

PROYECTO FINAL DE INGENIERÍA

SOCIAL MEDIA INTELLIGENCE ORIENTADO AL ÁMBITO POLÍTICO

Bustamante, Josefina – LU1019231

Coria, Agustín– LU1015257

Ingeniería en Informática

Tutor:

Salomone, Federico Aníbal, UADE

Agosto 20, 2016



**UNIVERSIDAD ARGENTINA DE LA EMPRESA
FACULTAD DE INGENIERÍA Y CIENCIAS EXACTAS**

1. Agradecimientos

Agradecemos a nuestro tutor, Federico Salomone, por su constante aliento, ayuda y dedicación. A pesar de la distancia, aportó de forma constante con su consejo para nuestra investigación.

También queremos agradecerle a Ricardo Ferré, por orientarnos en la selección del tema sobre el cual desarrollamos este escrito.

2. Resumen

Social Media Intelligence orientado al Ámbito Político

El presente estudio expone un conjunto de herramientas y técnicas para organizaciones políticas que buscan utilizar el gran caudal de datos generados por las redes sociales para su beneficio y el de su comunidad objetivo. Previo al comienzo de la investigación, se provee un entendimiento sobre la complejidad que existe en el cuantioso volumen de datos que se persiste en el campo digital y la necesidad de disgregarla y procesarla.

Se brinda una reseña de los estudios y los esfuerzos ajenos relacionados a la materia de investigación, formulando una mejor noción sobre el contexto académico en el que surge el estudio. Se detallan los conceptos y conocimientos propios del campo de estudio del Business Intelligence necesarios para poder comprender el desarrollo de la investigación, para luego sí brindar un detalle del conocimiento novedoso respecto al subcampo de nuestro análisis, el Social Media Intelligence. Se introduce, con un caso de las elecciones Presidenciales y Legislativas de 2012 de Obama, el concepto de Política 3.0 que implica un cambio en el paradigma de comunicación de una organización política con su población objetivo.

Posteriormente, se describen y evalúan las técnicas de Minería de Datos, de Text Mining y Network Mining, dando a conocer el detalle de estas vías para obtener información partiendo de datos abundantes, no estructurados, dispersos y en muchas ocasiones contaminados en las redes sociales. Además, se detalla el funcionamiento de la red social que hoy es termómetro de las opiniones de una sociedad, Twitter. Luego, se describe paso a paso cómo emular la técnica de Text Mining, sobre Twitter, con procesos en las herramientas de más renombre del mercado, RapidMiner y Knime. Se brindan los resultados de la ejecución de los procesos, elaborados previamente, sobre casos reales para luego comparar los beneficios y desventajas de las dos herramientas utilizadas.

Se expone cómo las entidades públicas utilizan los datos que brinda el campo digital para focalizar con mayor eficiencia los receptores de sus publicidades en la red social de YouTube, con el objetivo de incidir con mayor eficiencia en las opiniones y nociones de una población.

Finalmente, se cierra con observaciones y conclusiones sobre la eficiencia y la conveniencia de la aplicación de las técnicas descriptas en el proyecto, haciendo también una reflexión sobre la experiencia en su desarrollo.

3. Abstract

Social Media Intelligence in the Political Field

The purpose of this study is to expose the variety in tools and techniques that exist for political organizations that aim to use the vast amount of data that is generated by social networks for their benefit and their community's. We start up by presenting an understanding of the complexity that exists in the extensive quantity of data that is stored in the digital field and the need to untangle and process it.

A summary of the academic studies and efforts related to the subject in question is provided, thus forming a better notion of the academic context upon which this study emerges. The concepts and the knowledge of the Business Intelligence field of study necessary to further understand the development of this investigation is supplied, followed up by a detailed elaboration of our academic focus, Social Media Intelligence. Furthermore, the concept of Politics 3.0 is introduced, which implies a change in the communication paradigm between a political entity and its objective population.

A description and examination of Data Mining techniques of Text Mining and Network Mining is made, offering a brief of these methods of obtaining information from the great extent of data that is scrambled, disperse and in many occasions polluted with noise in social networks. A detail is given of the operating of the social network that serves as today's societies thermometer in terms of opinion, Twitter. In addition, there is a step by step description of how to emulate Text Mining, feeding off Twitter, with processes in the market's leading tools, Knime and RapidMiner. The results of executing the previous processes to then compare the benefits and disadvantages of the two tools are exposed.

Moreover, there is an explanation of how political entities use the data of the digital field to further target the consumers of their advertisements on the social network of YouTube, aiming to influence with greater ease in the opinions and conceptions of a population.

To close, observations and conclusions in terms of the efficiency and convenience of the use of the described techniques in our project are detailed, together with a reflection upon the experience it was to compose it.

4. Contenido

1. Agradecimientos	2
2. Resumen	3
3. Abstract	5
4. Contenido	6
5. Glosario	8
6. Introducción	17
6.1 Relevancia del tema	17
6.2 Objetivos	18
6.3 Descripción de Estructura	19
7. Estado del Arte	20
7.1 Marco teórico	20
7.1.1 Business Intelligence	20
7.1.2 Social Media Intelligence	20
7.1.3 Data Mining	22
7.1.4 Text Mining	23
7.1.5 Network Mining	24
7.1.6 Data Warehousing	25
7.1.7 Data Mart	26
7.1.8 API	26
7.1.9 Cookies	26
7.1.10 Google AdWords	27
7.1.11 Google Display Network	27
7.1.12 Twitter	27
7.2 Investigaciones hasta el momento	29
8. Hipótesis	31
9. Metodología	33
9.1 Caso Política 3.0	34
9.2 Técnicas analíticas predictivas	36
9.2.1 Text Mining	36
9.2.2 Network Analysis	38
9.2.3 Caso de ejemplo: elecciones de Brasil, año 2014	39
9.3 Combinación entre Text Mining y Network Mining	43
9.4 Nuevo enfoque - Combinación entre Sentiment Analysis y Network Analysis	43
9.5 Consideraciones sobre las técnicas desarrolladas	45
9.6 Análisis de herramientas de Minería de Datos	46
9.6.1 Knime	47
9.6.2 RapidMiner	52
9.6.3 Comparación de herramientas y conclusiones	59

9.6.4 Caso real con Knime y RapidMiner	62
9.7 Posicionamiento y focalización de publicidad en YouTube	68
10. Resultados	72
11. Conclusiones	78
12. Bibliografía	81

5. Glosario

<p>Algoritmo</p>	<p>En matemática, lógica y ciencias de la computación, un algoritmo es un conjunto prescrito de instrucciones o reglas, definidas, ordenadas y finitas, que permiten realizar una actividad mediante pasos sucesivos. (Brassard y Bratley, 1997).</p>
<p>API</p>	<p>Es la abreviatura de Application Programming Interface. Es un conjunto de reglas (código) y especificaciones que las aplicaciones pueden seguir para comunicarse entre ellas, sirviendo de interfaz entre programas diferentes de la misma manera en que la interfaz de usuario facilita la interacción humano-software. (Merino, 2014).</p>
<p>Aplicación móvil</p>	<p>Hace referencia a un sistema informático diseñado para ser ejecutado en dispositivos móviles, que permite al usuario efectuar una tarea específica de cualquier tipo. Por lo general se encuentran disponibles a partir de plataformas de distribución como Android, iOS, Windows Phone y Blackberry OS. Se suele abreviar el término utilizando la palabra app, del inglés, application. (Santiago <i>et al</i>, 2015).</p>
<p>Atributo</p>	<p>Son las características por medio de las cuales se puede describir una entidad. En el contexto en cuestión, es toda palabra utilizada como adjetivo, de la cual se puede definir su polaridad.</p>
<p>BI</p>	<p>Abreviatura de Business Intelligence. Es el conjunto de estrategias y aspectos relevantes enfocados a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa. (Turner, 2016).</p>
<p>Bots</p>	<p>Un bot, como abreviación de robot, es un software informático, imitando el comportamiento de un humano. Un bot puede realizar funciones rutinarias y responder a cuestiones sobre el contenido de un sitio (conversacional). (Rodríguez Posada <i>et al</i>, 2010).</p>
<p>Candidato</p>	<p>En política, se denomina candidato, a la persona que se postula a</p>

	<p>ser elegida para algún cargo público, electo en unas elecciones, y normalmente incluido en las listas electorales. (Lijphart, 1995).</p>
Ciudadano	<p>Representa a toda persona considerada como miembro activo de un Estado, titular de derechos políticos, y sometido a sus leyes. (Real Academia Española, 2014).</p>
Cliente	<p>Representa a toda persona física o sociedad que genera volumen de datos para las empresas. Según el rubro al cuál se dedica la organización en cuestión, utilizaremos el término cliente / ciudadano / usuario.</p>
Cloud	<p>La nube, que proviene del inglés, Cloud Computing, es el nombre que se le dió al procesamiento y almacenamiento masivo de datos en servidores que alojen la información del usuario. Favorece el acceso instantáneo y en todo momento y lugar, a los datos. (Marazzi, 2016).</p>
Clustering	<p>Es la división de los datos en grupos de objetos similares. Este proceso ayuda a la agrupación estructural de un conjunto de datos. (Han y Kamber, 2006).</p>
Comunidad	<p>El concepto hace referencia a la característica de común, por lo que permite definir a diversas clases de conjuntos: de los individuos que forman parte de una región o de personas vinculadas con intereses comunes. Gracias al desarrollo de la tecnología, y de Internet, se han conformado las comunidades virtuales. Las redes sociales, los foros, los sistemas de mensajería instantánea y los blogs son sitios que permiten la creación de este tipo de comunidades. (Definición De, 2016).</p>
Conocimiento	<p>El conocimiento es un conjunto de información almacenada mediante la experiencia, o el aprendizaje, o a través de la introspección. (Definición De, 2016).</p>
Data / Datos	<p>Del inglés, se traduce como datos. Es una representación simbólica de un atributo o variable cuantitativa o cualitativa. El dato no tiene sentido en sí mismo, sino que se utiliza en la toma de decisiones o en la realización de cálculos, a partir de un procesamiento adecuado, y</p>

	teniendo en cuenta su contexto. (Definición De, 2016).
Decisión ejecutiva / Decisión estratégica	Las decisiones estratégicas son las que definen el futuro de las empresas y las organizaciones. Típicamente estas decisiones implican comprometer cantidades considerables de recursos y los resultados a obtener son inciertos y se conocerán años después de que se tomó la decisión. (Ley Borrás, 2013).
Drag and drop	Desplazar un elemento de un lugar a otro arrastrándolo con el mouse. (Glosario.net, 2013).
Eclipse	Es una plataforma de software compuesto por un conjunto de herramientas de programación open source multiplataforma basadas en navegadores. (Eclipse.org, 2016).
Entidad pública	Entidades creadas por el gobierno para prestar servicios públicos. Pertenecen al estado, tienen personalidad jurídica, patrimonio y régimen jurídico propios.
Explotación	Es una sub-disciplina de la Informática que aporta herramientas de minería de datos para extraer el conocimiento no trivial que se encuentra (implícitamente) en los datos disponibles de diferentes fuentes de información dentro del marco de un proyecto de ingeniería. (Pollo-Cattaneo <i>et al.</i> 2012).
Extracción	La extracción de datos, está destinada a mostrar cualquier correlación dentro de un volumen de datos importante del sistema de información con el fin de detectar alguna tendencia. La extracción de datos se basa en técnicas de inteligencia artificial (red neuronales) para mostrar vínculos ocultos entre los datos. (CCM, Comunidad Informática, 2016).
Feed	Una fuente web o canal web (en inglés web feed) es un medio de redifusión de contenido web. Se utiliza para suministrar información actualizada frecuentemente a sus suscriptores. Son muy utilizados en los blogs de Internet, así como en prensa electrónica. La gran difusión de este tipo de servicio web ha estimulado el interés en otros sectores que

	comienzan a ver en las fuentes web una forma efectiva de difundir y compartir información. (Nature Publishing Group, 2016).
Followers	En Twitter, este concepto significa el número de usuarios que siguen una cuenta. Esto significa que cualquier aporte introducido en dicha cuenta, se mostrará en la página de inicio de cada uno de estos fans. (Intef, 2016).
Frecuencia	Número de veces que aparece, sucede o se realiza una cosa durante un período o un espacio determinados. (Oxford Dictionaries, 2016).
GUI	La interfaz es la mediadora entre dos sistemas de diferente naturaleza: el hombre y la máquina; ya que además de facilitar la comunicación y la interacción entre ambos, sirve de traductor, pues estos dos sistemas “hablan” lenguajes diferentes: verbo-icónico en el caso del hombre y binario en el caso de la computadora. (Lewis y Rieman, 1993).
Hashtag	Se traduce como etiqueta, es una cadena de caracteres, formada por una o varias palabras concatenadas, y precedidas por el símbolo numeral (#). El fin es que, tanto el usuario como el sistema lo identifique de forma rápida. Se utiliza comúnmente en redes sociales. (Twitter, 2016).
Individuo	Persona considerada independientemente de las demás. Un ser autónomo que se define por su capacidad racional y su fuerza de voluntad. (Definición.mx, 2016).
Información	Está constituida por un grupo de datos ordenados, que sirven para construir un mensaje. La información permite resolver problemas y tomar decisiones, ya que su aprovechamiento racional es la base del conocimiento. (Definición De, 2016).
Input	Conjunto de datos que se introducen en un sistema o un programa informáticos. (WordReference, 2016).
Java	Es un lenguaje de programación orientado a objetos. (Definición De, 2016).

Léxico	<p>Conjunto de palabras que constituyen una lengua. En el contexto de aplicación, un léxico consiste en duplas de palabra – polaridad, siendo el segundo un valor cuantitativo/cualitativo de la palabra referida/definida que invoca sentimientos positivos o negativos en el usuario. (Significados, 2016).</p>
Líder	<p>El líder es aquella persona que es capaz de influir en los demás. El líder, sin disponer necesariamente de autoridad jerárquica, tiene también capacidad de decidir la actuación del grupo en base a la influencia que ejerce, que viene determinada por la "autoridad moral" que ejerce sobre el resto del equipo. (Liderazgo y Mercadeo, 2006).</p>
Metadata	<p>Un metadato es un dato estructurado sobre la información, o sea, información sobre información, o de forma más simple, datos sobre datos. Los metadatos en el contexto de la web, son datos que se pueden guardar, intercambiar y procesar por medio del ordenador y que están estructurados de tal forma que permiten ayudar a la identificación, descripción, clasificación, y localización del contenido de un documento o recurso web y que, por tanto, también sirven para su recuperación. (Lamarca Lapuente, 2013).</p>
Metodología	<p>La metodología hace referencia al camino o al conjunto de procedimientos racionales utilizados para alcanzar el objetivo o la gama de objetivos o tareas que requieran habilidades, conocimientos o cuidados específicos. (Aullón de Haro, 2012).</p>
Nodo	<p>En estructuras de datos dinámicas un nodo es un registro que contiene un dato de interés y al menos un puntero para referenciar (apuntar) a otro nodo. (Castells, 1996).</p>
Online	<p>Se traduce como "en línea". El concepto se utiliza en el ámbito de la informática para nombrar a algo que está conectado o a alguien que está haciendo uso de una red (generalmente, Internet). (Definición De, 2016).</p>
Organización	<p>Es un grupo social formado por personas, tareas y</p>

	administración, que interactúan en el marco de una estructura sistemática para cumplir con sus objetivos. (Definición De, 2016).
Output	Este término es de uso frecuente en el ámbito de la informática para referirse a los datos resultantes de un proceso. (Definición De, 2016).
Patrón	Aquella serie de variables constantes, identificables dentro de un conjunto mayor de datos. En informática, pueden diseñarse complejos modelos matemáticos para crear patrones más complejos. No van a resolver un problema específico, pero proporcionan una especie de esquema arquitectónico que puede ser reutilizado con el fin de acelerar el proceso de desarrollo de un programa. (Gamma y Helm, 1994).
Perfil	Es la forma en la cual las redes sociales identifican a sus miembros. Incorpora información de las personas, que varía de una red a otra, pero que incluye datos básicos e información adicional que cambia según el tipo y finalidad de la red. (ISMS Forum Spain, 2011).
Plugin	Es aquella aplicación que, en un programa informático, añade una funcionalidad adicional o una nueva característica al software. (Definición De, 2016).
Polaridad	Uno de los objetivos del análisis sentimental es determinar la actitud del usuario, respecto a un tópico, o la polaridad de un conjunto de usuarios en relación a un tema específico. Se entiende por polaridad en este contexto, a si la actitud del que expresa una opinión es positiva, neutral o negativa en relación al tópico en análisis. (Turney, 2002).
Político	Referido a la persona que se dedica a la política, interviniendo o aspirando a intervenir en el gobierno de un Estado, comunidad o municipio.
Predictivo	El análisis predictivo agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza los datos actuales e históricos reales para hacer predicciones acerca del futuro o acontecimientos no conocidos. (Nyce, 2007).

Proceso	Puede informalmente entenderse como un programa en ejecución. Formalmente un proceso es "Una unidad de actividad que se caracteriza por la ejecución de una secuencia de instrucciones, un estado actual, y un conjunto de recursos del sistema asociados". (Tanenbaum, 2009).
Publicación (Post)	Es un término inglés que puede traducirse como mensaje o anotación. En el ámbito de Internet, post se utiliza como sinónimo de artículo o nota en las publicaciones digitales como los blogs o los foros online. (Definición De, 2016).
Query/Queries	Query significa consulta. Es una búsqueda o pedido de datos almacenados en una base de datos. (Alegsa, 2016).
Recurso	Medio del que una persona se sirve para conseguir un fin o satisfacer una necesidad. (Real Academia Española, 2014).
Red Social (Canal Social)	Representa un tipo de sitio en Internet que favorece la creación de comunidades virtuales, en las cuales es posible acceder a servicios que permiten armar grupos según los intereses de los usuarios, compartiendo contenido (fotos, videos, texto en general). (Definición De, 2016).
Retweet	Es la acción de tuitear un mensaje de otro usuario. Estoy compartiendo la publicación de otro usuario, y lo hago de manera explícita, acompañando el texto con la abreviatura RT. (Significados, 2016).
Servidor	Es una aplicación en ejecución capaz de atender las peticiones de un cliente y devolverle una respuesta en concordancia. (Wiley, 2011).
Sistema de gestión de bases de datos	En inglés DataBase Management System, son un tipo de software muy específico, dedicado a servir de interfaz entre la base de datos, el usuario y las aplicaciones que la utilizan. Se compone de un lenguaje de definición de datos, de un lenguaje de manipulación de datos y de un lenguaje de consulta. (Pérez Valdés, 2007).
SOCMINT	Es la abreviatura de Social Media Intelligence. Es el estudio de las redes sociales, blogs y comunidades online, como fuentes de

	información estratégica sobre el target de una organización, su competencia y mercados. Busca minimizar la incertidumbre en la toma de decisiones. (Epsilon Technologies, 2014).
Stakeholders	Aquellos grupos que pueden afectar o ser afectados por el logro de los propósitos de una organización. (Peterson y Ferrel, 2004).
Target	Significa “objetivo” y en el contexto presente se refiere al público objetivo de nuestras acciones. (Método Marketing, 2016).
Tuitear	Refiere al hecho de comunicar/enviar algo por medio de Twitter. (Real Academia Española, 2016).
Twitter timeline	Timeline de Twitter es la página principal de Twitter en la cual aparecen los mensajes de todos los usuarios de Twitter a los cuales sigue un usuario. Es la página principal ante la cual es presentado un usuario al ingresar a la red social. (Ayuda en la web, 2016).
Tópico / Tendencia	Hace referencia a una expresión o una idea muy empleada. En las redes sociales, se conocen como trending topics, haciendo alusión a las palabras o frases más repetidas en un momento concreto. Twitter fue la red social pionera en relación a este término. (Trending topic, 2016).
Trolls	Describe a una persona que publica mensajes provocadores o irrelevantes en una comunidad online, como ser un foro, blog, o similar, con la principal intención de molestar o provocar una respuesta emocional negativa en los usuarios y lectores, con fines diversos o, alterar la conversación normal en un tema de discusión. (Fundeu, 2016).
Tweet	Es una publicación o actualización de estado, realizada en Twitter, cuya extensión máxima es de 140 caracteres. Puede contener letras, números, símbolos y enlaces. (Significados, 2016).
Username	Se traduce como ‘nombre de usuario’; es un nombre ficticio, conformado por una cadena de caracteres, con el cual uno se identifica en un foro, aplicación, servicio online, red, que al registrarse requiere la definición de una contraseña. La asociación de ambas permite al usuario acceder a su cuenta. (Botella Alonso, 2013).

<p>Usuario</p>	<p>Representa a quien usa ordinariamente algo. En informática, se asocia a cualquier individuo que utiliza una computadora, un sistema operativo, servicio o sistema. En el contexto actual, usamos el término referido todo actor que interactúa a través de las redes sociales. (Real Academia Española, 2014).</p>
<p>Vírico</p>	<p>Capacidad de expansión veloz. Relacionado con el término "virus". (The Free Dictionary, 2016).</p>
<p>Votante</p>	<p>Persona que emite su voto en una elección o consulta, siendo el voto la manifestación de la opinión, parecer o voluntad de la persona, para aprobar o rechazar una medida, un candidato o partido. Término normalmente utilizado en política. (Lijphart, 1995).</p>
<p>Warning</p>	<p>Mensaje de advertencia ante una anomalía, que implica que el flujo podría fallar. Determina que no se está utilizando el proceso / lenguaje de la forma esperada. (Arquitectura de Sistemas, 2016).</p>
<p>Wordcloud</p>	<p>Es una representación visual de las palabras que conforman un texto, en donde el tamaño es mayor para las palabras que aparecen con más frecuencia. (Hassan-Montero y Herrero-Solana, 2016).</p>
<p>Workflow</p>	<p>Es el estudio de los aspectos operacionales de una actividad de trabajo: cómo se estructuran las tareas, cómo se realizan, cuál es su orden correlativo, cómo se sincronizan, cómo fluye la información que soporta las tareas y cómo se le hace seguimiento al cumplimiento de las tareas. (Workflow Patterns, 2016).</p>

6. Introducción

6.1 Relevancia del tema

La existencia de las redes sociales en el mundo actual tiene una presencia significativa, como así también lo tiene su influencia en las acciones que tomamos en nuestras vidas diarias. En la Era de la Información que se trasciende, la creación y la circulación de contenido ya no yacen tan solo en manos de las corporaciones masivas, sino que se han convertido en parte de la vida de todo individuo.

