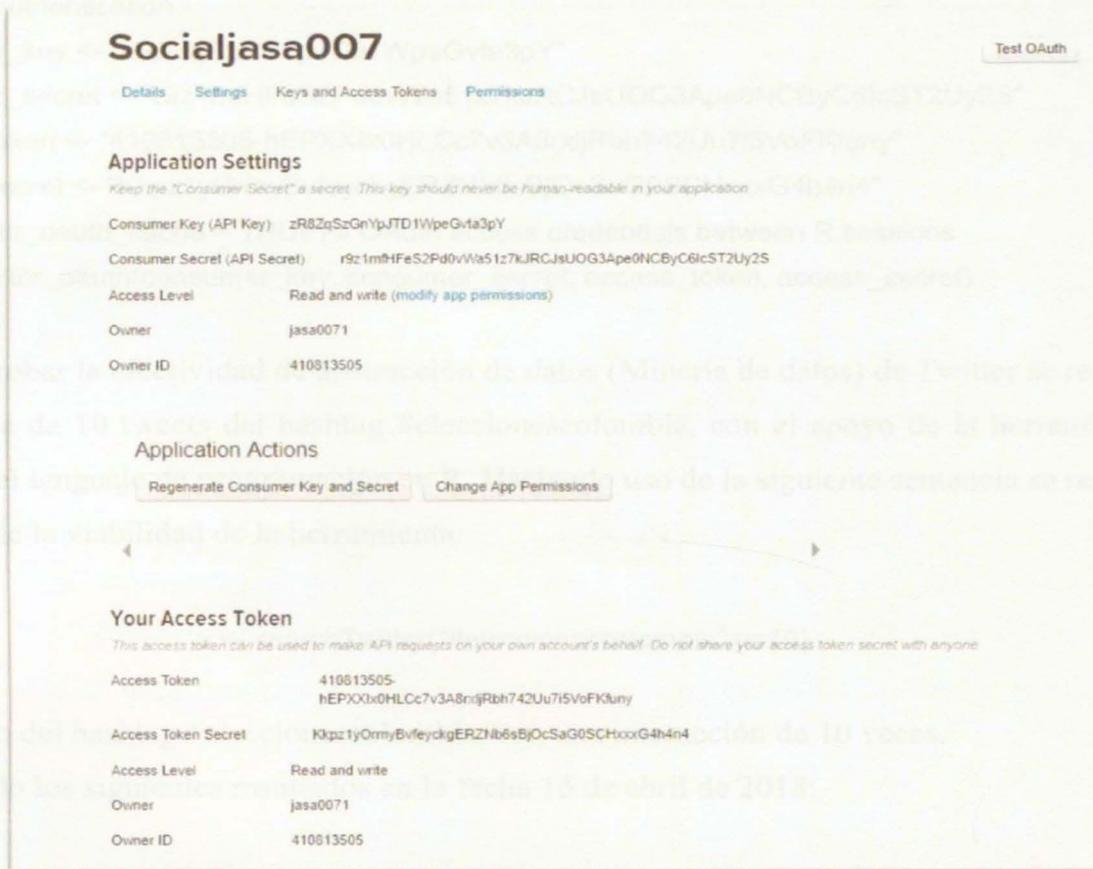


Figura 4. Métodos de Autenticación.

Nota: Elaboración propia

- Estos datos son de suma importancia, se debe generar un archivo en donde almacene estas llaves privadas, con el fin de realizar la comunicación de Twitter con nuestro aplicativo. Ver Figura 4. Métodos de Autenticación de la Red Social de Twitter)

- Key
- Secret
- Access token
- Access token secret

Una vez se tiene almacenada la información de las llaves se podrá obtener una aplicación autorizada por el usuario con acceso a los recursos de Twitter. La función ROAuth en twitterR realiza las interacciones necesarias para la comunicación:

```
# OAuth authentication
consumer_key <- "zR8ZqSzGnYpJTD1WpeGvta3pY"
consumer_secret <- "r9z1mfHFeS2Pd0vWa51z7kJRCJsUOG3Ape0NCByC6lcST2Uy2S"
access_token <- "410813505-hEPXXIx0HLCc7v3A8rxljRbh742Uu7i5VoFKfuny"
access_secret <- "Kkpz1yOrmyBvfeyckgERZNB6sBjOcSaG0SCHxxxG4h4n4"
options(httr_oauth_cache = TRUE) # OAuth access credentials between R sessions
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

Para comprobar la efectividad de abstracción de datos (Minería de datos) de Twitter se realiza la búsqueda de 10 tweets del hashtag #eleccionescolombia, con el apoyo de la herramienta RStudio y el lenguaje de programación en R. Haciendo uso de la siguiente sentencia se realiza el estudio de la viabilidad de la herramienta:

```
a <- searchTwitter("#eleccionescolombia",n=10)
```

Se hace uso del hashtag #eleccionescolombia, con una interacción de 10 veces.

Encontrando los siguientes resultados en la fecha 16 de abril de 2018:

```
[[1]]
```

```
[1] "Julsv_M: La oposición ya se dio cuenta que en los marihuaneros está un a muy buena opción de sumar votos a su causa? No son p... https://t.co/kCjT2aBRud"
```

```
[[2]]
```

```
[1] "Lapizaso: RT @RubenPinzonHdez: \u26a0;ALERTA!! #FraudeALaVista #EleccionesColombia #ColombiaDecide\n-----\n@SenorCaicedo @PetroVoluntario @LaryzEden @aD..."
```

```
[[3]]
```

```
[1] "biolucho2005: RT @jairopulgarin: #EleccionesColombia \nEmbriagados de felicidad, los jefes políticos del Polo tomaron la decisión de seguir con Fajardo...."
```

```
[[4]]
```

```
[1] "aydecris: RT @colombiaccesibl: Pendientes mañana del debate sobre #discapacidad en la ciudad de Cali.\n#EleccionesColombia https://t.co/Ns0LdLdB1k"
```

```
[[5]]
```

```
[1] "diegopfaff: RT @FERJEDY: #DesmoteParamilitarYa:MALDITOS\n#DesmoteGuerilleroYA: MALDITOS\n#DesmoteDeCorruptosYA:MALDITOS\n#DesmoteDePolitiqueirosYA:MAL..."
```

```
[[6]]
```

```
[1] "waltermunevar: RT @jairopulgarin: #EleccionesColombia \nEmbriagados de felicidad, los jefes políticos del Polo tomaron la decisión de seguir con Fajardo...."
```

```
[[7]]
```

```
[1] "sorendion: RT @jairopulgarin: #EleccionesColombia \nEmbriagados de fel
icidad, los jefes políticos del Polo tomaron la decisión de seguir con Faja
rdo...."
```

```
[[8]]
```

```
[1] "oreja777: RT @jairopulgarin: #EleccionesColombia \nEmbriagados de fel
icidad, los jefes políticos del Polo tomaron la decisión de seguir con Fajar
do...."
```

```
[[9]]
```

```
[1] "MimeGalindo: RT @jairopulgarin: #EleccionesColombia \nEmbriagados de f
elicidad, los jefes políticos del Polo tomaron la decisión de seguir con Fa
jardo...."
```

```
[[10]]
```

```
[1] "jairopulgarin: #EleccionesColombia \nEmbriagados de felicidad, los jef
es políticos del Polo tomaron la decisión de seguir con Fajar... https://t.co
/IJTOnyYpPA"
```

Del primer Tweet encontrado se pueden rescatar 17 campos, de los cuales se podrían analizar todos los metadatos a través de un sondeo profundo a los campos que se desee tener mayor interacción. Uno de esos campos que es de incumbencia para la investigación lo podemos encontrar en la localización del Tweet, a través de la longitud y latitud desde donde fue generado. (Ver Figura 5. Atributos de la consulta de Twitter)

