TESINGIN-PT1.0.1 – Análisis & modelamiento de datos psicográficos.
– Tesis / Proyecto de Título –[1]

Felipe Alfonso González López
Tesis final carrera de Continuidad en, Ingeniería en Informática, IACC Chile., 2020-2022

Análisis & modelamiento de datos psicográficos en internet utilizando machine learning para organizaciones, empresas privadas & gobiernos.

Instituto Superior de Artes y Ciencias de la Comunicación, IACC
Av. Salvador 1318, Metro Santa Isabel, Providencia, Santiago.
Chile.
f.alfonso@res-ear.ch – felipe.alfonso.glz@gmail.com – https://twitter.com/felipealfonsog
https://glrsrch.com - https://bit.ly/2zLMkEE3 - https://linkedin.com/in/felipealfonsog

Nombre Profesora Guía: María Lourdes Geizzelez Luzardo
Nombre Del Alumno: Felipe Alfonso González López.

Septiembre – Noviembre 2022

Palabras Clave: Big Data, Análisis de Datos, Machine Learning, Software, Internet.

Índice

Dedicatoria

Dedicado a mi querida Madre Gloria López Apablaza, que siempre ha creído en mí. A mi Padre Alfonso González Marquez, que siempre fue una inspiración también. A mi hermano Ponchi, a mi tía Veronica López, a mi prima Grisel B. López, y mi perrita, Aimy.

Abstract

Éste es un tema de gran relevancia y tendencia en Internet hoy en día, ya que de esto depende la continuidad operacional y a veces también de tipo estratégica de organizaciones y gobiernos si no se cuenta con herramientas que permitan desarrollar piezas de software para el análisis de datos; que en determinadas ocasiones podrían llevar al punto en que una empresa u organización o incluso partido político pueda caer en una escalada de fallos. Estos fallos pueden ser de tipo comunicacionales con los mismos usuarios que se conectan con la empresa, organización, etc. Toda esta idea permite guiar a los mismos usuarios a tomar ciertas determinaciones y decisiones para que influyan en la sociedad, así como lo es en "la sociedad de Internet". A esto se le denomina manejo y manipulación de masas en Internet de manera coloquial, aunque es derechamente análisis psicográfico, y gracias a la ciencia de datos, la Big Data y Machine Learning o aprendizaje de máquinas y también a piezas de software que permiten procesar y entregar un análisis completo de datos psicográficos y de comportamiento en los usuarios en Internet y/o "Sentiment Data", que pueden estar en redes sociales como Twitter (Que para esta investigación será utilizada a modo de ejemplo), o en cualquier otra forma de interacción en Internet.

1. El Problema

Hoy en día es complejo encontrar sistemas que permitan administrar y controlar gran cantidad de datos e información que fluye en Internet y este es el problema con el que se enfrentan muchas organizaciones debido a que existen masas que se mueven, a gran velocidad y en un flujo muy grande. Esta gran cantidad de datos por lo general es algo que se transforma en un gran dilema en organizaciones como partidos políticos o empresas privadas o incluso gobiernos. Ofrecer piezas de software que analizan datos, y que modelen toda esta información que se produce, y se genera de una forma muy gigantesca, requiere escalar estos datos de manera adecuada y fajo ciertos modelamientos, para que las organizaciones o gobiernos tomen buenas decisiones.

1.1 Planteamiento Del Problema

- ¿Para qué?

Para dar seguimiento, control y administración del flujo de datos en las masas en Internet que puedan influir en el producto, imagen o idea de una organización o gobierno.

- ¿Por qué?

¿Por qué es muy necesario comprender cómo funcionan los datos en Internet hoy en día?, simple, esto permite que se tomen determinaciones y decisiones positivas para una organización.

- ¿Beneficiarios?

Se benefician los gobiernos, organizaciones y empresas privadas en general. Permite llevar un control de procesos de comunicaciones, ideas y sus flujos, entre usuarios y las organizaciones.

- ¿De qué modo?

Creando piezas de software que permiten analizar mediante ciencia de datos y aprendizaje de máquina; también todo esto es para acelerar estos procesos. Estas piezas de software permitirán entregar un análisis general y completo de lo que se necesita visualizar.

- Proyección Social

Esto permite que los usuarios se sientan valorados además ayuda al orden en las organizaciones y a la optimización de estas y tener una claridad total respecto a cuáles van hacer las determinaciones y decisiones que un grupo social necesite.

- ¿Qué resuelve?

Resuelve problemas derivados en la construcción de ideas y de comunicaciones hacia determinados grupos psicosociales, grupos de usuarios en Internet.

- ¿Qué permitirá?

Permitir dar una solución rápida y oportuna a problemas en donde se requiera una solución rápida para dirigir determinados temas en grupos sociales, personas etc., que estén involucrados en determinados medios o redes en Internet.

1.1.1 Objetivo General

Desarrollar piezas de software que permitan facilitar la compresión de datos en Internet relativo a grandes masas de usuarios.

Todo esto permitirá facilitar la compresión de datos en Internet relativo a grandes grupos de usuarios, y que permitan también realizar análisis y modelamiento en relación a la ciencia de datos detrás del análisis que se realizará a grupos y personas en Internet en general todos los disponibles (Usuarios de redes sociales, etc.). Todo esto con el fin de entregar ideas precisas para que empresas, organizaciones o gobiernos tomen un curso de acción definido y claro para posicionar sus intenciones.