Remitiéndose a datos de Abril del presente año, existen 1.5 billones de usuarios en Facebook, 320 millones de usuarios en Twitter, 555 millones en Tumblr y 400 millones en Instagram. (Statista, 2016). Queda en evidencia que nos encontramos frente a un volumen desmesurado de datos públicos, que podemos obviar, como también recopilar e interpretar para conocer a nuestro público, cualquiera sea el fin de dicho estudio.

Las redes sociales se encuentran contaminadas con ruido y datos intrascendentes que pueden hacer de la gestión de opiniones, una tarea dificultosa. Mediante el conocimiento de las acciones que derivan en las distintas opiniones, una entidad puede revelar información valiosa oculta en datos provenientes de las redes sociales para, en consecuencia, apoyar sus decisiones estratégicas. Por lo tanto, con el objeto de obtener una posición superior a la de la competencia, es una buena práctica invertir en herramientas analíticas de explotación de datos. A pesar de que las herramientas de medición no son perfectas, ayudan a las organizaciones a comprender la tendencia generalizada y conseguir una mayor efectividad empresarial y mejores resultados. (Marketing Directo, 2016).

Considerando a un partido político como organización, resulta pertinente exponer que en Argentina, el 30% de los votantes tiene menos de 30 años, y el 50,2% no llega a cumplir los 40 años de edad. (Telam, 2016). Si, en contraposición, se analiza el porcentaje de aquellos ciudadanos que utilizan las redes sociales como medio de comunicación, los valores son interesantes y podrían servir como punto de partida para analizar el público sobre el cual un candidato político requiere generar fidelidad.

Se utilizará el concepto actual de 'Política 3.0' para referirse al tópico en cuestión. Implica el uso de las redes sociales y de Internet para la difusión de un mensaje político y la

construcción de una imagen de marca en los medios digitales. El gran cambio tiene que ver con el canal de comunicación bidireccional que ha generado el uso de las redes. Los políticos abrieron sus cuentas de Twitter, Facebook, entre otros medios de comunicación social, y contrataron agencias de publicidad digital. En cuestión de pocos años, la presencia en Internet de los principales partidos políticos, se ha convertido en algo natural. La mayor ventaja del medio digital es su cobertura y la facilidad para propagar la información. Una pequeña inversión en marketing puede implicar la posibilidad de lograr que un tópico se vuelva trending topic en un período muy corto de tiempo. (González Blanco, 2012).

Las principales diferencias entre campañas políticas 1.0, 2.0 y 3.0, radican en el canal de comunicación, en la forma en la cual la transmiten y en el uso que le dan al feedback que reciben del público objetivo.

6.2 Objetivos

Entre los objetivos del presente estudio yace el demostrar que existen medios estratégicos no tradicionales que, de ser utilizados correctamente, pueden proveer información para deducir conocimiento que derive en ventajas competitivas superiores a las de los medios tradicionales.

Históricamente, a modo de ejemplo, la publicidad y los eventos sirvieron para abastecer a la población de estímulo y las encuestas, los censos y los cuestionarios para obtener una retroalimentación de lo percibido a causa de lo primero. La frecuencia en la obtención de datos vía estos medios antiguos es hoy ampliamente superada por la de las redes sociales, que como se mencionó con anterioridad, ya están instauradas en la vida de la población.

En los últimos años surgieron medios novedosos para tomar provecho de dicha masa de datos, los cuales se engloban dentro del nuevo movimiento de SOCMINT, el cual persigue desgranar y detallar el presente escrito. Se detallarán herramientas y algoritmos, exponiendo la teoría que los fundamenta y los casos prácticos que los justifican. También, se explayarán experiencias prácticas en el asunto, para comprender con casos tangibles la aplicación en el mundo real. Todo esto teniendo en consideración la meta final, el verificar, o no, si el beneficio obtenido en la aplicación del conjunto de herramientas provee tanta potencia y tan gran utilidad como lo anuncia la comunidad de expertos.

El fin es el señalar que existe una gran cantidad de datos disponibles actualmente, de forma pública, disponible para su explotación que suele ser obviada y podría ser de gran utilidad.

Los objetivos primarios son:

- Concluir y exponer, dentro de un marco teórico, sobre la eficiencia, ventajas, desventajas del uso de SOCMINT dentro del ámbito político.
- Describir y analizar la explotación de redes sociales como Twitter y YouTube en el contexto político.
- Detallar las distintas herramientas y técnicas de Business Intelligence (BI) usadas actualmente dentro del ámbito SOCMINT en el contexto que nos compete.

El objetivo final respecto al lector de nuestro trabajo es el de proveer los conocimientos necesarios para poder comprender una variedad de técnicas y algoritmos que existen actualmente para procesar y analizar datos de las redes sociales referidos a índoles políticas y poder aplicarlas en su contexto.

6.3 Descripción de Estructura

Damos inicio al presente trabajo en la sección de Resumen, donde brindamos una breve síntesis de lo que investigaremos, ofreciendo a su vez una traducción en la sección subsiguiente de Abstract. En Contenidos buscaremos enunciar las páginas en las que detallamos cada sección. Luego, daremos un adelanto de lo que contendrá nuestra investigación en la sección de Introducción, seguido por una descripción de los estudios ya existentes realizados sobre la materia en Estado del Arte. En Hipótesis describiremos cuales son nuestras deducciones a validar en el desarrollo de este proyecto, para luego en Metodología exponer en detalle el contenido de nuestra investigación. Derivaremos en la sección de Resultados donde discutiremos si hemos podido validar nuestras conjeturas descritas en Hipótesis. Finalmente, en la Conclusión enunciamos nuestras experiencias en el desarrollo del presente estudio.

7. Estado del Arte

7.1 Marco teórico

7.1.1 *Business Intelligence*

Business Intelligence hace referencia al conjunto de herramientas, tecnologías y metodologías orientadas a transformar datos en información e información en conocimiento para apoyar las necesidades de las organizaciones y las decisiones estratégicas dentro del contexto empresarial. (The Data Warehouse Institute, 2002). Para empresas que manejan grandes cantidades de clientes y de procesos se torna una tarea de suma dificultad el gestionar sus datos, ya sea por el volumen de los mismos o la dificultad en su integración. La variedad en la tecnología empleada, la creciente presencia de estándares, la utilización de la automatización de procesos y demás factores influyen en los tipos de datos que se maneja e impactan también en la dificultad de su gestión. El desafío yace en el poder, a través del catálogo y análisis de los datos, obtener información agregada de ellos y poder interpretar patrones, tendencias, debilidades a corregir y fortalezas a explotar. (Ranjan, 2009).

7.1.2 *Social Media Intelligence*

Con el auge de las redes sociales, y con ello la llegada de la Era la Información, se vuelve más presente la necesidad de tomar provecho del gran caudal de datos presente en la nube que conforman los medios sociales. Se ha vuelto indispensable para las organizaciones alinear su estrategia a los factores que le incumben en los medios digitales para así disminuir la incertidumbre en la toma de sus decisiones. Para satisfacer dicha necesidad surge el concepto de Social Media Intelligence, conformado por las etapas de Recolección de Datos, Análisis y Filtrado de Datos, Analítica Avanzada y Presentación de la Información. Dichas etapas son las consideradas desde una perspectiva técnica que sirve al fin del presente documento, y dejando de lado las fases vistas desde el ángulo estratégico del negocio mismo.

Etapas del Social Media Intelligence

La etapa de Recolección de Datos consiste en la definición de las fuentes de datos que se utilizan en el proceso analítico y de cuyos datos se extrae la información en las etapas subsiguientes. La pregunta es ¿De dónde se obtiene la respuesta a los problemas y las preguntas planteadas en la organización? Dentro del SOCMINT, las fuentes de datos son las redes sociales, pero existen vías distintas para obtener su contenido. Resulta fundamental establecer los datos deseados, analizar las formas para obtenerlo (mediante APIs, Cookies, Queries, entre otros) y elegir adecuadamente una de ellas.

A su vez, se debe decidir sobre el medio en el cual se almacena lo extraído, pudiendo ser una base de datos, en archivos o bien el abastecimiento o feed directo de los datos al proceso. El soporte en la variedad de los datos depende de la herramienta tecnológica que se utilice y la voluntad o no a adaptarla al formato del dato, en la mayoría de las herramientas de BI, existe soporte para bases de datos relacionales, archivos .csv, archivos de texto, entre otros. (Santiago, 2014).

En los datos suele existir mucha información que es incorrecta respecto al dominio de la realidad que pretende cubrir, y un número menor, pero en ocasiones también relevante, de datos inconsistentes. Para que los datos sean útiles y generen nuevo conocimiento, deben ser completos y consistentes, por ello resulta necesario verificar que no se encuentren duplicados, incompletos, incorrectos, faltantes, con valores nulos o metadatos pocos claros. (Orallo Hernández *et al* 2004). Con Análisis y Filtrado de Datos se refiere a todos aquellos procedimientos que se aplican para depurar, ante anomalías como las mencionadas previamente, e integrar los datos obtenidos de forma de dejar tan solo lo precisado y en la forma requerida.

Análítica Avanzada es la etapa en la que el procesado de los datos, integrados previamente, se lleva a cabo. Se aplican los algoritmos, herramientas y técnicas, logrando los flujos de datos necesarios para obtener la información deseada por el negocio. Luego, en la etapa de Presentación de la Información se implementan los procesos necesarios para presentar la información, ya producida, a los distintos stakeholders.

7.1.3 Data Mining

Data Mining, o Minería de Datos en español, es el proceso de extracción de información significativa de grandes bases de datos, con el objetivo de revelar inteligencia del negocio a través de factores ocultos, tendencias y correlaciones. De este modo, el usuario puede realizar predicciones que resuelven problemas del negocio, proporcionando una ventaja competitiva. Las herramientas de Data Mining pronostican la situación futura de una organización, a partir de la sumarización de grandes volúmenes de datos que la misma ha ido almacenando con el paso del tiempo. Esto ayuda a la alta dirección en la toma de decisiones de forma proactiva. (Pérez López, 2007).

Según algunos autores, la Minería de Datos es aquella parte de la estadística que se usa para problemas que se presentan actualmente en el Análisis de Datos. Los problemas actuales se diferencian de los clásicos en que el número de datos a analizar es mucho mayor y, como consecuencia, las técnicas estadísticas clásicas no pueden ser aplicadas. La razón principal por la cual existe una amplia disponibilidad de datos en toda organización se debe al uso de sistemas computarizados, nuevas técnicas de captura de datos, tarjetas ‘contactless’, lectores de huellas digitales, entre otros, y por el avance en la tecnología de almacenamiento y su consiguiente reducción de costos. (Maestría en Explotación de Datos y Descubrimiento del Conocimiento, 2012). El Data Mining permite resumir los datos en segmentos de información útiles, la cual puede ser utilizada para incrementar beneficios, reducir costos y tomar decisiones. Con estas técnicas es posible hacer evidente las relaciones ocultas entre eventos, las cuales resultan complejas de identificar con los métodos tradicionales. Además, permiten trabajar con grandes cantidades de observaciones sin ningún inconveniente.

La utilización de Data Mining sirve en términos generales para hallar los siguientes tipos de relaciones (Universidad Carlos III Madrid, 2016):

- Clases: Las observaciones se asignan a grupos predeterminados. El proceso de clasificación consiste en asignar un conjunto de datos a grupos fijos.
- Clusters: Se construyen grupos de observaciones similares según un criterio prefijado. El proceso de clustering, agrupamiento en español, consiste en subdividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un

grupo esté lo más cercano posible a otro elemento, y grupos diferentes estén lo más lejos posible entre sí.

- Asociaciones: Las observaciones son usadas para identificar asociaciones entre variables.
- Patrones secuenciales: Se trata de identificar patrones de comportamiento y tendencias.

Algunas áreas de aplicación de Data Mining son las siguientes:

- Toma de decisiones. Ejemplos: banca, finanzas, seguros, marketing, política.
- Procesos industriales.
- Investigación científica. Ejemplos: medicina, epidemiología, bioinformática, psicología.
- Soporte al diseño de bases de datos.
- Mejora en la calidad de los datos.
- Mejora en el área de empresas de Consulting.

7.1.4 Text Mining

La Minería de Texto, en inglés Text Mining, se origina en gran parte a partir de la investigación existente sobre Data Mining, razón por la cual tienen muchas similitudes en su arquitectura de alto nivel. Ambos tipos de sistemas tienen rutinas de pre procesamiento, algoritmos de descubrimiento de patrones y herramientas de visualización similares. (Feldman y Sanger, 2007).

La Minería de Texto es una disciplina englobada dentro de las técnicas de acceso, recuperación y organización de información y consiste en un conjunto de técnicas que permiten extraer información relevante y desconocida de manera automática dentro de grandes volúmenes de información textual, no estructurada, normalmente en lenguaje natural. Permite el descubrimiento de nuevo conocimiento y patrones a partir de un conjunto de textos, su objetivo consiste en descubrir tendencias, desviaciones y asociaciones a partir de una gran cantidad de información textual.

Una de las principales características consiste en que, por lo general, la información no está estructurada. La extracción de información de una base de datos resulta más sencilla,

ya que las mismas están diseñadas para que sea posible el tratamiento automático de la información. Por ende, es de gran utilidad, considerando que alrededor de un 80% de la información de las organizaciones está almacenada en forma de texto no estructurado. (Eíto Brun y Senso, 2004).

El valor principal del Text Mining radica en su posibilidad de acelerar el ciclo de toma de decisiones al acortar el tiempo necesario para identificar, encontrar o descubrir información relevante en contenido textual. A partir de esta premisa, permite asignar más tiempo a la etapa de analizar y comprender la información descubierta. (Maysa Consultores, 2016).

Tradicionalmente se ha utilizado para realizar Análisis Sentimental sobre datos de las redes sociales. El poder del Text Mining yace en su eficacia para extraer valor de la cantidad de datos existentes en ellas.

El Análisis Sentimental, que se puede realizar a partir del Text Mining resulta un gran reto para las tecnologías del lenguaje, ya que obtener buenos resultados es mucho más difícil de lo que se cree. La tarea de clasificar automáticamente un texto escrito en un lenguaje natural en un sentimiento positivo o negativo, opinión o subjetividad (Pang and Lee, 2008), es a veces tan complicada que incluso es difícil poner de acuerdo a diferentes personas sobre la clasificación a asignar a un texto dado. La interpretación personal de un individuo es diferente de la de los demás, y además se ve afectada por factores culturales y experiencias propias de cada uno. La tarea es aún más difícil cuanto más corto sea el texto, y peor escrito esté, como es el caso de los mensajes en redes sociales como Twitter o Facebook.

7.1.5 Network Mining

Dentro de la Minería de Datos se encuentra la técnica analítica predictiva de Network Mining, el concepto nace de la mano de las redes de índole social, redes de datos en las que distintos agentes lógicos mantienen relaciones interactivas entre sí. (Zhu, 2014).

Desde el principio del siglo XXI, la ciencia de Network Mining ha tenido un avance vertiginoso con colaboraciones de parte de sociólogos, matemáticos y profesionales de sistemas. En un afán por comprender con mejoría las estructuras sociales que se forman en las comunidades digitales, se han desarrollado numerosas teorías de redes sociales, métricas, topologías y modelos matemáticos. (Newman, 2006).

Los datos obtenidos de redes sociales, a partir de la técnica que se describe en esta sección, se diferencian de aquellos hallados en bases de datos relacionales tradicionales por poseer dos características principales, un comportamiento interactivo y una estructura de red. La primera destaca la posibilidad de encontrar patrones de comportamientos, resaltar colaboraciones y reconocer relaciones y alineamientos entre usuarios. La segunda característica resalta la existencia de patrones estructurales dentro de los datos que permiten identificar redes/mapas de vínculos sociales entre los usuarios lógicos.

El esfuerzo de los estudios más recientes, dentro del campo del Network Mining, se centra en Link Mining, en español Minería de Vínculos, y Detección de Comunidades.

En Link Mining se persigue predecir las relaciones que pueden existir entre nodos de una red, los cuales pueden representar usuarios, clientes y productos. Sus vínculos, a su vez, simbolizan relaciones sociales, colaboraciones, intercambios, entre otros. (Chen *et al*, 2012).

El Link Mining se puede ejecutar con tan solo conocer información sobre la topología de una red social; el producir dicha topología dependerá de la fuente de datos que se obtenga, en ocasiones requerirá una intervención de quien la esté analizando. Entre las técnicas utilizadas para predecir vínculos faltantes o futuros se encuentran las de “common neighbors”, Jaccard’s coefficient, Adamic Adar measure, and Katz measure. Con el agregado de los atributos de los nodos y los vínculos, la efectividad de la minería puede incrementar significativamente.

La técnicas de Detección de Comunidades también son un ámbito que está tomando mucha relevancia en el Business Intelligence. (Fortunato, 2010). Al representar redes en grafos compuestos por nodos, se pueden aplicar algoritmos de partición de grafos para encontrar subredes sociales. (Chen *et al*, 2012).

7.1.6 Data Warehousing

Es una técnica para consolidar y administrar datos de variadas fuentes con el propósito de responder preguntas de negocios y tomar decisiones. Una definición muy conocida sobre el término es la siguiente: Un Data Warehouse es una colección de datos orientados a temas integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales. (Inmon, 2002).

7.1.7 Data Mart

Un Data Mart persigue el mismo propósito que un Data Warehouse, pero es limitado en cuanto a su alcance. Dicha limitación es intencionada, teniendo un determinado número de usuarios y siendo confeccionado para un área funcional específica, se logra abaratar el costo en su construcción. El inconveniente que se encuentra comúnmente al afrontar un Data Mart yace en la consistencia de sus datos. Su función es apoyar a otros sistemas para la toma de decisiones. (Inmon, 2002).

7.1.8 API

Una Application Programming Interface, en español Interfaz de Programación de Aplicaciones, es una vía para acceder componentes de software de un sistema de forma local o remota, utilizando mecanismos de comunicación estandarizados. Es de señalar que el software hoy se construye en base a abstracciones: una API expone servicios y funciones ofrecidas por una aplicación de software para que otras aplicaciones clientes los consuman, sin necesidad de conocer su implementación interna. Las APIs desde su origen han sido un factor determinante en la evolución de la comunicación entre sistemas y su programación. (Blanchette, 2008).

7.1.9 Cookies

Una Cookie es un archivo almacenado por un explorador o servidor web que contiene metadata referente a la presencia de un usuario en una página web. A modo de ejemplo, un usuario en un explorador web podría seleccionar su idioma de preferencia en una página; la próxima vez que se conecte a la misma no precisaría de seleccionarlo nuevamente, dado que el servidor web de la página sencillamente haría una consulta a dicha Cookie y cargaría la página considerando el idioma. El contenido que se guarda dentro de una Cookie dependerá de quien haya programado el servidor web, puede incluir qué secciones de una página se visitaron, cuándo se visitó la página, qué artículos de compra se incluyeron en un

carrito de compras y hasta qué usuario y contraseña se utilizaron. En inicio, las Cookies surgieron sencillamente para ayudar con datos respecto a preferencias básicas de los usuarios; al tiempo los desarrolladores y analistas notaron que mientras más información guardasen de sus usuarios, mejor podrían acomodarse a sus necesidades. (Woodward, 2015).

Hoy en día es común que en el explorador del usuario que se conecta a una página sólo se guarde un identificador único de una Cookie, un ID, y del lado del servidor web se guarden varios registros haciendo referencia a ese ID. A la hora de cargar la página para el usuario, el servidor web obtendrá el ID de la Cookie del usuario y buscará los registros que precise para amoldar la página.

7.1.10 Google AdWords

AdWords es un servicio de publicidad online de Google cuyo fin es atraer a los usuarios de Internet a las publicidades de sus clientes. Quien utilice la plataforma para dar a conocer su organización, producto o servicio sencillamente ingresa una colección de palabras relevantes a su publicidad y los usuarios que pretende atraer. Cuando en una búsqueda de un usuario se incluyan términos indicados por los clientes de AdWords, se mostrarán las publicidades de dichos clientes en el buscador. El mismo usuario, al ingresar a una página encontrada con esos términos, también verá publicidades de los clientes de AdWords. La probabilidad de mostrarse de una publicidad se basa en un sistema de puntaje, calculado en torno a la utilidad o no de la publicidad para el usuario, y también en el monto por click que el cliente haya ingresado. (Google Support, 2016).

7.1.11 Google Display Network

Es una colección de más de un millón de páginas web, aplicaciones móviles, blogs y demás destinos online que muestran publicidad de Google AdWords y TrueView, brindando la oportunidad de alcanzar más del 80% de los usuarios de Internet. (Google Support, 2016).

7.1.12 Twitter

Es una plataforma de microblogging. La aplicación es gratuita, web y móvil. Reúne las ventajas de los blogs, las redes sociales, y la mensajería instantánea. Permite a los usuarios

estar en contacto en tiempo real con personas de su interés, a través de breves mensajes de texto que se denominan updates o tweets.

Un usuario de Twitter puede crear y publicar mensajes de 140 caracteres, mejor conocido como tweet, sobre el tópico que desee desde un navegador web o una aplicación móvil. Dicho usuario posee un username en particular que lo identifica, asociado a su perfil, e interactúa a través del mismo, en vez de hacerlo por su nombre o apellido como en otras redes sociales. La forma de invocar una comunicación con dicho usuario es precediendo el username con el símbolo arroba (@). La red social permite que un usuario A (follower) pueda seguir a otros usuarios X (followed) de su gusto; en consecuencia, A verá en su Twitter Timeline los tweets de X. A su vez, otros usuarios pueden hacer referencia o interactuar con dichos tweets, ya sea a modo de reply, retweet o like.

