Inteligencia de Fuentes Abiertas aplicadas al contexto colombiano

Intelligence of Open Sources applied to the Colombian context

Integrantes:
Ricardo Andrés Pinto Rico
Martín José Hernández Medina
Cristian Camilo Pinzón Hernández

Director:
Daniel Orlando Díaz López

Escuela Colombiana De Ingeniería Julio Garavito
Ingeniería De Sistemas
Proyecto De Grado 1
Bogotá
2018

Glosario

Monitoreo: El término monitoreo podría definirse como la acción y efecto de monitorear. Pero otra posible acepción se utilizaría para describir a un proceso mediante el cual se reúne, observa, estudia y emplea información para luego poder realizar un seguimiento de un programa o hecho particular.

Datos: El dato es una representación simbólica de un atributo o variable cuantitativa. Los datos describen hechos empíricos, sucesos y entidades.

Ciber: Elemento prefijal que surge a partir de la palabra cibernáutica y que entra en la formación de nombres con el significado de 'cibernético' y más concretamente de 'informática'.

Inteligencia: Facultad de la mente que permite aprender, entender, razonar, tomar decisiones y formarse una idea determinada de la realidad.

Ciber inteligencia: Es la adquisición y análisis de información para identificar, rastrear, predecir y contrarrestar las capacidades, intenciones y actividades de los ciber actores (atacantes), y ofrecer cursos de acción con base en el contexto particular de la organización, que mejoren la toma de decisiones.

Fuentes abiertas: Aplicación informática, de la que una Administración Pública posee la titularidad de los derechos de propiedad intelectual y que ha sido desarrollada por ella misma o por un tercero mediante la pertinente contratación, que se pone a disposición de otra Administración Pública, sin contraprestación económica y sin un convenio específico, con el objetivo de una mayor transparencia o una mejor incorporación ciudadana a la Sociedad de la información

Perfilamiento de adversarios: Es la metodología de poder identificar cuáles son las principales características y motivaciones de un atacante, entender sus motivos, oportunidades y medios, así como entender su modus operandi.

Aprendizaje automático: El aprendizaje automático o aprendizaje de máquinas (del inglés, (Machine Learning) es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. ... Es, por lo tanto, un proceso de inducción del conocimiento.

MON: En el derecho penal de Estados Unidos, los medios, el motivo y la oportunidad son una suma común de los tres aspectos de un delito que deben establecerse antes de que la culpabilidad pueda posiblemente determinarse en un proceso penal. Respectivamente, se refieren a: la capacidad del acusado para cometer el delito (medio), la razón por la que el acusado cometió el delito (motivo) y si el acusado tuvo la oportunidad de cometer el delito (oportunidad).

Análisis de sentimientos: Es el proceso por el que determinamos si una frase o acto de habla contiene una opinión, positiva o negativa, sobre una entidad concreta o sobre un concepto en específico.

Ciencia de datos: La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados, 1 lo cual es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático y la analítica predictiva.

Criminal: Todo individuo que comete un crimen o que está implicado en algún tipo de delito. Como adjetivo calificativo, el término también puede aplicarse a organizaciones delictivas como así mismo a aquellas que luchan en contra de las primeras. Finalmente, también puede ser criminal un acto o hecho que interrumpe el designio de la ley e implica el cumplimiento de algún tipo de delito.

Ciber Criminal: En líneas generales son personas que realizan actividades delictivas en internet como robar información, acceder a redes privadas, estafas, y todo lo que tiene que ver con los delitos e ilegalidad.

Víctima: Persona o animal que sufre un daño o un perjuicio a causa de determinada acción o suceso.

OSINT: El acrónimo anglosajón OSINT se refiere Open Source Intelligence o Inteligencia de fuentes abiertas. Las fuentes de información OSINT, es un término acuñado y muy empleado entre militares, fuerzas del orden y personal de inteligencia de las agencias gubernamentales.

Open Source Intelligence (OSINT a partir de ahora) es la disciplina responsable de la adquisición, procesamiento y posterior transformación en inteligencia de información obtenida de fuentes públicas como prensa, radio, televisión, internet, informes de diferentes sectores y, en general, cualquier recurso de acceso público.

La efectividad de este tipo de adquisición de inteligencia parece haber sido demostrada, su uso y gestión son objeto de constante estudio y mejora, formando parte de la metodología para recopilar y analizar información de entidades públicas y privadas.

Dado el tamaño de la Internet profunda, es muy fácil que, aunque hayamos definido nuestro conjunto de fuentes abiertas, estamos perdiendo mucha de la información disponible. Esto se debe a la imposibilidad de acceder a toda la información de forma manual y porque nos es imposible conocer todas las fuentes de información que existen.

Por lo tanto, el uso de herramientas para explotar fuentes conocidas y localizar aquellas que no conocemos es fundamental.

Las fuentes de OSINT se pueden dividir en seis diferentes categorías de flujo de información:

Medios: imprima periódicos, revistas, radio y televisión desde y entre países.

Internet: publicaciones en línea, blogs, grupos de discusión, medios ciudadanos (es decir, videos de teléfonos celulares y contenido creado por el usuario), YouTube y

otros sitios web de redes sociales (es decir, Facebook, Twitter, Instagram, etc.). Esta fuente también supera a una variedad de otras fuentes debido a su puntualidad y facilidad de acceso.

Datos del gobierno público, informes del gobierno público, presupuestos, audiencias, guías telefónicas, conferencias de prensa, sitios web y discursos. Aunque esta fuente proviene de una fuente oficial, son de acceso público y se pueden usar abierta y libremente.

Publicaciones profesionales y académicas, información adquirida de revistas, conferencias, simposios, trabajos académicos, disertaciones y tesis.

Datos comerciales, imágenes comerciales, evaluaciones financieras e industriales y bases de datos.

Literatura gris, informes técnicos, preimpresiones, patentes, documentos de trabajo, documentos comerciales, trabajos inéditos, disertaciones y boletines informativos.

Entidad del Estado: Unidad delimitada territorialmente que en unión de otras entidades conforman a una nación. En los sistemas federales las entidades pueden participar en las actividades gubernamentales nacionales y actuar unilateralmente, con un alto grado de autonomía, en las esferas autorizadas en la Constitución, incluso en relación con cuestiones decisivas y, en cierta medida, en oposición a la política nacional, ya que sus poderes son efectivamente irrevocables.

Adversarios: El adversario puede ser un individuo que, por situaciones específicas, aparece como un contrario para otra persona ya que tiene intereses opuestos a los suyos.

Activos: El activo son los bienes, derechos y otros recursos de los que dispone una empresa, pudiendo ser, por ejemplo, muebles, construcciones, equipos informáticos o derechos de cobro por servicios prestados o venta de bienes a clientes. También, se incluirían aquellos de los que se espera obtener un beneficio económico en el futuro

Amenaza: Se conoce como amenaza al peligro inminente, que surge, de un hecho o acontecimiento que aún no ha sucedido, pero que de concretarse aquello que se dijo que iba a ocurrir, dicha circunstancia o hecho perjudicará a una o varias personas en particular.

TABLA DE CONTENIDOS

TABLA DE IMÁGENES

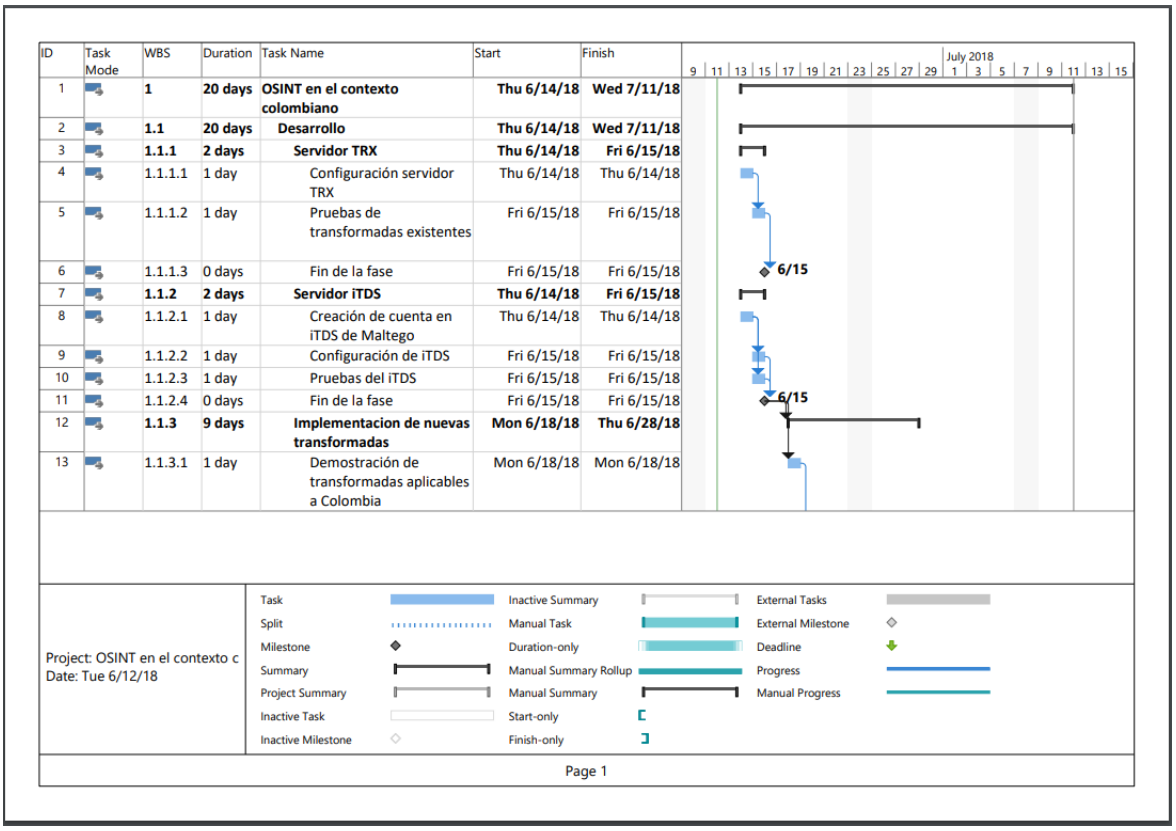
1 Objetivo general del proyecto

Generar un conjunto de artefactos de inteligencia de fuentes abiertas que esté adaptada al contexto colombiano que pueda ser de apoyo a las autoridades colombianas para proteger los activos de la Nación.

2 Objetivos específicos

1. Desarrollo y depuración de un conjunto de artefactos capaces de traer información desde las diversas fuentes abiertas.
2. Desarrollo de un software que extraiga información relevante de los contactos de LinkedIn del usuario.
3. Despliegue de una arquitectura de inteligencia utilizando un servidor TRX para la recolección de información.
4. Desarrollo y depuración de modelos de análisis de sentimiento que permitan perfilar mejor a un objetivo de inteligencia.

3 Cronograma



ID	Task Mode	WBS	Duration	Task Name	Start	Finish
1		**1**	**20 days**	**OSINT en el contexto colombiano**	**Thu 6/14/18**	**Wed 7/11/18**
2		**1.1**	**20 days**	**Desarrollo**	**Thu 6/14/18**	**Wed 7/11/18**
3		**1.1.1**	**2 days**	**Servidor TRX**	**Thu 6/14/18**	**Fri 6/15/18**
4		1.1.1.1	1 day	Configuración servidor TRX	Thu 6/14/18	Thu 6/14/18
5		1.1.1.2	1 day	Pruebas de transformadas existentes	Fri 6/15/18	Fri 6/15/18
6		1.1.1.3	0 days	Fin de la fase	Fri 6/15/18	Fri 6/15/18
7		**1.1.2**	**2 days**	**Servidor iTDS**	**Thu 6/14/18**	**Fri 6/15/18**
8		1.1.2.1	1 day	Creación de cuenta en iTDS de Maltego	Thu 6/14/18	Thu 6/14/18
9		1.1.2.2	1 day	Configuración de iTDS	Fri 6/15/18	Fri 6/15/18
10		1.1.2.3	1 day	Pruebas del iTDS	Fri 6/15/18	Fri 6/15/18
11		1.1.2.4	0 days	Fin de la fase	Fri 6/15/18	Fri 6/15/18
12		**1.1.3**	**9 days**	**Implementacion de nuevas transformadas**	**Mon 6/18/18**	**Thu 6/28/18**
13		1.1.3.1	1 day	Demostración de transformadas aplicables a Colombia	Mon 6/18/18	Mon 6/18/18