1.1.2 Objetivos Específicos

Analizar y actualizar el flujograma del proceso de análisis con la empresa u organización.
Construir un determinado software para el cliente y/o piezas de software que funcionen por ejemplo en una terminal en el servidor y que permitan construir modelos estadísticos.
Verificar el funcionamiento de los scripts en los distintos terminales en el servidor.

- Analizar la factibilidad de los datos en relación a las piezas de software que se van a crear para generar un seguimiento, y así para generar a su vez también análisis y se den cumplimiento a los estándares que se requieran en su minuto.

- Estudiar la factibilidad técnica y económica para el diseño de software y/o piezas de software que estarán funcionando en un servidor 24/7 eventualmente o cada cierta cantidad de tiempo bajo una determinada operación realizada por un ingeniero, y el cual permitirá realizar todos los

requerimientos necesarios para el cumplimiento con la organización.

- Desarrollar un software o piezas de software que permitan tanto a la empresa que desarrolla las piezas de software como al cliente, el monitorear y vigilar el análisis que se estarán realizando mediante Bots, así como a la organización a la cual se le entregará el servicio para realizar el análisis de datos, etc.

1.1.3 Hipótesis

Uno de los recursos más importante en Internet son los usuarios, las personas han permitido que en el tiempo se haya desarrollado un área conocida como ciencia de datos, en donde se analizan grandes cantidades de información; esto en la ingeniería de datos, permite precisamente analizar el comportamiento de los usuarios y estos pasan a ser un recurso muy importante, tal como en una mina, puede ser el oro o el cobre. Para realizar todo este análisis se requieren procedimientos y piezas de software. Todos estos serán realizados por una empresa que entregue estos servicios de manera eficiente e idónea.

Existen elementos que siempre van generando cambios en los usuarios como pueden ser las estructuras y algoritmos en las diferentes redes sociales o los mismos algoritmos que se pueden desarrollar en las piezas de software que permitirían generar un análisis completo de determinados incidentes que requieran manejar las organizaciones, empresas, gobiernos y partidos políticos, enfocándose en las masas en Internet.
Al encontrar situaciones anormales, refiriéndose a estos como datos que requieran ser reenfocados o re-direccionados, se podrán tomar estas como situaciones positivas ya que permitirán hacer un análisis en base a los errores y con estos se generará, mediante un algoritmo utilizando matemática y estadística, una mejor utilidad para generar reportes y así permitir que se resuelva una necesidad para una empresa o gobierno. Mediante estos sistemas se logrará realizar todos estos procesos permitiendo inferir si así fuese necesario lo que sería la revisión, la acción y el análisis, así como su gestión o cualquier ejecución necesaria para concluir de manera exitosa alguna acción tomada por la empresa privada, organización o gobierno.

En definitiva:
La premisa fundamental es que el sistema, y en toda su implementación entregará datos y modelos para hacer los ajustes de comunicación que afectan a determinados usuarios en Internet. Las personas o usuarios han permitido que un campo conocido como ciencia de datos se desarrolle con el tiempo, donde se almacenan grandes cantidades de información y a su vez permiten el análisis de estos.
En ingeniería de datos, esto permite precisamente analizar su comportamiento, y se convierte en un recurso importante, tal como se dijo antes de manera retórica, como oro o cobre en una mina.
Para realizar todos estos escaneos, se necesitan procedimientos. Todo realizado por una empresa que lo entrega de manera eficiente y adecuada.

1.2 Justificación De La Investigación

Una específica y certera metodología de investigación seleccionada permitirá reunir y obtener los datos correspondientes para su posterior procesamiento. El resultado entregará la información del estado de los datos en el momento actual como sistema de información, lo ideal es que exista un método a través de piezas de software que realice todos estos procesos de manera automática usando aprendizaje de máquina o machine learning.
Los principales detalles que se van a necesitar en torno a los respectivos análisis de datos y el procedimiento a realizar serán seguidos en una constante, esto para dar solución a la problemática planteada, y que hoy en día es un mundo emergente, en nuestra sociedad tecnologizada.
La orientación de esta investigación se enfocará en buscar falencias en los métodos como las empresas se relacionan con sus usuarios. Se involucra también a la organización para el seguimiento de los futuros procesos.
Además de implementar estos procedimientos y sistemas, a esta se realizaría un tipo de investigación cuantitativa ya que se presentarán los resultados de todo lo que se analizaría en relación a los usuarios que están involucrados con la organización y la naturaleza de las investigaciones de tipo descriptiva, ya que se describe en detalle todos los elementos que componen el sistema implementado y que se exponga, dentro de la organización.
En este sentido también, se plantea que en este caso el enfoque de esta investigación es cuantitativa pues se basa en el estudio de la realidad de estos procesos, cuantificando los elementos que intervienen en el sistema que se quiere implementar y aportando una posible solución. Por otro lado, esta investigación es del tipo descriptiva / explicativa pues, por un lado, se miden las dimensiones del problema planteado, describiendo lo que se investiga y por otro lado se debe explicar el fenómeno donde poder ofrecer una solución a la dinámica que se plantee.