- Reply (respuesta): Implica responder de forma directa a un mensaje que alguien escribió en Twitter.
- Retweet (publicar nuevamente): Permite re publicar el tweet de otra persona en tu perfil, ya sea agregando un comentario o no.
- Like (me gusta): Implica aprobación o apoyo sobre un tweet de otro usuario.

Un perfil puede pertenecer a un usuario común o a un usuario verificado, y a su vez, dichos usuarios pueden ser individuos o organizaciones. Twitter verifica un perfil para confirmar la veracidad de una figura pública en la red social (usuario verificado). Como en toda red social existen perfiles falsos, es decir, que no se identifican con una persona u organización existente físicamente.

Otra alternativa de interacción entre usuarios es a través de menciones. Una mención consiste en un arroba (@) concatenado con un nombre de usuario.

A su vez, existen perfiles privados o públicos.

- Los perfiles **privados** son aquellos cuyos tweets solo pueden ser accedidos por sus seguidores.
- Los **públicos** son aquellos a cuyos tweets puede acceder cualquier usuario de la red social.

También, existen cadenas de caracteres precedidas por numerales (#), denominados hashtags. Permiten destacar términos referentes a temas en tendencia en la red social. Cuando

un usuario desee buscar tweets sobre un tema en particular, podrá sencillamente filtrar utilizando dicho hashtag. (Twitter, 2016).

7.2 Investigaciones hasta el momento

Con el fin de exponer las investigaciones existentes sobre Social Media Intelligence, y su aplicación en el marco político, se ha analizado la información que figura pública en Internet sobre su influencia la Argentina y en contraposición la del exterior, la cual es bastante más abundante en contenido.

En primera instancia se destacan estudios en Argentina que se utilizan como fundamento; existen diferentes compañías que se dedican a consultoría sobre el tópico en nuestro país, aunque es notable que lo que existe público es escaso y poco técnico, motivo por el cual resulta interesante la presente investigación. Luego se hará hincapié en lo que existe en el exterior del país sobre el tema, cuya expansión y desarrollo es mucho mayor, y más profunda.

Becom1 es una consultora Argentina, fundada en el 2009, que se dedica al marketing digital, con experiencia en planificación de medios y campañas, utilizando herramientas de BI. (Becom1, 2016). Existe poca información pública técnica, aunque tuvimos la posibilidad de entrevistar al CEO, Mariano Tato, el cual estuvo involucrado en campañas políticas de las últimas elecciones presidenciales. Compartió con nosotros algunas estadísticas interesantes, por ejemplo, que un argentino medio pasa 4.5 horas por día en las redes sociales, con lo cual analizar las salidas de fuentes tan grandes de información, es vital para la toma de decisiones; además, que la utilización beneficiosa de las redes sociales en campañas políticas puede hacer variar los resultados de unas elecciones hasta un 5%, lo cual implicaría generar una diferencia, que podría influir de forma determinante en los resultados, considerando que el 50% del electorado argentino valora la utilización de redes sociales. “Los candidatos se han dado cuenta que las redes sociales son lo más permeable que tiene el electorado para poder ser influenciado. Ya no son un satélite o algo complementario de las campañas políticas: son herramientas críticas y fundamentales, sobre todo a la hora de emitir comunicados para transmitir un mensaje inequívoco, directo, levantado por los medios de forma inmediata, y para propagarlo con velocidad.” (Tato, 2016).

Otra compañía que se dedica al rubro es **Iluminati Lab**, específicamente a la consultoría de estrategia y operaciones. (Iluminati Lab, 2016). Fundada en 2010, con sede en Olivos, Buenos Aires. Cuando nos encontrábamos en pleno desarrollo de la tesis, encontramos publicaciones muy interesantes de la compañía en cuestión, las cuales nos sirvieron para reafirmar las conclusiones de la investigación sobre el tópico. Con experiencia hace 6 años en el rubro, participaron en más de veinte procesos electorales, crisis institucionales y cuestiones de imagen pública como también en distintos movimientos y causas de índole social e incidentes ambientales. Nos aporta valor en relación a la porción del proceso que no nos dedicamos a estudiar, que es la exposición de los resultados, y sirve para dar soporte a las conclusiones que llegamos en una porción de menor escala. Sin embargo, no conseguimos entrevistarlos, para poder indagar en las herramientas que utilizan para extracción de datos de redes sociales y otras fuentes, pero si nos aportaron valor en relación a los resultados y su difusión, por ende en posteriores secciones retomaremos los estudios sobre la misma.

En el exterior, el panorama es distinto, ya que existen múltiples compañías dedicadas al rubro, y herramientas diversas que ofrecen la posibilidad de analizar grandes volúmenes de datos no estructurados, con el fin de sacar conclusiones que puedan ser utilizadas para la toma de decisiones.

Considerando el amplio espectro de información que abunda en Internet sobre el tema, y a la vez la falta de tecnicismo explícito orientado a política, la exposición del desarrollo puede ser de utilidad como punto de partida para interesados en el área. El interés se centra en aplicar el uso de estas técnicas analítico predictivas en Argentina, como una forma de pasar de una política 2.0 a una 3.0, basándose en el conocimiento de los electores, la comunicación directa, la ampliación de la cobertura, el trato personalizado a cada segmento social y la propagación de la información.

8. Hipótesis

Vivimos en la Era de la Información en la cual niños, jóvenes, adultos y también ancianos acceden a las redes sociales y comparten opiniones de forma pública. Las organizaciones pueden mantenerse al margen de los cambios que han sucedido en la última década, y continuar aplicando los medios antiguos de entender lo que el cliente o el usuario o el ciudadano necesita, según el caso. Otra alternativa sería adaptarse de forma dinámica a lo que él mismo expresa, para poder adecuar mejor el proceso y lograr el objetivo deseado.

Muchas empresas, en el ámbito comercial, utilizan Social Media Intelligence para la explotación de la información de índice cuantitativa y cualitativa, que se desprende del uso, gustos, requerimientos, opiniones y apreciaciones que los clientes poseen sobre la empresa o sus productos. Utilizan dichos datos como parámetros a la hora de tomar decisiones, y esto les permite aproximar la eficacia de las campañas de marketing. Sin duda, las empresas que más probabilidades de éxito tienen, son aquellas que aprenden a modificar sus actividades y decisiones a tiempo, de forma dinámica, a partir de las preferencias de su público.

Actualmente, una de las principales dificultades de toda organización, que tiene una participación activa en Internet, y principalmente en las redes sociales, es su transformación en un 'negocio social', el cual utiliza las nuevas tecnologías y herramientas para orientar el público de forma beneficiosa para el negocio, aprovechando la información que se encuentra explícita y pública.

El enfoque de nuestro trabajo de investigación se centrará principalmente en las siguientes hipótesis.

(1) Las organizaciones actuales almacenan un volumen desmesurado de datos, el cual ha aumentado en forma exponencial en el último tiempo. Además de la información almacenada internamente, existen muchos datos públicos, en las redes sociales, que podrían extraerse y explotarse, con el fin de aportar conocimiento para la toma de decisiones.

(2) Existen herramientas de Social Media Intelligence, disponibles y orientadas para la toma de decisiones. Seleccionaremos herramientas que sean compatibles con el tipo de análisis necesario para un partido político con el fin de demostrar su eficacia en el área, y su potencia para la dirección de una organización política.

Nuestro enfoque para el desarrollo de este trabajo será sobre la extracción y explotación de dicha información y no tanto sobre el análisis táctico empresarial o comercial

de la misma; nuestro principal fin radica en la posibilidad de acortar la brecha entre la dificultad de analizar un gran volumen de información, y la necesidad de tomarla en cuenta para decidir sobre las necesidades y el futuro de una comunidad.

Existe poca información disponible respecto a la aplicación de técnicas y herramientas en Argentina, razón por la cual la presente investigación podría ser un aporte interesante a la comunidad.

9. Metodología



FIGURA 1: SOCMINT y el ámbito Político (Politica 3.0)

Luego de analizar las redes sociales y su influencia en el ámbito político, se considera que se incorporaron como herramientas de comunicación por tres razones principales:

- La primera es la velocidad, dado que cualquier mensaje político puede ser subido desde cualquier lugar del mundo y casi instantáneamente descargado en cualquier otro sitio físico, debido a las potentes capacidades de los buscadores.
- La segunda razón es la versatilidad, porque los medios digitales admiten textos, fotos, videos, mensajes de carácter público y privado.
- Finalmente, la facilidad de su uso constituye la tercera razón. La comunicación política ya no es de arriba hacia abajo (top-down) solamente, desde los partidos a los ciudadanos, sino que ahora también es horizontal y de abajo hacia arriba (bottom-up), dado que cualquier persona conectada puede convertirse en una fuente de comunicación.

El éxito en el uso de las redes sociales depende, en gran medida, de su grado de actualización e interacción entre páginas o cuentas y seguidores. Los ciudadanos necesitan sentir que forman parte, que sus opiniones cuentan, así como construir una identidad alrededor de un proyecto político compartido. Además, su participación es crucial dado que los nuevos medios requieren usuarios activos no sólo en lo referente a compartir contenidos políticos, sino también para ayudar en el proceso de comunicación de los mensajes de campaña.

Hasta el caso de Barack Obama, el cual se desarrollará a la brevedad, las redes sociales sólo se habían mostrado como poderosas herramientas de convocatoria y como fuentes de recolección de datos. Lo siguen siendo. No obstante, se han desarrollado también como instrumentos para alcanzar y organizar a los votantes en foros de debate, lugares donde se realizan y reciben ataques políticos y espacios de creación de liderazgos virtuales y también personales. Viralizan la difusión de los actos de campaña convencionales; estimulan la participación cuando permiten que los ciudadanos aporten sus puntos de vista, críticas y apoyos al gobierno o a una determinada campaña; permiten una suerte de “coproducción”: los votantes pueden compartir o aportar contenidos, postear o retwittear.

9.1 Caso Política 3.0

La campaña de Barack Obama de 2012 significó un verdadero punto de inflexión en la utilización de las redes sociales para los fines de comunicación política. Sirvió como incentivo para que muchos jefes de campaña y políticos emplearan el uso de las redes sociales con posterioridad, por concebirlas como una garantía de triunfo. De aquí surge el concepto de “Política 3.0”.

Luego de dos años de haber ganado su primera elección presidencial, Obama sufre su primera gran derrota en los comicios de las Elecciones Legislativas Parciales de 2010. Si bien la derrota fractura los ánimos y aspiraciones de la coalición de Obama, deja cierto aire de consolución en Dan Wagner, Director de Targeting, quien predice con un alto grado de eficacia la derrota padecida.

Dan inicia su participación en la organización política en el 2009, un año antes de la derrota, recolectando información de los votantes, analizándola y proveyendo sus hallazgos a un comité encargado de dirigirse a los votantes vía email o telefónicamente. Tomando consciencia que la información utilizada consistía de encuestas que proveían datos atómicos

sobre preferencias y comportamientos, solicita al Área de Tecnología que desarrollen un sistema capaz de recolectar los datos, almacenarlos en una base de datos y proveerlos a sus modelos estadísticos. Así conforma la primera versión de su aclamado sistema, Survey Manager.

En otoño de 2009, ante la vacante de uno de los 100 asientos en el Senado del Congreso, surge una elección especial para ocuparlo. Los medios y consultoras señalan que Martha Coakley, del mismo partido Demócrata que Obama, ganaría las elecciones con seguridad. Pero, el Survey Manager predice que el resultado sería el contrario y acierta, el partido Republicano le había quitado una banca al Demócrata y Wagner lo había predicho.

La notoriedad de su eficacia incrementa exponencialmente al predecir con 5 meses de anticipación y un promedio de 2,5% de error los resultados de las Elecciones Legislativas Parciales del 2010. Para ese entonces ya había incluido en los modelos estadísticos del Survey Manager los porcentajes estimados que los programas de movilización del partido Demócrata sumarían a sus votos y sabía que no alcanzaría. La palabra de Dan a partir de ese momento sería considerada ley dentro del Comité Nacional Demócrata.

La importancia del logro de Wagner superó ampliamente su habilidad de declarar varios meses antes el éxito de las elecciones. Su enfoque generó un quiebre en relación al seguimiento de la opinión pública, a partir de la selección de muestras pequeñas de individuos, pero lo suficientemente representativas como para sacar conclusiones.

El análisis de Wagner parte originalmente de un conjunto de analistas que pensaban en el electorado como individuos, y trabajaban con el fin de generar proyecciones en relación a sus opiniones y comportamientos, hasta revelar una imagen más concreta de cada uno. Sus técnicas llevaron a la creación de un nuevo modo de pensamiento, en el cual los votantes salían de la vinculación concreta con su geografía, para pasar a considerarse como individuos, en grupos segmentados en relación a edad, género, y atributos relacionados, muchas veces utilizados con fines comerciales.

Llegado el 2012, Wagner es Director del Área de Analítica (Chief Analytics Officer) y lidera a un grupo de informáticos que mediante sistemas de software redefinen uno de los pilares de las campañas políticas: el conocimiento sobre el electorado. Proveen sistemas capaces de estimular y medir la participación de los individuos en términos de su campaña. Hasta ese momento la información respecto a los datos demográficos e intereses de los individuos y aquella relacionada a la participación de los individuos en la campaña política se

mantenían separadas, principalmente por provenir de áreas o empresas consultoras distintas. Ante la innegable eficacia de las herramientas de individualización de Wagner resulta inevitable unificar la información y poseer perfiles completos con todos los detalles posibles de cada individuo. (Issenberg, 2012).

En las elecciones del 2012 se cruzan los datos descritos en el párrafo anterior con datos de las redes sociales, aumentando aún más la focalización sobre los individuos, y así generan la posibilidad de realizar un preciso microtargeting y una segmentación sofisticada y personalizada de los votantes. Esto permite a los agentes de la campaña involucrarse de forma directa con el electorado, principalmente mediante las redes sociales, y entender qué tipo de argumentos es conveniente utilizar en cada caso, así resultando en el resurgimiento electoral de Obama. Wagner, su equipo y su tecnología son señalados como coprotagonistas en la victoria de la coalición en las elecciones Presidenciales y Legislativas del 2012. (Rutledge, 2013).

9.2 Técnicas analíticas predictivas

El caso de estudio de Barack Obama sirve como fundamento para la utilización de la información proveniente de las redes sociales para la toma de decisiones vinculadas a política. A partir esta premisa se expondrán las siguientes técnicas de extracción y análisis de datos.

9.2.1 Text Mining

Como se ha descrito en el Marco Teórico, Text Mining comprende la identificación y exploración de patrones significativos en datos textuales no estructurados. Tradicionalmente se ha utilizado para realizar Análisis Sentimental sobre datos de las redes sociales.

Entendiendo los fundamentos de esta técnica analítico predictiva, y cómo es posible identificar términos positivos, neutrales y negativos, se introducirán las bases de aplicarla a Twitter para poder obtener en grandes volúmenes de datos, conclusiones sumariadas.

Con el fin de medir el nivel sentimental de un usuario sobre un tópico, es necesario utilizar como elemento de entrada un léxico, el cual involucra duplas de palabra – polaridad, invocando sentimientos positivos o negativos en el usuario. Si bien el léxico resulta vital para poder aplicar Text Mining, también es de suma importancia poder identificar el contexto en el

que las palabras son empleadas, objeto de estudio de otra técnica que con posterioridad desarrollaremos. Se pueden definir distintos léxicos, aplicables a distintos entornos.

Para poder cuantificar el nivel sentimental de un usuario en Twitter se precisa analizar todos sus tweets, replies y retweets.

Twitter se destaca de las demás redes sociales por la característica de la información que se encuentra en ella, siendo pública, textual y en consecuencia, utilizada por gran número de profesionales y otros usuarios, que tienen la posibilidad de mantenerse informados de hechos actuales en tiempo real. Hoy puede verse como una fuente directa de acceso a temas populares y tendencias del momento. Es utilizada por medios y organizaciones como herramienta para conocer las opiniones de muestras de determinadas poblaciones en función de temáticas determinadas.

No obstante, Twitter también puede ser considerado un canal de comunicación e información, principalmente. Cada minuto, son miles de mensajes los que se vierten en esta plataforma, de hecho, uno de los colectivos profesionales que más están utilizando este medio es el de los periodistas. Es un medio perfecto para obtener información actualizada sobre cualquier tema. Sin embargo, no siempre se trata de información contrastada.

Con el fin de dar un primer acercamiento a un ejemplo de Análisis Sentimental, que con posterioridad se realizará mediante herramientas explícitas, se expondrán ejemplos de tweets en los cuales se puede interpretar términos como positivos o negativos; cabe aclarar que la interpretación humana de estos términos y su clasificación puede variar. Las palabras en verde hacen referencia a una connotación positiva, las amarillas, neutral, y las rojas, negativa. Las mayúsculas acentúan una opinión, y existen términos como “muy”, “mucho”, “poco”, “grave”, “a veces”, “siempre”, que del mismo modo, resaltan o subrayan con más o menos potencia lo que el resto de las palabras de la misma frase expresan. La técnica en cuestión se basa en el análisis de los términos textuales, sin tomar en cuenta el contexto.

Los ejemplos siguientes se refieren a tweets en Argentina, relacionados con política, y específicamente vinculados con las dos figuras con mayor popularidad en el momento actual, Cristina Fernández de Kirchner y Mauricio Macri. Es habitual el uso de abreviaciones y/o hashtags que hacen referencia a personas determinadas, lugares o eventos, y que en cierto entorno o red, todos sus miembros entienden a qué o quién se refieren.

@martinzmercedes: #LaMatanza #CFK INAUGURÓ un frente de un edificio que de noche es una zona **sin luz** y con **#inseguridad**.

@lazarolloreans: El país con **problemas muy graves** y #Macri **desesperado** por cumplir con @GrupoClarín. **No sabe** que se está poniendo la sogá al cuello.

9.2.2 Network Analysis

El Network Analysis se basa en la relación entre individuos apoyándose en la comunicación sobre tópicos específicos, como conectores entre ellos. No solo identifican redes complejas de relaciones e influencias sociales, además existen herramientas que sirven para graficar dichos vínculos, sumarizarlos, proveer mediciones y sacar conclusiones que actúen como soporte en la toma de decisiones.

Con este enfoque es posible obtener una clasificación de los individuos según sus roles en la comunidad de usuarios, con dos posturas extremas la del líder y la del seguidor. En Twitter, estas posturas se identifican como *follower* y *following*, en relación a la influencia que generan en la comunidad y a su vez relativo a un tópico particular.

Las redes sociales se consideran redes complejas, ya que presentan características topológicas no triviales, con patrones de conexión entre sus elementos que no son ni puramente regulares ni puramente aleatorios.

Considerando que las personas, organizaciones o grupos interactúan con otros, dentro de su red, identificaremos tres tipos principales de redes sociales:

- Redes egocéntricas: están conectadas con un único nodo o individuo.
- Redes socio-céntricas: son redes cerradas de forma determinada. Un ejemplo de uso común de este tipo de red es el de empleados en una organización.
- Redes de sistemas abiertos: son redes en las cuales las líneas límite no se encuentran bien definidas. El tipo de red socio-política que analizamos en este escrito, es un ejemplo de red de un sistema abierto.

El análisis de las redes sociales examina la estructura de las relaciones entre las distintas entidades sociales. Estas entidades son las personas, pero también pueden ser grupos sociales, organizaciones políticas, redes financieras, residentes de una comunidad, los ciudadanos de un país, entre otras.

9.2.3 Caso de ejemplo: elecciones de Brasil, año 2014

Se introduce un ejemplo político real sobre la utilización de la técnica, el cual resulta útil para realizar una prueba de concepto; se realizó en Brasil, durante las elecciones presidenciales del año 2014, con el fin de encontrar patrones, a partir de una clasificación regional, analizando los trending topics de Twitter. (Santos, 2015).

Inicialmente, se hizo una extracción de datos de trending topics de Twitter para 14 ciudades brasileras, Brasilia, Belén, Belo Horizonte, Curitiba, Porto Alegre, Recife, Río de Janeiro, Salvador, San Pablo, Campiñas, Fortaleza, Goiania, Manaus y San Luis. Se generaron consultas con la API REST de Twitter para obtener los 10 trending topics para estas 14 ciudades, en un lapso de 20 minutos (limitado por ciertas restricciones de la API utilizada). El filtro aplicado en las consultas para las ciudades seleccionadas se hizo especificando su Yahoo GeoPlanet WOEIDs (Where On Earth IDs). (Yahoo Developer Network, 2016). Para esta prueba de concepto, se utilizó Python y una librería de Twitter, para obtener todos los datos del día de la elección, 26 de octubre de 2014, además de los dos días anteriores. Para cada día se hicieron 70 consultas distintas, para ayudar a identificar al instante los trending topics.

Lo siguiente es un JSON, en respuesta a una consulta; el ejemplo se basa en una consulta del 26 de octubre a las 12:40 am, y solo muestras datos para Belo Horizonte.

```
[{"created_at": "2014-10-26T02:32:59Z",
  "trends":
    [{"url": "http://twitter.com/search?q=%23GolpeNoJN",
      "name": "#GolpeNoJN", "query": "%23GolpeNoJN", "promoted_content": null},
      {"url": "http://twitter.com/search?q=%23SomosTodosDilma",
        "name": "#SomosTodosDilma", "query": "%23SomosTodosDilma", "promoted_content": null},
      {"url": "http://twitter.com/search?q=%23EAecio45Confirma",
        "name": "#EAecio45Confirma", "query": "%23EAecio45Confirma", "promoted_content": null},
      {"url": "http://twitter.com/search?q=Uilson",
        "name": "Uilson", "query": "Uilson", "promoted_content": null},
      {"url": "http://twitter.com/search?q=%22Lucas+Silva%22",
        "name": "Lucas Silva", "query": "%22Lucas+Silva%22", "promoted_content": null},
      {"url": "http://twitter.com/search?q=%22Marcelo+Oliveira%22",
        "name": "Marcelo Oliveira", "query": "%22Marcelo+Oliveira%22", "promoted_content": null},
      {"url": "http://twitter.com/search?q=Cruzeiro",
        "name": "Cruzeiro", "query": "Cruzeiro", "promoted_content": null},
      {"url": "http://twitter.com/search?q=Tupi",
        "name": "Tupi", "query": "Tupi", "promoted_content": null},
      {"url": "http://twitter.com/search?q=%22Real+x+Bar%C3%A7a%22",
        "name": "Real x Bar\u00e7a", "query": "%22Real+x+Bar%C3%A7a%22", "promoted_content": null},
      {"url": "http://twitter.com/search?q=Wanessa",
        "name": "Wanessa", "query": "Wanessa", "promoted_content": null}
    ],
  "as_of": "2014-10-26T02:40:03Z",
  "locations": [{"name": "Belo Horizonte", "woeid": 455821}]
}]
```

FIGURA 2: Respuesta JSON. (Santos, 2015).