```

data
  a
    List of 10
      :Reference class 'status' [package "twitterR"] with 17 fields
      ..$ text : chr "La oposición ya se dio cuenta que en los marihuaneros está una
      ..$ favorited : logi FALSE
      ..$ favoriteCount: num 0
      ..$ replyToSN : chr(0)
      ..$ created : POSIXct[1:1], format: "2018-04-17 02:21:45"
      ..$ truncated : logi TRUE
      ..$ replyToSID : chr(0)
      ..$ id : chr "986067158206746629"
      ..$ replyToUID : chr(0)
      ..$ statusSource : chr "<a href=\"http://twitter.com/download/iphone\" rel=\"no
      ..$ screenName : chr "Julsv_M"
      ..$ retweetCount : num 0
      ..$ isRetweet : logi FALSE
      ..$ retweeted : logi FALSE
      ..$ longitude : chr(0)
      ..$ latitude : chr(0)
      ..$ urls :'data.frame': 1 obs. of 5 variables:
      .. ..$ url : chr "https://t.co/kCjT2abRUD"
      .. ..$ expanded_url: chr "https://twitter.com/i/web/status/986067158206746629"
      .. ..$ display_url : chr "twitter.com/i/web/status/9"
      .. ..$ start_index : num 117
      .. ..$ stop_index : num 140
      ..and 53 methods, of which 39 are possibly relevant:
      .. getCreated, getFavoriteCount, getFavorited, getId, getIsRetweet, getLatitude,
      .. getLongitude, getReplyToSID, getReplyToSN, getReplyToUID, getRetweetCount, g
      .. getRetweeters, getRetweets, getScreenName, getStatusSource, getText, getTrun
  
```

Figura 5. Atributos de una consulta de Twitter.

Nota: Elaboración propia

3.1.6. Exploración de la Información

La limpieza y preparación de los datos constituye una parte fundamental del esfuerzo en la exploración de los datos. Es de aclarar que ningún conjunto de datos es perfecto; siempre le faltará información, habrá datos que se malinterpreten o estén incorrectos. Algunos campos de datos estarán sucios e inconsistentes.

Por lo tanto, es de vital importancia en esta etapa del proceso, examinar los datos antes de comenzar a modelar; en el peor de los casos, puede suceder que el modelo que se esté construyendo arroje predicciones incorrectas. En este documento, demostraremos algunas de las cosas que pueden salir mal con los datos, y exploraremos algunas maneras para abordar esos problemas utilizando el lenguaje estadístico R (<https://cran.r-project.org/>) antes de continuar el análisis. A lo largo de este proceso (Exploración de la información), se mantendrá el objetivo de usar el aprendizaje automático para construir un modelo predictivo. (De Jonge & Van Der Loo, 2013).

El propósito del preprocesamiento de datos es extraer datos útiles de la red social Twitter sin procesar y luego transformar estos datos en el formato necesario para el descubrimiento de patrones.

Debido a la gran cantidad de información irrelevante en los registros web de la red social, el registro original no se puede usar directamente en el procedimiento de minería, por lo tanto, en la fase de exploración de la información, los datos sin procesar deben ser limpiados, analizados y convertidos para su posterior paso de modelado. Al limpiar los datos, podemos crear la base de datos de acuerdo a nuestras necesidades que incluya la información sobre la identificación de usuario, identificación de sesión, ruta terminación, etc.

El preprocesamiento de datos involucra la eliminación de ruido como las anotaciones de los usuarios, URL, signos de puntuación, hashtags, controles y caracteres especiales.

3.1.7. Manipulación y Modelado de los datos

3.1.7.1. Minería de datos

Básicamente, la minería de datos en redes sociales es un análisis sistemático de la información generada de las redes sociales. Es imperativo aprovechar los enormes datos de las redes sociales

con la ayuda de la tecnología actual, que no está exenta de desafíos. Procesar un gran conjunto de datos medidos en petabytes es desafiante, y cosas como la relación señal / ruido S/N deben ser tomadas en consideración. Se estima que alrededor del 20 por ciento de tales flujos de datos de redes sociales contienen información relevante.

El conjunto de herramientas y técnicas que se utilizan para extraer dicha información es llamado técnicamente como minería de datos y en el contexto de las redes sociales es llamado minería de redes sociales (SMM⁷). SMM puede aportar conocimiento sobre cuánto alguien está influenciando (patrones tendenciales) a otros en la Web. SMM puede ayudar a las entidades a identificar puntos de dolor de su cliente en tiempo real. A su vez, esto se puede usar para una planificación proactiva. (Ravindran, Sharan Kumar Garg, 2015)

La minería de redes sociales tiene sus raíces de muchos campos, como en la estadística, el aprendizaje autónomo, la recuperación de información, los patrones de reconocimiento y bioinformática. (Luis, Mejías, & Auxiliadora, 2018)

La enorme cantidad de datos que se generan diariamente es sorprendente, pero las técnicas actuales permiten que nuevas soluciones de minería de datos y modelos computacionales escalables sean de ayuda para futuros conceptos y teorías fundamentales de computación.

En la minería de datos, los datos son la materia prima. En el momento que se les atribuye algún significado especial pasan a convertirse en información valiosa para la toma de decisiones.

En el argot popular informático los ingenieros definen la minería de datos como un sinónimo de otro término popularmente utilizado, descubrimiento de conocimientos a partir de datos, o KDD, mientras que otros ven la minería de datos simplemente como un paso esencial en el proceso de descubrimiento de conocimiento. (Han, Kamber, & Pei, 2011). El proceso de descubrimiento de conocimiento es se muestra en la Figura 6 como una secuencia iterativa de la metodología SEMMA.

Este diagrama de flujo muestra una extracción de datos dominante y en donde se pueden realizar, posteriormente, tanto las formas de datos o la transformación de datos sociales. Por lo general, las redes sociales están compuestas por comunidades de usuarios que conforman un grupo de usuarios en línea que interactúan entre sí y se relacionan dentro de la misma comunidad o grupo de usuarios a través de sus perfiles o nombres. (Zaki, Meira, & Meira, 2013)

⁷ Por sus siglas en inglés “Social media marketing”. Son las medidas implementadas por la empresas para promover los servicios dentro de una red social.

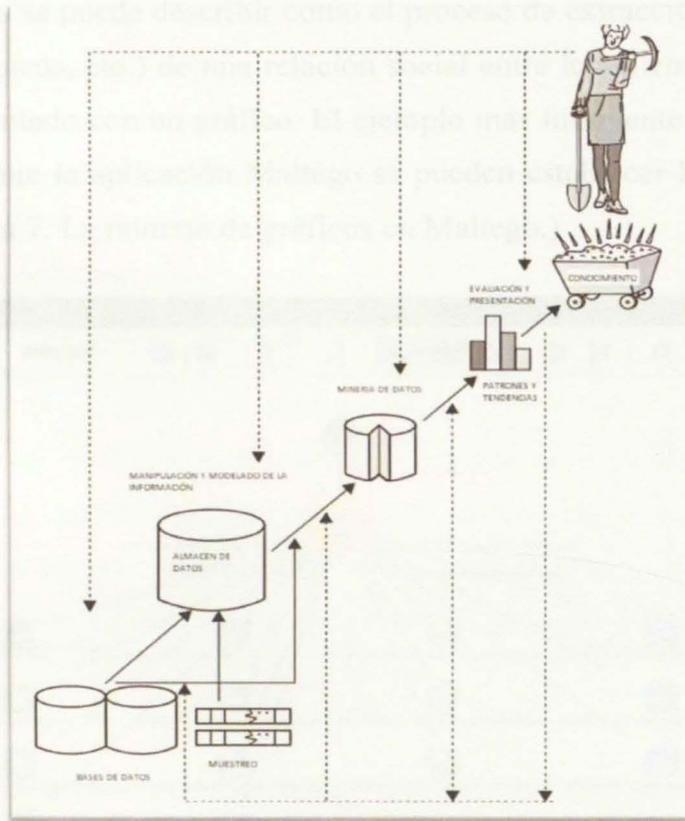


Figura 6. La minería de datos como un paso en el proceso de descubrimiento de conocimiento.

Fuente : (Han et al., 2011)

Básicamente en la actualidad existen dos técnicas aplicables para al procesamiento de la minería de datos en redes sociales. Para este caso de estudio, se utilizarán como ejemplos las redes sociales Facebook y Twitter como fuentes de datos.

3.1.7.2. Minería en Gráficos

Los diagramas de red son una estructura de datos dominante y en donde se pueden rescatar, esencialmente, todas las formas de datos o la información de redes sociales. Por lo general, las redes sociales están compuestas por comunidades de usuarios que constituyen un grupo de nodos, en dichos gráficos los nodos que se encuentran dentro de la misma comunidad o grupo tienden a compartir características comunes.(Zaki, Meria, & Meira, 2013)

La minería de gráficos se puede describir como el proceso de extracción de conocimiento útil (patrones, valores atípicos, etc.) de una relación social entre los miembros de la comunidad, que puede ser representado con un gráfico. El ejemplo más influyente de minería de gráficos es Twitter que mediante la aplicación Maltego se pueden establecer las relaciones entre sus miembros. (Ver Figura 7. La minería de gráficos en Maltego.)

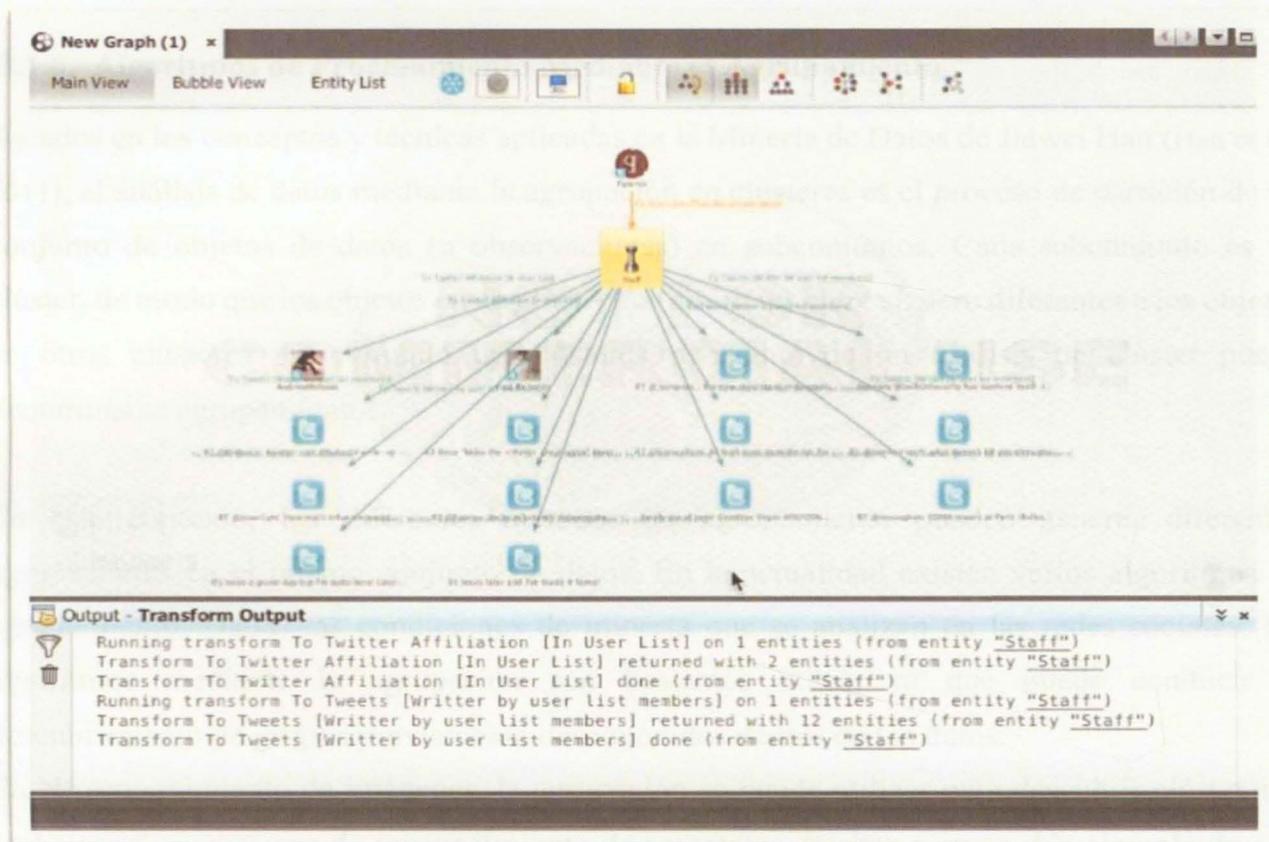


Figura 7. La minería de gráficos en Maltego.

Nota: Elaboración propia

3.1.7.3. Minería de texto

La minería de texto tiene como objetivo la extracción de los datos no estructurados presentes en las redes sociales. Los objetivos principales de este tipo de minería son los blogs y microblogs⁸ como Twitter. Puede ser ajustable a otras redes sociales como Facebook, que contienen enlaces a publicaciones, blogs y otros artículos de noticias.(Schab et al., 2018)

⁸ Los sitios web de microblogging han evolucionado para convertirse en fuente de tipo variado de información. Esto es debido a naturaleza de los microblogs en los que las personas publican en tiempo real mensajes sobre sus

Ella se define como el proceso de descubrimiento de patrones interesantes y nuevos conocimientos en una colección de textos, es decir, la minería de texto es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos (Hearst, 1999; Kodratoff, 1999).

3.1.8. Algoritmos de Procesamiento Mediante el Agrupamiento

Basados en los conceptos y técnicas aplicadas en la Minería de Datos de Jiawei Han (Han et al., 2011), el análisis de datos mediante la agrupación en clústeres es el proceso de partición de un conjunto de objetos de datos (u observaciones) en subconjuntos. Cada subconjunto es un clúster, de modo que los objetos en un clúster son similares entre sí, pero diferentes a los objetos en otros clústeres. El conjunto de clústeres resultante de un análisis de clúster puede denominarse agrupamiento.

En este contexto, los diferentes métodos de agrupamiento pueden generar diferentes agrupaciones en el mismo conjunto de datos. En la actualidad existen varios algoritmos de agrupamiento. Dadas las condiciones de minería que se analizan en las redes sociales, los algoritmos mediante la agrupación son bastantes útiles, ya que puede conducir al descubrimiento de grupos previamente desconocidos dentro de los datos.

En el reconocimiento de imágenes, la agrupación se puede utilizar para descubrir clústeres o "subclases" en sistemas de reconocimiento de caracteres escritos a mano. Un ejemplo de ello suele suceder en un escenario en que se tiene un conjunto de datos de dígitos escritos a mano, donde cada dígito está etiquetado como 1, 2, 3, y así sucesivamente. En este ejemplo puede haber una gran variación en la forma en que las personas escriben el mismo dígito. Se toma como ejemplo el número 2, algunas personas pueden escribir con un círculo pequeño en la parte inferior izquierda, mientras que otras pueden no hacerlo. Al usar método de agrupación para

opiniones sobre una variedad de temas, discutir problemas actuales, quejarse y expresar resultados positivos sentimiento de los productos que utilizan en la vida cotidiana. Actualmente, las compañías que fabrican tales productos realizan sondeos a sitios de microblogs para obtener una perspectiva general del sentimiento por su producto. Muchas veces la gran mayoría de las empresas estudian las reacciones de los usuarios y responden a los usuarios mediante los sitios de microblogs. (Han et al., 2011)

determinar las subclases para el número "2", cada una de los grafos representa una variación en la forma en que se pueden escribir el número dos (2). El uso de los múltiples modelos basados en las subclases puede mejorar la precisión general del reconocimiento.

En la minería de datos, el análisis de clúster o de agrupamiento se puede utilizar como una herramienta independiente para adquirir información sobre la distribución de los datos, observar las características de cada clúster, o para centrarse en un conjunto particular de clústeres para su posterior análisis. Alternativamente, puede servir como un paso de preprocesamiento para otros algoritmos, como la caracterización, la selección de subconjuntos de atributos y la clasificación, que luego operaría en los clústeres detectados y los atributos o características seleccionados.

A diferencia de los métodos de clasificación y la regresión, que analizan conjuntos de datos etiquetados (serie de entrenamientos), los algoritmos de agrupación analizan objetos de datos sin consultar las etiquetas de clase. La agrupación en clúster puede usarse para generar etiquetas de clase para un grupo de datos. Los objetos se agrupan basándose en el principio de maximizar la similitud intraclase y minimizar la similitud interclase. En otras palabras, se forman grupos de objetos para que los objetos dentro de un clúster tengan una gran similitud en comparación unos con otros, pero son diferentes a los objetos en otros grupos. Cada grupo formado se puede ver como una clase de objetos, de los cuales se pueden derivar reglas. El agrupamiento también puede facilitar la formación de taxonomía, es decir, la organización de observaciones en una jerarquía de clases que agrupan sucesos similares en un conjunto.

El análisis de agrupaciones se puede efectuar en los datos de la población de Colombia para identificar subpoblaciones homogéneas de personas que van a votar hacia la presidencia. Estos clústeres pueden representar grupos objetivo individuales para el sistema de votación.

Para ilustrar la metodología de agrupamiento, se va a suponer que se cuenta con la población de raza negra en Colombia, la cuales pueden estar ubicadas en diferentes zonas del país, como la costa caribe, costa pacífica o del interior de la nación; todos estos individuos van a efectuar el derecho al voto, pero en diferentes ubicaciones y no siempre por los mismos candidatos. La Figura 8 visualiza un gráfico en 2-D de la población de habitantes con respecto a las ubicaciones de las personas en una ciudad. Tres grupos de puntos de datos son evidentes.

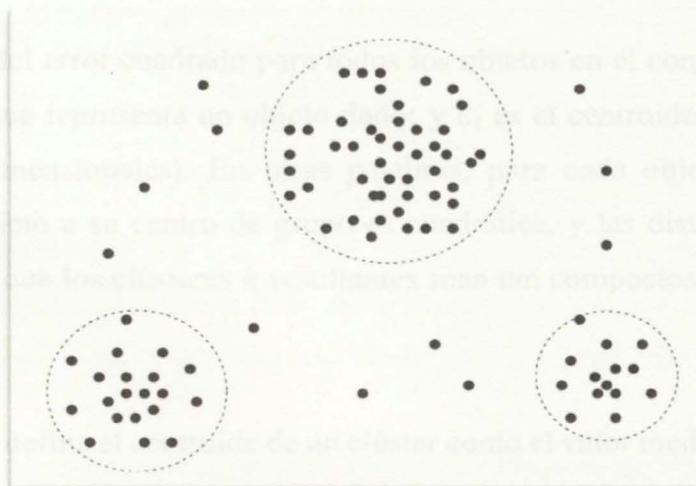


Figura 8. Un gráfico en 2-D de los datos de la población de Colombia con respecto a las ubicaciones de los habitantes en una ciudad, visualizando tres grupos de datos diferentes. Fuente : (Han et al., 2011)

3.1.8.1. Modelado con el Algoritmo K-Means

Teóricamente y de acuerdo a los estudios de los Algoritmos de Aprendizaje: KNN & K-means (García & Gómez, 2015) en un conjunto de datos, D , que contiene n objetos en el espacio euclidiano. Los métodos de particionamiento distribuyen los objetos en D en k clúster, C_1, \dots, C_k , es decir, $C_i \subset D$ y $C_i \cap C_j = \emptyset$ para $(1 \leq i, j \leq k)$. Por medio del algoritmo K-means se puede evaluar la calidad de la partición de modo que los objetos dentro de un clúster sean similares entre sí, pero diferentes a otros conjuntos. Es decir, la función objetivo apunta a una alta similitud intracluster y baja similitud intercluster.

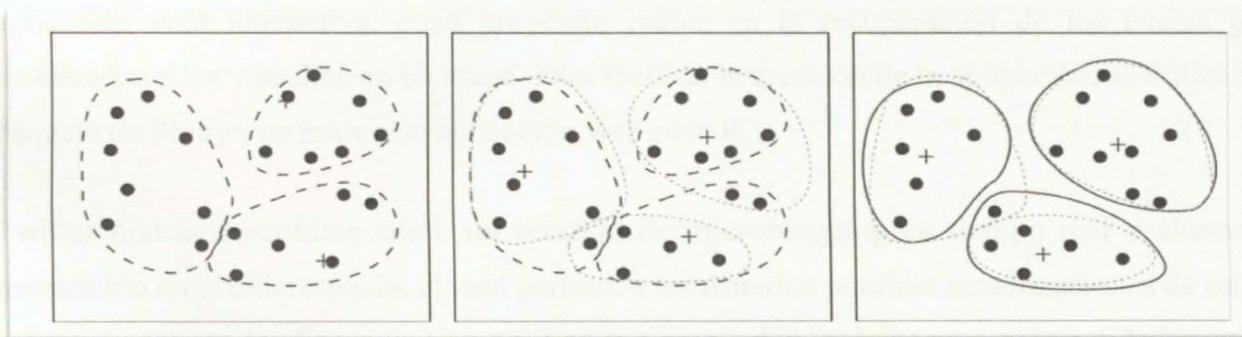
El algoritmo de partición está basado en la técnica del centroide, el cual utiliza el centro como punto de partida para un clúster, C_i , matemáticamente, el centroide de un clúster es su punto central. El centroide se puede definir de varias maneras, ya sea por la media de los objetos (o puntos) asignado al clúster, como también por la diferencia entre un objeto $p \in C_i$ y C_i , el representativo del conglomerado, se mide por la distancia (p, c_i) , donde la distancia (x, y) es la distancia euclidiana entre dos puntos (x, y) . La calidad del clúster C_i se puede medir mediante la variación dentro del clúster, que es la suma del error cuadrático entre todos los objetos en C_i y el centroide C_i , definido como:

$$E = \sum_{i=1}^k \sum_{p \in C_i} dist(p, C_i)^2$$

Donde E es la suma del error cuadrado para todos los objetos en el conjunto de datos; p es el punto en el espacio que representa un objeto dado; y C_i es el centroide del grupo C_i (tanto p como C_i son multidimensionales). En otras palabras, para cada objeto en cada grupo, la distancia desde el objeto a su centro de grupo es cuadrática, y las distancias se suman. Esta función intenta hacer que los clústeres k resultantes sean tan compactos y tan separados como sea posible.

El algoritmo k-means define el centroide de un clúster como el valor medio de los puntos dentro del clúster. Se procede de la siguiente manera. Primero, se selecciona aleatoriamente k objetos del espacio D , cada uno de los cuales representa inicialmente una media o centro del grupo. Para cada uno de los objetos restantes, un objeto se asigna al clúster al que es más similar, según la distancia euclidiana entre el objeto y la media del clúster. El algoritmo k-means luego mejora iterativamente la variación dentro del clúster.

Para cada clúster, el algoritmo calcula la nueva media utilizando los objetos asignados al clúster en la iteración anterior. Todos los objetos se reasignan utilizando los medios actualizados como los nuevos centros de clúster. Las iteraciones continúan hasta que la asignación sea estable, es decir, los grupos formados en la ronda actual son los mismos que los formados en la ronda anterior. (García & Gómez, 2015) (Ver Figura 9. Algoritmo de agrupamiento K-means)



a) Clúster Inicial

b) Primera Interacción

c) Interacción Final Clúster

Figura 9. Algoritmo de agrupamiento K-means.

Fuente : (Han et al., 2011)

3.1.9. Valoración de la Información

La valoración o evaluación de los resultados se verificará de acuerdo a la efectividad del algoritmo de clustering K-Means al agrupar el mayor número de elementos con la mayor similitud, diferenciándolos de otras agrupaciones. Este tipo de agrupamiento puede proporcionar conocimiento si los datos son perfectamente agrupados de acuerdo a la selectividad de los parámetros establecidos en la búsqueda y a la reducción de los errores de clasificación grupal. Este tema se ahondará en el capítulo 3.4.3 Valoración(Assess).

3.2. Capítulo II. Efectividad de abstracción de datos (Minería de datos) en la red social

Twitter

Como se ha venido describiendo a lo largo de este capítulo, la finalidad de este apartado es demostrar que tan efectivo es la abstracción de datos en Twitter mediante la utilización de la API para desarrolladores.

La red social Twitter es una de las aplicaciones más populares en la actualidad, esto es debido a su versatilidad y a la facilidad para obtener datos para aplicaciones y proyectos web. A través de su API de búsqueda, es posible recuperar contenido mediante la emisión de una consulta a Twitter basada en los datos suministrados. Los resultados se pueden analizar o visualizar utilizando herramientas auxiliares.

Para este capítulo de abstracción de datos, el siguiente paso consiste en la creación de una aplicación web interactiva, cuyo propósito radica en la recuperación de los tweets geolocalizados y los visualice en un mapa. Para facilitar la creación de la aplicación, se utilizará el paquete de Shiny⁹, un marco de aplicación web para R.

Twitter podría describirse como un servicio de microblogging en tiempo real y altamente reconocido en el ciberespacio, el cual permite a los usuarios publicar actualizaciones de estado breves, tweets en las líneas de tiempo. Los tweets pueden incluir una o más entidades en sus

⁹ Shiny hace que sea increíblemente fácil crear aplicaciones web interactivas con R. "Reactivo" automático el enlace entre las entradas y salidas y los widgets pre construidos hacen posible construir aplicaciones receptivas y potentes con un esfuerzo mínimo. (Chang, 2018)

140 caracteres de contenido y en algunos casos pueden referenciar lugares que se asignan a ubicaciones en el mundo real. Una comprensión de los usuarios, tweets, y los cronogramas son particularmente esenciales para el uso efectivo de la API de Twitter. (Checa García, 2013)

Los tweets son la esencia de Twitter, y aunque se los considera en muchos aspectos como irrelevantes, los 140 caracteres del contenido de texto se encuentran asociados con la actualización de estado de un usuario, así como también una gran cantidad de metadatos alojados allí. Además del contenido textual de un tweet en sí, los tweets vienen empaquetados con dos piezas adicionales de metadatos que son de gran ayuda para el estudio a realizar: entidades y lugares.

Las entidades de tweet son esencialmente los usuarios, hashtags, las URLs, y los medios que pueden estar asociados con un tweet, y los lugares son ubicaciones en el mundo real que pueden adjuntarse a un tweet. Hay que tener en cuenta que un lugar puede ser la ubicación real en la que el tweet fue creado, pero también podría ser una referencia al lugar descrito en un tweet.

Finalmente, las líneas de tiempo son las colecciones ordenadas cronológicamente de tweets. Abstractamente, se podría decir que una línea de tiempo es cualquier colección particular de tweets que se muestran en orden cronológico; sin embargo, comúnmente se verá un par de líneas de tiempo, de inicio y de usuario, que son particularmente visibles; por ejemplo desde la perspectiva de un usuario de Twitter arbitrario, la línea de tiempo de inicio es la vista que ve cuando ingresa a su cuenta y mira todos los tweets de los usuarios que lo está siguiendo, mientras que una línea de tiempo de usuario en particular reside en la colección de tweets de un cierto usuario.

Twitter ha estado a la vanguardia tecnológica en la creación de API's simples e intuitivas. Aun así, hay grandes bibliotecas disponibles para mitigar aún más el trabajo en efectuar las solicitudes hacia los servidores.

3.2.1. Características de Shiny para R

De acuerdo al Framework de aplicaciones Web del repositorio de CRAN para el paquete "Shiny" (Winston Chang, 2018) se obtienen las siguientes características:

- A. Las aplicaciones web utilizan pocas líneas de código; no se requiere JavaScript.
- B. Las aplicaciones se "activan" automáticamente de la misma manera que las hojas de cálculo en vivo. Las salidas cambian instantáneamente a medida que los usuarios modifican las entradas, sin requerir una actualización del navegador.
- C. Las interfaces de usuario pueden construirse completamente usando R, o pueden escribirse directamente en HTML, CSS y JavaScript para mayor flexibilidad.
- D. Funciona en cualquier entorno R (Consola R, Rgui para Windows o Mac, ESS, StatET, RStudio, etc.)
- E. Puede usar Widgets de salida pre construidos para mostrar gráficos, tablas y resultados impresos de objetos R.
- F. Utiliza un modelo de programación reactiva que elimina el código de manejo de eventos desordenados, por lo que puede centrarse en el código que realmente importa.

3.2.2. Instalación

Shiny está disponible en CRAN, se puede instalar de la manera habitual desde la consola R, de la siguiente manera:

```
install.packages("shiny") (RStudio, Inc., 2018)
```

3.2.3. Acceso a la API de Twitter

Como primera medida, se debe crear una cuenta de Twitter. Una vez que se haya registrado, se debe redirigir a la dirección URL: apps.twitter.com, en donde se efectúa los pasos realizados anteriormente en el capítulo de abstracción de datos, esta aplicación servirá de ayuda para obtener acceso a la API de Twitter (abreviación de Interfaz de programación de aplicaciones). En la sección "Teclas y tokens de acceso" de la nueva aplicación se le proporcionará lo siguiente: Consumer Key, Consumer Secret, Access Token, and Access Token Secret.

Lastimosamente hay algunos límites¹⁰ de velocidad para la API de búsqueda de Twitter (Ver Tabla 1.) que debe tener en cuenta. Acorde a los estudios realizados por el Ingeniero Santacruz

¹⁰ Twitter filtra sus publicaciones en función de la calidad de los Tweets y de las cuentas. El material que perjudica la calidad de la búsqueda o que genera una experiencia de búsqueda deficiente para otras personas puede eliminarse automáticamente de la búsqueda de Twitter. (Twitter, 2018)

(2016) y teniendo en cuenta las pruebas efectuadas en este Capítulo para validar la efectividad de acceso de datos, se ha vislumbrado que solo un pequeño porcentaje de los tweets públicos están geo-localizados (alrededor de 3 – 5%) por lo que no es posible recuperar algunos tweets que dispongan de los campos de latitud y longitud como se hubiere deseado.

En el Search API y el REST API existe una limitación para la versión “ESTANDAR” de 180 peticiones a la hora por usuario o por dirección IP en caso que la petición no se encuentre autenticada frente al servidor, si la petición se encuentra autenticada el numero al cual podrá acceder se incrementará a 300 peticiones. (Ver Tabla 2.) Es necesario configurar el tamaño de las páginas para poder aprovechar el máximo de peticiones efectuadas. (Alonso García González Director de Reporte Técnico Dra Alejandra García Hernández, 2013)

Tabla 1.

Límite de páginas por API de Twitter.

MAXIMO TAMAÑO			
API	PETICION	PÁGINA(ESTANDAR)	MAXIMO TOTAL(PREMIUM)
Search ¹¹	Search	180 tweets	1500 tweets
REST ¹²	STATUS	180 tweets	1500 tweets
REST	Friends/ids	5000 id usuarios	Todos
REST	Seguidores/ids	5000 id usuarios	Todos

Fuente : (Alonso García González Director de Reporte Técnico Dra Alejandra García Hernández, 2013)

Todos los Tweets de Twitter se alojan en los servidores y bases de datos de la anterior compañía, y existe una limitante de tiempo para obtenerlos.

¹¹ El SEARCH API suministra los tweets con una profundidad en el tiempo de 7 días que se ajustan a la búsqueda (QUERY). Es posible filtrar por cliente, lenguaje y geolocalización. (Alonso García González Director de Reporte Técnico Dra Alejandra García Hernández, 2013)

¹² El REST API ofrece a los desarrolladores el acceso al core de los datos de Twitter. Todas las operaciones que se puedan hacer vía Web es posible realizarlas desde la API. (Alonso García González Director de Reporte Técnico Dra Alejandra García Hernández, 2013)

Tabla 2.

Persistencia de datos de las peticiones de la API de Twitter.

API	LIMITACION TEMPORAL(ESTANDAR)	LIMITACION TAMAÑO(PREMIUM)	DE
Streaming	Únicamente tiempo real	No aplica streaming	
Search	Menos de 7 días	1500 últimos tweets	
REST	NO	3200 últimos tweets	

Fuente : (Alonso García González Director de Reporte Técnico Dra Alejandra García Hernández, 2013)

El API para desarrolladores de Twitter no es lo suficiente confiable para efectuar una gran cantidad de búsquedas en el mismo día, desde el mismo usuario de autenticación (Ver Figura 10), para ello es necesario realizar una programación que evada la autenticación con varios usuarios y desde una IP diferente, lo anterior hace que los programas que se construyan alrededor de esta temática sean bastantes complejos. (Ver Tabla 3.)

Tabla 3.

Mensaje de error.

<i>Mensaje de error, para un parámetro de búsqueda en red social Twitter.</i>
<pre>Warning message: In doRppAPICall("search/tweets", n, params = params, retryOnRateLimit = retryOnRateLimit, : 1500 tweets were requested but the API can only return 7</pre>

Nota: Elaboración propia

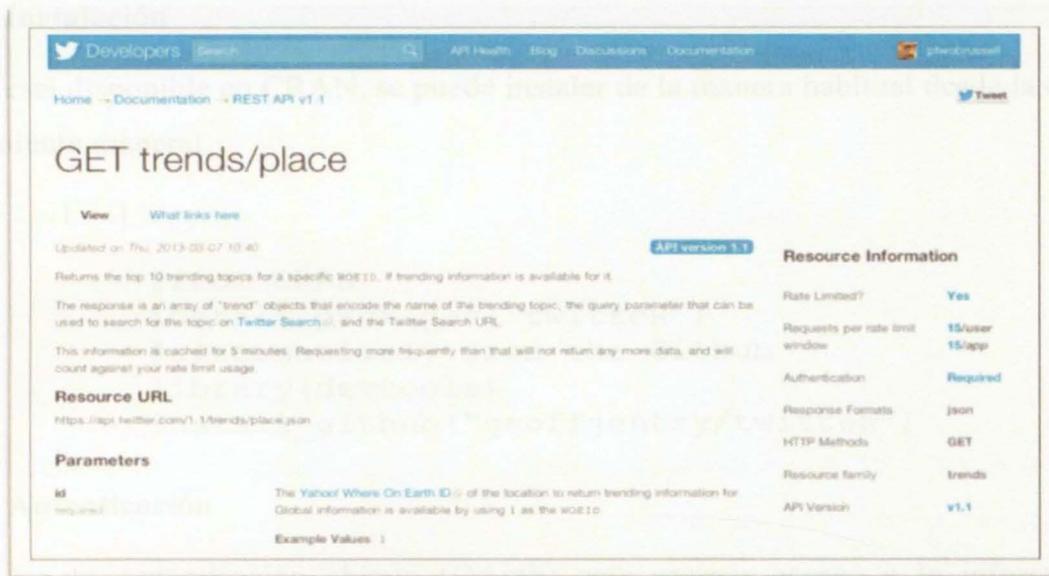


Figura 10. Los límites para los recursos de la API de Twitter se identifican en el documento en línea para cada petición de la API; el recurso API que se visualiza permite 15 solicitudes por "Ventana de límite de velocidad", que actualmente se define en 15 minutos. (Russell, 2015)

Nota: Elaboración propia

3.2.4. Preparación del código para realizar la aplicación

En una sesión R (para este caso se familiarizo con la herramienta RStudio) se tiene que instalar los siguientes paquetes: twitteR, Shiny y leaflet. Luego se debe que crear el código fuente en lenguaje R. Como la finalidad es verificar la eficiencia y la facilidad de obtener los datos de la red social Twitter, se colocará todo el código R necesario en un solo archivo, para la prueba se va a llamar geoprueba.R.

3.2.5. Twitter: Twitter basado en R

Twitter es un paquete R que proporciona acceso a la API de Twitter. La mayoría de las funcionalidades de la API son compatibles, direccionadas hacia las llamadas API que son más útiles en el análisis de datos que en la interacción diaria. (Jeff Gentry, 2015)

R usa la biblioteca twitteR, un cliente de Twitter que administra la comunicación con la API de Twitter. Este aplicativo fue creado por Jeff Gentry.

3.2.5.1. Instalación

TwitterR está disponible en CRAN, se puede instalar de la manera habitual desde la consola R, de la siguiente manera:

```
#from CRAN
install.packages("twitterR")
#alternatively from the Github
library(devtools)
install_github("geoffjentry/twitterR")
```

3.2.5.2. Autenticación

Twitter usa la Autenticación abierta (OAuth) para otorgar acceso a la información. La autenticación abierta es un método de autenticación basado en tokens. Las credenciales que se van utilizar son las siguientes, son únicas para cada uno de los aplicativos:

```
# OAuth authentication
consumer_key <- "zR8ZqSzGnYpJTD1WpeGvta3pY"
consumer_secret <- "r9z1mfHFeS2Pd0vWa51z7kJRCJsUOG3Ape0NCByC6lcST2Uy2S"
access_token <- "410813505-hEPXXIx0HLCc7v3A8rxIjRbh742Uu7i5VoFKfuny"
access_secret <- "Kkpz1yOrmyBvfeyckgERZNB6sBjOcSaG0SCHxxxG4h4n4"
```

Se utiliza la función `setup_twitter_oauth` para configurar la autenticación. La función `setup_twitter_oauth ()` toma las cuatro credenciales de twitter que se genera a partir de la API configurada anteriormente.

```
#set up to authenticate
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

3.2.5.3. Consulta de palabras claves

Realizar una consulta es cuestión de efectuar una pregunta a los servidores de Twitter. Para poder acceder a los datos que se requieren, se deben enviar consultas significativas a Twitter. Con Twitter, tenemos acceso a millones de datos con información de las tendencias más

significativas de las cuentas. En estas líneas de código se ejecutará la consulta haciendo uso del hashtag #bogota

```
shinyApp(
  ui = fluidPage (
    fluidRow (
      column (4, textInput ("searchkw", label = "search:", value = "#bogota")),
```

3.2.6. Librería para la creación de mapas

Leaflet es una de las bibliotecas de código abierto basado en JavaScript más populares para mapas interactivos. Es utilizado por sitios web que van desde The New York Times y The Washington Post a GitHub y Flickr, así como a especialistas de SIG como OpenStreetMap, Mapbox y CartoDB.

Este paquete de R facilita la integración y el control de los mapas (Vladimir Agafonkin, 2011).

3.2.6.1. Instalación

Leaflet está disponible en CRAN, se puede instalar de la manera habitual desde la consola R, de la siguiente manera:

```
install.packages("leaflet")
# Para instalar desde GitHub, ejecutar el comando
# devtools::install_github ("rstudio/leaflet")
```

3.2.6.2. Utilización del paquete Leaflet

Para la creación del mapa se debe seguir la secuencia lógica de los siguientes pasos, primero se crea un widget de mapa llamando `leaflet ()`. Posteriormente se agregan las capas (es decir, las características que se deseen) al mapa mediante el uso de funciones de capa (por ejemplo, `addTiles`, `addMarkers`, `addPolygons`). Una vez concluido este proceso, se puede visualizar el widget de mapa.

De la siguiente manera:

```
# Creación de un mapa en leaflet
mapTweets <- reactive ({
```

```
map = leaflet() %>% addTiles() %>%
  addMarkers(as.numeric(dataInput()$longitude), as.numeric(dataInput()$latitude), popup =
dataInput()$screenName) %>%
  setView(input$long, input$lat, zoom = 11)
})
output$myMap = renderLeaflet(mapTweets())
```

Para el ejemplo de verificación de la eficacia de la búsqueda y localización de un Tweet en la red social de Twitter por medio del hashtag #Bogota, se logra visualizar en la Figura 11 que se logra capturar un Tweet en la ciudad de Bogota con sus respectivas coordenadas, dicha búsqueda fue efectuada con un rango de acción de 10 km. Cabe resaltar que el API de Twitter limita las búsquedas de geolocalización de los usuarios, y estos datos dependen si el usuario tiene habilitado el GPS en el momento de subir el Tweet.