1.2.1 Técnica De Investigación

En este proyecto, la técnica de investigación será de tipo no experimental ya que se realizará un estudio, que se aplicará a los datos con el problema actual de la organización, y el diseño del sistema a implementar, en la cual consistirá en observar los procedimientos que se tienen actualmente para dar solución al problema.

2. Marco Teórico

Ahora, me referiré al marco teórico del proyecto.
La investigación muestra que la aplicación para el análisis de datos, debe desarrollarse en un sistema operativo como Mac o Linux, utilizando terminales con Python, bajo un sistema operativo idealmente Linux (Debian), idealmente bajo mi mirado es mucho mejor un sistema operativo basado en UNIX, ya que es más robusto Windows bajo mi investigación y análisis, todo esto ya que necesita la capacidad de extenderse a diferentes partes del país o del planeta, para esto se toman referencias, además es ideal contar con un servidor de BBDD como MySQL, ya que idealmente se pueden guardar los datos recolectados en esta. MySQL ha mostrado a través de su historia también mucha robustez, es un eficiente motor de base de datos.
Las referencias a estos conceptos se presentan a

continuación:

Python[2]: Python es un lenguaje de alto nivel de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código, se utiliza para desarrollar aplicaciones de todo tipo, ejemplos: Instagram, Netflix, Spotify, Panda 3D, entre otros.2 Se trata de un lenguaje de programación multiparadigma, ya que soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida[¿cuál?], programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma. Administrado por Python Software Foundation, posee una licencia de código abierto, denominada Python Software Foundation License.3 Python se clasifica constantemente como uno de los lenguajes de programación más populares.

MySQL[3]: MySQL es un sistema de gestión de bases de datos relacional desarrollado bajo licencia dual: Licencia pública general/Licencia comercial por Oracle Corporation y está considerada como la base de datos de código abierto más popular del mundo,12 y una de las más populares en general junto a Oracle y Microsoft SQL Server, todo para entornos de desarrollo web.
MySQL fue inicialmente desarrollado por MySQL AB (empresa fundada por David Axmark, Allan Larsson y Michael Widenius). MySQL AB fue adquirida por Sun Microsystems en 2008, y ésta a su vez fue comprada por Oracle Corporation en 2010, la cual ya era dueña desde 2005 de Innobase Oy, empresa finlandesa desarrolladora del motor InnoDB para MySQL.
Al contrario de proyectos como Apache, donde el software es desarrollado por una comunidad pública y los derechos de autor del código están en poder del autor individual, MySQL es patrocinado por una empresa privada, que posee el copyright de la mayor parte del código. Esto es lo que posibilita el esquema de doble licenciamiento anteriormente mencionado. La base de datos se distribuye en varias versiones, una Community, distribuida bajo la Licencia pública general de GNU, versión 2, y varias versiones Enterprise, para aquellas empresas que quieran incorporarlo en productos privativos.
Las versiones Enterprise incluyen productos o servicios adicionales tales como herramientas de monitorización y asistencia técnica oficial. En 2009 se creó un Fork denominado MariaDB por algunos desarrolladores (incluido algunos desarrolladores originales de MySQL) descontentos con el modelo de desarrollo y el hecho de que una misma empresa controle a la vez los productos MySQL y Oracle Database.

Está desarrollado en su mayor parte en ANSI C y C++.4 Tradicionalmente se considera uno de los cuatro componentes de la pila de desarrollo LAMP y WAMP. MySQL es usado por muchos sitios web grandes y populares, como Wikipedia, Google (aunque no para búsquedas), Facebook, Twitter, Flickr y YouTube.

Debian GNU/Linux[4]: Debian GNU/Linux es un sistema operativo libre, desarrollado por miles de voluntarios de todo el mundo, que colaboran a través de Internet.
La dedicación de Debian al software libre, su base de voluntarios, su naturaleza no comercial y su modelo de desarrollo abierto la distingue de otras distribuciones del sistema operativo GNU. Todos estos aspectos y más se recogen en el llamado Contrato Social de Debian. Debian se caracteriza por no tener las últimas novedades en GNU/Linux, pero sí tener el sistema operativo más estable posible. Esto se logra por medio de paquetes y librerías antiguas, pero con muchos meses de pruebas, asegurando la máxima estabilidad por cada versión que es lanzada por la comunidad de Debian.
Nació en el año 1993, de la mano del proyecto Debian, con la idea de crear un sistema GNU usando Linux como núcleo. El proyecto Debian es la organización responsable de su mantenimiento en la actualidad, y también desarrolla sistemas GNU basados en otros núcleos (Debian GNU/Hurd, Debian GNU/NetBSD y Debian GNU/kFreeBSD).

Uno de sus principales objetivos es separar en sus versiones el software libre del software no libre. El modelo de desarrollo es independiente a empresas, creado por los propios usuarios, sin depender de ninguna manera de necesidades comerciales. Debian no vende directamente su software, sino que lo pone a disposición de cualquiera en Internet, aunque sí permite a personas o empresas distribuir comercialmente este software mientras se respete su licencia.
Debian GNU/Linux puede utilizar distintos mecanismos de instalación, como son: DVD, CD, USB, e incluso directamente desde la red (este último depende de la velocidad de la red del usuario).