Para crear la red utilizando estos trending topics, se definieron las siguientes reglas:

- Cada ciudad es un vértice (o nodo) en la red.
- Si al menos hay un tema de tendencia en común entre dos ciudades, existe un enlace entre dichos nodos.
- Cada arista se pondera en función del número de tópicos de tendencia en común entre esas dos ciudades. Cuantos más temas en común hayan entre ambas ciudades, mayor sea el peso de su enlace.

En el ejemplo anterior, si se tuviese una arista entre Campiñas y Fortaleza, su peso sería 11, por tener 11 trending topics comunes a ambas ciudades.

Topología de red

Es el tipo de disposición en la que se distribuyen los nodos y aristas de la red. Para el caso de la red social que se analizó, la topología no cambia drásticamente a lo largo de los tres días, ya que los nodos de la red permanecen fijos. Sin embargo, se pueden detectar diferencias en los pesos de las conexiones entre los nodos, ya que el número de trending topics en común entre las ciudades, si varía.

Predicción de las elecciones utilizando datos de trending topics de Twitter

Con el fin de predecir resultados de las elecciones, se tuvo en cuenta no solamente los temas de tendencia en común entre las ciudades, sino también, cómo el contenido de estos temas se refiere en relación al posible apoyo de los diferentes partidos políticos principales, es decir, PT (Partido de los Trabajadores) y PSDB (Partido Social Demócrata Brasileiro).

Es posible, según la herramienta utilizada, aplicar tanto Text Mining como Network Mining, para el análisis, y obtener resultados de la combinación de ambas técnicas. Lo posterior se refiere exclusivamente a Text Mining, pero podría potenciarse, generando un porcentaje mayor de certeza en los resultados.

Se crearon listados de palabras y frases, para indicar inclinación positiva, neutral o negativa, en relación a alguna de las partes (partidos candidatos). Esta es de las tareas más complejas del proceso, considerando que el léxico de cada zona difiere. Para el caso de prueba, se aplicó un enfoque simplificado, por ende los resultados obtenidos podrían tener un porcentaje mucho más alto de veracidad, de aplicar un léxico como input con un trabajo más minucioso.

Habiendo hecho dicha clasificación, por cada nodo, se contabilizó el número de sus enlaces que incluyen términos indicando apoyo al partido PT, y el número de enlaces indicando apoyo a PSBD. Utilizando a Fortaleza, como ciudad de ejemplo, se obtuvieron los siguientes resultados:

- Fortaleza['PT'] = 56
- Fortaleza['PSBD'] = 37

De estos resultados se concluyó, que Fortaleza tiene una preferencia por el partido PT sobre PSBD.

Basándose en el algoritmo, el análisis generó resultados sorprendentemente similares a los de las elecciones, y más aún, si se considera que la entrada de términos positivos y negativos fue realizada desde una óptica poco profunda para un caso de estudio. Se muestra a continuación un mapa de los resultados aproximados de Twitter en relación a los resultados reales:

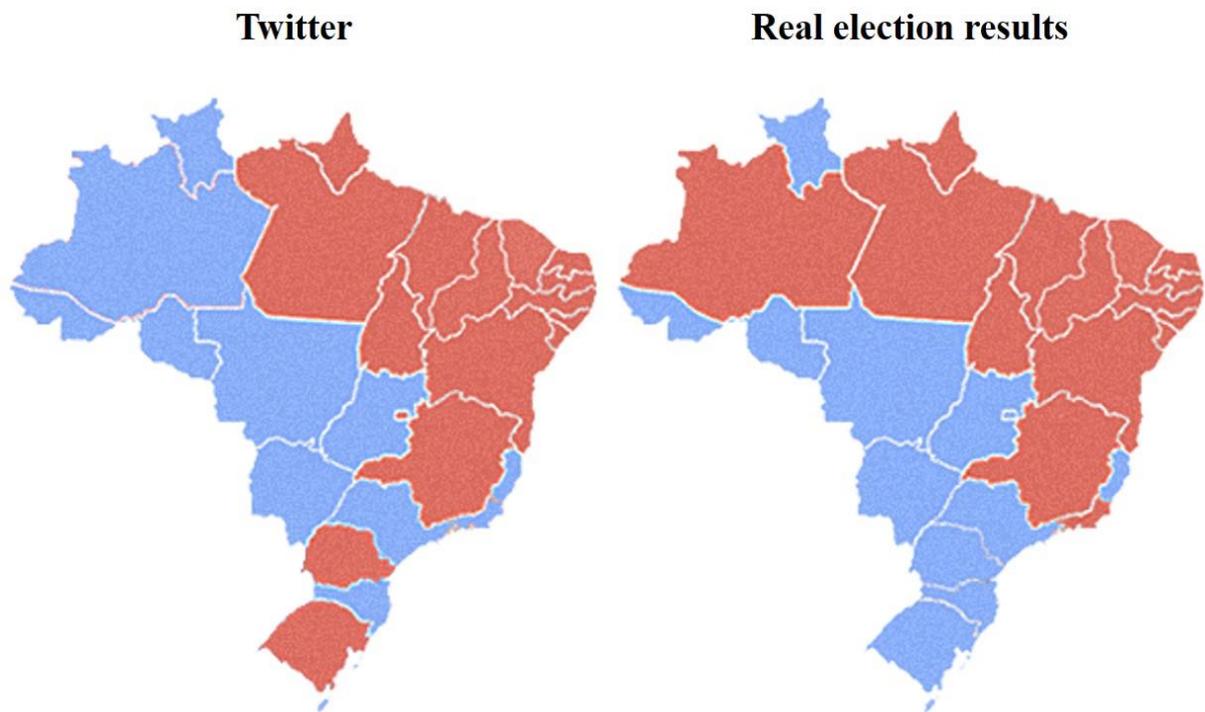


FIGURA 3: Contraste entre resultado de minería y realidad. (Santos, 2015).

Los resultados podrían mejorarse, utilizando algoritmos y sistemas de medición más sofisticados, pero sin duda, puede observarse un paralelismo interesante entre ambos mapas. Resulta oportuno destacar que, en este caso, el objetivo fue obtener conclusiones acerca de los posibles candidatos que ganarían las elecciones; pero es posible obtener información valiosa en pasos previos en un proceso de elecciones, que podría ser de utilidad para tomar decisiones acerca de cómo comunicar ciertas decisiones, o sobre el contenido que el público quiere o necesita escuchar y de esta forma, influir en los resultados.

Las siguientes métricas son de utilidad para inferir la importancia o influencia de un nodo y podrían aportar una mejora en la precisión de los resultados:

- **Centralización de nodos:** métricas para identificar los nodos con mayor influencia en una red.
- **Coefficiente de clustering:** el coeficiente de relación de un nodo mide la extensión con la cual los nodos vecinos se encuentran conectados entre sí. Puede servir también para evaluar la influencia de un nodo en el resto.
- **Grado de centralización:** es el más simple, se basa en el número de enlaces o conexiones a un nodo.

Sin embargo, sin este grado de sofisticación, los resultados sirven como evidencia para mostrar la potencia de las técnicas descritas a lo largo de este escrito.

9.3 Combinación entre Text Mining y Network Mining

Cada una de las técnicas tiene sus fines específicos. En la sección previa, se ejemplificaron la combinación de ambas, pero resulta interesante hacer hincapié en sus diferencias, y por qué su combinación genera tanta potencia. Para compararlas, se expondrán conceptos ya analizados, para generar un paralelismo concreto que aporte valor al lector.

En Text Mining se hace énfasis en la traducción textual de la información en sentimientos vía un proceso controlado, tomando en consideración un evento o suceso específico. No obstante, no contempla la relación que existe entre quien expresa un sentimiento y el resto de la comunidad que reacciona ante el mismo, así como tampoco la relevancia de quienes demuestran su valor sentimental. Tampoco considera el impacto cuantitativo en términos de lectores y actores que responden.

A su vez, Network Mining resulta muy útil para identificar como distintos individuos de una comunidad interactúan entre sí, resaltando aquellos que tienen influencia en otros y quienes asumen roles de seguidores. Esta técnica no puede por sí misma, proveer información sobre el contexto analizado.

Dado que estas dos técnicas han sido analizadas y desarrolladas en el ámbito académico, de forma independiente, no se consideró la opción de su combinación, o al menos, no con la potencia que podría darse para analizar un caso real con fines específicos. Resulta de mucha utilidad dicha combinación, para sacar provecho de sus beneficios y mitigar sus limitaciones.

9.4 Nuevo enfoque - Combinación entre Sentiment Analysis y Network Analysis

Uno de los objetivos del marketing es identificar usuarios negativos y neutrales, y en consecuencia estimularlos o volverlos positivos. Sin embargo, trabajar con todos los usuarios posibles suele resultar poco práctico y ambiguo. Existe una nueva tendencia que consiste en identificar a los líderes (quienes más influencia tienen según los demás) y actuar sobre ellos. Una mejor estrategia aún, en relación a los recursos, es la de identificar los líderes negativos y/o neutrales.

En el caso del análisis de usuarios en Twitter, primariamente resulta interesante aplicar Network Mining para obtener los pesos de autoridad de los distintos usuarios y resaltar así quienes tengan papeles de líderes en la comunidad estudiada. El siguiente paso implica aplicar Text Mining para identificar la media de la positividad/negatividad de cada uno de los líderes. Una técnica de mucha utilidad es la de confeccionar un gráfico de dispersión, posicionando en el eje X el puntaje de seguidor, en el eje Y el de líder y diferenciando cada punto con una gama de colores de acuerdo a su positividad (por ejemplo, verde = positivo, rojo = negativo).

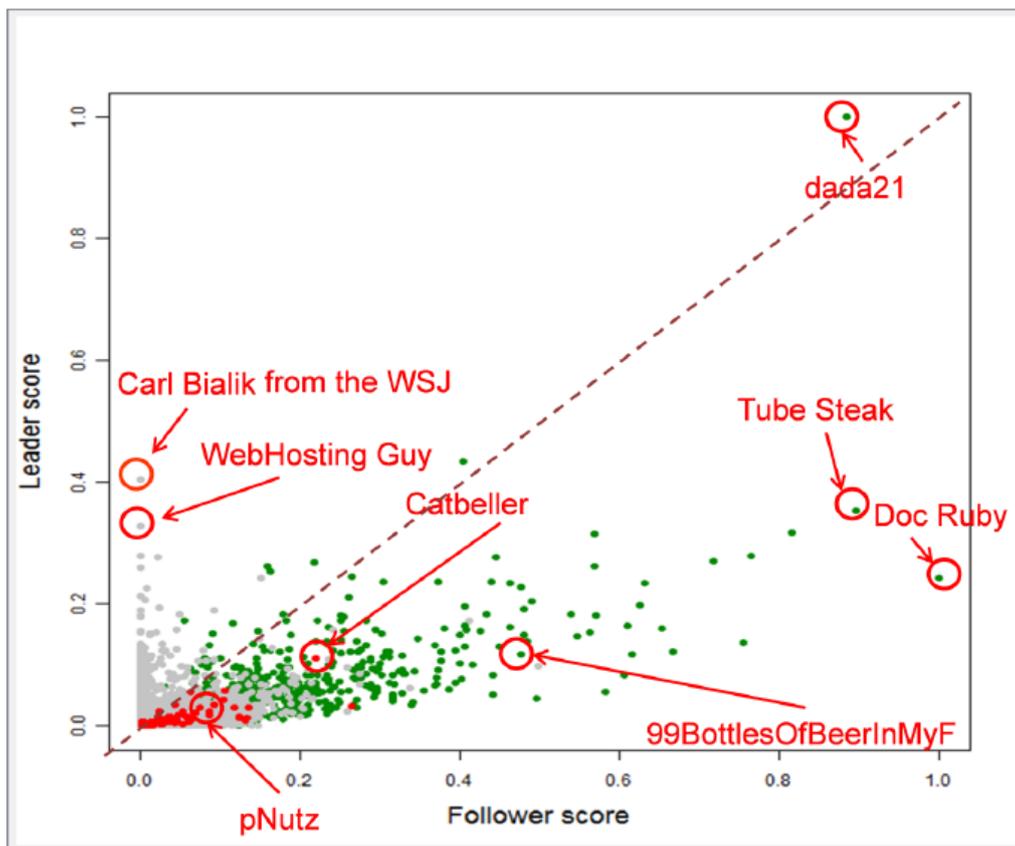


FIGURA 4: Ejemplo de resultado Sentiment Mining y Network Mining. (Thiel, 2012).

Dicha combinación de técnicas se puede utilizar como punto de partida para progresar en el análisis, apoyándose en el uso de otros métodos analíticos que usen como input de datos lo obtenido en este gráfico. Por ejemplo, se podrían extraer los usuarios líderes

negativos y neutrales y mediante ‘text mining’ confeccionar un word cloud obteniendo los términos más utilizados por ellos.

9.5 Consideraciones sobre las técnicas desarrolladas

Como todo proceso en el cual intervienen personas, existe una alta propensión a la manipulación de datos, con fines diversos. En la práctica resulta inevitable observar información ambigua, que de no descartarse, podría influir en los resultados obtenidos del análisis de los datos.

Cómo descartar el ruido

Uno de los cambios más importantes en relación a la política es su activa participación en las redes sociales. No es una mera herramienta de comunicación de un mensaje, como lo era en otras épocas la publicidad, sino que además, las redes sirven para crear estados de ánimo, tendencias, fijar eventos, e incluso influyen en los contenidos de los periódicos, radios y televisión.

Existe un juego sucio con el fin de obtener el control en las redes e imponer un discurso sobre los rivales, ya sea mediante ejércitos de robots o mismo empleados que se encuentran entrenados para desestimar los discursos digitales del rival.

Los robots, también conocidos como “bots”, son perfiles que simulan pertenecer a una persona y que son controlados mediante un software que permite que miles de ellos difundan a la vez el mismo mensaje utilizando el mismo hashtag o retwitteando el mismo post para viralizarlo con mucha velocidad comparado con lo que miles de usuarios harían en varios días.

Estos “bots” son fáciles de reconocer, dado que poseen generalmente comportamientos muy pautados. Pueden ser rastreados para evitar confundirlos con perfiles reales. En política son muy utilizados para alcanzar un trending topic, inflar artificialmente los temas de debate o desbaratar las conversaciones en las redes de los adversarios. De todos modos, esta práctica está muy mal vista por la comunidad de usuarios y más aún en el ámbito

político. Suelen ser detectadas con rapidez y denunciadas públicamente. Esto genera que sea posible filtrar el ruido generado por los robots. Existen sitios y blogs dedicados exclusivamente a la exposición de bots. (Hidalgo, 2016).

En las redes sociales y específicamente en Twitter, ha aparecido otro síndrome conocido como “trolls”. Son personas dedicadas a interrumpir, desviar o tergiverzar las conversaciones en las redes, inundándolas de ruido. Suelen infiltrarse en los foros de los rivales, generar desinformación, alterando la convivencia de diferentes opiniones.

Estas técnicas tienen como desventaja que no crean verdadero debate político, distorsionando auténticas conversaciones e impidiendo que Internet pueda utilizarse como herramienta para conocer el estado de ánimo del electorado. El vandalismo queda impune en una red global que solo se rige por cifras de usuarios y clicks en los enlaces. (Hidalgo, 2015). Un dato interesante a destacar es que en muchos casos ha ocurrido que los partidos que arrasaron en las redes, terminaron perdiendo en las urnas.

Es posible que una de las razones del estancamiento de Twitter, que ha mantenido sus 320 millones de usuarios activos durante un lapso significativo de tiempo, sea el alto porcentaje de viralización ficticia generada por robots y trolls. “Mientras Internet se rija de acuerdo a una economía basada en el número de clicks, puede que los trolls no ganen del todo, pero no será posible acabar con ellos”. (Phillips, 2014).

9.6 Análisis de herramientas de Minería de Datos

En la sección actual se desarrollará un análisis profundo de dos herramientas que son de utilidad para aplicar las técnicas analíticas predictivas que se describen.

Se seleccionaron estas herramientas por distintas razones. En primera instancia, el enfoque se basó en buscar herramientas que sean de utilidad para explorar las redes sociales y sacar patrones. Se hizo hincapié en el análisis de datos, por sobre reportería, gráficos, dashboards. Cabe aclarar que muchas herramientas se centran en la porción del proceso que analiza los resultados y orienta la publicidad y el marketing a partir de los datos provistos, y esto queda exento del análisis, por el alcance que se definió en primer lugar. Además, y aunque lo siguiente es menos esencial, se eligieron herramientas con una interfaz amigable, versión disponible gratuita, y extensa documentación online, para facilitar la capacitación, y tener dónde recurrir ante dificultades en su uso e implementación.

Con el objeto de explicar el uso de las herramientas seleccionadas, ‘Knime’ y ‘RapidMiner’, se utilizará un ejemplo de extracción de datos de Twitter y el análisis sentimental de los mismos. El foco estará puesto tanto en la extracción, como en la explotación, y se mostrará cómo es posible visualizar tanto volumen de información luego de ser procesada, pero los resultados no son el objetivo de estudio, sino una mera representación de la potencia que genera para la toma de decisiones.

En relación al ejemplo, se obtendrán Tweets a través de cuentas de usuarios de Twitter, relacionados con temas de interés político.

Resulta pertinente dar un panorama general del funcionamiento de las herramientas: cómo se estructura la información, que tipo de documentos de entrada permite recibir, con que APIs permite conectarse, y qué modos de procesamiento de información ofrece. El objetivo es dar las bases teóricas para explicar el ejemplo, facilitar el entendimiento del análisis, logrando un mejor seguimiento de la investigación desarrollada.

9.6.1 *Knime*

Knime es una plataforma de Minería de Datos que permite el desarrollo de modelos en un entorno visual. Está construida bajo la plataforma Eclipse, y programada esencialmente en Java. Su interfaz es gráfica y se comprende de nodos que se combinan de manera interactiva generando un flujo de trabajo. El carácter abierto de la herramienta hace posible su extensión mediante la creación de nuevos nodos que implementen algoritmos a la medida del usuario, y permite incorporar de manera sencilla código desarrollado en R o Python. Además, es una herramienta de software libre que puede ser descargada y utilizada gratuitamente.

La herramienta integra diversos componentes para aprendizaje automático y Minería de Datos a través de su concepto de fraccionamiento de datos modular. La interfaz gráfica de usuario permite el montaje fácil y rápido de nodos para preprocesamiento de datos, para el análisis de datos, modelado y visualización.

Componentes básicos de Knime

Un nodo es una caja previamente construida que realiza una tarea simple y única, y se la puede combinar en un flujo de trabajo para realizar una tarea más compleja. Los nodos

pueden agruparse en meta-nodos configurados de forma de trabajar como una unidad. Cada nodo recibe inputs (elementos de entrada) que pueden ser documentos/archivos que proveen datos desde el exterior, o mismo nodos, con información ya procesada dentro de la misma herramienta.

El orden en el cual se conectan los nodos no es indistinto, y requiere un estudio profundo de la herramienta, para cada proceso en particular, ya que cada nodo espera recibir cierto tipo de input, y de recibir algo que no puede procesar, arrojará una alerta, en inglés, warning, y el flujo no podrá ejecutarse.

Twitter API

Para el siguiente análisis se utilizó la API de Twitter; su utilidad radica en que permite conectarse desde la herramienta a cuentas de usuario personales, a partir de una API-key y API-secret, y un access-token y un access-token-secret (ambas claves de fácil obtención con una cuenta básica en Twitter), y luego hacer la búsqueda deseada desde dicha cuenta. Tanto el Twitter API Connector como el Twitter Search son nodos que se agregan al flujo de trabajo (workflow) con ‘drag and drop’ y tienen configuración propia para funcionar de acuerdo a lo que el usuario espera.

Para hacer este mismo análisis es posible obtener los datos de Twitter con herramientas alternativas, generando un ‘.csv’ o un ‘.xls’ con columnas identificadas para usuario, tweet, fecha, número de retweets, cantidad de followers, cantidad de follows, entre otros. Sin embargo, se tomó la decisión de aprovechar que la herramienta puede conectarse con la API de Twitter para obtener los datos de manera limpia, sin herramientas intermedias.

Nodos para Text Mining

Los nodos utilizados para el alcance de la investigación se circunscriben a los relacionados con Text Mining y a Network Mining, ambos en relación a redes sociales.

Strings to Documents: Convierte los strings dados como elemento de entrada en una tabla de elementos, en documentos (también generados en una tabla).

Column Filter: Filtra las columnas que se especifican en la configuración del nodo. Posee una tabla con las columnas que se pueden incluir/excluir, y permite al usuario tomar dicha decisión.

Subjectivity Corpus: Es un meta nodo que posee muchos nodos con los que filtra información. En nuestro ejemplo lo utilizamos para ponerle como input un archivo con palabras positivas y negativas identificadas según su polaridad, para luego poder compararlo con los datos obtenidos de Twitter y sacar conclusiones sobre los mismos.