Figura 1 Cronograma Página #1

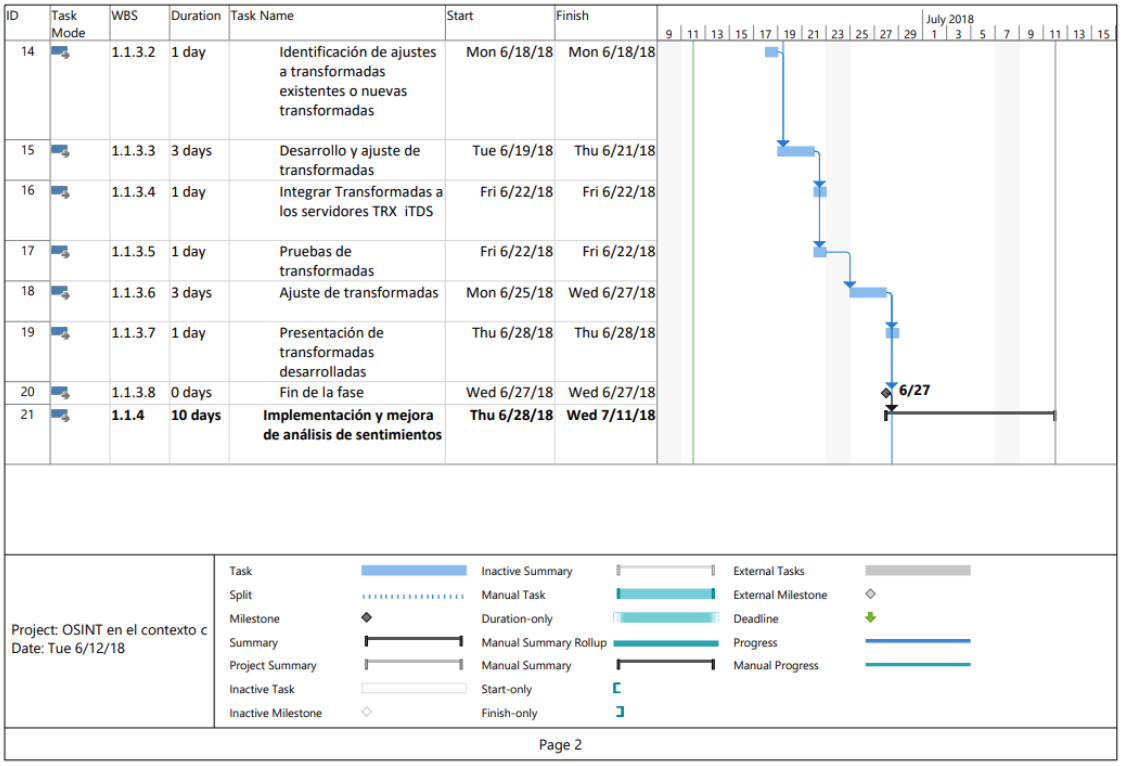


Figura 2 Cronograma Página #2

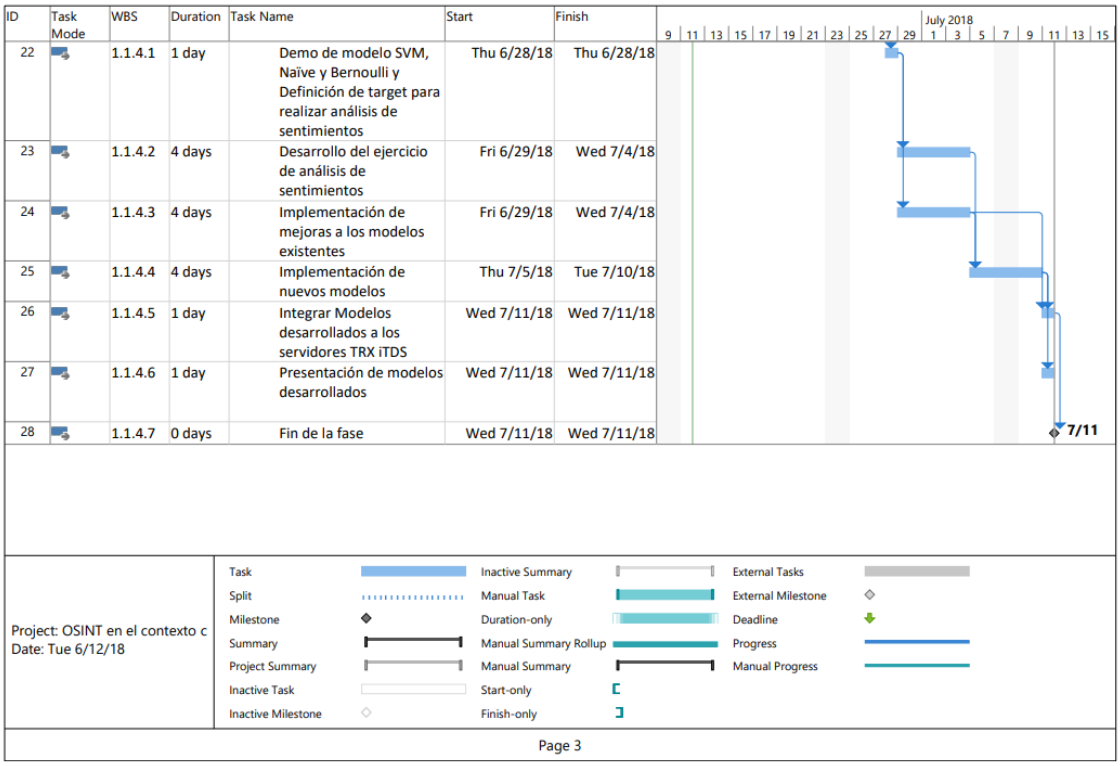


Figure 3 Cronograma Página #3

Fecha	Planeado	Ejecutado
25 de enero	● Investigar todo lo relacionado con OSINT para responder preguntas como: ¿Qué es?, ¿Para qué se usa?, etc. ● Investigar qué herramientas de OSINT están disponibles en actualmente en el mercado. ● Elaborar una tabla comparativa de las diferentes herramientas encontradas, para poder determinar las ventajas y desventajas y poder seleccionar 3 herramientas que serán utilizadas para el desarrollo del proyecto.	● A partir de la investigación realizada sobre OSINT para el uso de información de fuentes abiertas, se respondió las preguntas planteadas sin ningún inconveniente, dichas preguntas ayudaron a entender el contexto de toda esta información, así pudiendo concluir que las fuentes abiertas son recursos importantes en la tarea de realizar un producto de inteligencia. ● Se realizó una investigación a profundidad de las mejores 10 herramientas para realizar OSINT, se encontraron varias herramientas pagas, código abierto y extensión de la misma que podían usarse sin ningún tipo de inconveniente. ● Se seleccionaron las principales características (funcionales y no funcionales) de las herramientas, para así realizar una tabla comparativa para determinar cuáles herramientas solucionan parcialmente las necesidades que nosotros necesitábamos.

31 de enero	● Investigar a fondo el funcionamiento de las 3 herramientas de OSINT anteriormente seleccionadas. ● Elaborar una presentación y un informe que muestre todas las funcionalidades más importantes de cada una de estas herramientas. ● Elaborar una lista de características funcionales y no funcionales que debería tener una herramienta OSINT. ● Realizar un producto de ciber inteligencia sobre un target en específico y elaborar una presentación mostrando la información obtenida usando las 3 herramientas de OSINT seleccionadas anteriormente.	● Se realizó una investigación de las herramientas “Maltego”, “SpiderFoot” e “Intel Techniques”, donde se respondía cómo funcionaban y qué características ofrecían. ● Se elaboró dicha presentación donde claramente se expone las principales características de cada herramienta y se muestra una pequeña demostración en vivo de cada una de las herramientas. ● Se elaboró un listado de características funcionales y no funcionalidades a partir de la información recolectada de estas 3 herramientas seleccionadas. ● Se usaron las 3 herramientas seleccionadas para realizar un producto de ciber inteligencia al “target” Daniel Díaz.

8 de febrero	● Realizar una investigación exhaustiva del “target” Orlando Rico Rodríguez (ORR), cada uno debe usar la herramienta OSINT elegida y presentar toda la información posible obtenida con dicha herramienta. ● Realizar un producto de ciber inteligencia sobre un target en específico y elaborar una presentación mostrando la información obtenida usando las 3 herramientas de OSINT seleccionadas anteriormente.	● Cada uno realizó un documento donde publicó la información obtenida a partir de los datos inicialmente dados, haciendo uso de las herramientas de OSINT seleccionadas. ● Se realizó un producto de ciber inteligencia uniendo los tres documentos anteriormente mencionados sobre el target ORR y se entregó al comando conjunto cibernético de Colombia.
15 de febrero	● Se seleccionó la herramienta de OSINT “Maltego” y se pidió investigar cómo realizar una transformada. ● Luego de investigar cómo se elaboraba una transformada, se solicitó construir la primera transformada e integrar la a la herramienta “Maltego”.	● Se realizó una profunda investigación en manuales, documentación de la herramienta y se logró encontrar videos explicativos de cómo se puede realizar dichas transformadas. ● Se seleccionó la primera fuente de datos públicos de Colombia, en este caso fue de la página de internet “registraduria.gov.co” donde se obtuvo el lugar de expedición de la cédula, a partir del dato “Cedula”. Dicha transformada se integró a “Maltego” de forma satisfactoria.

22 de febrero	● Investigación sobre ¿Qué es minería de datos?, ¿Qué es análisis de sentimiento? ● Investigar cuales son los métodos y características del análisis de sentimiento y elaborar una presentación donde se plasme un resumen de lo más relevante del tema. ● Primer entregable de herramientas que permita el análisis de sentimiento.	● Al realizar una investigación sobre qué es la minería de datos y el análisis de sentimientos se obtuvieron los conocimientos básicos para empezar a abordar el tema mucho más a fondo. ● Se concluye que existen 2 métodos al momento de realizar minería de datos y análisis de sentimientos los cuales se profundizó en los métodos descriptivos y los métodos predictivos, también se destacó 4 características clave que eran: localización de palabras clave, afinidad léxica, métodos estadísticos y técnicas a nivel de concepto. ● Se investigó proyectos de análisis de sentimiento de código libre (preferiblemente en Python) y se procedió a realizar pruebas y entender el funcionamiento de los mismos, se presentaron 3 demostraciones una por cada proyecto escogido.

1 de marzo	● Realizar una fuerte comparación de los tres proyectos elegidos y determinar cuáles eran sus ventajas y desventajas. ● Entender el funcionamiento de cada proyecto y determinar cuál era la precisión de cada uno de los proyectos. ● Búsqueda de más fuentes de información colombiana.	● Se realizó una comparativa de los proyectos de análisis de sentimientos elegidos (todos ellos de código libre y escritos en Python) los cuales fueron “Text Blob”, “Aylien”, “NLTK y Spanish DAL”, para cada proyecto se mostró sus fortalezas y debilidades al momento de “clasificar” una frase. ● Se estudió a profundidad el código fuente que cada integrante encontró en sus proyectos, ya que eran códigos escritos por otros programadores. ● Se buscaron diferentes páginas web colombianas que brindaran información pública que fuera útil para la generación de un producto de inteligencia colombiano.
8 de marzo	● Investigar cómo se realiza el proceso de cálculo de sentimiento a las frases introducidas por el usuario. ● Elaborar una presentación donde se pueda evidenciar el proceso y el paso a paso de cada uno de esos modelos.	● Se investigó qué librerías usaban los tres proyectos de análisis de sentimiento, también los archivos usados para entrenar (como un diccionario). ● ya previamente analizado los proyectos se procese a plasmar los procesos en presentaciones realizadas a Daniel Orlando Díaz para explicar y decidir si se seguían con esos proyectos.

15 de Marzo	● Elaboración de un conjunto de transformadas haciendo uso de todas las fuentes abiertas de Colombia ya previamente consultadas. ● Elaboración de un video donde se muestre el funcionamiento de maltego y el uso de las transformadas realizadas.	● Se elaboraron un total 35 transformadas que consultaban todo tipo de páginas web colombianas, como Datos abiertos, Registraduría, Policía, Ejercito, Sisben, etc. ● Tras ya tener implementadas algunas transformadas para el proyecto se quiso dejar evidencia del funcionamiento de alguna de estas transformadas y que servían ingresando datos colombianos.
22 de Marzo	● Investigar cómo construir un conjunto de entrenamiento para los diferentes modelos de análisis de sentimientos a partir de frases.	● Tras investigar sobre el análisis de sentimientos se decidió a profundizar en cómo crear un conjunto de entrenamiento para los modelos descriptivos de tal manera que fueran entrenados para dar el análisis de un sentimiento de una frase en español.
10 de Abril	● Elaboración de cada uno de los modelos para el análisis de sentimientos (Bayes Naive, Máquinas de vectores de soporte, Bernoulli) ● Explicar el procedimiento que se realizaba desde la “limpieza del texto” hasta la “clasificación” de la frase.	● Luego de haber definido los pasos necesarios para hacer análisis de sentimiento, se investigó y se eligió tres modelos de análisis de sentimiento y se procedió a implementarlos en Python, para la correcta integración a las herramientas de OSINT, también se realizaron pruebas y se comparó la precisión de cada modelo.

		● Tras revisar en varias fuentes de información tales como: páginas web, blogs, artículos científicos, libros dedicados al machine learning, se logra encontrar un modelo básico de procesamiento de una frase para poder obtener el sentimiento .
19 de Abril	● Realizar un informe de OSINT similar al de ORR. Revisar si tienen redes sociales (twitter) y hacer análisis de sentimientos si la tienen. ● Buscar 3 perfiles de twitter (personas o grupos) con alto contenido de post asociados a cualquier temática que permitan hacer análisis de sentimientos. Por ejemplo: Periodistas, jefes de prensa, community managers, columnistas, caricaturistas.	● Se realizó una búsqueda en profundidad con las herramientas de OSINT a tres “target” que pertenecían a las entidades del gobierno y se procedió a realizar un producto de ciber inteligencia, también se les realizó un análisis de sentimiento a los comentarios u opiniones de estos tres individuos usando los modelos de análisis de sentimiento anteriormente implementados. ● se encontraron varias cuentas de twitter que cumplían con los requerimientos anteriormente mencionados, ya que tenían alto contenido relacionado con varios temas como lo era el fracking, las elecciones, la homosexualidad el aborto

		etc., de tal manera que se pudieran probar los diferentes modelos ya implementados y además buscar posibles atacantes.
26 de Abril	● Realizar nuevamente un informe de OSINT similar al de ORR con diferente “target”, esta vez enfocado a personas con alto rango en el ejército. Revisar si tienen redes sociales (twitter) y hacer análisis de sentimientos si la tienen.	● Se realizó una búsqueda en profundidad con las herramientas de OSINT a tres “target” que pertenecían ejército y tenían investigaciones en contra. y se procedió a realizar un producto de ciber inteligencia, también se les realizó un análisis de sentimiento a los comentarios u opiniones de estos tres individuos usando los modelos de análisis de sentimiento anteriormente implementados.
3 de Mayo	● Preparación para la participación en el 8vo seminario de seguridad de la Escuela Colombiana de Ingeniería Julio Garavito .	● Se realizó búsqueda de perfiles que tengan la posibilidad de hacer ataques a la infraestructura crítica de Rusia para generar retrasos en la ejecución del mundial de Rusia 2018.

10 de mayo	● Participación en OWASP latam tour 2018 Evento realizado en la Escuela Colombiana de Ingeniería, con la ponencia “Open Source Intelligence” mostrando los beneficios de las fuentes abiertas para la ciberseguridad por medio de herramientas de código libre como lo es Maltego.	● Se formalizó una presentación concreta con todo lo realizado en el transcurso del proyecto, mostrando en dicho evento el avance del proyecto.
5 de junio	● Desarrollo de la primera versión del artículo para PGR2. ● Investigar cómo obtener datos de LinkedIn usando su api REST.	● Se desarrolló la primera versión del artículo donde se incluye lo primordial de la investigación y desarrollo del proyecto. ● Se desarrolló la primera versión del extractor de LinkedIn usando el método scrapping, dado que el api necesita una previa autorización para extraer información de otros perfiles.
8 de junio	● Realización primera versión del extractor de LinkedIn ● Mejoramiento de la primera versión del extractor de información de LinkedIn. ● Levantamiento de información para implementar un servidor ITDS y TRX.	● Dicha primera versión extraía información de cada graduado como su trabajo actual, último registro académico, la descripción de la persona y su fotografía, aún restaba extraer los logros y aptitudes de cada graduado al igual que su historial académico y laboral completo.

		● Se implementó la versión 2 del extractor de LinkedIn con interfaz gráfica. ● Se levantó información para poder implementar el servidor, donde se especifican los requerimientos mínimos que debe tener la máquina.
12 de junio	● Elaboración del cronograma a seguir para realizar la debida implementación de los servidores.	● Se elaboró un cronograma donde se especificaron las fechas exactas de implementación de dichos servidores, y tareas posteriores, ● Se elaboró el cronograma en Microsoft Project. ● Se realizó la entrega el extractor de información LinkedIn.
14 de junio	● Configuración de servidor TRX dentro de las instalaciones del Comando Conjunto Cibernético CCOC. ● Creación de cuenta en servidores de Paterva	● Se realizó la instalación de un servidor Ubuntu 16.04 con características TRX y Apache instalado, a pesar de eso, las configuraciones de las transformadas y la instalación de las mismas se pudo realizar exitosamente hasta el 7 de julio por cuestiones de incompatibilidad, falta de permisos. La creación de la cuenta se realizó exitosamente

15 de junio	● Prueba de transformadas existentes, configuración del servidor ITDS y pruebas en el mismo	● Se realizó la adaptación de las transformadas a formato del backend correspondiente al servidor TRX y se crearon los seeds asociados a cada transformada dentro del servidor ITDS.
18 de junio	● Demostración de transformadas aplicadas a Colombia ● Identificación de ajustes que se deben realizar a las transformadas ya existentes y propuesta de nuevas transformadas	● Se realizó una sesión de demostración de las transformadas ya desarrolladas a personal del Comando Conjunto Cibernético y se recibió retroalimentación de manera que se propusieron 9 nuevas transformadas que acceden a fuentes de información colombiana.
19 de junio	● Desarrollo y ajuste de transformadas acordadas	● Se realizó parcialmente dicho desarrollo pero se tomó más de un día.
22 de junio	● Integración de las nuevas transformadas al servidor TRX ● Prueba de las transformadas recientemente integradas	● Se realizó una visita al Comando Conjunto Cibernético para la labor planificada pero nuevamente se tuvieron problemas de asignación de permisos a la máquina sobre la cual está el servidor TRX, debido a eso se retrasó la labor y no se pudo completar sino hasta el 7 de julio junto con las transformadas iniciales

28 de junio	● Presentación de transformadas desarrolladas al personal del Comando Conjunto Cibernético ● Implementación de modelos de análisis de sentimiento	● Al finalizar el desarrollo y la adaptación al servidor TRX (7 de julio) las transformadas ya estaban disponibles para añadir a maltego desde cualquier cliente y se realizó la entrega del código fuente de las mismas en versión TRX y en versión local normal, adicional a eso se envió un manual de uso y un manual de instalación que documentan la manera como se deben aprovechar dichas transformadas. ● El desarrollo de análisis de sentimiento se extendió hasta el 18 de julio
18 de julio	● Entrega modelos análisis de sentimiento ● Segunda fase extractor de LinkedIn	● Se finalizó la estandarización de los tres modelos para hacer que las fases de pre procesamiento fueran lo más similares posible y que de esta manera fueran más comparables. ● La segunda fase del extractor contemplaba el historial laboral y académico al igual que los logros y aptitudes pero aún restaba el mejoramiento de la interfaz gráfica
26 de julio	● Última versión del extractor de LinkedIn	● Integración de una interfaz gráfica desarrollada en Java para permitir una mejor visualización.

Tabla 1 Cronograma Proyecto

4 Estado del arte

OSINT tools overview

Para realizar una labor de OSINT existen diversas herramientas que ayudan al desarrollo adecuado de la investigación y el perfilamiento, aunque, se debe mencionar que cada una de las presentadas a continuación está orientada a un tipo de uso diferente, por lo tanto, la mayoría son complementarias entre ellas.