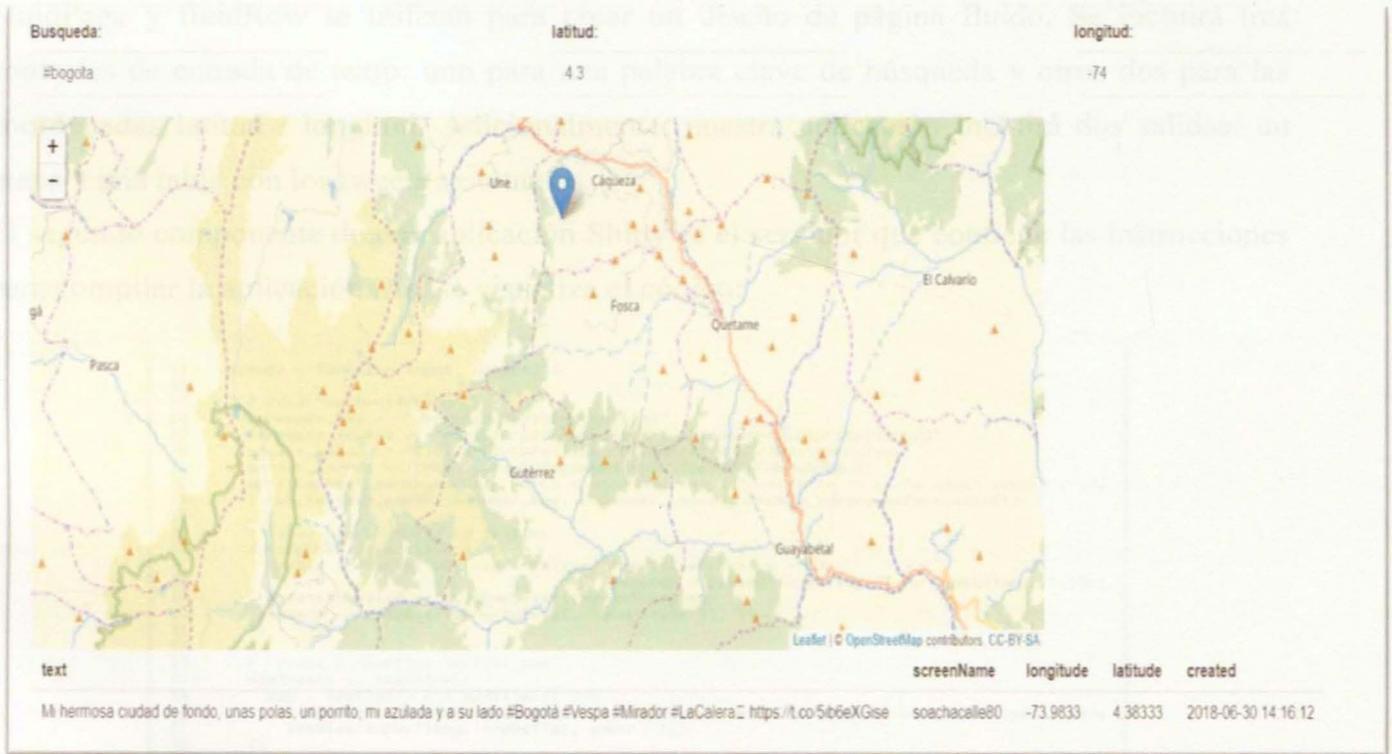


Figura 11. Localización del tweet con el hashtag #Bogotá

Nota: Elaboración propia

Una aplicación en Shiny tiene dos componentes. La primera parte, llamada UI, define la apariencia visual o la interfaz de usuario de la aplicación. Tal como se visualiza el código:

```

1 library(leaflet)
2 library(twitter)
3
4 shinyApp(
5   ui = fluidPage(
6     fluidRow(
7       column(4, textInput("searchkw", label = "search:", value = "#bogota")),
8       column(4, textInput("lat", label = "latitude:", value = 4.30)),
9       column(4, textInput("long", label = "longitud:", value = -74)),
10      column(8, leafletOutput("myMap")),
11      column(12, tableOutput('table'))
12    )
13  ),
14

```

Figura 12. Línea de código Apariencia visual #Bogotá.

Nota: Elaboración propia

FluidPage y fluidRow se utilizan para crear un diseño de página fluido. Se incluirá tres controles de entrada de texto: uno para una palabra clave de búsqueda y otros dos para las coordenadas latitud / longitud. Adicionalmente, nuestra aplicación incluirá dos salidas: un mapa y una tabla con los tweets resultantes.

El segundo componente de una aplicación Shiny es el servidor que contiene las instrucciones para compilar la aplicación. Así se visualiza el código:

```

15- server = function(input, output) {
16-
17-   # OAuth authentication
18-   consumer_key <- "zR8ZqSzGnYpJTDlwpeGvta3py"
19-   consumer_secret <- "p9z1mfHfEs2Pd0vwa51z7k JRC JsuOG3Ape0NCByC6IcST2uy2S"
20-   access_token <- "410813505-HEPXtX0HLc7v3A8rx1jRbh742Uu745vofKfuny"
21-   access_secret <- "KkpzlyOrmyBvfeYckgERZNB6sBj0cSaG0SCHxxxG4h4n4"
22-   options(httr_oauth_cache = TRUE) # enable using a local file to cache OAuth access crede
23-   setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
24-
25-   # Issue search query to Twitter
26-   dataInput <- reactive({
27-     tweets <- twListToDF(searchTwitter(input$searchkw, n = 10,
28-                                       geocode = paste0(input$lat, ",", input$long, ",10km")
29-                                       tweets$created <- as.character(tweets$created)
30-                                       tweets <- tweets[!is.na(tweets[, "longitud"])])
31-   })
32-
33-   # Create a reactive leaflet map
34-   mapTweets <- reactive({
35-     map = leaflet() %>% addTiles() %>%
36-       addMarkers(as.numeric(dataInput()$longitud), as.numeric(dataInput()$latitud), popu
37-       setView(input$long, input$lat, zoom = 11)
38-   })
39-   output$myMap = renderLeaflet(mapTweets())
40-
41-   # Create a reactive table
42-   output$table <- renderTable(
43-     dataInput(), c("text", "screenName", "longitud", "latitud", "created")
44-   )
45- }
46-

```

Figura 13. Línea de código Adquisición de Tweets en el mapa hashtag #Bogotá.

Nota: Elaboración propia

3.3. Capítulo III. Estableciendo los criterios de seguridad y búsqueda de información

Una vez el aplicativo se halla autenticado con el API de twitter, como se describió anteriormente, se realiza el código que efectuará la búsqueda en mapa. `Datainput` es una función reactiva que emite una consulta a Twitter basada en una cadena de búsqueda (`input $ searchkw`) proporcionada por el usuario. El objetivo de este apartado es conocer los tweets geo-codificados, es necesario proporcionar el radio de búsqueda y las coordenadas de latitud / longitud en el formato de latitud, longitud y radio. En esta aplicación de ejemplo, la cantidad máxima de tweets a devolver ($n = 10$) y el radio de búsqueda (10 km) se establecen por simplicidad, y por los mismos límites que coloca Twitter al realizar búsquedas que contengan gran envergadura.

La última línea de entrada de `datos: ! is.na ()` se usa para eliminar los tweets que no están geo-codificados.

Para finalizar, la creación del mapa y la tabla en función de los resultados de la consulta se verán almacenadas en `dataInput`. Las ventanas emergentes en el mapa visualizan el nombre de pantalla (es decir, el nombre de usuario de twitter) mientras que la tabla de salida contiene el tweet, la ubicación y la fecha en que se creó el tweet, además del nombre de usuario.

Se utilizaron fragmentos de línea de código del Ingeniero Ali Santacruz¹³ el cual fue de gran ayuda para contrastar la efectividad de la adquisición de Tweets. (Santacruz, 2018)

Las aplicaciones de mapas con Shiny son particularmente eficaces y pueden ser útiles para la abstracción de datos en redes sociales, tal y como se logró demostrar en el presente capítulo; que sumado a una metodología de detección de patrones y procesamiento estadístico podría arrojar información estratégica. Esta metodología será explicada en el capítulo 3.4.

¹³ Ali Santacruz es un geomático apasionado por las tecnologías de la información geográfica y la programación con R. Es ingeniero graduado de la Universidad Nacional de Colombia como Ingeniero Forestal (2004) y como Máster en Geomática (2010). (Santacruz, 2018)

3.3. Capítulo III. Estableciendo los criterios de selectividad y búsqueda de información en el ciberespacio para parametrizar un peligro geoestratégico para Colombia.

Uno de los elementos comunes para realizar las búsquedas en una red social y que hacen parte de una tendencia, reside en las características del hashtag¹⁴ #criteriodeselectividad, por lo tanto se utilizará como el mecanismo de búsqueda para recuperar los principales tweets y de esta manera realizar un posterior análisis. Pero antes de establecer los criterios de selectividad debe existir una metodología para concretar los peligros geoestratégicos a los que está expuesta Colombia frente a los entes internacionales.

Es así como desde el punto de vista militar, se describen las amenazas relacionadas con la Seguridad Nacional, las cuales se encuentran conexas con los peligros geoestratégicos para Colombia y dadas sus características estratégicas pueden verse expuestas por los diferentes elementos geopolíticos, económicos, sociales, tecnológicos o geográficos.

Para Colombia, siendo un territorio de gran extensión cobra vital importancia los espacios geográficos vacíos en los cuales el Estado no ejerce influencia, así como también el ciberespacio, el espacio marítimo, el espacio aéreo y ultraterrestre. En el contexto externo “internacional-estratégico”, estos espacios comunes en los que confluyen varios actores son objeto de pugna y confrontación.

En el 2018, el Twitter oficial del Departamento de Seguridad Español y DSN-Estrategia de Seguridad Nacional 2017 establece al ciberespacio, el espacio marítimo, el espacio aéreo y ultraterrestre como escenarios de vital importancia para el estudio geoestratégico de la seguridad y defensa. (Ver Figura 14)

¹⁴Los hashtags son como palabras clave que las personas utilizan para marcar el tema del contenido que están compartiendo en las redes sociales. Cada hashtag es transformado en un hipervínculo que dirigen la búsqueda de las personas, las cuales, también, marcaron su contenido con un hashtag en específico. Cuando selecciones una palabra para un contenido, este automáticamente es adicionado al grupo de búsqueda de ese tema. Los hashtags realizan la conexión con la información de otra página Web y ofrecen la posibilidad de la intertextualidad. (Luiza Drubscky, 2018)



Figura 14. Espacios comunes globales. Tomado de la cuenta de Twitter oficial del Departamento de Seguridad Español.

Es también relevante destacar la exposición de infraestructuras críticas¹⁵ a las amenazas, dado el impacto que ejercen para la provisión de los servicios fundamentales de la población Colombiana, sectores como eléctrico están catalogados como de muy alto impacto en la ICCN¹⁶, en razón a que sus procesos funcionales guardan una alta dependencia con las TIC. Otras plataformas críticas cibernéticas se encuentran distribuidas en el sector financiero, transporte e hidrocarburos, los cuales encuentran riesgos cibernéticos de alto impacto en la prestación de sus servicios esenciales, así mismo basan su operación en plataformas digitales que en la mayoría de los casos se conectan a internet. (Comando Conjunto Cibernético, 2017).

¹⁵ La globalización y el uso de las tecnologías de la información y las comunicaciones, así como las tecnologías de operación en todas las áreas del conocimiento y la sociedad, traen consigo nuevos retos, escenarios y riesgos en el ecosistema digital, los cuales pueden afectar la sostenibilidad y competitividad de los países. Es así, que Colombia desde el año 2011, ha visto la dependencia e importancia de las TICs en los servicios esenciales de la Nación y su infraestructura crítica, por lo tanto, mediante el CONPES 3701 de 2011 y el 3854 de 2016, estableció a la ICCN como un activo nacional a ser protegido de las amenazas en el ciberespacio, en aras de garantizar la sostenibilidad y prosperidad económica y social del país. (Comando Conjunto Cibernético, 2017)

¹⁶ Infraestructura Crítica Cibernética Nacional. Comando Conjunto Cibernético.

Estas amenazas se describirán desde el contexto de seguridad y defensa nacional (ver Figura 15), con el fin de llegar a establecer los criterios de selectividad en las búsquedas de la red social, y de esta manera agilizar los procesos de búsqueda de Ciberinteligencia¹⁷ para la obtención de insumos para la toma de decisiones, de la siguiente manera:



Figura 15. Amenazas para Seguridad Nacional.

Fuente : (Departamento de Seguridad Nacional, 2017)

3.3.1. Conflictos armados

Los conflictos armados siguen siendo una de las amenazas más latentes para la Seguridad Nacional, en especial para el contexto de tensión geopolítica que se sigue viviendo en Latinoamérica. Esto es debido al incremento de las capacidades de proyección militar, terrestre, aérea y naval, de diversos Estados como el de Venezuela, así como la preocupación por explotar

¹⁷ Ciberinteligencia: Es aquella inteligencia capaz de prevenir riesgos de activismo en las redes (fuentes abiertas) que derivan en ciberataques e incluso en ataques físicos, demostrando que no hay ya una barrera entre el mundo virtual y el real, en donde también se pueden percibir fenómenos de cibercrimen, ciberfraude, desprestigio reputacional y demás riesgos emergentes que se orquestan desde las redes sociales y en el Internet profundo. (Javier Santos, 2015)

las capacidades en otros dominios como el ciberespacio o el espacio aéreo y ultraterrestre. (Departamento de Seguridad Nacional, 2017)

Sumado a la problemática de Colombia de conservar un conflicto armado irregular interno por más de cincuenta años en donde se ven relacionados con un sinnúmero de factores. Entre ellos se encuentran la persistencia del problema agrario; la irrupción y la propagación del narcotráfico; las limitaciones y posibilidades de la participación política; las influencias y presiones del contexto internacional; la fragmentación institucional y territorial del Estado. (Javier Santos, 2015).

Ahora bien, en el ámbito internacional estos conflictos son cohesionados por mecanismos de influencia como el poder¹⁸ inteligente¹⁹, un ejemplo de ello son las amenazas relacionadas con la armas de destrucción masiva, armamento de alta precisión, proliferación de armamento de destrucción masiva, que combinados con el dominio del ciberespacio (“ciberataques”) pueden llegar a efectuar operaciones de influencia y desinformación a cualquier actor del Sistema Internacional. Estas amenazas se han denominado bajo el nombre de conflictos híbridos, aquellos que incorporan operaciones de información, subversión, presión económica y financiera junto a acciones militares. Estas acciones, perpetradas tanto por actores estatales como no-estatales, tienen por objeto la movilización de la opinión y la desestabilización política. (Departamento de Seguridad Nacional, 2017).

Así mismo, la inestabilidad del Estados frente a los Departamentos de frágil gobernanza, en zonas de frontera, proporcionan espacios para que diversos grupos armados ocupen estos territorios. Derivándose en una mayor exposición del Estado Colombiano frente a otras

¹⁸ El profesor Joseph Nye, de la Universidad de Harvard, es el creador del concepto de poder duro y blando, el término de poder suave, también llamado poder blando, se refiere a la capacidad de conseguir que los demás hagan lo que uno quiere sin recurrir a la represión o al dinero, por el contrario el poder duro es la habilidad de un actor para influenciar a otro por medio de amenazas o recompensas.