En relación a este elemento de entrada, se presenta una dificultad: la mayoría de archivos con términos que figuran en Internet etiquetados como positivos o negativos, se encuentran en inglés. Para poder analizar los tweets de input al proceso, que en los casos de Argentina son en español, no son de utilidad. Por ejemplo, una búsqueda con la consulta “Macri”, generaría resultados con certeza en español, y no habría punto para comparar estos términos.

Por este motivo, y también porque para un análisis profundo, fue requisito agregar muchas palabras relacionadas exclusivamente con política, o palabras de uso cotidiano, que no existen o no suelen aparecer en el diccionario (‘slang words’); se creó manualmente un documento de términos.

Dictionary Tagger: Este nodo reconoce palabras y las etiqueta según el valor y el tipo; recibe como input los datos de Twitter y también el archivo de palabras positivas y negativas, y etiqueta a las palabras positivas encontradas. Luego otro nodo del mismo tipo hace el mismo procedimiento pero etiquetando las palabras con connotación negativa.

Document Scoring: Es un meta nodo que analiza y filtra a partir de las siguientes tareas:

- **Bag of words creator:** Separa los documentos proveídos como entrada en palabras únicas a partir de los espacios.
- **Terms Frequency:** Agrega una columna a cada término con la cantidad de ocurrencias del mismo en cada documento.
- **Group By:** Agrupa la tabla a partir de la columna especificada en la configuración del nodo.
- **Missing Values:** Filtra o reemplaza los campos vacíos.
- **User Scoring:** Acumula frecuencia de palabras positivas y negativas en relación al criterio del usuario/puntaje.

Numeric Binner: Definición sobre la cantidad de términos necesarios para considerar que un texto tiene una polaridad positiva o negativa, que resulta de partir de la diferencia entre ambas.

En este punto resulta pertinente hacer una aclaración, ya que existió la necesidad de reemplazar ciertos valores, considerando que la longitud máxima de un tweet es de 140 caracteres. Como es posible ingresar filas con contenido mucho más extenso que el de un post en Twitter, no es lo mismo definir una diferencia de 2 palabras positivas como parámetro para determinar que la polaridad del tweet es positivo, que definir 10. Es preciso ser minucioso a la hora de elegir estos valores, porque en caso contrario, las conclusiones de un análisis pueden resultar ambigüas, o sumamente neutrales; en un caso como el anterior, el problema radicaría en que la escala estaría desfasada para el input en cuestión y no así el proceso generado.

Color Manager: Los colores pueden ser asignados según valores nominales o numéricos. Los valores son computados durante la ejecución. Permite seleccionar colores que identifican los diferentes tipos de atributos.

Scatter Plot: Define en X los valores de una de las columnas y en Y los valores de la otra, y grafica a partir del output del color manager y a partir de la configuración dada para este nodo (en relación al tamaño del ítem, el número de columnas a mostrar, y muchas otras variantes). El ejemplo de Scatter Plot implica un input de 1000 tweets que contuvieran la palabra 'Trump', y un documento de input con palabras en inglés. Lo atractivo de este output, es la posibilidad de visualizar de forma simple e inteligible los resultados, y sacar conclusiones a grandes rasgos sobre el tema en cuestión, sin tener mucha experiencia sobre el tema o el uso de la herramienta.

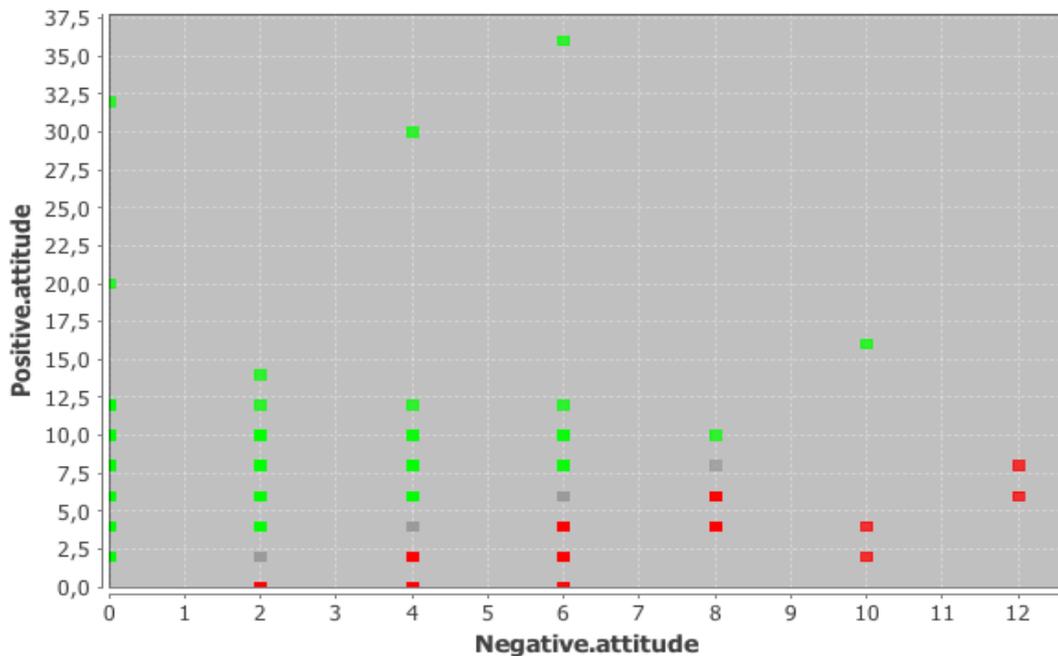
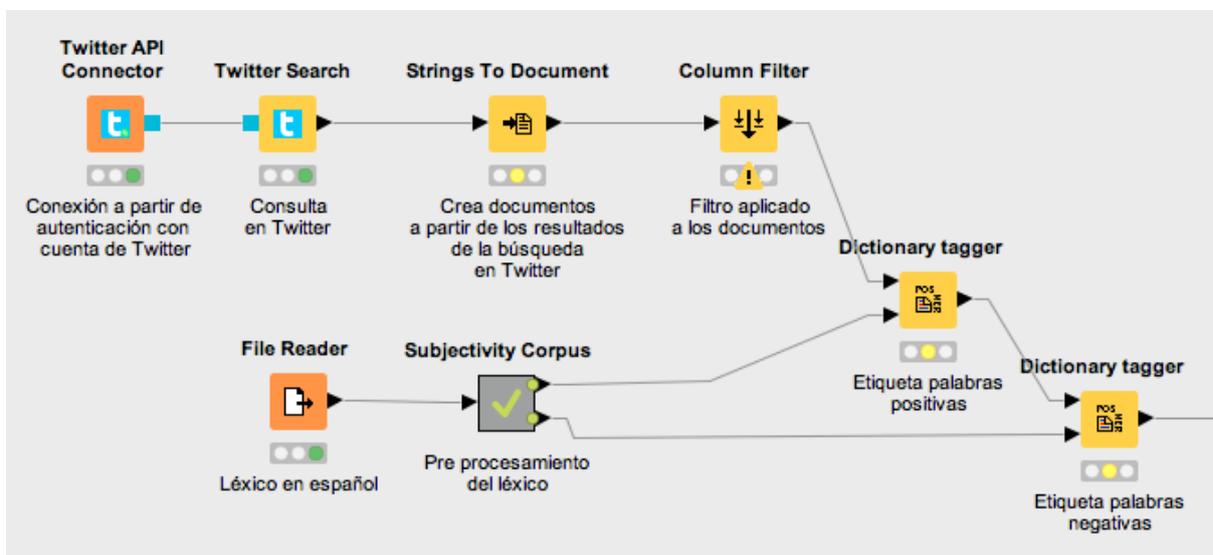


FIGURA 4: Ejemplo de Scatter Plot en Knime.

La siguiente figura ilustra el proceso que se generó, utilizando los componentes descritos con anterioridad de ‘Análisis Sentimental con Text Mining’ para obtener finalmente el resultado en un Scatter Plot. Evitando asumir que el lector conoce la herramienta, existe la posibilidad de utilizar diferentes filtros para el mismo análisis, y son múltiples las posibilidades de documentos que podrían servir como entradas del proceso, además de formas posibles de ilustrar los resultados.



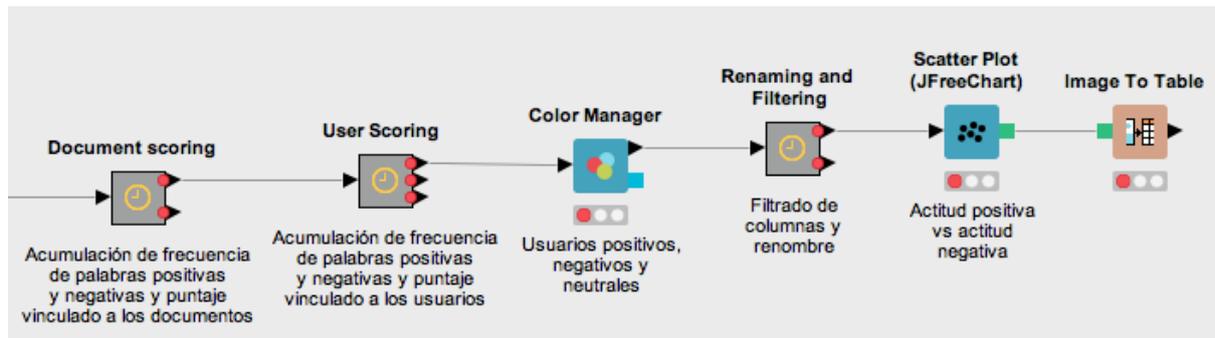


FIGURA 5: Análisis Sentimental con Text Mining.

Una limitación que se halló al utilizar la API de Twitter es que restringe el número máximo de Tweets que se pueden analizar en una búsqueda determinada; por ende, si se quiere hacer un análisis que implica mucho volumen de información en un período extendido de tiempo, es probable que se use alguna técnica intermedia para obtener la información, generando un ‘.csv’ y a partir del mismo, aplicando algún nodo como ‘*StringsToDocuments*’, se transforme dicha entrada en algo comprensible para el resto del proceso. Otra opción factible sería, volver a correr la query cada cierto tiempo, para reunir los datos, con por ejemplo, un bucle de nodos y un nodo de espera (podrían agruparse previamente por ID, para evitar duplicaciones). (Thiel, 2012).

‘Text processing’ con Knime

El plugin provee nuevos tipos de datos y nodos que permiten analizar gramaticalmente (en inglés, “parse”) documentos de texto, representar estos documentos con tipos de datos Knime y operar con ellos de múltiples maneras. Este plugin es una herramienta poderosa para integrar, procesar y gestionar datos de texto con KNIME.

9.6.2 RapidMiner

RapidMiner es una aplicación desarrollada en Java utilizada con fines comerciales, de negocio o académicos. La plataforma provee las herramientas necesarias para desarrollar análisis de negocio, minería de datos y análisis predictivo. Presenta al usuario todas las facetas de la minería de datos: preparación de los datos, procesado de los datos y obtención de patrones, visualización y evaluación/optimización de la minería.

La herramienta provee una interfaz gráfica de usuario (GUI) que permite diseñar y ejecutar distintos flujos de trabajo. Dichos flujos de trabajo son denominados procesos en la aplicación, y están conformados por distintos operadores.

Cada operador lleva a cabo una tarea específica, recibiendo una entrada de datos, procesándola y generando, en consecuencia, una salida que podrá servir de entrada para el siguiente operador. A su vez, un proceso puede contener sub-procesos que realizan una conjunción de tareas de manera transparente, en forma de caja negra, recibiendo una entrada y devolviendo una salida. La entrada de un operador puede ser, además de la salida de un proceso previo, un archivo, documento o una conexión a una interfaz de programación de aplicaciones (API) que provea un flujo de datos.

Alternativamente, las funcionalidades de los distintos operadores de RapidMiner pueden ser invocadas desde otras aplicaciones en forma de API y también presenta un modo por línea de comandos donde ofrece distintos operadores pre-programados en forma de funciones.

Los componentes que se describirán a continuación, al igual que se aclaró con la herramienta anterior, se acotan a los utilizados en la investigación para los casos que conciernen, con el fin de demostrar que RapidMiner sirve de utilidad como herramienta para el análisis sentimental y predictivo de casos orientados a política, para la toma de decisiones a partir de la extracción de datos de redes sociales.

Componentes básicos de la herramienta

Los siguientes operadores y herramientas se emplearon en el proceso de obtención de datos para realizar un análisis sentimental respecto a la búsqueda de términos políticos en la red social Twitter:

Twitter Connection: La herramienta posee un submenú para la creación y gestión de conexiones a distintas aplicaciones, brindando al usuario de los detalles técnicos y drivers necesarios para realizar las mismas. Implica la opción de una conexión a Twitter, la cual brinda la posibilidad de hacer una búsqueda, con una variedad de filtros en la red social. Para ello, es necesario ingresar el detalle de una cuenta de Twitter y proveer los permisos necesarios. El dato que solicita es un access-token obtenido con tan solo tener un usuario de Twitter.

Twitter search: Permite al usuario realizar una búsqueda y obtener tweets que coincidan con la misma, especificando de ser necesario, detalles de una búsqueda avanzada.

Data to Documents: Tomando como entrada una colección de datos, genera un documento por cada unidad dentro de la misma, utilizando los saltos de líneas como divisores. En el caso del proceso elaborado de análisis sentimental de Twitter, la colección de datos que provee el operador que antecede consiste en Tweets separados por saltos de línea, derivando entonces en una salida de un documento por Tweet.

Multiply: Este operador copia el objeto que esté presente en su entrada y lo replica en todos sus puertos de salida. Su función es justamente la de multiplicar un objeto y proveerlo como entrada para múltiples procesos. En el proceso generado, dicha multiplicación servirá dado que se analiza la positividad y la negatividad de los mismos en sub-procesos separados.

Process Documents from Files: Es un subproceso que genera un conjunto de vectores, cada uno representando una palabra, en función de un documento como entrada. El mismo permite definir internamente, mediante operadores auxiliares, como procesará el documento y que formato le dará a los vectores resultantes.

Los operadores que son de interés respecto al Análisis Sentimental son los siguientes:

Transform cases: Transforma el contenido de cada vector en minúsculas o mayúsculas para simplificar el análisis futuro. En el caso del desarrollo elaborado, el sub-proceso presente se utilizó para recibir de entrada un léxico con palabras negativas y en otra ocasión un léxico con palabras positivas, utilizando un semi-colon.

Process Documents: A fines del caso creado, se utilizó el presente nodo, el cuál es un subproceso, que tiene como entrada la lista de términos, representados por vectores, la cual

fue generada por el operador previo. Otro elemento de entrada son los Tweets, representados por documentos, generados con el operador Data to Documents. Seleccionando la opción de Term Ocurrences, en español ocurrencia de términos, el operador generará una instancia de TextObject (objeto de texto) por cada documento. Incluirá el documento original y atributos, vistos como filas, con la cantidad de ocurrencias de cada vector/término del léxico de entrada en el documento.

Generate Aggregation: Este operador genera un atributo nuevo con el resultado de la función, especificada en su configuración, sobre los atributos de su objeto de entrada seleccionados. Entre las funciones soportadas se encuentran las de generar una sumatoria, obtener el máximo, realizar un conteo, sacar un promedio, una multiplicación, entre otras.

En el actual proceso, por ejemplo, se seleccionaron todos los atributos con valores enteros resultados de lo procesado en el operador anterior y luego se eligió la opción de sumatoria (representada por “sum”). Se asignó un nombre al atributo generado y, por cuestiones experimentales, se especificó conservar todos los atributos originales (tildando la opción de “keep all”), de la siguiente manera:

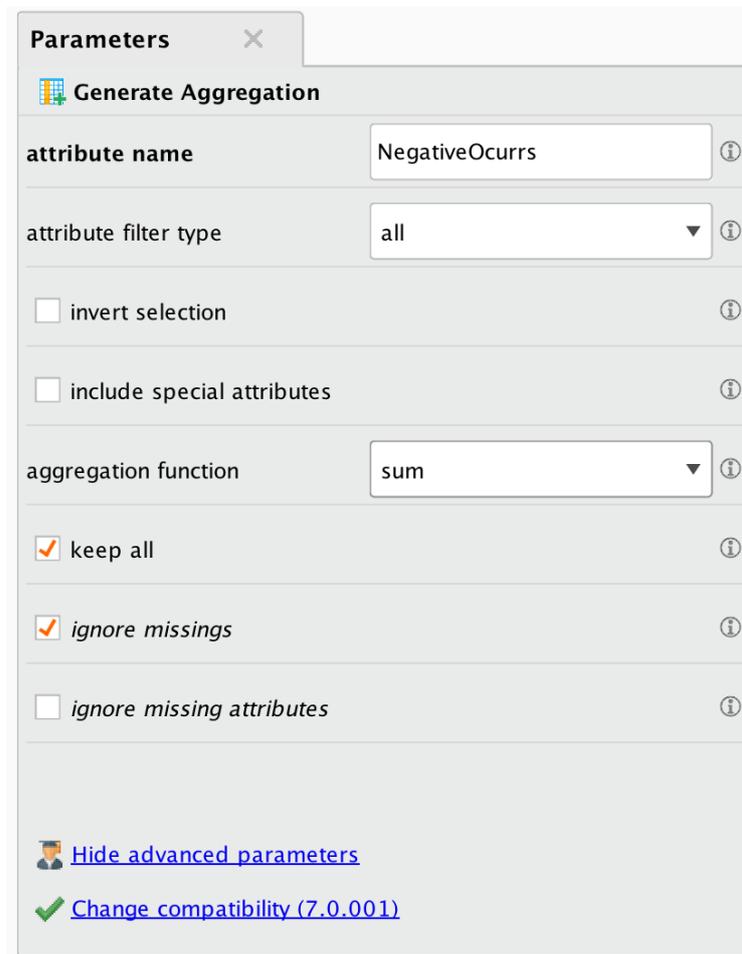


FIGURA 6: Propiedades del operador Generate Aggregation de RapidMiner.

Este operador se utilizó para generar tanto la sumatoria de ocurrencias de términos negativos, como la de positivos.

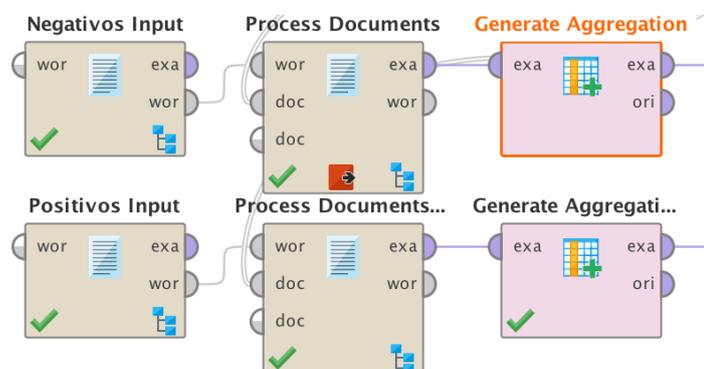


FIGURA 7: Flujo de datos Generate Aggregation en RapidMiner.

Select Attributes: Permite seleccionar un subconjunto de atributos de un objeto de entrada. En el ejemplo mostrado, dado que no es de interés quedarse con los atributos contadores de términos originales, se seleccionó proyectar tan solo los atributos resultantes de las sumatorias y los tweets/documentos originales.

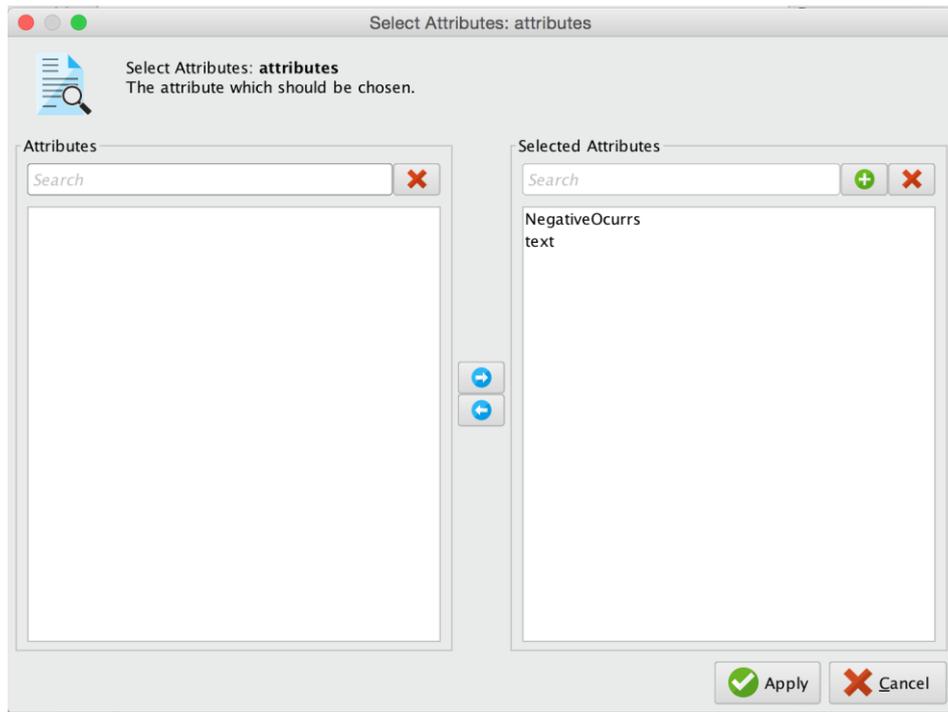


FIGURA 8: Propiedades del operador Select Attributes de RapidMiner.