Herramienta	Licencia	Búsqueda	Plataforma
Maltego	MIT	Dominios, nombre de usuario, archivos, url, correos, etc..	Linux, Windows, Mac
Metagoofil	GNU 2.0	Metadatos de documentos	Linux, Windows
The Foca	GPL 3.0	Metadatos de documentos	Linux, Windows
Shodan	MIT	Servidores, routers, dispositivos web, etc…	Web
The Harvester	GPL 2.0	Correos	Linux, Windows, Mac
Recon-NG	GNU 2.0	Correos	Linux
Spiderfoot	GPL 2.0	Dominios, nombre de usuario, archivos, url, correos, etc..	Linux, Windows
Intel Techniques	N/A	Información personal	Web

Tabla 2 Estado del arte herramientas OSINT

La primera herramienta que se mencionara es The Harvester, la cual se orienta a encontrar correos electrónicos de un dominio proporcionado por el usuario buscando en servidores como Google, Bing, LinkedIn, etc... Esta información puede ser útil para obtener posibles puntos de entrada a la información de una organización, por ejemplo, utilizando el dominio de la misma obtener varios correos asociados a ella.

Si se está hablando de recolección de información a partir de dominios, se debe mencionar a ReconNG, que tiene un módulo para el escaneo de dominios que muestra gran cantidad acerca de éstos y se pueden usar de manera similar a The Harvester, obteniendo correos de una organización, además, ReconNG permite obtener ubicación de un dominio, nombre de su administrador y algunos datos más similares a Who Is, lo cual es útil para perfilar a una organización y sus integrantes.

Si se está hablando de OSINT no se debe dejar de lado a la herramienta FOCA, la cual permite encontrar metadatos existentes en documentos de Open Office, Microsoft Office y documentos PDF y cruzar dichos metadatos para obtener información relevante, por ejemplo, si se está realizando inteligencia a una persona y se encuentran archivos relacionados, se puede identificar la fecha en que fue creado y modificado y el nombre del usuario que realizó dichos cambios.

No se debe dejar de lado la Herramienta Shodan la cual es un motor de búsqueda que sirve para encontrar servidores, routers o cualquier tipo de dispositivo conectado a internet por medio protocolos como HTTP, SSH, Telnet, etc… y se apoya en un conjunto de filtros de búsqueda cuyo acceso depende del tipo de suscripción que se tenga. Por ejemplo, se puede realizar una búsqueda por keyword “Password:” y Shodan puede traer dispositivos que tengan dentro de su código fuente una asignación de contraseña y por consiguiente de acceso a dicho dispositivo.

Por último, se menciona la herramienta Metagoofil, la cual ayuda a que el usuario pueda extraer metadatos de documentos de formatos como DOC, XLS, PPT, PDF, etc… Por ejemplo, si se está realizando inteligencia a un determinado objetivo y se encuentran archivos de tipo ofimático como documentos o hojas de cálculo se puede identificar la fecha en que fue creado y modificado y el nombre del usuario que realizó dichos cambios.

Maltego

En este caso vamos a hablar de una de las muchas herramientas que facilita la obtención de información por las distintas fuentes abiertas y públicas, estamos hablando específicamente de Maltego, una poderosa herramienta que tiene un potencial enorme a la hora de encontrar información sobre personas, empresas, organizaciones o un “target” en específico.

Si bien el reconocimiento es la primera etapa de cualquier ataque, es importante tener en cuenta que estos mismos conocimientos (las distintas metodologías para la obtención de información de una persona u organización en específico) son utilizados por analistas del sector público y privado, como la inteligencia empresarial, para ayudar a los inversores y las empresas a tomar decisiones importantes. Muchas empresas pagan mucho dinero por el mismo tipo de habilidades de investigación (la capacidad de recolección, evaluación, análisis, integración e interpretación de toda la información disponible de la víctima, permitiendo su transformación en conocimiento, de forma que resulte útil a la hora de tomar decisiones con el menor nivel de incertidumbre posible.) que los hackers emplean al seleccionar un objetivo.

Maltego permite cruzar datos para obtener perfiles en redes sociales, servidores de correo, etc... Además, permite iniciar búsquedas a partir de dominios, IPs, ubicaciones geográficas, correos, nombres, teléfonos e incluso frases. Usa grafos para representar las relaciones que existen entre los diferentes tipos de entidades halladas, todo con el fin de ofrecer una interfaz que sea sencilla y fácil de usar, está llena de opciones que pueden ser muy útiles para realizar todo tipo análisis de ciberinteligencia.

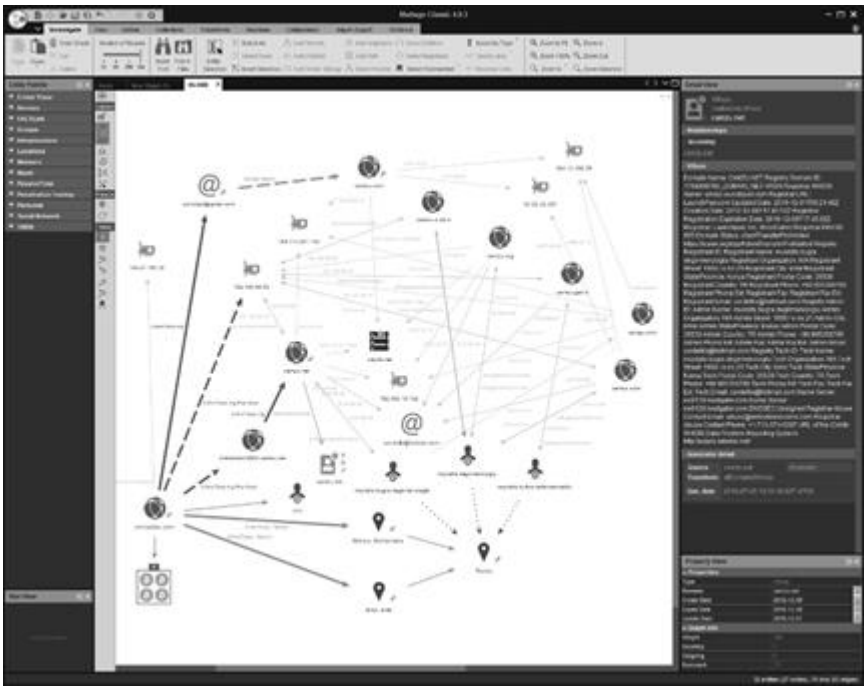


Figure 4. Herramienta OSINT: Interfaz gráfica de Maltego en su versión 4.

¿Pero esta información pública es útil? ¿Tiene alguna importancia o sentido para alguien? Estas preguntas pueden surgir por toda la información desclasificada y pública que abunda en internet o por la sencilla razón de que no se tiene claro un contexto donde pueda usarse, para esto veamos un claro ejemplo donde esta información puede llegar a usarse para fines malignos, un claro ejemplo es cuando un ciberdelincuente usa toda esta información valiosa para preparar un ataque, ya que con la misma puede, por ejemplo, valerse de perfiles psicológicos y tomar personalidades falsas en las redes sociales. Vale recordar que, en ocasiones, los ciberdelincuentes buscan atacar redes corporativas y lo logran haciéndose pasar por empleados a través de la recolección de información.

Otro ejemplo, puede ser a la hora de buscar establecer contacto con una empresa, esta herramienta puede proporcionarnos datos muy útiles como direcciones de correo electrónico asociados a un dominio de la organización, números telefónicos, lo que nos facilita el contacto con esta empresa o persona.

Una entidad es una representación abstracta de cualquier tipo de información que se encuentra en la vida real, como por ejemplo nombres de usuario, correos, nombres completos, teléfonos, direcciones, redes sociales, etc. Actualmente Maltego ofrece alrededor de 36 tipos de transformadas, en la Tabla 2 pueden observarse las principales entidades.

Entidad	Descripción
Afiliación	Membresía de una red social
Afiliación Flickr	Membresía de la red social Flickr
Afiliación Twitter	Membresía de Twitter
Alias	Un alias para una persona
Área circular	Un área circular en algún lugar de la Tierra
Dispositivo	Un dispositivo como un teléfono o cámara
Nombre DNS	Nombre del servidor del sistema de nombres de dominio
Documento	Un documento en Internet
Dominio	Un dominio de internet
Dirección de correo electrónico	Un buzón de correo electrónico
Archivo	Un archivo almacenado internamente en el gráfico
Coordinación de GPS	Una ubicación en un marco de coordenadas del Sistema Geodésico Mundial
Hashtag	Hashtag de Twitter
Imagen	Una representación visual de algo
Dirección IPv4	Una dirección IP versión 4
Ubicación	Una ubicación en la Madre Tierra

Persona	Entidad que representa un ser humano
Número de teléfono	Un número de teléfono
Frase	Cualquier texto o parte del mismo
Puerto	Un puerto de red TCP / UDP
Sentimiento	Esto representa el sentimiento hacia una entidad.
Servicio	Servicio de red (combinación de puerto y banner)
Código de localización	Representa un código de seguimiento para un servicio web.

Tabla 3 Principales Entidades en Maltego

Maltego ofrece alrededor de 146 transformadas y un total de 15 categorías (Tabla 2), dichas transformadas pueden usarse para la prevención a posibles ataques de cibercriminales, con la información relevante proveniente de fuentes abiertas, podemos identificar los posibles canales utilizados por el cibercriminal para cometer los abusos o ataques. Entender los nuevos métodos de ataque, así como los patrones de conducta de los cibercriminales. Detectar las amenazas que se originan en Internet y que pueden poner en riesgo a una persona o una compañía. Detectar de forma temprana las amenazas a las infraestructuras.

Categoría	**Descripción**
Convertir a dominio	Convierte sitios web, MX, NS y nombres DNS a sus dominios
DNS del dominio	Encuentra nombres DNS para un dominio usando varias transformaciones
DNS desde IP	Prueba un par de transformaciones para obtener DNS o datos históricos inversos

Detalle del propietario del dominio	Encuentra información sobre el whois del dominio, como números de teléfono y direcciones de correo electrónico
Dominio usando MX NS	Determina qué dominios están asociados con la IP en términos de MX compartido o NS
Direcciones de correo electrónico del dominio	Obtener direcciones de correo electrónico usando el motor de búsqueda, PGP y whois
Direcciones de correo electrónico de una persona	Intenta obtener la dirección de correo electrónico de la persona
Archivos y documentos del dominio	Encuentra archivos y documentos relacionados con el dominio
Archivos y documentos de la frase	contiene transformaciones que buscarán archivos y documentos utilizando la frase
Encontrar IP en la web	Obtiene tanta información relacionada con esta IP
Detalles del propietario de IP	Un archivo almacenado internamente en el gráfico
Información de NS	Encuentra bloques de red y otros dominios que utilizan este NS
Enlaces dentro y fuera de comentarios	Enlaces hacia y desde un sitio web
Direcciones de correo electrónico relacionadas	Correos asociados
Resolver a IP	Resuelve a dominio

Tabla 4 Principales Categorías de Transformadas en Maltego.

SpiderFoot

Siguiendo con las herramientas que facilita la obtención de información por las distintas fuentes abiertas y públicas, esta vez hablaremos sobre SpiderFoot, Una excelente herramienta que tiene mucho que explotar, esta brinda la posibilidad de encontrar nombres de dominios, direcciones IP, Nombre de host, subdominios y subredes [2].

Ya que estamos hablando de OSINT, lo primordial en este proceso de inteligencia es obtener la mayor cantidad de información pública (sin cruzar la fina línea entre un proceso de obtención de información legal e ilegal) de nuestro “objetivo” y clasificar lo que nos interesa y no, no se nace con esta capacidad de análisis de la información de saber reconocer un “Falso positivo”, es un mejoramiento continuo que se obtiene al realizar muchos informes de inteligencia de fuentes abiertas.

Todo este proceso de recolección de información tiene un único fin, perfilar a nuestro “objetivo” que se está investigando y la forma en la que SpiderFoot realiza este perfilamiento es por medio de relaciones entre las entidades de información, comúnmente conocido como relación de grafos, que lo que permite es brindarle al usuario una vista muy limpia de toda la información que pudo relacionar para que de manera intuitiva el analista de ciber inteligencia pueda realizar un producto de inteligencia en la figura 4 podemos ver los diferentes tipos de búsqueda de información tales como:

- Modo All: obtener todo sobre el objetivo, todos los módulos de SpiderFoot estarán habilitados (lentos) pero se obtendrá y analizará toda la información posible sobre el objetivo.
- Modo Footprint: comprender qué información expone este objetivo a Internet, comprender el perímetro de la red del objetivo, las identidades asociadas y otra información que se obtiene a través del rastreo y el uso del motor de búsqueda.
- Modo Investigate: es lo mejor para cuando sospecha que el objetivo es malicioso, pero necesita más información, se realizará algunos footprinting básicos además de consultas de listas negras y otras fuentes que pueden tener información sobre la malicia de su objetivo.
- Modo Passive: Cuando no quiere que el objetivo siquiera sospeche que están siendo investigados, se recopilará tanta información sin tocar el objetivo o sus afiliados, por lo tanto, solo se habilitarán los módulos que no toquen el objetivo.

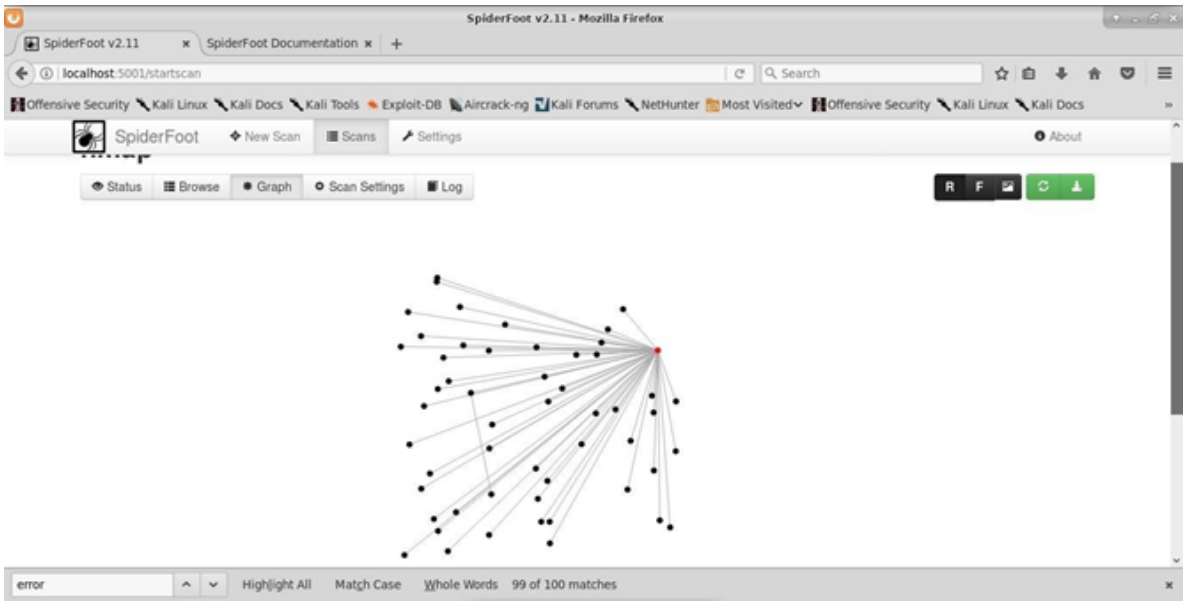


Figure 5 Herramienta OSINT: Interfaz gráfica de SpiderFoot en su versión 2.12.

