CIS1710CP07

SENTINEL: Analítica sobre percepción de corrupción en Facebook

Manuela Estefanía Forero Pedreros
Jeffrey Torres Arango
Sebastián Gracia Valderrama

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
CARRERA DE INGENIERÍA DE SISTEMAS
BOGOTÁ, D.C.
2017

CIS1710CP07

SENTINEL: Analítica sobre percepción de corrupción en Facebook

Autores:

Manuela Estefanía Forero Pedreros

Jeffrey Torres Arango

Sebastián Gracia Valderrama

MEMORIA DEL TRABAJO DE GRADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TÍTULO DE INGENIERO DE SISTEMAS

Director

Ing. Alexandra Pomares Quimbaya Ph. D.

Jurados del Trabajo de Grado

Efrain Ortiz Pabón

Carlos Andrés Barreneche Jurado

Página web del Trabajo de Grado

http://pegasus.javeriana.edu.co/~CIS1710CP07/

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA DE SISTEMAS

BOGOTÁ, D.C.

Noviembre,2017

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
CARRERA DE INGENIERÍA DE SISTEMAS

Rector Magnífico de la Pontificia Universidad Javeriana

Jorge Humberto Peláez Piedrahita, S.J.

Decano de la Facultad de Ingeniería

Ing. Jorge Luis Sánchez Téllez

Director de la Carrera de Ingeniería de Sistemas

Ing. Mariela J. Curiel H. Ph. D.

Director de Departamento de Ingeniería de Sistemas

Ing. Efraín Ortiz Pabón

Artículo 23 de la Resolución No. 1 de Junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

AGRADECIMIENTOS

Manuela Forero Pedreros

Gracias a mi familia por su apoyo, sin ellos no estaría donde estoy en este momento. A mi mamá que fue y será siempre el motor de mi vida y la que me inspira cada día.

Sebastián Gracia Valderrama

A Jairo, María Elena, Carolina y Nydia, todo lo bueno que hay en mí empezó por ustedes.

Jeffrey Torres Arango

A mi familia y a Dios. Los amo.

A nuestra directora, Alexandra Pomares,

por su apoyo y por ser nuestra guía en este trabajo.

RESUMEN

En este documento describimos cómo aplicando diferentes técnicas de análisis y minería de datos a publicaciones y comentarios extraídos de Facebook, se puede obtener información acerca del estado actual de la percepción de líderes de opinión, partidos políticos, instituciones y casos de corrupción en Colombia. Los resultados obtenidos muestran que los datos generados en redes sociales son un insumo importante para encontrar nuevos indicadores de comportamiento y percepción de los usuarios y, asimismo, permite apoyar iniciativas anticorrupción. El proyecto desarrollado demuestra un gran potencial y se espera que sirva como base para futuras investigaciones y mejoras en la lucha anticorrupción.

ABSTRACT

This document describes how we can obtain information about the current state of corruption, political parties, politics and institutions in Colombia by applying different data mining and other analysis techniques to Facebook posts and comments. The results showed that the data generated in Social Networks are an important resource to find new indicators of behavior and perception of users and, in addition, they enable new anticorruption initiatives. The project shows great potential and is expected to serve as a foundation for future research and improvements in the fight against corruption.

Contenido

Lista de tablas

Lista de Ilustraciones

CAPÍTULO 1: INTRODUCCIÓN

En la actualidad, un tema que se encuentra recurrente no solo en los gobiernos mundiales sino en la concepción de la ciudadanía es la corrupción. Diferentes casos a nivel mundial se han detonado, generando alta incertidumbre sobre la manera de gobernar, no solo cuestionando a los líderes políticos que llegan al poder sino la debilidad institucional para garantizar la justicia. Un informe realizado por Transparencia Internacional *(Transparency, 2017)* sobre la percepción de la corrupción en 2016 sitúa a Colombia como uno de los países con mayor índice de corrupción, teniendo un puntaje de 37 unidades, donde 0 indica nivel alto de corrupción y 100 ausencia de corrupción, postulándolo como uno de los países más corruptos tomando la posición 90 sobre 176 países analizados. Esto demuestra que existe una alta desconfianza en las instituciones públicas, posibles casos de soborno y extorsión. Es importante realizar una observación más cercana de la población para identificar la percepción de corrupción de los ciudadanos frente a los gobiernos y de esa forma aportar a la lucha contra este fenómeno.

En Colombia existe una institución que tiene como objetivo luchar contra este fenómeno llamado el Observatorio de Transparencia y Anticorrupción de Colombia. Esta institución se autodefine como una herramienta que mide y analiza los fenómenos de la corrupción, a partir de la interacción entre entidades, ciudadanos y organizaciones públicas y privadas del orden nacional y territorial, para contribuir a elevar el nivel de transparencia en la gestión pública. El trabajo del observatorio gira en torno a tres ejes importantes: observar, educar y dialogar (*Observatorio de Transparencia y Anticorrupción, 2017)*. Estos tres ejes tienen las siguientes características:

- ***Observar***: consiste en recopilar, analizar y visualizar indicadores sobre transparencia y anticorrupción
- ***Educar***: se encarga de brindar herramientas para la promoción de la transparencia y la lucha contra la corrupción
- ***Dialogar***: se encarga de facilitar espacios para el diálogo entre ciudadanía, academia y servidores públicos.

A partir de estos tres ejes y de la problemática que existe debido a la corrupción, se decidió construir un sistema capaz de monitorear los casos de corrupción en Colombia, por medio de la extracción y análisis de los datos de la red social Facebook abierta puesto que las redes sociales son un predictor significativo de los comportamientos de participación política de los usuarios, tanto en Internet como en el mundo real *(Gil*

de Zúñiga, Jung, & Valenzuela, 2012). A dicho sistema se le dio el nombre de Sentinel y busca fortalecer los dos primeros ejes definidos por el Observatorio: observar y educar.

Sentinel es un sistema de información que está compuesto de tres partes: extracción de datos de Facebook, un componente de analítica encargado de hallar información relevante sobre corrupción y un componente de visualización en el cual se pueden ver los resultados obtenidos con el proceso de analítica. Sentinel busca apoyar al Observatorio de Transparencia y Anticorrupción mediante la aplicación de análisis y minería de datos, utilizando técnicas como análisis de sentimientos, correlaciones de palabras y detección de datos atípicos.

Desde el inicio del trabajo de grado se contactó al Observatorio de Transparencia y Anticorrupción de Colombia con el propósito de desarrollar la propuesta planteada por el equipo de trabajo. Una vez identificados los objetivos de los ejes, se planteó desarrollar un sistema que enriqueciera la información disponible para el Observatorio, utilizando como fuente de datos la red social Facebook abierta y de esta forma poder crear un sistema capaz de monitorear, analizar y visualizar la información obtenida.

El propósito de este documento es mostrar el proceso de desarrollo y los resultados obtenidos a partir del sistema Sentinel. El documento está compuesto por los siguientes capítulos:

Descripción General: Describe la formulación del problema identificado, los objetivos que se deben cumplir para solucionar la problemática, entregables y finalmente la metodología implementada para la identificación y desarrollo de la solución.

Contexto del proyecto: Explica los conceptos teóricos necesarios para entender la problemática y desarrollar la solución.

SENTINEL: Un sistema para el monitoreo de casos de corrupción en Facebook: Describe todas las actividades realizadas dentro de la metodología CRISP-DM para el desarrollo del proyecto.

Construcción del software de generación de modelos de Sentinel: Describe todo el diseño de la solución propuesta, su especificación funcional, diseño y arquitectura del sistema, implementación y pruebas de software.

Resultados: Presenta los resultados obtenidos del proyecto de investigación y su nivel de aceptación respecto a los objetivos planteados en la propuesta de Trabajo de Grado.

Conclusiones: Presenta las conclusiones obtenidas como parte del desarrollo del Trabajo de Grado, el impacto esperado del mismo y las actividades que pueden llegar a ser parte del trabajo futuro a realizar.

CAPÍTULO 2: DESCRIPCIÓN GENERAL

1. Motivación del proyecto

La Alianza CAOBA surgió debido a una invitación directa del Ministerio de las Tecnologías de Información y las Comunicaciones y el Departamento Administrativo de Ciencia, Tecnología e Innovación –Colciencias-, de diferentes empresas del sector público y privado de Colombia. Este centro de excelencia apoya el uso de las tecnologías de Big Data y Data *Analytics* (BD&DA) a través de diferentes frentes que incluyen la formación del talento humano, la investigación aplicada y el desarrollo de productos cuya propuesta de valor está fundamentada en la generación de soluciones alrededor de las tecnologías del BD&DA *(Alianza Caoba, 2016).*

RART (Real Time Social Data Mining) es un *framework* desarrollado por Jaime Andrés Mendoza y Daniel Alejandro Calambás en 2016 como un trabajo de grado liderado por la Ing. Alexandra Pomares y la Alianza CAOBA. El principal objetivo del trabajo de grado fue desarrollar un software que recolectara, de forma continua, información relacionada a reacciones en las publicaciones de uno o varios *fans pages* dentro de Facebook. Por lo tanto, la finalidad del proyecto fue generar un conjunto organizado de datos para apoyar el proceso de *Data Analytics (RART, 2016)*.

Al conocer sobre las capacidades de RART, el grupo de trabajo buscó la manera de utilizar este *framework* para crear un sistema que pudiera aportar al desarrollo del país a partir del análisis de la información recolectada. Entre las diferentes temáticas y aplicaciones identificadas por el grupo de trabajo con RART, resultó de gran interés el tema de noticias sobre corrupción. Para esto, se tuvo en cuenta que, hoy en día, los medios de comunicación buscan publicar en las redes sociales todo el material informativo y de tendencia con el propósito de propagar las noticias tan amplio como sea posible *(Bandari, Asur, & Huberman, 2012)*. A partir de esto, nace el proyecto Sentinel.

El proyecto Sentinel aborda el análisis de las temáticas de corrupción usando como fuente de información las publicaciones realizadas en la red social *Facebook*, con el fin de desarrollar un sistema capaz de monitorear, analizar y visualizar el fenómeno de la corrupción en esta red social.

Habiendo seleccionado el tema, se procedió a investigar sobre posibles instituciones y expertos del negocio que pudieran estar interesados en contar con un sistema capaz de monitorear la actividad de medios

y usuarios en Facebook, enfocándose en el tema de corrupción. En esta búsqueda, el grupo de trabajo contactó al Observatorio de Transparencia y Anticorrupción. El Observatorio es "una herramienta para la medición y análisis del fenómeno de la corrupción, a partir de la interacción entre entidades, ciudadanos, y organizaciones públicas y privadas del orden nacional y territorial, para contribuir a elevar el nivel de transparencia en la gestión pública" *(Observatorio de Transparencia y Anticorrupción, 2017)*. El Observatorio está compuesto por diferentes entidades, entre estas la Secretaría de Transparencia, principal punto de contacto y apoyo para el desarrollo de este proyecto.

Sentinel nació, inicialmente, con la idea de tener un sistema de información capaz de monitorear la actividad de los medios y de los usuarios en Facebook cuando publican o comentan sobre corrupción. Pero, gracias al apoyo de la Secretaría de Transparencia, Sentinel evolucionó para convertirse en un sistema que no solo monitorea casos de corrupción, sino la percepción de los usuarios sobre diferentes líderes de opinión, partidos políticos e instituciones, así como ser capaz de medir el sesgo de los medios a la hora de publicar sobre diferentes entidades y casos de corrupción.

2. Oportunidad

Basándose en la motivación del proyecto se encontró una gran oportunidad de fortalecer las fuentes de información con las que cuenta el Observatorio, debido a que no tienen en cuenta los datos y opiniones generadas en redes sociales, las cuales son una gran fuente de información. Según la Secretaría de Transparencia, el fenómeno de corrupción es complicado de medir debido a que es algo naturalmente oculto. Por esta razón, Sentinel es un proyecto que promete gran utilidad para el Observatorio, debido a que al tener la información hallada con Sentinel, se puede tener un historial del comportamiento de los medios y usuarios con respecto al tema de corrupción, facilitando y apoyando el trabajo del Observatorio y la Secretaría de Transparencia *(Castañeda, Ortiz, & Pérez, 2017)*

La Secretaría de Transparencia asegura que este proyecto puede mejorar la forma en que hacen monitoreo de medios. Además, permitirá contrastar los resultados hallados con Sentinel contra las encuestas de percepción de corrupción realizadas por LAPOP (también conocidas como Barómetro de las Américas). Estas encuestas se realizan todos los años desde 2004 y buscan tener una aproximación más detallada de la cultura política de grupos específicos de la población colombiana, y poder contrastar sus opiniones y actitudes políticas con aquellas del resto de la población *(Garcia, Montalvo, & Seligson, 2015)*. Comparando la información obtenida en las encuestas con la información obtenida con Sentinel, se pueden sacar conclusiones más informadas y completas sobre el fenómeno de la corrupción *(Castañeda, Ortiz, & Pérez, 2017)*.

3. Objetivo General

Desarrollar un modelo de analítica que permita enriquecer la información disponible para el Observatorio de Transparencia y Anticorrupción de Colombia utilizando como fuente de datos la red social Facebook abierta.

4. Objetivos Específicos

1. Definir el modelo de analítica de corrupción a partir de publicaciones en *fan pages* de medios de comunicación, empresas, personajes públicos y partidos políticos.
2. Diseñar e implementar un componente web que presente los resultados mediante gráficas de manera clara y concisa.
3. Validar la utilidad y exactitud de los modelos generados en la fase de modelado junto con el experto del negocio.

5. Entregables y estándares

Se desarrolló un componente de analítica que permitió obtener nueva información a partir de las diferentes publicaciones y comentarios extraídos de Facebook. Además, se desarrolló un aplicativo web que permite visualizar los resultados obtenidos en el proceso de minería y análisis de datos. Para poder llevar a cabo la realización de la solución propuesta, fue necesario seguir las fases de desarrollo las cuales requieren de una correcta documentación.

En la tabla 1 se incluyen las entregas del proyecto:

Entregable	**Estándares asociados**	**Justificación**
Software Project Managemet Plan	*16326-2009 - ISO/IEC/IEEE International Standard Systems and Software Engineering--Life Cycle Processes--Project Management*	El propósito del plan de proyecto es identificar el alcance del proyecto, estimar la cantidad de trabajo involucrado y crear una correcta calendarización para la ejecución del mismo. Entonces se busca describir las tareas que llevarán a culminar el proyecto.
Software Design Description	*1016-2009 - IEEE Standard for Information Technology--Systems Design--Software Design Descriptions*	Busca describir el producto de software a desarrollar, para que el equipo de desarrollo tenga unos lineamientos claros sobre qué hacer y cómo desarrollar la arquitectura del proyecto.

Software Requirement Specification	*830-1998 - IEEE Recommended Practice for Software Requirements Specifications*	Describe los requerimientos funcionales y no funcionales, el diagrama de casos de uso y diagramas de comportamiento tales como de secuencia
Componente de analítica de datos	-	Hace parte de la solución propuesta de analítica de datos en donde se aplicarán técnicas de minera y análisis de datos
Componente web para visualizar	-	Hace parte de la solución propuesta para mostrar los resultados obtenidos del componente de analítica.
Manual de Usuario	-	Hace parte de la solución propuesta. Esta será una guía para que el usuario conozca las funcionalidades que puede realizar el aplicativo.

Tabla 1 Entregas del proyecto

6. Metodología.

Considerando la naturaleza del proyecto, fue necesario utilizar dos metodologías. La primera fue CRISP–DM (*Cross-Industry Standard Process for Data Mining*): CRISP-DM es una metodología bien fundamentada para el desarrollo de proyectos de minería de datos; la segunda SCRUM, una metodología de desarrollo ágil de software. Las fases de CRISP-DM se pueden observar en la ilustración 1.

CRISP-DM guio la identificación de los problemas de negocio y la definición de los tipos de modelos requeridos para satisfacer las necesidades de información del Observatorio, así como también su evaluación final. De manera complementaria, y considerando que CRISP-DM incluye una fase en la que se generan los modelos de análisis, y las herramientas de minería existentes no eran suficientes para producir los modelos esperados por parte de los usuarios de negocio, el proyecto requirió el desarrollo de un componente de software (también referido como componente de visualización) que permitiera generar dichos modelos requeridos con su respectiva visualización.

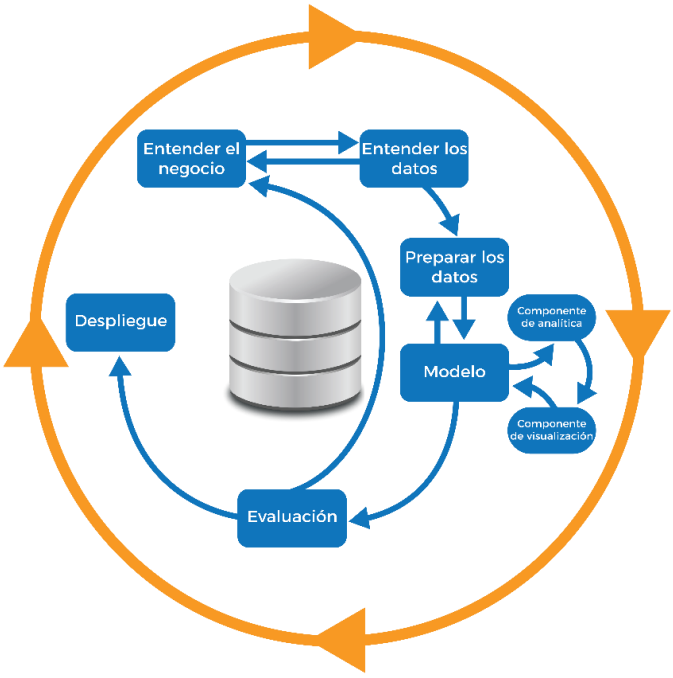

Ilustración 1 Metodología CRISP del proyecto Sentinel (Chapman & Clinton, 2000)

Basándose en las necesidades del negocio, se establecieron *sprints* por semana cuyo *backlog* son los requerimientos priorizados. El uso de la metodología SCRUM se lleva a cabo en la fase de Modelo (fase 4) de la metodología CRISP-DM, debido a que fue necesario realizar el desarrollo del componente de visualización en paralelo con el desarrollo de los algoritmos de análisis y minería, para asegurar el adecuado funcionamiento y visualización de los resultados. A continuación, en la ilustración 2, se identifican las diferentes fases que componen a la metodología SCRUM:

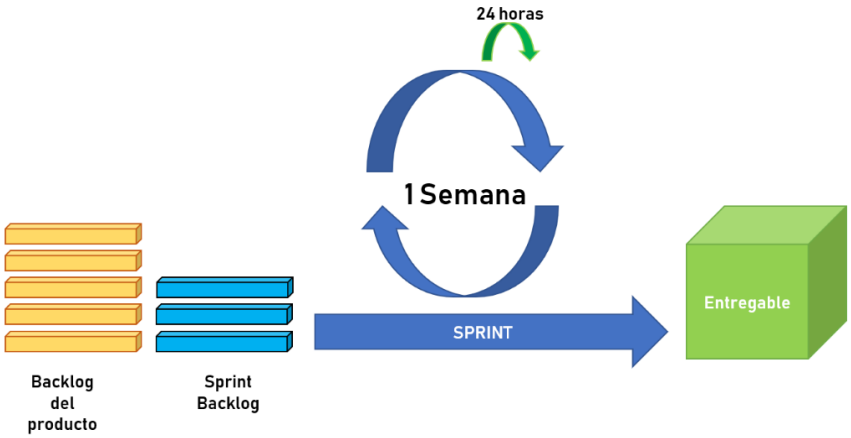

Ilustración 2 Metodología SCRUM ("What is Scrum?", 2017)

CAPÍTULO 3: CONTEXTO DEL PROYECTO

1. Contextualización

El trabajo de grado propuesto tiene como finalidad aportar a la misión del Observatorio de Transparencia y Anticorrupción de Colombia. El objetivo del Observatorio es medir y analizar el fenómeno de la corrupción, a partir de las interacciones entre entidades, ciudadanos, y organizaciones públicas y privadas para elevar el nivel de transparencia en la gestión pública *(Observatorio de Transparencia y Anticorrupción, 2017)*.