Join: Este operador es utilizado para unir, mediante uno o más atributos en común, dos objetos del mismo tipo. En el caso detallado, fue de interés unir el objeto de 'sumatoriasPostivas' por tweet/documento con el de 'sumatoriasNegativas' de los mismos tweets, mediante el atributo que represente el tweet, de la siguiente forma:

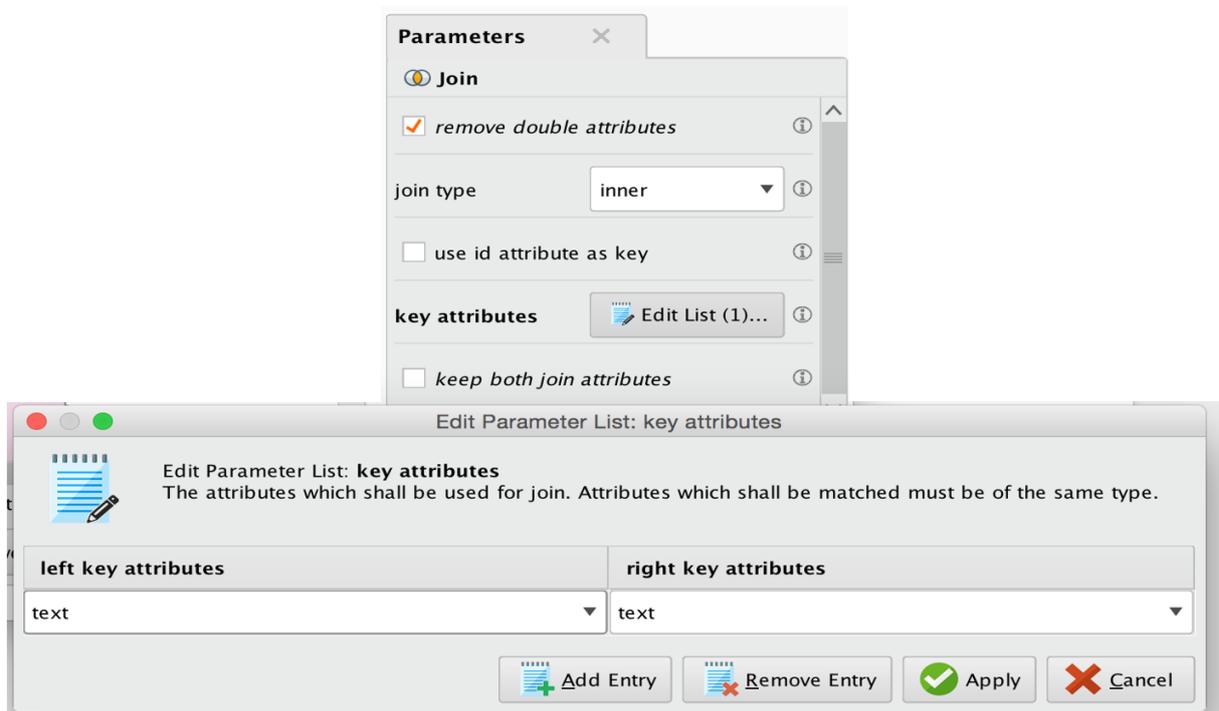


FIGURA 9: Propiedades del operador Join de RapidMiner.

Generate Attributes: El operador le presenta al usuario una forma sencilla de crear atributos nuevos en base a expresiones matemáticas; en el caso se utilizó para producir la resta entre la sumatoria positiva y la negativa:

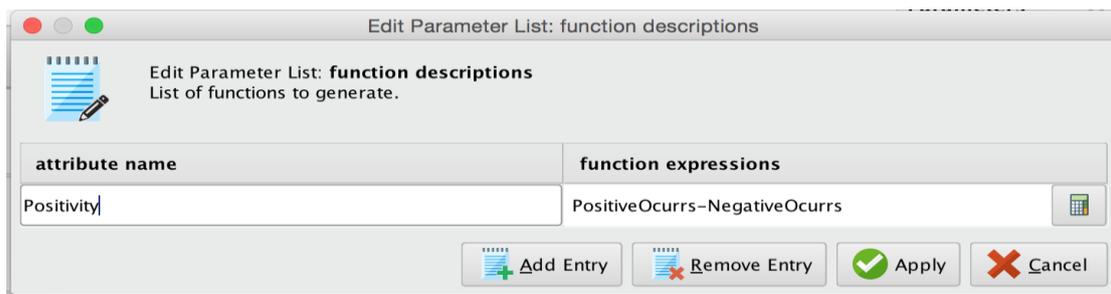


FIGURA 10: Propiedades del operador Generate Attributes de RapidMiner.

El proceso completo resultante es el de la Figura 11.

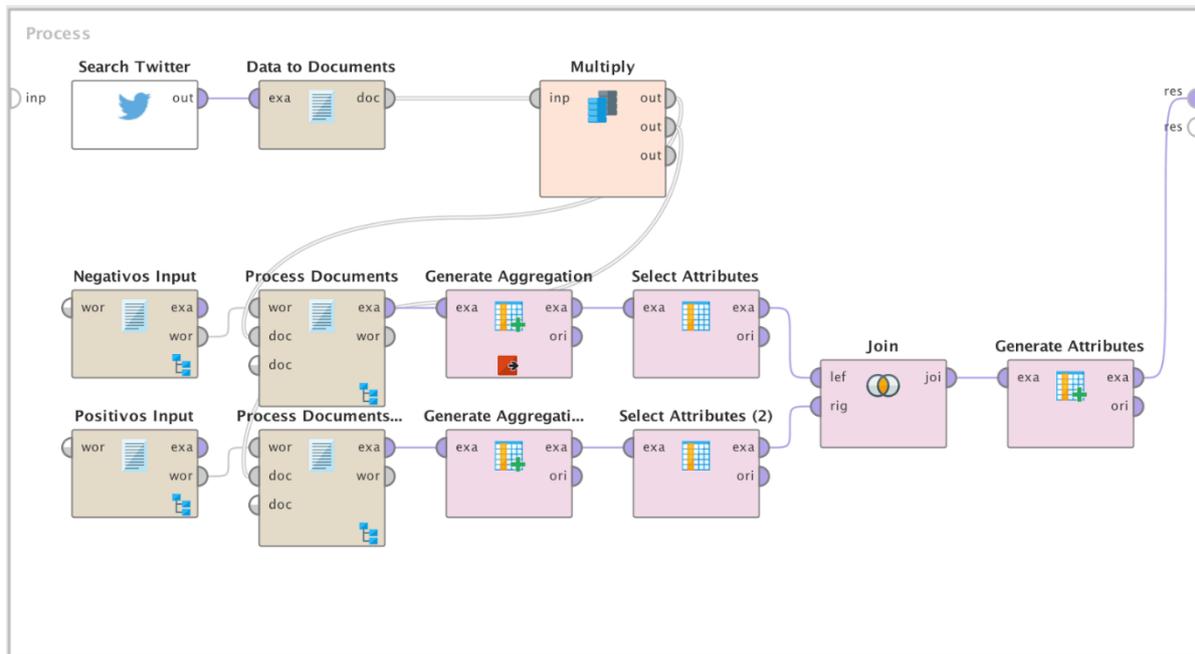


FIGURA 11: Proceso de Text Mining con RapidMiner.

9.6.3 Comparación de herramientas y conclusiones

Vinculado a las herramientas que se investigaron y analizaron, se concluyó que ambas sirven para extraer datos de redes sociales, procesarlos, seleccionar diferentes filtros customizables por el usuario, y obtener la exposición de los resultados de múltiples formas.

Knime es open-source, para análisis y creación de modelos predictivos. El modelo básico es gratuito, también lo son gran variedad de plug-ins. A su vez, ofrece actualización y soporte online en versiones pagas. Se encuentra disponible para Windows, Mac y Linux.

RapidMiner presenta una interfaz mucho más simple y permite una interacción del usuario más fluida. Es también utilizada para el análisis de texto y análisis predictivo. Ofrece una versión de prueba de 14 días y una versión libre inicial para estudiantes. También, está disponible para Windows, Mac y Linux.

Son herramientas muy potentes, con muchas más funcionalidades de las que se pudo comprobar de forma práctica. Sin embargo, los casos específicos que se analizaron, que implicaron la extracción de posts de Twitter, permitieron la comparación de tiempos de

procesamiento, que a gran escala en volumen de datos influirían en la decisión sobre cuál seleccionar. Además se encontraron diferencias en las interfaces gráficas de ambas, en la experiencia de usuario, en la cantidad de información disponible acerca de su uso que se puede observar en la siguiente tabla.

		KNIME	RAPIDMINER
Tiempo de procesamiento de Twitter Search	1.000 tweets	11,5 segundos	3 segundos
	5.000 tweets	43 segundos	39 segundos
	17.000 tweets	2 minutos 20 segundos	2 minutos
Límite máximo de Tweets por Search		Entre 18.000 y 45.000 en una ventana de 15 minutos.	Entre 18.000 y 45.000 en una ventana de 15 minutos.
Búsqueda de nodos/operadores de procesamiento		Poco efectivo. No filtra en orden alfabético por caracteres ingresados (considera que la palabra se encuentre internamente en la descripción del proceso).	La búsqueda resulta muy efectiva, sumado a esto se tiene un orden en carpetas según la funcionalidad que cumple cada nodo.
Descripción de cada nodo		El paso a paso es muy explicativo y con ejemplos claros.	No incluye un paso a paso y muchas veces tampoco ejemplos.

Interfaz gráfica	Compleja.	Simple y amigable.
Documentación sobre herramienta	Foro muy completo, mucho debate sobre cualquier cosa que googleas.	Muchos menos datos online.
Plantillas y Ejemplos	Mucha cantidad.	Tiene una oferta reducida de procesos ya armados.
Formatos de Datos	Para conocer el input y output que ofrece cada nodo, es necesario entrar a su descripción.	RapidMiner indica gráficamente el formato de los datos de input y de output de los operadores.
Complejidad en la customización	Más compleja, pero con infinitas más posibilidades.	Ofrece posibilidades de customización, pero hay ciertos aspectos que no se permite modificar, como ser los datos que obtiene en una Twitter Search.

TABLA I: Comparación de herramientas Knime y RapidMiner.

En relación al tiempo de procesamiento, en primera instancia se percibió una diferencia llamativa; Knime tarda en procesar 3.000 tweets varios minutos, porque por default, trae toda la información de cada Tweet, inclusive la imagen de perfil del usuario, lo cual genera una carga mucho mayor. Al identificar dicha falencia, se empataron los criterios para ambas herramientas, y la comparación resultó mucho más lógica, trayendo en ambos casos datos con texto simple.

Una limitación que proveen ambas herramientas es la cantidad máxima posible de Tweets que se pueden extraer en una ventana de tiempo, la cual varía si la cuenta con la cual se autentica la API es de un usuario particular, o de una aplicación. Cuando se requiere analizar un volumen grande de datos, esto puede dificultar el análisis, o mismo requerir de

una persona física que cada cierto tiempo reinicie la búsqueda. Una opción alternativa que se analizó, es la de utilizar una herramienta de extracción distinta, y proveerle al proceso de los datos en un archivo .csv. En Google Drive, existen add-ons que se pueden instalar con alguna finalidad específica: Twitter Archiver se utiliza en Google Sheets, ante una consulta busca resultados en Twitter.

En relación a la potencia de ambas herramientas, Knime posee la posibilidad de agregar muchos más subprocesos, con cualidades distintas para filtrar los datos, y ser mucho más específicos a la hora de agrupar contenido, eliminar duplicados, seleccionar léxico como elemento de entrada, seleccionar cantidad de términos que determina la polaridad de una fila, entre otros. RapidMiner, por el contrario, es más limitada en las cuestiones previamente mencionadas. Por más que fue posible analizar un caso a la par que Knime, el proceso se creó fue mucho más simple y con menos procesamiento, hasta obtener el resultado.

9.6.4 Caso real con Knime y RapidMiner

Con el fin de vincular lo analizado previamente a un caso político real, se ingresó en ambas herramientas el hashtag ‘#ImpunidadParaNinguno’, trending topic en el contexto político argentino. Se limitó el número a una cantidad de 5000 tweets, obteniendo un 25% negativos, 2% positivos y 73% neutrales para Knime, y 28%, 3% y 69% en RapidMiner. Como se pudo observar en la comparación de ambos porcentajes, se hallaron similitudes importantes. Para empatar lo más posible los inputs del proceso, se aprovisionó a ambos el mismo léxico, formateándolo para que ambas herramientas puedan recibirlo y leerlo. El léxico seleccionado fue en español, obtenido de un paper sobre el tema en España. Al mismo, se le agregaron manualmente muchos términos propios de la lengua, y utilizados en política, analizando tweets en dicho contexto. (Pérez-Rosas et al, 2012). Esto ayudó a que ambas llegaran a resultados similares.

Row ID	S Author	I Uniqu...	D Sum(...	D Sum(P...	D Good...	S Good...	S Good...
Row630	Guillermo Lau...	3	0	8	8	good	positive
Row74	- Axis	7	4	10	6	good	positive
Row51	- Alejandra	2	0	4	4	good	positive
Row52	- Alejandro#...	2	0	4	4	good	positive
Row66	- Anny3	6	4	8	4	good	positive
Row69	- Ariadna	4	2	6	4	good	positive
Row140	- Gustavo	4	2	6	4	good	positive
Row187	- Lili	2	0	4	4	good	positive
Row205	- Manchala	2	0	4	4	good	positive
Row247	- Olajodida	4	2	6	4	good	positive
Row334	- carolina	2	0	4	4	good	positive
Row351	- grace	2	0	4	4	good	positive
Row365	- la.lic	2	0	4	4	good	positive
Row443	Alejandro Las...	1	0	4	4	good	positive
Row452	Ana Lia Trevis...	2	0	4	4	good	positive
Row475	Argentina de ...	2	0	4	4	good	positive
Row478	Ariel Misciagna	1	0	4	4	good	positive
Row502	Calamardo d...	2	0	4	4	good	positive
Row515	Carmen Gentile	2	0	4	4	good	positive
Row557	De Alsina	1	0	4	4	good	positive
Row563	Dra.Catalina ...	2	0	4	4	good	positive
Row568	ESCORPIANA,...	2	2	6	4	good	positive
Row575	El Holandés	2	0	4	4	good	positive
Row812	Marullo Claudia	6	4	8	4	good	positive
Row1067	fernando boneo	2	0	4	4	good	positive
Row1080	ingrid c	4	2	6	4	good	positive

FIGURA 12: Resultado de usuarios positivos en caso de estudio ImpunidadParaNinguno con herramienta Knime.

Row ID	S Author	I Uniqu...	D Sum(...	D Sum(P...	D Good...	S Good...	S Good...
Row12	- #ChauBotin	1	0	2	2	good	neutral
Row13	- #ChauBotin!	1	0	2	2	good	neutral
Row14	- #Justiciadel...	1	0	2	2	good	neutral
Row23	- ...Cris...	4	4	6	2	good	neutral
Row24	- :G:.	1	0	2	2	good	neutral
Row39	- APABULLADA	1	0	2	2	good	neutral
Row41	- ARGENTINA	1	0	2	2	good	neutral
Row42	- Aby	1	0	2	2	good	neutral
Row44	- Adrian	1	0	2	2	good	neutral
Row55	- Alibita	1	0	2	2	good	neutral
Row57	- Amada	1	0	2	2	good	neutral
Row63	- Andreita	1	0	2	2	good	neutral
Row65	- Anhedonia	1	0	2	2	good	neutral
Row67	- Any	1	0	2	2	good	neutral
Row68	- ArgentinoDi...	2	2	4	2	good	neutral
Row70	- Ariana	2	2	4	2	good	neutral
Row77	- BOSTERO	1	0	2	2	good	neutral
Row81	- Betty	1	0	2	2	good	neutral
Row82	- Bibiana.md...	1	0	2	2	good	neutral
Row85	- Burchi	1	0	2	2	good	neutral
Row87	- CLOET	1	0	2	2	good	neutral
Row95	- Chino	1	0	2	2	good	neutral
Row97	- ClauDeLasP...	1	0	2	2	good	neutral
Row98	- ConditaBost...	1	0	2	2	good	neutral
Row100	- CosasArgen...	1	0	2	2	good	neutral
Row102	- Cristian	1	2	4	2	good	neutral

FIGURA 13: Resultado de usuarios neutrales en caso de estudio ImpunidadParaNinguno con herramienta Knime.

Row ID	S Author	I Uniqu...	D Sum(...	D Sum(P...	D Good...	S Good...	S Good...
Row679	Juan Pablo Ro...	2	4	0	-4	bad	negative
Row680	Juan Solveyra	2	4	0	-4	bad	negative
Row685	Julio Vatt	2	4	0	-4	bad	negative
Row692	Kenny Vive	3	6	2	-4	bad	negative
Row712	Leandro Zam...	1	4	0	-4	bad	negative
Row715	Leonor Salazar	1	4	0	-4	bad	negative
Row718	Lic. Nico Tara...	2	4	0	-4	bad	negative
Row723	Liliana Laura ...	2	4	0	-4	bad	negative
Row724	Liliana Valsagna	1	4	0	-4	bad	negative
Row737	Luis Mazza	2	4	0	-4	bad	negative
Row739	Luján Smart	2	4	0	-4	bad	negative
Row742	M Gondra 22F	1	4	0	-4	bad	negative
Row756	Maite Martinez	3	6	2	-4	bad	negative
Row764	Marcela Medina	2	4	0	-4	bad	negative
Row768	Marcelo Villan...	2	4	0	-4	bad	negative
Row769	Maria Jose Sa...	1	4	0	-4	bad	negative
Row781	Maria Torre	2	4	0	-4	bad	negative
Row784	Marian Cavan...	1	4	0	-4	bad	negative
Row785	Mariana Nana	1	4	0	-4	bad	negative
Row788	Mariavaggem...	2	4	0	-4	bad	negative
Row789	Mariel L.	2	4	0	-4	bad	negative
Row790	Mariela Berdina	2	4	0	-4	bad	negative
Row791	Mariela G.	1	4	0	-4	bad	negative
Row832	Myriam de la ...	1	4	0	-4	bad	negative
Row836	NANCY MAZZEO	2	4	0	-4	bad	negative
Row838	Nacho Ramón	1	4	0	-4	bad	negative

FIGURA 14: Resultado de usuarios negativos en caso de estudio ImpunidadParaNinguno con herramienta Knime.

Las anteriores figuras muestran los resultados obtenidos en Knime, en forma de tabla con contenidos, para Tweets con polaridad positiva, neutral y negativa. Resulta interesante su observación para entender como los datos se exponen de forma simple de interpretar para un usuario, considerando que el mismo puede tener o no experiencia en relación a la herramienta.

Posteriormente es posible observar un gráfico de dispersión para los mismos datos de las tablas anteriores y como se distribuyen los usuarios espacialmente según su polaridad.

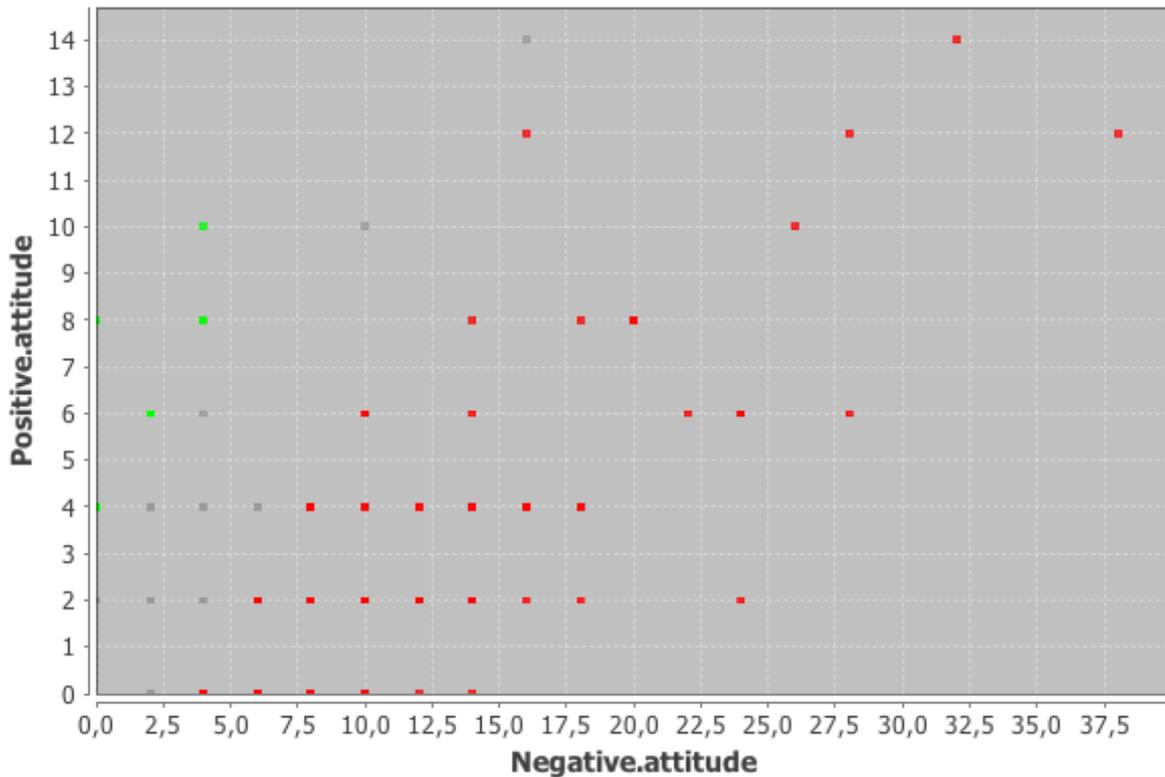


FIGURA 15: Gráfico de dispersión final para caso de estudio #ImpunidadParaNinguno con herramienta Knime.

Es importante destacar que la conclusión implica que existe un mayor porcentaje de tweets que demuestran descontento del público con el tópico en cuestión. Al mismo tiempo, en los medios, el presidente ha optado por hablar al respecto para calmar los ánimos y prometer que definitivamente no habrá “impunidad para ninguno”. (Romero, 2016).

NACIONAL | NUEVO ESCENARIO

Mauricio Macri presentó a su gabinete

El presidente electo prometió que 'no habrá impunidad para ninguno' de los integrantes en eventuales hechos de corrupción y aclaró que los problemas del país 'no se resuelven de un día para el otro'.