Pero a todas estas, ¿Para qué sirve la información obtenida por SpiderFoot? ¡¡Absolutamente todo!! Bueno, está bien, casi todo, pero hay muchas formas en que SpiderFoot puede usarse por diversión o en su trabajo actual. Aquí hay unos ejemplos:

- Due Diligence de una compañía (la investigación de una empresa o persona previa a la firma de un contrato o una ley con cierta diligencia de cuidado. Puede tratarse de una obligación legal, pero el término comúnmente es más aplicable a investigaciones voluntarias, esto es due diligence. [7])
- Reclutamiento (En muchas empresas para evitar procesos de selección exhaustos y largos, realizan una búsqueda de las personas que cumple con las características que se necesitan. [8])
- Inteligencia de amenazas (ayuda a predecir incidentes en base a una serie de investigaciones y evidencias [9])

Estos son los clientes más comunes que podrían utilizar SpiderFoot para sus investigaciones de inteligencia:

- Agencias del Estado
- Cualquier compañía o una organización
- Posibles ciberdelincuentes

El contenido de la internet profunda (Deep Web) es de 400 a 550 veces mayor de lo que se puede encontrar en la internet superficial [3]. Información relacionada con narcotráfico, el crimen organizado, los paquetes de malware, la pedofilia, etc. [10]. Para que SpiderFoot pueda acceder a esta información alojada en la Deep Web simplemente se necesita descargar y ejecutar el cliente Tor "autónomo" y habilitar las conexiones de control para que SpiderFoot pueda controlarlo, ya que varios módulos SpiderFoot tienen mecanismos para detectar cuando el bloqueo de otros CAPTCHA de o está siendo implementado por diversas fuentes de inteligencia (por ejemplo, Google, Bing, etc.) Cuando SpiderFoot detecta esto, se utiliza la biblioteca de vástago para enviar un comando NEWNYM a Tor. Tor luego configura un nuevo circuito para conexiones posteriores. SpiderFoot intentará cambiar los circuitos hasta tres veces antes de darse por vencido. y así puede ser cosechada toda esta información y entregada en tiempo real.

Ya dando una introducción formal a SpiderFoot, entremos un poco a contestar siguientes las preguntas ¿Qué es SpiderFoot? ¿Qué es un módulo en SpiderFoot? ¿Qué es SpiderFoot?

SpiderFoot es una herramienta de automatización de inteligencia de fuentes abiertas (OSINT). Su objetivo es automatizar el proceso de recopilación de la mayor inteligencia posible sobre un objetivo determinado.

SpiderFoot se puede utilizar ofensivamente, es decir, como parte de una prueba de penetración de recuadro negro para recopilar información sobre su objetivo, o de forma defensiva para identificar qué información proporciona libremente su organización para que los atacantes la usen en su contra. Utiliza más de cincuenta fuentes de datos tales como motores de búsqueda, paginas previamente ya quemadas, servidores públicos etc. Para construir esta información, con más y más fuentes de datos agregadas con cada lanzamiento, desde 2012.

¿Qué es un módulo en SpiderFoot? Un módulo es el proceso “equivalente a las Transformadas en Maltego” de pasar de un tipo de dato o información a otro muy distinto pero que permiten ser relacionados y SpiderFoot no posee más componentes ya que todo está contenido en su código fuente, lo curioso es que cuando un módulo descubre una pieza de datos, esa información se transmite a todos los demás módulos que están 'interesados' en ese tipo de datos para procesarlos. Esos módulos actuarán sobre ese dato para identificar nuevos datos y, a su vez, generar nuevos eventos para otros módulos que puedan estar interesados, y así sucesivamente.

Por ejemplo, el módulo llamado sfp_dns puede identificar una dirección IP asociada con el objetivo, notificando a todos los módulos interesados. Uno de esos módulos interesados sería el módulo sfp_ripe, que tomará esa dirección IP e identificará el bloque de red del que forma parte, BGP ASN, y así sucesivamente.

Estos son algunos módulos que maneja SpiderFoot ya que posee más de 101 módulos diferentes:

Módulo	**Nombre**	**Descripción**
sfp_abusech.py	abuse.ch	Compruebe si un host / dominio, IP o netblock es malicioso de acuerdo con abuse.ch.
sfp_accounts.py	Cuentas	Busque posibles cuentas asociadas en casi 200 sitios web como Ebay, Slashdot, reddit, etc.
sfp_adblock.py	Comprobación de AdBlock	Compruebe si las páginas vinculadas serían bloqueadas por AdBlock Plus.
sfp_ahmia.py	Ahmia	Busca en el motor de búsqueda de Tor 'Ahmia' las menciones del dominio de destino.
sfp_alienvault.py	AlienVault OTX	Obtener información de AlienVault Open Threat Exchange (OTX)
sfp_alienvaultiprep.py	Reputación de IP AlienVault	Compruebe si una IP o netblock es malicioso de acuerdo con la base de datos de Reputación de IP de AlienVault.
sfp_archiveorg.py	Archive.org	Identifica versiones históricas de archivos / páginas interesantes de Wayback Machine.
sfp_badipscom.py	badips.com	Compruebe si un dominio o IP es malicioso de acuerdo con badips.com.
sfp_base64.py	Base64	Identifique las cadenas codificadas en Base64 en cualquier contenido y URL, a menudo revelando información oculta interesante.
sfp_bingsearch.py	Bing	Algunos ligeros Bing raspado para identificar subdominios y enlaces.
sfp_bingsharedip.py	Bing (IP compartidas)	Busque Bing para los hosts que comparten la misma IP.
sfp_binstring.py	Extractor de	Intentar identificar cadenas en contenido binario.

	cadenas binarias	

Tabla 5 Módulos de SpiderFoot.

OSINT Intel Techniques

Intel Techniques es un sitio web creado por Michael Bazzel que provee un conjunto de servicios orientados a OSINT.

Servicios pagos:

- Entrenamiento Online: Capacitación para aprender sobre OSINT gracias a contenidos como videos y libros certificando a quien adquiera este servicio.

- Conferencias: Se presenta la perspectiva de Michael respecto al delito informático, estas sesiones incluyen demostraciones reales en lugar de presentaciones estáticas.

- Entrenamiento presencial: Se ofrecen dos cursos presenciales dictados por instructores capacitados por Michael, el primero de estos cursos es sobre OSINT y el otro curso es acerca de privacidad y seguridad. En el portal se especifica el contenido diario de cada uno de los cursos.

- Otros servicios: como eliminación de datos en línea, supervisión de credenciales, investigaciones en línea y consulta personal.

Servicios gratuitos:

- Herramientas de búsqueda: conjunto de enlaces a páginas que sirven para búsqueda de información pública.
- Enlaces hacia libros y conferencias.
- Acceso a foros.
- Blog informativo
- Podcasts

Es preciso conocer un poco acerca de Michael Bazzel para saber la calidad del trabajo que ha realizado con Intel Techniques. Michael Bazzell pasó 18 años como investigador del crimen informático del gobierno estadounidense, años en los

cuales, fue asignado al Grupo de trabajo contra delitos cibernéticos del FBI, donde se centró específicamente en OSINT, casos de cibercrimen y métodos de eliminación de datos personales.

El uso más frecuente de Intel Techniques es el de su módulo de herramientas de búsqueda, el cual ofrece un conjunto de enlaces a páginas de búsqueda de información pública específica, por ejemplo, hay un módulo especializado en búsquedas de Facebook, donde gracias a un username de dicha red social, es posible obtener datos públicos del perfil como grupos a los que pertenece, lugares visitados, publicaciones a las cuales ha dado like, etc…

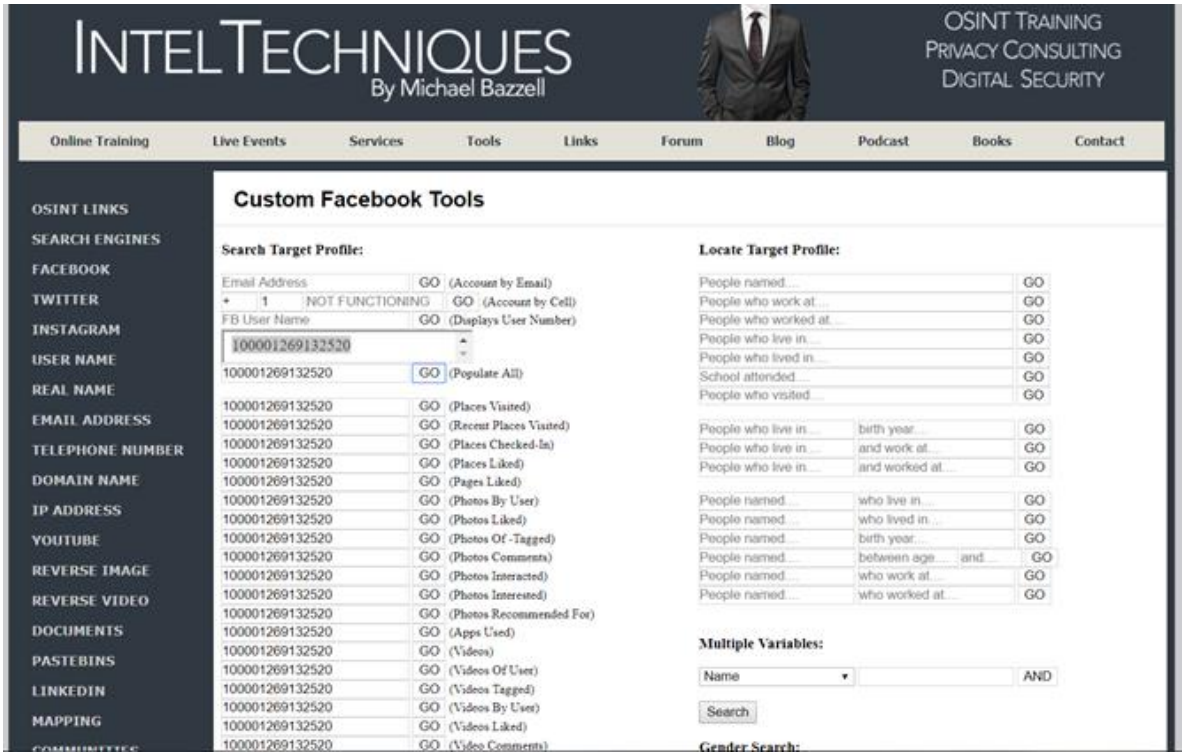


Figure 6 Interfaz de Intel Techniques.

Además de la búsqueda en Facebook, hay enlaces a páginas de reversa de imágenes, búsqueda de documentos usando Google Hacking, diversos servicios de WHOIS, etc…

Minería de texto:

Se debe partir del punto de vista de que el lenguaje natural utilizado por los seres humanos para comunicarse está dentro de la categoría de datos no estructurados, dada esta premisa se puede afirmar que el lenguaje natural escrito puede ser un punto de partida para realizar minería de texto extrayendo el significado de dicho texto.

Actualmente a grandes rasgos se podrían definir las siguientes utilidades contenidas dentro de la minería de datos:

Categorización de textos: asignar una clasificación o categoría al texto que se está procesando, por ejemplo, definir si el texto que va a ser tratado pasará por un proceso de identificación de idioma, análisis de sentimiento, etc.

Similaridad semántica de textos: Mediante este proceso se busca conocer qué tan similares semánticamente hablando son dos piezas de texto, por ejemplo, esta utilidad puede ser aplicada a buscadores indexadores de internet en la búsqueda de información.

Resumen de documentos: Busca encontrar una porción de datos que representen de manera generalizada las principales ideas presentadas en el texto extrayendo para esto las partes más informativas de dicho escrito.

Extracción de conceptos o entidades: Consiste en localizar elementos de un texto que están categorizados y definidos, es decir, encontrar dentro de un documento nombres de personas, organizaciones, lugares, etc...

Desambiguación de un significado: Busca la manera de poder seleccionar el sentido en el cual una palabra es utilizada dentro de una frase cuando la misma tiene muchos posibles significados.

Análisis de sentimientos: La meta dentro del análisis de sentimiento es determinar la polaridad de una frase o documento mediante distintos métodos como datos históricos, diccionario de palabras clasificadas, estadística, machine learning, etc...

Análisis de sentimiento:

Para trabajar el análisis de sentimiento, actualmente se tienen tres tipos de profundidad o de volumen de información que puede ser analizada:

Documento: Expresar o resumir la polaridad u opinión general de un documento

Frase: Analizar frases independientes y extraer la polaridad de las mismas

Entidad: Obtener un sentimiento dado a partir de un nombre propio reconocido ya sea personal o de tipo organización, así como también de abstracciones del mundo real tales como “Atención al cliente”, “Personal comercial”,

Teniendo en cuenta la profundidad del análisis es importante mencionar también aspectos que actualmente siguen siendo motivo de investigaciones como el hecho de que los sentimientos, la mayor parte de las veces son subjetivos y no se puede generalizar esta metodología ya que existen siempre casos que pueden estar fuera de la lógica de un modelo de clasificación. Para el problema anteriormente mencionado se ha pensado una estrategia como el aprendizaje automático supervisado o no supervisado.

El análisis supervisado se obtiene a partir de un conjunto de datos etiquetado con la polaridad del documento o frase, dicho etiquetado se puede realizar automatizadamente con opiniones escritas por personas públicamente ya sea en redes sociales, comentarios, etc…

En este análisis se tienen en cuenta variables como:

Términos del documento, su frecuencia y relevancia, lo cual se calcula gracias a técnicas estadísticas

Categorías gramaticales de las palabras tales como adjetivos, nombres, verbos, sustantivos, etc…

Expresiones y palabras con sentimiento directo (bueno, malo, horrible)

Palabras modificadoras de polaridad (no, nunca, poco, nadie)

El enfoque no supervisado se basa en comparar palabras consecutivas con patrones sintácticos predefinidos y prefijados. Dichos patrones tenían una polaridad en base a su distancia (en cuanto a sentido de la palabra) a palabras positivas y negativas, una vez hecho esto se calculaban todas las polaridades de estos patrones para calcular la totalidad del documento

5 Publicaciones o conferencias

Para la publicación del artículo que se genere en este Proyecto de Grado se han revisado las siguientes tres revistas científicas:

1. Journal of Engineering and Education
    a. Descripción:
        i. The Journal publica manuscritos en una amplia variedad de áreas de investigación en el campo de la educación de

ingeniería. Una descripción de las áreas de investigación actuales en educación de ingeniería se puede encontrar en el informe especial, "La agenda de investigación para la nueva disciplina de la educación en ingeniería"

   b. Universidad responsable:
      i. Clemson University.
   c. ISSN:
      i. 1069-4730
      ii. Online ISSN: 2168-9830
   d. URL:
      i. https://revistas.ucc.edu.co/index.php/in

2. TECKNE, Innovación e Investigación en Ingeniería
   a. Descripción:
      i. La Revista ITECKNE, Innovación e Investigación en Ingeniería, es una revista de carácter científico y tecnológico, editada semestralmente por la División de Ingenierías y Arquitectura de la Universidad Santo Tomás desde el año 2002. ITECKNE, es un medio de divulgación que busca promover la publicación científica que contribuya al desarrollo de la ciencia y la industria en un contexto nacional e internacional.
   b. Universidad responsable:
      i. Universidad Santo Tomás
   c. ISSN:
      i. 1692-1798
   d. URL:
      i. http://www.ustabuca.edu.co/ustabmanga/revista-iteckne
3. Vínculos
   a. Descripción:
      i. La revista Vínculos es una revista institucional de ingeniería en telemática y tecnología en sistematización de datos de la facultad tecnológica de la Universidad Distrital Francisco José de Caldas. El primer fascículo apareció en el año 2004 en su segundo volumen, hasta entonces ha mantenido su periodicidad de publicación.
   b. Universidad responsable:
      i. Universidad Distrital Francisco José de Caldas
   c. ISSN:
      i. 1794-211X
   d. URL:

i. http://revistas.udistrital.edu.co/ojs/index.php/vinculos

6 Análisis y diseño de transformadas

6.1 Cómo se ejecutan las transformadas

¿Qué es una Transformada?

Consiste en el proceso de pasar de un tipo de datos "X" a otro tipo de datos "Y", mirando las relaciones que tienen entre sí, como por ejemplo con una cedula puedo traer el nombre completo de una persona o con un correo puedo traer las redes sociales asociadas a dicho correo.

Maltego ofrece alrededor de 146 transformadas y un total de 15 categorías, dichas transformadas pueden usarse para la prevención a posibles ataques de cibercriminales, con la información relevante proveniente de fuentes abiertas, podemos identificar los posibles canales utilizados por el cibercriminal para cometer los abusos o ataques. Entender los nuevos métodos de ataque, así como los patrones de conducta de los cibercriminales. Detectar las amenazas que se originan en Internet y que pueden poner en riesgo a una persona o una compañía. Detectar de forma temprana las amenazas a las infraestructuras.