3. Marco Metodológico

Primero, el desarrollo de una encuesta no se limita a la definición del contenido temático o su distribución, por lo tanto, es imperativo llevar estos contenidos a través de una metodología que se oriente hacia el logro de los objetivos planteados.
En este presupuesto de investigación, las bases deben ser el detalle de los procedimientos necesarios para obtener la información requerida, así de esta manera poder estructurar o resolver el problema planteado. El proceso de investigación comienza cuando uno se hace una pregunta de la que no sabe la respuesta y que debe ser contestada. Para hacer esto, se debe planificar dicha investigación, lo que implica atravesar un conjunto de fases preliminares, definir el tipo de estudio y delinear los componentes. En el caso del proyecto en cuestión, el enfoque para resolver el problema planteado respecto a la creación de piezas de software para monitorear los respectivos modelos y análisis es ante todo una “conversación con los usuarios involucrados”, de los trabajadores y la dirección, para poder integrar elementos necesarios para el desarrollo informático, luego realizar juntos el diseño visual de la aplicación si fuera necesario, y luego técnicamente analizar la programación para su uso en desarrollo, servidores, sistemas de comunicación, etc. Con esto claro se define la creación de todo desarrollo.

[2] Descripción tomada desde Wikipedia - Referencia: https://bit.ly/2GuKE4N
[3] Descripción tomada desde Wikipedia - Referencia: https://bit.ly/3eh70IK

[4] Descripción tomada desde Wikipedia - Referencia: https://bit.ly/3RwbF80

3.1 Características Del Sistema A Desarrollar Durante La Investigación

Son piezas de software que se enfocan en el análisis de datos, estas piezas de software pueden bien funcionar bajo una terminal, siendo ejecutadas en un servidor Mac OSX y fundamentalmente en Debian GNU/Linux.

3.1.1 Metodologías Y Fundamentaciones De Desarrollo Planteadas

En el proyecto se utilizará el modelo iterativo del ciclo de vida en cascada. Este modelo pretende reducir el riesgo de que las necesidades del usuario y el producto final puedan surgir de malentendidos durante la fase de recopilación de requisitos.

Consiste en la iteración de varios ciclos de vida en cascada. Al final de cada iteración se le entrega al cliente una versión mejorada o con mayores funcionalidades de cualquier producto informático ofrecido. En este caso este, tal como se ya he mencionado antes, corresponde a piezas de software que tendrán un funcionamiento en Terminales de un servidor Debian GNU/Linux.
El cliente es aquel que después de cada iteración evalúa el producto y lo corrige o sugiere mejoras, respecto a los resultados entregados, no al funcionamiento que se pueda entregar en sí misma, sino basándonos en determinadas piezas de software producidas.

Estas iteraciones se repetirán hasta obtener piezas que satisfagan cualquier tipo de necesidad, pero por sobre todo a las piezas de software privativas, ya que el código fuente de cada una de las específicas piezas de software no se entregan para este caso, solo se entrega como producto, los resultados que estos entregaran.
En concreto esta metodología nos entrega el modo ideal para poder trabajar en nuestro sistema ya que cada pieza de software en sí, es una versión en sí misma. Tal como lo es por ejemplo la producción del Kernel de GNU/Linux y su constante desarrollo.

Este modelo generalmente se usa en proyectos en los que los requisitos no son claros para el usuario, por lo que es necesario crear diferentes prototipos para presentarlos y obtener el cumplimiento del cliente o de una investigación. Una de las principales ventajas que ofrece este modelo es que los requisitos no tienen que definirse completamente durante el desarrollo, sino que se pueden refinar en las iteraciones.

Al igual que otros modelos similares, tiene la ventaja de realizar desarrollos en ciclos pequeños, lo que permite mejores riesgos, mejor gestión de las entregas de cada versión en las piezas de software.

3.1.2 Tipo De Investigación

La elección de una metodología de investigación específica y precisa permitirá recopilar y obtener los datos correspondientes para su análisis posterior.
El resultado proporcionará la información sobre el estado de los datos en el momento actual como información del sistema, idealmente hay un método a través del software que ejecuta automáticamente todos los procesos utilizando el aprendizaje automático.
Los principales detalles que van a ser necesarios estarán en torno a los respectivos datos y el procedimiento a seguir constantemente, esto para solucionar el problema de manejo de grandes cantidades de datos, en lo que hoy es un mundo emergente en nuestra sociedad.
La orientación de esta investigación se centrará en la investigación sobre los métodos de relación de las empresas con sus usuarios y de investigación.
La organización también está involucrada en el seguimiento de los procesos
En este sentido, también es menester especificar que en este caso, el objeto de esta investigación es cuantitativa, ya que se basa en la realidad de estos procesos, cuantificando los elementos que intervienen en el sistema a poner en marcha y una posible solución.
Por otro lado, esta investigación es del tipo descriptiva-explicativa porque, por un lado, miden las dimensiones del problema planteado, describiendo a quién se investiga y, por otro lado, el fenómeno debe explicar dónde se puede ofrecer una solución a la dinámica que surge.

3.1.3 Recolección De Datos

La recopilación de datos se refiere al enfoque sistemático de recopilar y medir información de varias fuentes para obtener una imagen completa en diferentes áreas de interés. La recopilación de datos permite a un individuo o una empresa responder preguntas relevantes, evaluar resultados y anticipar mejor probabilidades y futuros.
La precisión de la recopilación de datos es esencial para garantizar la integridad del estudio, las decisiones comerciales sólidas y la calidad.