El presidente electo, Mauricio Macri, presentó hoy su Gabinete y prometió que 'no habrá impunidad para ninguno' de los integrantes en eventuales hechos de corrupción y aclaró que los problemas del país 'no se resuelven de un día para el otro'. El jefe de Estado

FIGURA 16: Título de artículo “Mauricio Macri presentó a su gabinete”. (Romero, 2016).

La razón principal por la cual existe un porcentaje tan alto de tweets con polarización neutral se debe a que cada post en Twitter tiene pocos caracteres, por ende para que exista una buena diferencia entre términos positivos y negativos, tiene que ocurrir que coincidan con los del léxico y a su vez que sean lo suficientemente enfáticos como para determinarlos.

Sin embargo, es importante notar que la diferencia entre el porcentaje de positividad y de negatividad implica una brecha significativa y contundente, que permite a simple vista entender lo que le genera al público.

9.7 Posicionamiento y focalización de publicidad en YouTube

Hemos debatido en ocasiones que la inmensa cantidad de datos facilitada por la tecnología actual ha potenciado enormemente el termómetro social que una entidad política puede tener sobre una población objetivo. Esto no solo se traduce a una mejor noción de la percepción de una sociedad; también se traduce a una mejor individualización a la hora de dar a conocer su imagen vía la publicidad online.

Vimos como con algoritmos de Minería de Datos, y herramientas basadas en ellos, una entidad pública puede obtener información muy ajustada sobre patrones, relaciones y comportamientos de una comunidad en términos de distintos contextos. La información obtenida con estos algoritmos y el consecuente conocimiento generado es utilizado para alimentar diversas acciones estratégicas buscando lograr una mejor imagen o relación con una población; una de estas vías es la publicidad online.

El poseer información producida por las técnicas de los capítulos previos respecto a opiniones y nociones de distintos segmentos de individuos, permitirá definir tanto el contenido de la publicidad a compartir, como los espectadores potenciales a quienes dirigimos. De saber que porcentualmente cierto subconjunto de una comunidad tiene una mala percepción de tal asunto podría alentar a contenido que busque aclarar el mismo solo para el subconjunto en cuestión.

En la publicidad, el medio con mayor potencia para dar a conocer ideas es el video, siendo este visual, sonoro y animado. Existe en la Internet una red social en particular donde la publicidad vía este medio es elemento de su funcionamiento y existencia, YouTube. YouTube permite a usuarios registrados subir sus videos a sus canales para que otros usuarios registrados o anónimos los visualicen. A su vez, los usuarios registrados pueden suscribirse a canales, comentar sobre contenido y compartirlos en otras redes sociales, dándole una propiedad social notoria. Posee más de mil millones de usuarios y más de 4 mil millones de vistas por día. Asimismo, los videos de YouTube pueden ser vistos desde otras páginas, actualmente 323 días de extensión de contenido es visto a través de Facebook cada minuto. (Smith, 2016). Según la página que analiza el tráfico online Alexa.com, YouTube es actualmente la segunda página más frecuentada, dejando en evidencia el objetivo de espectadores potenciales que una publicidad en la misma podría alcanzar. (Alexa – Actionable Analytics for the Web, 2016).

En época de procesos electorales es común ver YouTube inundada con publicidades de partidos u organizaciones políticas, también es común ver publicidades de YouTube en páginas de deportes o noticias. Es frecuente ver que las publicidades proyectadas por persona suelen ser del mismo partido político o referir al mismo tópico social de nuestro interés, a continuación se explicará por qué las publicidades que se nos muestran poseen tal homogeneidad y suelen estar tan en línea con nuestros intereses.

Las Cookies juegan un rol central en la publicidad de Youtube. TrueView es el subsistema de AdWords utilizado para publicitar en video un negocio o entidad pública en YouTube o cualquier página que esté dentro del Google Display Network. Allí el primer paso es el de subir un video a YouTube, el cual hará de unidad publicitaria a ser mostrada. Luego se indica el presupuesto con el que se contará y, lo que más aporta a nuestro fin, se selecciona el “target”, objetivo en español, al cual se le mostrará el video publicitario. (West, 2013). Ahora bien, dicho objetivo se define en función de diversos parámetros: (Google Support, 2016).

- Grupo demográfico: Se selecciona la edad, género y estado civil de la audiencia que se busca alcanzar.
- Intereses: Se selecciona entre categorías predefinidas para alcanzar personas interesadas en ciertas temáticas aun cuando estén visitando páginas sobre otras índoles.
- Remarketing de videos: Permite alcanzar usuarios en base a sus interacciones pasadas con videos previos, el canal de videos o las publicidades anteriores del cliente.
- Posicionamiento: Apuntar a canales únicos, páginas o ubicaciones dentro de las mismas. Por ejemplo, se podría posicionar la publicidad en blogs o páginas de noticias con un mucho tráfico.
 - Las ubicaciones incluyen:
 - Canales de socios de YouTube.
 - Videos dentro de YouTube.
 - Distintas páginas, donde se verá la publicidad teniendo a YouTube.com como editor y publicador.
- Tópicos: Distinto al Interés, un Tópico refiere a contenido dentro de YouTube o el Google Display Network. Mediante la definición de tópicos, la publicidad permite

alcanzar videos o páginas relacionadas a lo indicado. Por ejemplo, si se selecciona el tópic "Automotores", se le mostrará la publicidad a usuarios observando videos o páginas sobre Autos.

- **Keywords/Palabras Claves:** Dependiendo del formato de video publicitario, se podrá mostrar el mismo basándose en palabras claves o frases relacionadas a videos o canales de Youtube o páginas que estén asociadas a las palabras claves indicadas.

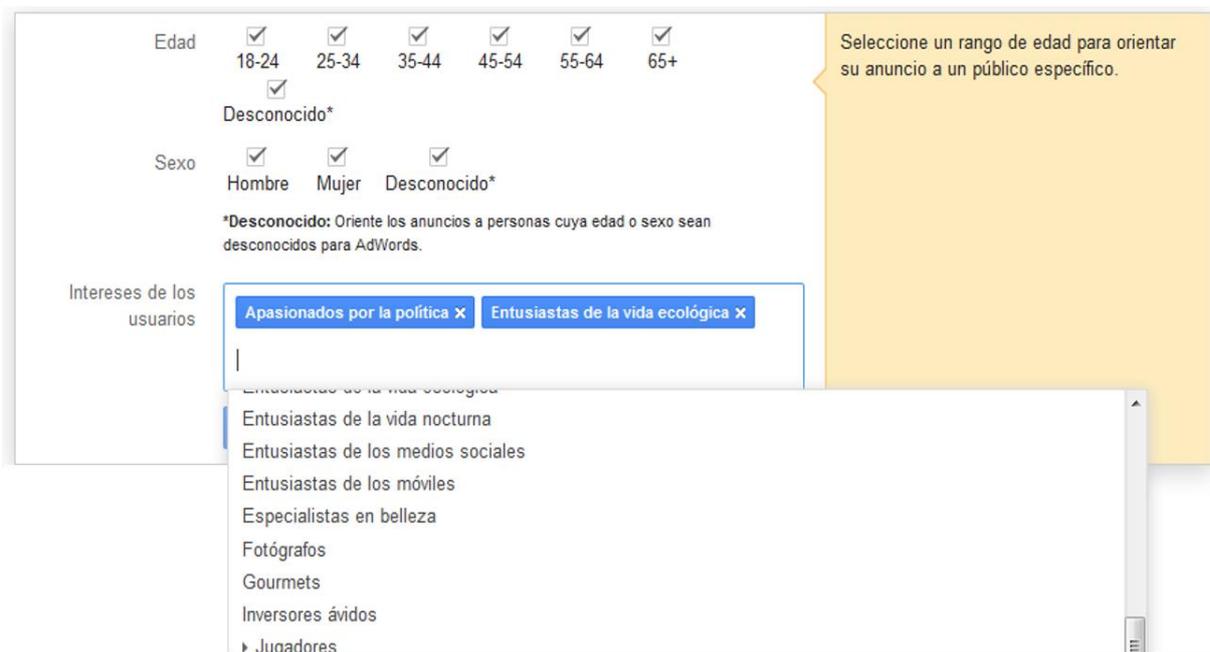


FIGURA 17: Opciones de focalización de objetivo en TrueView para YouTube.

Aquí la pregunta en cuestión es la siguiente: ¿cómo hace TrueView para conocer los datos demográficos y de interés mencionados?

Es justamente allí donde entra en juego la potencia de las Cookies que se describieron en el Marco Teórico.

Una cookie publicitaria suele tener el siguiente contenido: (Geary, 2012).

```

tiempo (time): 06/Aug/2008 12:01:32
id_usuario (userid): 0000000000000001
id_publicidad (ad_id): 1003
id_ubicacion_publicidad (ad_placement_id): 105
tiempo_reproducido (time_played): 29
    
```

es_publicidad_presionada (is_ad_clicked): 0
ip_cliente (client_ip): 123.45.67.89
url_referencia (referral_url): "http://youtube.com/categories"

Esto le indica a YouTube, respectivamente, el tiempo y la fecha en la cual se vio una publicidad, el identificador único que se le ha otorgado al explorador del usuario, el identificador único de la publicidad, el identificador único de la ubicación en la página en la que se vio la publicidad, el tiempo reproducido del video publicitario, si se presionó el link de la publicidad, la dirección ip del usuario, en que página se encontraba el usuario cuando vio la publicidad.

Porque se ha registrado la dirección IP del usuario, se puede tener una buena noción de la ubicación geográfica sin la necesidad que el mismo esté logueado. En caso que el usuario esté logueado en YouTube, se la agregan datos respecto al mismo a la cookie. (The Guardian 'DoubleClick (Google): What is it and what does it do?', Joana Geary, 2012)

TrueView, como otras plataformas de publicidad online, a través de la utilización de Cookies, permite tener características de los individuos entre los cuales filtrar el objetivo seleccionado en base a lo obtenido con herramientas de Minería de Datos.

Entonces, no solo es posible obtener una visión acertada de lo que opina un individuo, sino también alcanzarlo con contenido audiovisual que le sea relevante para intentar lograr un estímulo y una reacción, e incluso posteriormente medir el impacto de lo mismo. Las herramientas existentes en la tecnología actual entonces permiten medir cómo influir la percepción de una sociedad, facilitando la labor de mantener una organización de forma social y disminuyendo la incertidumbre involucrada.

10. Resultados

La utilidad estratégica que presenta el poder obtener los perfiles de individuos de interés para una organización, vía las redes sociales, no tiene precedentes. Antiguamente se debía acudir a encuestas telefónicas o presenciales, cuestionarios, censos y demás técnicas que consumían esfuerzos tanto por su ejecución como por su procesado posterior. La llegada de la Era de la Información, donde la cantidad desmedida de datos resulta tangible, persistente y en muchas ocasiones pública, brinda la posibilidad de disminuir el esfuerzo formidablemente. Sumado a esto, el potencial que existe en la obtención de información resulta enigmático.

Vemos en la sección de Metodología que el esfuerzo en el tratado de los datos no se ve disminuido por sí solo, es una consecuencia de las técnicas empleadas. No ofrece resultados significativos el aprovechar lo generado por las redes sociales mediante análisis unitarios de Tweets o publicaciones en blogs y consecuentes provisiones a planillas de Excel. Los datos poseen ruidos, en ocasiones son inutilizables y pueden resultar inservibles en ciertos contextos o por sí solos. El emplear herramientas que obtienen los datos, los filtran y los procesan vía la materialización de algoritmos estudiados sí disminuye la obtención y el filtrado de los datos de forma notable. Hemos visto en la experiencia del uso de Knime y RapidMiner que el esfuerzo requerido en el empleo, configuración y prueba de algoritmos en estas herramientas no requiere un esfuerzo cuantioso y la recompensa aporta una significativa cuota de valor.

En el caso expuesto en la sección 8.6.4, del hashtag #ImpunidadParaNinguno, pudimos comprobar con claridad la potencia que nos presentan las herramientas en el tratado de grandes volúmenes de datos. Para el caso, se utilizan 5.000 tweets, considerando un promedio de 67,9 caracteres cada uno, llegamos a un aproximado de 339.500 caracteres. (Parikh, 2014). Adicionalmente, siendo que en el idioma español se posee una media de 8,8 caracteres por palabra, se pueden estimar 38.579 palabras. (Koh, 2016). Es decir, se tarda aproximadamente 40 segundos en procesar 38.579 palabras emitidas por 5.000 personas distintas y obtener resultados de los ánimos en torno al tópico en cuestión. Se reduce sin ningún espacio a la duda el tiempo que emplearía procesar tal cantidad de datos de forma manual o mediante procedimientos rudimentarios. La generación del flujo que incluye nodos

u operadores para procesar los datos puede implicar un cierto tiempo dependiendo de la experiencia del profesional, e inclusive requiere de capacitación para el uso de la herramienta seleccionada. Sin embargo, luego de tener el proceso armado, la tarea operativa de correrlo no requiere de experiencia en el área, simplemente de saber dónde ingresar ciertos datos y como correrlo.

Habiendo transcurrido el proceso de desarrollo de esta investigación, creemos considerable la potencia que existe en la aplicación de las técnicas analítico-predictivas. Mediante el uso de estas herramientas una organización política puede hoy tener una percepción acertada respecto a los ánimos de una sociedad en torno a cualquier tópico, individuo o evento presente en la comunidad. Identificamos que el estudio que desarrollamos se centra en una visión acotada en relación a este análisis que surge efecto en el público objetivo. Queremos destacar que las herramientas analizadas sirven para observar por períodos prolongados las reacciones del público y cuanto más información se almacena, mayor es la chance de sacar conclusiones que aporten valor a la toma de decisiones.

Un ejemplo que sirve para tomar noción de esto es la siguiente exposición de resultados sobre la opinión del público acerca de nuestro presidente, Mauricio Macri: el análisis se realizó durante seis meses consecutivos, el primer semestre de su gestión en el cargo, y resulta interesante porque se extiende en un período considerable de tiempo. Nuestro análisis se acotaba a una visión más micro, sobre un tópico en particular, y en un período corto, por ende aplicar la misma técnica regularmente, varias veces al día y por varios meses, llevaría a conclusiones como las de la figura 18.



FIGURA 18: Ejemplo de resultados de herramientas de Minería de Datos sobre presencia de Mauricio Macri en redes sociales. (Illuminati Lab, 2016).

Podemos visualizar en la exposición de resultados, un *wordcloud* con los insultos más utilizados en las redes sociales, sobre la figura política. Inclusive vemos, distinguidos por porcentaje de apariciones, los términos positivos y negativos más frecuentes en los medios sociales.

Otro caso que demuestra la importancia de analizar las opiniones expresadas por redes sociales sobre distintos tópicos de interés, es el siguiente respecto al tópico Panamá Papers. Los tweets con mayor impacto muestran el liderazgo de dichos usuarios, ya que poseen muchos retweets y por ende a más lectores han llegado. Los hashtags con mayor presencia muestran tópicos o figuras políticas que son tendencia en el período analizado. Por último, podemos observar los términos relacionados con las dos figuras políticas más populares en nuestro país.

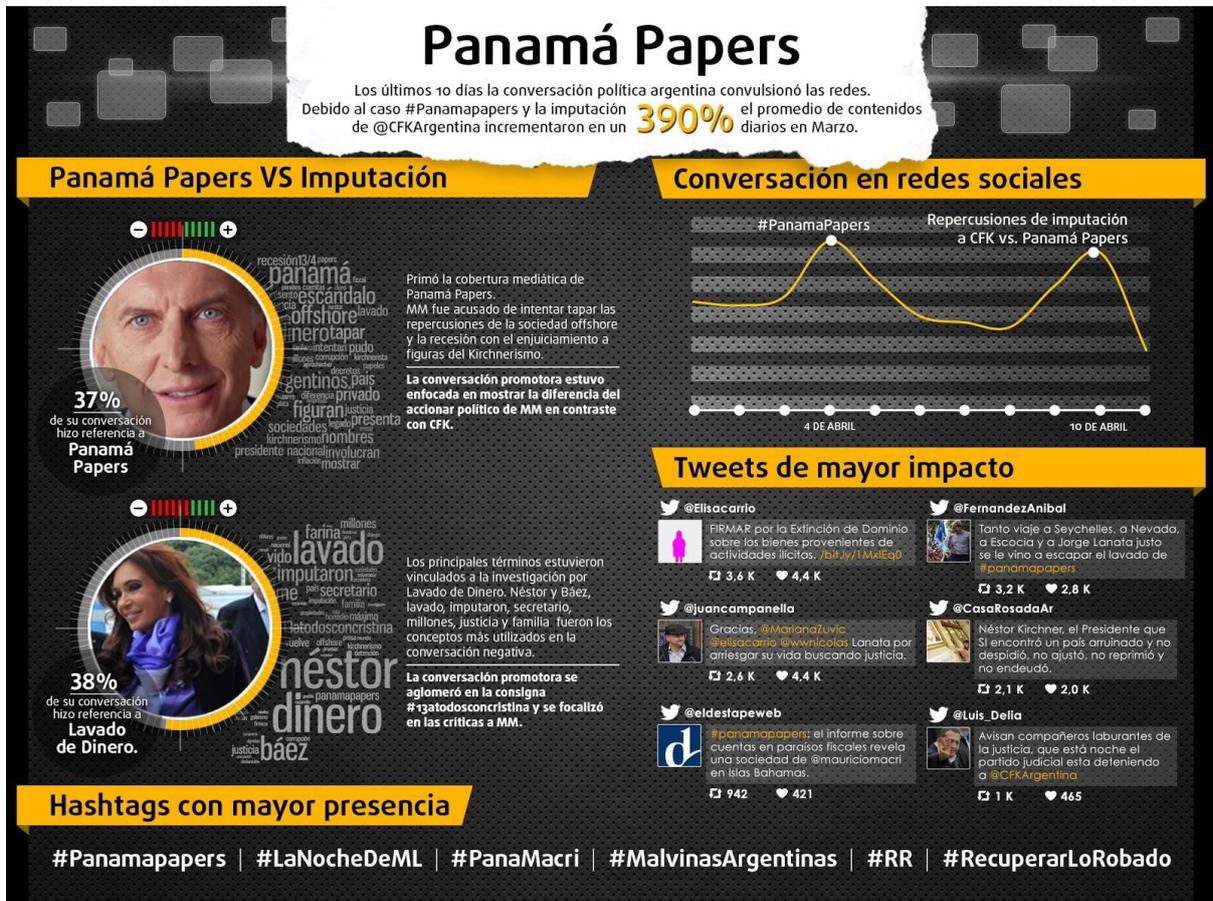


FIGURA 18: Ejemplo de resultados de herramientas de Minería de Datos sobre presencia de Mauricio Macri y Cristina Kirchner en función de tópicos en redes sociales. (Illuminati Lab, 2016).

Creemos importante alentar el estudio de la exposición de resultados para futuras investigaciones, ya que por más que es un tópico menos novedoso, es esencial para la observación de los mismos, y para que profesionales que no se dedican a sistemas puedan interpretarlos y aconsejar a los políticos a partir de ellos.

Nos referimos durante el desarrollo de este escrito, a “la toma de decisiones” en el ámbito político, como un objetivo en relación al uso de las herramientas de SOCMINT. Creemos que el análisis exhaustivo del público permite, en primera instancia, conocer los perfiles individuales de las personas. En segunda instancia, permite segmentar la población por características que difieren lo meramente geográfico, por edad, grupo social, profesión a la que se dedican, género, entre otras. Por último, permiten entender su reacción ante decisiones, eventos, tópicos.

Creemos que conocer al público permite en algunas ocasiones tomar mejores decisiones y, en otras, comunicar de mejor forma las decisiones tomadas. Saber qué es lo que espera la mayoría, no necesariamente influye en la forma de gestionar un país, porque es claro que cada individuo suele velar por su propio beneficio, por ende, excede a lo que es justo o mejor. Pero sí es importante para un político entender lo que el público espera de él, para así actuar de forma empática y generar simpatizantes en todos los segmentos sociales.

Retomando el caso analizado en la Metodología, en la comparación de las herramientas de SOCMINT, el análisis de los datos que se encuentran públicos sirve para predecir resultados en elecciones, para orientar al público e inducirlo a modificar su elección, como también, para utilizar lo que espera en beneficio de la organización y dirigir sus discursos con “el diario del lunes”.

Es importante aplicar la gestión de las redes sociales con tiempo, ya que generar un cambio o entender lo que el público espera no es algo que se puede hacer de un día para el otro. De hecho, para generar un compromiso real, es necesario trabajar con tiempo en el discurso, en los puntos más importantes en los que el partido se apoya. Dejaremos, de todos modos, la parte estratégica para el estudio de otros campos de interés en el ámbito.

Sin embargo, resulta interesante el análisis que realizamos en la Metodología vinculado a las herramientas, como comprobación de resultados obtenidos a partir de las mismas, que en la práctica afectaron el discurso dado en los medios por parte de nuestro presidente. No podemos comprobar de forma explícita que los utiliza, pero sí que se condice con lo que el público está esperando de él. Si esto lo aplicamos en una escala aún más grande, sin limitarnos a un proceso de tweets obtenidos en 40 segundos, sobre un tópico muy específico como lo es “#ImpunidadParaNinguno”, podemos observar que en su campaña actual, “el cambio” fue en primera instancia una forma de acercarse a la gente, por haber estudiado al público y entender, que más allá del grupo social al que pertenecieran, muchos sectores no estaban conforme con la política del país. Entonces, a partir de este argumento, buscó homogeneidad. Con el tiempo pasó a ser un slogan, es ahí donde el candidato necesita fortalecer su posición como identidad, es decir, lograr que su identidad y su imagen digital sean una sola: la articulación entre lo que se expresa y lo que se interpreta. En todas estas etapas del proceso, utilizar la información resultante del análisis de los datos de las redes sociales, sirve de utilidad para dirigir el accionar del político.