Durante los últimos años, la cantidad de datos generada ha aumentado a gran escala en diferentes campos. Según un informe de la *International Data Corportation* (IDC), en 2011, la cantidad de datos generada y copiada fue de 1.8 ZB y según predicciones se espera que esta cantidad se duplique cada dos años. A raíz de este incremento masivo de datos generados, nace el término ***Big Data***, que se utiliza para describir grandes sets de datos *(Chen, Mao, & Liu, 2014)*.

Debido al gran crecimiento de la web 2.0 *(Alexander & Levine, 2006)* y la demanda/influencia de las redes sociales sobre las personas, se están creando grandes cantidades de contenido que influencian la opinión de la gente en aspectos políticos *(Pinquart & Sörensen, 2000) (Domínguez, 2009)*. Con estas opiniones se pueden realizar diferentes tipos de análisis del comportamiento de estos individuos sobre un tema en particular y las reacciones sobre las noticas que tienen que ver con este tema.

Sentinel aprovecha el hecho de que las redes sociales son la principal razón por la que se utiliza la web en Colombia *(Casa Editorial El Tiempo, 2016)*, siendo Facebook la red social con la que más usuarios colombianos cuenta, con 25.000.000 de usuarios en el país *(MinTic, 2017)*. Para el día 17 de marzo de 2017, solamente las páginas de Facebook de El Espectador, El Tiempo y Semana sumaban 7.457.466 seguidores. Adicionalmente, la infraestructura de las redes sociales puede soportar una variedad de aplicaciones de analítica de datos, tales como: búsqueda, análisis de texto, análisis de imágenes, análisis de sentimientos y correlaciones *(Aggarwal, 2011)*.

El trabajo de grado busca aprovechar la cantidad de usuarios activos en Colombia que utilizan la red social Facebook y aplicar técnicas de análisis y minería a los datos extraídos de la misma, haciendo uso del *framework* RART. Esto con el fin de poder proveer una herramienta que permita hacer un monitoreo de la corrupción, líderes de opinión, partidos políticos e instituciones colombianas, y de esta manera desarrollar un componente de analítica para finalmente entregar resultados de manera visual en el componente de visualización.

2. Análisis de contexto

A raíz del surgimiento del gobierno electrónico (*e-government*)[1], los ciudadanos han tenido mayores oportunidades de tomar acción y verse involucrados directamente con el gobierno. En un estudio realizado en 2009, denominado *Online anti-corruption tools in Latin America (Matheus & Ribeiro, 2009)*, por Ricardo Matheus y Manuella Maia Ribeiro de la Universidad de Sao Paulo, se mencionan y se describen varias herramientas de anticorrupción en línea, específicamente en América latina. Estas herramientas son:

- *Denuncia la Corrupción (Xalapa, México):* *Denuncia la Corrupción* es una aplicación móvil construida en Xalapa, México, con el objetivo de combatir la corrupción mediante el envío de información por parte de los ciudadanos. La aplicación puede recibir casos de corrupción, permitiendo a los usuarios ser voces activas en la denuncia de casos, pero este aplicativo no permite hacer seguimiento de los reportes una vez se ha ingresado en la aplicación.
- *Denuncia Corrupción (Guadalajara, México):* *Denuncia Corrupción* es una herramienta muy parecida a *Denuncia la Corrupción*, que promueve la participación de los usuarios para luchar contra la corrupción mediante la denuncia de casos. Una de las debilidades de esta herramienta es que no garantiza el anonimato del usuario que está denunciando. Además, esta herramienta tampoco realiza un análisis sobre la evolución de los casos de corrupción.
- *Denuncia en línea (Guayaquil, Ecuador):* *Al igual que l*as herramientas anteriores, la aplicación registra casos de corrupción que son reportados por medio de la actividad. Sin embargo, no garantiza el anonimato de los usuarios que denuncian y al igual que las demás herramientas tampoco ofrece actualizaciones y análisis sobre el estado de corrupción ni de las denuncias.
- *Investigación y Análisis (Cartago, Costa Rica):* *Investigación y Análisis* permite a los usuarios reportar casos de corrupción. A pesar de contener las palabras "investigación" y "análisis" en el nombre de esta herramienta, la realidad es que no provee ningún análisis sobre la corrupción ni casos de corrupción. Al igual que las demás herramientas, no se provee ningún mecanismo para mostrar la evolución.

Al buscar herramientas similares en Colombia, el grupo de trabajo no encontró ninguna aplicación o herramienta que permita hacer algo parecido a las anteriormente mencionadas. Por el contrario, en la página

[1] El Gobierno Electrónico es la aplicación de las tecnologías de la información y la comunicación (TIC) al funcionamiento del sector público, con el objetivo de incrementar la eficiencia, la transparencia y la participación ciudadana. OEA. (2009, agosto 1). OEA - Organización de los Estados Americanos: Democracia para la paz, la seguridad y el desarrollo. Recuperado el 13 de noviembre de 2017, a partir de http://www.oas.org/es/sap/dgpe/guia_egov.asp

oficial del Observatorio de Transparencia y Anticorrupción, aparece un aviso con el mensaje: "El Observatorio de Transparencia y Anticorrupción no tramita denuncias por casos de corrupción. Para ello, por favor escribir al e-mail de la Secretaría de Transparencia de la Presidencia: contacto@presidencia.gov.co".

Además de esto, el grupo de trabajo observó que cada entidad pública, en su página oficial, especifica cómo se debe denunciar un caso de corrupción para esa entidad, proceso que es generalmente llevado a cabo por correo o de manera presencial. La página oficial de la Presidencia de la República menciona un proceso más general: "La herramienta fundamental que le permite a los ciudadanos colombianos ejercer control sobre actos de corrupción es el derecho de petición. Se trata de un mecanismo que le permite a cualquier persona solicitar información a las entidades públicas para su posterior trámite con los órganos de control" *(Presidencia de la República, 2016).*

En la página oficial del Observatorio de Transparencia y Anticorrupción se pueden ver algunos indicadores de corrupción: http://www.anticorrupcion.gov.co/paginas/Indicadores.aspx. Estos indicadores incluyen: índices de desempeñó fiscal (IDF), índices de gobierno abierto (IGA), indicadores de sanciones e indicadores de transparencia. Sin embargo, la mayoría de los indicadores están en formato CSV (también están disponibles en la página de Datos Abiertos del Gobierno), lo cual no resulta tan fácil de interpretar para un usuario del común. Sentinel busca complementar la información ya disponible en la página del Observatorio, permitiendo visualizar los resultados obtenidos durante el desarrollo del proyecto de manera gráfica para llegar a un mayor número de usuarios y para que los resultados encontrados sean más fáciles de interpretar.

Herramienta	*Permite denunciar casos de corrupción*	*Muestra indicadores de corrupción*	*Permite medir la actividad de medios frente a la corrupción*	*Permite medir la actividad de usuarios frente a la corrupción*	*Se garantiza anonimidad*	*Aplicación web*	*Aplicación Móvil*
Denuncia la Corrupción	✓	×	×	×	✓	×	✓
Denuncia Corrupción	✓	×	×	×	×	✓	×

Denuncia en línea	✓	×	×	×	×	✓	×
Investigación y Análisis	✓	×	×	×	×	✓	×
Página del Observatorio de Transparencia y Anticorrupción	×	✓	×	×	✓	✓	×

Tabla 2 Cuadro comparativo de trabajos relacionados

Como se puede apreciar en la Tabla 2, la mayoría de herramientas anticorrupción existentes se centran en la denuncia de casos de corrupción. La única herramienta que ofrece información sobre el estado actual de corrupción en el país es la página del Observatorio. Sentinel tiene el objetivo de enriquecer la información disponible para el Observatorio, teniendo en cuenta las publicaciones de los medios y los comentarios de los usuarios en la red social Facebook. Como se mencionó en la sección 1.4 (Justificación del Problema), las redes sociales son una gran fuente de información debido a que proponen un entorno activo en el que los usuarios pueden opinar y contrastar.

Vale la pena aclarar que Sentinel no es una "competencia" del Observatorio de Transparencia y Anticorrupción. Todo lo contrario, busca apoyar la labor del Observatorio al permitir obtener opiniones de los usuarios en redes sociales y medir la actividad de los medios con respecto a casos de corrupción.

Uniendo la información obtenida con Sentinel junto con los datos que tiene el Observatorio de Transparencia y Anticorrupción, se pueden derivar conclusiones más completas e informadas sobre el fenómeno de corrupción *(Castañeda, Ortiz, & Pérez, 2017).*

Teniendo en cuenta el estado actual de las herramientas anticorrupción y la información obtenida a partir de las entrevistas con la Secretaría de Transparencia, se decide crear Sentinel: un sistema que es capaz de monitorear la red social Facebook para conocer la percepción que tienen los usuarios sobre noticias de corrupción y además monitorear lo que los medios y líderes de opinión están hablando en materias de corrupción.

Características con las que cuenta Sentinel:

- ✓ *Permite medir la actividad de medios frente a la corrupción*
- ✓ *Permite medir la actividad de líderes de opinión frente a la corrupción*

- ✓ *Permite medir la actividad de usuarios frente a la corrupción*
- ✓ *Se garantiza anonimidad de los usuarios del sistema*
- ✓ *Aplicación web*

Características con las que NO cuenta Sentinel:

- × *Denunciar casos de corrupción*
- × *Mostrar indicadores de corrupción*

No se incluye la funcionalidad de denunciar casos de corrupción debido a que ya existe un proceso para cada entidad de la cual se quiera denunciar un caso. Por otro lado, no se muestran indicadores de corrupción debido a que esto ya se puede ver en la página del Observatorio. A pesar de que Sentinel no es una aplicación móvil nativa, se diseñó teniendo en cuenta el acceso desde dispositivos móviles (*responsive design*), por lo cual los usuarios pueden ingresar a la página de Sentinel desde su dispositivo móvil y la página tendrá un despliegue adecuado.

En los siguientes capítulos se presenta Sentinel: los resultados obtenidos en cada fase de las metodologías utilizadas, el modelo de analítica desarrollado y las técnicas implementadas. Además, se muestran las partes del componente de visualización, así como las pruebas realizadas del sistema, tanto funcionales como con el usuario final: la Secretaría de Transparencia.

CAPÍTULO 4: SENTINEL: Un sistema para el monitoreo de casos de corrupción en Facebook

El desarrollo de la solución siguió la metodología *CRISP-DM (Chapman & Clinton, 2000)* para determinar las necesidades de analítica frente al tema de corrupción. En este capítulo se presentan los resultados de este proceso para cada fase de esta metodología.

1. Entendimiento del Negocio

Para poder llevar a cabo la realización del proyecto planteado por el grupo de trabajo fue necesario contar con la ayuda de expertos, uno de ellos fue el Observatorio Nacional de Transparencia y Anticorrupción colombiano debido a su iniciativa, su gran interés y alto de nivel de análisis del fenómeno de la corrupción en Colombia (*Observatorio de Transparencia y Anticorrupción, 2017).*

¿Cómo opera el Observatorio?

El Observatorio es liderado por la Secretaría de Transparencia de la Presidencia de la República, con el apoyo del Área de Información y Sistemas de la entidad. Para el proceso de formulación, cálculo, análisis

y presentación de indicadores, el Observatorio cuenta con una Mesa Técnica, convocada por la Secretaría de Transparencia, en la que participan representantes de las distintas entidades integrantes de la Comisión Nacional de Moralización. A través de estos encuentros se definen los protocolos para el intercambio de la información necesaria para alimentar los indicadores, se revisan los resultados arrojados por los indicadores existentes y se proponen nuevos frentes de trabajo (*Observatorio de Transparencia y Anticorrupción, 2017).*

¿Quiénes componen el Observatorio?

El Observatorio fue concebido como una herramienta a disposición de la Comisión Nacional de Moralización, y como nodo de articulación institucional y punto de convergencia de estas entidades con la sociedad civil y la comunidad *(Observatorio de Transparencia y Anticorrupción, 2017).*

Ilustración 3 Composición del Observatorio

Datos de la entidad	
Nombre	Observatorio de Transparencia y Anticorrupción
Ubicación	Carrera 8 # 12B - 61 Piso 10 Edificio BIC - Bogotá D.C

Tabla 3 Datos del Observatorio Transparencia y Anticorrupción

Punto de contacto	
Paula Castañeda	paulacastenada@presidencia.gov.co

Tabla 4 Datos de contacto del Observatorio Transparencia y Anticorrupción

Los objetivos principales de negocio del observatorio con el análisis que se va a realizar son:

Objetivo general

Crear un instrumento capaz de monitorear los casos de corrupción, la actividad de los medios y de los usuarios de Facebook frente a estos sucesos, donde se ven involucradas diferentes entidades, para brindar los insumos necesarios que permitan hacer un diagnóstico de la corrupción en esta red social.

Objetivos específicos

- El instrumento debe permitir mejorar la forma de hacer monitoreo de medios en la red social Facebook.
- El instrumento debe permitir contrastar la percepción de la corrupción en la red social Facebook contra las encuestas realizadas en Colombia (LAPOP) para poder determinar si se debe tratar cada una de forma diferente.

Una vez evaluada la situación actual del Observatorio Nacional de Transparencia y Anticorrupción el grupo de trabajo identificó la oportunidad de proveer una percepción social frente a la corrupción y de esta manera apoyar la toma de decisiones para poder crear propuestas, alternativas y ser agentes de cambio en un ambiente virtual y social frente a la corrupción en Colombia.

El Observatorio Nacional de Transparencia y Anticorrupción decidió apoyar al grupo de trabajo con información y lineamentos para llevar a cabo un trabajo de calidad y con impacto social, con ánimos de enriquecer y fortalecer su valor, para así poder contar con una perspectiva social y directa con la ciudadanía.

A continuación, se listarán los recursos físicos con los que cuenta el proyecto para su elaboración y ejecución.

NOTA: Algunos recursos listados a continuación provienen de diferentes fuentes, por esta razón se especifica el nombre del recurso, la descripción y el dueño.

Archivos		
Nombre	**Descripción**	**Dueño**
Lexicones de sentimientos (CAOBA, 2017)	Desarrollado por la Alianza CAOBA.	Alianza CAOBA
Casos representativos de corrupción (2000-2015) (Ortiz, 2017)	Es una colección de los casos más representativos de corrupción colombianos identificado por el Observatorio Nacional de Transparencia y Anticorrupción.	Observatorio Nacional de Transparencia y Anticorrupción – Secretaria de Transparencia

Análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes sociales (Caicedo, Carrillo, Forero, & Urueta, 2017)	Fue un trabajo de grado realizado por estudiantes de Ingeniería industrial bajo la asesoría de Jorge Alvarado. En este trabajo se desarrolla un diccionario de sentimientos enfocado a las redes sociales y el tema de política.	Luis Eduardo Caicedo Ortiz, Angie Carrillo Chappe, Catalina Forero Arévalo, Julián David Urueña García, Juan Camilo Urueña. Director TG: Alvarado Valencia, Jorge Andrés
RART 2.0 (Facebook Extractor)	Basado en el trabajo de grado RART desarrollado por el grupo de trabajo, este extractor de datos está desarrollado e implementado en Python y permite extraer constantemente los datos de Facebook *(Calambás Marin & Mendoza, 2016).*	Grupo de trabajo.

Tabla 5 Recursos provistos por diferentes entidades

Adicionalmente, se contó con la ayuda y asistencia de tres recursos humanos de la Secretaría de Transparencia, los cuales dieron los lineamientos necesarios para llevar a cabo el entendimiento y contexto del problema de negocio.

Vocabulario

Basándose en los lineamientos, información recopilada por el grupo de trabajo y con la asistencia del Observatorio Nacional de Transparencia y anticorrupción, fue posible identificar un vocabulario propio del negocio en donde se describen entidades y su detalle, esto con el objetivo propio de poder diagramar y representar por medio de un modelo de dominio el mundo de la corrupción sobre el cual el trabajo de grado desarrollará el proyecto.

Vocabulario (Concepto)	Descripción
Reacción	Identifica las reacciones que los diferentes tipos de cuenta, perfil o página, de Facebook pueden dar a una publicación o comentario; *Like, Love, Angry, Sad.*
Opinión	Representa el punto de vista propuesto por una persona sobre una publicación. En Sentinel, la opinión puede ser positiva, negativa o neutral.

Comentario	Representa una cadena de caracteres la cual contiene el comentario hecho por un perfil o página.
Persona Natural	Representa a aquella persona que no es una entidad.
Caso de Corrupción	Representa los casos de corrupción más notables en el país; estos casos de corrupción fueron propuestos por el Observatorio Nacional de Transparencia y Anticorrupción y deben poder ser configurables.
Entidad	Representa la asociación de personas de cualquier tipo; mayor nivel de abstracción.
Página	Representa una posible herencia del concepto Cuenta; identifica una Fan Page en Facebook.
Publicación	Representa la publicación de una página o perfil; puede ser un texto, imagen, link, entre otros.
Partido	Representa una posible herencia del concepto Entidad; esta identifica a los partidos políticos
Pública	Representa una posible categorización del concepto Entidad; instituciones públicas o del estado.
Privada	Representa una posible categorización del concepto Entidad; instituciones privadas o compañías.
Tiempo	Representa el tiempo; periodos de tiempo, momentos puntuales en un año, meses, semanas, días. Este concepto se ve involucrado y relacionado en todo el modelo y es de vital importancia.
Involucrado	Representa el involucrado en uno o varios casos de corrupción. Este involucrado se representa como una Entidad.
Cuenta	Representa el tipo de cuenta en Facebook, puede ser Perfil o Página.
Perfil	Representa una posible herencia del concepto Cuenta; identifica un perfil de una persona natural.

Tabla 6 Descripción del vocabulario del negocio.

Modelo de dominio

Una vez identificado y descrito el vocabulario se procede a realizar por medio de un diagrama de clases el modelo de dominio, en donde se identifican los conceptos clave en el dominio de la corrupción en el ámbito colombiano, su rol y su comportamiento entre ellos. El grupo de trabajo hace un mapeo de los conceptos propios del negocio y los componentes de la red social Facebook que Sentinel va a utilizar.