Por ejemplo, los datos se pueden recopilar a través de dispositivos móviles, visitas a sitios web, respuestas en redes sociales, programas de fidelización y encuestas en línea para obtener más información sobre los clientes.
Existen diferentes métodos y técnicas de recopilación que pueden resultar útiles.
La elección del método depende de la estrategia, el tipo de variable, la precisión deseada, el punto de recolección y las habilidades del entrevistador. Las entrevistas son uno de los métodos más comunes. Si se decide hacerlo, se requiere especial atención en las preguntas que le harán, que también pueden ser bajo una entrevista cara a cara, vía telefónica e incluso por correo electrónico o vía redes sociales.

- Los cuestionarios son una herramienta útil para la recopilación de datos.
- Para obtener los resultados esperados, deben llevarse a cabo con cuidado.
- Es por esto que antes de escribirlo, es importante que el investigador defina los objetivos de la investigación.
- Existen dos formatos de cuestionario: cuestionarios que se aplican cuando se desea conocer la opinión de sus experiencias y sentimientos sobre un tema específico.

- Por el contrario, en el cuestionario cerrado, los investigadores controlan que preguntan y quieren saber, lo que puede obligar a limitar las respuestas de los participantes.
- Si preferimos la observación in situ para conocer a sus clientes, se pueden realizar de acuerdo a otras metodologías.
- Teniendo en cuenta cómo la información se encuentra, será de gran ayuda a la hora de analizarla.
- Poder medir y presentar informes de datos precisos y reales también es muy importante para una correcta toma de decisiones.

3.1.4 Instrumento De Recolección De Datos Utilizado

Cuestionario con las siguientes preguntas:
Cuéntanos sobre tu impresión y cómo te sientes frente a la organización.
La primera pregunta en el cuestionario del cliente o usuarios, incluye toda la información básica sobre estos, que consta de:

- *Nombre del cliente o empresa o de usuarios - personas.*

Datos importantes de la persona (Se realiza una investigación completa psicográfica del usuario, donde vive, que dice, como lo dice, etc.)
Datos de contacto, nombres de usuario, etc.

- ¿Cuáles son sus clientes objetivo?

Esto debe incluir todos los datos demográficos aplicables, sentimientos expresados, imágenes publicadas, etc. Comprender el tipo de público del cliente dará una idea de qué elementos de modelamiento usar.

- ¿Tienes un sistema actualmente?

Tendrá que evaluar si ya hay algún sistema que pueda ser funcional con los objetivos a realizar en la organización, compararlo con los objetivos de la empresa para ver si necesita un ajuste o una reconstrucción completa.

- ¿Qué es lo que requiere mayor atención?
- La respuesta del cliente a esta pregunta lo ayudará a comprender los elementos más importantes y que tendrán más importancia para ellos. Definir el propósito, comprender sus debilidades actuales y crear una lista detallada de características de la dirección hacia donde se apuntará, lo ayudará a construir una base sólida para un proyecto exitoso.
- ¿Por qué se requiere en su organización sistemas como el que se ofrecerá e integrará?

Al igual que la pregunta anterior en el cuestionario del cliente, esta pregunta lo ayuda a comprender las debilidades de sus ideas y a ver qué no funciona para el cliente. Le ayudará a comprender el propósito de los nuevos objetivos. Podría ser que solo necesite agregar una nueva característica diferente, o podría necesitar crearse en una nueva plataforma con características diferentes.

- ¿Qué características tendrán las ideas que se potenciarán en los usuarios?

Esta respuesta debe ser lo más detallada posible. Las características incluyen:

- Sentimientos de cercanía hacia nuestras ideas
- Influir en grupos sociales específicos
- Acumular datos y modelos sobre los tipos de usuarios a influir.

De nuevo, se anima al cliente o a la organización a considerar su público y los objetivos de sus ideas, al crear la lista de funciones necesarias hacia donde enfocarse.

¿Qué ideas similares te llaman la atención o cual es el grupo a combatir socialmente en Internet?
Esto mostrará las ideas que les gustan al cliente, y brindará ejemplos de características que podrían ser difíciles de describir. Puede ser especialmente útil para el cliente señalar las características de su 'competencia'.

¿Cuál es la fecha límite para comenzar a entregar datos estadísticos y modelos que signifiquen algo importante para que los Bots comiencen con su tarea de recolección y análisis de datos?
Se puede usar esta información en el cuestionario del cliente para determinar si las necesidades de este y en cómo se pueden cumplir antes de la fecha límite. Es posible que se deba proporcionar una línea de tiempo para mostrar qué se puede hacer antes de la fecha límite y qué se puede agregar más adelante y cuándo.
Este cuestionario se aplicará en reuniones con los jefes de la organización y departamentos involucrados.

3.1.5 Población Y La Muestra

Población se refiere al universo, conjunto o totalidad de elementos sobre los que se investiga o hacen estudios.
Muestra es una parte o subconjunto de elementos que se seleccionan previamente de una población para realizar un estudio.
Para nuestra investigación la población la componen los usuarios en Internet, involucradas en determinadas ideas que una organización o empresa represente, las cuales son:

- Usuarios que expresen ideas de apoyo a la organización o gobierno.
- Usuarios que expresen ideas de rechazo a la organización o gobierno.
- Usuarios potenciales para influir y atraer a la organización.