Luego de haber definido que es una transformada, explicaremos el proceso más general que realiza una transformada el cual consta de 3 partes fundamentales como se ilustra en la figura 1:

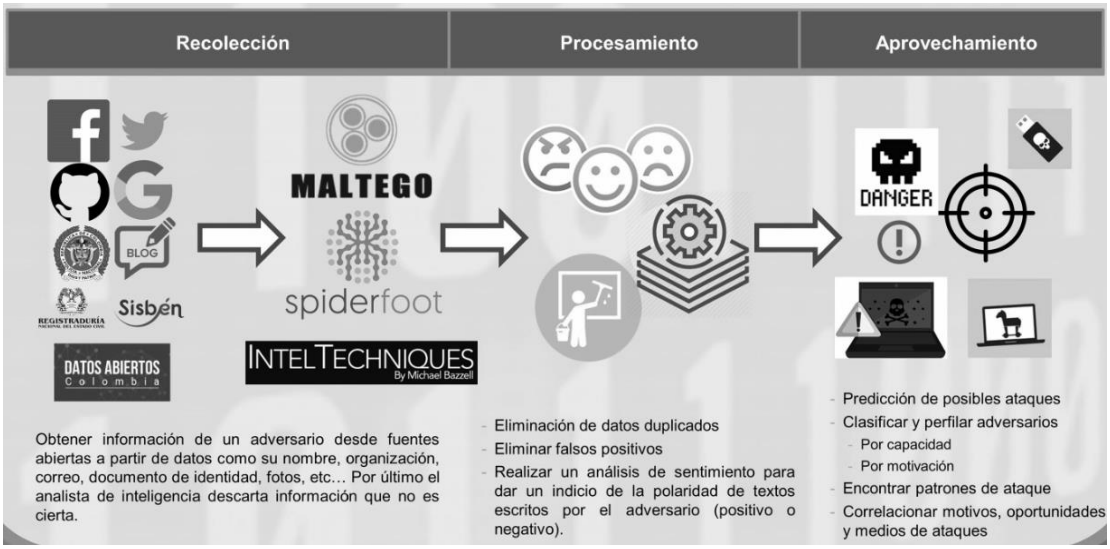


Figure 7 Arquitectura del proyecto, Fuente. Elaboración propia

- La recolección de la información (Gathering)
- El procesamiento de la información (Processing)
- Sacar provecho de la información (Taking advantage)

De tal manera que nos vamos a enfocar en cada una de ellas dando una mejor y clara definición de cada una.

- La recolección de la información (Gathering)
  Este proceso es nuestra primera fase, como su nombre lo indica, es la recolección de toda la información pública de nuestro objetivo, ¿qué quiere decir toda la información pública?, todo tipo de información que pueda están en servidores públicos, páginas web, blogs, redes sociales y en nuestro caso la obtención de información de páginas de fuentes abiertas colombianas como lo son la página de datos abiertos de Colombia, la página de registraduría Nacional del Estado Civil, etc. Todo esto con el fin de correlacionar esta información encontrada, y también aprender a utilizar las herramientas de inteligencia, aprender como es el proceso con el cual trae toda la información.

- El procesamiento de la información (Processing)
  ¿Qué es lo que se quiere hacer con toda esa información? Lo primero es eliminar información duplicada para que la correlación que se realiza sea fácil y poder disminuir información que puede ser útil o no. Lo segundo que se procede a hacer en este proceso es la eliminación de falsos positivos dado que la información que se encuentre no esté completamente relacionada con nuestro objetivo dado que nuestro objetivo pueda tener algo en común con

otra persona, empresa, etc. y esto hace que la información encontrada sea incierta.

Otro caso que se realiza en este proceso es el análisis de datos no estructurado, dado que la información en el internet no está toda completamente estructurada, la herramienta de inteligencia tiene como tarea poder interpretar esta información no estructurada tratarla y exponerla al analista de inteligencia, de forma que sea fácil y entendible para él, que sirva de ayuda para mejorar el análisis que se está realizando. Y, por último, pero no menos importante tenemos el análisis de sentimiento de la información pública (comentarios en redes sociales, blogs, etc.) que nuestro objetivo quiere transmitir a sus lectores frente a un tema en específico.

- Sacar provecho de la información (Taking advantage)
  Este es el paso final y el paso que le interesa a nuestro analista, una de las formas en la cual se le saca provecho a esta información es la predicción de posibles ataques, esto debido a que, a la forma en la cual una persona o empresa escribe sus comentarios de manera que quiere generar o tener un motivo o debido a los medios que posee para poder realizar un ataque de forma que al final lo que se desea es perfilar al atacante según su capacidad o según su motivación.

Luego de haber especificado la arquitectura que siguen las transformadas es necesario definir algunos conceptos básicos para poder entender cómo consultan y traen la información de todas estas fuentes abiertas, para esto se definirán ¿Qué es una entidad?,¿Que es un servidor semilla?,¿Que es un servidor TAS? etc.

¿Qué es una Entidad?

Una entidad es una representación abstracta de cualquier tipo de información que se encuentra en la vida real, como por ejemplo nombres de usuario, correos, nombres completos, teléfonos, direcciones, redes sociales, etc. Actualmente Maltego ofrece alrededor de 36 tipos de transformadas.

¿Qué es un Seed Server?

Un servidor semilla (seed server) es el encargado de re direccionar las peticiones del usuario al correspondiente servidor TAS, es decir el servidor semilla sabe cuál servidor TAS tiene la transformada asociada, luego de identificarlo este envía la solicitud realizada por el usuario a dicho servidor TAS.

¿Qué es un TAS Server?

TAS (Servidor aplicado de transformadas) es un servidor que contiene las transformadas que son ejecutadas por petición del usuario de Maltego.

Ahora se procede a ser mucho más específicos con el proceso de ejecución de las transformadas en la herramienta de OSINT Maltego como se ilustra en la figura 2:

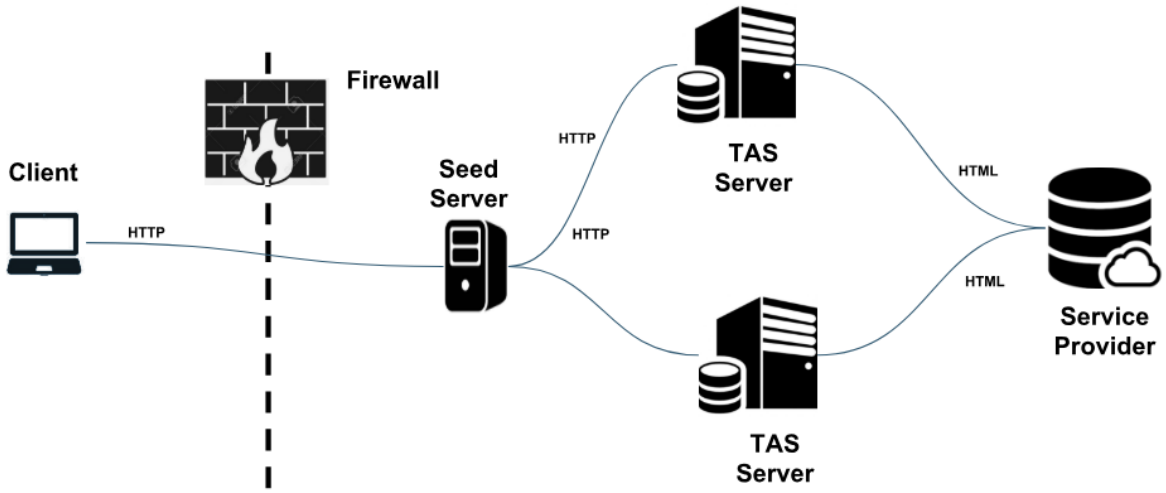


Figure 8 Esquema gráfico del funcionamiento de Maltego, Fuente. Elaboración propia

Para esto el cliente selecciona la entidad que proporciona la información necesaria para que la transformada pueda ejecutarse (Figura 3) que en este caso es la cédula de la persona a la cual quiere obtenerle el lugar de expedición de la cédula, luego de “arrastrar” la entidad al área de trabajo.

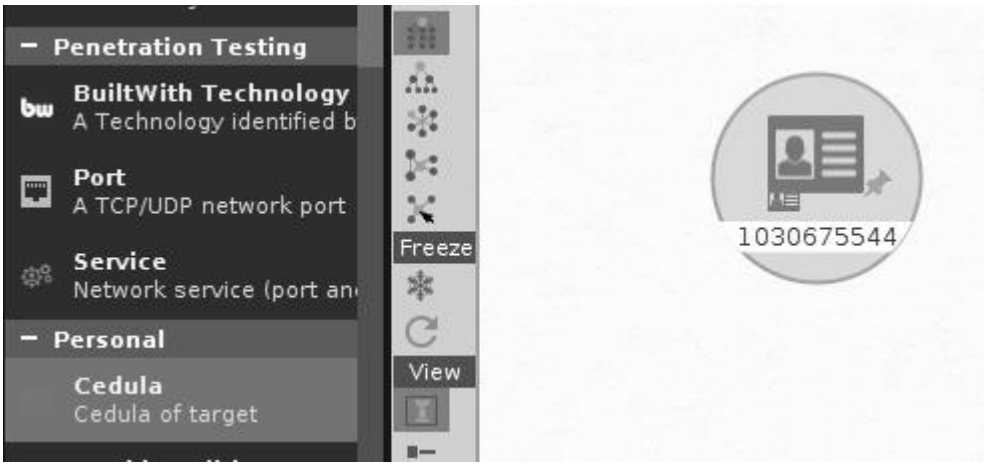


Figure 9 Entidad cédula en Maltego, Fuente. Elaboración propia

Se procede a darle “clic” derecho y se mostrará el listado de posibles transformadas asociadas a dicha entidad (Figura 4), cabe destacar que en esta sección salen tanto las transformadas locales (desarrolladas por los integrantes) y públicas (ofrecidas por Maltego).

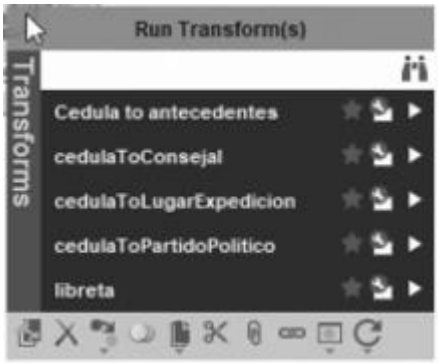


Figure 10 listado de transformadas para la entidad cédula en Maltego, Fuente. Elaboración

En este ejemplo se procederá a ejecutar la transformada de antecedentes judiciales de una persona, se procede a seleccionar el icono en forma de flecha al lado del nombre de la transformada, esto desencadenará el proceso descrito en la figura 2.

El usuario enviará una petición de la transformada seleccionada y la enviará por medio del protocolo HTTP a los servidores semilla como se puede apreciar en la figura 5.

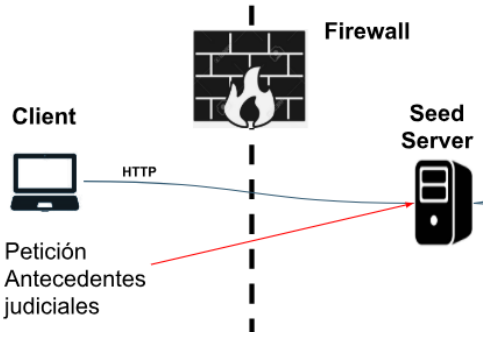


Figure 11 Petición entre el cliente y el servidor semilla de Maltego, Fuente. Elaboración propia

El servidor semilla recibe la petición del usuario y re direcciona al servidor TAS que contiene la transformada encargada de traer la información como se puede observar en la figura 6.

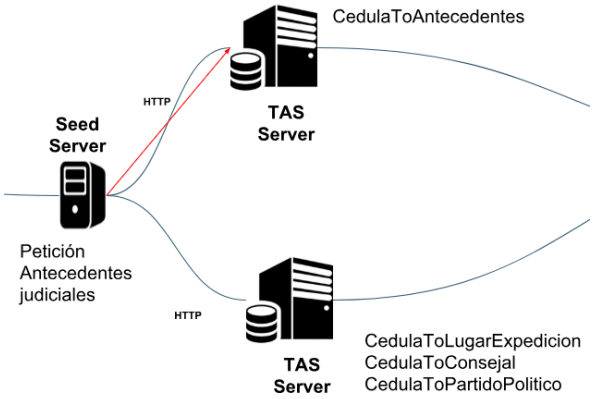


Figure 12 Petición entre el servidor semilla y los servidores TAS de Maltego, Fuente. Elaboración propia

Por último, el servidor TAS ejecuta la transformada asociada a la consulta y trae la información consultando las fuentes abiertas de las entidades del gobierno y retorna la información al usuario en formato XML figura 7 y 8.

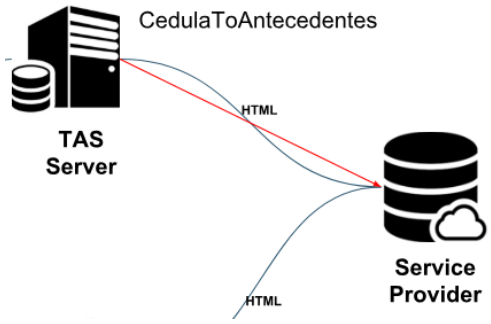


Figure 13 Servidor TAS ejecutando la transformada extrayendo información con Maltego

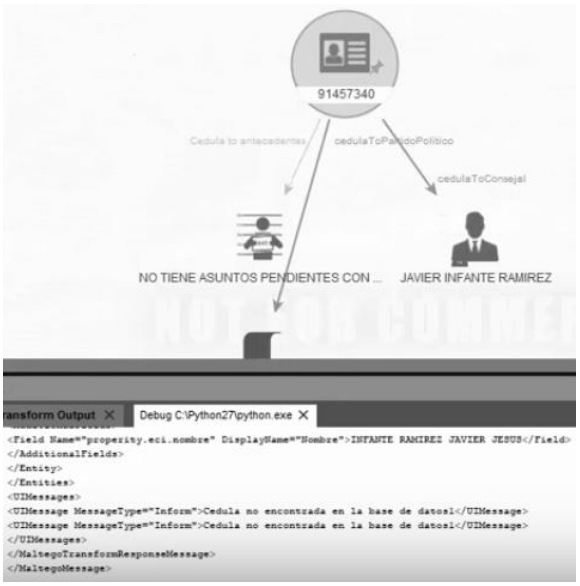


Figure 14 Petición entre el servidor semilla y los servidores Maltego, Fuente. Elaboración propia

Este proceso es realizado cada vez que se realiza la ejecución de una transformada por parte de un usuario en “Maltego”, en la figura 9 puede verse una posible representación de un grafo construido a partir de una entidad “correo electrónico”, pudiendo ver claramente las relaciones que existen entre la información obtenida desde las fuentes abiertas colombianas.

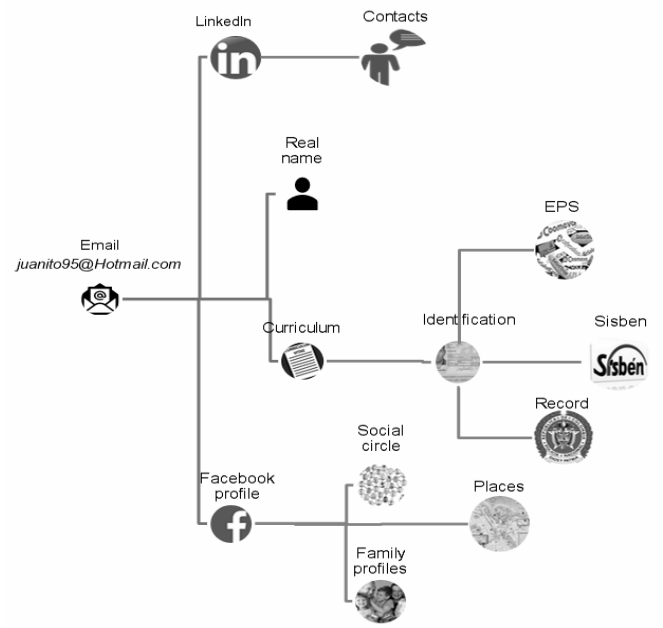


Figure 15 Representación gráfica de una Transformada y una Entidad en Maltego

7 Análisis y diseño de modelos de análisis de sentimientos

7.1 El proceso de construcción de los modelos de análisis de sentimientos

En este primer caso se implementó el proceso Bayes Naive para el análisis de sentimiento a frases escritas por un posible atacante, con esto se quiere determinar la postura que tiene frente a un tema en específico, a continuación, se explicará el modelo y la serie de pasos necesarios para su respectiva clasificación.

