Sistema de análisis de sentimientos para la identificación de patrones de comportamiento relacionados con la Ciberseguridad utilizando técnicas de Data Mining en la red social Twitter

Aizaga Tamayo, Steven Xavier y Rojas Benítez, Juan Carlos

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Sistemas e Informática

Trabajo de titulación, previo a la obtención del título de Ingeniero en Sistemas e Informática

MSc. Tapia León, Freddy Mauricio

7 de septiembre del 2021

URKUND

Urkund Analysis Result

Analysed Document:	Solo_Tesis_Aizaga_Rojas_V12.docx (D111114346)
Submitted:	8/9/2021 6:43:00 PM
Submitted By:	fmtapia@espe.edu.ec
Significance:	2 %

Sources included in the report:

Terán Francisco_Tarea 5 Revisión de Literatura Escritura.docx (D100622433)
https://offerdos.com/twitter/user/ElixirOP
http://201.159.223.180/bitstream/3317/10404/1/T-UCSG-PRE-ECO-CECO-241.pdf
https://www.researchgate.net/publication/272463313_Sentiment_Analysis_and_Text_Mining_for_Social_Media_Microblogs_using_Open_Source_Tools_An_Empirical_Study
http://repositorio.espe.edu.ec/bitstream/21000/23409/1/T-ESPE-044176.pdf
https://repositorio.uam.es/bitstream/handle/10486/688212/delvalle_contreras_javier_tfg.pdf?sequence=1&isAllowed=y

Instances where selected sources appear:

9

Firmado electrónicamente por:
FREDDY
MAURICIO
TAPIA LEON

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE CIENCAS DE LA COMPUTACIÓN
CARRERA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Sistema de análisis de sentimientos para la identificación de patrones de comportamiento relacionados con la Ciberseguridad utilizando técnicas de Data Mining en la red social Twitter**" fue realizado por los señores ***Aizaga Tamayo Steven Xavier y Rojas Benítez Juan Carlos*** el cual ha sido revisado y analizado en su totalidad por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 24 de septiembre del 2021.

FREDDY
MAURICIO
TAPIA LEON

Ing. Freddy Tapia León MSc.
C.I.: 1714745690

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE CIENCAS DE LA COMPUTACIÓN
CARRERA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

RESPONSABILIDAD DE AUTORÍA

Nosotros, Aizaga Tamayo Steven Xavier y Rojas Benítez Juan Carlos, con cédulas de ciudadanía 1720690021 y 1717222820, respectivamente, declaramos que el contenido, ideas y criterios del trabajo de titulación: **"Sistema de análisis de sentimientos para la identificación de patrones de comportamiento relacionados con la Ciberseguridad utilizando técnicas de data mining en la red social Twitter"** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 7 de septiembre del 2021

STEVEN XAVIER AIZAGA TAMAYO

Aizaga Tamayo Steven Xavier
C.C.:1720690021

JUAN CARLOS ROJAS

Rojas Benítez Juan Carlos
C.C.: 1717222820

DEPARTAMENTO DE CIENCAS DE LA COMPUTACIÓN
CARRERA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

AUTORIZACIÓN DE PUBLICACIÓN

Nosotros, **Aizaga Tamayo Steven Xavier y Rojas Benítez Juan Carlos**, con cédulas de ciudadanía 1720690021 y 1717222820, respectivamente autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: "**Sistema de análisis de sentimientos para la identificación de patrones de comportamiento relacionados con la Ciberseguridad utilizando técnicas de data mining en la red social Twitter**" en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 7 de septiembre del 2021

.....
Aizaga Tamayo Steven Xavier
C.C.:1720690021

.....
Rojas Benítez Juan Carlos
C.C.: 1717222820

DEDICATORIA

Dedico mi tesis a Dios, por darme la paciencia y fortaleza necesaria para que a pesar del tiempo transcurrido nunca perdí la esperanza de poder culminar con una meta que me propuse hace mucho tiempo atrás, a mi padre por su amor, su comprensión, su sacrificio y sobre todo por nunca perder la fe en mí, aunque se fue sin verme culminando esta etapa, pero sé que desde el cielo él está celebrando y saltando como lo hacía cuando estaba conmigo, a mi madre por su apoyo constante para lograr una meta más, y finalmente a mi amada esposa por su apoyo y constancia y mis hermosas hijas Lady, Emilia y mi bebe que son y serán siempre el motor que me impulsa para seguir adelante y a pesar del tiempo lograr culminar una importante meta en mi vida. Los amo.

-Juan Carlos Rojas

DEDICATORIA

Dedico esta tesis a mis padres ya que ellos fueron la razón principal por la que me propuse empezar y culminar la carrera, les agradezco por su apoyo incondicional y las lecciones aprendidas durante el proceso. También le dedico esta tesis a mi hermana, porque de ella aprendí varias cosas que me ayudaron a continuar con mi carrera.

-Steven Aizaga

AGRADECIMIENTO

Un agradecimiento especial a mis profesores que a lo largo de mi carrera han fomentado el trabajo, la responsabilidad y sobre todo la constancia, al Ing. Freddy Tapia por darme la oportunidad de poder realizar este Tesis y por confiar en mi cuando le dije que no me iba a echar para atrás, a mi compañero y amigo Steven Aizaga por darme la mano y apoyarme. Agradezco a todas las personas que estuvieron presentes en mi vida universitaria, ya que, de alguna u otra manera me enseñaron diferentes cosas que me han forjado como persona.

-Juan Carlos Rojas

AGRADECIMIENTO

Un agradecimiento total a mis maestros universitarios que gracias a su enseñanza he logrado avanzar en la carrera y culminarla.

Un agradecimiento especial a Graciela Guerrero por su confianza y enseñanza, a Tatiana Gualotuña por su amistad y por aceptarme como parte de su laboratorio de investigación para permitirme aprender cosas nuevas, a Fernando Solís por su amistad y enseñanzas en Programación 2, a Ramiro Delgado por su amistad y confianza, a Rodrigo Fonseca por su conocimiento y amistad, a Fabian Ordoñez por su paciencia, conocimiento y amistad, a Henry Coral por sus enseñanzas y amistad, a Diego Paz por su amistad, a Edison de la Torre por su conocimiento y amistad, a Edison Lascano por su amistad y enseñanzas.

A mi enamorada y compañera de carrera Vicky por ayudarme siempre que pudo.

A mis compañeros/as y amigos/as de carrera Daniel, Cristina, Jonathan, Luis, David, Yury, Pepo, D4VE, Jean Karlo, Miguel, Bryan, Wlady, Kelly, Jordan, Denis, Bryan Miguel, Aldo, Carlos, Dolménica, gracias por todo lo vivido a lo largo de la carrera.

A mi tutor Freddy Tapia por su amistad, confianza y paciencia, a mi compañero de tesis Juan Rojas por su colaboración y amistad, ya que sin ellos el desarrollo de esta tesis no hubiera sido posible.

-Steven Aizaga

Índice

Capítulo I 16
Introducción 16
Antecedentes 16
Planteamiento del problema 17
Justificación 20
Objetivos 22
Objetivo general 22
Objetivos específicos 22
Alcance 23
Hipótesis 25
Capítulo II 26
Marco Metodológico 26
Estado del arte. 26
Características del estado del arte 34
Metodología de la investigación. 35
Marco teórico 37
Variable Independiente 38
Variable Dependiente 40
Capítulo III 43
Análisis, Diseño e Implementación 43
Análisis de la solución 43
Affin (Normas efectivas para palabras) 43
ConceptNet 45
Algoritmo KNN (k-Nearest Neighbors) 47
Distancia Euclidiana 48
Virus Total (API) 50
Fase de Iniciación 51
Servidor de base de datos 52
Servidor API Rest 52
Aplicación Web 52

Fase de producción 52
Desarrollo del servidor API Rest 53
Desarrollo de la aplicación web 54
Desarrollo de script de recolección de datos en Twitter 56
Desarrollo de script de limpieza de datos en R 58
Desarrollo de los diccionarios basados en Affin para el análisis de tendencias. 59
Fase de estabilización 61
Integración del cliente con el Api Rest 61
Fase de pruebas 62
Pantalla general 62
Pantalla de análisis 62
Modal de tendencias 66
Capítulo IV 67
Pruebas y Resultados 67
Introducción 67
Factores de análisis 67
Tweets únicos por día. 68
Tendencia Ecuador 68
Tendencia Efraín Rúales 69
Tendencia Gratis 70
Tendencia Banco Pichincha 71
Cantidad de URLs por día 72
Tendencia Ecuador 72
Tendencia Efraín Rúales 73
Tendencia Gratis 74
Tendencia Banco Pichincha 75
Análisis KNN y distancia euclidiana 76
Capítulo V 83
Conclusiones, Recomendaciones y Líneas de trabajos futuros 83
Conclusiones 83
Recomendaciones 84

Líneas de trabajo futuro.....84
Bibliografía 85

Índice de Tablas

Tabla 1 *Preguntas de investigación* 24
Tabla 2 *Artículos primarios y palabras clave* 28
Tabla 3 *Estudios primarios*.....30
Tabla 4 *Resultados obtenidos – Tendencia "Banco Pichincha"*.....76
Tabla 5 *Resultados obtenidos al variar el valor de k - Tendencia "Banco Pichincha"* 77
Tabla 6 *Resultados importantes obtenidos mediante Knn - Tendencia "Banco Pichincha"* 78
Tabla 7 *Resultados obtenidos mediante distancia euclidiana - Tendencia "Pastebin"* 79
Tabla 8 *Resultados obtenidos al variar el valor de k - Tendencia "Pastebin"*.....80
Tabla 9 *Resultados aplicados KNN – Tendencia "Pastebin"* 81

Índice de Figuras

Figura 1 *Diagrama de Ishikawa sobre ataques*.....20
Figura 2 *Ciclos de la Metodología Design Science Research*.....37
Figura 3 *Red de categorías de las variables*.....38
Figura 4 *Ejemplo ConceptNet*.....47
Figura 5 *Representación Distancia Euclidiana*.....49
Figura 6 *Representación en el plano cartesiano de los puntos de nuestro ejemplo*.....50
Figura 7 *Análisis Virus Total API*.....51
Figura 8 *Implementación de servicios Rest*.....54
Figura 9 *Secciones disponibles* 55
Figura 10 *API Twitter - desarrollador* 56

Figura 11 *Extracto de script para obtener datos de Twitter* 57
Figura 12 *Script de Python* 57
Figura 13 *Conexión MongoDB* 58
Figura 14 *Consumo de Endpoints* 61
Figura 15 *Tendencias recolectadas* 62
Figura 16 *Resultado del Análisis* 63
Figura 17 *Resultados después del análisis.* 64
Figura 18 *Resultado Positivos, negativos y neutros* 65
Figura 19 *Análisis de URL* 66
Figura 20 *Tendencia de Twitter* 66
Figura 21 *Tweets diarios tendencia Ecuador* 68
Figura 22 *Tweets diarios tendencia EfrainRuales* 69
Figura 23 *Tweets diarios tendencia Gratis* 70
Figura 24 *Tweets diarios tendencia Banco Pichincha* 71
Figura 25 *Total, de URL en tendencia "Ecuador"* 72
Figura 26 *Total, de URL en tendencia "EfrainRuales"* 73
Figura 27 *Total, de URL en tendencia "Gratis"* 74
Figura 28 *Total, de URL en tendencia "Gratis"* 75
Figura 29 *Matriz de confusión y resultados de sensibilidad y especificidad Tendencia "Banco Pichincha"* 78
Figura 30 *Matriz de confusión y resultados de sensibilidad y especificidad Tendencia "Pastebin"* 81

Resumen

Con el crecimiento de usuarios en redes sociales que interactúan entre ellos alrededor del mundo, el número de ataques informáticos es elevado lo cual nos permitió plantear este trabajo de titulación con el fin de proponer un sistema de análisis de sentimientos para la identificación de patrones que puedan comprometer la seguridad de la información de las personas, conociendo que actualmente en el Ecuador no existe ninguna organización ni entidad Gubernamental que se encargue de recibir las incidencias ni dar solución a estas es importante tener en cuenta que estos ataques han ido en aumento debido a la pandemia que está en el planeta y al incremento en la utilización de plataformas informáticas para el trabajo, estudio, etc., todo este proceso se realizara mediante la utilización de algoritmos de data mining.

El estudio planteado se enfocará en la red social Twitter ya que es una de las redes sociales más utilizadas por los atacantes ya que es vulnerables en cuanto a control de usuarios y a las publicaciones realizadas, ya que podemos encontrar un amplio contenido, incluido tweets que tratan de robar información de usuarios mediante varias técnicas y enfocadas a personas con poco o nada de conocimiento sobre ciberseguridad o se dejan llevar por el morbo y la curiosidad.

Palabras clave

- **MINERIA DE DATOS**
- **CIBERSEGURIDAD**
- **REDES SOCIALES**
- **TWITTER**
- **ANÁLISIS DE SENTIMIENTOS**

Abstract

With the growth of users in social networks that interact with each other around the world, the number of computer attacks is high, which allowed us to propose this degree work in order to propose a sentiment analysis system for the identification of patterns that may compromise the security of people's information, knowing that currently in Ecuador there is no organization or Government entity that is in charge of receiving incidents or solving them, it is important to bear in mind that these attacks have been increasing due to the pandemic that is on the planet and the increase in the use of computer platforms for work, study, etc., all this process will be carried out through the use of data mining algorithms.

The proposed study will focus on the social network Twitter since it is one of the social networks most used by attackers since it is vulnerable in terms of user control and the publications made, since we can find extensive content, including tweets that deal with of stealing user information through various techniques and focused on people with little or no knowledge about cybersecurity or are carried away by curiosity and curiosity.

Keywords

- **TEXT MINING**
- **CYBERSEGURITY**
- **SOCIAL MEDIA**
- **TWITTER**
- **SENTIMENT ANALYSIS**

Capítulo I
Introducción

Antecedentes

En la última década la tecnología ha avanzado de forma constante y acelerada, haciendo que más personas tengan acceso a internet y en particular a las redes sociales, en enero del 2021 según el análisis mundial de “Hootsuite” publicado en “We are social” (Kemp, 2021b), existen 7,83 billones de usuarios de internet y 4,20 billones usan redes sociales en el mundo, lo cual representa el 53,60% de usuarios que interactúan o tienen cuenta con alguna red social.

La crisis presentada a inicios del 2020 causada por la COVID-19 ha evidenciado la dependencia de la sociedad hacia la tecnología y más aún a la compartición de información de forma masiva, ya sea en redes sociales, portales web, compras en línea, transacciones financieras, actividades educativas, tramites gubernamentales y un sin fin de actividades que necesariamente incurren en compartir algún tipo de información personal o empresarial, por consiguiente, más sensibles a amenazas cibernéticas.

Según (BID & OEA, 2020), dice que los daños económicos causados por ataques cibernéticos pueden sobrepasar el 1% del PIB en algunos países, y alcanzaría el 6% del PIB en casos de ataques a infraestructura critica.

Es una realidad que la pandemia ha marcado un punto de inicio para una transformación digital a gran escala, por lo que ahora somos una sociedad enteramente dependiente de una infraestructura digital y como consecuencia de esto el número de ataques cibernéticos ha incrementado, según (Cisco, 2020) . A nivel mundial, hubo un crecimiento del 776% en ataques cibernéticos de 2018 a 2019, y el número total de ataques DDoS se duplicará de 7,9 millones en 2018 a 15,4 millones en 2023.

Por lo que los problemas de seguridad que actualmente está afrontando el mundo son inmensos ya que con el uso masivo del internet varias empresas, países, y entidades privadas han sufrido algún tipo de ataque informático o su seguridad ha sido vulnerada, por esta condición los ataques cibernéticos se encuentran en la categoría de riesgos "tecnológicos". Donde también aparecen otros peligros, como las consecuencias adversas derivadas de los avances tecnológicos, como son los incidentes relacionados con los fraudes o robo de información, así como la interrupción de redes de información e infraestructuras críticas.

Con el aumento de usuarios en las redes sociales los ciber delincuentes aprovechan la curiosidad y la falta de educación en ciberseguridad para atacar, como lo reporta ESET en su blog We Live Security (Lubeck, 2020) "Una nueva campaña de ingeniería social activa a través de WhatsApp se aprovecha de estos tiempos de confinamiento a raíz de la pandemia del COVID-19 con el objetivo de robar datos personales de los usuarios. Bajo la consigna "Quédate en casa", las potenciales víctimas reciben un mensaje en el que se ofrece un supuesto bono como forma de ayuda, aunque sin hacer referencia a una entidad, empresa u organismo como responsable de esta iniciativa.".

Planteamiento del problema

En Ecuador, según el análisis publicado en Hootsuite, realizado por (Kemp, 2021a), existen 14 millones de usuarios de internet y de redes sociales, es decir el 78,8% de la población ecuatoriana tiene una participación en internet. Además, según un estudio realizado a un segmento de la población por investigadores de la Universidad Católica de Cuenca (Jara- Obregón et al., 2017) nos dice que" un 71% de los encuestados reconoció que sí tienen conocimiento del cometimiento de delitos a través de redes sociales, un 23% reconoció haber sido víctima de amenazas, acosos, extorsión y hacking de sus cuentas de

usuario. El 96% de los que reconocieron haber sido víctimas de estos delitos indicó que no dieron parte a las autoridades porque prefirieron ignorarlo o creyeron que no sería necesario, el restante 4% que sí dio parte a las autoridades y se obtuvieron que un 75% indicaron que no se recibió respuesta inmediata y el 25% restante que denunció el hecho indica que demoró mucho.”.