¹⁹ Poder inteligente, definido por el profesor Joseph Nye como la capacidad de un actor de combinar elementos de poder duro y blando de maneras que se refuerzan mutuamente de modo que los propósitos del actor avancen de manera efectiva y eficiente. Sostiene que el avance del poder inteligente se ha convertido en un imperativo de seguridad nacional, impulsado tanto por los cambios estructurales a largo plazo en las condiciones internacionales como por los fracasos a corto plazo de la actual administración. Los debates actuales sobre la diplomacia pública y el poder blando adolecen de fallas al abordar las dimensiones conceptuales, institucionales. (Wilson, 2008)

amenazas como el narcotráfico, la inmigración²⁰ o el terrorismo. Esto involucra mayores compromisos de colaboración y participación en operaciones internacionales, en especial con las relacionadas en proveer estabilidad fuera de sus fronteras. Por tal motivo el objetivo estratégico número uno del gobierno del Presidente Juan Manuel Santos durante el lapso 2015-2018 fue la de contribuir con las capacidades de la Fuerza Pública a la terminación del conflicto y la construcción de la paz. (Alberto José Mejía Ferrero et al., 2017)

3.3.2. Terrorismo

El terrorismo, fundamentalmente de carácter subversivo, ha ocupado escenarios cada vez más complejos. El terrorismo subversivo del ELN²¹ proyecta su ideología radical y actúa a nivel global, donde ha protagonizado interminables atentados. Se han caracterizado por su vertiginosa versatilidad y su adaptación a los cambios y estrategias seguidas contra ellos.

Los atentados en lugares de concentración de la Fuerza Pública, medios de transporte o infraestructuras críticas nacionales, en especial al sector petrolero, representan una tendencia o patrón actual del terrorismo del ELN, que opera en pequeños grupos organizados como de manera individual, con el fin de producir el mayor número de víctimas y un fuerte impacto mediático dentro de la población colombiana.

Aunque el número de atentados terroristas ha presenciado en los colombianos una disminución en los últimos años, siguen siendo para el Estado un problema del cual no ha sido fácil de solucionar. La Figura 16 visualiza la disminución de los ataques terrorismo desde el año 2014, producto de las negociaciones del Estado Colombiano con el extinto grupo terrorista FARC. (Carlos et al., 2018)

²⁰ Durante los últimos cuatro años, pero con especial intensidad en 2017, se ha experimentado una masiva inmigración de venezolanos hacia Colombia. Aunque no existe una cifra exacta sobre el número de dichos inmigrantes, la oficina de Migración Colombia ha estimado que serían unas 800.000 personas. (Departamento de Seguridad Nacional, 2017)

²¹ El ELN es una abreviatura de su nombre completo “Ejército de Liberación Nacional”, y en pocas palabras es una de las tantas organizaciones de guerrilla colombiana que hace presencia principalmente en su país. Esta guerrilla organizada también se conoce como un pensamiento marxista-leninista, así como pro revolución cubana.



Figura 16. Ataques Terroristas en Colombia.

Fuente: Ministerio de Defensa Nacional.

El desarrollo tecnológico se ha convertido en el eslabón más dinámico para el acceso a los recursos del grupo terrorista ELN. Utilizando la desinformación masiva y el empleo de redes sociales para adoctrinar y evadir sus responsabilidades frente a actos terroristas, evidenciado con la constante difusión de propaganda terrorista y propagación de formas de radicalización.(Alberto José Mejía Ferrero et al., 2017)

3.3.3. Crimen Organizado

El crimen organizado es una amenaza de naturaleza transnacional, versátil y multidimensional²². Consistente en un fenómeno con capacidad desestabilizador, que por su naturaleza destructiva debilita el Estado, perjudicando la estabilidad económica. Entre sus

²² La Seguridad Multidimensional del Hemisferio (incentivada por la Organización de los Estados Americano - OEA) incorpora una combinación de problemáticas políticas, económicas, medioambientales, militares y de seguridad humana, como respuesta a las múltiples amenazas transnacionales como el crimen transnacional organizado. (Stein, 2009)

características se pueden mencionar los tipos relacionados con los tráficos ilícitos (narcotráfico), además del contrabando, testaferrato y el uso de paraísos fiscales²³.

En Colombia se han emitido los lineamientos de política del Ministerio de Defensa Nacional para enfrentar a los Grupos que se caracterizan como Grupos Armados Organizados (GAO), a través de la Directiva permanente 0015 del 22 de Abril de 2016.(MINISTERIO DE DEFENSA NACIONAL, 2016). Esta directiva reemplaza la anterior directiva permanente 0014 de 2011, y reconoce el crecimiento de las Bandas Criminales (BACRIM), y las reclasifica en Grupos Delictivos Organizados (GDO) y Grupos Armados Organizados (GAO). La Figura 17 visualiza los lineamientos de la D.P 0015 del Ministerio de Defensa Nacional.

El desenvolvimiento del conflicto armado en Colombia, constata como este fenómeno tiene estrecha vinculación internacional con el crimen organizado y el terrorismo, lo que fortalece sus efectos y consecuencias sobre la población. El crimen organizado potencializa, otras amenazas a la seguridad y defensa amplificado su poder a través del empleo de la tecnología, recurso creciente para desarrollar actividades delictivas.



Figura 17. Visualización de los lineamientos de la D.P 0015 del Ministerio de Defensa Nacional.

Nota: Elaboración propia

²³ Los ‘Panama Papers’, fue un escándalo conocido en el mundo porque destapaba la magnitud de la fuerte presencia que mantenían compañías, políticos y personas naturales en los paraísos fiscales, ha tenido ya su efecto económico en Colombia, pues le ha permitido a la bolsa de los impuestos nacionales un recaudo de 85.000 millones de pesos, recursos con los que no contaba el país. (Martha Morales Manchego, 2018)

Colombia está expuesta al crimen organizado y a sus diversas manifestaciones, en gran medida, a su posición geoestratégica como punto de acceso a Centroamérica, es un paso obligado para los migrantes hacia Estados Unidos o países del Hemisferio Sur.

Por tal motivo el objetivo estratégico número cuatro del gobierno del Presidente Juan Manuel Santos durante el lapso 2015-2018 consiste en combatir las nuevas y tempranas expresiones de crimen organizado que amenacen la seguridad y el funcionamiento transparente del Estado, usando todas las capacidades de la Fuerza Pública. (Alberto José Mejía Ferrero et al., 2017).

3.3.4. Espionaje y Ciberespionaje

El espionaje fue durante la guerra fría el principal elemento para anticiparse a los peligros estratégicos por parte del Estado, en la actualidad esta categorizada como una amenaza de primer orden para la seguridad, la cual se ha visto modernizada velozmente a las alternativas que ofrece la tecnología. En este sentido, el ciberespacio establece quizás el papel más relevante a nivel de espionaje y es utilizado por todos los Estados del Sistema internacional, grupos o individuos, que mediante plataformas y capacidades sofisticadas se valen de programas que proporcionan acceso a grandísimos volúmenes de información o datos sensibles para su beneficio.

El espionaje industrial²⁴, tiene como objetivo acceder al conocimiento tecnológico y estratégico de la competencia, acogiendo una posición dominante con respecto a la otra. Todos los días del año, las empresas soportan de forma regular este tipo de ataque, la cual busca quebrantar el sistema económico y afectar al bienestar de los ciudadanos. En vista de este fenómeno desfavorable para unos, pero favorable para otros, los gobiernos han empezado a efectuar un enfoque colaborativo entre el sector público y privado. Es así, como en los últimos años se ha manifestado un incremento de agresiones procedentes de los servicios de inteligencia

²⁴ Según los expertos en seguridad nacional e informática originarios de la India, el espionaje industrial se considera una prioridad para el ciberespionaje chino. La razón es que reduciría gastos y tiempo en el esfuerzo chino para construir un ejército e industria militar modernos. (Ministerio de defensa de España, 2009)

extranjeros contra los intereses nacionales, y que se materializan tanto a través de procedimientos clásicos como de inteligencia en el ciberespacio(ciberespionaje²⁵).

3.3.5. Amenazas en el Ciberespacio

De acuerdo al contexto Español (Departamento de Seguridad Nacional, 2017) el ciberespacio, es aquel espacio accesible a cualquier Estado del Sistema Internacional, así como también, a cualquier particular que disponga de un computador y una conexión a Internet; este servicio es inconcebible de excluir, dadas las connotaciones de hiper-conectividad a las que nos enfrentamos en la actualidad.

Es así, como este escenario se caracteriza por no tener fronteras físicas, ausencia general de la soberanía y jurisdicción por parte de los Estados, en la actualidad es complejo establecer la atribución de las acciones delictivas en el ambiente Ciber, debido a su débil regulación. Por tal motivo, las amenazas inmersas en el ciberespacio son de índole fundamental para la Seguridad Nacional.

La regulación, una adecuada gestión del ciberespacio, la cooperación internacional y la asistencia mutua es vital para contrarrestar los fenómenos que afectan este escenario.

La utilización del ciberespacio como medio para la ejecución de actividades ilícitas, acciones de desinformación, propaganda mediática (Fake news) y actividades producto del crimen organizado, recaen directamente sobre la Seguridad Nacional, evidenciando el complejo contexto al que nos enfrentamos. (Figura 18 Vulnerabilidades del ciberespacio)

²⁵ El ciberespionaje persigue información de altísimo valor y debe considerarse una amenaza crítica. (Candau, Ciberseguridad, & Centro, 2015)



Figura 18. Vulnerabilidades del ciberespacio.

Fuente : (Departamento de Seguridad Nacional, 2017)

3.3.6. Soberanía Territorial

3.3.6.1. Soberanía en el Espacio Marítimo

El espacio marítimo, es de gran importancia para Colombia, es un país privilegiado al poseer dos mares (Mar caribe y Océano Pacífico) y dos costas que la comunican con el mundo. Su territorio marítimo y costero corresponde al 45% del total del territorio.(CERÓN, 2012).

Dadas sus connotaciones geoestratégicas, Colombia reside como un país con un gran valor estratégico para la comunidad internacional.

El espacio marítimo además de ser por si solo un mecanismo de influencia, posee elementos característicos como los son las rutas marítimas, las cuales son vitales para efectuar las actividades comerciales y de transporte. Igualmente, el aprovisionamiento energético es basado en gran parte por este medio, y la interconexión de información digital de Colombia con el planeta se efectúa por cables submarinos.

Los factores que amenazan la seguridad marítima se agrupan principalmente en acciones como la piratería, el terrorismo, el narcotráfico, actos contra la conservación del patrimonio cultural

subacuático²⁶, la inmigración irregular por vía marítima o la explotación incontrolada de los recursos marinos, como por ejemplo la pesca.

De la misma forma, y acuerdo a textos de Cadena, W. R., & Devia, C. A. (2012), la delimitación de espacios marítimos entre Estados es un problema frecuente que puede generar ambientes de crisis o conflicto debido a la trascendencia geopolítica que ejerce el acceso y control de estos espacios. Para Colombia, sus fronteras marítimas fueron reguladas a partir de tratados bilaterales suscritos, fundamentalmente, en las décadas de 1970 y 1980. Esta situación se complementa con la no ratificación de la Convemar tanto por Colombia como por parte de otros estados vecinos y con el surgimiento de diferendos limítrofes que han llevado, en algunos casos, a escenarios de confrontación y procesos internacionales. Colombia afronta dos diferendos marítimos en esta materia con Venezuela y Nicaragua.(Afanador, 2012).

3.3.6.2. Soberanía Espacio Aéreo y ultraterrestre

El espacio aéreo es un escenario del teatro de operaciones inmiscuido en todo tipo de violaciones a la seguridad por parte de actores estatales y no estatales. Los descatos contra las normas de aviación, las violaciones a los sistemas de control de navegación o el tráfico ilícito por medio del espacio aéreo son ejemplos de cómo este recurso se ha puesto a disposición de organizaciones fraudulentas para aprovechar su potencial.(Departamento de Seguridad Nacional, 2017)

Del mismo modo, con la incursión de las aeronaves no tripuladas (UAV²⁷), ya sea para ejecutar operaciones ofensivas o realizar actividades ilícitas por parte de Estados u organizaciones no estatales, evidencia la importancia del gobierno Colombiano para proteger el espacio aéreo.

Por tal motivo el objetivo estratégico número cinco del gobierno del Presidente Juan Manuel Santos durante el lapso 2015-2018 consiste en garantizar la soberanía e integridad del territorio nacional, protegiendo los intereses nacionales. (Alberto José Mejía Ferrero et al., 2017).

²⁶ En 1708 Galeón San José, de bandera española cayó y naufragó bajo el fuego de los navíos ingleses que lo acechaban, estaba a aproximadamente a 19 kilómetros de Cartagena de Indias, Colombia, a cercanías de la isla de Barú, se cree está a 250 metros de profundidad, y que al momento de su hundimiento llevaba una carga de especias, telas, oro y plata, con un estimado de valor entre 3 y 5 mil millones de dólares. (SALAZAR, 2013)

²⁷ UAV (Unmanned Aerial Vehicle): Vehículo aéreo no tripulado.

3.3.6.3. Protección y Defensa para la Infraestructuras crítica cibernética de Colombia (ICCN)

Derivado del Plan Nacional para la protección y Defensa de la ICCN como producto del resultado de un trabajo intersectorial, coordinado y articulado con el fin de buscar la resiliencia cibernética para Colombia, el gobierno enfoca sus esfuerzos para mitigar todo riesgo que atente contra la infraestructura estratégica del país.

Estas infraestructuras se encuentran compuestas de instalaciones de grandes proporciones, redes, sistemas integrados, equipos físicos, sistemas computacionales y redes de comunicaciones, en cuya responsabilidad reside la operación y el funcionamiento de los servicios básicos para la población de Colombia. Estos servicios son obligatorios para las funciones básicas de salud, seguridad, bienestar social y económico de la Nación. (Figura 19 Sectores Estratégicos ICCN)



Figura 19. Sectores Estratégicos ICCN.

Fuente : (Departamento de Seguridad Nacional, 2017)

Al haber agrupado las amenazas de acuerdo a una priorización basada en los objetivos estratégicos que se ejercen a nivel de Seguridad Nacional mediante el Ministerio de defensa, se

pueden concatenar tres tablas dependientes la una con las otras, para que se sometan los criterios de búsquedas más relevantes y de esta manera efectuar la minería de datos enfocada en los peligros geoestratégicos descritos en la Tabla 4.

Tabla 4.
Criterios de selectividad de peligros Geoestratégicos.