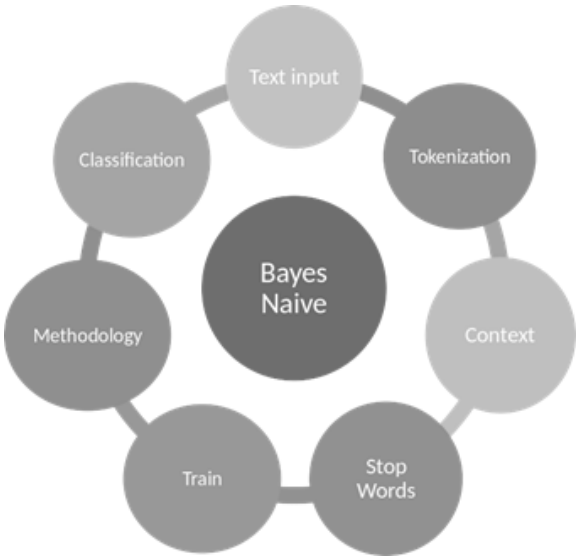


Figure 16 Proceso Bayes Naïve para análisis de sentimiento.

Este proceso consta de 6 pasos los cuales son necesarios para el análisis de sentimiento, los pasos son los siguientes:

Tokenización

En este paso se procede a “filtrar” la frase la cual pasa por tres importantes pasos:

- Eliminación de caracteres, se procede a eliminar cualquier carácter que no sea importante y que no modifique el sentimiento de la frase, caracteres como por ejemplo @ {} []?¡¨*?¡!” #!&%$¬||+*-
- En este momento no se contempla frases en un contexto de preguntas, ironía o exclamación.
- Mayúsculas a minúsculas, en este proceso se procede a pasar todas las palabras de la frase que estén en mayúsculas a minúsculas, con el fin de evitar tener palabras repetidas que solo varían por este factor, por ejemplo, la palabra “malo” y “Malo” tienen el mismo sentimiento por lo que no hace falta mantener las dos palabras de forma individual.
- Separación en una lista, en este paso se procede a separar la frase en una lista para fácil manejo del modelo y entrenamiento.

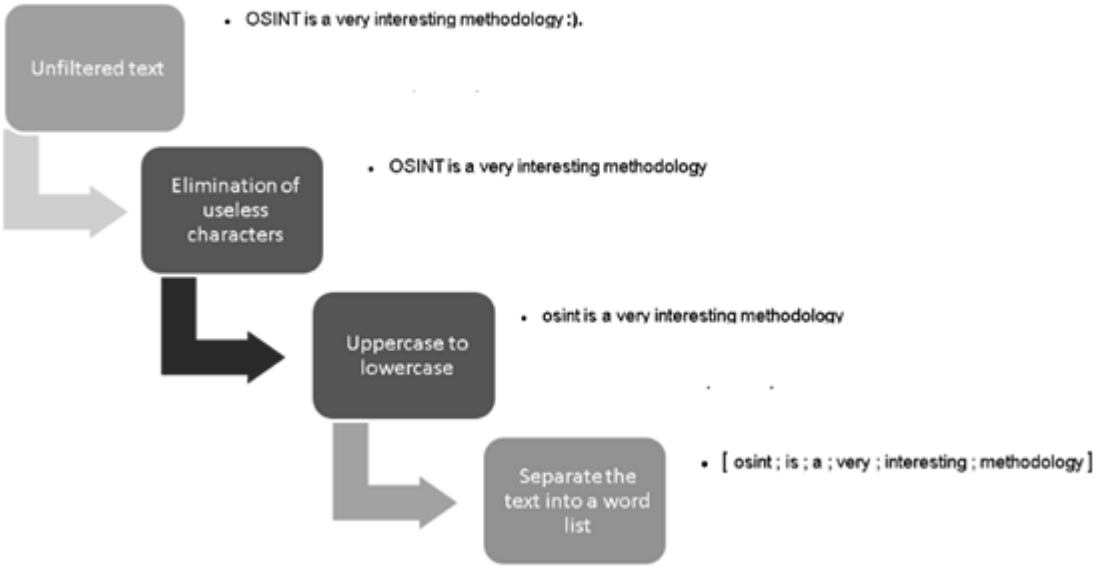


Figure 17 Proceso Bayes Naïve - Tokenización.

Contexto

En este paso se procede a darle contexto a la frase, realizando un análisis por palabras para determinar si las mismas "potencian" a otras palabras, como por ejemplo la palabra "muy" es potenciadora de sentimiento de otras palabras, por lo que es importante tener "muy malo" a tener las palabras individuales.

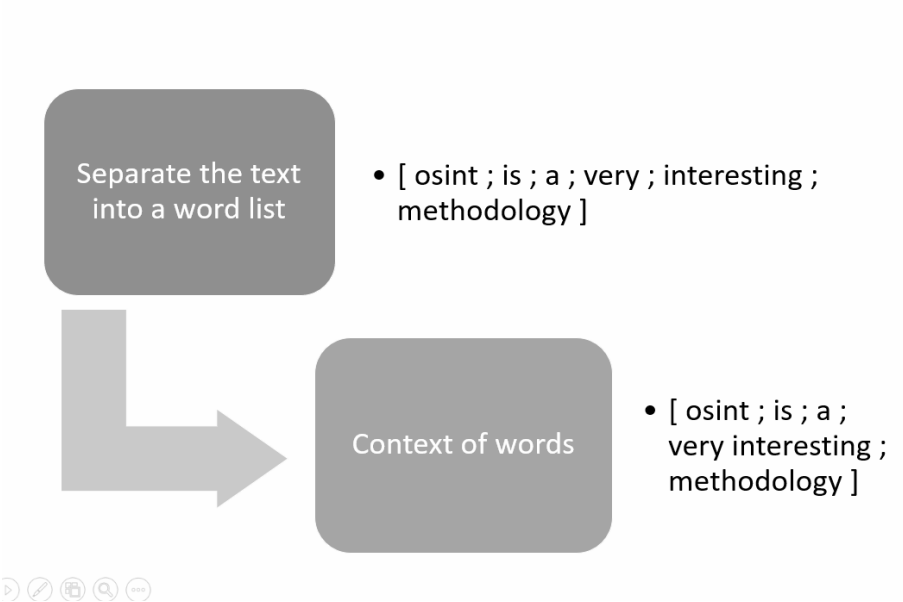


Figure 18 Proceso Bayes Naïve - Contexto.

Stop words

En este paso se procede a eliminar de la lista de palabras anteriormente procesado las palabras llamadas “Stop words”, es decir palabras que no alteran el sentimiento de la frase, tales como: además, donde, algunas, etc.

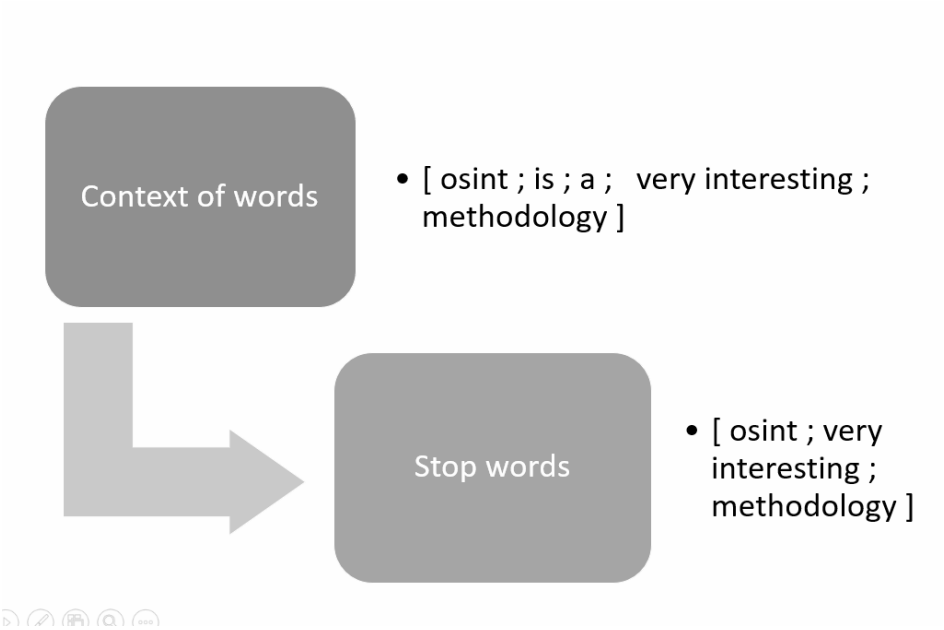


Figure 19 Proceso Bayes Naïve – Stop words.

¿**Training Set**

En este paso se procede a usar la lista de la frase ya procesada y se le asigna una etiqueta, manejando un estilo “matriz”.

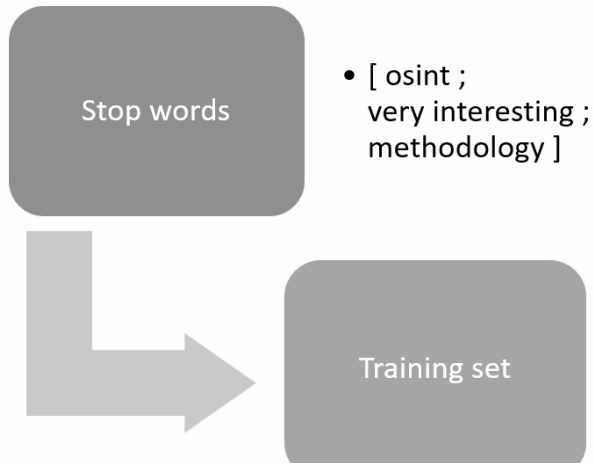


Figure 20 Proceso Bayes Naïve – Training Set.

Metodología

Para calcular el sentimiento de la frase se procede a seleccionar cada palabra de la frase y se procede a contar cuantas veces sale dicha palabra en las frases del training set, por ejemplo se cuenta la palabra “OSINT” en las frases que fueron asignadas como “positivas” o ”negativas”, luego de contar la palabra se procede a calcular la probabilidad asociada a la palabra dividiendo la cantidad de ocurrencias de la palabra en el training set sobre la cantidad total de palabras en el training set (se realiza tanto para “positivo” y “negativo”) y la probabilidad se acumula en la probabilidad total de la frase.

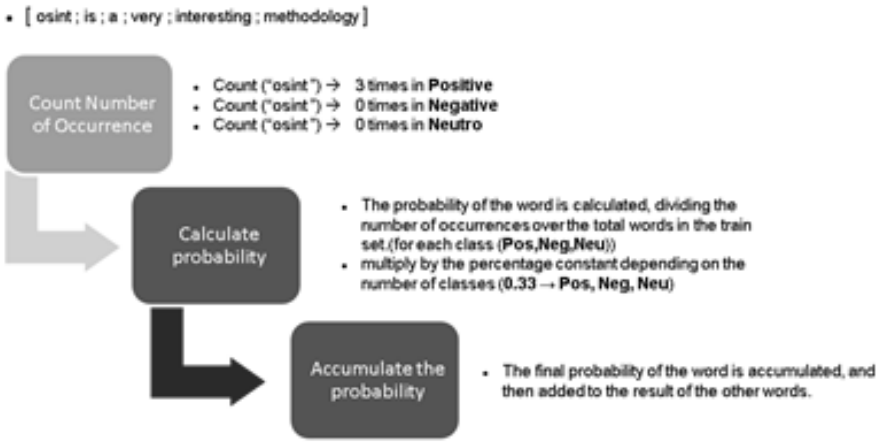


Figure 21 Proceso Bayes Naïve – Metodología.

Clasificación

Después de haber calculado la probabilidad de todas las palabras de la frase, se procede a elegir la probabilidad acumulada máxima (dada por la fórmula de Bayes Naive) por cada clase (positivo, negativo Figura 20-A) y esa será la debida clasificación de la frase.

$$P = \frac{p(C)}{V_{C^n}} \sum_{i=0}^{n} count(d_i, C) \quad (1)$$

Dónde:

• P es la probabilidad final de la frase

• C es la clase de frase que puede ser positiva o negativa

• p (C) es la probabilidad de que una palabra sea parte de una frase positiva o negativa

• count (di, C) es el número de ocurrencias de la palabra di para cada clase C (positiva o negativa)

• Vc es el número total de palabras pertenecientes a la clase C (positiva o negativa) existente en el conjunto de entrenamiento

• n es el número total de palabras en el conjunto de entrenamiento

Fórmula 1. Fórmula Bayes Naive

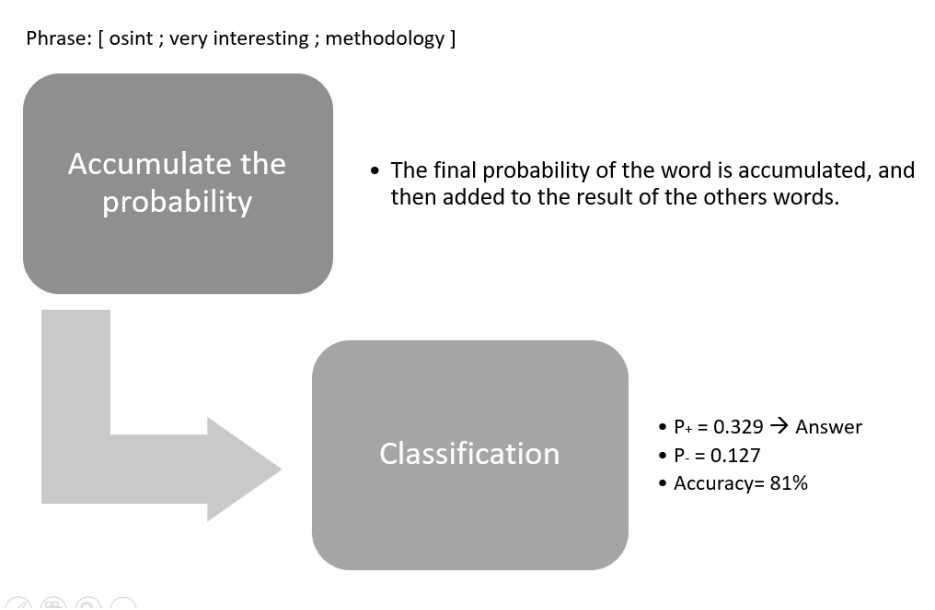


Figure 22 A Proceso Bayes Naïve – Metodología.

Case 2: Modelo 2 SVM

En este segundo caso se implementó el proceso Máquinas de vectores de soporte (SVM del inglés Support Vector Machine) para el análisis de sentimiento a frases escritas por un posible atacante, con esto se quiere determinar la postura que tiene frente a un tema en específico, a continuación, se explicará el modelo y la serie de pasos necesarios para su respectiva clasificación.

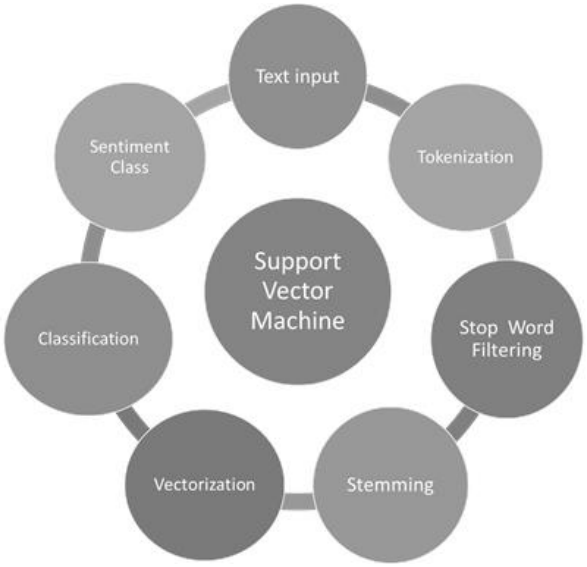


Figure 23 Proceso SVM para análisis de sentimiento

Este proceso consta de 6 pasos los cuales son necesarios para el análisis de sentimiento y los pasos son los siguientes:

Tokenización

En este paso lo que se busca es limpiar la frase que nos ingresa, se realiza de la siguiente forma:

- Reemplazar mayúsculas por minúsculas

  En este paso lo que se busca es dejar toda la frase, valga la redundancia en minúsculas para hacer un mejor tratamiento de cada palabra en pasos posteriores. Ejemplo: Hola, ¿Cómo has estado? “Me imaGino que muy Bien.” :) a hola, ¿cómo has estado? “me imagino que muy bien.” :)

- Remover caracteres que no provean información relevante

  En este paso lo que se busca es remover todos los símbolos y signos de puntuación tales como: la coma, punto y coma, punto, dos puntos, puntos suspensivos, signos de interrogación, signos de admiración, paréntesis, guion y comillas que no me dan información al momento de realizar un análisis de sentimientos. Ejemplo: hola, ¿cómo has estado? “me imagino que muy bien.” :) a hola cómo has estado me imagino que muy bien.