Es común que campañas de propagación de malware[1] se desplieguen cuando existen sucesos importantes ya que suelen tener más impacto, como se evidencio en Ecuador durante los meses de abril y octubre de 2019, fechas en las que Julian Assange, fundador de Wikileaks deja la Embajada de Ecuador en Londres y el paro de transportistas por el retiro del subsidio a los combustibles, respectivamente.

En el caso de Assange se tuvo la campaña #OpAssange la cual aparte de publicar información privada de sitios gubernamentales mediante “bins de información”, también publicaban exploits[2] dentro de pastebins[3], a lo cual sumamos los estados de ánimo como la euforia y curiosidad que les caracteriza a las personas, lo que da como resultado una mezcla potencialmente dañina lo que puede ocasionar que el usuario ingrese o acceda a sitios en donde su información pueda ser afectada.

Según la publicación de la revista Vice en su sección de tecnología (Franceschi-Bicchierai, 2020) “*Pastebin is one of the most famous websites that allows anyone, even without being registered, to “paste” any kind of text and make it public. Over the years, it became a repository for all kinds of unsavory data, such as the personal details of people*

[1] Malware (software malicioso) es un término que se utiliza para describir cualquier programa o código creado con la intención de dañar una computadora, red o servidor. (Crowdstrike, 2019)

[2] Un exploit es un programa informático, una parte de un software o una secuencia de comandos que se aprovecha de un error o vulnerabilidad para provocar un comportamiento no intencionado o imprevisto en un software, hardware o en cualquier dispositivo electrónico.(PANDASECURITY, 2021)

[3]Pastebin es un servicio o aplicación web que permite publicar o subir texto para que esté visible a todo el mundo.(Adeva, 2021)

who got doxed by hackers, leaked passwords, hacker manifestos, and even malware payloads. Naturally, this meant it was a treasure trove for security researchers investigating data breaches or hunting hackers".

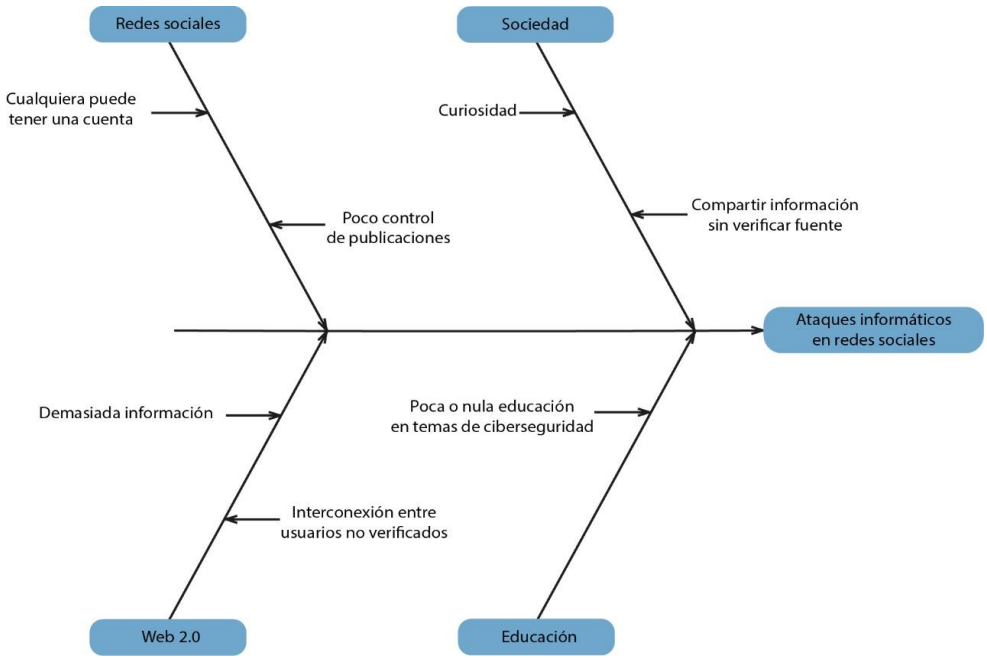
Por lo general estos enlaces de Pastebin se pueden encontrar regados por todo Twitter por lo que con solo realizar una búsqueda podemos obtener información de brechas de datos o encontrarnos con un troyano camuflado en un archivo de texto.

Todo esto se contrasta con el informe de gestión de riesgos y vulnerabilidad de NopSec (Zurkus, 2016), la cual es encargada de proporcionar soluciones automatizadas de medición de control de seguridad de TI, "*Twitter se está convirtiendo en una de las principales plataformas para investigadores y atacantes que buscan difundir exploit de prueba de concepto (POC). Las vulnerabilidades asociadas con malware activo se tuitean nueve veces más que las vulnerabilidades con solo un exploit y 18 veces más que todas las demás vulnerabilidades*". Lo que convierte a Twitter en un el lugar con un índice de existencia de ciberataques muy alto, por lo que un simple tweet es un puente hacia un atacante que puede afectar la vida personal de un individuo o de una empresa, ya sea exponiendo su vida privada a internet, robando su dinero, suplantando su identidad o dañando su imagen.

Todas estas causas hacen que los ataques informáticos en redes sociales sean mucho más fáciles de realizar, todo esto lo vemos reflejado en el diagrama de Ishikawa en la Figura 1.

Figura 1

Diagrama de Ishikawa sobre ataques



Justificación

La Ciberseguridad a nivel mundial se ha convertido en un tema de mucha importancia, ya que han aumentado los riesgos, ciberataques y ciber amenazas.

En el 2019, más de 770 millones de correos electrónicos y 21 millones contraseñas únicas fueron expuestas en Pastebin para después ser alojado en el servicio en la nube MEGA[4]. Se convirtió en la colección individual más grande de robo de credenciales personales en la historia, llamado "Colección # 1", fue descubierta como una pequeña

[4] Es el sucesor del servicio de archivos en la nube Megaupload y Megavideo. (Wikipedia, n.d.)

porción de una mayor fuga de datos de 1 TB, dividido en siete partes y distribuido a través de un esquema de intercambio de datos. (Primicias24.com/tecnología, 2019).

Airbus, el segundo fabricante más grande de aviones comerciales en el mundo sufrió una violación de datos, los cuales dieron a conocer los datos personales de algunos de sus empleados. (Cyber Security Report, 2020).

El aumento del uso de tecnologías como: el Internet de las cosas, Inteligencia Artificial, Big Data, Cloud Computing, entre otras; están generando dos principales riesgos: riesgos de seguridad y de privacidad de los datos, lo que ubica a la ciberseguridad en un nivel de relevancia alto. (Reset, Una Idea Bancolombia).

Según el Índice Global de Ciberseguridad (IGC), de la Unión Internacional de Telecomunicaciones (UIT), de julio del 2020, en el que se mide el compromiso de los Estados frente al tema de seguridad informática, Ecuador se encuentra en el sexto puesto de 19 países de América Latina. (Mintel, 2020).

La Ciberseguridad en el Ecuador es un reto para el 2021, ya que a cada minuto se genera contenidos en redes sociales, correo electrónico y varias aplicaciones que generan datos personales, financieros y privados de alta importancia y de alta seguridad.

El presente proyecto está enfocado en el análisis de sentimientos para la identificación de patrones de comportamiento relacionados con la ciberseguridad utilizando técnicas de Data Mining en la red social Twitter, lo cual nos permitirá tener información más acertada de posibles ataques, vulneración de información en la red Twitter.

Poder realizar un análisis de patrones en la interacción de usuarios en la red social Twitter, nos permite ayudara a conocer cuáles son las intenciones o fines que tienen los usuarios.

El Equipo de Respuesta de Incidentes de Seguridad CSIRT de CEDIA actualmente no cuenta con ningún tipo de aplicación de medición o registro de datos de ataques de ciberseguridad relacionado a redes sociales, con este precedente el presente proyecto servirá para realizar las pruebas y validaciones necesarias de funcionalidad.

Objetivos

Objetivo general

Implementar un sistema de análisis de sentimientos mediante técnicas de minería de datos para la identificación de patrones de comportamiento que puedan afectar la seguridad de los datos de los usuarios de la red social Twitter.

Objetivos específicos

- Realizar una revisión preliminar de literatura para evaluar la factibilidad del proyecto mediante la búsqueda en bases de digitales.
- Identificar la problemática actual sobre incidentes de ciberseguridad en Ecuador, por medio del análisis de la literatura.
- Implementar una estrategia de minería de datos basada en la clasificación de sentimientos para identificar patrones de comportamiento en publicaciones de la red social Twitter que puedan afectar la ciberseguridad.
- Determinar cuáles son los algoritmos adecuados para realizar proceso de minería que se busca, mediante una revisión preliminar de literatura con el objetivo de determinar los posibles patrones de comportamiento.

Alcance

Esta investigación comprende la implementación de un sistema de análisis de sentimientos para la identificación de patrones de comportamiento dentro de la red social Twitter.

Nuestro sistema se basará en el análisis de sentimientos ya que al estar relacionado con la ciberseguridad debemos tomar en cuenta que los ataques de phishing pretenden que la persona crea en algo que no es real, por lo que se utilizan técnicas de psicología social para confundirla o hacerle creer que podría obtener un beneficio al confiar en la información que se le da. Lo cual despertaría sentimientos positivos o negativos en la victima los cuales le harían caer en la trampa.

El tipo de aprendizaje que se va a utilizar es una mezcla del método de los k vecinos más cercanos (KNN; K-Nearest-Neighbors, por sus siglas en inglés) y el algoritmo euclidiano de cálculo de distancia, el cual se basa en un aprendizaje supervisado.

Como caso de estudio se utilizarán hashtags que tengan que ver con ataques de ingeniera social, malware y bins de información. Con el fin de validar el sistema propuesto, se desarrollará un prototipo de plataforma informática que permita a las personas encargadas del CSIRT de CEDIA el realizar consultas sobre posibles ataques hacia instituciones del estado, empresas o personas.

Para delinear de forma adecuada el alcance de la investigación planteada, se proponen varias preguntas de investigación tal como se muestra en la Tabla 1.

Tabla 1

Preguntas de investigación.

Preguntas de investigación
RQ1 ¿Con qué frecuencia se han presentado estudios en torno a la problemática relacionada con ciberseguridad y redes sociales?
RQ2 ¿Existen estudios que utilicen minería de datos para la identificación de patrones de comportamiento ligado a la ciberseguridad?
RQ3 ¿Cuáles son los ataques más comunes relacionados con delitos informáticos en Ecuador?
RQ4 ¿Cuál es la normativa legal vigente que protege a la ciudadanía de ataques relacionados a ciberseguridad?
RQ5 ¿Cuáles son los principales patrones de comportamiento en la propagación de noticias falsas, phishing, spear phishing, enlaces infectados, bins de información?
RQ6 ¿Qué soluciones (método, herramienta, modelo, guía, técnica, framework) proponen los estudios realizados en relación con el tema de investigación propuesto?
RQ7 ¿Que algoritmos son los más utilizados para la búsqueda de patrones con minería de datos?
RQ8 ¿Qué herramientas son las más adecuadas para el análisis de patrones de comportamiento?
RQ9 ¿Qué tan efectivo es el prototipo en comparación con el método tradicional de reporte de incidentes informáticos?
RQ10 ¿Cuántas URLs maliciosas encontramos dentro del análisis realizado a los tweets?

Hipótesis

Un sistema de análisis de sentimientos nos permitirá identificar patrones de comportamiento que creen escenarios propicios para el robo de datos a usuarios de la red social Twitter.

En base a esta hipótesis se identificó las siguientes variables:

Variable independiente: Análisis de sentimientos.

Variable dependiente: Identificar patrones de comportamiento que creen escenarios propicios para el robo de datos a usuarios de la red social Twitter.

Capítulo II
Marco Metodológico

Estado del arte.

Con el objetivo de identificar métodos existentes sobre análisis de patrones, se procedió a realizar una revisión preliminar de literatura, este proceso se lo realiza según las siguientes fases:

I. Motivación
II. Criterios de inclusión y exclusión
III. Selección de artículos para el grupo de control y extracción de palabras claves
IV. Creación y pilotaje de la cadena de búsqueda
V. Selección de artículos primarios
VI. Elaboración del estado del arte.

I. Motivación

Esta revisión preliminar de literatura pretende encontrar trabajos relacionados con la investigación propuesta que nos permitan entender a profundidad los métodos, modelos y fases de desarrollo de la minería de datos y la identificación de patrones, para así poder generar un prototipo acorde a los lineamientos propuestos.

II. Criterios de inclusión y exclusión.

a. Criterios de inclusión

i. Estudios a partir del año 2015, debido al creciente avance de la tecnología y es necesario obtener artículos actualizados.
ii. Estudios cuyo propósito sea el uso de técnicas de minería de datos para encontrar patrones de comportamiento.
iii. Estudios cuyo propósito sea la adquisición de datos para el análisis de sentimientos con el propósito de clasificar los datos obtenidos.
iv. Estudios que propongan soluciones validadas sobre análisis de sentimientos en redes sociales, basado en el análisis de texto o de emoticones.
v. Estudios que muestren el estado actual de la minería de datos y sus usos en la obtención de resultados.
vi. Estudios cuyo propósito sea segmentar en grupos a usuarios potencialmente dañinos.

b. Criterios de exclusión
i. Estudios enfocados en la obtención de datos en redes sociales para el análisis enfocado al marketing.
ii. Estudios que tengan que ver con el análisis o perfilamiento de usuarios por ideologías.

III. Selección de artículos para el grupo de control y extracción de palabras claves

En este paso procedemos a listar los artículos de más relevancia con los cuales vamos a obtener nuestras palabras clave para armar una cadena de búsqueda lo más

específica posible. Los artículos se obtuvieron de Google mediante búsquedas avanzadas para filtrar la información a obtener.

Tras una revisión por parte del equipo de investigación realizamos una validación cruzada de los términos que encontramos y logramos obtener 9 artículos en el grupo de control.

Los artículos seleccionados para formar el grupo de control y palabras claves obtenidas de cada estudio se detallan en la siguiente tabla:

Tabla 2

Artículos primarios y palabras clave.

Código	Título	Cita	Palabras clave
EC1	Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study	(M. M. Choudhary & Choudhary, 2015)	Text Mining, Sentiment Analysis, Open Source, Twitter Data Analysis, Social Data Mining, R Packages.
EC2	Fake News Detection on Social Media: A Data Mining Perspective	(Shu et al., 2017)	Social media, fake news, data mining, model
EC3	Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data	(Feng et al., 2019)	Big Data Analytics (BDA), Data Mining, Data Visualization, Neural Network k, Time Series Forecasting
EC4	Generating Real Time Cyber Situational Awareness Information Through Social Media Data Mining	(Rodriguez & Okamura, 2019)	Data Mining, Machine Learning, Cyber Situational Awareness, Social Media,

Código	Título	Cita	Palabras clave
			Cybersecurity, Text Mining
EC5	Topic Model Based Opinion Mining and Sentiment Analysis	(Vamshi et al., 2018)	Opinion Mining; Sentiment Analysis; Machine Learning; Natural Language Processing; Topic models
EC6	Opinion Mining Using Live Twitter Data	(Aslam et al., 2019)	Twitter, Opinion Mining, Sentiment Analysis, Twitter Data, Polarity
EC7	Cybercrime Profiling: Text mining techniques to detect and predict criminal activities in microblog posts	(Alami & Elbeqqali, 2015)	Cybercrime, Semantic Web, Social Media, Text Analysis, Text Mining, Similarity, NCD Normalized Compression Distance, Profiling, Suspicious Profile
EC8	Sentiment Analysis of Text Reviewing Algorithm using Data Mining	(M. Choudhary & Choudhary, 2018)	sentiment analysis, text mining, opinion mining, online reviews
EC9	A Review on Social Audience Identification on Twitter using Text mining methods	(Dastanwala & Patel, 2016)	Data mining, Text mining, Social media, Twitter, Audience segmentation

IV. Creación y pilotaje de la cadena de búsqueda

Con las palabras claves que se obtuvo en el paso anterior se conformaron varias cadenas de búsqueda de las cuales se obtuvieron varios resultados. Después estos resultados se filtraron creando varias versiones de la cadena de búsqueda.

Todas las cadenas se las realizó con el uso de la herramienta “Advance Search” que ofrece la interfaz Web de la IEEE.

V. Selección de artículos primarios

De los 53 artículos, se revisaron títulos y resúmenes de cada uno de estos artículos, obteniendo los siguientes resultados.

Estos estudios primarios que se seleccionaron se muestran en la siguiente tabla:

Tabla 3

Estudios primarios.