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
Conflicto Armado	Proyección militar	#VenezuelaAntiimperialista #OperaciónPatriaSegura #NoALaInvasionExtranjera #ProyectoCassandra #Nuclear #FANB #PatriaGrande #MTRRVenceremos #mtrrvenceremosumbv #195AniversarioArmadaFANB Tags <ul style="list-style-type: none"> • Artillería pesada + armamento + frontera • Enviando armamento a Nicaragua • Maduro +moviliza armamento + bélico • Armamento + aviones combate + Venezuela hacia la frontera • Venezuela misiles + armamento antiaéreo • Armamento ruso
	Narcotráfico	#IndependenciaYPaz2018 #cocainavenezuela #Sinaloa

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
	Tráfico Ilícito	<p>#TareckElAissami</p> <p>#DiosdadoCabello #Cartel de los Soles #corrupto</p> <p>Tags</p> <ul style="list-style-type: none"> Venezuela + Colombia + rutas aéreas y marítimas + transportar droga + Honduras
	Armas de Destrucción Masiva	<p>#AcciónMultidimensionalFANB</p> <p>#DefensaMultidimensional</p>
	Armamento de alta precisión	#AcciónMultidimensionalFANB
	Ciberataques	<p>#cibeataques</p> <p>#Seguridad #Hackers</p>
	Inmigración venezolanos	<p>#Inmigración #Venezolanos</p> <p>#inmigrantes #estadosunidos</p> <p>#venezolanos #visas</p>
	Invasión militar Venezolana, Nicaragüense	<p>#YOSIQUIEROQUEINTERVENGAN</p> <p>#UnidadDeLosPatriotas</p> <p>#UnionCivicoMilitarPorLaPatria</p> <p>#IndependenciaONada</p> <p>#VenezuelaAzul</p> <p>Tags</p> <ul style="list-style-type: none"> Nicolás Maduro rechaza Intervención EE.UU

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
Terrorismo	Terrorismo ELN, BACRIM	<p>#TumacoEsEl HaitíCbiano</p> <p>#Catatumbo</p> <p>#CatatumboEnPaz</p> <p>#Catatumboresiste</p> <p>#Hezbollah</p> <ul style="list-style-type: none"> • Ataques de la guerrilla izquierdista del ELN
	Atentados al sector petrolero	<p>#MediosRegistranACP</p> <p>Tags</p> <ul style="list-style-type: none"> • Atentados + sector petrolero
Crimen Organizado	Crimen Transnacional	<p>#NATO</p> <p>Tags</p> <ul style="list-style-type: none"> • Crimen organizado transnacional • Crimen transnacional • Control del extractivismo estratégico • Coca • Cargamentos de coca
	Contrabando	<p>#SistemaModal</p> <p>#Venezuela a #Colombia</p> <p>Tags</p> <ul style="list-style-type: none"> • Contrabando de extracción desde • #Contrabando con el Lago de Maracaibo • Contrabando de gasolina
	Testaferrato	<p>Tags</p> <ul style="list-style-type: none"> • FARC + testaferrato + supermercados • TESTAFERRATO + FERROCARRIL PACIFICO

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
	Paraísos Fiscales	#PanamáPapers #EstafaPro #PanamaPapers #PuntoHerenciaFranco #Odebrecht #PanamaPapers #wikileaks #ParadisePapers
	Migración de Haitianos; Venezolanos	#Inmigracion #Haiti #Venezuela #Soros #Haitianos #ColoquioEuroLatinoamericano2018 #GrecolAlc
Espionaje y Ciberespionaje	Espionaje Industrial	#DefensaDemocracia
	Inteligencia en el ciberespacio (Ciberinteligencia)	#cibersegurança #ciberseguridad #ciberinteligencia #OSINT #cracker #cibercultura #ciberintelligence #hubots #bots #perfilesfalsos #Daesh
Amenazas en el ciberespacio	Fake news	#fakenews #AlertaDeFakeNews #NoMasFakeNews
	Robo de datos	Tags <ul style="list-style-type: none"> • Colombia + actividades piratería informática
	Ransomware	#Ransomware

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
	DDOS (Ataques de denegación de servicios)	<p>#DdoS #Attack #Schema</p> <p>#CyberSecurity #Zombie #Tech</p> <p>#Technology #CyberAttack #Hack</p> <p>#Breach #Threat #CyberWarfare</p> <p>#Malware #Ransomware</p> <p>#Cyberwarning #Phishing #SpyWare</p>
	Hack de dispositivos móviles en los hogares	<p>#hacksmartphone</p> <p>#RussianHacker</p> <p>#RussianBear</p>
	Ciberataques ICCN	<p>#ciberataques</p> <p>#SCI (Sistemas de Control Industrial)</p> <p>#infraestructura +</p> <p>#ciberataques -</p> <p>#sci #Dos #DdoS</p> <p>Tags</p> <ul style="list-style-type: none"> • Ciberataques + infraestructura crítica • Ciberataques + sabotear infraestructura crítica
Soberanía Espacio Marítimo territorial y	Rutas marítimas	<p>@ANI_Colombia.</p> <p>Tags</p> <ul style="list-style-type: none"> • Colombia + rutas marítimas. • Rutas aéreas + marítimas + distribución de estupefacientes.
	Cables Submarinos	<p>#MásConectadosQueAntes</p> <p>#MinTICRindeCuentas</p> <p>#asambleafedesoft</p> <p>Tags</p> <ul style="list-style-type: none"> • Cables submarinos de fibra óptica

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
	Plataformas Offshore	<p><u>#offshore</u></p> <p><u>#offshore #petróleo #Colombia</u></p> <p><u>#Colombia #Oil #Gas</u></p> <p><u>#Wind #Windfarm #CleanEnergy</u></p> <p><u>Tags</u></p> <ul style="list-style-type: none"> • Plataformas + Chuchupa • Plataformas petrolíferas ‘offshore’
	Transporte Marítimo	<p><u>#TransporteMaritimo</u></p> <p><u>#MaritimeDay #Transport #Logistics</u></p> <p><u>#Colombia</u></p> <p><u>Tags</u></p> <ul style="list-style-type: none"> • Comercio y transporte Marítimo • Transporte Marítimo + Fluvial + Industria Colombia
	Piratería	<p><u>#Pirateria</u></p> <p><u>#Antipirateria</u></p> <p><u>Tags</u></p> <ul style="list-style-type: none"> • OTAN + fortalecer + piratería de transnacionales • Piratería + México o Colombia
	Patrimonio cultural subacuático	<p><u>#Hilo #GaleonSanJose</u></p> <p><u>#SalvemoselGaleon</u></p>
	Pesca	<u>#PescaArtesanal San Andrés</u>

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
	Secesión de Territorio	<p>#Meridian84</p> <p>#Miskitos</p> <p>#CornIslands</p> <p>#Colombia #ELN #Fallos #CJ.</p> <p>#REVISIONdelFallo2012CU #SAI</p> <p>#EstatutoRaizal</p> <p>#SanAndresArchipelago</p> <p>#colombia #sanandres #providencia</p> <p>#sanandrésarchipelago</p> <p>Tags</p> <ul style="list-style-type: none"> • Separation of San Andres Islands & Old Prvidence and Santa Catalina from Corn Islands & Moskito Cays • Separación Antioquia + San Andres islas de Colombia
	Tráfico ilícito Aéreo	<p>Tags</p> <ul style="list-style-type: none"> • rutas aéreas + marítimas + distribución de estupefacientes.

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
Infraestructura Ciber Cibernética Nacional	UAV, Drones	<p>#drones #Colombia</p> <p>#drones/#uav/#rpas</p> <p>#Ejército de Tierra #rpas/#drones/#uav</p> <p>#aerialphotography</p> <p>#dronepilot</p> <p>#ABMS (Advanced Battle Management System)</p> <p>Tags</p> <ul style="list-style-type: none"> • Eradicate coca + drones • Aparatos voladores no tripulados + fotografía + fumigación + ganadería + seguridad + control movilidad. • DRONES + ISRAEL • Drones + erradicación + cultivos ilícitos

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
Infraestructura Crítica Cibernética Nacional	Sector Agua	<p><u>#Hidroituango</u></p> <p><u>#HidroituangoEsUnDesastre</u></p> <p><u>#RioCauca #HidroItuango</u></p> <p><u>#ProyectoItuango</u></p> <p><u>#RiosVivosCol</u></p> <p><u>#NaturalZone #Hidroeléctrica</u></p> <p><u>#Colombia #Bosque #Aqua</u></p> <p><u>#Ambiente</u></p> <p><u>#Represas #CañondelCauca</u></p> <p><u>#SuspendanHidroituango</u></p> <p>Embalse del <u>#Guavio</u> <u>#Gachalá</u></p> <p><u>#Cundinamarca.</u></p> <p><u>Tags</u></p> <ul style="list-style-type: none"> • Colapso en hidroeléctrica • Hidroeléctrica + Sogamoso • Nivel + embalses + generación Hidroeléctrica + Colombia • Hidroeléctrica + Quimbo
	Sector Energía	<p><u>Tags</u></p> <ul style="list-style-type: none"> • Ataques + comunidad + Yopal. • Ataques + torres + energía eléctrica • Torre + energía + Guajira • Ataques + Oleoductos + Torres de Energía + Infraestructura vial. • Ataques terroristas + energía Tumaco • Ataques + energía eléctrica + gas

Ver (Figura 10. Diagrama de Flujo Metodología para el análisis de anticipación de peligros por amenazas).

Criterios de Selectividad en Redes Sociales		
Amenaza y Vulnerabilidad	Palabra Clave	Hipervínculo – Hashtag a la Red Social Twitter
	Sistema de Transporte en la frontera	<p>#Ureña</p> <p>#EnVillaDelRosario</p> <p>#TRANSPORTE</p> <p>Tags</p> <ul style="list-style-type: none"> • Problema transporte, ciudadanos frontera • Colombia + Venezuela transporte Binacional + frontera • transporte + carga + frontera con Colombia • CIERRE + FRONTERA + COLOMBIA • Transporte + ganado + frontera + Ecuador y Colombia.

Nota: Elaboración propia

3.4. Capítulo IV. Elaborar la metodología de análisis de los datos recopilados para el procesamiento de resultados, que permitan generar un comportamiento tendencial para llevar a cabo una alerta para la toma de decisiones.

De acuerdo a la metodología SEMMA (Sample, Explore, Modify, Model, Assess), argumentada previamente en el marco teórico, se modelará la secuencia lógica para el procesamiento de los datos recopilados en las redes sociales, con el fin de desglosar y analizar los patrones tendenciales que permitan la construcción de alertas para la toma de decisiones estratégicas.

Ver (Figura 20. Diagrama de Flujo Metodología para el modelo analítico de anticipación de peligros geoestratégicos).

Metodológicamente, se distribuirá el proceso mediante la ejecución de tres fases, establecidas así:

- A. Muestreo (Sample) y Exploración (Explore)
- B. Manipulación (Modify)
- C. Modelado (Model)
- D. Valoración (Assess)

3.4.1. Muestreo (Sample) y Exploración (Explore)

El muestreo y la exploración es la fase inicial del proceso de recolección de la información relevante alojada en la plataforma social de Twitter, se debe tener en cuenta los criterios de selectividad usados en el Capítulo III para lograr establecer los parámetros de búsqueda más acertados para la prevención de peligros geoestratégicos.

Teniendo en cuenta los parámetros de selectividad, el siguiente paso consiste en la abstracción de datos de la Red social Twitter mediante el API de desarrolladores descritos en el Capítulo II.

- A. El parámetro de selectividad escogido es: #VenezuelaAntiimperialista
- B. Se efectúa una muestra de diez interacciones, $\rho = 10$, para el ejemplo. Para efectuar minería de datos, es necesario realizar una captura de 1500 tweets, esto debe realizarse con la compra de la versión "PREMIUM" de la API de Twitter. Pero para ello es imperativo trabajar sobre una VPN (*Virtual Private Network*) que constantemente cambie la IP de la fuente, y por supuesto cambie el método de autenticación con la plataforma Twitter.
- C. Mediante las librerías de RStudio, se almacenan los datos obtenidos de la red social Twitter en una base de datos en "RSQLite²⁸", la cual permitirá organizar todos los campos de consulta que en un futuro se deseen contrastar.

²⁸ Incorpora el motor de base de datos 'SQLite' en R y proporciona una interfaz que cumple con el paquete

'DBI'. (Kirill Müller, 2018)

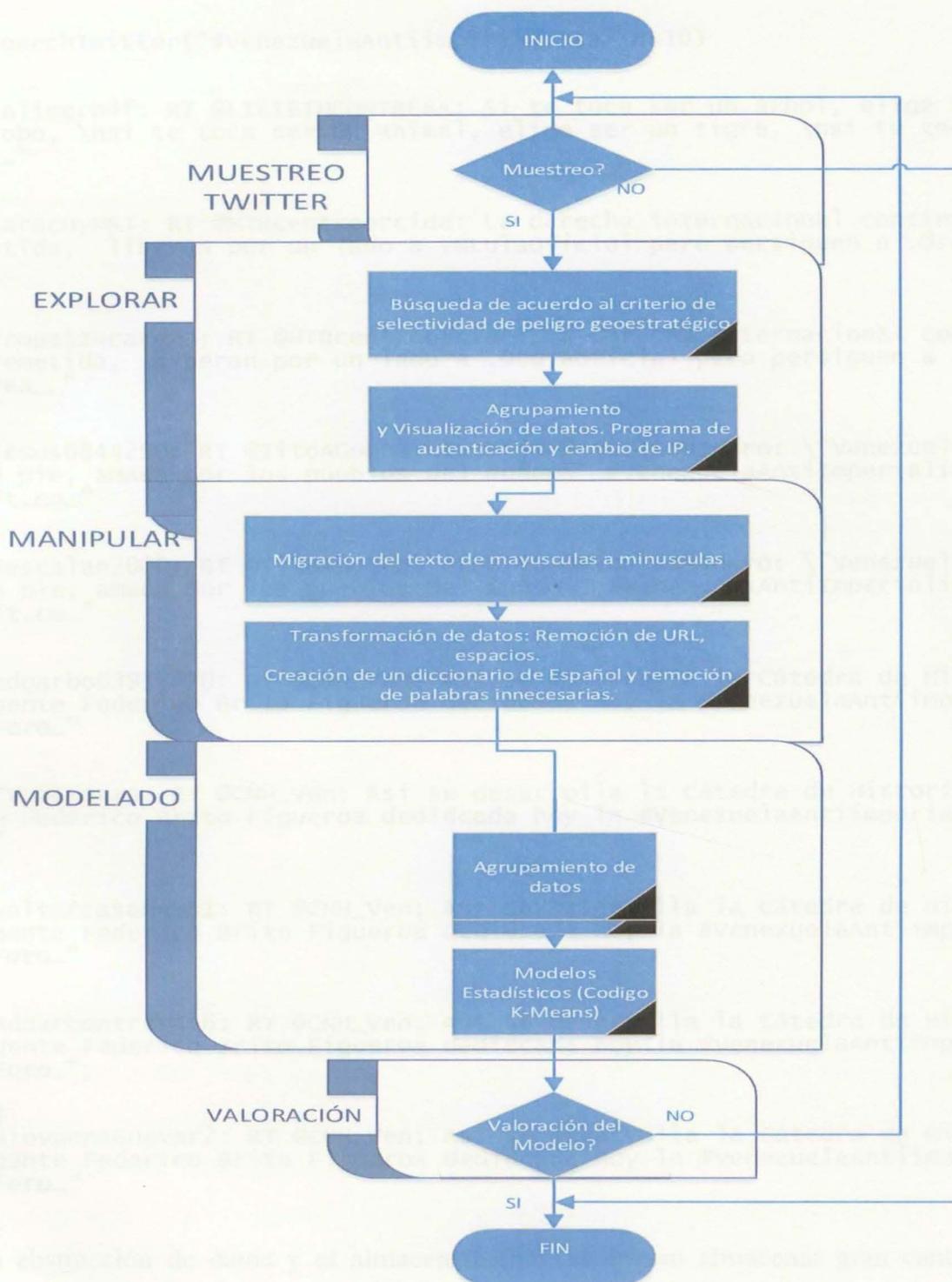


Figura 20. Diagrama de Flujo Metodología para el modelo analítico de anticipación de peligros geoestratégicos.