- Separar la frase en una lista de palabras

Para hacerle la frase fácil de entender al modelo lo que se realiza es cambiar su representación, es decir, pasar de una frase a una representación de una lista. Ejemplo: hola cómo has estado me imagino que muy bien a [hola; cómo; has; estado; me; imagino; que; muy; bien;]

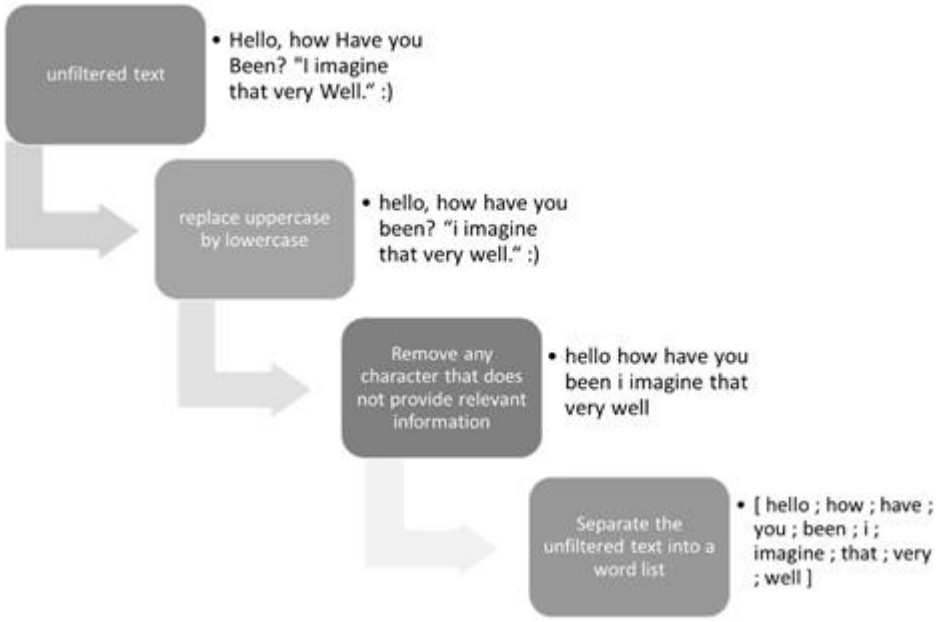


Figure 24 Proceso SVM - Tokenización.

Filtro de Stop Words

En este paso se procede a eliminar de la lista de palabras anteriormente las palabras llamadas "Stop words", es decir palabras que no generan un sentimiento por si solas, tales como: un, una, unas, unos, uno, sobre, todo, también, tras, otro, algún, etc. Ejemplo: [hola; cómo; has; estado; me; imagino; que; muy; bien;] a [hola; imagino; muy; bien;]

Figure 25 Proceso SVM– Stop Words.

Stemming (raíz de la palabra)

Para este proceso se empieza a reducir una palabra a su raíz o (en inglés) a un stem, con el fin de que al modelo le resulte fácil entrenarse y calcular el sentimiento, tratando de buscar el sentimiento de una palabra que tiene un mismo sentimiento en plural o en singular. Ejemplo: [hola; imagino; muy; bien;] a [hol; imagin; muy; bien;]

Figure 26 Proceso SVM– Stemming.

Vectorización

Lo primero que se realiza es cambiar nuevamente de representación, ahora lo hacemos de la forma de una matriz donde las columnas son las palabras en su raíz, y la fila es el número de ocurrencias que esa palabra aparece en la frase. Ejemplo:

| hol | imagin | muy | bien |

| 1 | 1 | 1 | 1 |

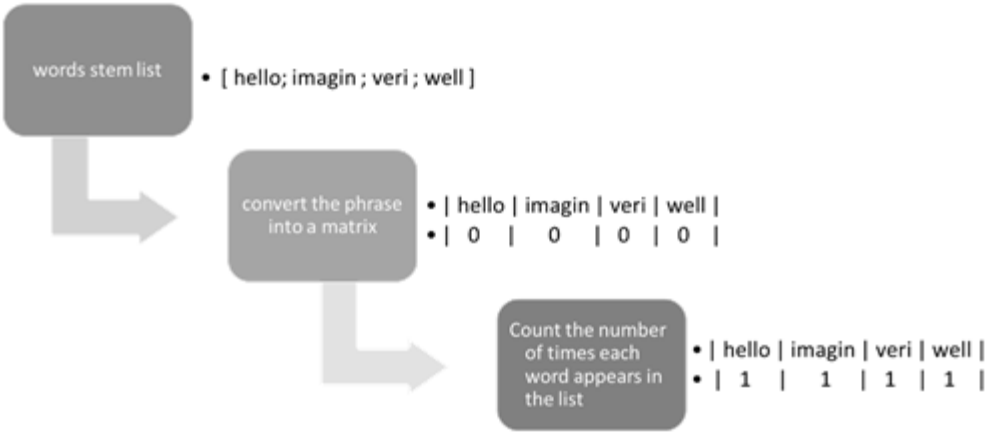


Figure 27 Proceso SVM– Vectorización.

Clasificación

El proceso de clasificación es un poco más complejo de entender, para darse la idea es clasificar nuestra frase vectorizada en un Multi-plano el cual está dividido por un único plano que divide entre frases negativas y frases positivas previamente etiquetadas y vectorizadas para que el SVM sepa qué tipo de frase aproximadamente es la que le ingresan.

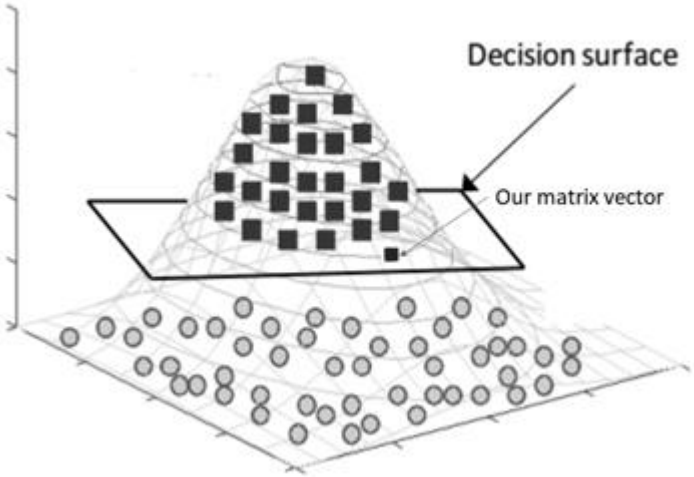


Figure 28 Proceso SVM– Representación gráfica del multi-plano.

El tipo de plano que separa entre positivo o negativo es escogido según las siguientes ecuaciones donde Xi y Xj representan puntos que definen del plano que separa y clasifica los diferentes vectores en el plano (decisión Surface, figura 24):

$$K(X_i, X_j) = \begin{cases} X_i \times X_j & Linear \\ (\gamma X_i \times X_j + C)^{d} & Polynomial \\ \exp\left(-\gamma (X_i - X_j)^2\right) & RBF \\ \tanh(\gamma X_i \times X_j + C) & Sigmoid \end{cases} \quad (2)$$

Donde:

- C es la constante de capacidad
- Núcleo ϕ se utiliza para transformar datos de entrada en el espacio de funciones
- γ y ɗ representan parámetros para la gestión de datos no separables

Fórmula 2. Fórmula Maquinas de Vectores de Soporte

Clasificación de sentimiento

Una vez clasificada nuestra frase, a máquina de vectores de soporte puede determinar el sentimiento de la frase según su localización espacial en el multi-plano, se retorna en este caso un sentimiento ya sea negativo o positivo, para nuestro ejemplo fue sentimiento positivo con un porcentaje de 85% según el modelo entrenado.

Modelo 3 Bernoulli

Por último, se va a mostrar el procedimiento de análisis de sentimiento con el modelo de Machine Learning de Bernoulli paso por paso, pero omitiendo procesos explicados anteriormente:

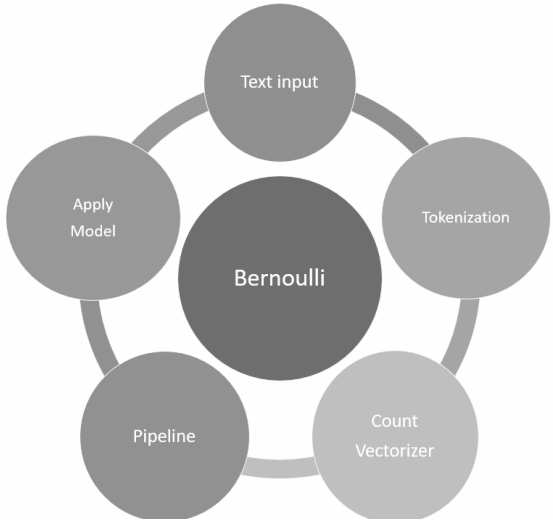


Figure 29 Proceso Bernoulli para análisis de sentimiento.

Tokenización

Es el proceso mediante el cual se hace el pre procesamiento de una frase para ser introducida al modelo, en este proceso se hace una limpieza de signos de puntuación y palabras que no aportan sentimiento además de extraer las raíces de las palabras. A continuación, un ejemplo:

- [El, gatito, está, muy, feliz, el, día, de, hoy] (Entrada normal convertida en arreglo)

- [el, gatito, esta, muy, feliz, el, dia, de, hoy] (Eliminar signos de puntuación, tildes)
- [gatito, muy, feliz, dia, hoy] (Eliminar palabras que no tienen sentido propio)
- [gato, muy, feliz, dia, hoy] (Extraer raíz)

Count Vectorizer

El count vectorizer es un elemento que anteriormente se ha mencionado y a grandes rasgos es un objeto que contabiliza las apariciones de una palabra dentro de tweets positivos y negativos del training set para determinar la probabilidad de ser positiva o negativa.

Pipeline

El Objeto Pipeline es un contenedor dentro del cual se asignan varios parámetros, los cuales son, el count vectorizer que ya se tiene creado y almacenado, la función de Tokenización y el modelo según el que se va a entrenar, en este caso Bernoulli.

La finalidad del objeto Pipeline es poderlo serializar y guardar como archivo para que, una vez entrenado el modelo con el training set y con la tokenización realizada no sea necesario repetir este proceso cada vez que se va a clasificar una frase.

Aplicar modelo

Concretamente el modelo de Machine Learning de Bernoulli es de tipo binomial y en él, se tiene la siguiente estructura:

- Contabilizar apariciones
- Calcular probabilidad

A continuación, se presenta un ejemplo para entender mejor el concepto.

Training set:

Frases	Sentimiento
Chinese Beijing Chinese	Positivo
Chinese Chinese Shangai	Positivo

Chinese Macao	Positivo
Tokyo Japan Chinese	Negativo

Tabla 6 y Training set - Ejemplo

Test set:

Frases	**Sentimiento**
Chinese Chinese Chinese Tokyo Japan	**¿?**

Tabla 7 Test set - Ejemplo

Tabla 4. Test set y Training set - Ejemplo

La fórmula que permite calcular la probabilidad es la siguiente:

$$P(t|c) = (Tct + 1) / (\sum ct + 2)$$

Dónde:
• Tct es el número de ocurrencias de t en el conjunto de entrenamiento
• ct es la cantidad de elementos que tiene la clase C dentro del conjunto de entrenamiento.

Fórmula 3. Fórmula Bernoulli

A continuación, la aplicación de la fórmula para entenderla mejor:

Se tendrá que la clase c es positivo y la clase ĉ es negativo.

P(Chinese|c) = (5+1) / (3+2) =4/5

P(Japan | c) =P(Tokyo | c) = (0+1) / (3+2) =1/5

P(Beijing | c) =P(Macao | c) =P(Shangai | c) = (1+1) / (3+2) =2/5

P(Chinese|ĉ) = (1+1) / (1+2) =2/3

P(Japan | ĉ) =P(Tokyo | ĉ) = (1+1) / (1+2) =2/3

P(Beijing | ĉ) =P(Macao | ĉ) =P(Shangai | ĉ) = (0+1) / (1+2) =1/5

(La ĉ simboliza la clase opuesta a c)

Una vez evaluadas, se procede a calcular el total, es decir, la probabilidad del caso que se quiere clasificar (Chinese Japan Tokyo) con Chinese solo una vez porque este modelo no tiene en cuenta cantidad de apariciones por su condición binomial:

Para la clase c:

P(c | test) =P(c)×P(Chinese | c) ×P(Japan | C) ×P(Tokyo | C) ×(1-P(Beijing | C)) ×(1-P(Shangai | C)) ×(1-P(Macao | C))

P(cetes)=0.005

Para la clase ĉ:

P(ĉ | test) =P(ĉ)×P(Chinese | cĉ) ×P(Japan | ĉ) ×P(Tokyo | ĉ) ×(1-P(Beijing | ĉ)) ×(1-P(Shangai | ĉ)) ×(1-P(Macao | ĉ))

P(ĉ|test) =0.022

Como la probabilidad con la clase ĉ es mayor, se clasifica como negativo.

8 Prototipos desarrollados

1. El primer prototipo desarrollado fue el conjunto de transformadas aplicadas a un contexto colombiano para una herramienta de inteligencia de fuentes abiertas, para esto se realizaron 35 transformadas que traen información de las distintas fuentes abiertas de Colombia (Tabla 8), dichas transformadas se integraron a la herramienta de OSINT Maltego con el fin de ofrecer una opción totalmente adaptada a nuestro país para el desarrollo de productos de ciber inteligencia.