Código	Título	Cita
EP1	"Real-time monitoring of Twitter traffic by using semantic networks"	Bisio, F., Meda, C., Zunino, R., Surlinelli, R., Scillia, E., & Ottaviano, A. (2015). Real-time monitoring of Twitter traffic by using semantic networks. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, 966–969.
EP2	"Weak signals as predictors of real-world phenomena in social	Charitonidis, C., Rashid, A., & Taylor, P. J. (2015). Weak signals as predictors

Código	Título	Cita
	media"	of real-world phenomena in social media. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, 864–871.
EP3	“Sentiment Analysis for Topics based on Interaction Chain Model”	Gu, N., Sun, D. Y., Li, B., & Li, Z. (2016). Sentiment Analysis for Topics based on Interaction Chain Model. Proceedings - 2015 European Intelligence and Security Informatics Conference, EISIC 2015, 133–136.
EP4	“Domain-Oriented Topic Discovery Based on Features Extraction and Topic Clustering”	Lu, X., Zhou, X., Wang, W., Lio, P., & Hui, P. (2020). Domain-oriented topic discovery based on features extraction and topic clustering. IEEE Access, 8, 93648–93662.
EP5	"Geospatial Event Detection by Grouping Emotion Contagion in Social Media"	Lwowski, B., Rad, P., & Choo, K.-K. R. (2018). Geospatial Event Detection by Grouping Emotion Contagion in Social Media. IEEE Transactions on Big Data, 6(1), 159–170.
EP6	“IoCMiner: Automatic Extraction of Indicators of Compromise from Twitter”	Niakanlahiji, A., Safarnejad, L., Harper, R., & Chu, B. T. (2019). IoCMiner: Automatic Extraction of Indicators of Compromise from Twitter. Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019, 4747–4754.

VI. Elaboración del estado del arte

EP1 (Bisio et al., 2015) "Real-time monitoring of Twitter traffic by using semantic networks"

Este artículo nos habla de la comparación entre una red semántica integral (ConceptNet) y una base de datos semántica existente (EuroWordNet), el análisis se lo realizó tomando en cuenta 3 elementos, clúster cardinality (CC), relevant scored (RS) count y Relevant Attracted (RA) count, los resultados de la investigación muestran que la red semántica integral mejora el análisis del tráfico de Twitter mientras que la base de datos semántica es efectiva siempre y cuando se conozca el tema a analizar.

EP2 (Charitonidis et al., 2015) "Weak signals as predictors of real-world phenomena in social media"

Este artículo nos habla del análisis de datos en redes sociales para detectar pequeñas señales o indicadores que puedan avisar de un posible evento en el mundo real, es decir analizar temas importantes para encontrar señales fuertes prevalentes en una red social para poder pronosticar eventos futuros.

EP3 (Gu et al., 2016) "Sentiment Analysis for Topics based on Interaction Chain Model"

Este artículo nos habla de un nuevo método de análisis de sentimientos basado en cadenas de interacción, es decir, se organizan los mensajes para después encontrar temas comunes entre estos y poder reclasificar estos temas para evaluar el sentimiento de cada uno mediante el algoritmo de polaridad SBV basado en semántica. Los resultados demuestran que el algoritmo heurístico propuesto extrae temas significativos por lo que el análisis de sentimientos es efectivo.

EP4 (Lu et al., 2020) "Domain-Oriented Topic Discovery Based on Features Extraction and Topic Clustering"

Este artículo nos habla de métodos de extracción de características orientadas al dominio mediante palabras clave ITFIDF-LP, mediante características de palabras en un tema LDA-SLP y en un vector de similitud de producto. Este análisis se lo realizó con 4 sets de datos:

- Corpus de noticias wiki de código abierto en inglés.
- Datos obtenidos con un crawler[5] de 8 blogs de seguridad informática, plataformas de noticias entre otras.
- Temas relacionados a la ciberseguridad extraídos de artículos con la etiqueta "malware".
- Temas relacionados a la ciberseguridad extraídos de artículos con la etiqueta "vulnerabilidad".

Los resultados muestran que los métodos propuestos son de utilidad para identificar tendencias sobre temas de seguridad informática lo cual ayuda a que se tomen acciones oportunas.

EP5 (Lwowski et al., 2018) "Geospatial Event Detection by Grouping Emotion Contagion in Social Media"

En este artículo se utiliza la API de Twitter para obtener información identificada por geolocalización y analizada semánticamente ya que mucha de esta tiene emoticonos que pueden dificultar su análisis puesto que se usa en diferentes contextos, se propone una herramienta que permita monitorear zonas geoespaciales para saber cuál es el estado de ánimo de esas personas.

[5] Comúnmente conocido como rastreador, es un programa que analiza los documentos de los sitios web.

A su vez se propone el uso de inteligencia artificial mediante teoría de grafos y estadística para identificar tendencias en durante momentos de extrema emoción.

EP6 (Niakanlahiji et al., 2019) "IoCMiner: Automatic Extraction of Indicators of Compromise from Twitter"

Este artículo nos habla de un framework para extraer ciber inteligencia de amenazas (CTI) en Twitter, utilizando teoría de grafos, aprendizaje automático y minería de texto. IoCMiner busca la reputación del usuario para darle credibilidad a sus tweets y así tomarlo como fuente de CTI. Los resultados mostraron que solo el 10% de las URLs encontradas mediante IoCMiner estaban registradas en bases de datos de lista negra.

Características del estado del arte

Existe una cantidad considerable de estudios en cuanto a la aplicación de técnicas de minería de datos y ciberseguridad, la mayoría habla en resumen de las técnicas y tiene algo concreto con lo que se pueda trabajar, pero no se proponen sistemas que pueden asistir a un CSIRT a evidenciar los ataques que se propician en la red social Twitter para que puedan tomar acciones al respecto.

Un estudio que resulto interesante es el modelo de análisis de sentimientos basado en temas y en cadenas de interacción donde podemos organizar mensajes y después encontrar los temas comunes entre ellos para reclasificar y evaluar el sentimiento de cada uno.

Otro estudio interesante fue el que propone el análisis de pequeñas señales que ayuden a predecir fenómenos sociales a gran escala, todo esto a partir de analizar temas específicos y revisar si existe algún mensaje o propuesta de movilización o de ataque para así poder frenarlo.

Al realizar la revisión de los artículos se ha llegado a varias conclusiones con respecto al análisis de datos, se pueden realizar comparación de redes semánticas que pueden ser integrales o existentes, en donde los resultados son favorables en las redes semánticas integrales, también señales o indicadores los cuales pueden pronosticar algún evento futuro.

La organización de datos de los mensajes obtenidos para después encontrar temas comunes y reclasificarlos mediante el algoritmo SBV, mediante esto se obtienen el análisis de sentimientos más efectivo y especifico.

Realizando análisis a los datos mediante la extracción de palabras clave las cuales estarán definidas por diccionarios o corpus, con esto se identifican tendencias o acciones futuras.

Todos los análisis de información, sentimientos, posibles escenarios a ocurrir son valorados por varios estudios, comparaciones, organización, reclasificación de caracteres y datos específicos, los cuales utilizan algoritmos y técnicas que permitirán obtener una respuesta especifica o más segura de cualquier posible evento.

Metodología de la investigación.

Metodología Design Science

Según (van Aken & Romme, 2012) la metodología Design Science, no trata de desarrollar proposiciones "verdaderas" sobre la realidad sino desarrollar propuestas que enseñan a las personas sobre cómo crear mejores realidades.

Esta metodología busca extender los límites de las capacidades humanas y organizacionales creando artefactos nuevos e innovadores. En la metodología, el conocimiento y la comprensión de un dominio del problema y su solución se consiguen en la construcción y aplicación del arte hecho (Hevner, March, Park, & Ram, 2004).

Es fundamentalmente una metodología que nos permita una resolución de problemas, la cual busca crear innovaciones que definan las ideas, las prácticas, las capacidades técnicas y los productos a través de los cuales el análisis, diseño, implementación, manejo y uso de los sistemas de información puedan logrados o alcanzados de manera efectiva.

Esta metodología es usada comúnmente en Ciencias de la Computación ya que permite diseñar una solución basada en resultados que permita resolver un problema importante mediante la aplicación de tecnología.

Consta de 4 artefactos:

i. Constructos
    i. Nos permiten comunicar los problemas y soluciones.
    ii. Vocabulario y símbolos.
ii. Modelos
    i. Representa la conexión entre el problema, solución y sus componentes.
    ii. Abstracciones y representaciones.

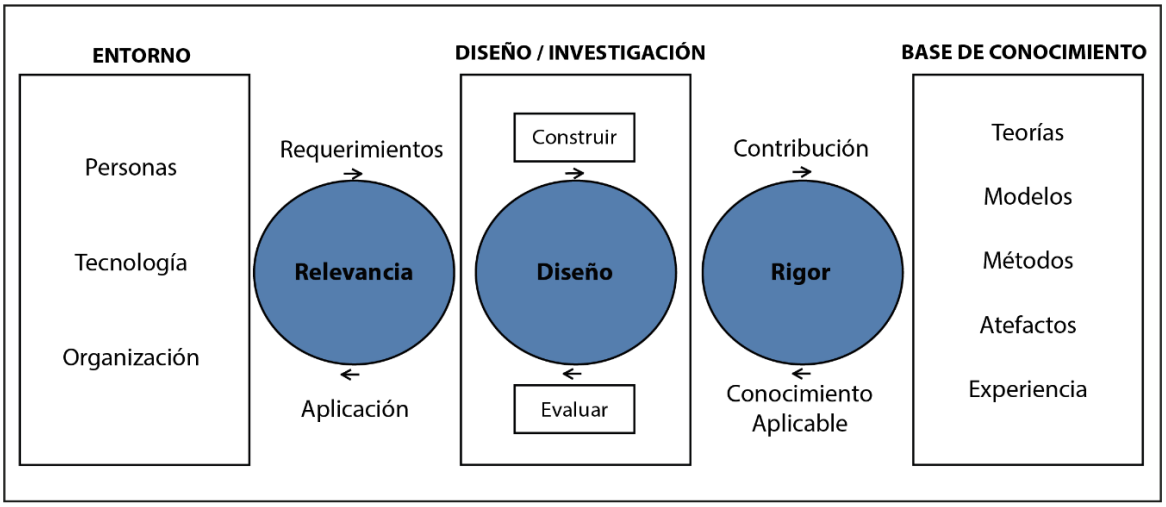
iii. Métodos
    i. Proveen guías sobre cómo encontrar soluciones a problemas.
    ii. Algoritmos y prácticas.
iv. Instanciaciones
    i. Muestran que constructos, métodos o modelos pueden ser implementados.
    ii. Prototipos y sistemas implementados.

En el Design Science Research se toma en cuenta 7 pasos según (Cataldo, 2015), los cuales nos guiarán a lo largo del proceso de desarrollo del tema propuesto, estos son:

1. Relevancia del problema.
2. Diseño como artefacto.
3. Rigor de la investigación.
4. Diseño como un proceso de búsqueda.
5. Evaluación del diseño.
6. Contribuciones a la investigación.
7. Comunicación de la investigación.

Figura 2

Ciclos de la Metodología Design Science Research

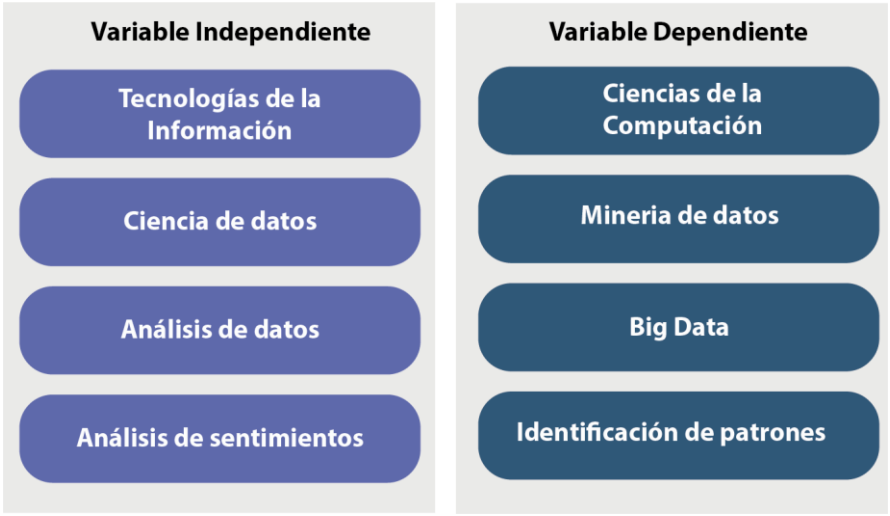


Marco teórico

Con el fin de evidenciar el potencial del proyecto de investigación, se ha realizado una descripción de ciertas áreas específicas y generales.

Figura 3

Red de categorías de las variables



Variable Independiente

Tecnologías de la información.

Para Julio Cabero, catedrático y especialista en Nuevas Tecnologías, las TIC: "*En líneas generales podríamos decir que las nuevas tecnologías de la información y comunicación son las que giran en torno a tres medios básicos: la informática, la microelectrónica y las telecomunicaciones; pero giran, no sólo de forma aislada, sino lo que es más significativo de manera interactiva e Inter conexionadas, lo que permite conseguir nuevas realidades comunicativas*"(Julio Cabero Almenara, 2002).

Para Consuelo Belloch las TICS : La dimensión social de las TIC se vislumbra atendiendo a la fuerza e influencia que tiene en los diferentes ámbitos y a las nuevas estructuras sociales que están emergiendo, produciéndose una interacción constante y bidireccional entre la tecnología y la sociedad(Ortí, 2012).

El avance de las tecnologías de la información nos ha permitido crear herramientas de productividad impensables años atrás, como asistentes inteligentes que hagan cosas

por nosotros mientras nos enfocamos en lo importante. Pero a su vez, se han creado riesgos importantes dentro del uso de la tecnología.

Ciencia de datos.

Las ciencias de datos son el procesamiento y análisis de los datos masivos se sustentan en una serie de técnicas y capacidades individuales y organizacionales, que son el objeto de la disciplina conocida como ciencia de datos(Lerena, 2019).

La ciencia de datos consiste en la aplicación de métodos científicos para construir algoritmos y sistemas que permiten detectar patrones y descubrir conocimiento útil para la toma de decisiones. Involucra procesos de integración y análisis de datos de distintas fuentes y en una variedad de formatos, a fin de construir modelos que ayudan a identificar y comprender fenómenos complejos (Ramirez-Morales et al., 2018).

Con la cantidad de información que se genera actualmente en el mundo, la ciencia de datos se ha convertido en una de las necesidades de empresas cuyo flujo de datos es muy extenso, ya que les permite el análisis de características, como cantidad, origen, etc., que nunca habían tomado en cuenta, lo que les permite obtener una vista más amplia de sus negocios.

Análisis de datos

Según (Stedman, 2021) El análisis o analítica de datos es el proceso de examinar conjuntos de datos para encontrar tendencias y sacar conclusiones sobre la información que contienen. También se utiliza por científicos e investigadores para verificar o refutar modelos, teorías e hipótesis científicas.

El análisis de datos se lo aplica en casi todos los escenarios científicos, ya que, por su versatilidad, nos permite aplicar desde los algoritmos más fáciles hasta los más

complicados obteniendo, para cada caso, resultados importantes que ayudan a la investigación.

Análisis de sentimientos

Según Danny Zambrano et al: El análisis de sentimientos en texto particularmente permite por ejemplo clasificar la polaridad de un texto dado, es decir si la opinión expresada en un documento o una oración es positiva, negativa, o neutra. También podemos obtener clasificaciones más detalladas que buscan, por ejemplo, estados emocionales tales como "enfado", "tristeza", o "felicidad" (Zambrano et al., 2017).

Los estudios sobre sentimientos han estado presentes en diferentes campos y con fines distintos. Desde los primeros trabajos cognitivistas de Arnold (1960), muchas han sido las líneas de investigación seguidas para el análisis de las emociones y los sentimientos, casi tantas como los términos adoptados dentro de los diversos campos para referirse a la manifestación de las emociones, opiniones, gustos y valoraciones (Aguado-de-Cea et al., 2013).

Estos análisis se lo utilizan comúnmente en la industria en general ya que analizar la emoción que ocasiona un producto en una persona les da una pauta para saber qué productos deben seguir a la venta y cuales ya no, haciendo que sean más eficientes en el mercado.

Variable Dependiente

Ciencias de la computación.

El objeto fundamental de estudio de las ciencias de la computación son los algoritmos y, en su caso, su implementación (Viso Gurovich, 2006).

Las ciencias de la computación es el arte de mezclar ideas humanas y herramientas digitales para aumentar la capacidad de resolver problemas (Code, 2015).

Si bien es cierto las ciencias de la computación nos brindan un set completo de herramientas con las cuales nosotros podemos resolver problemas de forma automática y eficiente, casi nadie se da cuenta que a la par, existen organizaciones o grupos de personas que se dedican a explotar estas herramientas para sacar provecho propio de las vulnerabilidades de las personas.

Minería de datos.

Una definición tradicional es la siguiente: Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos (Vallejos, 2006).

Desde el punto de vista empresarial, lo definimos como: La integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión (Vallejos, 2006).

La minería de datos es un proceso relativamente nuevo a pesar de que ya hace un buen tiempo atrás existían ensayos y experimentos basándose en esta técnica. Actualmente nos encontramos en un punto donde tanto la minería de datos y la inteligencia artificial se unen para poder crear herramientas que satisfagan las necesidades de las empresas y personas.

Big Data

El termino Big Data se refiere a la evolución y uso de tecnologías que provean a un usuario indicado, en el tiempo indicado, la información correcta, desde una masa de datos que ha ido creciendo exponencialmente en nuestra sociedad. El desafío no solo es lidiar con volúmenes enormes de información sino también la dificultad de manejar formatos heterogéneos que a su vez son complejos (Riahi & Riahi, 2018).