Recientemente muchos medios señalaron que el triunfo de Mauricio Macri en las elecciones presidenciales se debió en gran medida a las redes sociales, en especial a Facebook. Tanto así que muchos medios periodísticos califican a Macri como el primer presidente “Facebook” del mundo porque superó las interacciones en redes sociales de Obama. (Santa Maria, 2016). Habiendo analizado el punto de inflexión que generó Obama a nivel mundial, y la aplicación emergente en nuestro país de las técnicas mencionadas, resulta interesante la comparación entre ambos candidatos y su influencia en las redes sociales como herramienta para sus campañas.

Coincidimos con la expresión sobre un experto en el tópico, “ninguna elección política se gana con las redes sociales, pero sin redes sociales no se gana ninguna elección política”. (Santa Maria, 2016).

11. Conclusiones

En el transcurso del presente proyecto nos encontramos con diversas complicaciones, frenos y oportunidades. Para concluir, expondremos nuestra experiencia a partir del desarrollo del presente trabajo de investigación.

Vinculado al tema que seleccionamos como foco de estudio, consideramos interesante el tópico, ya que existía poca información en Internet al respecto, sobre todo en relación al uso de herramientas de Social Media Intelligence en el campo político. Desarrollamos el trabajo durante las elecciones presidenciales en nuestro país, lo que potenció nuestro interés en investigar cómo se conformaban las campañas políticas, cómo seleccionaban su contenido, y cómo hacían uso de las redes sociales como fuentes de información acerca de los individuos, y como medios para comunicarse de forma horizontal con su público.

La falta de información académica que existe respecto a los temas tratados nos significó tanto un desafío como un aliento. En muchas ocasiones tuvimos que limitar nuestro estudio a lo disponible y en otras tuvimos que sacar nuestras propias conjeturas respecto a elaboraciones vagas ya existentes. Resultamos entonces en la consecución de uno de nuestros objetivos con el trabajo, el proveer al ámbito académico y comercial un elaborado preciso y claro sobre una temática poco explotada.

Respecto al alcance de nuestra investigación, nos mantuvimos exentos de estudiar la porción del proceso que compete a áreas relativas al marketing, como lo es la exposición de los resultados obtenidos de las herramientas, o relativo a la consultoría en relación a la toma de decisiones políticas. Consideramos oportuno alentar a la investigación sobre estos ámbitos a estudiantes dedicados a áreas afines, como continuación del proceso de investigación que iniciamos, y dando soporte a profesionales y jefes de campaña.

En relación a los medios de extracción y procesado de datos, nos encontramos con dificultades a la hora de elegir cuáles estudiaríamos, ya que son muchas las opciones y las formas de armar flujos de trabajo para obtener resultados sumariados de grandes volúmenes de datos. En un principio, intentamos hacer el análisis que luego hicimos con las herramientas, de forma manual, para corroborar que implicaba demasiado tiempo, y que un partido en pleno proceso de campaña política podía optimizar los recursos y disminuir el esfuerzo empleado. Para esto, extrajimos tweets a partir de herramientas online, y analizamos

el sentimiento que percibimos en cada uno, trasladando los resultados a una planilla en Excel. Y en esta instancia observamos que la opinión de uno difería en muchos casos de la del otro, y que era difícil tener una opinión parcial y objetiva. Es por esto que consideramos apropiado avocarnos al estudio de herramientas que hicieran este trabajo por nosotros, dado por comprobada la dificultad de hacerlo de forma manual, y la subjetividad de opinar post por post el sentimiento reflejado en los mismos.

Una limitación con la que nos encontramos a la hora de utilizar las APIs de las redes sociales para extracción de datos es la restricción, más particularmente en Twitter, respecto al factor temporal de los datos. Twitter solo permite barrer un historial de tweets de un máximo de dos semanas mediante su API, por ello a la hora de seleccionar que datos analizar tuvimos que considerar datos presentes o conservar datos pasados para no perder el hilo del estudio.

En relación al análisis de las herramientas, nos implicó previamente una selección de las mismas. En este punto nos encontramos con dudas acerca de qué tipo de herramientas seleccionar. Hicimos un estudio bastante intensivo de las mismas, y en ocasiones no pudimos probar algunas, por ser pagas, lo cual implicaba decidir seleccionarla o descartarla simplemente leyendo información sobre la misma, haciendo menos verificable su eficiencia y conveniencia. Le pedimos consejos a nuestro tutor, luego de haber hecho un filtro entre nosotros sobre dónde hacer foco, y aquí descartamos algunas, como por ejemplo QlikView, que estaba más orientada a reportaría, gráficos, dashboards, por sobre la extracción de la información, o el procesado de la misma. Y sobre este tema, también creemos que merece explicar por qué seleccionamos esta porción del proceso, y no la exposición de los resultados, o cómo utilizarlos para tomar decisiones.

Decidimos en un principio hacer énfasis en la extracción de los datos de las redes sociales, lo cual compete más a nuestra área de estudio de Business Intelligence, al igual que la parte del procesado de la información, que también puede hacerse con herramientas de SOCMINT. Es muy interesante cómo, a partir de información procesada, se toman decisiones, pero nos excede a la experiencia que ganamos en el transcurso de nuestra carrera como Ingenieros, y teníamos el fin de aplicar algo de lo estudiado en los últimos años. El procesado de los datos lo seleccionamos con el fin de aprovechar la potencia de las herramientas y demostrar que sirve en política, cuando también podría haber sido en otras áreas, como la comercial, o para terrorismo, medicina, entre otras. Y como previamente mencionamos, estábamos entrando en un año de elecciones, y eso influyó en nuestra selección del área de

estudio. Lo consideramos como una oportunidad de comprobar si en nuestro país se estaban utilizando este tipo de herramientas y técnicas, y de no ser así, alentar su uso y su conocimiento.

Tenemos la suerte de haber conocido expertos en el tema, inclusive un contacto trabaja en una compañía de generación de contenido la cual, en combinación con una consultora que se dedica a campañas políticas digitales, manejaron la campaña de Larreta el año pasado. Con lo cual, pudimos involucrarnos con lo que existe hoy y hacer pública la necesidad de su implementación.

Creemos firmemente que si algo es poderoso y aporta valor, no requiere ser un “secreto”, y aplicarse solamente en partidos específicos, sino que es útil que todos conozcan las técnicas y ofrezcan competencia a nivel país. La oposición es buena y necesaria, y si candidatos no compiten por no saber cómo utilizar las redes sociales en su beneficio, es importante que no corran con desventaja y puedan adaptarse a la nueva Política 3.0.

12. Bibliografía

ALEGSA [en línea]. [consulta 13 ago. 2016] <<http://www.alegsa.com.ar/Dic/query.php>>

ALEXA – ACTIONABLE ANALYTICS FOR THE WEB [en línea]. [consulta 5 ago. 2016] <<http://www.alexa.com/topsites>>

ARQUITECTURA DE SISTEMAS [en línea]. [consulta 14 jul. 2016] <http://www.it.uc3m.es/abel/as/DSP/L2/CCErrors_es.html>

AYUDA EN LA WEB [en línea]. [consulta 1 ago. 2016] <<http://www.ayudaenlweb.com/microblogging/twitter/que-es-el-timeline-de-twitter>>

BECOM1 [en línea]. 2016. [Consultado 12 ago. 2016] <<http://becom1.com>>

BLANCHETTE, Jasmin. 2008. The Little Manual of API Design [en línea]. [consulta 16 jun. 2016]. <<http://people.mpi-inf.mpg.de/~jblanche/api-design.pdf>>

BRASSARD, Gilles y BRATLEY, Paul. 1997. [consulta 4 jul. 2016]. <https://es.wikipedia.org/wiki/Algoritmo#cite_note-Brassard-1>

BOTELLA ALONSO, Juan José. 2013. [consulta 4 jul. 2016] <<http://es.ccm.net/forum/affich-354867-saber-que-es-nombre-de-usuario>>

CASTELLS, Manuel. 1996. La era de la información. Economía, sociedad y cultura [en línea]. [consulta 6 jul. 2016] <http://www.felsemiotica.org/site/wp-content/uploads/2014/10/LA_SOCIEDAD_RED-Castells-copia.pdf>

CCM - COMUNIDAD INFORMÁTICA [en línea]. [consulta 16 jul. 2016] <<http://es.ccm.net/contents/198-extraccion-de-datos-y-olap>>

CHEN, Hsinchun, CHIANG, Roger H. L., STOREY, Veda C. 2012. Business Intelligence and Analytics: From Big Data to Big Impact [en línea]. [consulta 10 jun. 2016] <<https://ai.arizona.edu/sites/ai/files/MIS611D/chen-bi-december-2012.pdf>>

DEFINICIÓN DE [en línea]. [consulta 3 jul. 2016] <<http://definicion.de/>>

DEFINICIÓN.MX [en línea]. [consulta 2 jul. 2016] <<http://definición.mx/>>

ECLIPSE.ORG [en línea]. [consulta 13 ago. 2016] <<https://eclipse.org/>>

EÍTO BRUN, Ricardo, SENSO, José A. Minería Textual [en línea]. [consulta 2 jun. 2016] <<http://eprints.rclis.org/11491/1/Artmineriapdf.pdf>>

EPSILON TECHNOLOGIES [en línea]. [consulta 4 jul. 2016] <<http://www.epsilontec.com/que-es-el-social-media-intelligence-definicion/>>

FELDMAN, Ronen, SANGER, James. The Text Mining Handbook [en línea]. [consulta 6 ago. 2016] <<http://www.roelsbeestenboel.nl/text.pdf>>

FORTUNATO, S. *Community detection in graphs*. 2010. p. 3–5, p. 75–174. ISBN 978-3-642-01205-1

FUNDEU [en línea]. Fundación del Español Urgente y del Diccionario Panhispánico de Dudas. [consulta 13 ago. 2016] <<http://www.fundeu.es/recomendacion/trol-no-troll/>>

GAMMA, E., HELM, R. Design Patterns. Elements of Reusable Object-Oriented Software. [en línea] [consulta 8 jul. 2016] <<http://www.uml.org.cn/c++/pdf/DesignPatterns.pdf>>

GEARY, Joanna. DoubleClick (Google): What is it and what does it do? [en línea]. 2012. [consulta 6 ago. 2016] <<https://www.theguardian.com/technology/2012/apr/23/doubleclick-tracking-trackers-cookies-web-monitoring>>

GLOSARIO.NET [en línea] [Consulta 13 ago. 2016] <[http://tecnologia.glosario.net/terminos-tecnicos-internet/arrastrar-y-colocar-\(drag-and-drop\)-127.html](http://tecnologia.glosario.net/terminos-tecnicos-internet/arrastrar-y-colocar-(drag-and-drop)-127.html)>.

GONZÁLEZ BLANCO, Larry. 2012. La Política 3.0. [consulta 1 may. 2016]. <<http://es.slideshare.net/aserasociados/politica-30>>

GOOGLE SUPPORT [en línea]. 2016. [consulta 20 jul. 2016] <<https://support.google.com>>

HAN, J. KAMBER M. *Data Mining: Concepts and Techniques*. 2006. ISBN 9380931913.

HASSAN-MONTERO, HERRERO-SOLANA, V. 2016. [consulta 16 jul. 2016] <http://www.yusef.es/improving_tagclouds.pdf>

HIDALGO, Carlos. ¿Están ganando los troles la batalla de Internet? [en línea]. 2015. [consulta 20 jul. 2016] <<http://www.bez.es/426696332/Estan-ganando-los-troles-la-batalla-de-Internet.html>>

HIDALGO, Carlos. Los ejércitos de troles y bots crecen en número, pero menguan en importancia política [en línea]. 2016. [consulta 20 jul. 2016]

<<http://www.bez.es/115832542/Los-ejercitos-de-troles-y-bots-crecen-en-numero-pero-menguan-en-importancia-politica.html>>

ILUMINATI LAB [en línea]. 2016. [Consultado 12 ago. 2016] <<http://illuminatilab.com>>

INMON W. H. *Building the Data Warehouse*. 3ª. ed. Nueva York: John Wiley & Sons, Inc. 2002. ISBN: 0-471-08130-2

INTEF [en línea]. 2016. [consulta 2 jul. 2016] <http://www.ite.educacion.es/formacion/materiales/155/cd/modulo_5_2_Face_Twitt/siguiendo_seguidores_y_tweets.html>

ISMS Forum Spain [en línea]. [consulta 4 jul. 2016] <http://www.protegetuinformacion.com/perfil_tema.php?id_perfil=12&id_tema=164>

ISSENBERG, Sasha. 2012. How Obama's Team Used Big Data to Rally Voters [en línea]. [consulta 10 ago. 2016] <<https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters>>

KOH, Yoree. What Twitter Gains by Expanding the 140-Character Tweet. *The Wall Street Journal* [en línea]. 2016. [consulta 10 ago. 2016]. <<http://blogs.wsj.com/digits/2016/01/06/what-twitter-gains-by-expanding-the-140-character-tweet>>

LAMARCA LAPUENTE, María Jesús. 2013. [consulta 13 ago. 2016]. <<http://www.hipertexto.info/documentos/metadatos.htm>>

LEY BORRÁS, Roberto. 2013. [consulta 4 jul. 2016] <<http://decidir.org/serv01.htm>>

LIDERAZGO Y MERCADEO [en línea]. 2006. [consulta 14 jul 2016] <[http://www.hacienda.go.cr/centro/datos/Libro/Que%20es%20un%20Lider\[1\].pdf](http://www.hacienda.go.cr/centro/datos/Libro/Que%20es%20un%20Lider[1].pdf)>

LIJPHART, A. 1995. [consulta 4 jul. 2016]. <https://es.wikipedia.org/wiki/Candidato_pol%C3%ADtico#cite_ref-1>

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DEL CONOCIMIENTO [en línea]. [consulta 12 feb. 2016]. <<http://datamining.dc.uba.ar/datamining/>>

MARAZZI, Axel. 2016. [consulta 2 jul. 2016] <<http://www.conexionbrando.com/1389864-que-es-la-nube-para-que-sirve-y-cuales-son-los-servicios-que-tenes-que-conocer>>

MARKETING DIRECTO [en línea]. [consulta 11 may. 2016] <<http://www.marketingdirecto.com/digital-general/social-media-marketing/social-media-intelligence-con-las-redes-sociales-ya-no-basta-con-oir-las-empresas-deben-aprender-a-extraer-inteligencia-de-las-redes/>>

MAYSA CONSULTORES [en línea]. [consulta 1 ago. 2016] <http://www.maysaconsultores.com.ar/recursos/Maysa_Text_Mining.pdf>

MERINO, Marcos. 2014. [consulta 2 jul. 2016]. <<http://www.ticbeat.com/tecnologias/que-es-una-api-para-que-sirve/>>

MÉTODO MARKETING [en línea]. [consulta 13 ago. 2016]. <<http://www.metodomarketing.com/que-es-target/>>

NATURE PUBLISHING GROUP [en línea]. [consulta 14 jul. 2016] <<http://www.nature.com/webfeeds/index.html>>

NEWMAN, Mark, BARABÁSI, Albert-Láslo y WATTS, Duncan J. *The Structure and Dynamics of Networks*, 2006. ISBN 9781400841356

NYCE, Charles. 2007. [Consulta 16 jul. 2016] <<https://www.cgi.com/sites/default/files/white-papers/Predictive-analytics-white-paper.pdf>>

ORALLO HERNÁNDEZ, José, RAMÍREZ QUINTANA, José M. y FERRI RAMÍREZ, César. *Introducción a la Minería de Datos*. Editorial Pearson, 2004. ISBN 8420540919

OXFORD DICTIONARIES [en línea]. [consulta 14 jul. 2016] <<http://www.oxforddictionaries.com/es/>>

PARIKH, Ravi. Distribution of Word Lengths in Various Languages [en línea]. 2014. [consulta 10 ago. 2016] <<http://www.ravi.io/language-word-lengths>>

PÉREZ LÓPEZ, César. *Minería de Datos: Técnicas y herramientas*. Editorial Paraninfo, 2007. ISBN 8497324927

- PÉREZ-ROSAS, Verónica, BANEÁ, Carmen y MIHALCEA, Rada. Learning Sentiment Lexicons in Spanish [en línea]. 2012. [consulta 15 jun. 2016]. <<http://web.eecs.umich.edu/~mihalcea/papers/perez.lrec12.pdf>>
- PÉREZ VALDÉS, Damián. 2007. [consulta 4 jul. 2016]. <<http://www.maestrosdelweb.com/que-son-las-bases-de-datos/>>
- PETERSON, Robert A., FERREL, O. C. La evaluación del concepto de stakeholders según Freeman [en línea]. 2004. [consulta 14 jul. 2016] <http://www.iese.edu/es/files/La%20evaluaci%C3%B3n%20del%20concepto%20de%20stakeholders%20seg%C3%BAn%20Freeman_tcm5-39688.pdf>
- PHILLIPS, Whitney. [en línea]. 2014. [consulta 20 jul. 2016] <<https://billions-and-billions.com/2014/02/13/my-book-officially-forthcoming-with-mit-press/>>
- POLLO-CATTANEO, M., GARCÍA MARTÍNEZ, R., BRITOS, P. Elementos para una ingeniería de explotación de información. [en línea]. 2012. [consulta 14 jul. 2016] <<http://www.unla.edu.ar/sistemas/gisi/papers/UTN-FRBA-Proyecciones-10-1-67-84-2012.pdf>>
- RANJAN, Jayanthi. 2009. [consulta 21 feb. 2016]. <<http://www.imt.edu/FacultyandResearch/JayanthiRanjan.aspx>>
- REAL ACADEMIA ESPAÑOLA [en línea]. 2014. [consulta 2 jul. 2016] <<http://www.rae.es>>
- ROMERO, Sergio. Mauricio Macri presentó a su gabinete. *El Tribuno* [en línea]. 2016. [consulta 10 ago. 2016] <<http://www.eltribuno.info/mauricio-macri-presento-su-gabinete-n647090>>
- RUTLEDGE, Pamela. How Obama Won the Social Media Battle in the 2012 Presidential Campaign [en línea]. 2013. [consulta 10 ago. 2016] <<http://mprcenter.org/blog/2013/01/how-obama-won-the-social-media-battle-in-the-2012-presidential-campaign>>
- SANTA MARIA, Luigi. Los candidatos presidenciales y sus campañas en las redes sociales. *Staff Creativa* [en línea]. 2016. [consulta 8 ago. 2016] <<http://www.staffcreativa.pe/blog/candidatos-presidenciales-camapanas-redes-sociales>>

SANTIAGO, Hugo. Metodología HEFESTO, Fases del Proceso de BI [en línea]. 2014. [consulta 4 jul. 2016] <<http://pentahoparatodos.blogspot.com.ar/2014/03/metodologia-hefesto-fases-del-proceso.html>>

SANTIAGO, Raúl et al., 2015. [consulta 4 jul. 2016]. <https://es.wikipedia.org/wiki/Aplicaci%C3%B3n_m%C3%B3vil#cite_ref-1_1-0>

SANTOS, Elder. Data Mining for Predictive Social Network Analysis [en línea]. 2015. [consulta 20 jul. 2016]. <<https://www.toptal.com/data-science/social-network-data-mining-for-predictive-analysis>>

SIGNIFICADOS [en línea]. [consulta 4 jul.] <<http://www.significados.com>>

SMITH, Craig. By the Numbers: 125+ Amazing YouTube Statistics [en línea]. 2016. [consulta 4 ago. 2016] <<http://expandedramblings.com/index.php/youtube-statistics>>

STATISTA [en línea]. [consulta 12 jul. 2016] <<http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>>

TANENBAUM, Andrew S. *Sistemas Operativos Moderno*. 2009. ISBN 6074420467

TDWI. The Data Warehouse Institute [en línea]. [consulta 6 mar. 2016] <<https://tdwi.org/Home.aspx>>

TELAM [en línea]. [consulta 2 may. 2016]. <<http://www.telam.com.ar/notas/201305/17023-la-edad-de-los-que-deciden.html>>

THE FREE DICTIONARY [en línea]. [consulta 21 jul. 2016] <<http://es.thefreedictionary.com/v%C3%ADrica>>

THIEL, Killian. Creating Usable Customer Intelligence from Social Media Data [en línea]. 2012. [consulta 25 mar. 2016] <https://www.knime.org/files/knime_social_media_white_paper.pdf>

TRENDING TOPIC [en línea]. [consulta 4 jul. 2016] <<http://www.trending-topic.com/>>

TURNER, Dawn M., 2016. What is Venture Management?. [consulta 4 jul. 2016] <<http://www.ventureskies.com/>>

TURNEY, Peter. 2002. [consulta 4 jul. 2016] <https://en.wikipedia.org/wiki/Sentiment_analysis#cite_ref-PangA102_2-0>

TWITTER [en línea]. [Consulta 3 jul. 2016] <<https://twitter.com/>>

UNIVERSIDAD CARLOS III MADRID [en línea]. Introducción a Data Mining. [consulta 1 ago. 2016]. <<http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/introduccion-DM.pdf>>

WEST, Tara. How to Advertise on YouTube [en línea]. 2013. [consulta 5 ago. 2016] <<https://www.koozai.com/blog/social-media/video-marketing/how-to-advertise-on-youtube>>

WILEY, John & sons. 2011. Windows Server Administration Fundamentals. Microsoft Official Academic Course. [consulta 5 may. 2016] <<http://www.wiley.com/WileyCDA/WileyTitle/productCd-EHEP001832,miniSiteCd-MOAC.html>>

WORDREFERENCE [en línea]. [consulta 3 ago. 2016] <<http://www.wordreference.com/es/>>

WORKFLOW PATTERNS [en línea]. [consulta 16 jul. 2016] <<http://www.workflowpatterns.com/>>

WOODWARD, Martin. Internet Cookies [en línea]. 2015. [consulta 20 jul. 2016] <<http://surfthenetsafely.com/surfsafely5>>

YAHOO DEVELOPER NETWORK [en línea]. [consulta 20 de jul. 2016] <<https://developer.yahoo.com>>

ZHU, Feida, SUN, Huan y YAN, Xifeng. Network Mining and Analysis for Social Applications [en línea]. 2014. [consulta 15 jun. 2016] <www.kdd.org/kdd2014/tutorials/T9_Tutorial_KDD2014.pdf>