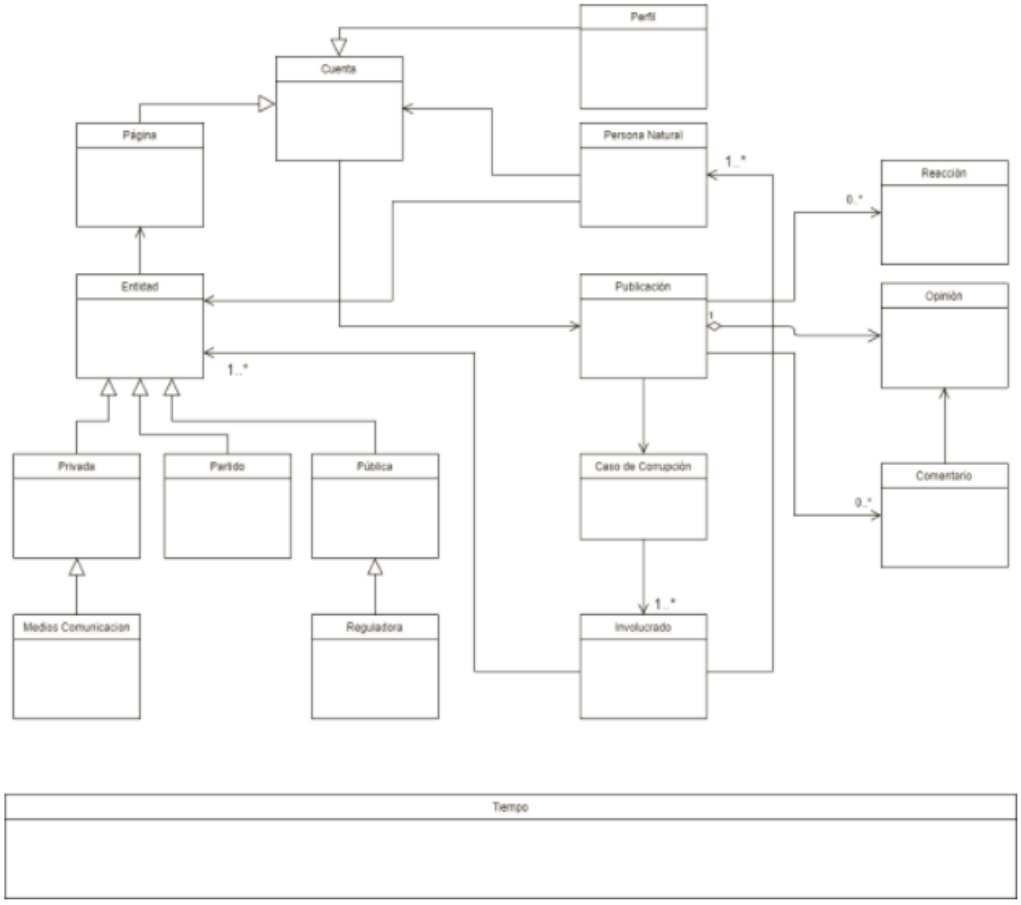

Ilustración 4 Modelo de dominio.

El modelo de dominio fue revisado por la Secretaría de Transparencia, la cual recalcó que la entidad *Partido* se debería considerar diferente a una entidad privada y/o pública (antes el grupo la consideraba como una entidad pública). Esto debido a que los partidos políticos se financian de manera privada, pero ejercen una función pública. La ilustración 3 representa el modelo de dominio final.

En el modelo de dominio se visualiza la relación entre las diferentes entidades (privadas, públicas, partidos), personas, casos de corrupción y usuarios de Facebook, así como toda su actividad (comentarios, publicaciones, reacciones). El tiempo es algo transversal al modelo, razón por la cual se extiende a lo largo del modelo de dominio. Por organización del modelo no se muestran las asociaciones de *Tiempo* con todas las entidades. El concepto de tiempo es de especial importancia debido a que al Observatorio le interesa ver

tendencias sobre las diferentes entidades, para así ver en qué períodos hay más actividad de los usuarios o publicaciones por parte de medios.

Determinar los objetivos de minería

Considerando los objetivos de negocio se plantearon cinco objetivos de minería:

- Describir el nivel de agrado y desagrado de los usuarios en Facebook por medio de algoritmos de análisis de sentimientos.
- Hacer un monitoreo de los casos de corrupción más relevantes en Colombia por medio de sus publicaciones, comentarios y reacciones a lo largo de un periodo de tiempo definido.
- Utilizar algoritmos descriptivos (correlación) para encontrar los temas más relacionados a la corrupción y diferentes entidades.
- Desarrollar un componente de analítica de datos, compuesto por diferentes módulos (algoritmos) enfocados a aplicar las siguientes técnicas:
    - Análisis de sentimientos
    - Asociación de palabras
    - Funciones agregadas
    - Funciones compuestas
- Desarrollar un componente que permita visualizar e interpretar los resultados obtenidos al aplicar las técnicas de analítica seleccionadas.

Criterio de éxito de minería

Se establecen dos criterios de éxito para considerar que el proyecto Sentinel fue satisfactorio:

1. Se logran responder las preguntas con prioridad 2 y 3 (ver Anexo B Sentinel - SRS). 3 es la prioridad más alta, y esta prioridad la tienen las preguntas que más les interesa conocer al Observatorio de Transparencia y Anticorrupción, por lo cual son las primeras en desarrollarse en los componentes de analítica y visualización.
2. El sistema se valida con las personas que han apoyado al grupo de trabajo en la elaboración del proyecto. Esta validación se hace con la encuesta TAM. El proyecto se considera exitoso si se obtiene un promedio mínimo de 3.0 en cada eje evaluado por la encuesta TAM.

1.1 Requerimientos derivados

Gracias a la construcción del modelo de dominio, fue posible identificar con los expertos del negocio, las preguntas a nivel de negocio que concretan el cumplimiento de los objetivos del proyecto. Estas preguntas fueron validadas por el Observatorio Nacional de Transparencia y Anticorrupción y el grupo de

trabajo. A continuación, se listan las preguntas que fueron categorizadas por concepto del modelo de dominio:

ID	Entidad
P - 001	¿Cuáles son las entidades asociadas a un caso de corrupción?
P - 002	¿Cuántas son las entidades asociadas a un caso de corrupción?
P - 003	¿Cuántas entidades privadas están asociadas a un caso de corrupción?
P - 004	¿Cuáles entidades privadas están asociadas a un caso de corrupción?
P - 005	¿Cuántas entidades públicas están asociadas a un caso de corrupción?
P - 006	¿Cuáles entidades públicas están asociadas a un caso de corrupción?
P - 007	¿Cuál es el porcentaje por reacción de una publicación hecha por un tipo de entidad Privada/Publica?
P - 008	¿Qué tipo de entidades publican más sobre casos de corrupción?
P - 009	¿Cuál es el sector (privado-publico) más involucrado en casos de corrupción?
P - 010	¿Cuál es la entidad con más seguidores?
P - 011	¿Cuál es la entidad que más ha tenido actividad (Publicaciones, opinión) durante un periodo de tiempo sobre un caso de corrupción?
P - 012	¿Cuál es la entidad más ha se ha visto involucrada en publicaciones durante un periodo de tiempo?
P - 013	¿Cuál es la entidad más ha se ha visto involucrada en publicaciones respecto a un caso de corrupción?
P - 014	¿Existen relaciones entre los dos sectores respecto a un caso de corrupción?
P - 015	¿Es posible detectar el sesgo de los medios de comunicación respecto a los casos de corrupción?
P - 016	¿Es posible detectar el sesgo de los medios de comunicación respecto a las publicaciones que hacen?

Tabla 7 Requerimientos de entidad.

ID	Caso de corrupcion
P - 017	¿Cuál es el caso de corrupción que recibe más reacciones(like-dislike) y opiniones?
P - 018	¿Dado un caso de corrupción, es posible comparar dos líderes de opinión basándose en like-dislike?
P - 019	¿Cuál es el caso de corrupción que más ha tenido actividad (Publicaciones, opinión) durante un periodo de tiempo?
P - 020	¿Cuál caso de corrupción genera el mayor malestar (dislike) en los usuarios?
P - 021	¿Cuál es el numero promedio de involucrados en un caso de corrupción?
P - 022	¿Cuál es el numero promedio de publicaciones respecto a un caso de corrupción en particular?
P - 023	¿Cuánto tiempo toma un caso de corrupción en ser 'olvidado' por los medios de comunicación?
P - 024	¿Cuánto tiempo toma un caso de corrupción en ser viral en un periodo de tiempo?
P - 025	¿Cuántas veces es mencionado un caso de corrupción?
P - 026	¿Es posible comparar los casos de corrupción? (Basándose en criterios; mas mencionados, más involucrados, mas like-dislike, entre otros.)
P - 027	¿Es posible comparar los casos de corrupción dado un periodo de tiempo? (Basándose en criterios; mas mencionados, más involucrados, mas like-dislike, entre otros.)
P - 028	¿Qué palabras se suelen mencionar cuando se habla de casos de corrupción? (Asociación de palabras)

Tabla 8 Requerimientos de caso de corrupción.

ID	Personas naturales
P - 029	¿Cuáles son los líderes de opinión que más comentan sobre corrupción?
P - 030	¿Cuántas reacciones positivas y negativas tienen los usuarios a las publicaciones de los lideres?
P - 031	¿Cuáles líderes de opinión están más involucrados con corrupción?
P - 032	¿En las publicaciones sobre corrupción los comentarios hechos por los usuarios, cuántas líderes de opinión son nombrados con más frecuencia en esos comentarios?
P - 033	¿Cuánto tiempo nombran a un líder de opinión política en comentarios y publicaciones?
P - 034	¿Qué tienen en común los comentarios hechos por los usuarios con más likes?
P - 035	¿Cuál es la reacción de los líderes de opinión ante noticias de corrupción?
P - 036	¿Cuáles líderes políticos están nombrados en organizaciones involucradas en corrupción?
P - 037	¿Cuáles lideres están más nombrados en los comentarios sobre noticias de corrupción?
P - 038	¿Cuáles son los sentimientos de los usuarios sobre los líderes políticos?

Tabla 9 Requerimientos de personas naturales.

ID	Publicacion
P - 039	¿Cuáles páginas publican más sobre corrupción?
P - 040	¿Cuál es la opinión de las páginas (positivo/negativo/neutral) acerca de los líderes políticos y entidades?
P - 041	¿Cuáles son las páginas cuyas publicaciones sobre un líder político o entidad se comparten más?
P - 042	¿Cuáles son las páginas cuyas publicaciones de corrupción se comparten más?
P - 043	¿Cuáles publicaciones de corrupción (ej: Odebrecht, Reficar, etc) generan más reacciones (like/angry/sad/etc)?
P - 044	¿Cuáles publicaciones de corrupción generan más comentarios?
P - 045	¿Cuáles son los casos de corrupción que más se publican?
P - 046	Comportamiento de publicación de casos de corrupción a través del tiempo
P - 047	Del total de publicaciones de una página, ¿cuál es la proporción de noticias de corrupción?
P - 048	¿Cuántas publicaciones de corrupción son del sector privado/público?
P - 049	¿Qué palabras se suelen mencionar cuando se habla de corrupción? (Asociación de palabras)
P - 050	¿Qué región del país está más involucrada en noticias de corrupción?
P - 051	¿De dónde son los corruptos que se mencionan en noticias?

Tabla 10 Requerimientos de publicaciones.

ID	Comentarios
P - 052	¿Los comentarios sobre corrupción son un reflejo de la realidad (comparar con encuestas)?
P - 053	¿Cuáles son las instituciones del Estado más atacadas por corrupción en los comentarios?
P - 054	¿Cuál es la opinión (positiva/negativa/neutral) de los usuarios sobre líderes políticos y entidades?

Tabla 11 Requerimientos de comentarios.

ID	Involucrados
P - 055	¿Cuáles son las personas o entidades más involucradas en casos de corrupción?
P - 056	¿Del total de entidades involucradas en un caso de corrupción, cuántas son Entidades públicas y privadas?
P - 057	¿Del total de entidades involucradas en un caso de corrupción, cuántas son entidades públicas reguladoras?
P - 058	¿Del total de entidades involucradas en un caso de corrupción, cuántas son entidades públicas, partidos?
P - 059	¿Del total de entidades involucradas en un caso de corrupción, cuántas son entidades privadas?

Tabla 12 Requerimientos de involucrados.

Priorización de requerimientos

Una vez identificadas las preguntas, su ID y su correcta categorización, el grupo de trabajo procedió a priorizar los requerimientos de la siguiente manera:

Razon	Valor
Bajo	1
Medio	2
Alto	3

Tabla 13 Tabla de prioridades.

El grupo de trabajo realizó la priorización de los requerimientos bajo los objetivos del negocio previamente identificados. Esta categorización se muestra en la tabla 15. Los requerimientos de valor 1 no están contemplados en el alcance del proyecto por motivos de tiempo y priorización. Se espera que en trabajos futuros sean desarrollados e implementados.

A continuación, se presenta la priorización final de estado Alto a Bajo:

ID	Valor	ID	Valor	ID	Valor
P-001	3	P-020	3	P-049	3
P-002	3	P-023	3	P-052	3
P-003	3	P-024	3	P-053	3
P-004	3	P-025	3	P-054	3
P-005	3	P-026	3	P-022	2
P-006	3	P-027	3	P-030	2
P-007	3	P-028	3	P-040	2
P-008	3	P-029	3	P-018	1
P-009	3	P-031	3	P-021	1
P-010	3	P-032	3	P-033	1
P-011	3	P-037	3	P-034	1
P-012	3	P-038	3	P-035	1
P-013	3	P-039	3	P-036	1
P-014	3	P-043	3	P-041	1
P-015	3	P-044	3	P-042	1
P-016	3	P-045	3	P-047	1
P-017	3	P-046	3	P-050	1
P-019	3	P-048	3	P-051	1

Tabla 14 Tabla de categorías.

1.2 Restricciones identificadas

Con ánimos de garantizar calidad y cumplimiento en el desarrollo del proyecto, se identifican restricciones y limitaciones por parte del equipo y expertos del negocio.

Tipo de restricción	Restricción
Datos	Se extraerán datos desde el 1 de enero de 2016 hasta el 28 de octubre de 2017.
Herramientas	Se utilizarán herramientas *open source* para el desarrollo de la plataforma.
Datos	Se extraerán datos en el idioma español.
Datos	Se extraerán datos única y exclusivamente de los perfiles y páginas públicas de Facebook.
Herramientas	Se utilizará R, Python para el desarrollo del módulo de analítica.
Herramientas/Datos	Se utilizará una base de datos no relacional para modelar los datos extraídos de Facebook.
Herramientas	Se utilizarán los siguientes sistemas operativos: Windows 10.

	Ubuntu 16.04. macOS Sierra.

Tabla 15 Restricciones del proyecto en términos de minería de datos.

2. Entendimiento de los datos

Los datos serán extraídos de la red social Facebook y para llevar a cabo la extracción es necesario primero entender qué tipo de datos se necesitan, su relación y naturaleza dentro la red social Facebook. Inicialmente se requieren de unos pre-requisitos antes de llevar a cabo el proceso de extracción:

1. Creación de una aplicación en Facebook *Developer* https://developers.facebook.com/
2. Solicitar un *Access Token* para poder autenticarse con Facebook y validar la identidad.

Una vez cumplidos los requisitos previamente mencionados, se procede entonces a utilizar la herramienta de extracción seleccionada. Para el caso del proyecto, el grupo de trabajo cuenta con un código desarrollado en Python que por medio de un protocolo HTTPS/REST consume la API de Facebook, conocida como *Graph API*, para obtener estados y publicaciones junto con otro conjunto de atributos descritos más adelante. Este código está basado en la herramienta RART*(RART, 2016)*.

Los datos son extraídos en formato JSON de Facebook *(Facebook, 2017)* y son posteriormente almacenados en una base de datos no relacional, conocida como *MongoDB (MongoDB, 2017).* En esta base de datos, los documentos se guardan en formato JSON. Estos documentos se almacenan en colecciones. Se puede pensar en las colecciones como el equivalente a tablas en una base de datos relacional.

A continuación, se describen los elementos generales de Facebook, con ánimos de entender la anatomía de la red social y su manera de representar los datos que la componen.

Elemento	Descripción
Fan Page	Página pública de Facebook que ha sido diseñada con ánimos de compartir información sobre un interés mutuo y captan la atención de varios usuarios *(Comunitarios, 2014).*
Comentario	Un comentario es una apreciación sobre cualquier tema puesto en análisis *(Comunitarios, 2014).*
Publicación	Es la idea de hacer que una información se conozca y salga del ámbito de lo privado a lo público *(Comunitarios, 2014).*

Reacción	Es la respuesta a una publicación; se conocen como emociones: positiva, negativa, humor, furor *(Comunitarios, 2014)*.
Like/Dislike	Reacción positiva o negativa de publicaciones y/o comentarios *(Comunitarios, 2014)*.

Tabla 16 Descripción de elementos de Facebook.

Por otro lado, el sistema debe tener la capacidad de extraer las publicaciones y comentarios que hablen de corrupción, por esta razón el grupo de trabajo utilizará expresiones regulares para poder identificar las publicaciones adecuadas. Para ver la base de conocimiento ver el anexo G. Base de conocimiento

Descripción inicial de datos

El sistema extraerá publicaciones y comentarios de los medios de comunicación, líderes de opinión y demás figuras cuya presencia en la red social sea fuerte y se necesite monitorear. Estas entidades fueron provistas por el Observatorio Nacional de Transparencia y Anticorrupción. Por esta razón, y con ánimos de garantizar la disponibilidad de tiempo para el procesamiento de datos, se extraerán los datos hasta el 28 de octubre de 2017.

Extrayendo publicaciones y comentarios de los 24 medios de comunicación provistos por el Observatorio y los 8 líderes de opinión que tienen página de Facebook, desde el 1 de enero de 2016 hasta el 28 de octubre de 2017, el espacio de almacenamiento que ocuparon los datos fue de 14.72 GB. Con base en esta información, se estima que mensualmente se extraen alrededor de 670 MB de publicaciones y comentarios, y al año alrededor de unos 8 GB.

Para que los datos sean apropiados para el desarrollo del proyecto, se desea solo extraer los elementos de Facebook presentados en la ilustración 5 y los atributos que lo componen identificando su nombre y una descripción de cada atributo:

Ilustración 5 Modelo de datos

1. **Posts**
    a. *_id*: representa el ID del post.
    b. *created_time*: representa la fecha de creación de la publicación.
    c. *message*: representa la cadena de caracteres de la publicación.
    d. *name*: representa el nombre de la publicación.
    e. *description*: representa una descripción de la publicación.
    f. *shares*: representa la cantidad de veces que una publicación fue compartida.
    g. *link*: representa la URL de la publicación (ej: URL que comparte un medio de una noticia).
    h. *reactions***:**
        i. sad: representa la cantidad de reacciones *sad* que recibió la publicación.
        ii. haha: representa la cantidad de reacciones *haha que recibió la publicación*.
        iii. love: representa la cantidad de reacciones *love* que recibió la publicación.
        iv. like: representa la cantidad de reacciones *like* que recibió la publicación.
        v. wow: representa la cantidad de reacciones *wow* que recibió la publicación.
        vi. *angry*: representa la cantidad de reacciones *angry* que recibió la publicación.
2. **Comments**
    a. *_id*: representa el ID del comentario.
    b. *created_time*: representa la fecha de creación del comentario.
    c. *message*: representa la cadena de caracteres del mensaje.
    d. *like_count*: representa la cantidad de veces que los usuarios le dieron *like* a ese comentario.

Como se puede detallar en el diagrama, todos los elementos tienen un identificador único denominado "*_id*", junto con una fecha y los atributos correspondientes a cada elemento.

Debido a que la diferencia entre "*message*", "*name*" y "*description*" de una publicación es difícil de comprender, a continuación, se muestra una publicación de Facebook y se identifican sus diferentes partes:

Ilustración 6 Modelo de datos

Vale la pena mencionar que no todas las publicaciones cuentan con estas tres partes. Es decir, no son obligatorias.

Atributos numéricos

La tabla 18 muestra los atributos numéricos extraídos a partir de publicaciones y comentarios.