Identificación de patrones

La identificación de patrones es la ciencia que se ocupa de los procesos sobre ingeniería, computación y matemáticas relacionados con objetos físicos y/o abstractos, con el propósito de extraer información que permita establecer propiedades de o entre conjuntos de dichos objetos. es un elemento importante(Nava, 2006).

El reconocimiento de patrones involucra diferentes etapas. Se comienza con la observación de la realidad e identificación de un sistema físico, continua con la definición de un sistema de medición, la obtención y validación de los datos y finaliza con la modelización matemática a través de un modelo de reconocimiento adecuado.

El reconocimiento de patrones conlleva la definición acerca de cómo se clasificarán los objetos. En ese sentido se destaca la selección de variables. Dentro de la selección de variables se puede trabajar sobre las características más adecuadas para la clasificación y/o el procesamiento (Gawron et al., 2014).

Capítulo III
Análisis, Diseño e Implementación

Análisis de la solución

Affin (Normas efectivas para palabras)

Es una lista de términos y palabras calificadas manualmente con una valoración de acuerdo con lo positivo, negativo o neutro que sean descritas. Cada valoración va desde el -5 al -1 en valoración negativa, 0 como neutral y desde 1 al 5 en valoración positiva, esta valoración fue propuesta por Finn Arup Nielsen entre 2009 y 2011.

Esta puntuación se la realizó ya que SentiStrength2 utiliza una puntuación de –5 como muy negativa y +5 como muy positiva, calificando la valencia dejando fuera palabras que representen subjetividad, objetividad, excitación y dominio, puntuando las palabras manualmente por parte del Autor. (Nielsen et al., n.d.)

Creada originalmente para realizar el análisis de sentimientos sobre la Conferencia de las Naciones Unidas sobre China (COP15) en el 2009, que originalmente contenía 1468 palabras diferente y algunas frases, distribuida en internet como AFFIN-96.

Actualmente AFFIN contiene alrededor de 2477 palabras, las cuales están comprendidas en palabras únicas y 15 frases relevantes.

Esta lista fue creada partir de varios conjuntos de palabras positivas, negativas y obscenas, las cuales se ampliaban gradualmente mediante la verificación de las publicaciones realizadas en el COP15, también con palabras de dominio público.

Utilizó Twitter para definir en qué contexto se empleó la palabra y mediante eso darle una valoración, Servicios Web de similitud de n-gramas de Microsoft Web

(“Agrupación de palabras en función de la similitud del contexto”) para descubrir palabras relevantes, y no distinguir categoría de palabras para evitar ambigüedades y palabras de alta excitación, pero de sentimiento variable, como, por ejemplo:

- Paciente
- Firme
- Mezquino
- Poderoso
- Franco
- Sorpresa

También mediante la examinación de lista de palabras de General Inquire y OpinionFinder con valoraciones de palabras de valencia positiva de +1 y negativa de –1, y sentimientos de fuerza de SentiStrength en su Web Service.

“*El análisis de la intersección entre la lista de dos palabras indicó que la puntuación de ANEW es mejor. El rendimiento ligeramente mejor de mi lista con todo el léxico puede deberse a la inclusión de jerga de Internet y palabras obscenas*.” (Nielsen et al., n.d.)

Para este proyecto se usó léxico AFFIN con una modificación en varias palabras para que estén acorde a los términos en oraciones y frases que son utilizadas en nuestro país Ecuador.

Para realizar un análisis de palabras hay que tener en cuenta que el significado de varias palabras puede ser diferentes a los comunes conocidos, estos significados pueden variar de acuerdo con la región y país de donde se quiera realizar el análisis, por tanto, también afecta en la valoración de estas.

Se propone para este proyecto el AFFIN modificado con las palabras que puedan tener un significado y valoración relevante y mediante este realizar el análisis de sentimientos que sea lo más cercano a la realidad y conforme lo necesitemos se modifique y sea más completo.

ConceptNet

ConceptNet se originó a partir del proyecto de crowdsourcing Open Mind Common Sense, que se lanzó en 1999 en el MIT Media Lab. Desde entonces, ha crecido para incluir conocimientos de otros recursos de colaboración colectiva, recursos creados por expertos y juegos con un propósito.

En los últimos años se ha empezado con el desarrollo de nuevas bases del conocimiento como lo es ConceptNet, desarrollado por Novell Linguistic Technology Department, a partir de WordNet 1.5 con el concepto de synset (conjunto de sinónimos), su arquitectura está basada en todas las especificaciones de WordNet 1.5 más varios elementos de estructura básicos propios.

Está diseñado para conservar de forma fácil las complejas redes semánticas multilingües, con nuevos significados, palabras y relaciones y almacenar los resultados de forma fácil.

La base de datos de ConceptNet está diseñada y programada para el exclusivo manejo de redes semánticas, con una base de datos sin comprimir lo que permite realizar cambios a la información y comprimiéndose para integrarse al producto final.

La base de datos comprimida funciona con motores de acceso rápido con funciones claras y sencillas, se conecta con otros módulos como una analizador y generador morfológico que nos permite tratar al léxico de forma paradigmática desde muchos puntos de vista.

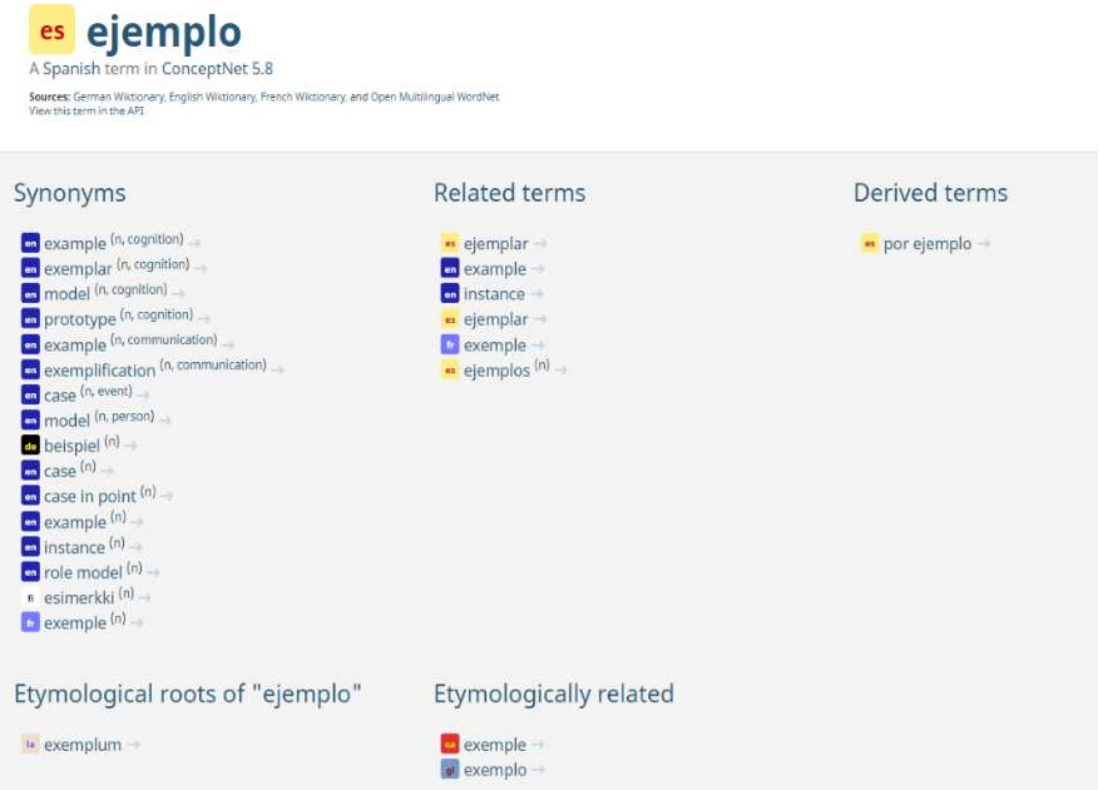
Cuando ConceptNet se combina con inserciones de palabras adquirido de la semántica distributiva (como word2vec), proporciona a las aplicaciones el entendimiento de que adquirir de la semántica distributiva solamente, ni de recursos más limitados como WordNet o DBPedia.

ConceptNet es una base de conocimiento que conecta palabras y frases de lenguaje natural, ConceptNet 5.5 se ha ampliado para incluir léxicos y conocimiento mundial de muchas fuentes diferentes en varios idiomas, nos permite conocer el significado y la contextualización de las palabras que se utilizan en cualquier tipo de análisis, capturando una amplia gama de conceptos y relaciones de sentido común.

A continuación, se representará un ejemplo de ConceptNet 5.5, en el cual podremos observar los diferentes contextos en los que ConceptNet nos ayuda a entender las palabras con sus diferentes sinónimos, términos relacionados, términos derivados, formas de palabras y su etimología.

Figura 4

Ejemplo ConceptNet



La utilización de ConceptNet en este proyecto es importante ya que facilitaría en la valoración y la contextualización que se le dará a cada palabra, frase y oración que se realizara el análisis, así también como entender el significado verdadero de las palabras conforme al contexto de cada región y significado de este.

Algoritmo KNN (k-Nearest Neighbors)

Este algoritmo busca valores más cercanos al punto que se está tratando de predecir. Se encuentra dentro del conjunto de algoritmos supervisados ya que nosotros definimos nuestro conjunto de datos de entrenamiento. Además, se lo denomina como basado en instancia ya que de por si no aprende explícitamente un modelo, sino que

memoriza los modos de entrenamiento que se encuentran en su base para predecir un punto.

Para especificar el número de vecinos más próximos tomamos un valor de k, con este podemos saber cuántos vecinos a la redonda podemos tener para poder inferir una predicción.(IBM, 2021)

Las características que pueden tomar las variables son:

- Nominales: Cuando sus valores representan categorías sin clasificación intrínseca. Por ejemplo: región, código postal, afiliación religiosa.
- Ordinales: Cuando sus valores representan categorías con alguna clasificación intrínseca. Por ejemplo: grado de satisfacción, calificaciones.
- Escalares: Cuando sus valores representan categorías ordenadas con una métrica significativa. Por ejemplo: edad, ingresos en miles de dólares.

En el caso de estudio planteado, se utilizará una escala de medición Ordinal ya que permite clasificar las características por grados de acuerdo con un criterio de orden. (Gene V. Glass, 1986)

Distancia Euclidiana

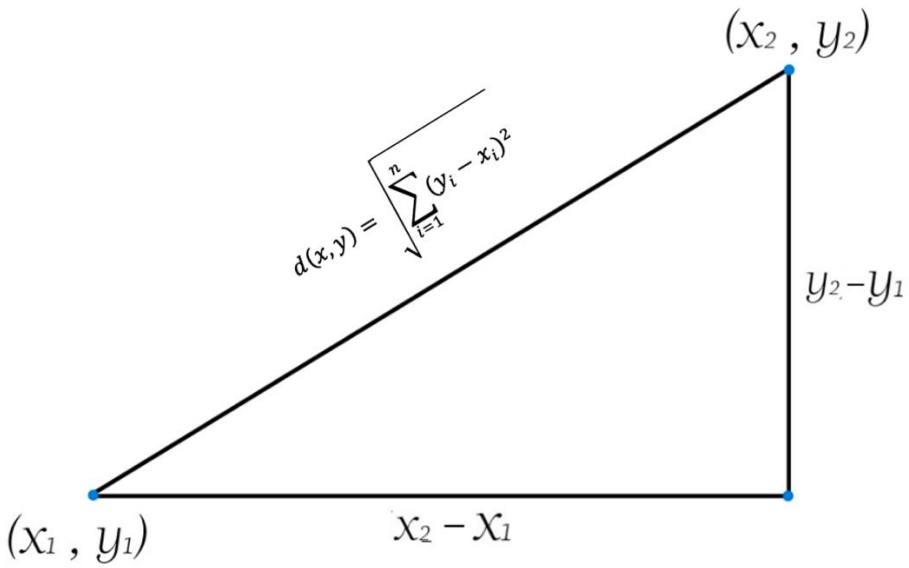
Es una fórmula que permite medir la distancia entre 2 puntos en un espacio de n dimensiones, a continuación, su descripción:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

De forma gráfica se lo representa de la siguiente manera:

Figura 5

Representación Distancia Euclidiana



La distancia euclidiana resulta ser útil en el estudio de grafos para obtener la similitud entre conjunto de datos.

En el prototipo se utilizará esta fórmula para pre clasificar los tweets evaluando la distancia con respecto al grupo de entrenamiento.

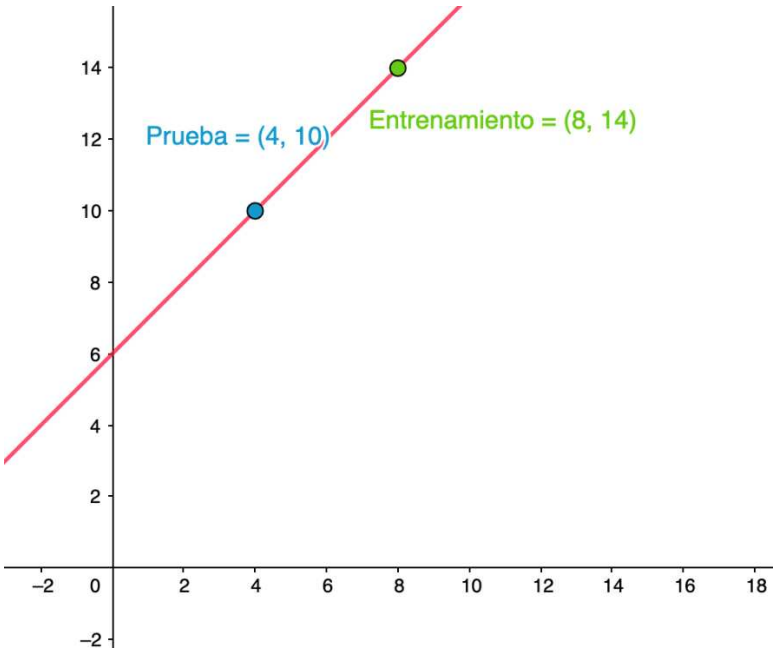
Por ejemplo, tenemos un set de datos de entrenamiento y un set de datos de prueba los cuales comparten las siguientes variables:

- Puntuación
- Longitud
- Ponderación

Los valores que usaremos para el cálculo serán tanto la puntuación como la longitud del tweet, representando (x, y) respectivamente y con esto obtendremos 2 puntos en el plano cartesiano.

Figura 6

Representación en el plano cartesiano de los puntos de nuestro ejemplo



Con estos datos calcularemos la distancia con la fórmula que presentamos anteriormente y el resultado se guardará temporalmente en memoria para utilizarlo con Knn y así obtener nuestra ponderación final.

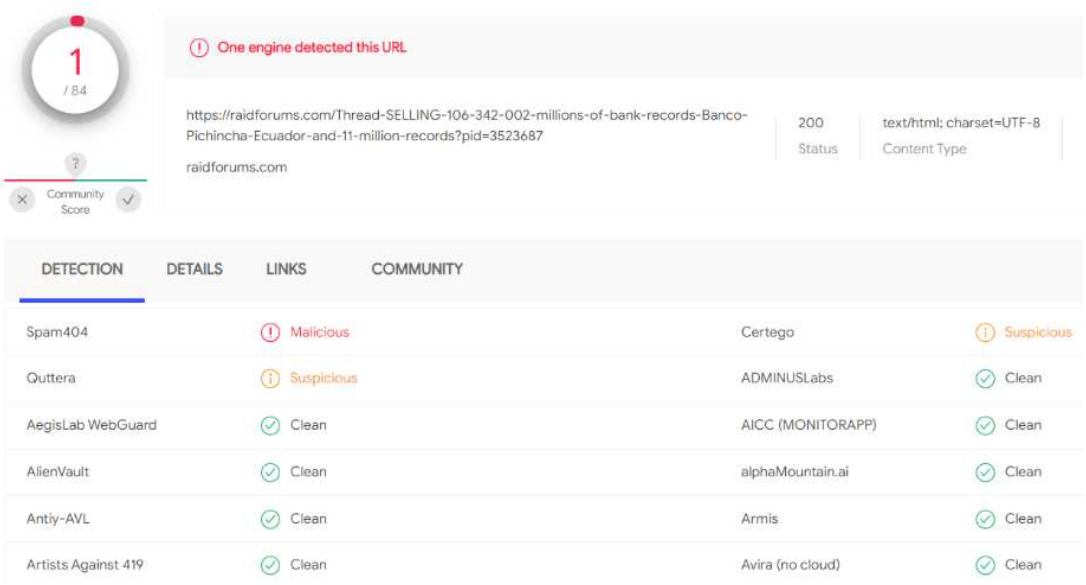
Virus Total (API)

Es una API que ~~nos~~ permite realizar el análisis de una URL, esta es escaneada en varias bases de datos de antivirus, phishing, programa maligno, etc.

Esta API es muy importante para el análisis de archivos y en este caso particular una URL, dado ~~a~~ que cuenta con un sin número de bases de conocimiento en donde se analizan y presenta de forma legible y grafica el resultado que si es maliciosa o está limpia.