Fuente de información colombiana	Servicio consultado	Nombre de la Transformada	Entrada / Salida	Ejemplo
Policía Nacional de Colombia	Verificación de antecedentes judiciales	Cedula to antecedentes	Entrada -> Cédula	Entrada -> 1049564433
			Salida -> Antecedentes judiciales	Salida -> NO TIENE ASUNTOS PENDIENTES CON LAS AUTORIDADES JUDICIALES
Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales (SISBEN)	Afiliación y verificación de puntaje	Cedula to sisben	Entrada -> Cédula	Entrada -> 1049564433
			Salida -> Estrato y puntaje de SISBEN	Salida -> Datos demográficos e información relacionada con el SISBEN (Puntaje: EN VERFICACIÓN)
Ejército Nacional de Colombia	Verificación de servicio militar	Cedula to libreta	Entrada -> Cédula	Entrada -> 1045678866
			Salida -> Situación militar	Salida -> libreta militar de Segunda Clase
	Información asociada con la	Cedula to Lugar Expedición	Entrada -> Cédula	Entrada -> 1045654433

Registraduría Nacional del Estado Civil	cédula colombiana		Salida -> Puesto de votación	Salida -> Bogotá Distrito Capital
Procuraduría General de la Nación	Verificación de antecedentes judiciales	Cedula to antecedentes procuraduría	Entrada -> Cédula	Entrada -> 1030456655
			Salida -> Antecedentes judiciales procuraduría	Salida -> NO POSEE DELITOS FRENTE A LA PROCURADURÍA
Datos Abiertos	Listado de concejales del municipio de San Andrés Santander	sanAndresCedulaToPartidoPolitico	Entrada -> Cédula	Entrada -> 91457084
			Salida -> partido político	Salida -> PARTIDO CONSERVADOR
		sanAndresCedulaToLocationConsejal	Entrada -> Cédula	Entrada -> 91457084
			Salida ->localización	Salida -> SAN ANDRES - SANTANDER
		sanAndresCedulaToPersonConsejal	Entrada -> Cédula	Entrada -> 91457084
			Salida ->Nombre, partido político, localización	Salida -> ISMAEL SIERRA BERMUDEZ, PARTIDO CONSERVADOR, SAN ANDRES - SANTANDER

		San Andrés Cedula to concejal	Entrada -> Cédula	Entrada -> 91457084
			Salida ->Nombre	Salida -> ISMAEL SIERRA BERMUDEZ
	Listado de concejales del municipio de coromoro	coromoroCorreoToDireccionPerson	Entrada ->Correo	Entrada -> antoninopinca@gmail.com
			Output -> Nombre, celular, partido político,	Output -> PINZON CACERES, 32056664553, CONSERVADOR
		coromoroTelefonoToCorreoDireccionPerson	Entrada ->teléfono	Entrada -> 32056664553
			Output -> Nombre, partido político, correo	Output -> PINZON CACERES, antoninopinca@gmail.com, CONSERVADOR
	Secretarios De Gobierno Municipios Del Valle Del Cauca 2016-2019	caucaTelefonoToCorreoDireccionPerson	Entrada ->teléfono	Entrada ->3148299098
			Output -> Nombre, localización, correo	Output -> ANA LUCIA GIRALDO, VERSALLES COLOMBIA, secgobierno@versalles-valle.gov.co

		caucaPersonToCorreoDireccionTelefono	Entrada -> person	Entrada ->ANA LUCIA GIRALDO
			Salida -> Correo, teléfono, Dirección	Output -> secgobierno@versalles-valle.gov.co, 3148299098, VERSALLES COLOMBIA
		caucaCorreoToPersonDireccionTelefono	Entrada -> Email	Entrada -> secgobierno@versalles-valle.gov.co
			Salida -> Celular, Localización , nombre	Salida -> 3148299098, VERSALLES COLOMBIA , ANA LUCIA GIRALDO
	Gabinete Gobernación de Antioquia	Nombre to cargo Antioquia	Entrada -> Nombre	Entrada -> JAVIER MAURICIO GARCÍA QUIROZ
			Salida -> Cargo	Salida -> Secretario de Educación
	INSCRIPCIÓN RECURSO HUMANO MANTENIMIENTO EQUIPOS BIOMÉDICOS	Cedula to nombre biomédicos	Entrada -> Cedula	Entrada -> 1098708750
			Salida -> Nombre	Salida -> JEFFERSSON ELIECER
	DISCAPACITADOS TEORAMA	Cedula to discapacidad	Entrada -> Cédula	Entrada -> 5,519,399

			Salida -> Discapacidad	Salida -> FISICO
	Presidentes de Juntas De Acción Comunal 2016-2019	telefonoToNombreDireccion	Entrada -> Teléfono	Entrada -> 3112351493
			Salida -> Nombre, ubicación	Salida -> ROSALVA DEVIA NARVAEZ, CLL 66 2W 27
Registro Único De Afiliados	Datos referentes a problemas de salud, pensiones, subsidios familiares, riesgos en términos de desempeño profesional, despidos, entre otros	Cedula to RUAF	Entrada -> Cédula	Entrada -> 1049564433
			Salida -> eps, pensión, subsubsidios, riesgos profesionales	Salida -> ALREDO GUTIERREZ ,PENSIONES PROVENIR, EPS FAMISANAR.
Sync.me	Verificar los números de teléfono celular	Celular to información	Entrada -> celular	Entrada -> 3245664556
			Salida -> Nombre, ubicación, operador	Salida -> ALREDO GUTIERREZ, Bogota,TECEL.
SECOP	Buscar multas, sanciones o	Cedula to Secop	Entrada -> Cédula	Entrada -> 1049564433

	inhabilidades por ID o no		Salida -> bienes, sanciones o inhabilidades	Salida -> NO POSEE MULTAS NI SANCIONES.
Registro Único Empresarial y Social	Razón social de una empresa	NIT to razón social	Entrada -> NIT	Entrada -> 860034811
			Salida ->Razón social	Salida ->Escuela Colombiana de Ingeniería
	Contratos públicos realizados por una entidad	NIT to contratos	Entrada -> NIT	Entrada -> 860034811
			Salida -> nombre entidad, municipio, seccional, fechas, valor, estado, tipo contratista, nit entidad	Salida -> INVIMA, Bogotá, Secretaría general, 2012/10/16, 3500000, en ejecución, 00000830000167 - 2
	Actualizaciones, renovaciones o	Nit to noticias	Entrada -> NIT	Entrada -> 860034811

	terminaciones de la cuenta registrada en Cámara de comercio		Salida ->fecha, operación, noticia	Salida -> 2018/04/12 - 18:49:54, RENOVACIÓN, Renovó su inscripción en el registro de los proponentes clasificándose como
SECOP	Buscar contratos con entidades públicas por ID o no	Cedula to contratos	Entrada -> Cédula	Entrada -> 1049564433
			Salida -> Toda la información sobre contratos públicos	Salida -> POSEE CONTRATOS CON LA REGISTRADURÍA Y EL INVIMA
SECOP	Traer los procesos de contratación en los que una persona compite u obtiene un contrato del nombre.	Nombre to contratos	Entrada -> Nombre	Entrada -> ANDRÉS PRIETO
			Salida -> Información contractual	Salida -> CONTRATOS CON IBBARA PRODUCCIONES, ALIMENTOS SAS, INDUSTRIA TEXTIL DE COLOMBIA.
Google Places API	Extraer información sobre determinado lugar como su dirección	nombreOrganizacion to google places	Entrada -> Lugar, nombre de la compañía	Entrada -> Universidad Nacional De Colombia

	exacta, nombre de la entidad, etc…		Salida -> Información de google maps	Salida ->puntaje, horarios, dirección localización, reseñas.
Administradora de los Recursos del Sistema General de Seguridad Social en Salud.	información sobre afiliados del Régimen Contributivo y el Régimen Subsidiado	Cedula to ADRES	Entrada -> Cédula	Entrada -> 1049564433
			Salida -> estado, entidad, régimen, fecha de afiliación efectiva, fecha de finalización de afiliación, tipo de afiliado	Salida -> ACTIVO, FAMISANAR E.P.S. LTDA - CAFAM – COLSUBSIDIO, CONTRIBUTIVO, 07/04/2005, 31/12/2999, COTIZANTE

Table 8 Lista de transformadas en el contexto colombiano

Modelo	Precisión	Descripción
Bayes Naive	59%	El algoritmo de clasificación Naive Bayes es un clasificador probabilístico. Se basa en modelos de probabilidades que incorporan fuertes suposiciones de independencia. Las suposiciones de independencia a menos no tienen ningún efecto sobre la realidad. Por lo tanto, se consideran ingenuas (naive en inglés). La metodología bayesiana especifica un modelo de probabilidad que contiene algún tipo conocimiento previo acerca de un parámetro investigativo, de este modo se acondiciona al modelo de probabilidad para realizar el ajuste de los supuestos.[1]
		El modelo de SVM lineal implementado se basa en la búsqueda de un hiperplano que separe de forma óptima a las frases

[1] Lesley Ofelia Mesa Páez, Miller Rivera Lozano, Jesús Andrés Romero Davila. (2011). Descripción general de la Inferencia Bayesiana y sus aplicaciones en los procesos de gestión. *La simulación al servicio de la academia*, Retrieved May 29, 2018, from http://www.urosario.edu.co/Administracion/documentos/investigacion/laboratorio/miller_2_2.pdf.

Máquinas de vectores de soporte	85%	de un sentimiento de la de otro, que eventualmente han podido ser previamente proyectados a un espacio de dimensionalidad superior. La característica fundamental es buscar el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo. De esta forma, los puntos del vector que son etiquetados con una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado.[2]
Bernoulli	83%	Es importante entender que todos los modelos presentados hacen referencia a análisis de sentimiento por frase y no por palabra. como se menciona en[3] las etapas de pre procesamiento son implementadas para poner en práctica el modelo de Bernoulli, esos pasos de pre

[2] López-Sarmiento, Danilo A., Manta-Caro, Héctor C., & Vera-Parra, Nelson E. (2013). Clasificador Basado en una Máquina de Vectores de Soporte de Mínimos Cuadrados Frente a un Clasificador por Regresión Logística ante el Reconocimiento de Dígitos Numéricos. *Tecno Lógicas*, (31), 37-51. Retrieved May 29, 2018, from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-77992013000200003&lng=en&tlng=es.

[3]H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, “A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle,” Proc. 50th Annu. Meet. Assoc. Comput. Linguist., no. July, pp. 115–120, 2012.

[3] López-Sarmiento, Danilo A., Manta-Caro, Héctor C., & Vera-Parra, Nelson E. (2013). Clasificador Basado en una Máquina de Vectores de Soporte de Mínimos Cuadrados Frente a un Clasificador por Regresión Logística ante el Reconocimiento de Dígitos Numéricos. *Tecno Lógicas*, (31), 37-51. Retrieved May 29, 2018, from http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-77992013000200003&lng=en&tlng=es.

		procesamiento son tokenización y vectorización, adicionalmente es necesario construir un pipeline que almacene la información del modelo y no recalcular el conjunto de entrenamiento. Concretamente el modelo de Machine Learning de Bernoulli es de tipo binomial y en él, se tiene la siguiente estructura: Contabilizar apariciones Calcular probabilidad.

Tabla 9 Lista de modelos de análisis de sentimientos

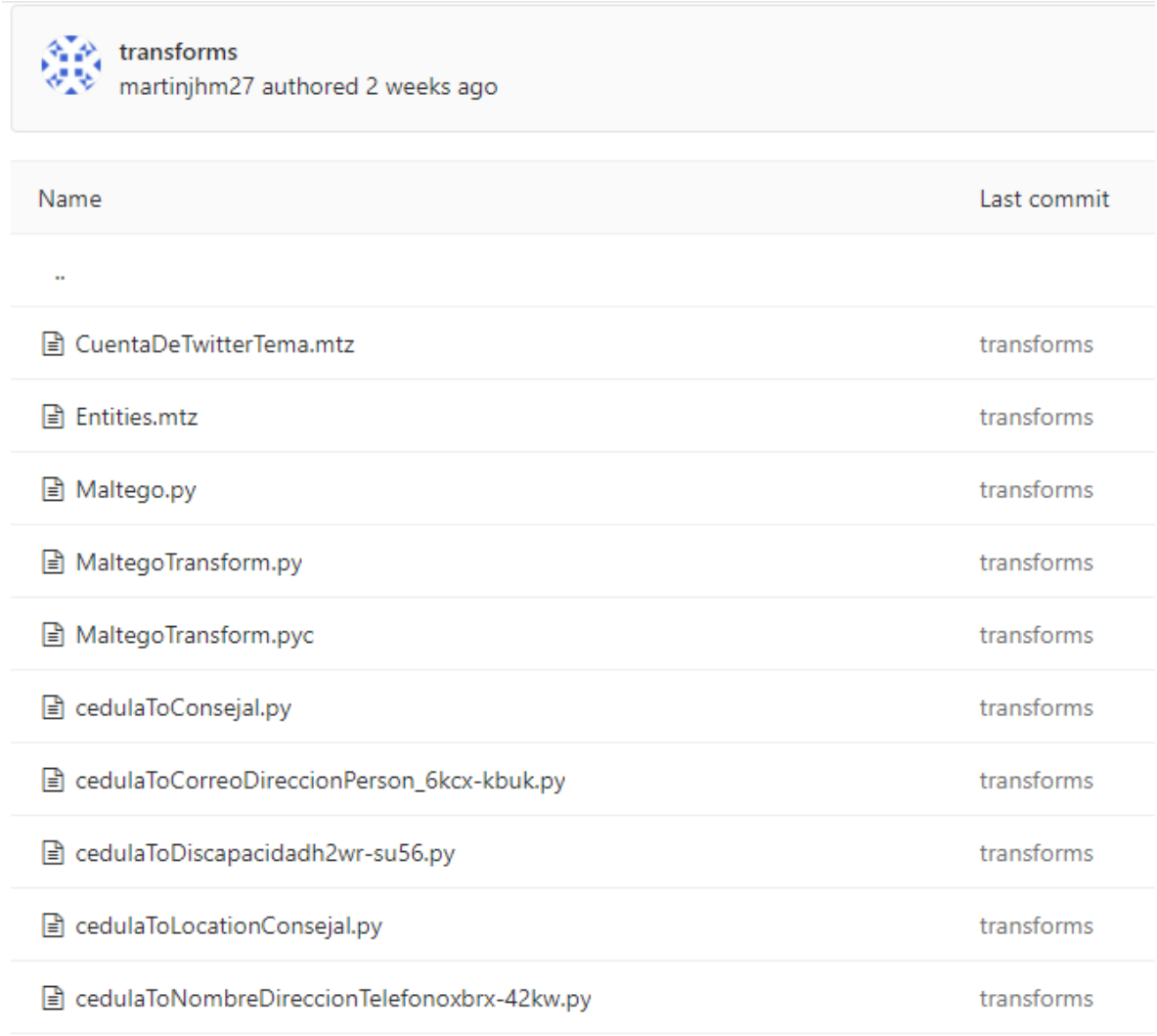


Figure 30 Conjunto de transformadas para un contexto colombiano, https://gitlab.com/ricardopinto08/OSINT

2. El tercer prototipo es la arquitectura de ciberinteligencia conformada por el servidor TRX y el servidor ITDS, las cuales el servidor TRX es un servidor montado con Ubuntu Server y está en producción dentro de las instalaciones físicas del Comando Conjunto Cibernético en el Ministerio de Defensa Nacional y el servidor iTDS es el que brinda Paterva. El servidor TRX sirve para almacenar las transformadas de manera privada y ejecutarlas de manera remota. El usuario ingresa los datos en la herramienta, los cuales van al servidor ITDS para luego ser ejecutados en el servidor TRX y retornar el resultado a manera de entidad.

- La IP pública del servidor es: 200.122.247.13
- Básicamente hay tres archivos que son de interés. Todos ellos se encuentran en / var / www / TRX /
    - debugTRX_Server.py o TRX.wsgi: Archivos de enrutamiento a cada una de las transformadas almacenadas en el TRX.
    - Biblioteca de transformadas (EJ. CCOCTRANSFORMS.py): Este archivo contiene todas las transformadas desarrolladas definidas en forma de funciones.
    - Biblioteca Maltego (maltego.py): Archivo que permite la interacción entre el maltego (Objetos definidos por XML) y las transformadas.

3. El cuarto prototipo es el extractor de LinkedIn, el cual es una herramienta que surgió a desde la Decanatura de Sistemas de la Escuela Colombiana de Ingeniería Julio Garavito que necesita postular anualmente candidatos a obtener el reconocimiento como graduados destacados de la Escuela. Para ello se desarrolló una herramienta que por medio de una cuenta de LinkedIn recolecta información de contactos como estudios, certificaciones, logros, etc…

9 Logros

- Se adquirió la capacidad de realizar informes de inteligencia basados en información pública de un adversario.
- Se construyó un conjunto de aproximadamente 35 transformadas que traen información desde fuentes de datos abiertas del gobierno:
    a. Policía Nacional
    b. Registraduría
    c. Datos Abiertos de Colombia
    d. Sisben
    e. Runt
    f. Libreta militar
    g. ADRES
- Se implementaron tres modelos de análisis de sentimientos para poder determinar la polaridad de un texto escrito por un adversario, los cuales fueron basados en Bayes Naive, Máquinas de vectores de soporte y Bernoulli.

10 Conclusiones

1. Se puede afirmar entonces que cualquier blog, página web de la empresa, periódicos en línea, redes sociales, foros e incluso bases de datos gratuitas constituyen la mayor parte de estas fuentes de información.
2. La privacidad de los datos públicos de cada blanco nunca se viola, no se violan las leyes colombianas relacionadas con la protección de la información de las personas.
3. OSINT tiene muchos usos, marketing, defensa cibernética, política, trabajo de contratación, etc.
4. El Estado colombiano tiene mucha información pública sobre las personas, esa información debe ser bien utilizada.

11 Trabajos futuros

1. Hacer desarrollo de una solución propia.
2. Integración de más transformadas pero relacionadas con otras herramientas que no están contempladas, como INTEL Techniques.
3. Búsqueda de información pública en la red tor.
4. Análisis de información no estructurada, como documentos.
5. Mejorar la precisión de los modelos descriptivos.
6. Implementación de otros modelos de Machine learning.
7. Desarrollar nuevas funcionalidades similares a las contenidas en las soluciones comerciales de inteligencia.