Atributo	Mínimo	1er cuartil	Mediana	Media	3er cuartil	Máximo	Desv. Estánd.
shares	0	1	6	201.7	34	462640	3400.667
Sad	0	0	0	22.18	2	34642	244.037
wow	0	0	0	15.29	4	16902	103.146
love	0	0	1	24.91	5	17263	193.894
Like	0	9	43	349.2	184	118635	1841.25
angry	0	0	0	33.2	4	39719	270.592
haha	0	0	0	24.94	3	22472	255.841
like_count	0	0	0	1.521	1	9798	18.866

Tabla 17 Atributos numéricos.

Exploración inicial de datos

En la exploración inicial de los datos, realizada en R *("R: What is R?", 2017),* se generaron varios gráficos para mostrar propiedades generales de los atributos numéricos (ej: varianza, asimetría, etc.). A continuación, se muestra un *boxplot* de la cantidad de veces que se compartieron noticias de los diferentes medios monitoreados (por legibilidad se muestran cuatro medios de los 24 analizados):

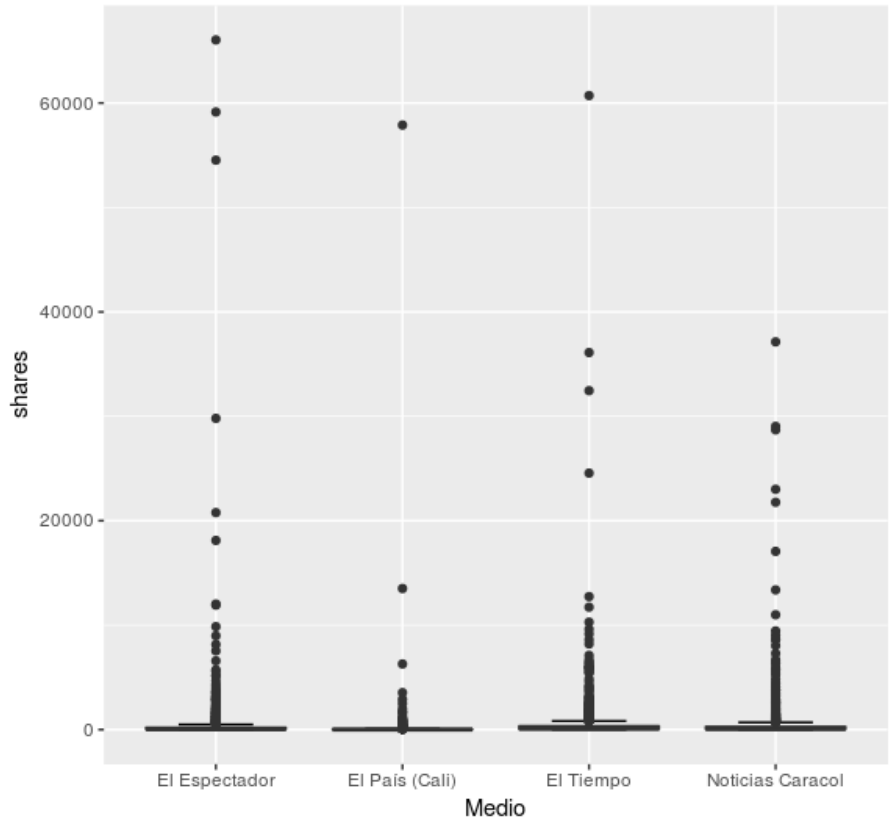

Ilustración 7 Cantidad de publicaciones que comparten los usuarios por página de noticias

Se nota un alto número de publicaciones *outliers*, debido a que se encuentran por fuera de los "bigotes" de las cajas. El tamaño pequeño de las cajas indica una baja varianza, existen muchos puntos cercanos a 0. Esto nos dice que, en general, los usuarios comparten poco las publicaciones de una página de noticias.

Para más detalle sobre la exploración de los datos, ver Sección 2.3 "Exploración inicial de datos" del Anexo A Sentinel - CRISP-DM

3. Preparación de datos

Las actividades realizadas para generar la vista minable y posteriormente realizar el modelado fueron:

1. Seleccionar los datos

Para realizar esta actividad se seleccionaron todas las variables que contenga texto, debido a que es la fuente principal de datos para el análisis. En todos los análisis que se van a realizar, el atributo *created_time* es necesario, debido a que, en la mayoría de casos, la visualización se realiza dentro de un rango de tiempo específico. Por ejemplo: el usuario puede ver los sentimientos hacia un líder político a través del tiempo.

2. Limpieza de datos

Esta actividad se dividió en dos partes: Básica y avanzada

Pre procesamiento básico:

- Eliminación de direcciones URL.
- Los caracteres desconocidos se asignan a su variante ASCII más cercana, utilizando el módulo Python *Unidecode*.
- Se remueven los *stop words* utilizando el módulo Python NLTK.
- Eliminación de signos de puntuación
- Eliminación de acentos (tildes, virgulillas, diéresis, etc.)

Preprocesamiento avanzado:

Se hace *stemming* utilizando el módulo Python NLTK. Específicamente se utiliza el *Snowball Stemmer* (Goodger, 2013) para el idioma español.

3. Construcción de datos

→ Se genera un nuevo atributo llamado "*stemmed*" para la colección de "*comments*" y "*posts*". En ese atributo guarda la versión *stemmed* de los mensajes (ya sean comentarios o publicaciones) para evitar volver a calcularlo de manera innecesaria.

→ Se genera un nuevo atributo llamado "*whole_sentence*", donde se guarda la versión preprocesada del texto (sin *stop words*, acentos, signos de puntuación, etc.), pero sin aplicar el *stemming*. En este atributo, se concatenan los atributos originales "*message*", "name" y "*description*" (de las publicaciones), para que quede todo en un solo atributo. El atributo "*whole_sentence*" se utiliza en el algoritmo de asociación de palabras.

4. Integración de datos

No se realiza ninguna integración de datos, debido a que la fuente única de datos es Facebook.

Resultados de la fase:

- Plan de preprocesamiento de datos (ej. *stemming*)
- Limpieza de datos
- Generación de atributos derivados

4. Modelo

La fase de modelado es una de las más importantes de la metodología CRISP *(Chapman & Clinton, 2000)* y fue la fase en donde se invirtió la mayor cantidad de tiempo. En esta fase, se creó el modelo de analítica, identificando y aplicando las técnicas de análisis y minería de datos. Además, se calibraron los parámetros de los algoritmos para obtener el mejor resultado posible. Basándose en las preguntas generadas y los requerimientos funcionales, se procedió a generar el siguiente modelo de analítica:

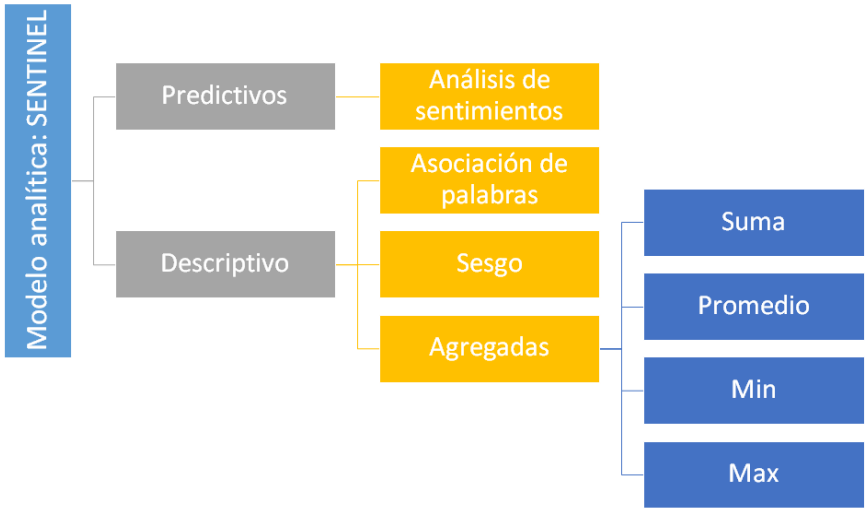

Ilustración 8 Modelo de analítica.

Una vez identificadas las agrupaciones generales, se proceden a seleccionar las técnicas de minería adecuadas para responder a las preguntas planteadas por el negocio. En la ilustración 9 se muestran los algoritmos que se implementaron o se utilizaron por cada categorización. Es necesario aclarar que para resolver una pregunta es posible que se necesite uno o más algoritmos.

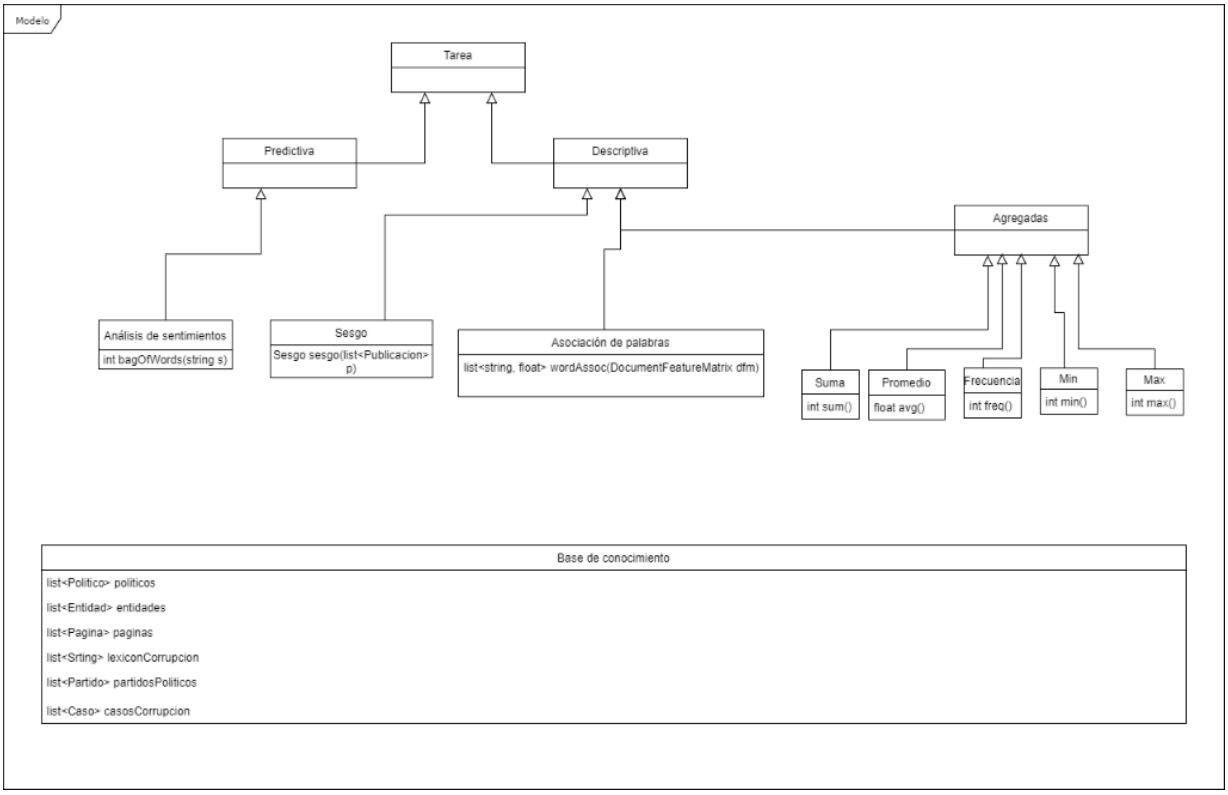

Ilustración 9 Modelo de analítica con algoritmos.

Basado en el modelo de analítica definido y tomando los requerimientos priorizados, se implementaron las técnicas y en paralelo se desarrolló el componente de visualización de las técnicas. Por medio del componente de visualización de Sentinel, el usuario final puede ver los resultados de los algoritmos previamente ejecutados por el componente de analítica.

4.1 Desarrollo de Análisis de Sentimientos

Descripción:

- **Objetivo:** determinar la polaridad de los comentarios sobre líderes de opinión, partidos políticos e instituciones a través del tiempo, para ver cómo evoluciona el nivel de agrado y desagrado de los usuarios de Facebook sobre estas entidades.
- **Algoritmo núcleo:** *Bag-of-words*
- **Entrada:** cadena de caracteres
- **Salida del algoritmo:** -1 (sentimiento negativo), 0 (sentimiento neutral), 1 (sentimiento positivo)
- **Salida en el componente de visualización:** línea de tiempo con la cantidad de sentimientos positivos, negativos y neutrales que se han hecho sobre un líder de opinión, partido político o institución a través del tiempo.

Inicialmente, el grupo de trabajo planteó utilizar una solución comercial de análisis de sentimientos; específicamente, IBM *Watson Tone Analyzer (IBM, 2017)* y *Google Cloud Natural Language API* (Cloud Natural Language API, 2017). Sin embargo, estas soluciones eran bastante costosas, por lo cual se recurrió a utilizar un algoritmo de análisis de sentimientos desarrollado por CAOBA en el artículo "*CSL: A Combined Spanish Lexicon - Resource for Polarity Classification and Sentiment Analysis*" *(Moreno-Sandoval et al., 2017).* Aun así, el grupo de trabajo quiso evaluar los resultados de ambas soluciones sobre una muestra clasificada manualmente por los integrantes del grupo y comparar su rendimiento. Los resultados de esta evaluación se muestran más adelante en la sección *4.5 Evaluación de los modelos* del presente capítulo.

El algoritmo de análisis de sentimientos se probó con diferentes diccionarios de sentimientos. Estos diccionarios son listas de palabras, cada una con un valor asociado. Un valor negativo indica que la palabra representa un sentimiento negativo, un valor positivo indica que la palabra representa un sentimiento positivo, y un valor de 0 indica que la palabra es neutral. Para utilizar este algoritmo, fue necesario remover *stop words* y aplicar *stemming*. El proceso se detalla en la sección anterior (Preparación de datos).

A continuación, se presentan capturas de pantalla de la salida del algoritmo de análisis de sentimientos en Sentinel, para diferentes entidades:

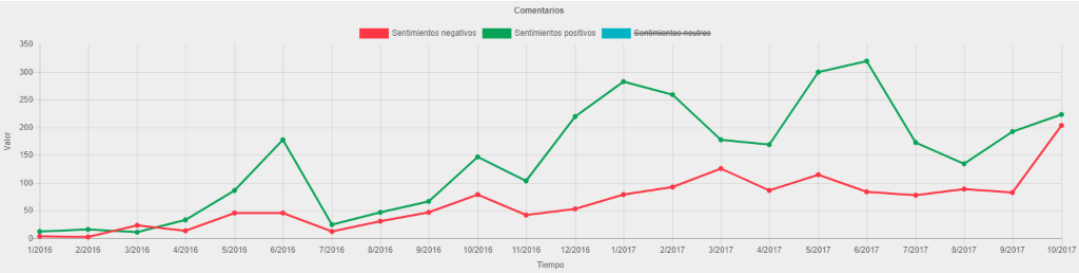

Ilustración 10 Análisis de sentimientos sobre la líder de opinión Claudia López. Línea verde: sentimientos positivos; línea roja: sentimientos negativos; línea azul: sentimientos neutrales (en este caso no se muestran porque los desactivó el usuario).

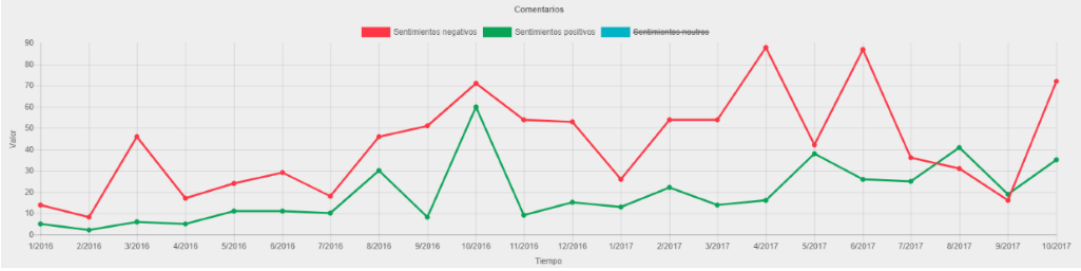

Ilustración 11 Análisis de sentimientos para el partido político Centro Democrático.

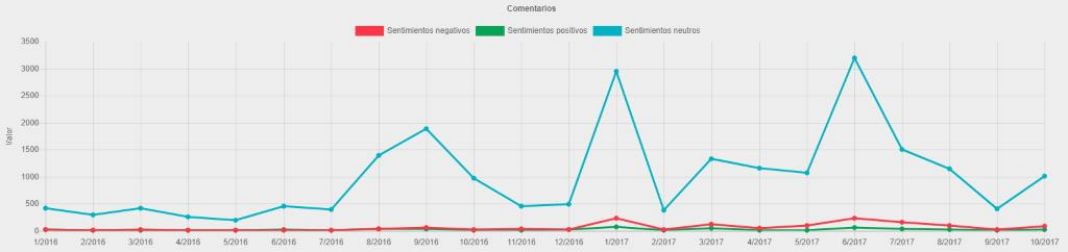

Ilustración 12 Análisis de sentimientos para la institución ONU

4.2 Desarrollo del Algoritmo de Sesgo

Descripción:

- **Objetivo:** medir la inclinación de los medios en base a la cantidad de publicaciones que hacen respecto a casos de corrupción, líderes de opinión y entidades.
- **Algoritmo núcleo:** rango intercuartil
- **Entrada:** cantidad de publicaciones que han hecho todos los medios sobre un caso de corrupción o alguna entidad particular, ya sea líder de opinión, partido político o institución. Esta entrada se espera que sea un arreglo, donde cada elemento contiene un medio y la cantidad de publicaciones que hizo ese medio sobre la entidad específica.
- **Salida del algoritmo:** valores atípicos (*outliers*) identificados, ya sea porque publican considerablemente más respecto a una entidad o caso de corrupción específica comparado al resto de medios o porque publican menos.
- **Salida en el componente de visualización:** una descripción indicando si un medio publica considerablemente más o menos sobre un caso de corrupción o alguna entidad.

En Sentinel, al comparar los medios, se muestra una descripción si pública más o menos sobre corrupción (utilizando las palabras otorgadas por el Observatorio), sobre algún caso de corrupción en particular, o sobre las diferentes entidades como líderes de opinión, partidos políticos e instituciones. En la ilustración 13, el usuario está comparando los medios "Noticias Caracol" y "La F.M.". Se observa que La F.M. publica considerablemente más sobre el caso de Odebrecht comparado a los demás medios mientras que Noticias Caracol publica en una cantidad considerada "normal" (no es un atípico).