Figura 7

Análisis Virus Total API



En este proyecto la utilización de este API es muy importante ya que al analizar las URL podemos agregar a la verificación de data realizada en Twitter y valorar que la información es maliciosa o no.

Fase de Iniciación

Para esta fase del desarrollo se va a profundizar en el uso que se va a dar a cada recurso o herramienta presentada dentro de diagrama de arquitectura mostrado.

Servidor de base de datos

El servidor de base de datos a usar es MongoDB en un servidor alojado localmente. Se seleccionó este servidor ya que nos permite crear colecciones de datos con los cuales podemos interactuar tanto para la limpieza de datos como para su visualización.

Servidor API Rest

Para el servidor API se utilizará el framework NodeJs, el cual nos permite crear aplicaciones web escalables (OpenJS Foundation, 2020) y que permite crear los Endpoints que necesitemos sin necesidad de utilizar muchos recursos del sistema en el que está alojado. Al ser un prototipo todas las peticiones se realizarán localmente con HTTP.

Aplicación Web

Se desarrollará en el framework NodeJs para facilitar su desarrollo y pruebas. A su vez se utilizará el IDE Visual Studio por su versatilidad y gran variedad de herramientas para el desarrollo de esta aplicación web.

Fase de producción

Según la metodología Scrum las fases de desarrollo pueden dividirse en varias por semana y a su vez dividirse en pequeños desarrollos que tienen su retroalimentación al final de la semana de trabajo, haciendo mejorías a lo largo del proceso (Cycle et al., 2007) , lo importante es que podamos tener las reuniones semanales o diarias sobre el desarrollo de la aplicación.

Desarrollo del servidor API Rest

Para el servidor API Rest utilizamos las siguientes librerías:

- Express: Es un framework web para crear APIs.
- MongoDB: Es una librería que permite conectar con la base de datos alojada localmente.

Nuestro prototipo tiene previsto usar 3 bases de datos de MongoDB, los cuales serán, análisis, limpieza y resultados, este último se concatenará con la tendencia que será analizada, por ejemplo “ResultadosElecciones”.

La conexión con la base de datos será inicializada en la sección principal del servidor, de forma que se instancie una sola vez y así podamos evitar un consumo desmedido de recursos.

La implementación de los servicios REST tienen que ver con los documentos que se guardaron en la colección “tesis”. Los cuales están clasificados de acuerdo con la visualización de datos en cada página de la aplicación web.

Figura 8

Implementación de servicios Rest

```
JS app.js > ...
  1  var path = require("path");
  2  var http = require("http");
  3  var bodyParser = require("body-parser");
  4  var express = require("express");
  5  var MongoClient = require("mongodb").MongoClient;
  6  var url = "mongodb://localhost:27017/";
  7  const axios = require('axios');
  8  var app = express();
  9  app.use(bodyParser.urlencoded({ extended: false }));
 10  app.use(express.static(__dirname + "/public"));
 11  var httpServer = http.createServer(app);
 12
 13  app.get("/", function (req, res) {···
 15  });
 16  app.get("/puntuacion", function (req, res) {···
 18  });
 19  app.get("/global", function (req, res) {···
 38  });
 39  app.get("/sortAsc", function (req, res) {···
 59  });
 60  app.get("/score", function (req, res) {···
 79  });
 80  app.get("/tweets", function (req, res) {···
 93  });
 94  app.get("/analyzed", function (req, res) {···
107  });
108
109  httpServer.listen(3000, function () {
110    console.log("HTTP on port 3000!");
111  });
```

Desarrollo de la aplicación web

El prototipo constara de 3 secciones las cuales serán divididas por el tipo de consulta a realizar, es decir tenemos la sección general, la sección de resultados y la sección de tendencias en Ecuador.

Todas las secciones constarán de un módulo de navegación en el cual se mostrarán las secciones disponibles mencionadas anteriormente.

Figura 9

Secciones disponibles

```
<nav class="navbar navbar-expand-lg navbar-light">
    <div class="container-fluid">
        <div class="collapse navbar-collapse" id="navbarSupportedContent">
            <ul class="navbar-nav me-auto mb-2 mb-lg-0">
                <li class="nav-item">
                    <a class="nav-link active" aria-current="page" href="/">General</a>
                </li>
                <li class="nav-item">
                    <a class="nav-link " aria-current="page" href="/puntuacion">Análisis</a>
                </li>
                <li class="nav-item">
                    <button type="button" class="btn btn-light" data-bs-toggle="modal" data-bs-target="#staticBackdrop">
                        Ver tendencias en Ecuador
                    </button>
```

La sección general tiene como objetivo mostrar al usuario el número de tweets que fueron recolectados, las fechas de inicio y fin de la recolección y una tabla donde se pueden ver los últimos 20 tweets. (Ver sección “Fase de pruebas")

La sección de análisis mostrará los tweets que quedaron después de la limpieza de datos, que pretende eliminar repetidos y obtener solo tweets orgánicos[6] sobre una tendencia en especial. En esta sección tendremos el número de tweets, la fecha de inicio y fin de la recolección de datos, los resultados del análisis que mostrarán los casos positivos, neutros y negativos, así como 2 gráficos sobre los casos detectados y en que ciudades del Ecuador se han generado estos tweets, y al final de la página tenemos una sección de búsqueda de enlaces maliciosos de los 10 tweets con mayor puntaje. (Ver sección “Fase de pruebas")

[6] Se refiere al contenido que se genera de forma espontánea sin necesidad de un estímulo económico.

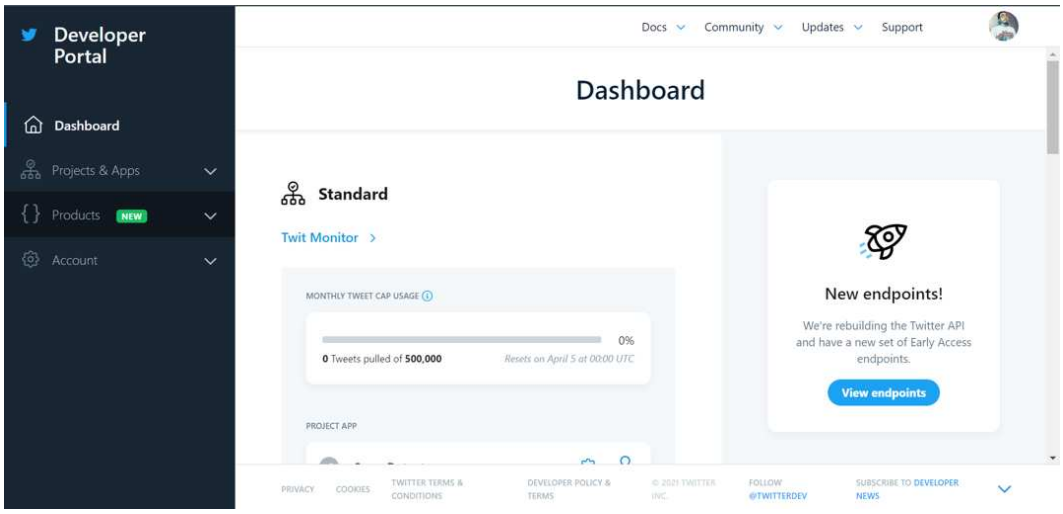
Desarrollo de script de recolección de datos en Twitter

Este script está basado en el uso de Tweepy[7], una librería para Python con la cual podremos acceder a la información de los Endpoints de Twitter de forma fácil. Basta con instalarla como dependencia[8] dentro de NodeJs.

Para usar el API de Twitter debemos obtener una cuenta de desarrollador, la cual por lo general tarda varios días en ser verificada. Esta API suele tener contenido filtrado, es decir previamente pasa por controles de seguridad.

Figura 10

API Twitter - desarrollador



Una vez tengamos las claves de acceso[9] al API de Twitter, las colocamos dentro del script y en la variable “Query” debemos escribir la tendencia de la cual obtendremos los datos, como se muestra en el ejemplo a continuación.

[7] Es una librería de Python para acceder a la API de Twitter.
[8] Es una librería requerida por un programa para funcionar correctamente.
[9] Para poder consumir un API se necesita tener claves de acceso a este recurso.

Figura 11

Extracto de script para obtener datos de Twitter

```
query="EfrainRuales"
# Language code (follows ISO 639-1 standards)
language = "es"
count=100
res_t="recent"
tweet_mode="extended"
# Calling the user_timeline function with our parameters
results = api.search(q=query, lang=language, result_type=res_t,
count=count,tweet_mode=tweet_mode )
collection = db[query]
```

Después de la recolección de datos se procedió a realizar la reestructuración de la data mediante scripts de Python (figura 12) y luego se eliminan los duplicados e información vacía mediante scripts en R Studio.

Figura 12

Script de Python

Desarrollo de script de limpieza de datos en R

El proceso de limpieza tiene la finalidad de eliminar, nombres de usuario, tweets repetidos y tweets que no contengan información.

Para la limpieza de datos se usó R como herramienta debido a su simpleza en cuanto a uso y codificación.

Para limpiar los datos se los agrupa en una variable general llamada data, en la cual se guarda temporalmente toda la información de la colección seleccionada.

Luego procedemos a buscar los tweets repetidos gracias a la librería dyplr[10], la cual mediante la función distinc[11], nos permiten realizar esta limpieza.

Para guardar los datos limpios, usamos la librería mongolite, la cual nos permite conectar a la base de datos Mongo tanto para lectura como para escritura.

Figura 13

Conexión MongoDB

```
library(tidyverse)
library(mongolite)
library(dplyr)
library(purrr)
library(stringr)
library(rjson)

#Conexion a mongo para lectura
ecuador <- mongo("ecuador", url = "mongodb://127.0.0.1:27017/tesis")
#Total de datos en bdd
total <- ecuador$count()
#Asignacion de valores a variables
data <- ecuador$find('{}')
textPurge <- data %>% distinct(text, .keep_all = TRUE)
#Conexion a mongo para escritura
purgaTexto <- mongo("ecuador",url = "mongodb://127.0.0.1:27017/limpieza")
purgaTexto$insert(textPurge)
```

[10] Es una librería para R que nos permite simplificar el manejo de grupos de datos.
[11] Es una función del paquete dyplr para obtener los valores únicos de un grupo de datos.

Desarrollo de los diccionarios basados en Affin para el análisis de tendencias.

Para esta investigación se tomaron ejemplos de Affin y SentiStrength para la ponderación e inclusión de palabras, de forma general se buscó en diccionarios los significados de ciertas palabras para poder ponderarlas del -5 al 5, tomando en cuenta la jerga popular.

Cada una de las palabras que componen los diccionarios fue analizada de forma minuciosa basándonos en los siguientes casos de prueba que vamos a analizar:

- General
    - Se utilizaron términos que puedan dar indicios sobre comportamiento malicioso o criminal de forma general sobre el tema Ecuador.
- Compras
    - Se utilizaron términos que puedan dar indicios de actividad maliciosa o criminal sobre descuentos, cadenas de supermercados, etc.
- Virus
    - Se utilizaron términos que puedan dar indicios de actividad maliciosa o criminal sobre virus, información personal, filtraciones de datos en Pastebin.

Estos diccionarios están basados en 4 escenarios:

- Ecuador
    - Se obtuvieron tweets que tengan la palabra “ecuador” incluida con el fin saber si existe algún tipo de sentimiento a favor o en contra.
- Gratis

    - Se obtuvieron tweets con la palabra “gratis” con el fin de encontrar algún comportamiento peligroso que tenga que ver con compras u ofrecimientos de servicios.
- Efraín Rúales
    - Se obtuvieron tweets con el nombre de Efraín Rúales con el fin de analizar la data y ver si existe alguna información importante sobre su asesinato.
- Pastebin:
    - Se obtuvieron tweets con la palabra “Pastebin” con el fin de saber si existe algún ataque en Pastebin que tenga que ver con Ecuador.

Estos 4 escenarios nos van a ayudar a definir si en el transcurso de 1 mes existió la posibilidad de un ataque o algún evento malicioso que pudiera ser detectado.

Todo esto se realizó en un archivo de formato JSON, el cual nos permitirá guardar la información de forma ordenada y simple, para más adelante proceder a ejecutar un script en Python y así poder obtener la clasificación de los tweets.

Desarrollo del script para analizar una tendencia en base al diccionario

Para el script de análisis decidimos usar JavaScript ya que cuenta con herramientas propias del lenguaje con las cuales podemos manipular de mejor forma la información.

De forma general este script suma todas las ponderaciones de las palabras que constituyen un tweet para así obtener su valor final. Este valor final será guardado en la base de datos con el prefijo “Resultados. La cual puede ser consultada mediante el Endpoints destinado para eso.

Fase de estabilización

En esta fase se integrarán todos los componentes de la aplicación, es decir, vamos a conectar al cliente con los Endpoints del servicio REST.

Integración del cliente con el Api Rest

Para poder consumir los Endpoints desarrollados previamente, vamos a realizarlo mediante Ajax desde el cliente, con esto podemos generar métodos HTTP para leer o guardar información en la base de datos.

Figura 14

Consumo de Endpoints

```
function getSearchTerm() {
  var searchTerm = $("#searchTerm").val();
  var temaTitle = document.getElementById("temaTitle");
  temaTitle.innerText = "Por tendencia: " + searchTerm;
  $.get("/global?bdd=" + searchTerm, function (response, status) {
    total = response.length;
    console.log("Elemento: "+response)
    console.log("Numero total: " + total);
    var headerTotal = document.getElementById("cardTitle");
    headerTotal.innerText = total;
    var headerFechaInicial = document.getElementById("cardInitDate");
    let initialDate = response[0].created;
    headerFechaInicial.innerText = initialDate.substring(0, 10);
    var headerFechaFinal = document.getElementById("cardFinalDate");
    let finalDate = response[total - 1].created;
    headerFechaFinal.innerText = finalDate.substring(0, 10);
  });
}
```

Fase de pruebas

Una vez terminado el desarrollo de la aplicación, procedemos a realizar las pruebas, con el cual podemos visualizar el funcionamiento de esta.

Pantalla general

En esta pantalla se visualizarán las tendencias recolectadas en la base de datos, el número total de tweets, la fecha de la primera recolección de datos y la fecha de la última recolección de datos. En la parte inferior encontraremos una tabla en la cual se desplegarán los últimos 20 tweets recolectados.

Figura 15

Tendencias recolectadas

Pantalla de análisis

En esta pantalla se visualizarán los resultados del análisis en la base de datos, el número total de tweets, la fecha de la primera recolección de datos y la fecha de la última recolección de datos, los casos positivos, neutros, y negativos, 2 gráficos informativos

sobre los casos encontrados y sobre las ciudades del Ecuador desde donde se realizaron estos tweets. Además, en la parte inferior de la página podremos ver una sección dedicada al top 10 de los tweets con el mayor puntaje positivo y su respetivo botón de analizar, el cual va a realizar una consulta en la base de datos de virus total para ver si el sitio web es propenso a contener malware o fue reportado como sitio malicioso.

Figura 16

Resultado del Análisis

Figura 17

Resultados después del análisis.