Nota: Elaboración propia

```
a <- searchTwitter("#VenezuelaAntiimperialista",n=10)
```

```
[[1]]
[1] "calimero4f: RT @LILIEETHCONTRERA: Si te toca ser un árbol, elige ser un
algarrobo, \nsi te toca ser un animal, elige ser un tigre, \nsi te toca ser
una pi..."
```

```
[[2]]
[1] "YaracuyMRT: RT @MTRcentrooccide: La derecha internacional continua su
arremetida, liberan por un lado a .@LulaOficial pero persiguen a .@rafaelc
orraa..."
```

```
[[3]]
[1] "tropaSINcarea: RT @MTRcentrooccide: La derecha internacional continua
su arremetida, liberan por un lado a .@LulaOficial pero persiguen a .@rafa
elcorrea..."
```

```
[[4]]
[1] "jesus6844290: RT @TitoAGomez: #LoDijo @NicolasMaduro: \"Venezuela segu
irá de pie, amada por los pueblos del mundo\" #VenezuelaAntiImperialista ht
tps://t.co..."
```

```
[[5]]
[1] "aescaln2000: RT @TitoAGomez: #LoDijo @NicolasMaduro: \"Venezuela segu
irá de pie, amada por los pueblos del mundo\" #VenezuelaAntiImperialista ht
tps://t.co..."
```

```
[[6]]
[1] "edgarbo63957770: RT @CNH_Ven: Así se desarrolla la Cátedra de Historia
insurgente Federico Brito Figueroa dedidcada hoy la #VenezuelaAntiimperiali
sta. Foro..."
```

```
[[7]]
[1] "TVLecturas: RT @CNH_Ven: Así se desarrolla la Cátedra de Historia insu
rgente Federico Brito Figueroa dedidcada hoy la #VenezuelaAntiimperialista.
Foro..."
```

```
[[8]]
[1] "waltercasamay01: RT @CNH_Ven: Así se desarrolla la Cátedra de Historia
insurgente Federico Brito Figueroa dedidcada hoy la #VenezuelaAntiimperiali
sta. Foro..."
```

```
[[9]]
[1] "eduarcontreras6: RT @CNH_Ven: Así se desarrolla la Cátedra de Historia
insurgente Federico Brito Figueroa dedidcada hoy la #VenezuelaAntiimperiali
sta. Foro..."
```

```
[[10]]
[1] "GiovannaGuevar2: RT @CNH_Ven: Así se desarrolla la Cátedra de Historia
insurgente Federico Brito Figueroa dedidcada hoy la #VenezuelaAntiimperiali
sta. Foro..."
```

Con la abstracción de datos y el almacenamiento, se logran almacenar gran cantidad de metadatos que fortalecen futuras búsquedas en el ciberespacio, como, por ejemplo, posición geográfica, sistema operativo del dispositivo, hora y fecha de escritura del Tweet, entre otras. (Ver Figura 21. Metadatos de un Tweet.)

created	double (S3: POSIXct, POSIXt)	2018-07-14 21:18:03
favoriteCount	double [1]	0
favorited	logical	FALSE
id	character [1]	'1018243249130131460'
isRetweet	logical	TRUE
latitude	character [0]	
longitude	character [0]	
replyToSID	character [0]	
replyToSN	character [0]	
replyToUID	character [0]	
retweetCount	double [1]	6
retweeted	logical	FALSE
screenName	character [1]	'YaracuyMRT'
statusSource	character [1]	'Twitter for Android ...
text	character [1]	'RT @MTRcentrooccide: La derecha internacional continua su arremetida. liberan ...
truncated	logical	FALSE
urls	list [0 x 4] (S3: data.frame)	A data.frame with 0 rows and 4 columns
refClassDef	S4 (methods::refClassRepresentati	S4 object of class refClassRepresentation
self	S4 [1] (twitter::status)	
created	double (S3: POSIXct, POSIXt)	2018-07-14 21:18:03
favoriteCount	double [1]	0
favorited	logical	FALSE
getClass	function (methods::refMethodDef)	function(...) { ... }

Figura 21. Metadatos de un tweet.

Nota: Elaboración propia

3.4.2. Manipulación (Modify)

La manipulación de los Datos, requiere de una pericia en el ámbito de la programación, puesto que es imperativo comenzar a efectuar un procesamiento y análisis de los datos adquiridos, con el fin de validar con más certeza que la información de salida sea realmente relevante para generar un producto de inteligencia estratégica. Con base en lo anterior es necesario efectuar los siguientes pasos:

A. Migrar el texto recolectado de mayúsculas a minúsculas, cuando este sea necesario.

Utilizando el paquete toLower en las librerías de R, se puede hacer uso de esta sencilla conversión, la cual facilitará la visualización de los textos. (Figura 22 Línea de Código para la conversión de mayúsculas a minúsculas)

```

b <- searchTwitter("#maburro",n=150)
df=twListToDF(b)
date<-Sys.Date()
date<-as.character(date)
name<-paste(date, ".RData")
#Guardamos en .RData
save(df, file =name)
#Creacion de un corpus:
library(tm)
# Construccion de un corpus, y se especifica el vector para los caracteres
myCorpus <- Corpus(VectorSource(df$text))
# Convirtiendo a Minusculas
myCorpus <- tm_map(myCorpus,
                    content_transformer(tolower))

```

Figura 22. Conversión de mayúsculas a minúsculas.

Nota: Elaboración propia

En el primer recuadro podemos observar las palabras en mayúsculas. (Ver Tabla 5) En el segundo recuadro se observa cuando se le ha aplicado la función toLower para convertir a minúsculas.

Tabla 5.

Resultado conversión de mayúsculas

```

[1] "RT @Mvandalico: @Guerrea13 @crittttico @el9verdugo En Nicarag
ua los que están des-gobernando son igual de pendejos que #maburro
... Tengamos..."

[1] "rt @mvandalico: @guerrea13 @crittttico @el9verdugo en nicarag
ua los que están des-gobernando son igual de pendejos que #maburro
... tengamos..."

```

Nota: Elaboración propia

B. Remover URL no necesarias para el proceso (Figura 23).

```

# remover URLs
removeURL <- function(x) gsub("http[^\s:]*", "", x)

```

Figura 23. Código remoción de la URL .

Nota: Elaboración propia

C. Remover espacios no necesarios para el proceso (Figura 24).

```
# Remover espacios |
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*",
+ "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeNumPunct))
```

Figura 24. Código remoción de espacios.

Nota: Elaboración propia.

La tabla 6 visualiza el siguiente resultado:

Tabla 6.

Resultado de la eliminación de los espacios

1] "rt mvandalico guerrea crittttico elverdugo en nicaragua los que es tan desgobnando son igual de pendejos que maburro tengamos"

Nota: Elaboración propia

D. Creación de un diccionario de palabras del lenguaje cotidiano en español para las redes sociales para su posterior sustitución y análisis. (Tabla 07. Diccionario de palabras)

Tabla 7.

Diccionario de pablaras para la corrección idiomática.

Diccionario para la corrección de ortografía	
Palabra en Redes Sociales	Sustitución básica
Abra	Habrá
Aber	Haber
acer	Hacer
Amiguis	amigos
Ay	Ahí
asia	Hacia
Alos	A los
bamos	vamos
Bb	bebe
bemos	vemos
ceis	seis

Diccionario para la corrección de ortografía	
Palabra en Redes Sociales	Sustitución básica
d	de
dises	Dices
Ej	es
Ezpero	Espero
escrivo	escribo
Sta	está
k	que
x	por
Xq	Porque
Ksa	casa
kuando	cuando
Ola	Hola
llo	yo
povres	pobres
vueno	bueno
Zalimos	salimos

Fuente: Elaboración propia

E. Remoción de palabras innecesarias

La Tabla 8 visualiza el código para ejecutar la remoción de letras o palabras que no aportan objetividad a la toma de decisiones.

Tabla 8.*Remoción de palabras innecesarias.*

Remoción de palabras innecesarias
<pre>myStopwords <- c(setdiff(stopwords("spanish"), c("de", "en", "el")), "rt", "del", "la", "da", "a", "las", "pero", "", "hola", "s", "cil", "n", "a", "dÃ", "polÃ", "rÃ", "u", "as", "s", "os", "d", "v", "o", "aquÃ", "paÃ", "aÃ", "vÃ", "ellos", "via", "los", "en", "dentro", "de", "mas", "el", "igual", "asi", "sera", "mejor", "casa", "s", "cil", "aÃ", "jul", "os")</pre>

Nota: Elaboración propia

3.4.3. Modelado (Model)

El modelado, se realizará mediante el algoritmo de K-Means, el cual fue explicado en el Capítulo III numeral 3.1.8 Algoritmos de Procesamiento; este algoritmo es uno de los más usados y conocidos para hacer clustering o agrupamiento de datos. La idea básica reside en la búsqueda de una partición óptima que minimice el criterio de la suma del cuadrado de los errores. De acuerdo al algoritmo proporcionado por Rstudio y las librerías de TM²⁹ para la programación neurolingüística se encontrarán los patrones y tendencias de acuerdo los modelos de búsqueda. (Alcobendas, 2018)

Estas líneas de código permiten agrupar los temas más relevantes de acuerdo al criterio de búsqueda establecido en el anterior punto. (Ver Tabla 9)

El algoritmo fue proporcionado por el autor Yanchang Zhao del libro R and Data Mining: Examples and Case Studies. (Zhao, 2013)

²⁹ Es un framework para aplicaciones de minería de texto para el lenguaje R. Inventado por el ingeniero Ingo Feinerer bajo el repositorio de CRAN el 19 de junio de 2018.

Tabla 9.*Código y Ejecución del algoritmo de k-means.*

Clustering Con El Algoritmo De K-Means
<pre> #Construcción de una matriz: tdm <- TermDocumentMatrix(myCorpus, control = list(wordLengths = c(1, Inf))) m <- as.matrix(tdm) #Eliminación de términos dispersos tdm2 <- removeSparseTerms(tdm, sparse=0.95) #Crea una sparse=máxima dispersión permitida en el rango. #Se usa para quitar elementos sin interés #head(inspect(tdm2)) m2 <- as.matrix(tdm2) # Transponemos la matriz m3 <- t(m2) ##Algoritmo de las k-medias #Fijamos semilla set.seed(122) # Grupos por el algoritmo de las k-medias # Tomamos k=7 k <- 7 kmeansResult <- kmeans(m3, k) # Centros round(kmeansResult\$centers, digits=3) #Revisamos tres primeras palabras de cada cluster for (i in 1:k) { cat(paste("cluster ", i, ": ", sep="")) s <- sort(kmeansResult\$centers[i,], decreasing=T) cat(names(s)[1:3], "\n") # Imprimir tweets de cada Cluster # print(rdmTweets[which(kmeansResult\$cluster==i)]) } </pre>

Para siete (7) Cluster, las matrices fueron expuestas de la siguiente manera:

```
round(kmeansResult$centers, digits=3)
```

	n	a	armygames	cielo	claro	codaifanb	conducción	continuacion	equipo	mediante		
1	0.200	0.200	1.000	1	1	1.000	0	0.2	0	1	0	
2	0.000	0.000	0.000	0	0	0.000	0	0.0	0	0	0	
3	2.000	1.000	1.000	1	1	0.911	1	1.0	1	1	1	
4	0.000	0.000	0.971	0	0	0.000	0	0.0	0	0	0	
5	0.000	0.000	0.000	0	0	0.000	0	0.0	0	0	0	
6	0.000	0.000	0.000	1	1	2.000	0	0.0	0	1	0	
7	0.133	0.067	0.200	0	0	0.000	0	0.0	0	0	0	

	participación	pro	vehiculos	ejbbrig	ejercitofanb	entren	internacionales	juegos	militares		
1	0.2	0	0.000	0	0.000	0	0.000	0	0	0	0
2	0.0	0	0.000	0	0.000	0	0.000	0	0	0	0
3	1.0	1	0.933	0	0.000	0	0.000	0	0	0	0
4	0.0	0	0.000	1	0.941	1	0.941	1	1	1	1
5	0.0	0	0.000	0	0.000	0	0.000	1	1	1	1
6	0.0	0	0.000	0	0.000	0	0.000	0	0	0	0
7	0.0	0	0.000	0	0.000	0	0.000	0	0	0	0

	realizan	soldados	cancha	entrenamiento	hoy	inquebrantable	rate	realiza	reconocimiento	voluntad		
1	0.000	0.000	0	0	0	0	0	0	0	0	0	
2	0.000	0.000	0	0	0	0	0	0	0	0	0	
3	0.000	0.000	0	0	0	0	0	0	0	0	0	
4	0.941	0.941	0	0	0	0	0	0	0	0	0	
5	0.000	1.000	0	0	0	0	0	0	0	0	0	
6	0.000	0.000	1	1	1	1	1	1	1	1	1	
7	0.000	0.000	0	0	0	0	0	0	0	0	0	

	airshow	arrived	aviadarts	belorussian	chinese	crews	international	part	ryazan	starting	take		
1	0	0.0	0	0	0.000	0	0	0	0.000	0.000	0	0	
2	1	1.0	1	1	1.000	1	1	1	1.917	0.917	1	1	
3	0	0.0	0	0	0.000	0	0	0	0.000	0.000	0	0	
4	0	0.0	0	0	0.000	0	0	0	0.000	0.000	0	0	
5	0	0.0	0	0	0.000	0	0	0	0.000	0.000	0	0	
6	0	0.0	0	0	0.000	0	0	0	0.000	0.000	0	0	
7	0	0.4	0	0	0.333	0	0	0	0.000	0.000	0	0	

	armyt	así	cmdteaviacion	entrenan	h	paracaidistas	venezuelapotencia		
1	0.000	0	0.000	0	0	0	0	0	0
2	0.000	0	0.000	0	0	0	0	0	0
3	0.000	0	0.000	0	0	0	0	0	0
4	0.000	0	0.000	0	0	0	0	0	0
5	0.981	1	0.981	1	1	1	1	1	1
6	0.000	0	0.000	0	0	0	0	0	0
7	0.000	0	0.000	0	0	0	0	0	0

Nota: *Elaboración propia*

Después de haber ejecutado el código anterior se logra determinar que existen un par de temas con mayor importancia: los ejercicios militares internacionales que ha efectuado el Ejército Bolivariano (ejercitofanb) en manos del CODAI (Comando de Defensa Aeroespacial Integral) en espacio aéreo venezolano con el apoyo de las armas provenientes de China. (Ver Tabla 10)

Tabla 10.*Resultado final clustering con el algoritmo de k-means*

RESULTADO FINAL CLUSTERING CON EL ALGORITMO DE K-MEANS
cluster 1: armygames cielo claro
cluster 2: ryazan airshow arrived
clúster 3: armygames
clúster 4: cara ejercitofanb internacionales
clúster 5: internacionales juegos militares
cluster 6: codaifanb cielo claro
cluster 7: arrived chinese armygames

Nota: Elaboración propia

3.4.4. Valoración (Assess)

La valoración de los resultados se verificará de acuerdo a la efectividad del algoritmo de clustering K-Means al agrupar el mayor número de elementos con la mayor similitud, diferenciándolos de otras agrupaciones. La probabilidad de que un elemento pertenezca a dos grupos distintos debe ser estudiada como el error del algoritmo al no realizar el debido agrupamiento.

La eficacia del algoritmo es medido acorde a la inclusión del tamaño de la muestra a abstraer, la cantidad de clúster ejecutados en el código y en la eliminación de palabras o letras no acordes con el tema en cuestión, ya que uno de los grandes inconvenientes que se vieron en la ejecución del código radicó en la utilización de librerías en idioma inglés(NLTK) y no en español, lo cual afectó el resultado final.

Uno de los mayores retos que se afrontó con respecto al código, fue la escogencia del número de clúster para efectuar la búsqueda de los patrones y tendencias, el porcentaje máximo de frecuencia de la aparición de un elemento dentro del grupo y el número mínimo de textos en que debe aparecer una palabra para ser incluida dentro del clúster.

Por tal motivo, es objeto de estudio realizar los cálculos para determinar el error de la inclusión de un elemento al grupo del clúster.

Un valor de precisión de 1 indica que el clúster es perfecto, es decir, que es idéntico a la clasificación del centroide inicial.

En la siguiente Tabla 11 se observa el resultado de variar el número de clúster(k).