La muestra que utilizaremos serán los distintos tipos de usuarios en Internet.

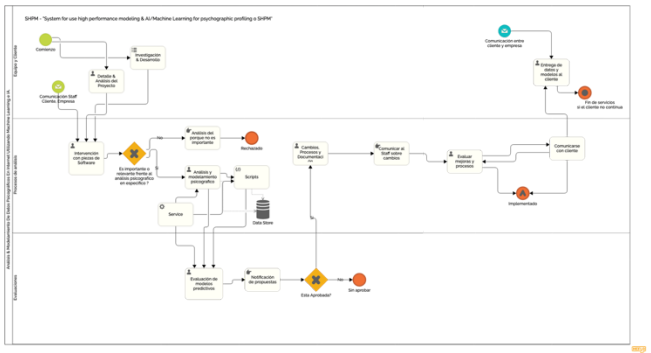
4. Análisis De Los Resultados

Primeramente, la técnica que se utilizará para la medición, es la observación es cuantitativa.
El objeto de estudio es: La cantidad de datos a analizar en una cantidad determinada de tiempo.

Para esto se tomará una encuesta hacia los trabajadores y se tomarán adicionalmente datos a jefatura encargada del cierre de estos las encuestas para recopilar datos ya sea antes o después de la implementación de los servicios respectivos para dar solución al problema planteado.
Todo lo anterior es para lograr obtener un completo análisis diario, mensual y anual. En caso de presentar algún inconveniente se podrá solucionar de forma inmediata ya que se contará con todos los sistemas de respaldos correspondientes, de tal forma que el registro sea eficaz y congruente.

4.1 Diagrama De Flujo De La Investigación Planteada

Figura 1
Diagrama de flujo, para el proyecto propuesto.



Nota. Es posible tener una visión completa desde el siguiente link: https://bit.ly/3EuAL3a

5. Explicación Sobre Diagrama De Flujo De La Investigación Planteada

La explicación al diagrama de flujo anterior es esencialmente dar la muestra del comienzo de los procesos de investigación o de proyectos y procesos, basándose inicialmente en reuniones con el cliente, acerca de detalles y análisis en sí del proyecto a futuro. Todo también en comunicación constante por email con el equipo y el cliente, que se tomará como recurso de vital importancia para los procesos de análisis y/o de evaluaciones, según corresponda. Es importante mencionar que, de manera constante, ya sea existan clientes o no, se puede desarrollar investigación y desarrollo en los temas planteados.

En la sección de procesos de análisis, estaremos ya frente a piezas de software que serán privativas, y funcionarán bajo terminales en un servidor donde existirá una base de datos que idealmente también almacenará información referente a datos estadísticos, de análisis, etc. Esto en sí es un conjunto de scripts que conformarán un servicio como conjunto, en específico.

Luego se evaluará lo analizado a menos que haya sido rechazado, realizado el respectivo análisis de porque se rechazó. Si no fue rechazado, continuamos con la evaluación de los modelos predictivos; se generan notificaciones, que de ser aprobadas pasan por cambios y mejoras en sus procesos y se realiza una respectiva documentación. Todo bajo total comunicación con el staff, respecto a las evaluaciones de mejoras en las versiones de los servicios. Ya implementado, con cambios y mejoras, todos los scripts están funcionando, y se comunicará al cliente sobre los análisis posteriores de estos. Si el cliente no desea continuar con los servicios, se determina un fin de estos. Pero se puede mantener procesos de investigación y desarrollo, bajo el mismo flujo.

5.1 Modelo, Funcionalidad Y Scripts

Este flujo lo que hará en su entorno, como piezas de software en distintos tipos, es analizar la segmentación psicográfica de usuarios en Internet, basándonos en conceptos que pueden estar siendo utilizados por ejemplo en la red social Twitter (En esta etapa planificamos que segmentos de usuarios estarán dentro de nuestro Target), en cuyo caso ya nos encontramos utilizado, una pieza de software que genera datos en torno a lo que puede ser el 'Sentiment Data' de usuarios (Aquí tal como lo indica, un ciclo de vida de un sistema de información, es importante el proceso de análisis).

Para analizar resultados nos basamos en un algoritmo que pueda determinar cuán positivo o negativos en una generalidad 'X' existen, frente a ciertos conceptos, ideas o personas, etc. Estos valores pueden anticipar cambios futuros, tal como los que se muestran en esta gráfica (Esto sirve para determinar en el ciclo de vida de un sistema de información, un correcto enfoque en la etapa de diseño y desarrollo):

Figura 2
Esquema sobre Segmentación y Factores Psicográficos.

Nota. Factores propuestos por Consunt. Adaptado de "Factores-Psicograficos.gif" [Imagen], por https://bit.ly/3X0yynk

Nos basamos también en la construcción de popularidad y subjetividad en ciertos mensajes, y a través de un algoritmo, y son clasificados para determinar dentro de una expresión si son o no - Bots. Eso permite tener un filtro mayor. Puede ser de suma importancia para nuestro cliente u organización, ya que de esa manera ellos pueden apoyarse en ciertos mensajes, para acercarse a una población 'A' o 'B' y determinar sus valores, así mismo la empresa o nuestro

cliente tomará mejores decisiones para acercarse a ciertos sentimientos masivos en Internet.