Figura 18

Resultado Positivos, negativos y neutros

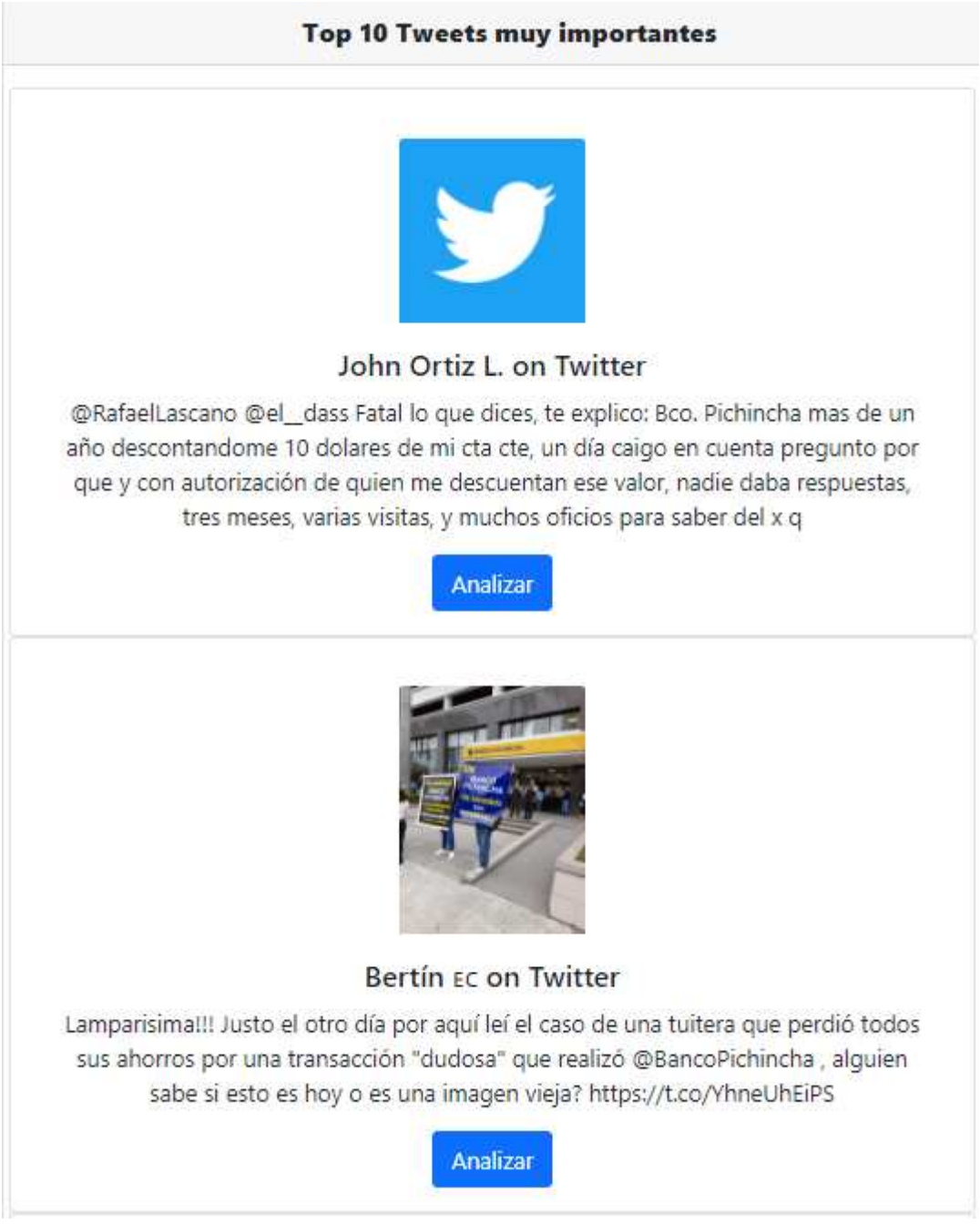
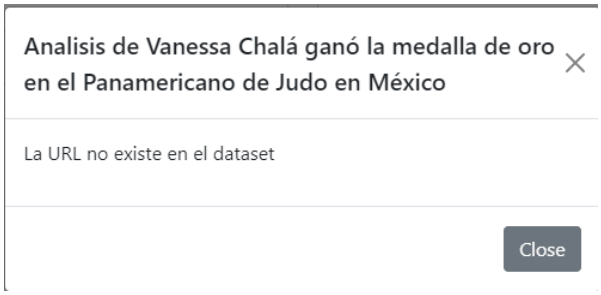


Figura 19

Análisis de URL

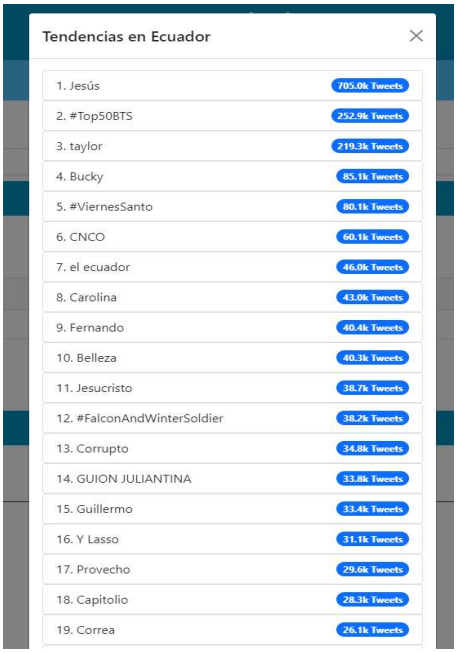


Modal de tendencias

En este modal se visualiza las tendencias en Ecuador en ese momento, es decir cada vez que le demos clic al botón de “Tendencias en Ecuador” este consultará con el API de Twitter y traerá las 20 tendencias.

Figura 20

Tendencia de Twitter



Capítulo IV
Pruebas y Resultados

Introducción

Este capítulo permitirá entender cómo se realizaron las pruebas de concepto para sustentar la investigación.

Se empleo un computador notebook con las siguientes características: Intel Core I7 de 3,40 GHz, memoria RAM de 16GB, con un almacenamiento de 1TB y SO Windows, con la cual se realizará tanto la obtención de información, limpieza, análisis y visualización de los resultados obtenidos, adicionalmente se realizarán pruebas con el API de virus total para saber si dentro de los 10 tweets con puntuación positiva existe algún enlace malicioso.

Para el análisis estadístico se utilizó a Tableau[12] como herramienta de soporte para obtener tanto los gráficos, como las relaciones entre las tablas y campos.

Factores de análisis

Para realizar los análisis nos basaremos en la tendencia Ecuador, por ejemplo: Efraín Rúales (Muerte de presentador de televisión ocurrida en enero 2021), Gratis (Black Friday, Promociones) y Banco Pichincha (Posible robo de datos personales de clientes). Se escogió estas tendencias ya que cada una es diferente entre sí, y permitirá obtener mejores resultados.

[12] Es un software de visualización y manejo de datos.

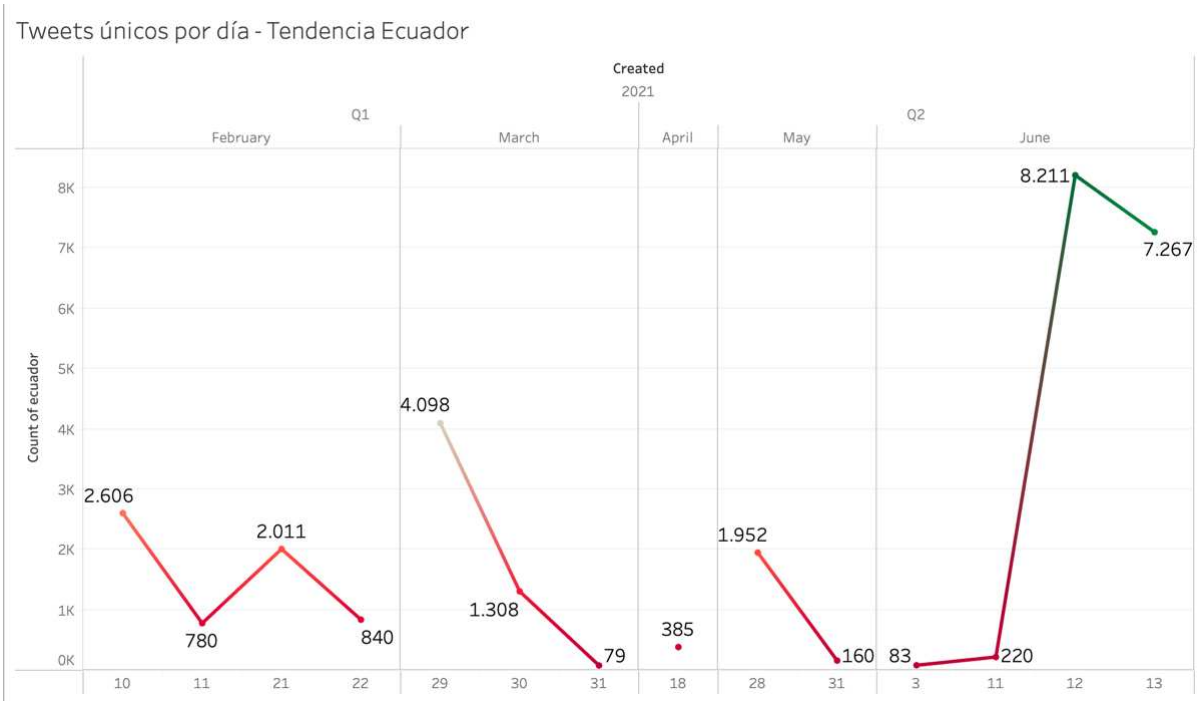
Tweets únicos por día.

Se segmentó la información por días y calculamos el total de tweets generados, luego se realizó un cruce entre los 2 campos[13] para obtener las siguientes gráficas.

Tendencia Ecuador

Figura 21

Tweets diarios tendencia Ecuador



En la gráfica podemos observar que el día 12 de junio tenemos un total de 8211 tweets únicos para la tendencia Ecuador. Este número elevado de tweets tienen que ver en parte por acontecimientos reportados en noticias, como un sismo de 4.2 grados en la

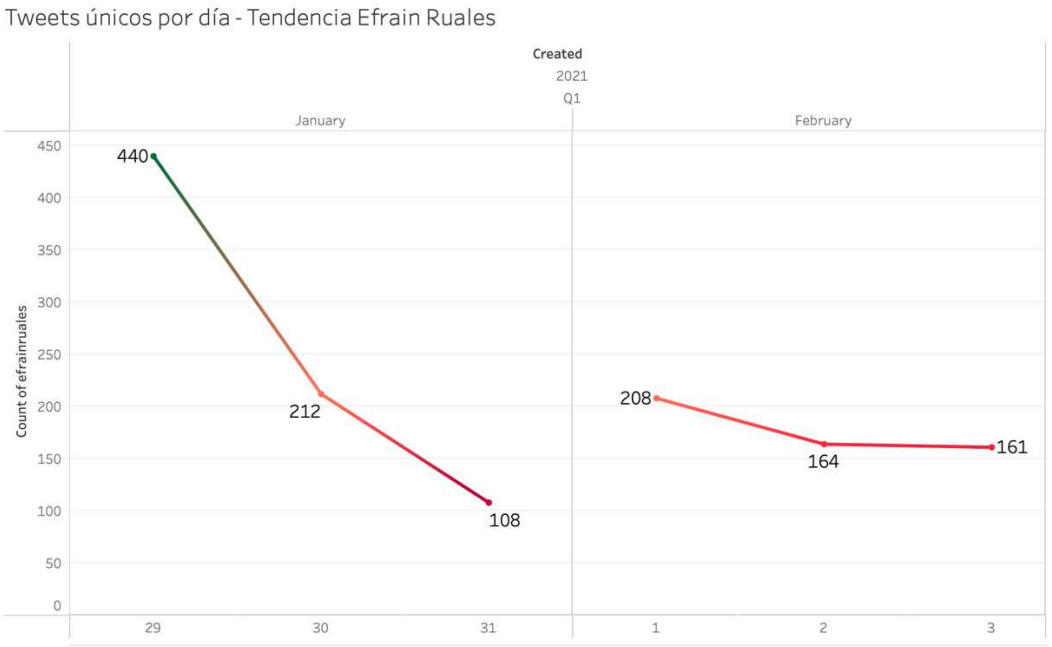
[13] Valores obtenidos de un análisis previo.

escala de Richter, la desaparición de una niña e incandescencia en el cráter del volcán Reventador.

Tendencia Efraín Rúales

Figura 22

Tweets diarios tendencia EfrainRuales

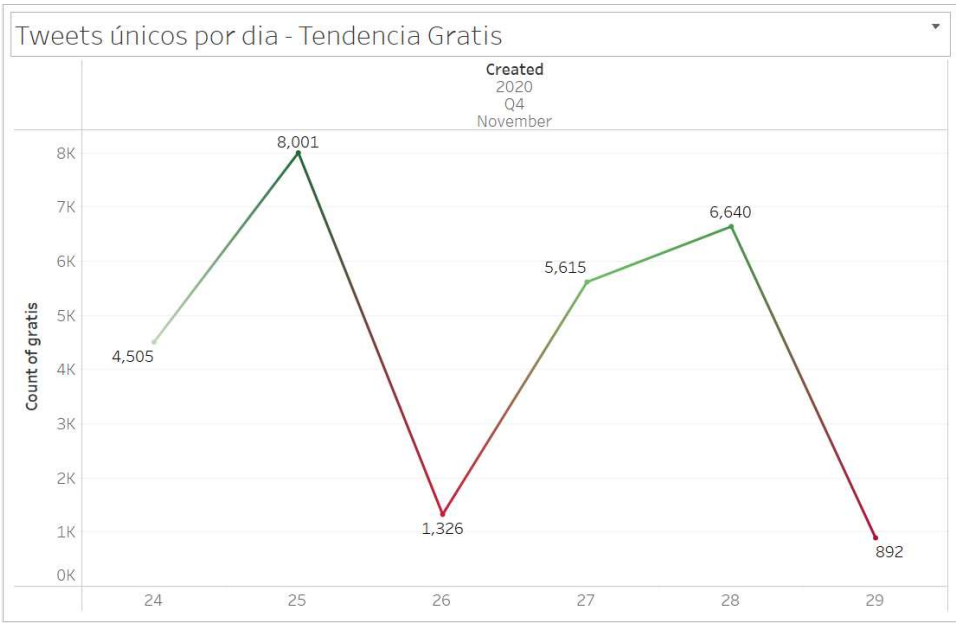


En la gráfica podemos ver como el 29 de enero de 2021 se obtuvieron 440 tweets únicos sobre la tendencia “EfrainRuales”, la cual con el pasar de los días fue decayendo. Y a pesar de que seguimos capturando datos hasta el mes de junio, no se mostraron más resultados relevantes después del mes de febrero.

Tendencia Gratis

Figura 23

Tweets diarios tendencia Gratis

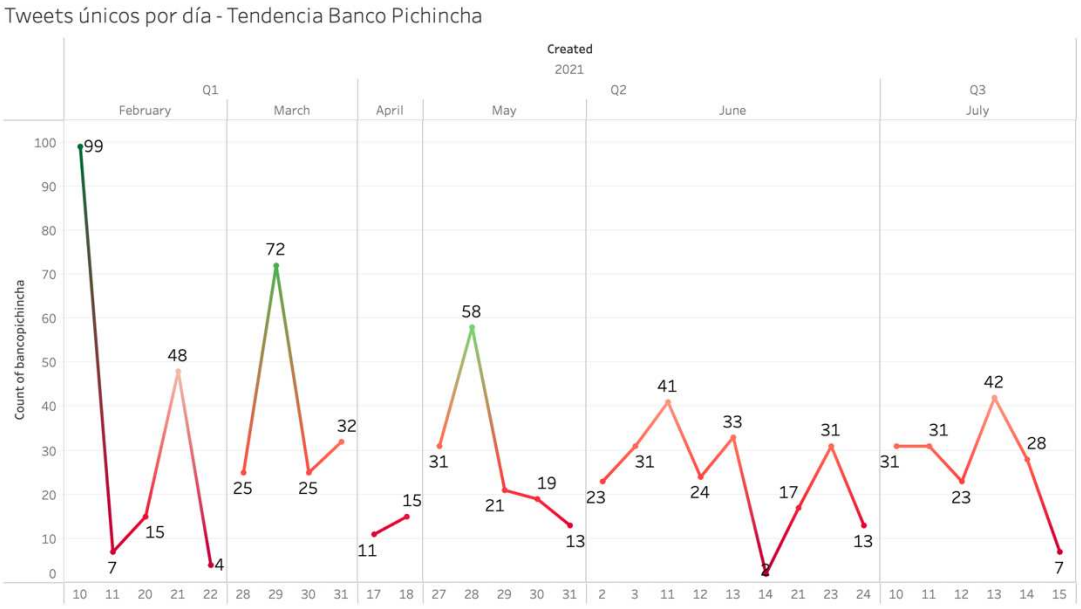


En la gráfica podemos ver que el día 25 de noviembre de 2020 la tendencia gratis tuvo 8001 tweets únicos y a lo largo del pasar de los días tuvo un comportamiento de altos y bajos, cabe aclarar que la información se recolectó en fechas cercanas a CyberMonday.

Tendencia Banco Pichincha

Figura 24

Tweets diarios tendencia Banco Pichincha



En la gráfica podemos ver que el 10 de febrero y 29 de marzo fueron los días con más interacciones en Twitter, tomando en cuenta de que en el mes de febrero existió un supuesto ataque informático hacia Banco del Pichincha, por lo que estuvimos recolectando información hasta julio, y así poder analizarla de forma conjunta para buscar algún posible patrón de comportamiento.

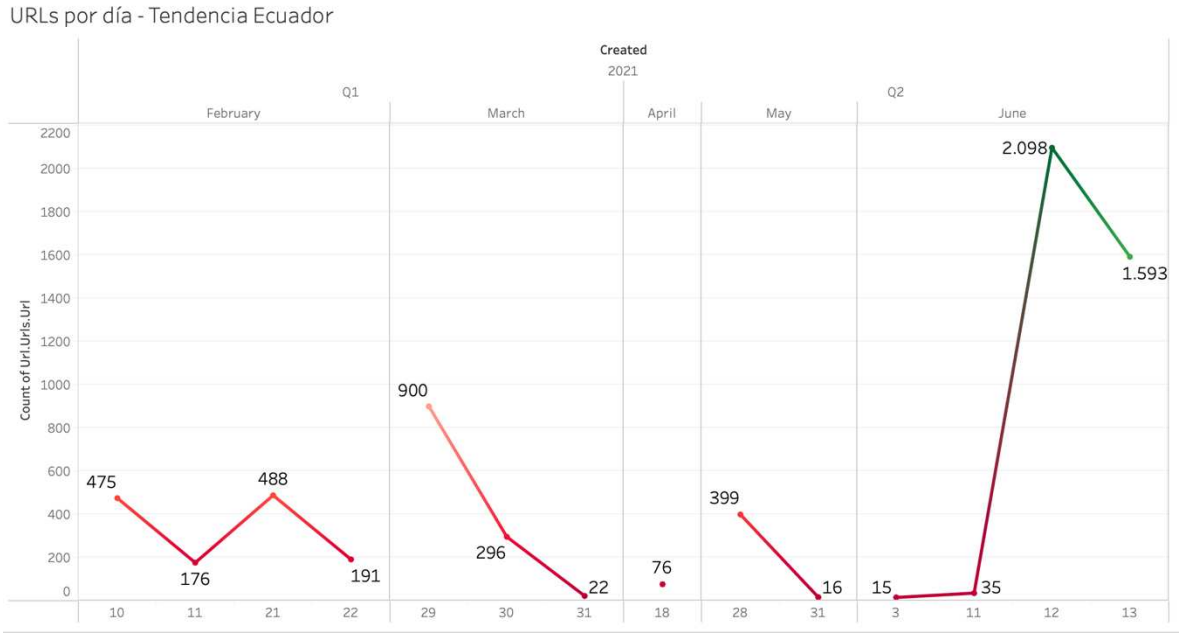
Cantidad de URLs por día

En las siguientes graficas se indica la cantidad total de URLs encontradas en el análisis realizo a los Twitter correspondiente a la tendencia.