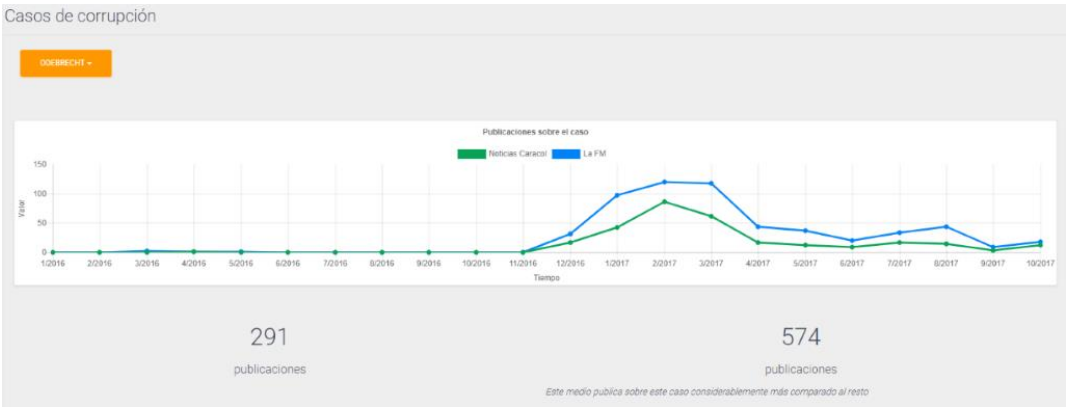

Ilustración 13 El medio La F.M. (línea azul) publicó considerablemente más noticias sobre Odebrecht comparado al resto de medios. Se muestran las publicaciones a través del tiempo y además una descripción debajo del número de publicaciones

Además de comparar dos medios, el usuario puede elegir ver todos los medios al tiempo, como se muestra a continuación en la ilustración 14:

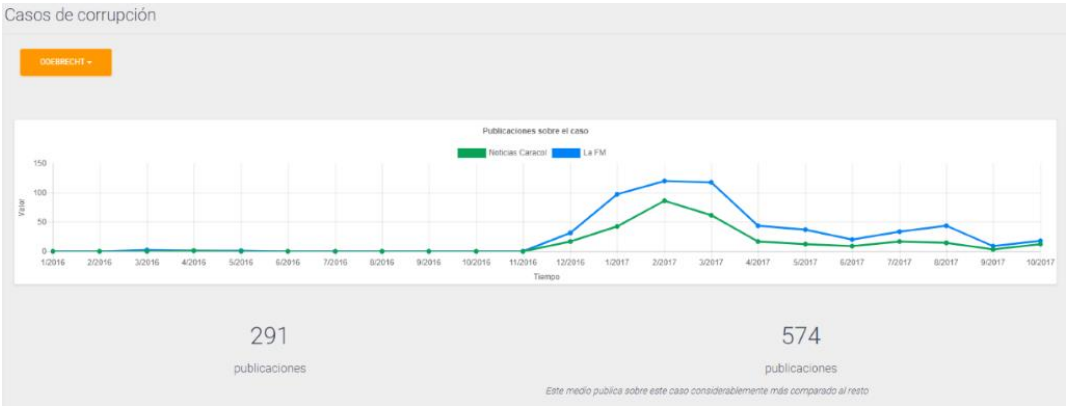

Ilustración 14 Comparación de todos los medios al tiempo. Debajo de la gráfica, a mano derecha, se muestra una descripción que indica que La F.M. es un atípico (publica sobre el caso de Odebrecht considerablemente más que el resto de medios).

El algoritmo de sesgo no solo se aplica para los medios, sino que también se utiliza para identificar cuáles líderes de opinión publican más en su página de Facebook sobre corrupción, comparado al resto, como se muestra en la ilustración 15:

Ilustración 15 Cantidad de publicaciones sobre corrupción hechas por Claudia López en su página de Facebook

4.3 Desarrollo de Asociación de Palabras

Descripción:

- **Objetivo:** encontrar los temas más relacionados a las diferentes entidades y casos/palabras de corrupción a partir de publicaciones de medios y comentarios de usuarios
- **Algoritmo núcleo:** Matriz de correlación
- **Entrada:** publicaciones y comentarios, además de las palabras que son de interés para el Observatorio, para poder encontrar la asociación de esas palabras claves dentro de las publicaciones/comentarios. Estas palabras son: los líderes de opinión, las instituciones, los casos de corrupción, los partidos políticos y las palabras de corrupción (diccionario de corrupción).
- **Salida del algoritmo:** para cada palabra clave, se obtiene una lista de palabras asociadas, con un valor de 0 (no hay ninguna asociación) a un valor de 1 (asociación muy fuerte).
- **Salida en el componente de visualización:** se muestra un grafo que muestra la relación entre las palabras. Los nodos son las palabras y las aristas representan la asociación entre las palabras. Entre mayor es el tamaño del nodo, mayor es la asociación entre las palabras que están conectadas.

Se construye una matriz de documento-palabra *(Document-Feature Matrix*), donde en las columnas se tienen todas las palabras (sin *stop words* ni signos de puntuación), y en las filas se tienen todos los documentos (ej. cada publicación es un documento, igual con los comentarios). Se normaliza la matriz basado en TF-IDF, de tal manera que cada celda de la matriz tiene un valor entre 0 y 1 (según su peso TF-IDF).

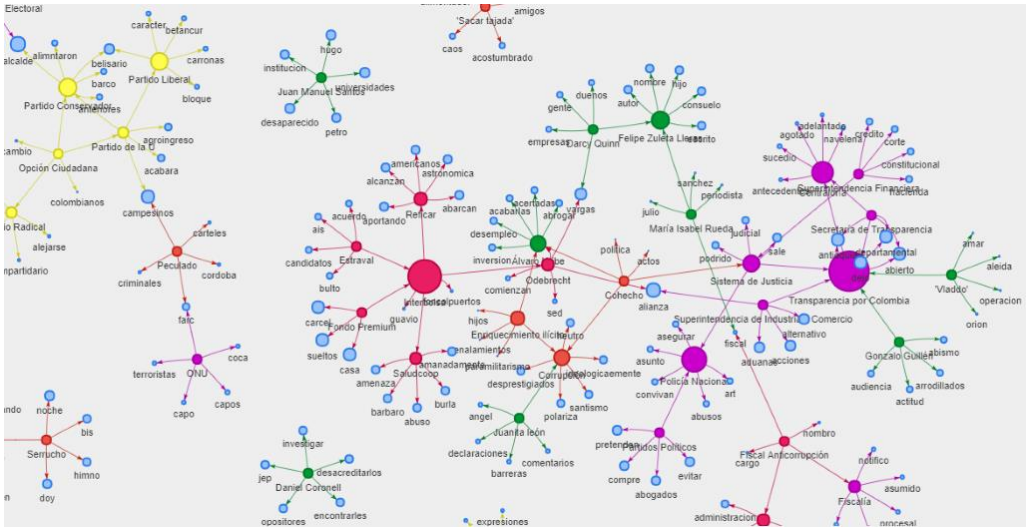

Ilustración 16 Visualización de Asociación de Palabras. Los nodos naranjas representan palabras relacionadas a corrupción, los amarillos partidos políticos, los verdes son líderes de opinión, los rosados son casos de corrupción y los morados instituciones

4.4 Funciones agregadas

Descripción:

- **Objetivo:** realizar cálculos sobre los datos extraídos para realizar mediciones básicas
- **Algoritmos núcleo:** Suma, Resta, Min, Max, Promedio
- **Entradas:** conjunto de datos a ser resumido. En el caso de Sentinel, estos datos pueden ser: publicaciones, comentarios y reacciones.
- **Salidas del algoritmo:** un valor que resuma los datos de entrada. Este valor puede ser la suma, resta, promedio, etc.; dependiendo de lo que se intenta visualizar.
- **Salidas en el componente de visualización:** gráficos de barras, líneas de tiempo.

En Sentinel, las funciones agregadas permiten a la Secretaría de Transparencia ver tendencias en cuanto a la cantidad de publicaciones, comentarios y reacciones a través del tiempo. Las tendencias que se muestran son sobre líderes de opinión, partidos políticos, instituciones y casos de corrupción. A continuación, se muestran algunas capturas de pantalla de las salidas de estas funciones agregadas en el componente de visualización:

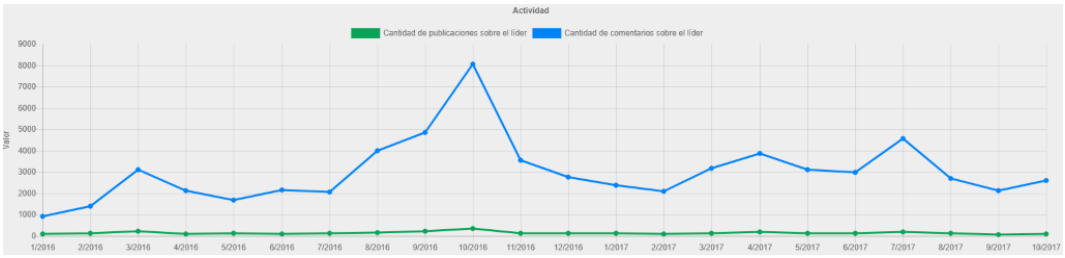

Ilustración 17 Cantidad de publicaciones (línea verde) y comentarios (línea azul) que se han hecho sobe el líder de opinión Álvaro Uribe Vélez a través del tiempo.

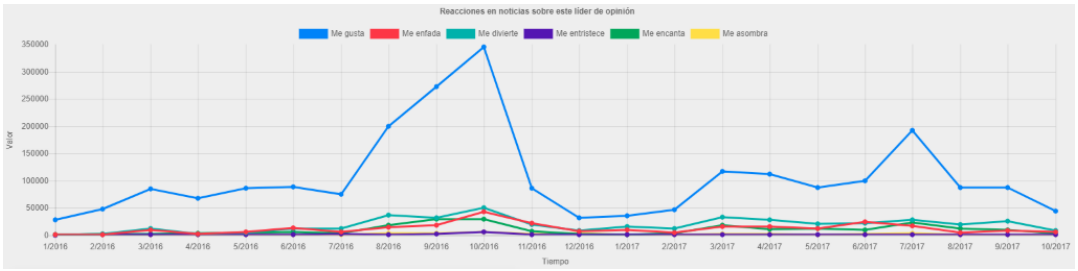

Ilustración 18 Cantidad de reacciones que se han hecho en las publicaciones sobre el líder de opinión Álvaro Uribe Vélez a través del tiempo.

Comentarios populares sobre Juan Manuel Santos este mes

No.	Mensaje	Likes
1	Que porquería, en lo que hemos quedado.. Los ASESINOS de las farc en el congreso!!!. 53 años de asesinatos, secuestros y terror y ahora se creen señores y quieren gobernar... Cárcel para esos bandidos de las farc y destierro para Juan Manuel santos	195
2	Ja, las farc siguen en el monte delinquiendo.... ahora tambien estan en el congreso legalizados x juan manuel santos enmermelador.	139
3	Desde Cesar Gaviria y hasta Juan Manuel Santos son todos caspas, no hay ningún inocente, todos son igual de pícaros y enmermelados, pero creo que el campeón es Álvaro Uribe, hace las embarradas y cuando salió de la presidencia vino a dar clases sobre como se mejora el país, sin haber dado ejemplo de rectitud y transparencia	112
4	Germán Vargas Lleras y pinzón que hacían parte del gobierno de Juan Manuel santos ahora como candidatos presidenciales se dieron cuenta que no les gusta el gobierno , que hipocresía, que mentiras y que cinismo para engañar alos colombianos en su intención de voto para las próximas elecciones presidenciales " el fin justifica los medios " Maquiavelo.	102
5	Los paramilitares están extraditados a los eeuu, en cambio Juan Manuel Santos - Presidente tiene a sus camaradas violadores y reclutadores de niños niñas en el congreso de Colombia W Radio Colombia 🧑⚖️✍️🛰️ extradición a los narcotraficantes FARC.	90

Ilustración 19 Los cinco comentarios más populares sobre el líder de opinión Juan Manuel Santos en octubre de 2017.

Ilustración 20 Cantidad de reacciones que ha recibido Juan Manuel Santos en las publicaciones que ha hecho en su página de Facebook

Ilustración 21 Nube de palabras de las publicaciones realizadas por el líder de opinión Juan Manuel Santos

4.5 Evaluación de los modelos

Evaluación de algoritmo de análisis de sentimientos

- Se realiza un muestreo aleatorio de 200 comentarios sobre líderes de opinión
- Los tres integrantes del grupo clasifican los comentarios manualmente de manera independiente.
- El criterio para decidir la clasificación final del sentimiento es el siguiente:
    - Si al menos dos personas clasificaron el sentimiento con la misma clase, se deja ese sentimiento
    - Si cada persona lo considera como un sentimiento diferente, se deja neutral
- Luego de clasificar los comentarios manualmente, se comparan con los resultados del algoritmo de análisis de sentimientos utilizando una matriz de confusión y se determina la precisión, exhaustividad y exactitud lograda con el algoritmo

En la Tabla 19, se muestra la matriz de confusión utilizando el diccionario político creado en el trabajo de grado "*Análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes sociales." (Caicedo, Carrillo, Forero, & Urueña, 2017)*

	Esperado (manual): negativo	Esperado (manual): neutral	Esperado (manual): positivo	
Predicho (algoritmo): negativo	1	4	0	5
Predicho (algoritmo): neutral	47	82	54	183
Predicho (algoritmo): positivo	0	6	6	12
	48	92	60	

Tabla 18 Resultados algoritmo de análisis de sentimientos con diccionario político

- **Exactitud:** 44.5%
- **Precisión Negativo:** 20.0%
- **Precisión Neutral:** 44.8%
- **Precisión Positivo:** 50%
- **Exhaustividad Negativo:** 2.08%
- **Exhaustividad Neutral:** 89%
- **Exhaustividad Positivo:** 10%

Debido a la baja exactitud del algoritmo de análisis de sentimientos utilizando *bag-of-words*, se prueba con las herramientas de *IBM Watson Tone Analyzer* y *Google Cloud Natural Language API*. Vale la pena aclarar que para utilizar *Watson Tone Analyzer*, fue necesario traducir los comentarios a inglés, mientras que la herramienta de Google tiene soporte nativo para análisis de sentimientos en español. A continuación, se muestran las matrices de confusión con el *cutoff* que generaron una exactitud mayor para cada herramienta:

IBM Watson Tone Analyzer, cutoff = 0.3

	Esperado (manual): negativo	Esperado (manual): neutral	Esperado (manual): positivo	
Predicho (algoritmo): negativo	30	22	25	77
Predicho (algoritmo): neutral	12	15	17	44
Predicho (algoritmo): positivo	14	15	50	79

	56	52	92	

Tabla 19 Matriz de confusión IBM Watson Tone Analyzer, cutoff = 0.3

- **Exactitud:** 47.5%
- **Precisión Negativo:** 39.0%
- **Precisión Neutral:** 34.1%
- **Precisión Positivo:** 63.3%
- **Exhaustividad Negativo:** 53.6%
- **Exhaustividad Neutral:** 28.9%
- **Exhaustividad Positivo:** 54.3%

Google Cloud Natural Language API, cutoff = 0.25

	Esperado (manual): negativo	**Esperado (manual): neutral**	**Esperado (manual): positivo**	
Predicho (algoritmo): negativo	27	18	6	51
Predicho (algoritmo): neutral	18	24	21	63
Predicho (algoritmo): positivo	11	10	65	86
	56	52	92	

Tabla 20 Matriz de confusión Cloud Natural Language API, cutoff = 0.25

- **Precisión Negativo:** 52.9%
- **Precisión Neutral:** 38.1%
- **Precisión Positivo:** 75.6%
- **Exhaustividad Negativo:** 48.2%
- **Exhaustividad Neutral:** 46.2%
- **Exhaustividad Positivo:** 70.7%

A partir de esta evaluación, se determina que la mejor herramienta para el caso de Sentinel es *Google Cloud Natural Language API*. Esta herramienta provee mayor precisión y exhaustividad que el algoritmo de *bag-of-words* y *Watson Tone Analyzer*, por lo cual se sugiere a la Secretaría de Transparencia utilizar la herramienta de Google dado el caso que cuenten con los recursos para costearlo.

Para ver más detalles de la evaluación de las diferentes herramientas con diferentes *cutoffs*, ver la sección *4.4 Evaluación de modelos* en el Anexo Sentinel - CRISP-DM

Evaluación de algoritmos descriptivos

- Se realiza un muestreo aleatorio de 300 comentarios y publicaciones sobre entidades (líderes de opinión, instituciones, partidos políticos) y palabras de corrupción (incluyendo casos de corrupción), haciendo una búsqueda con las expresiones regulares creadas por el grupo.
- 150 comentarios y publicaciones son sobre entidades y los otros 150 son sobre palabras y casos de corrupción.
- Se verifica que en realidad las publicaciones y comentarios extraídos correspondan a entidades y palabras/casos de corrupción, con fines de evaluar qué tan bien desarrolladas están las expresiones regulares. Se obtienen los siguientes resultados:
    - **Publicaciones y comentarios sobre entidades extraídas por el algoritmo:** 150
    - **Publicaciones y comentarios con palabras y casos de corrupción extraídas por el algoritmo:** 150
    - **Publicaciones y comentarios que realmente mencionaban entidades:** 139
    - **Publicaciones y comentarios que realmente mencionaban palabras y casos de corrupción:** 150

Con esto se obtiene una exactitud de **96.3%** en la extracción de entidades y palabras/casos de corrupción mencionados. Las expresiones regulares que tienen error son las que buscan entidades. Al indagar un poco más a fondo cuáles son las expresiones regulares que están generando errores, el grupo observa que son las del Partido Mira (porque comúnmente se refiere a este partido simplemente por Mira). Esta expresión regular disminuye la precisión debido a que los usuarios comentan seguido la palabra "mira" (del verbo mirar), por lo cual el grupo decide modificar la expresión regular para que busque "Partido Mira" en vez de "Mira". El resto de expresiones regulares se dejan como estaban inicialmente. Habiendo garantizado la calidad de las expresiones regulares generadas, se aplicaron los algoritmos descriptivos y funciones agregadas sobre la muestra realizada para verificar el correcto funcionamiento de los mismos.

5. Evaluación

En la reunión que se tuvo para presentar el sistema en funcionamiento, se pudieron derivar las siguientes conclusiones junto con la Secretaría de Transparencia:

- En la sección *Casos de Corrupción* de la visualización, se pudo analizar que los usuarios manifiestan su indignación en Facebook por medio de la reacción "Me divierte", indicando una postura sarcástica frente a los casos de corrupción que se monitorean.

- La reacción "Me gusta" no es muy diciente sobre la actitud de las personas hacia casos de corrupción. Es decir, los usuarios le dan "Me gusta" a una publicación más que cualquier otra reacción cuando se espera que generen más reacciones de "Me enfada" y "Me divierte".
- Algunos de los líderes de opinión monitoreados no tienen página en Facebook y por lo tanto no se puede obtener información más detallada sobre estos.
- Se observa un pico en la actividad de los líderes de opinión Álvaro Uribe y Juan Manuel Santos en el mes de noviembre de 2016, debido al proceso de paz.
- Para el caso de los partidos políticos, los comentarios negativos por lo general prevalecen sobre los comentarios positivos. Esto se puede contrastar con los resultados obtenidos en la encuesta LAPOP 2015, donde los partidos políticos cuentan el nivel de desconfianza más alto entre los encuestados.
- Desde un punto de vista general, se observa que la negatividad que los usuarios manifiestan es mayor a los comentarios o reacciones positivas. En la entrevista inicial con la Secretaría de Transparencia, Mauricio Ortiz tenía la hipótesis que en redes sociales hay mayor indignación *(Castañeda, Ortiz, & Pérez, 2017)*. A partir de los resultados obtenidos, todo parece indicar que hay un mayor nivel de indignación en Facebook respecto a líderes de opinión, partidos políticos e instituciones.
- Sentinel permite saber quiénes pueden ser aliados de la Secretaría de Transparencia para llevar a cabo investigaciones e iniciativas anticorrupción, priorizando estrategias dirigidas a ciertas áreas donde la percepción de corrupción de los usuarios sea más alta.

La Secretaría de Transparencia realizó un informe donde incluyen algunas conclusiones que pudieron sacar a partir de Sentinel y sugerencias para trabajos futuros. Para ver el informe, ir al *Capítulo 6 Resultados* de este documento.