Como cualquier pieza de software está sujeta a errores, aun cuando ya utilizamos librerías Open Source en Python como TextBlob, entre otras para agilizar procesos (En esta etapa los scripts que como en este ejemplo determinan ciertos datos, pasa a una etapa de integración, QA o procesos de Testing).

A continuación, se despliegan algunas porciones del código de este proyecto, y sus resultados en una terminal, utilizando VSCode, en un ambiente Mac OSX & bajo Testing en un servidor Debian GNU/Linux. (Utilizaremos la variable 'Virgin Airlines' para este caso, como ejemplo).

En este caso nuestra pieza de software está en una etapa o fase de mantenimiento dentro del ciclo de vida de un sistema de información.

Figura 3

Llamado a librerías.

```
import tweepy
from textblob import TextBlob
```

Figura 4

Autorización aceptada por la API[5] de Twitter.

```
auth = tweepy.OAuthHandler(consumer_key, consumer_key_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
public_tweets = api.search('virgin airlines')

for tweet in public_tweets:
            print(tweet.text)
            analysis = TextBlob(tweet.text)
            print(analysis.sentiment)
            if analysis.sentiment[0]>0:
                        print ('Positivo')
            else:
                        print ('Negativo')
            print("")
```

Figura 5 - 6 & 7

Flujos de ciclos en programación en Python, para evaluaciones de "Sentiment Data" en Twitter.

```
    tweet_list.append(tweet.text)
    analysis = TextBlob(tweet.text)
    score = SentimentIntensityAnalyzer().polarity_scores(tweet.text)
    neg = score['neg']
    neu = score['neu']
    pos = score['pos']
    comp = score['compound']
    polarity += analysis.sentiment.polarity
```

Figura 6

```
    if neg > pos:
        negative_list.append(tweet.text)
        negative += 1

    elif pos > neg:
        positive_list.append(tweet.text)
        positive += 1

    elif pos == neg:
        neutral_list.append(tweet.text)
        neutral += 1
```

[5] Interfaz de programación de aplicaciones. Una API representa la capacidad de comunicación entre componentes de software. Referencia: https://bit.ly/3O3ZoqA

Figura 7

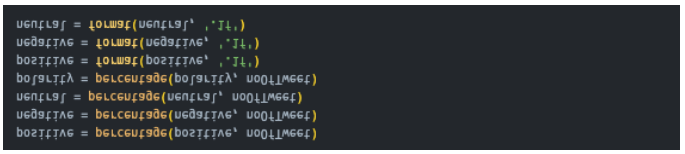


Figura 8 – 9 – 10 & 11

Ejemplos del tipo de resultados que obtenemos en una terminal.

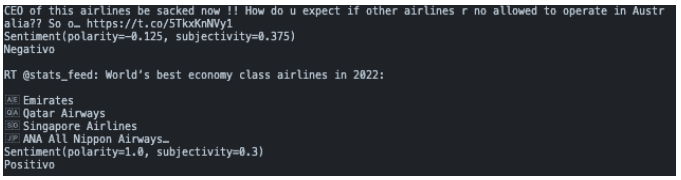


Figura 9

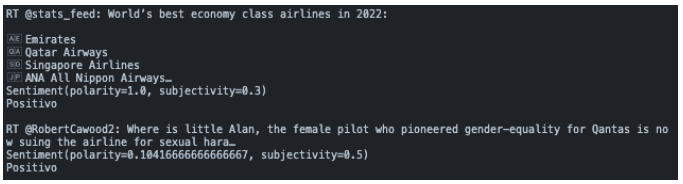


Figura 10

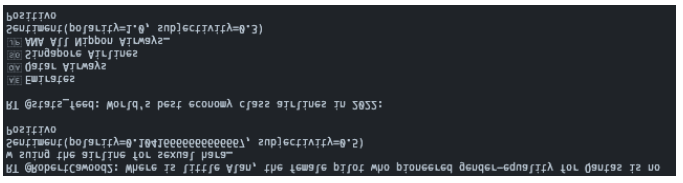
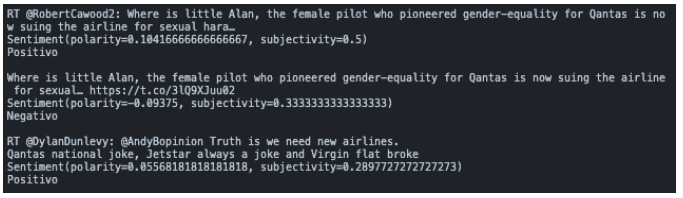


Figura 11



6. Proceso De Codificación, Requerimientos, Configuración & Tests[6]

Se utiliza para el desarrollo del proyecto, el lenguaje de programación Python, cuyas especificaciones en cualquier modo ya han sido detalladas anteriormente. Python es uno de los lenguajes más populares del mundo y con, de hecho, se creó el algoritmo de recomendación de Netflix y el que controla los autos sin conductor.

En términos de escalabilidad, Python tiene una ventaja sobre los lenguajes de programación como R en que ofrece un enfoque distinto para resolver diferentes problemas.
En términos de velocidad, Python también se destaca entre Matlab y Stata.

[6] Es importante señalar que, para estos casos, contamos con una cuenta autorizada por Twitter, para desarrollo.