Tendencia Ecuador

Figura 25

Total, de URL en tendencia "Ecuador"

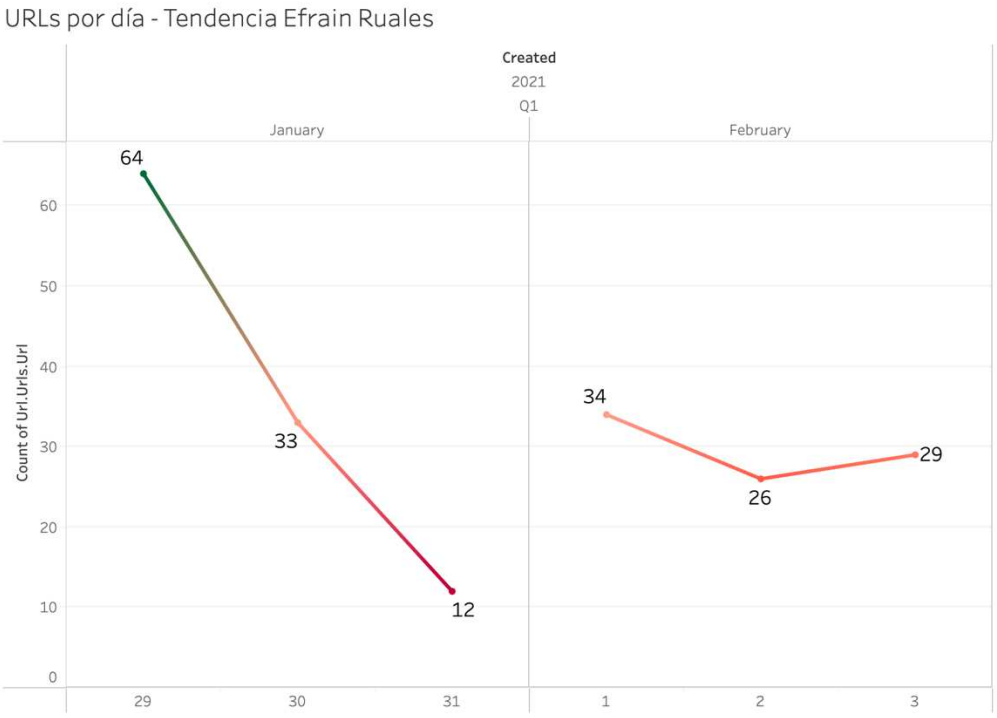


En esta grafica podemos ver que el número de URLs el día 12 de junio de 2021 es de 2098 en todo el día entre estas URLs tenemos noticias, anuncios de cuentas del gobierno y enlaces a páginas de interés en general.

Tendencia Efraín Rúales

Figura 26

Total, de URL en tendencia "EfrainRuales"



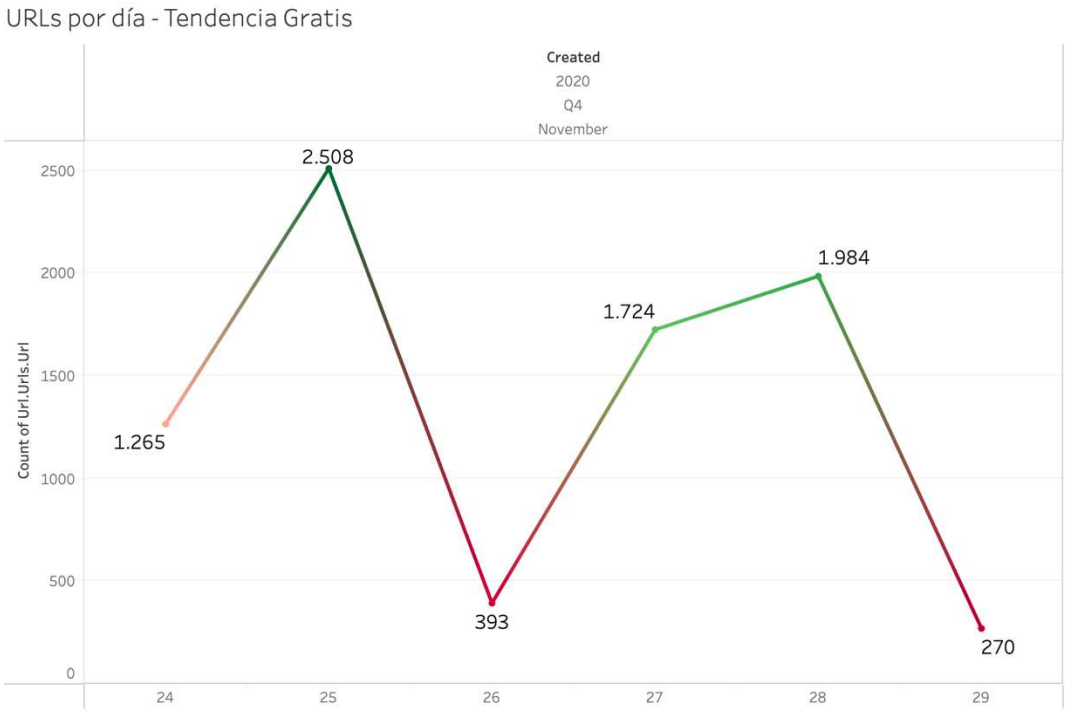
En esta gráfica se puede ver como el día 29 de enero se enviaron enlaces sobre la tendencia "Efraín Rúales", entre los cuales tenemos noticias, videos conmemorativos, enlaces a videos en vivo en Facebook, supuestas evidencias de los agresores.

En su mayoría para esta tendencia encontramos mensajes de apoyo a la familia, a sus compañeros de trabajo, videos de la persecución a los posibles victimarios, etc.

Tendencia Gratis

Figura 27

Total, de URL en tendencia "Gratis"

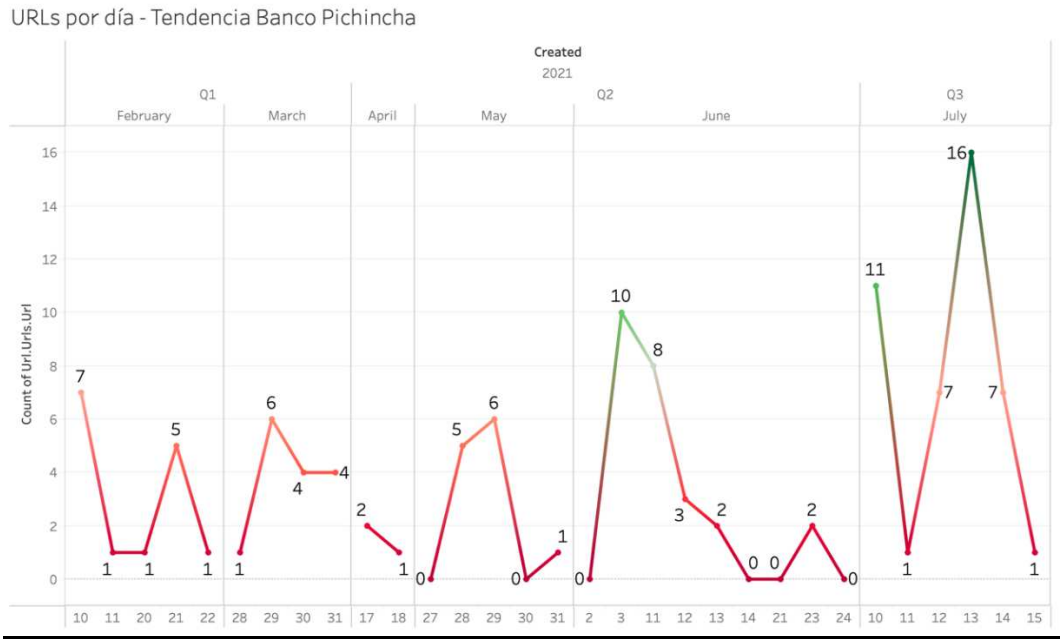


En esta grafica se observa que el día 25 de noviembre del 2020 la cantidad de promociones y enlaces que se enviaron fueron 2508, en este grupo existían descuentos en almacenes, anuncio de nuevos productos por el CyberMonday. No encontramos indicios de phishing o ingeniería social.

Tendencia Banco Pichincha

Figura 28

Total, de URL en tendencia “Gratis”



En esta grafica podemos ver que el día 13 de Julio se enviaron muchos enlaces, entre los cuales se encuentran investigaciones periodísticas, páginas web con supuestos datos privados de clientes del banco Pichincha, evidencias de la filtración de datos, comunicados del banco y demás. En esta categoría encontramos algunos enlaces de interese, los cuales los veremos más adelante.

Análisis KNN y distancia euclidiana

En cuanto al análisis realizado mediante el uso de KNN y la distancia euclidiana, cuyo proceso se encuentra detallado en el literal 3.1.4, obtuvimos los siguientes resultados.

Tabla 4

Resultados obtenidos – Tendencia "Banco Pichincha"

Contenido del tweet	Valor calculado por distancia euclidiana
Banco Pichincha cobra de un crédito letra 12 de 12 y resulta que aún te queda una última pendiente que te la clavan como 1 de 1 y eso que en la penúltima letra te dicen tiene que cancelar del seguro primero si no le puedo cobrar la cuota	50,45
NOVA = Banco Pichincha confirmen	42,95
Siempre detesté Banco Pichincha https://t.co/TomJrJ3eQI	42,11
Llevo más de tres meses sin trabajo estoy pagando mis deudas fiando al uno y al otro para pagar préstamos y tarjetas he enviado CV, pero ven mi edad y no me quieren dar trabajo todo es un caos.	41,01
Estaba indignado me acaban de suspender la cta. del Banco de Pichincha, todos mis ahorros de toda mi vida se fueron al carajo, confié en ellos y sin argumento alguno me suspenda la cuenta. He llorado a mares. Luego recordé que nunca he tenido cuenta en ese banco.	40,61
Optimista??? Con clientes del banco del pichincha que le robaron su dinero y. Nadie hace nada para ayudarlos. Con los. Profesores y médicos. Sin pago de su. Salario. Con la subida del precio de los combustibles con la delincuencia.	40,05
Leonardo Campana tiene menos ataque que el sistema de seguridad del Banco Pichincha	39,85
@MrLinkEc Qué pasó con el Banco Pichincha.	39,46
@ defensoría del pueblo mi cuenta del banco pichincha fue hackiada y el banco no me quiere responder por el valor sustraido por favor su ayuda porque nunca mendaron un correo si aceptaba el valor a transferir y el banco se lava las manos	38,01

Contenido del tweet	Valor calculado por distancia euclidiana
#Datos | Banco Pichincha, es investigado por lavado de dinero en Ecuador leer aquí: https://t.co/SiPqkqhDMs	37,85

Realizando el cálculo de la distancia euclidiana en los datos obtenidos con la tendencia “Banco Pichincha”, obtuvimos resultados importantes, como opiniones acerca del banco, una supuesta foto de perfil de la persona que “hackeo” al banco, además de una mención a @MrLinkEc, un investigador de seguridad.

Mediante el análisis por clasificación Knn, obtuvimos los siguientes resultados:

Tabla 5

Resultados obtenidos al variar el valor de k - Tendencia “Banco Pichincha”

Exactitud	Precisión	Valor de k[14]
0.7064	> 86%	1
0.6927	> 86%	2
0.7202	> 89%	3
0.7225	> 87%	4

Por lo que, se decidió utilizar k=3 para el análisis, con este valor se obtuvo una precisión mayor al 89%, es decir que el modelo puede discernir de mejor manera entre casos negativos y así obtener menos falsos positivos.

[14] K se define como un hiperparámetro para controlar el proceso de aprendizaje en un algoritmo seleccionado.

La matriz de confusión resultante con el valor de k especificado anteriormente es la siguiente:

Figura 29

Matriz de confusión y resultados de sensibilidad y especificidad Tendencia “Banco Pichincha”

```
          prediction
actual   CD  CI   D   I  MI   N  ND  NI  PI
    CD    2   0   0   3   2   0   3   0   8
    CI    0   0   0   3   8   1   0   1   0
    D     0   0   0   4   2   0   0   0   0
    I     1   0   0 182   7   0   9   1   5
    MI    4   0   0  12   5   0   5   1   7
    N     0   0   0   3   4   0   0   1   0
    ND    3   0   0   6   3   0   5   0   3
    NI    0   2   0   2   0   0   0 112   0
    PI    0   0   0   4   0   0   4   0   8

            Class: CD Class: CI Class: D Class: I Class: MI Class: N Class: ND Class: NI Class: PI
Sensitivity  0.200000  0.000000       NA   0.8311   0.16129 0.000000   0.19231    0.9655   0.25806
Specificity  0.962441  0.970046  0.98624   0.8940   0.92840 0.981609   0.96341    0.9875   0.98025
```

Tabla 6

Resultados importantes obtenidos mediante Knn - Tendencia "Banco Pichincha"

Contenido del tweet	Clasificación por Knn
NOVA = Banco Pichincha confirmen	MI (muy importante)
ECUATORIANOS a las AFUERAS del BANCO PICHINCHA ¿Dónde estgá el CUSQUI? https://t.co/1sGRdyk64V vía @YouTube	MI (muy importante)
Banco Pichincha indica que la seguridad de los recursos financieros de sus clientes "no está comprometida". Vea lo que se informó sobre el acceso no autorizado: https://t.co/bPRDWxdvpO https://t.co/qzLQLb9Q7P	MI (muy importante)
Buen día! El mensaje que nos muestra efectivamente no es oficial de Banco Pichincha sino un intento de FRAUDE. Por favor, reenvíenos el e-mail a	MI (muy importante)

Contenido del tweet	Clasificación por Knn
usrnocsm@pichincha.com para bloquear los links fraudulentos. Muchas gracias por alertarnos. Saludos Estafa ya que últimamente por la filtración de datos bancarios del banco Pichincha están estafando, así como robando dinero de las cuentas de los clientes	MI (muy importante)

Entre los tweets obtenidos en la tabla resultante se encuentran observaciones de los clientes y alertas del banco, en donde se explica que los mensajes recibidos no son enviados por la institución bancaria por lo que podría tratarse de una estafa.

Tabla 7

Resultados obtenidos mediante distancia euclidiana - Tendencia "Pastebin"

Contenido del tweet	Valor calculado por distancia euclidiana
A combo gaming/personal #pastebinmondayhttps://t.co/G9LJB3KbS4	65,17
Algunos sitios web de brechas de datos: https://t.co/CNXOvIo9sr https://t.co/DtzMSP4tBu	59,57
¡Otros 2000! https://t.co/QXKKqGjMS8	49,16
INFO: T.O.S https://t.co/x2FzuC69F9	42,19
Parece ser que parte del Monster Hunter Digital Event de mañana se filtró por parte de PraticalBrush en RedditSource: https://t.co/31rqkL2qOs https://t.co/vSIGUMRX4o	41,45
Código: https://t.co/41H1RKPgbn	39,01
mi lista de #tweaks en iOS 14.3para al que le interesehttps://t.co/e77obGNuXd#jailbreak #checkra1n #unc0ver #manticore	37,22
Hola, estimados para su conocimiento, uso y entretenimiento. NO olviden sus capturas. -Usar en su App de IPTV favorita. - Disfruten!!https://t.co/Dk6G6BV0nw	33,54
Phishing a clientes bancarios ????????#OpFreenom 03-02-2021142 dominios fraudulentos. Bloquear	31,02

Contenido del tweet	Valor calculado por distancia euclidiana
navegación a IP: 101.99.90.35https://t.co/BHY2pAZdW4#Phishing https://t.co/7Ua3EccZd3	
???? El reporte espacial n.° 793 ya está traducido al español. https://t.co/koISSSV2hS#traducciones #JSRenespañol https://t.co/XiaKtQoU6v	30,27

Al realizar el cálculo de la distancia euclidiana a los datos obtenidos de la tendencia Pastebin, obtuvimos, varia información sobre DLC[15] de videojuegos, un sitio web de filtración de datos, scripts y archivos de chat. Todo esto podría servir para realizar una investigación por archivo mediante la URL obtenida.