6. Despliegue

Una vez obtenidos los resultados y validaciones en la sección 5 (Evaluación), se procede a desarrollar una estrategia de despliegue. Se listan y describen todas las actividades para ejecutar el proceso de instalación y configuración si se llegase a necesitar en otros proyectos o actividades en un futuro. A continuación, se listan los requisitos y componentes necesarios para llevar a cabo el despliegue del sistema.

Requerimientos

Dentro de estas actividades se mencionan los elementos y prerrequisitos del sistema en donde se especifican los requerimientos mínimos para ejecutar los módulos del sistema, se mencionan los siguientes:

1. Requerimientos mínimos del Sistema Operativo y sus componentes.

2. Código fuente ubicado en Github: https://github.com/sentinelanalytics donde reside todo el desarrollo del sistema, allí se encontrarán los siguientes módulos:
    a. **facebook-scraper-py**: componente de extracción de datos.
    b. **analytics**: componente de analítica.
    c. **sentinel**: componente de visualización.

Instalación

Para realizar la instalación del sistema se requieren de los siguientes componentes:

1. MongoDB
2. Python 3.5.2
3. PIP
4. R
5. Node.js y NPM

Con los componentes mencionados anteriormente, ya se puede proceder a la ejecución de sistema, para esto es necesario ubicar la ruta en donde está ubicado el script *setup.sh* y ejecutarlo.
El procedimiento para realizar el despliegue e instalación de la herramienta se puede ver con mayor detalle en el anexo I Sentinel – Manual de instalación.

Plan de monitoreo y mantenimiento

El primer aspecto de mantenimiento que debe tener en cuenta la Secretaría de Transparencia (o cualquier otra persona o institución que desee utilizar el sistema) es en cuanto a las palabras claves que se desean buscar. A medida que se van encontrando nuevas palabras, ya sean entidades, palabras de corrupción, casos de corrupción, etc.; estas se deben ir agregando a la base de conocimiento, con fines de enriquecer los términos y entidades para poder llevar a cabo un análisis más completo sobre la corrupción.

Los archivos que contienen la base de conocimiento, están situados en el código fuente del componente de analítica, específicamente en la carpeta "base-conocimiento".

Se presentan a continuación dos de los archivos de la base de conocimiento:

1. *casos-corrupcion.all.txt*

    Este archivo es un archivo separado por comas, y en él se encuentran las expresiones regulares para identificar publicaciones y comentarios que mencionan casos de corrupción.

    Este archivo cuenta con el siguiente esquema:

        i. Expresión regular para identificar el caso de corrupción.

ii. Valor booleano, indicado si se debe hacer un encuentro exacto de la expresión regular.

iii. Caso de corrupción.

iv. Nombre del caso de corrupción.

Se presenta a continuación un ejemplo del archivo:

Odebrecht,false,odebrecht,Odebrecht fiscal anticorrupci[óòo]n,false,fiscal-anticorrupcion,Luis Gustavo Moreno (fiscal anticorrupción) Luis Gustavo Moreno,false,fiscal-anticorrupcion,Luis Gustavo Moreno (fiscal anticorrupción) ..

Tabla 21 Ejemplo de archivo

2. *Instituciones.all.txt*

Este archivo tiene el mismo esquema que el de casos de corrupción. La única diferencia es que se definen expresiones regulares para identificar a instituciones en vez de casos de corrupción. Igualmente, se conservan las reglas de agrupar varias formas de referirse a una institución.

Por otro lado, el sistema cuenta con dos archivos: 1) *config.medios.json* y 2) *config.lideres.json*. Estos dos archivos contienen la lista de páginas de medios y líderes de opinión respectivamente, de las cuales se desea extraer publicaciones y comentarios. Estos archivos son modificables ya sea para agregar páginas o eliminar páginas del monitoreo que se desea realizar.

Por motivos de simplicidad y extensión del documento no se describen todos los archivos de la base de conocimiento, pero se recomienda revisarlos en el Anexo G Base de conocimiento - Github

Para mayor detalle acerca de las acciones que se pueden realizar en Sentinel, remitirse al Anexo E Sentinel – Manual de Usuario

CAPÍTULO 5: CONSTRUCCIÓN Y DISEÑO DE SOFTWARE SENTINEL

A continuación, se presenta todo el diseño de la solución propuesta, su especificación funcional, diseño y arquitectura del sistema, implementación y pruebas de software.

Áreas de la propuesta de solución: Analítica y Desarrollo de Software

Teniendo en cuenta el objetivo definido, se desarrolló un sistema que aplique técnicas de análisis y minería de datos sobre las publicaciones, comentarios y reacciones extraídas de las páginas de Facebook de diferentes medios y líderes de opinión, para enriquecer la información disponible para el Observatorio de Transparencia y Anticorrupción. La ilustración 22 presenta los grandes componentes del sistema y consta de tres componentes:

1. **Extracción de datos (RART):** este componente se encarga de extraer los datos de las páginas de Facebook y almacenarlas en una base de datos no relacional.
2. **Analítica:** este componente aplica técnicas y algoritmos de análisis y minería sobre los datos almacenados por el componente de extracción de datos. Los resultados de estas técnicas también se almacenan en la base de datos.
3. **Visualización:** toma los resultados del componente de analítica y los muestra por medio de una interfaz web para que puedan ser interpretados por los expertos del negocio.

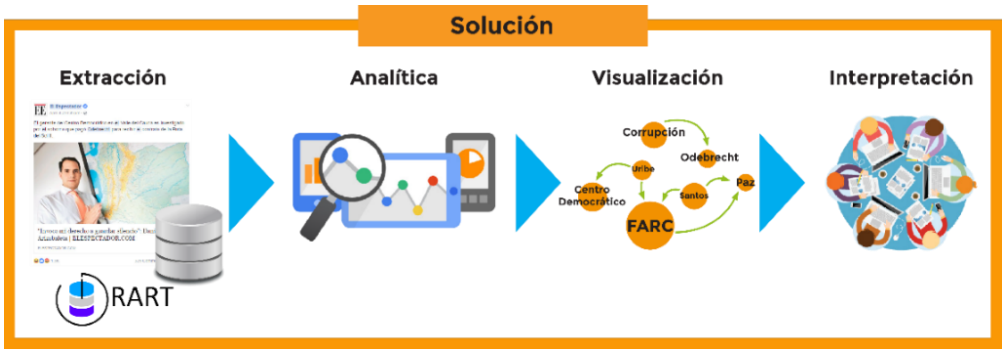

Ilustración 22 Solución propuesta

Como se mencionó en la sección 6 del C*apítulo 2: Descripción General*, es importante aclarar que la metodología usada para llevar a cabo el desarrollo del proyecto de minería y analítica de datos (CRISP-DM) no contempla desarrollo de software tradicional y debido a la naturaleza del mismo fue necesario emplear una metodología de desarrollo de software ágil para poder suplir las necesidades del sistema de información Sentinel, dicha metodología se llama SCRUM.

1. Especificación funcional

Una vez definidos los requerimientos del sistema se proceden a identificar los *user stories*, con ánimos de detallar la interacción del sistema con el usuario final.

ID	Como un…	quiero…	para…
US-001	Usuario	ver los sentimientos asociados que tienen los usuarios de Facebook hacia un líder de opinión o entidad	conocer la percepción de los usuarios.
US-002	Usuario	consultar cuáles son los temas que están más asociados a un líder de opinión.	saber cómo perciben al líder de opinión
US-003	Usuario	consultar las reacciones (*likes*, *happy*, *sad*, entre otros) sobre noticias de corrupción.	medir las opiniones de las personas sobre el tema
US-004	Usuario	consultar cuáles medios y líderes de opinión publican más/menos noticias sobre corrupción o líderes de opinión.	medir el sesgo de los medios
US-005	Usuario	consultar cuáles medios publican más noticias sobre un líder de opinión	medir la neutralidad de los medios
US-006	Usuario	consultar cuales son las entidades/políticos/partidos/personas más asociados a noticias de corrupción	tener una visión sobre los entes más corruptos
US-007	Usuario	consultar cuáles son los temas en que más se nombra la corrupción	identificar cómo los usuarios asocian la corrupción con distintas acciones o entidades.

US-008	Usuario	consultar las estadísticas tales como: número total de seguidores, cantidad de reacciones (*likes*, *happy*, *sad*, entre otros) de un líder de opinión.	identificar el apoyo de los usuarios a líderes de opinión.

Tabla 22 Definición de User Stories

Teniendo en cuanta los *user stories* del sistema, se procedió a identificar los requerimientos funcionales en la tabla 24:

ID	Requerimiento
RF-001	El sistema debe permitir extraer comentarios de Facebook.
RF-002	El sistema debe permitir extraer reacciones de Facebook.
RF-003	El sistema debe presentar por medio de gráficas los resultados obtenidos del componente de analítica.
RF-004	El sistema debe permitir analizar los sentimientos de los comentarios por medio de un algoritmo de análisis de sentimientos (*Bag of words*)
RF-005	El sistema debe permitir ver el sesgo de los medios por medio de un algoritmo que identifique datos atípicos.
RF-006	El sistema debe implementar algoritmos de asociación de palabras.
RF-007	El sistema debe implementar algoritmos de funciones agregadas, como suma, resta, valor mínimo, valor máximo y promedio.
RF-006	El sistema debe permitir al usuario seleccionar el caso de corrupción a monitorear.
RF-007	El sistema debe permitir al usuario seleccionar un líder de opinión a monitorear.
RF-008	El sistema debe permitir al usuario seleccionar una institución a monitorear.
RF-009	El sistema debe permitir al usuario suprimir los *datasets* mostrados en las gráficas.
RF-010	El sistema debe permitir ver los comentarios con más *likes* respecto a un líder de opinión.

Tabla 23 Requerimientos funcionales del sistema

2. Diseño

Con los requerimientos definidos y su priorización establecida, se procedió a diseñar la arquitectura de la solución bajo el lenguaje unificado de modelado UML.

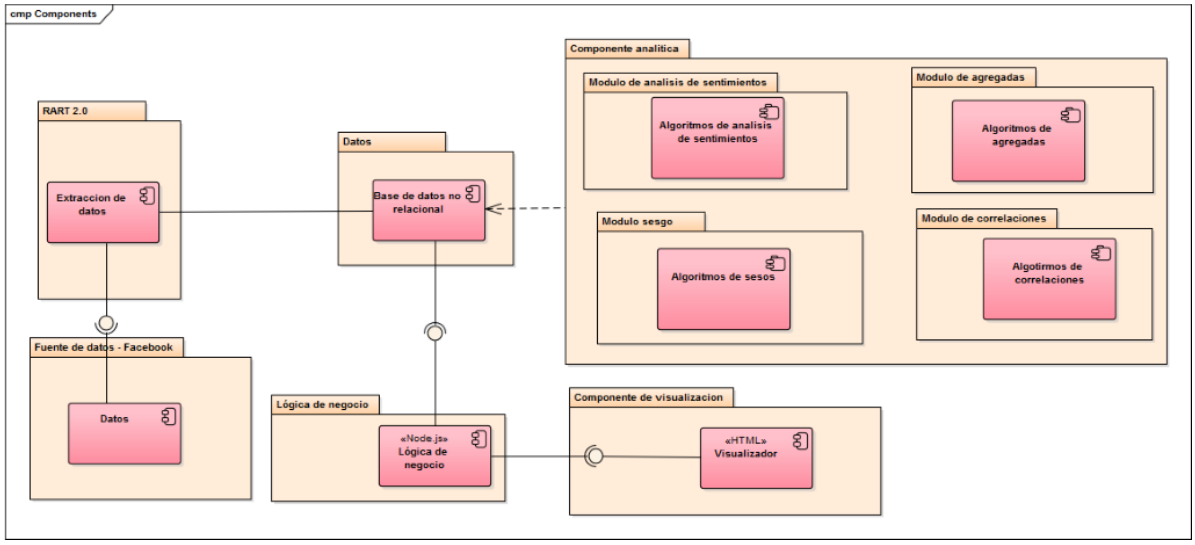

Ilustración 23 arquitectura del sistema

A continuación, se detallan las capas y sus componentes:

1. **Fuente de datos – Facebook**
    a. **Datos:** Representa la fuente principal de datos a extraer: el origen y forma inicial provendrá de la red social Facebook, el sistema accede constantemente a los datos abiertos (accesibles públicamente) de Facebook y por medio de HTTPS/REST se extraen los datos al sistema.
2. **RART 2.0**
    a. **Extracción de datos:** Es el componente encargado de extraer datos de Facebook, es un código desarrollado e implementado en Python, basado en RART, que permite por medio de métodos POST y GET extraer los datos de Facebook y los deja a disposición del sistema de manera local.
3. **Base de datos no relacional:** Debido a la forma y estructura de los datos, junto con la naturaleza del proyecto, los datos extraídos son en su mayoría no estructurados (ej: texto), por lo cual los datos deben ser tratados y almacenados en una base de datos no relacional. Además, los datos de Facebook ya están en formato JSON, por lo cual no se debe hacer ninguna conversión para almacenarlos.
4. **Componente de analítica**

Todos los módulos del componente de analítica fueron desarrollados e implementados en Python 3.5+ y R 3.1+

a. **Módulo de análisis de sentimientos**
    i. **Algoritmos de sentimientos:** es un algoritmo enfocado a realizar análisis de sentimientos sobre los comentarios extraídos.
b. **Módulo de sesgo**
    i. **Algoritmos de sesgos:** es un algoritmo enfocado y desarrollado a identificar valores atípicos en cuanto a cantidad de publicaciones de los datos extraídos.

    Bajo este contexto, el sesgo se mide por medio del rango intercuartil, en donde los valores atípicos son aquellos medios que publican menos o más sobre un caso de corrupción o alguna entidad.

c. **Módulo de agregadas**
    i. **Algoritmos de agregadas:** es una colección de algoritmos enfocados y desarrollados a realizar funciones básicas de agregados, tales como sumar, restar, contar, entre otros.
d. **Módulo de correlaciones**
    i. **Algoritmos de correlaciones:** es un algoritmo enfocados a identificar correlaciones sobre los datos extraídos.

5. **Componente de visualización**

El componente de visualización está definido bajo una arquitectura Cliente-Servidor, donde el servidor está implementado en Node.js y el cliente con las tecnologías HTML5, CSS3 y JavaScript. El servidor se encarga de buscar los resultados del proceso de analítica y enviarlos al cliente para su posterior visualización.

Diagrama de despliegue

Para el despliegue de la arquitectura se contarán con tres nodos principales como se muestra en la ilustración 24, donde se identifica el servidor de Sentinel, el dispositivo por el cual se visualizarán los resultados y Facebook como fuente principal de datos.

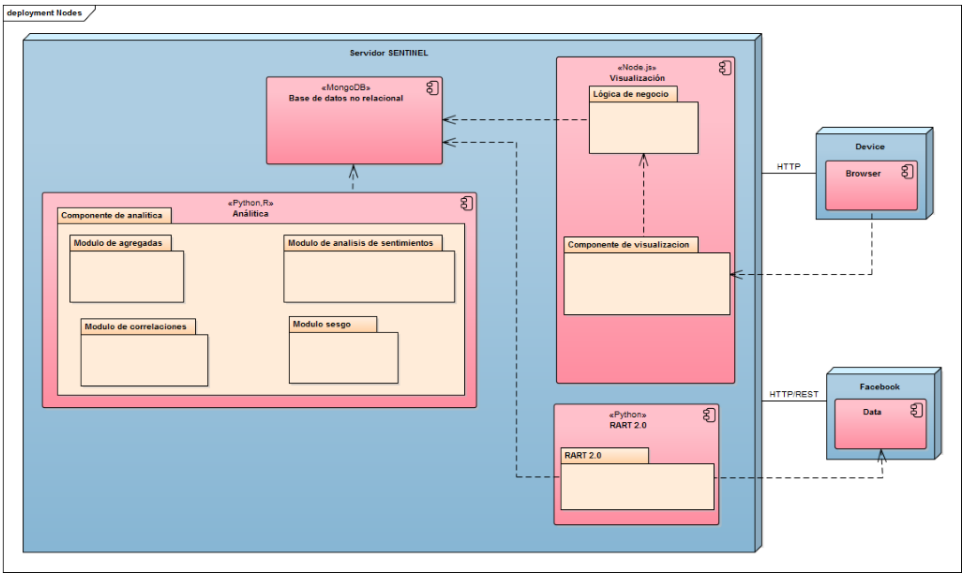

Ilustración 24 Diagrama de despliegue de la arquitectura

A continuación, se detallan los nodos:

1. **Servidor SENTINEL**: Este servidor almacena todos los componentes del sistema de información Sentinel.
    a. Base de datos no relacional.
    b. Analítica.
    c. RART 2.0.
    d. Visualización.
2. **Device**: representa el dispositivo que podrá visualizar los resultados finales de **Sentinel**, solo se necesita de un navegador web para desplegar la interfaz, ya sea desde un dispositivo móvil o un computador personal.
3. **Facebook:** Es la red social de donde se extraerán los datos.

3. Implementación

Para la visualización se utilizaron una serie de librerías web que proveen herramientas gráficas para poder visualizar los resultados obtenidos del componente de analítica. A continuación, se listan las librerías utilizadas:

a. **Vis.js**: Librería dinámica de visualización web (http://visjs.org/)
b. **Chartist.js**: Librería desarrollada en JavaScript que permite graficar datos en diferentes tipos de graficas en cavas HTML5 (https://gionkunz.github.io/chartist-js/)

c. **Chart.js**: Librería desarrollada en JavaScript que permite graficar datos en diferentes tipos de graficas en canvas HTML5 (http://www.chartjs.org/)
d. **Word_cloud**: Generador de nube de palabras desarrollado en Python. (https://github.com/amueller/word_cloud)

Para el desarrollo de todos los algoritmos del componente de analítica se utilizaron los siguientes lenguajes de desarrollo y paquetes:

a. Python 3.5.
    i. NLTK
    ii. NumPy
    iii. Unidecode
    iv. PyMongo
b. R Studio 3.3.3.
    i. TM
    ii. rmongodb

4. Pruebas

Pruebas Unitarias

Las precondiciones para obtener un correcto resultado en las pruebas fueron las siguientes:

1. El servidor host que contiene el componente de visualización esté encendido
2. La base de datos debe estar en ejecución.
3. Los resultados del componente de analítica deben estar almacenados en la base de datos.

Habiendo definido las precondiciones, se proceden a realizar las pruebas funcionales del sistema. Estas pruebas buscan garantizar el correcto funcionamiento del sistema. En este caso se realizan pruebas para garantizar que el aplicativo se esté ejecutando y además se prueba que los servicios web encargados de consultar los resultados de analítica estén respondiendo con los datos que en realidad deberían retornar.

Nota: la evaluación de los algoritmos de analítica se puede ver en la sección 4.4 Evaluación de Modelos.

En la tabla 25 se muestra la prueba de estado de las páginas, enviando métodos GET y POST para validar su respuesta y estado.