Tabla 11.
Pruebas de variación de Clúster

Numero de Clúster(k)	Porcentaje de Precisión
2	0.42
7	0.58
14	0.496
15	0.49
16	0.49

Nota: Elaboración propia

Acorde a los resultados arrojados en la Tabla 11, se puede inferir que al aumentar el número de clúster el porcentaje de precisión es menor, por lo cual no es apropiado usar un número muy elevado de clúster. Planteado de la forma como se realizó el análisis, el mejor resultado se obtiene cuando $k = 7$, tal como se efectuó anteriormente.

La Tabla 12 visualiza la frecuencia de repetición de todas las agrupaciones de clúster del muestreo efectuado para la búsqueda seleccionada, cabe resaltar que independiente del número de clúster elegido, el agrupamiento es capaz de organizar de mayor a menor los elementos que más tuvieron incidencia o tendencia en los Tweets.

Tabla 12.
Frecuencia repetición de las palabras clave.

Palabra Clave	Frecuencia de Repetición
codaifanb	120
cielo	87
claro	87
equipo	87
armygames	86
internacionales	86
juegos	86
militares	86
soldados	84
entrenan	52
paracaidistas	52
venezuelapotencia	52
armyt	51
cmdteaviacion	51
Participación	46
conducción	45
mediante	45
cancha	37
inquebrantable	37
rate	37

Palabra Clave	Frecuencia de Repetición
realiza	37
reconocimiento	37
voluntad	37
cara	34
ejercitofanb	34
ejbbrig	32
entren	32
realizan	32
ryazan	23
arrived	18
chinese	17
airshow	12
aviadarts	12
belorussian	12
crews	12
international	12
part	12
take	12
china	1

Nota: Elaboración propia

La precisión del resultado de los patrones y tendencias se puede ver supeditado al algoritmo K-Means y al afinamiento de la eliminación de palabras innecesarios dentro del contexto que estamos buscando, lastimosamente estos códigos fueron diseñados para palabras en inglés, lo

cual género que algunas de las palabras que contenían tildes no fueran procesadas correctamente. Para futuros trabajos se hace indispensable implementar un sistema de programación neurolingüística para el idioma español.

información legible, en un tiempo razonable, para un objetivo específico; los sistemas de mejoras en seguridad de redes ante los ataques de redireccionamiento que permiten hacer un análisis detallado de los datos.

El crear un modelo de redes sociales tecnológicas como lo define el artículo de organizar la información de las redes sociales y que sirve para la anticipación de peligros geoestratégicos para Colombia, es el punto de partida para lograr obtener información relevante en inteligencia militar.

La información obtenida de esta investigación, logra mostrar una variable significativa para la utilización de personas de las redes sociales para desempeñar las funciones de las personas que intervienen sobre ellas y de esta forma, lograr aportar a la zona de desarrollo económico sobre la inteligencia estratégica.

El marketing en redes sociales (MRS) es una forma particular de los sistemas de redes las cuales ayudan a clientes potenciales (Pantazidou S., 2014). El marketing digital que una parte de los productos para la minería de datos para comprender mejor la zona de desarrollo económico, en donde las empresas pueden anticiparse a los servicios de un grupo de personas para orientar sus esfuerzos de ventas hacia estos tipos de personas, esta investigación orientada en proyectos sobre la inteligencia estratégica en aquellos casos que afectan la seguridad y defensa de la nación y que es imperativo anticiparse desde el punto de vista de inteligencia militar como para el decisor político Nacional.

1. En su significado inglés "Social media marketing", son las medidas implementadas por la empresa para promover los servicios de uno de sus productos.

4. Conclusiones

La cantidad de datos digitales a la que nos vemos expuestos en la actualidad, ha llevado a que cada día se desarrollen mejoras a nivel computacional en un afán de tratar de obtener información legible, en un tiempo razonable, para un objetivo específico; sin embargo, estas mejoras no significan nada sino se enmarcan en modelamientos que permitan hacer un análisis ordenado de los datos.

Diseñar un modelamiento basado en metodologías reconocidas como SEMMA con el propósito de organizar la información de las redes sociales y que sirva para la anticipación de peligros geoestratégicos para Colombia, es el punto de partida para lograr obtener información coherente en inteligencia militar.

La información plasmada en esta investigación, logra incorporar una variable significativa para la utilización de patrones de las redes sociales para comprender las tendencias de las personas que interactúan sobre éstas y de esta forma, lograr aportar a la toma de decisiones orientadas hacia la inteligencia Estratégica.

El marketing en redes sociales (SMM³⁰) es una forma poderosa para que las empresas de todos los tamaños lleguen a clientes potenciales (Punithavathi K, 2018); el marco teórico vislumbró que una parte de los procesos para la minería de datos están enfocados hacia la toma de decisiones empresariales, en como las empresas pueden anticiparse a las tendencias de un grupo de personas para orientar sus esfuerzos de ventas hacia estos tipos de patrones, esta investigación encaminó su proyección hacia la inteligencia estratégica, en aquellos casos que afectan la seguridad y defensa de la nación y que es imperativo anticiparse desde el punto de vista de inteligencia militar como para el decisor político Nacional.

³⁰ Por sus siglas en ingles “Social media marketing”. Son las medidas implementadas por la empresas para promover los servicios dentro de una red social.

Es así, como el análisis de las redes sociales aplicado al ambiente de la inteligencia militar-estratégico, permite hacer visible lo que hasta hace poco era impensable visualizar: la estructura de las relaciones sociales en el ciberespacio y el comportamiento anticipatorio de fenómenos que anteriormente no se lograrían ver a simple vista, como un aporte significativo a una de las variables del proceso de análisis de información para la toma de decisiones estratégicas. Una de las aplicaciones prácticas directas del análisis y difusión de productos de inteligencia para la toma de decisiones, es la identificación de líderes de opinión o generadores de tendencias en las redes sociales. Esto fue posible, gracias al desglose de los objetivos estratégicos de Colombia y a la concatenación con los criterios de selectividad de las palabras claves para orientar el esfuerzo de búsqueda que finalmente concluyó con un producto de inteligencia estratégica.

En términos metodológicos, SEMMA demostró ser una hoja de ruta para organizar los procesos de la presente investigación, cada una de sus etapas fueron de gran uso para encausar la búsqueda, el análisis y el procesamiento de los datos. Indistintamente de las herramientas o algoritmos usados para efectuar la programación, fue un punto de partida que permitió al estado del arte buscar las mejores técnicas asociadas al objetivo general de anticiparse a los peligros geoestratégicos.

A nivel de programación, el API de Twitter, fue el principal elemento de captura de información, que en conjunto a un buen tratamiento de programación en un ambiente como el del lenguaje en R, lograron efectuar un análisis de datos para la caracterización de cualquier fenómeno geoestratégico y el proceso de toma de decisiones. Cabe destacar que la API de Twitter restringe y limita las consultas efectuadas en su versión “Estándar a 180 peticiones”, como se verificó en el capítulo II, y fue viable y positivo la recolección de datos, pero no lo suficiente para lograr un análisis con millones de bytes descargados, lo cual, hubiese enriquecido el producto de inteligencia hacia una mejor toma de decisiones. Asimismo, se identificó gracias al apoyo del Ingeniero Santacruz (2016) que solo un pequeño porcentaje de los tweets públicos están geo-localizados (alrededor de 3-5%) por lo que no es posible recuperar algunos tweets que dispongan de los campos de latitud y longitud como se hubiere deseado.

Uno de los grandes retos que se asumieron en la presente investigación, derivado de la gran cantidad de datos que circulan por Internet, sigue siendo el contrastar la veracidad de la información que se aloja dentro del ciberespacio, cuyo principal eje reside en los usuarios que deciden dar su opinión a través de las redes sociales; sin embargo, a través de los procesos de análisis de tendencias (algoritmo de agrupación K-means), los usuarios comunes pueden ser un punto de referencia para ser analizados dentro del contexto estratégico.

Esta línea de investigación no busca ni pretende realizar espionaje, cabe mencionar que este proyecto resguarda la privacidad o abusos vinculados al ejercicio mismo. La concepción del proyecto radica en el análisis del comportamiento colectivo que se pueda llegar a obtener con la adquisición de información por medio de fuentes abiertas.

Este trabajo de investigación es un derrotero para fortalecer la capacidad de Ciberinteligencia e inteligencia estratégica de la Armada Nacional de Colombia y responde al modelamiento analítico de un sistema que permita la anticipación de peligros geoestratégicos para Colombia, a través del análisis de patrones y tendencias en Redes Sociales. De igual forma, es un punto de partida para futuras investigaciones, a través del mejoramiento de las metodologías de modelado y análisis de información (k-means) como en la integración de más plataformas de redes sociales, tales como Facebook, Instagram, YouTube, ente otras, las cuales aportarían información relevante a ser contrastada para detectar la veracidad del producto final.

5. Referencias Bibliográficas

- Achayra, S. M. (2003). *Data Mining: Multimedia, soft computing and bioinformatics*. John Wiley & Sons.
- Afanador, W. R. C. (2012). Conflictos de delimitación marítima y la convemar. una mirada desde Colombia. *Prolegómenos. Derechos y Valores*, 199–223. Recuperado a partir de <http://www.redalyc.org/articulo.oa?id=87625443011>
- Alberto José Mejía Ferrero, G., Leonardo Santamaría Gaitán, A., Eduardo Bueno Vargas, C., Rodolfo Palomino López, G., Manuel Neira Núñez, L., Carlos Villegas Echeverri, L., ...
- Henry John Blain Garzón, A. (2017). *POLÍTICA DE DEFENSA Y SEGURIDAD*. Recuperado a partir de https://www.dimar.mil.co/sites/default/files/attach/politica_de_defensa_y_seguridad_2015-2018_diagramada_feb_17_16.pdf
- Alcobendas, G. R. (2018). *Técnicas estadísticas en análisis de redes sociales*. Universidad de Sevilla.
- Alonso García González Director de Reporte Técnico Dra Alejandra García Hernández, J. (2013). Revisión sistemática de herramientas y métodos para las diferentes fases del análisis de redes sociales en línea. Reporte Técnico Maestro en Ingeniería de Software Presenta. Recuperado a partir de <https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/427/1/ZACTE31.pdf>
- Boyd, R. (2012). *Getting Started with OAuth 2.0*. (M. L. and S. Wallace, Ed.) (O'Reilly M). Sebastopol, CA 95472: O'Reilly Media.
- Candau, J., Ciberseguridad, J. D. E., & Centro, D. E. L. (2015). “ El ciberespionaje persigue información de altísimo valor y debe considerarse una amenaza crítica ”, 4–7.

Carlos, L., Echeverri, V., Carolina, D., Cardona, P., Andrea, L., Viancha, N., ... Guerra, R. (2018). *Logros de la Política de Defensa y Seguridad Todos por un Nuevo País*.

Recuperado a partir de

https://www.mindefensa.gov.co/irj/go/km/docs/Mindefensa/Documentos/descargas/estudios_sectoriales/info_estadistica/Logros_Sector_Defensa.pdf

CERÓN, C. A. C. (2012). *Elementos del territorio marítimo-costero: sus instrumentos de gestión y administración territorial. Bases para una política pública de ordenamiento territorial carlos armando chávez cerón universidad nacional de colombia*. Universidad Nacional de Colombia. Recuperado a partir de

<http://bdigital.unal.edu.co/12059/1/carlosarmandochavez.2012.pdf>

De Jonge, E., & Van Der Loo, M. (2013). An introduction to data cleaning with R.

Recuperado a partir de https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf

Departamento de Seguridad Nacional. (2017). Amenazas y desafíos para la seguridad nacional. *Estrategia de Seguridad Nacional 2017*, 54–77. Recuperado a partir de http://www.dsn.gob.es/sites/dsn/files/ESN2017_capitulo_4.pdf

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques*, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems). Recuperado a partir de <http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf>

Javier Santos. (2015). Ciberinteligencia en las redes sociales. *La Ventana de KPMG*.

Recuperado a partir de

https://static1.squarespace.com/static/5493ed68e4b0b73d7c7ea649/t/55ef1615e4b0d5617cb61b5b/1441732117973/El+Mundo+INNOVADORES_ionando.pdf

Kirill Müller. (2018). “SQLite” Interface for R [R package RSQLite version 2.1.1].

Recuperado el 17 de julio de 2018, a partir de <https://cran.r-project.org/web/packages/RSQLite/index.html>

Martha Morales Manchego. (2018). Así van las investigaciones de Panama Papers en

Colombia - Sector Financiero - Economía - ELTIEMPO.COM. Recuperado el 14 de julio de 2018, a partir de <http://www.eltiempo.com/economia/sector-financiero/asi-van-las-investigaciones-de-panama-papers-en-colombia-201292>

Ministerio de defensa de España. (2009). *La Violencia Del Siglo Xxi. Nuevas Dimensiones De La Guerra.*

Ministerio de Defensa Nacional. (2016). República de Colombia Ministerio de Defensa

Nacional asunto: expedir los lineamientos del Ministerio de Defensa Nacional para caracterizar y enfrentar a los grupos armados organizados (gao) derogar la Directiva Permanente 014 de 2011 que establece la estr, 9. Recuperado a partir de

https://www.mindefensa.gov.co/irj/go/km/docs/Mindefensa/Documentos/descargas/Prensa/Documentos/dir_15_2016.pdf

Montequín, R., Teresa, M., Cabal, Á., Valeriano, J., Fernández, M., Manuel, J., & Valdés, G.

(2015). Metodologías Para La Realización De Proyectos De Data Mining. *Universidad de Oviedo*, 257–265. Recuperado a partir de

http://www.aepro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf

Punithavathi K, D. H. (Enero de 2018). Social media marketing. *International Journal of Commerce and Management Research*, Volume 4, Page No. 40-43. ISSN: 2455-1627

- Ravindran, Sharan Kumar Garg, V. (2015). *Mastering Social Media Mining with R*.
- Russell, M. A. (2015). *Mining the Social Web. CEUR Workshop Proceedings* (Vol. 1542).
<https://doi.org/10.1017/CBO9781107415324.004>
- SALAZAR, M. G. A. (2013). *Monografía Colombia: el rescate de un patrimonio economico y cultural, el facultad de estudios a distancia relaciones internacionales y estudios politicos bogota 2013 2 contenido*. Universidad Militar Nueva Granada . Recuperado a partir de
<https://repository.unimilitar.edu.co/bitstream/10654/11384/1/AranaSalazarMarioGerman2013.pdf>
- Sánchez, J. M. (2016). *Universidad De Oviedo Trabajo Fin De Máster Adaptación De Estándares De Dirección De Proyectos Particularizados Para La Minería De Datos Fecha*. Recuperado a partir de
<http://digibuo.uniovi.es/dspace/bitstream/10651/43633/7/TFMjulioCesarMerida01RUO.pdf>
- Schab, E., Rivera, R., Bracco, L., Coto, F., Cristaldo, P., Ramos, L., ... Herrera, N. E. (2018). *Minería de Datos y Visualización de Información*. Recuperado a partir de
http://sedici.unlp.edu.ar/bitstream/handle/10915/67382/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y
- Stein, A. (2009). *El concepto de Seguridad multidimensional*. Recuperado a partir de
<https://www.peaceportal.org/documents/130226273/130473587/El+Concepto+de+Seguridad+Multidimensional/61ea9879-48a6-464e-985c-8b75354eda21>
- Wilson, E. J. (2008). Hard Power, Soft Power, Smart Power. *The ANNALS of the American*

Academy of Political and Social Science, 616(1), 110–124.

<https://doi.org/10.1177/0002716207312618>

Zaki, M. J., Meria, W., & Meira, W. (2013). *Data Mining and Analysis: Fundamental Concepts and Algorithms*.

Zhao, Y. (2013). *R and data mining : examples and case studies*. Academic Press.

BIBLIOTECA CENTRAL DE LAS FF.MM.
"TOMAS RUEDA VARGAS"



201002335