Mediante el análisis por clasificación Knn, obtuvimos los siguientes resultados:

Tabla 8

Resultados obtenidos al variar el valor de k - Tendencia "Pastebin"

Exactitud	Precisión	Valor de k[16]
0.1832	> 79%	1
0.1374	> 75%	2
0.1641	> 74%	3
0.1565	> 70%	4

Por lo que, se decidió utilizar k=1 para el análisis, con este valor se obtuvo una precisión mayor al 78%, es decir que el modelo puede discernir de mejor manera entre

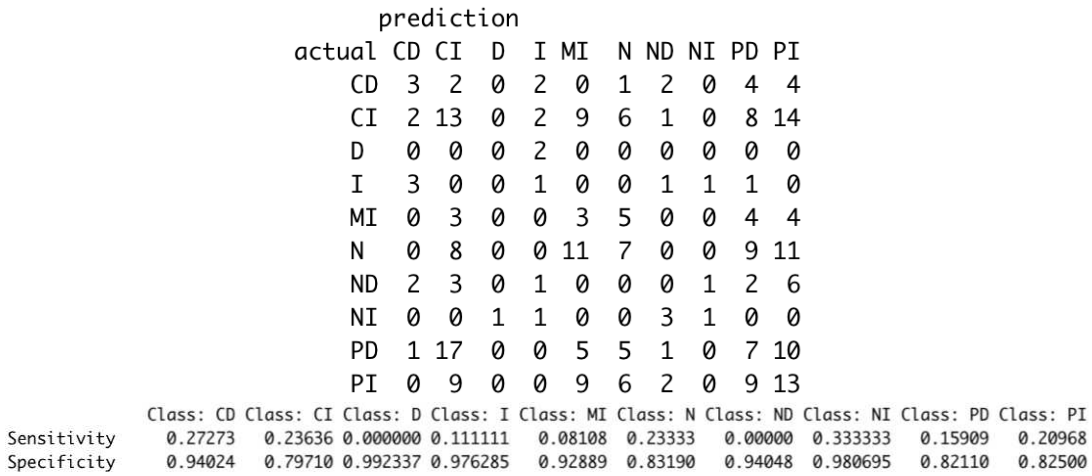
[15] DLC, proviene del acrónimo inglés Downloadable content, es decir contenido descargable
[16] k se define como un hiperparámetro para controlar el proceso de aprendizaje en un algoritmo seleccionado.

casos negativos y así obtener menos falsos positivos, pero disminuye su exactitud en cuanto a los valores de la diagonal en la matriz de confusión.

La matriz resultante con el valor de k especificado anteriormente es la siguiente:

Figura 30

Matriz de confusión y resultados de sensibilidad y especificidad Tendencia "Pastebin"



prediction

actual	CD	CI	D	I	MI	N	ND	NI	PD	PI
CD	3	2	0	2	0	1	2	0	4	4
CI	2	13	0	2	9	6	1	0	8	14
D	0	0	0	2	0	0	0	0	0	0
I	3	0	0	1	0	0	1	1	1	0
MI	0	3	0	0	3	5	0	0	4	4
N	0	8	0	0	11	7	0	0	9	11
ND	2	3	0	1	0	0	0	1	2	6
NI	0	0	1	1	0	0	3	1	0	0
PD	1	17	0	0	5	5	1	0	7	10
PI	0	9	0	0	9	6	2	0	9	13

	Class: CD	Class: CI	Class: D	Class: I	Class: MI	Class: N	Class: ND	Class: NI	Class: PD	Class: PI
Sensitivity	0.27273	0.23636	0.000000	0.111111	0.08108	0.23333	0.00000	0.333333	0.15909	0.20968
Specificity	0.94024	0.79710	0.992337	0.976285	0.92889	0.83190	0.94048	0.980695	0.82110	0.82500

Tabla 9

Resultados aplicados KNN – Tendencia "Pastebin"

Contenido del tweet	Clasificación por Knn
¡Otros 2000! https://t.co/QXKKqGjMS8	MI (muy importante)
???? El reporte espacial n.° 793 ya está traducido al español. https://t.co/koISSSV2hS#traducciones #JSRenespañol https://t.co/XiaKtQoU6v	MI (muy importante)
????????DESCARGA ESSENTIA VITAE - A POKEMON STORY BETA 1????????⭐Link al Pastebin con los enlaces: https://t.co/2dUoE3IsXg ⭐ https://t.co/bkjvp2D3Nf	MI (muy importante)
Servers activos de qakbot a día de hoy. https://t.co/5dTyB1zs39Recordar que son	MI (muy importante)

Contenido del tweet	Clasificación por Knn
máquinas infectadas funcionando como proxis #dbleaks BDs función pública del gobierno de Colombia vuln 1- https://t.co/iC9MthsmMb 2- https://t.co/sR76oZUtDy #Sin1pecrew #OpColombia #opcolombia #OpCol #Anonymous #ParoNacional13J #Colombia #ParoNacional12J https://t.co/q5w4qJDr6T	MI (muy importante)

Se obtuvieron resultados consistentes con el algoritmo de la distancia euclidiana, es decir, enlaces de DLCs de videojuegos, descargas de películas, posibles leaks[17], entre otros.

[17] Fuga o entrega de documentos sensibles a los medios de comunicación

Capítulo V
Conclusiones, Recomendaciones y Líneas de trabajos futuros

Conclusiones

- El Combinar algoritmos de minería de datos, permitió obtener resultados diferentes pero aproximados a la realidad. Ya que, en el caso de la tendencia “Banco Pichincha” se obtuvo información sobre el supuesto ataque informático que sucedió en febrero de 2021.
- En ciertos temas el prototipo no funciona correctamente ya que se debe incrementar y mejorar la calidad el diccionario de datos para la categoría que se quiere analizar.
- Los datos tienen que analizarse inmediatamente ya que se pudo ver como ciertos tweets no se encontraban disponibles después de 1 a 3 semanas.
- El uso de R Studio permitió la realización de este trabajo de forma más adecuada, ya que existen varias librerías que se adaptan a las herramientas propuestas y permitieron obtener un análisis eficiente.
- El uso del API estándar de Twitter es limitado en cuanto al tiempo de actualización de tweets, ya que se debe esperar 15 minutos por cada petición para obtener resultados actualizados y por lo tanto se pierde información.

Recomendaciones

- Para un análisis de datos mas efectivo se recomienda obtener un volumen mayor a 12000 tweets ya que la base de conocimientos será más extensa.
- Se recomienda automatizar el proceso de obtención de tweets mediante el uso del API en tiempo real.
- Se recomienda el uso de herramientas de integración continua como Codebuild[18], para facilitar el desarrollo e integración del proyecto.
- Se recomienda el uso de Lambdas[19] (AWS) para disminuir los costos de operación del proyecto.

Líneas de trabajo futuro

- En un futuro podríamos pensar en actualizar los scripts de obtención de información con el uso de observables, para poder subscribirnos a un flujo de datos continuo y así no perder información importante.
- Además, sería factible pasar la infraestructura a la nube, para abaratar costos y aumentar la eficiencia de nuestro sistema, ya que en nuestro desarrollo estuvimos limitados por la capacidad de cómputo de nuestros ordenadores domésticos.

[18] AWS CodeBuild es un servicio de integración continua que ejecuta un flujo de compilación de código, incluyendo pruebas e instalación de paquetes.
[19] AWS Lambda es un servicio informático que no requiere de un servidor aprovisionado y ejecuta acciones basadas en eventos.

- Al realizar este trabajo mediante un algoritmo supervisado, creemos conveniente pasarlo a un algoritmo no supervisado con inteligencia artificial, tomando en cuenta las limitantes en cuanto a costos.

Bibliografía

Adeva, R. (2021). *Pastebin: qué es y las mejores alternativas para compartir texto o código online*. https://www.adslzone.net/listas/mejores-webs/compartir-texto-codigo-online/

Aguado-de-Cea, G., Barrios Rodríguez, M. A., Bernardos, M. S., Campanella, I., Montiel-Ponsoda, E., Rodríguez, V., & Muñoz-García, Ó. (2013). Análisis de Sentimientos en un Corpus de Redes Sociales. *31st AESLA International Conference*.

Alami, S., & Elbeqqali, O. (2015, December). Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts. *2015 10th International Conference on Intelligent Systems: Theories and Applications, SITA 2015*. https://doi.org/10.1109/SITA.2015.7358435

Aslam, A., Qamar, U., Khan, R. A., Saqib, P., Ahmad, A., & Qadeer, A. (2019). Opinion mining using live Twitter data. *Proceedings - 22nd IEEE International Conference on Computational Science and Engineering and 17th IEEE International Conference on Embedded and Ubiquitous Computing, CSE/EUC 2019*, 36–39. https://doi.org/10.1109/CSE/EUC.2019.00016

BID, & OEA. (2020). CIBERSEGURIDAD, Riesgos, Avances y el camino a seguir en America Latina y El Caribe. *Bid- Oea*, *1*, 116–119. https://publications.iadb.org/publications/spanish/document/Reporte-Ciberseguridad-2020-riesgos-avances-y-el-camino-a-seguir-en-America-Latina-y-el-Caribe.pdf

Bisio, F., Meda, C., Zunino, R., Surlinelli, R., Scillia, E., & Ottaviano, A. (2015). Real-time monitoring of Twitter traffic by using semantic networks. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 966–969. https://doi.org/10.1145/2808797.2809371

Cataldo, A. (2015). Design Science Research (DSR): Una breve introducción. *II Workshop RedSTI*, *February*, 56.

Charitonidis, C., Rashid, A., & Taylor, P. J. (2015). Weak signals as predictors of real-world phenomena in social media. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 864–871. https://doi.org/10.1145/2808797.2809332

Choudhary, M., & Choudhary, P. K. (2018). Sentiment analysis of text reviewing algorithm using data mining. *Proceedings of the International Conference on Smart Systems and Inventive Technology, ICSSIT 2018*, 532–538. https://doi.org/10.1109/ICSSIT.2018.8748599

Choudhary, M. M., & Choudhary, P. K. (2015). Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study. *International Journal of Computer Applications*, *112*(5), 44–48.

Cisco. (2020). *Informe anual de Internet de Cisco - Informe anual de Internet de Cisco (2018-2023) Informe técnico - Cisco*. Informe Anual de Internet de Cisco (2018-2023) Informe Técnico. https://www.cisco.com/c/en/us/solutions/collateral/executive-

perspectives/annual-internet-report/white-paper-c11-741490.html
Code. (2015). *Introducción al Arte de las Ciencias de la Computación*.
Crowdstrike. (2019). WHAT IS MALWARE (MALICIOUS SOFTWARE). *Crowdstrike*. https://www.crowdstrike.com/cybersecurity-101/malware/?utm_campaign=dsa&utm_content=nola&utm_medium=sem&utm_source=goog&utm_term=&gclid=Cj0KCQjwh_eFBhDZARIsALHjIKdoUzrsgRoUqhzPZrQa7NwjNc-LHXnvDuWwaDPrw7vHeMjFwWG3XFEaApJFEALw_wcB
Cycle, S., Team, T. S., Terms, C. U., Backlog, P., Backlog, S., Testing, U., Owner, P., Meeting, S. P., Meeting, S. R., Software, A., Ecosystems, D., Work, D. A., & Meetings, A. (2007). *What is SCRUM ?* 1–7. https://www.scrum.org/resources/what-is-scrum
Dastanwala, P. B., & Patel, V. (2016). A review on social audience identification on twitter using text mining methods. *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016*, 1917–1920. https://doi.org/10.1109/WiSPNET.2016.7566476
Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y., & Liu, Q. (2019). Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. *IEEE Access*, *7*, 106111–106123. https://doi.org/10.1109/ACCESS.2019.2930410
Franceschi-Bicchierai, L. (2020). *VICE - Pastebin Made It Harder To Scrape Its Site And Researchers Are Pissed Off*. https://www.vice.com/en_us/article/y3m83v/pastebin-made-it-harder-to-scrape-its-site-and-researchers-are-pissed-off
Gawron, O., González, N., & Lage, F. (2014). Análisis de métodos para el reconocimiento de patrones en ECG. *Memórias Do WTA 2014*, 42–45.
Gene V. Glass, J. C. S. (1986). Metodos estadísticos aplicados a las ciencias sociales. In *Metodos estadísticos aplicados a las ciencias sociales*. Prentice-Hall. https://www.urbe.edu/UDWLibrary/InfoBook.do?id=3101
Gu, N., Sun, D. Y., Li, B., & Li, Z. (2016). Sentiment Analysis for Topics based on Interaction Chain Model. *Proceedings - 2015 European Intelligence and Security Informatics Conference, EISIC 2015*, 133–136. https://doi.org/10.1109/EISIC.2015.23
IBM. (2021). *Análisis vecino más cercano - Documentación de IBM*. SPSS Statistics. https://www.ibm.com/docs/es/spss-statistics/SaaS?topic=features-nearest-neighbor-analysis
Jara- Obregón, L. S., Ferruzola-Gomez, E., & Rodríguez-López, G. (2017). Delitos a través redes sociales en el Ecuador: una aproximación a su estudio. *RIDTEC | Vol. 13, n.° 2*, *13*(2).
Julio Cabero Almenara. (2002). Impacto de las nuevas tecnologias de la informacion y la comunicacion en las organizaciones educativas. In *Grupo Editorial Universitario*.
Kemp, S. (2021a). *Digital in Belgium: All the Statistics You Need in 2021 — DataReportal – Global Digital Insights*. Kepios Pte. Ltd., We Are Social Ltd. and Hootsuite Inc. https://datareportal.com/reports/digital-2021-ecuador
Kemp, S. (2021b). *Digital2021_GlobalReport_en.pdf*. https://hootsuite.widen.net/s/zcdrtxwczn/digital2021_globalreport_en
Lerena, O. (2019). *Métodos y aplicaciones de la ciencia de datos para las políticas de CTI, vol. 1 - Redes sociales, minería de textos y clustering*.
Lu, X., Zhou, X., Wang, W., Lio, P., & Hui, P. (2020). Domain-oriented topic discovery based on features extraction and topic clustering. *IEEE Access*, *8*, 93648–93662. https://doi.org/10.1109/ACCESS.2020.2994516
Lubeck, L. (2020). *Programa "Quédate en casa": un engaño que busca robar información de los usuarios | WeLiveSecurity*. https://www.welivesecurity.com/la-

es/2020/04/29/programa-quedate-casa-engano-busca-robar-informacion-usuarios/
Lwowski, B., Rad, P., & Choo, K.-K. R. (2018). Geospatial Event Detection by Grouping Emotion Contagion in Social Media. *IEEE Transactions on Big Data*, *6*(1), 159–170. https://doi.org/10.1109/tbdata.2018.2876405
Nava, S. (2006). Reconocimiento De Patrones. *Luis Enrique Erro No. 1, Tonantzintla, Pue., C.P. 72840, M´ Exico,* 4.
Niakanlahiji, A., Safarnejad, L., Harper, R., & Chu, B. T. (2019). IoCMiner: Automatic Extraction of Indicators of Compromise from Twitter. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 4747–4754. https://doi.org/10.1109/BigData47090.2019.9006562
Nielsen, F. Å., Informatics, D. T. U., & Dinamarca, U. T. De. (n.d.). *análisis de sentimiento en microblogs*.
OpenJS Foundation. (2020). *About Node.js*. OpenJS Foundation. https://nodejs.org/en/about/
Ortí, C. B. (2012). *LAS TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN (T.I.C.)*. Universidad de Valencia.
PANDASECURITY. (2021). *Exploit: ¿sabes qué es y cómo funciona? - Panda Security*. https://www.pandasecurity.com/es/security-info/exploit/
Ramirez-Morales, I., Mazon-Olivo, B., & Pan, A. (2018). *Capítulo 1: Ciencia de datos en el sector agropecuario* (pp. 12–44).
Riahi, Y., & Riahi, S. (2018). Big Data and Big Data Analytics: concepts, types and technologies. *International Journal of Research and Engineering*, *5*(9), 524–528. https://doi.org/10.21276/ijre.2018.5.9.5
Rodriguez, A., & Okamura, K. (2019). Generating real time cyber situational awareness information through social media data mining. *Proceedings - International Computer Software and Applications Conference*, *2*, 502–507. https://doi.org/10.1109/COMPSAC.2019.10256
Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600
Stedman, C. (2021). *¿Qué es Análisis o analítica de datos? - Definición en WhatIs.com*. https://searchdatacenter.techtarget.com/es/definicion/Analisis-de-Datos
Vallejos, S. (2006). *Universidad Nacional del Nordeste Facultad de Ciencias Exactas, Naturales y Agrimensura*.
Vamshi, K. B., Pandey, A. K., & Siva, K. A. P. (2018, August). Topic Model Based Opinion Mining and Sentiment Analysis. *2018 International Conference on Computer Communication and Informatics, ICCCI 2018*. https://doi.org/10.1109/ICCCI.2018.8441220
van Aken, J. E., & Romme, A. G. L. (2012). A Design Science Approach to Evidence-Based Management. In *The Oxford Handbook of Evidence-Based Management*. https://doi.org/10.1093/oxfordhb/9780199763986.013.0003
Viso Gurovich, E. (2006). Introducción a ciencias de la computación. *Facultad de Ciencias, UNAM*.
Wikipedia. (n.d.). *Mega (sitio web) - Wikipedia, la enciclopedia libre*. Retrieved July 12, 2021, from https://es.wikipedia.org/wiki/Mega_(sitio_web)
Zambrano, D., Roman, D., & Zambrano, M. (2017). Innovación para el Análisis de Sentimientos en Texto, una revisión de la técnica actual aplicando metodologías de crowdsourcing. *Economía y Desarrollo*, *158*.

Zurkus, K. (2016). *Social media, the gateway for malware*. CSO.