Estado HTTP	
Página	**Respuesta**
Página principal	*HTTP Response 200 OK*
Página líderes	*HTTP Response 200 OK*
Página instituciones	*HTTP Response 200 OK*
Página casos	*HTTP Response 200 OK*

Página medios	*HTTP Response 200 OK*
Página descubre	*HTTP Response 200 OK*
Página aviso	*HTTP Response 200 OK*

Tabla 24 Estado HTTP de páginas

En la tabla 26 se valida el estado y respuesta del algoritmo de sesgo

Sesgo		
Sesgo para...	**Debe retornar...**	**Resultado**
Corrupción	Year, month, entity=corrupcion	OK
Casos	Year, month, entity=casos	OK
Instituciones	Year, month, entitiy=instituciones	OK
Partidos	Year, month, entitiy=partidos	OK
Líderes	Year, month, entitiy=lideres	OK
Corrupción-instituciones	Year, month, entitiy=corrupcion-instituciones	OK
Corrupción-partidos	Year, month, entitiy=corrupcion-partidos	OK
Corrupción-líderes	Year, month, entitiy=corrupcion-lideres	OK
Líderes-corrupción	Year, month, entitiy=lideres	OK

Tabla 25 Estado y respuesta Web Service de sesgo

En la tabla 27 se valida el estado y respuesta del algoritmo de análisis de sentimientos

Análisis de sentimientos		
Análisis de sentimientos para...	**Debe retornar...**	**Resultado**
Líder de opinión		OK
Instituciones	Friendly_name=SIC, lider=sic, type=comments, entity=instituciones	OK
Partido político	Friendly_name=Centro Democrático, lider=centro-democratico, type=comments, entity=partidos	OK

Tabla 26 Estado Web Service de análisis de sentimientos

En la tabla 28 se valida el estado y respuesta del algoritmo de asociación de palabras

Asociación de palabras		
Asociación de palabras para...	**Debe retornar...**	**Resultado**
Comentarios	Year, month, nodes, edges, type=comments	OK
Publicaciones	Year, month, nodes, edges, type=posts	OK

Tabla 27 Estado y respuesta Web Service de asociación de palabras

En la tabla 29 se valida el estado y respuesta de los algoritmos de funciones agregadas.

Agregadas		
Agregadas para...	**Debe retornar...**	**Resultado**
Corrupción	Year, month, entity=corrupcion, type=activity_count	OK
Casos	Year, month, entity=casos, type=activity_count	OK
Instituciones	Year, month, entitiy=instituciones, type=activity_count	OK
Partidos	Year, month, entitiy=partidos, type=activity_count	OK
Líderes	Year, month, entitiy=lideres, type=activity_count	OK
Corrupción-instituciones	Year, month, entitiy=corrupcion-instituciones, type=post_count	OK
Corrupción-partidos	Year, month, entitiy=corrupcion-partidos, type=post_count	OK
Corrupción-líderes	Year, month, entitiy=corrupcion-lideres, type=post_count	OK
Líderes-corrupción	Year, month, entitiy=lideres, type=post_count	OK

Tabla 28 Estado y respuesta Web Service de funciones agregadas.

Los resultados mostrados anteriormente resumen las pruebas de las funcionalidades del sistema. Para ver de manera detallada las pruebas realizadas ver anexo J Sentinel – Pruebas del sistema

Encuesta TAM

Para validar el sistema con el usuario final, el grupo de trabajo desarrolló una serie de preguntas bajo los lineamientos de validación TAM *(Venkatesh & Hillol, 2008)* en donde el principal propósito es realizar una serie de preguntas en donde se validan cuatro ejes principales:

1. **Variables externas (Involucramiento):** Qué tanto el usuario final estuvo involucrado en el desarrollo del sistema
2. **Utilidad percibida:** Qué nivel de utilidad encuentra el usuario final frente al sistema.
3. **Percepción de facilidad de uso:** Qué tan fácil fue para el usuario final la interacción con el sistema.
4. **Actitud hacia el uso:** Qué tan dispuesto está el usuario final al momento de usar el sistema.

Previamente a la realización de la encuesta, a los usuarios finales se les capacitó y mostró el correcto funcionamiento y uso del sistema desarrollado. La encuesta realizada tiene la siguiente escala de clasificación:

Valor	Descripción
1	Totalmente en desacuerdo
2	Moderadamente en desacuerdo
3	Neutral
4	Moderadamente de acuerdo
5	Totalmente de acuerdo

Tabla 29 Encuesta escala de clasificación

Para ver la encuesta TAM realizada ver el Anexo D Encuesta TAM.

A continuación, se presentan los resultados de la encuesta:

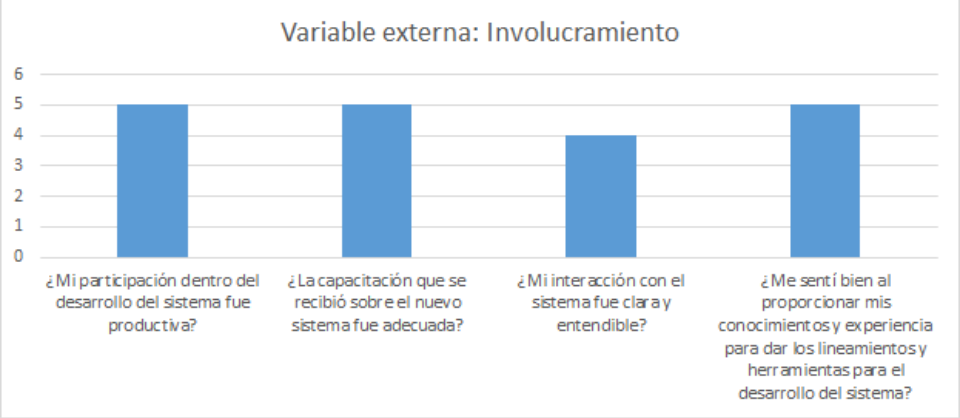

Ilustración 25 Variables externas: Involucrados

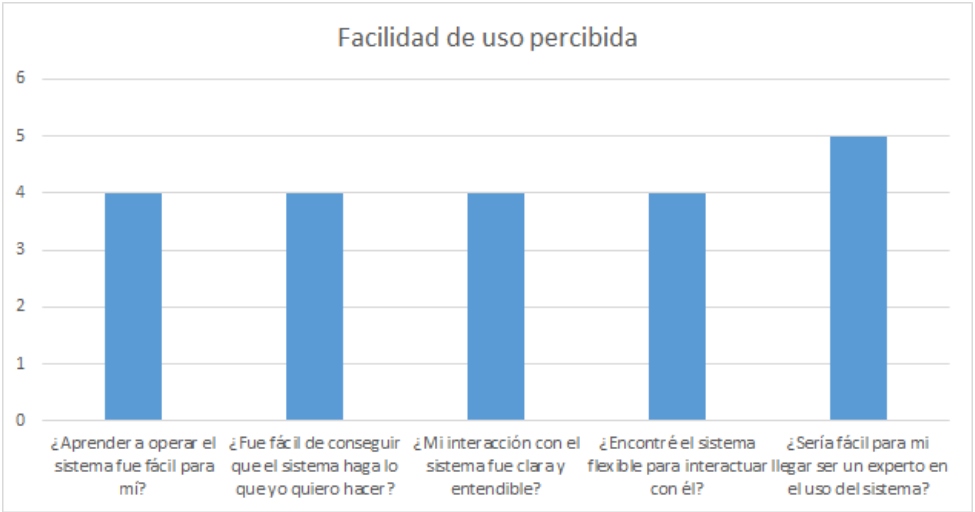

Ilustración 26 Facilidad de uso percibida

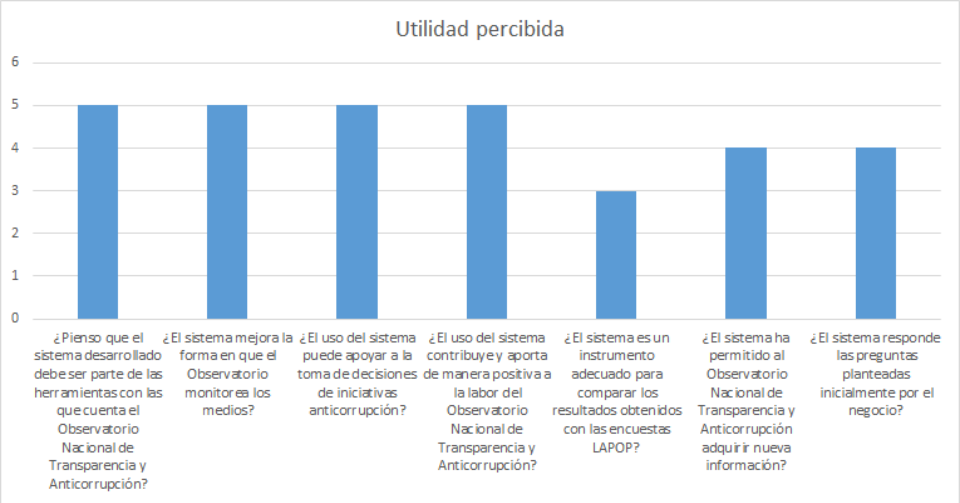

Ilustración 27 Utilidad percibida

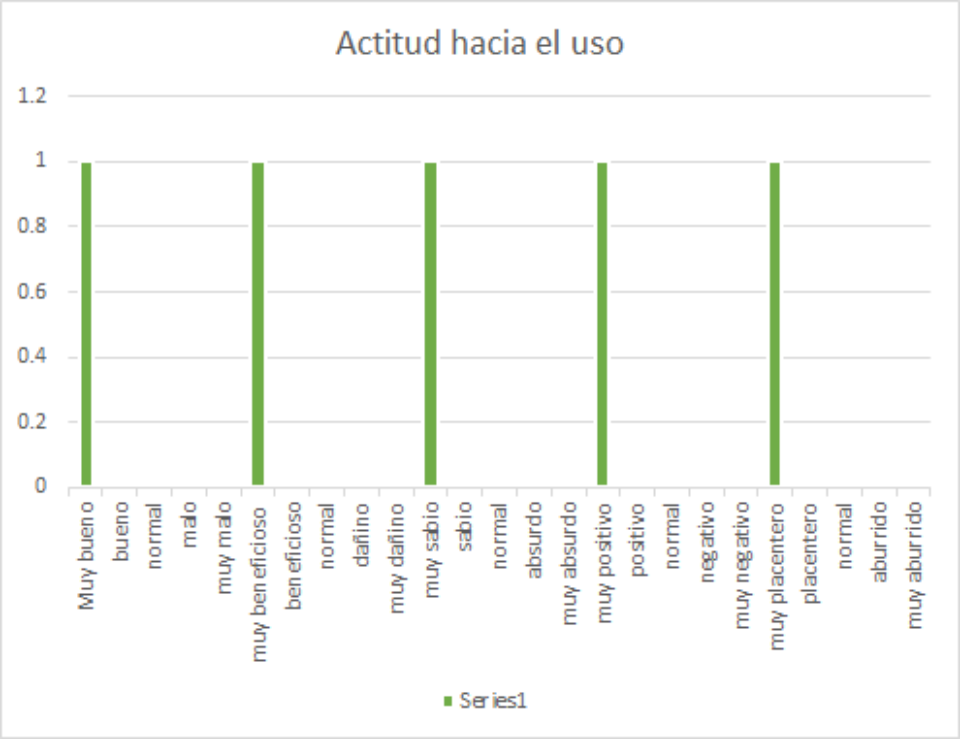

Ilustración 28 Actitud hacia el uso

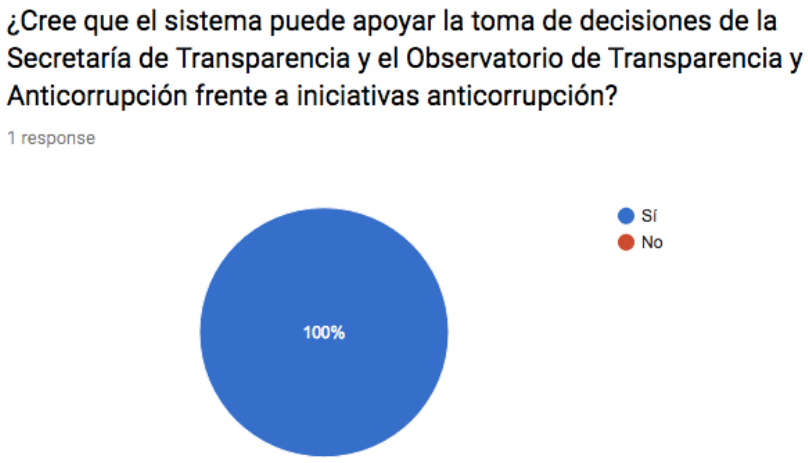

Ilustración 29 Nivel satisfacción Sentinel

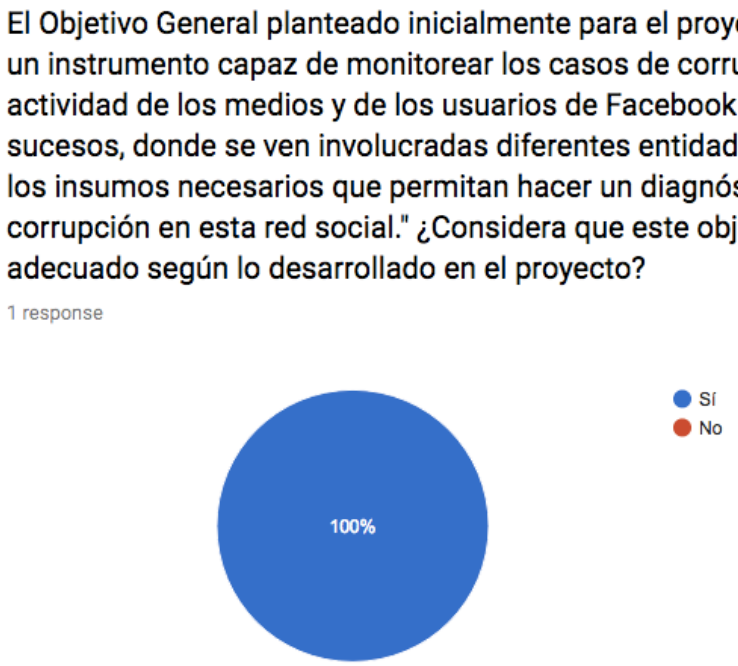

Ilustración 30 satisfacción objetivo general

Como se puede observar en los resultados de las encuestas TAM se puede concluir los siguiente:

1. En términos generales, el usuario encuentra el sistema fácil de usar y navegar; se desplaza con comodidad y fluidez por la interfaz del sistema, sin embargo, hay espacio para la mejora y el arreglo de ciertas secciones de la interfaz como lo es por el ejemplo el grafo que visualiza la asociación de palabras.
2. El usuario resalta la utilidad y beneficio de usar el sistema en las labores diarias, y además considera que el sistema puede ayudar en iniciativas anticorrupción.
3. El involucramiento del usuario en el desarrollo del sistema fue productivo y constante.

4. El usuario afirma que el sistema cumple con el objetivo general planteado al incido de la formulación del trabajo de investigación.

CAPÍTULO 6: RESULTADOS

Cumplimiento del Objetivo General

Se definió como objetivo general: "Desarrollar un modelo de analítica que permita enriquecer la información disponible para el Observatorio de Transparencia y Anticorrupción de Colombia utilizando como fuente de datos la red social Facebook abierta."

Se desarrolló el sistema de información Sentinel y se implementó un modelo de analítica que por medio del entendimiento de la corrupción en Colombia describe y modela de manera adecuada los factores y variables que identifican el comportamiento de la percepción de la corrupción, basándose en su principal fuente de datos la cual proviene de la red social Facebook. En efecto, dicho modelo de analítica fue validado por el experto del negocio, Secretaría de Transparencia, y permite enriquecer la información puesto que ayuda a diagnosticar y analizar los indicadores de corrupción en la ciudadanía. Dicho lo anterior, se puede afirmar que el objetivo general se cumple satisfactoriamente.

Cumplimiento de Objetivos Específicos

Objetivo	Resultado
Definir el modelo de analítica de corrupción a partir de publicaciones en *fan pages* de medios de comunicación, empresas, personajes públicos y partidos políticos.	Se desarrolló un modelo de analítica el cual describe las relaciones de los diferentes conceptos que involucran a la corrupción y los componentes de Facebook. Ver Modelo de analítica.
Diseñar e implementar un componente web que presente los resultados mediante gráficas de manera clara y concisa.	Se diseñó e implementó un componente web que presenta los resultados del componente de analítica de manera visual con el uso de gráficas. La usabilidad del sistema fue evaluada en la encuesta TAM. Ver TAM. Ver Anexo E Sentinel – Manual de Usuario.
Validar la utilidad y exactitud de los modelos generados en la fase de modelado junto con el experto del negocio	Una vez generados los modelos se procedió a validar la utilidad y exactitud de los mismos. El experto de negocio validó dichos modelos y destacó su utilidad, y el grupo de trabajo se encargó de validar la exactitud de los modelos. Ver Anexo H Conclusiones y recomendaciones herramienta SENTINEL – Observatorio Nacional de Transparencia y Anticorrupción. Ver evaluación de modelos.

Tabla 30 Cumplimiento de Objetivos Específicos

Conclusiones y recomendaciones herramienta SENTINEL

La Secretaría de Trasparencia realizó un informe donde expresan algunas recomendaciones y conclusiones sobre Sentinel, además expresan que existe un potencial enorme en la iniciativa y que así mismo hay amplio espacio para agregarle nuevas funcionalidades. Dicho informe se muestra a continuación:

Actualmente, la Secretaría enfoca sus esfuerzos en analizar los indicadores de percepción, única aproximación existente a la medición del fenómeno de la corrupción. El indicador más conocido es el calculado por Transparencia Internacional, en el que Colombia está en el puesto 90 de 176 países partícipes de la medición. Cabe destacar que la percepción es una medición subjetiva, y es necesario analizar los factores que la afectan, sus manifestaciones y si es un medio efectivo de medir la corrupción en un país.

Sentinel podría ser entonces una herramienta útil para observar, desde las redes sociales, cómo los medios de comunicación, los líderes de opinión y las instituciones, pueden alterar a través de sus publicaciones la percepción de corrupción en los ciudadanos y generar reacciones de apoyo o indignación.

Este mapeo de los factores que más alteran la percepción y la participación de los ciudadanos en las discusiones relacionadas con corrupción, nos permite entrever los focos a los que debe dirigirse la formulación de política pública, conforme a las necesidades y demandas de la ciudadanía. A su vez, se genera un desafío importante para lograr que las noticias no procuren solamente un aumento en la percepción de corrupción, sino que incrementen las acciones concretas de vigilancia y control social.

A continuación, esbozamos algunas conclusiones y recomendaciones sobre las secciones del aplicativo

1. **Descubre:**

Esta sesión nos fue útil para conocer a profundidad el discurso que se genera en redes sociales alrededor del fenómeno de la corrupción. Esta sección del aplicativo podría mejorar si:

- Se filtran aquellas palabras que no son relevantes para el análisis.
- El grafo se despliega más claramente para una mejor interpretación de las relaciones entre palabras y fuentes.
- Es posible filtrar por categoría para verlas de forma aislada; *e.g.* si se puede tomar el nodo de líderes de opinión y filtrar por líder cuáles son las palabras más asociadas a cada uno. Si bien ya es posible entrever algunas de estas cosas en el grafo general, limitar la visión a determinadas características puede hacer que su interpretación sea más ágil.
- Se pueden ver estadísticas sobre la cantidad de veces que se usa una determinada palabra por cada publicación; *e.g.* si al colocar el puntero sobre una de las palabras (círculos azul claro) se puede ver el número de menciones.

2. **Medios:**

Esta sección nos permitió reconocer la visibilidad por caso en medios regionales y nacionales, y sacar datos relevantes sobre la permanencia de una noticia de corrupción luego de que es publicado el caso por primera vez. Esta sección del aplicativo podría mejorar si:

- Es posible obtener estadísticas por año o por mes.
- Se puede identificar cuál es la proporción de las publicaciones que están directamente relacionadas con casos de corrupción frente al total de publicaciones que genera un determinado medio, líder de opinión, u otros; *e.g.* "de *X* número de noticias generales que publica *Y* medio, *Z* noticias se asocian a temas de corrupción".
- Se puede identificar cuál es la proporción de las publicaciones sobre un caso de corrupción en particular frente al total de publicaciones de actos de corrupción que genera un determinado medio, líder de opinión, u otros; *e.g.* "de *X* número de noticias sobre casos de corrupción en general que publica *Y* medio, *Z* noticias son sobre un determinado caso en particular".

3. **Casos de corrupción:**

Esta sección nos permitió concluir que efectivamente existe, en la mayoría de los casos, una relación directamente proporcional entre las publicaciones relacionadas con corrupción y los comentarios de los individuos en redes sociales. Cabe destacar que esto confirma la hipótesis de que la actividad ciudadana en redes tiende a ser exclusivamente reaccionaria; esto es, que difícilmente se sigue comentando sobre un determinado caso de corrupción si se dejan de publicar noticias sobre el mismo en páginas de medios, líderes de opinión, u otros (entre otros). Esta sección del aplicativo podría mejorar si:

- Se genera una gráfica que compare los datos de varios casos de corrupción a la vez; esto es, una gráfica donde se pueda observar cuál caso de corrupción fue el que tuvo más publicaciones y cómo se compara con los demás (lo mismo se puede hacer en materia de comentarios).

4. **Instituciones:**

Esta sección provee información valiosa para realizar estudios comparados con otras encuestas tales como el Barómetro de las Américas. Es aconsejable, sin embargo, mejorar la forma en que se presenta la información; esto, como en otros casos anteriores, se puede lograr si:

- Se genera una gráfica que compare los datos de varias instituciones a la vez; esto es, una gráfica donde se pueda observar cuál institución tuvo el mayor número de publicaciones sobre corrupción y cómo se compara con los demás (lo mismo se puede hacer en materia de comentarios).

5. **Partidos políticos:**

Esta sección permite a la Secretaría hacer un diagnóstico sobre cuáles son las fuerzas políticas que mayor actividad tienen en términos de publicaciones y comentarios acerca del fenómeno de la corrupción. Para proyectos posteriores, esta información podrá servir para analizar si una mayor o menor actividad en esta materia se correlaciona con un mayor número de iniciativas o proyectos presentados para luchar contra la corrupción, o si de hecho se relaciona con un mayor o menor número de investigaciones sobre sus miembros sobre casos de corrupción. Tal como en otras secciones, ese aspecto se puede mejorar si:

- Se genera una gráfica que compare los datos de varios partidos políticos a la vez; esto es, una gráfica donde se pueda observar cuál partido político tuvo el mayor número de publicaciones sobre corrupción y cómo se compara con los demás (lo mismo se puede hacer en materia de comentarios).

6. **Líderes de opinión:**

En este apartado pudimos observar la importancia del tema de la corrupción para los líderes de opinión, y el apoyo o reprobación de los ciudadanos a sus posturas. Esta sección del aplicativo podría mejorar si:

- Es posible realizar cruces entre las publicaciones y comentarios de los líderes de opinión con el resto de categorías.

Las conclusiones hechas por la Secretaría de Transparencia demuestran el potencial de Sentinel. Sin embargo, en cuanto a las sugerencias realizadas, debido al alcance establecido inicialmente, los recursos disponibles y las restricciones de ciertas herramientas, no se pudo llevar a cabo la ejecución de todas las sugerencias hechas. En la reunión de validación del sistema, la Secretaría de Transparencia expresó la necesidad de saber qué publicaciones y comentarios se estaban haciendo sobre las diferentes entidades y casos de corrupción a través del tiempo (y no solamente en el último mes como se hacía inicialmente), por lo cual el grupo de trabajo implementó esta sugerencia. En cuanto a la filtración de palabras en el grafo, se creó un archivo de texto con *stop words* para que la Secretaría pueda filtrar las palabras que no son relevantes para el análisis que ellos realizan.

CAPÍTULO 7: CONCLUSIONES

Desde el inicio del proyecto, la pregunta generadora se planteó de la siguiente manera: ¿Cómo aprovechar los datos obtenidos en Facebook para generar conocimiento útil sobre la situación actual de corrupción en Colombia? Después de culminar el desarrollo del trabajo de investigación y teniendo en cuenta los comentarios hechos por el Observatorio, se puede afirmar que los datos extraídos de la red social Facebook toman un papel crucial para la lucha contra la corrupción en Colombia puesto que, por medio del sistema desarrollado y la aplicación de algoritmos de análisis y minería de datos, fue posible encontrar información que permite analizar y visualizar el fenómeno de la corrupción, apoyando de esta manera la labor de entidades gubernamentales que velan por la transparencia y luchan contra la corrupción a nivel nacional

1. Conclusiones

Este proyecto permitió darle una aproximación diferente al análisis de opiniones y percepción en redes sociales sobre corrupción y política en general (líderes de opinión, instituciones, partidos políticos). Existen varios trabajos que se han enfocado en el tema de minería de opiniones y política en Twitter *(Pak & Paroubek, 2010) (Tumasjan, Sprenger, Sandner, & Welpe, 2010) (Hridoy, Ekram, Islam, Ahmed, & Rahman, 2015) (Younus et al., 2011) (Razzaq, Qamar, & Bilal, 2014)*, por lo cual Sentinel es una oportunidad para explotar la información disponible en Facebook, teniendo en cuenta que es la red social más utilizada en Colombia. Sentinel es un ejemplo de cómo se puede aportar, desde la ingeniería de sistemas, al desarrollo y bienestar del país y que, gracias a este sistema y la guía de los expertos de negocio, se puede descubrir nueva información que puede apoyar iniciativas anticorrupción. Es importante resaltar que, Sentinel no es un sistema que automáticamente saca conclusiones por si mismo, sino que requiere de

interpretación por parte del usuario final, preferiblemente de un experto del negocio para poder sacar conclusiones con un fundamento sólido.

Por otro lado, a partir del desarrollo de Sentinel se puede concluir que es viable extraer grandes cantidades de datos de Facebook para llevar a cabo una serie de análisis del impacto que tiene esta red social en el ámbito político en Colombia. Esto es posible gracias a la apertura que tienen los medios de comunicación en las redes sociales y la alta participación de los usuarios. El uso de APIs públicas como la de Facebook, proveen todas las herramientas necesarias para la extracción rápida de datos de calidad y gracias a este tipo de herramientas se pueden realizar proyectos como Sentinel. Hay que tener en cuenta que extraer un gran volumen de datos en un tiempo corto no significa que también se puedan analizar rápidamente, sobre todo si la mayoría de datos extraídos son texto, como lo es el caso de Sentinel.

La alta dimensionalidad del texto trae consigo sus propios retos, como la cantidad de espacio que ocupa en memoria y preprocesar los datos de tal manera que permitan realizar un análisis adecuado del texto. A esto se le suma el hecho de que lo que nosotros como seres humanos, consideramos de sentido común y fácilmente identificable (por ejemplo, el sarcasmo), no resulta tan fácil de identificar para algoritmos de análisis de sentimientos como bag-of-words. Esto fue un caso recurrente al analizar los comentarios hechos por los usuarios. Existe una variedad de herramientas de análisis de sentimientos que pueden llegar a tener un buen rendimiento, como, por ejemplo: Google Cloud Natural Language API. Sin embargo, el grupo de trabajo se encontró con el problema de que no todas las herramientas están disponibles en español y que, en términos generales, fue difícil hallar recursos para realizar análisis de sentimientos en español.

Una de las grandes ventajas de Sentinel es que los insumos utilizados para llevar acabo el análisis y minería de datos no son solo las noticias hechas por los medios sino también los comentarios realizados por los usuarios. Esto permite hacer un monitoreo más detallado de la percepción de corrupción y demás entidades, en vez de ceñirse a lo que publican los medios. La constante participación de los usuarios sobre publicaciones de casos de corrupción genera una gran variedad de opiniones que permiten encontrar información y resultados que no se acostumbran a ver con encuestas tradicionales. Por esta razón, Sentinel permite entregar a las entidades gubernamentales, como lo es el Observatorio, información que refleja el punto de vista y comportamiento de la ciudadanía y esto genera gran impacto, pues las nuevas iniciativas o políticas públicas serán planteadas no solo desde el punto de vista político-administrativo sino también tendrán en consideración la opinión del ciudadano.

2. Análisis de impacto del proyecto

Con el desarrollo e implementación del sistema de información Sentinel, se presentan dos impactos importantes a nivel social y tecnológico:

Impacto social

Dentro de los deberes establecidos por el gobierno colombiano, se establece que la ciudadanía debe participar de manera activa y tener una constante colaboración para garantizar el buen funcionamiento de la administración y justica de la nación (Georgetown, 2017). Por esta razón, el trabajo de investigación propone una herramienta capaz de monitorear la percepción de corrupción de la ciudadanía en un ambiente virtual como lo son las redes sociales, con el objetivo de diagnosticar y generar indicadores de percepción de la corrupción. De esta manera, el sistema desarrollado no solo aporta a la lucha contra la corrupción, sino que además propone vías de involucramiento académico con el apoyo de las tecnologías de información. Entonces, Sentinel se postula como un agente de cambio a nivel social y tecnológico que vela y aporta a los objetivos de las entidades gubernamentales que tienen como principal motivación la construcción de un mejor país para futuras generaciones.

Impacto tecnológico

Con el desarrollo del sistema de información Sentinel, se muestra cómo es posible utilizar análisis y técnicas de minería de datos para afrontar problemas de carácter político-social. Se busca generar interés para continuar desarrollando proyectos y trabajos de investigación cuya problemática pueda ser abordada desde la perspectiva de minería de datos. Adicionalmente, se busca construir canales de apoyo mutuo entre la disciplina de la ingeniería de sistemas y las entidades propias del Estado y de esta manera trabajar juntos por un mejor país.

3. Trabajo futuro

En el informe realizado por la Secretaría de Transparencia, se resaltan algunas sugerencias para Sentinel, que quedan como trabajo futuro. Además de proponer nuevas funcionalidades en el informe, en la reunión de validación del sistema que se tuvo con la Secretaría de Transparencia, se propusieron las siguientes funcionalidades:

- ✓ Poder generar reportes con la herramienta.
- ✓ Se propone crear una ontología de corrupción, con fines de tener una base de conocimiento más robusta y tener mayor exhaustividad en la búsqueda de entidades y palabras o casos relacionados a corrupción.
- ✓ Mejorar el módulo de análisis de sentimientos. El grupo de trabajo inicialmente pretendía utilizar una solución comercial (*IBM Watson Tone Analyzer/Google Cloud Natural Language API*), pero no contó con los recursos monetarios necesarios. El grupo considera que el tema de análisis de sentimientos es algo complicado debido a que los mismos integrantes del grupo tuvieron problemas para determinar la polaridad de los comentarios. Como trabajo futuro, se propone probar con

diferentes algoritmos y técnicas (ej: redes neuronales, Naive Bayes, etc.), para medir cuál se ajusta mejor para este proyecto.

- ✓ Permitir realizar denuncias de casos de corrupción dentro de Sentinel, y aplicar procesos de analítica a las denuncias en sí para conocer cuáles entidades o cuáles son los temas de corrupción más mencionados en denuncias de casos de corrupción.
- ✓ El Observatorio reconoció la importancia del sistema desarrollado, pero identificó que sería de gran utilidad realizar el mismo monitoreo sobre la red social Twitter con ánimos de obtener conclusiones que abarquen todo el espectro de las redes sociales en la población de país. Para ver detalles sobre cómo se podría realizar la integración de esta red social, ver Anexo I Sentinel - Manual de instalación.
- ✓ Finalmente, se propone contactar a otros observatorios tanto de transparencia y política (ej: Transparencia por Colombia, Observatorio de la Democracia de la Universidad de los Andes) como de medios (ej: Observatorio de Medios de la Universidad de la Sabana y de la Universidad Sergio Arboleda). Estos pueden ser grandes aliados para dar lineamientos y validar la utilidad del sistema desarrollado. Con estos aliados se pueden derivar nuevas conclusiones y se pueden obtener más recomendaciones para mejorar el sistema.

REFERENCIAS

1. Transparency, I. (2017). Corruption Perceptions Index 2016. Recuperado el 10 de noviembre de 2017, a partir de https://www.transparency.org/news/feature/corruption_perceptions_index_2016
2. Observatorio de Transparencia y Anticorrupción. (2017, octubre 1). Qué es el Observatorio. Recopilado Mayo 7, 2017, a partir de http://www.anticorrupcion.gov.co/Paginas/Qu%C3%A9-es-el-Observatorio.aspx
3. Alianza Caoba. (2016). ¿Qué es CAOBA? Recuperado el 15 de abril de 2017, a partir de http://alianzacaoba.co/que-es-caoba/
4. Calambás Marin, D. A., & Mendoza Mendoza, J. A. (2016). RART. Recopilado Marzo 29, 2017, a partir de http://pegasus.javeriana.edu.co/~CIS1630AP06/entregables.html
5. Bandari, R., Asur, S., & Huberman, B. A. (2012). The Pulse of News in Social Media: Forecasting Popularity. Recuperado a partir de https://arxiv-org.ezproxy.javeriana.edu.co/abs/1202.0332
6. Castañeda, P., Ortiz, M., & Pérez, S. (2017). *Entrevista con El Observatorio Nacional de Transparencia y Anticorrupción*. Observatorio Nacional de Transparencia y Anticorrupción
7. Garcia, M., Montalvo, J., & Seligson, M. (2015, junio). Cultura política de la democracia en Colombia, 2015: Actitudes democráticas en zonas de consolidación territorial. Vanderbilt University.
8. Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0
9. Alexander, B., & Levine, A. (2006). Web 2.0. *A New Wave of Innovation for Teachning and Learning*, 32–44.
10. Pinquart, M., & Sörensen, S. (2000). Influences of socioeconomic status, social network, and competence on subjective well-being in later life: A meta-analysis.
11. Domínguez, D. C. (2009). Democracy 2.0: politics inside social networks. *Pensar la Publicidad*, *3*(2), 31–48.
12. Casa Editorial El Tiempo. (n.d.). Especial: ¿Qué tan digital eres? - Especial. Recopilado Marzo 19, 2017, a partir de http://www.eltiempo.com/multimedia/especiales/estadisticas-del-uso-de-internet-en-colombia/16758954/1
13. MinTic. (2017). Cifras - Ministerio de Tecnologías de la Información y las Comunicaciones. Recuperado el 19 de marzo de 2017, a partir de http://www.mintic.gov.co/portal/604/w3-article-4425.html
14. Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining Text Data* (2012 edition). Springer.
15. Aggarwal, C. C. (Ed.). (2011). *Social Network Data Analytics* (2011 edition). Springer.

16. Aggarwal, C. C. (2015). *Data Mining: The Textbook* (2015 edition). New York, NY: Springer.
17. Aggarwal, C. C. (2011). An Introduction to Social Network Data Analytics. En C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 1–15). Springer US. https://doi.org/10.1007/978-1-4419-8462-3_1
18. Matheus, R., & Ribeiro, M. M. (2009). Online anti-corruption tools in Latin America. En *Proceedings of the 3rd international conference on Theory and practice of electronic governance* (pp. 381–382). ACM.
19. Presidencia de la República. (2016). Colombia contra la corrupción. Recuperado el 8 de noviembre de 2017, a partir de http://especiales.presidencia.gov.co/Documents/20160511-colombia-contra-corrupcion/colombia-contra-la-corrupcion.html
20. Observatorio de Transparencia y Anticorrupción. (2017, octubre1). Instituciones para la Lucha Contra la Corrupción. Recuperado el 1 de octubre de 2017, a partir de http://www.anticorrupcion.gov.co/Paginas/Instituciones.aspx
21. Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2012). Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation. *Journal of Computer-Mediated Communication*, *17*(3), 319–336. https://doi.org/10.1111/j.1083-6101.2012.01574.x
22. Venkatesh, V., & Hillol, B. (2008). Technology Acceptance Model 3 and Research Agenda on Intervations. Decision sciences, 39(2), 273-315
23. Georgetown. (2017, December 11). Deberes del ciudadano. Recuperado el 12 de NobiembreNovember 12, 2017, a partir de http://pdba.georgetown.edu/Comp/Derechos/deberes.html
24. IBM. (2017, diciembre 17). Tone Analyzer | IBM Watson Developer Cloud. Recuperado el 28 de marzo de 2017, a partir de https://www.ibm.com/watson/developercloud/tone-analyzer.html
25. Moreno-Sandoval, L. G., Beltrán-Herrera, P., Vargas-Cruz, J. A., Sánchez-Barriga, C., Pomares-Quimbaya, A., Alvarado-Valencia, J. A., & García-Díaz, J. C. (2017). CSL: A Combined Spanish Lexicon - Resource for Polarity Classification and Sentiment Analysis (pp. 288–295). Presentado en 19th International Conference on Enterprise Information Systems. Recuperado a partir de http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=J41cKicqYUA=&t=1
26. SCRUM: What is Scrum? (2017, julio 11). Recuperado el 11 de mayo de 2017, a partir de http://www.scrum.org/resources/what-is-scrum
27. Chapman, P., & Clinton, J. (2000). CRISP - DM 1.0. Step-by-step data mining guide. SPSS.
28. R: What is R? (2017). Recuperado el 11 de mayo de 2017, a partir de https://www.r-project.org/about.html

29. Goodger, D. (2013, febrero 21). Stemmers. Recuperado el 15 de noviembre de 2017, a partir de http://www.nltk.org/howto/stem.html
30. Cloud Natural Language API. (2017). Cloud Natural Language API | Google Cloud Platform. Recuperado el 15 de noviembre de 2017, a partir de https://cloud.google.com/natural-language/
31. Caicedo, L., Carrillo, A., Forero, J., & Uruena, J. (2017, octubre 9). Análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes sociales. Recuperado el 9 de octubre de 2017, a partir de https://repository.javeriana.edu.co/handle/10554/20516
32. Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Presentado en Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA). Recuperado a partir de http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf
33. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. En *Fourth International AAAI Conference on Weblogs and Social Media*. Recuperado a partir de http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441
34. Hridoy, S. A. A., Ekram, M. T., Islam, M. S., Ahmed, F., & Rahman, R. M. (2015). Localized twitter opinion mining using sentiment analysis. *Decision Analytics*, *2*(1), 8. https://doi.org/10.1186/s40165-015-0016-4
35. Younus, A., Qureshi, M. A., Asar, F. F., Azam, M., Saeed, M., & Touheed, N. (2011). What Do the Average Twitterers Say: A Twitter Model for Public Opinion Analysis in the Face of Major Political Events. En *2011 International Conference on Advances in Social Networks Analysis and Mining* (pp. 618–623). https://doi.org/10.1109/ASONAM.2011.85
36. Razzaq, M. A., Qamar, A. M., & Bilal, H. S. M. (2014). Prediction and analysis of Pakistan election 2013 based on sentiment analysis. En *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)* (pp. 700–703). https://doi.org/10.1109/ASONAM.2014.6921662

ANEXOS

a. Sentinel - CRISP-DM
b. Sentinel - SRS
c. Sentinel - SDD
d. Encuesta TAM
e. Sentinel – Manual de Usuario
f. Desarrollo - Github
g. Base de conocimiento - Github
h. Conclusiones y recomendaciones herramienta SENTINEL – Observatorio Nacional de Transparencia y Anticorrupción
i. Sentinel – Manual de Instalación
j. Sentinel – Pruebas del sistema