Algunas de las características importantes de Python son:

- La sintaxis es bastante fácil de usar y, por lo tanto, quién puede aprender Python en menos tiempo.
- Python es, además, un lenguaje versátil y fácil de usar.
- En términos de escalabilidad, Python tiene una ventaja sobre los lenguajes de programación como R en el sentido de que ofrece un enfoque para resolver diferentes problemas.
- En términos de velocidad, Python también se destaca entre Matlab y Stata.
- Tiene una gran biblioteca.
- Una biblioteca o una biblioteca es un conjunto de las cuales están enlazadas entre sí.
- Puede usarse una y otra vez para programas.
- Tiene una comunidad muy sólida que ayuda a actualizar bibliotecas y marcos.
- Las bibliotecas y los marcos se pueden descargar y son gratuitos.

Python es un lenguaje de programación interpretado, es decir, primero se convierte en código de bytes que contiene instrucciones de bajo nivel y después lo ejecuta el intérprete de Python.
Es multiplataforma, lo que significa que una vez escrito en Python, puede ejecutarse en cualquier sistema operativo: Windows, Mac, Linux, etc.

6.1 Codificación Categórica

La codificación categórica es una técnica para codificar datos categóricos.
Es bueno tener en cuenta que los datos categóricos son conjuntos que contienen etiquetas de variables en lugar de valores. Muchos algoritmos de aprendizaje automático no pueden manejar variables categóricas. Por lo tanto, es importante codificar los datos adecuadamente para poder preprocesar estas variables.
Dado que necesita ajustar y evaluar, necesita codificar los datos categóricos y convertir cualquier variable de entrada y salida en valores numéricos.
De esta forma, el modelo podrá comprender y extraer información generando el resultado deseado.
Los datos categóricos varían en función del número de valores posibles.
La mayoría de las variables categóricas son nominales.
Estas variables ayudan a categorizar y etiquetar atributos.

6.1.1 Librería Más Utilizada

Otra necesidad de primera línea en análisis psicográfico es generar visualizaciones. En este sentido, es imposible evitar la presencia de Matplotlib y Seaborn. Ambas bibliotecas se usan mucho para la ciencia de datos, siendo Matplotlib la más antigua y la más popular, y Seaborn es un paquete floreciente que se basa precisamente en el código de Matplotlib.
Por tanto, el uso de ambas librerías es una sinergia relevante para la ciencia de datos.

6.1.2 Visual Studio Code Para Programar En Python

Visual Studio Code o más conocido como VSCode es un editor de código fuente de Microsoft que se puede usar en Windows, así como en macOS y Linux. Además, es un editor de código abierto que está disponible en GitHub.
Tiene características muy interesantes para el desarrollo de código como resaltado de sintaxis y autocompletado, integración con el sistema de control de versiones Git y depuración desde el propio editor. Al igual que con otros editores, como Atom o Sublime, también admite la capacidad de instalar compilaciones de terceros que agregan funcionalidad adicional. Instalar extensiones o add-ons en VSCode es tan simple como hacer clic en el botón correspondiente en el menú o acceder a ellas directamente con el atajo de teclado Ctrl+Shift+X.
Esto abre una nueva sección en el lado izquierdo del programa que contiene un motor de búsqueda, para que pueda buscar extensiones por nombre, y una lista agrupada en tres categorías: instaladas, populares y recomendadas.

6.1.3 Python Y Testing

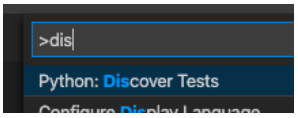
Esta extensión 'Python', desarrollada por Microsoft, agrega muchas funciones de Python a VSCode, como autocompletado y formateo de código, tal como se había mencionado, herramientas de depuración, nuestro propio código Python y gestión del entorno, entre otras. Otra consideración a tener en cuenta es activar el test que utilizaré, para elegir entre Unittest, Pytest o Nose.
Podemos hacer esto fácilmente accediendo a la configuración escribiendo Python Testing en la barra de búsqueda. En mi caso como utilizo Unittest marco la casilla correspondiente para activarlo o directamente en la configuración JSON, de VSCode. Para este caso se utilizará la prueba de una pieza de código relacionada con el login de usuarios a un micro-sistema en una Terminal.

6.1.4 Ejecutar Y Depurar (Debugging) Pruebas

Habilitar prueba unitaria en VScode.
Para habilitar Unittest en VSCode, ejecutamos el comando Discover test:

Figura 12

Búsqueda de Add-on en VSCode para encontrar un Testing Add-on.

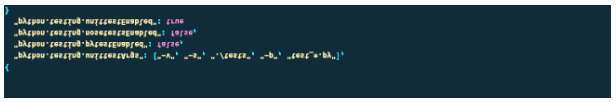


Esto nos pedirá que configuremos el marco de prueba utilizado, Pytest o nosetests, Unittest en este ejemplo, estoy usando Unittest.

Una vez configurado, el archivo de configuración .vscode/settings.json estará actualizado así:

Figura 13

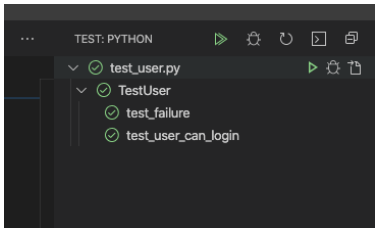
Edición de data JSON en la configuración de VSCode.



Esta configuración define los argumentos de Unittest utilizados y define qué marco de prueba de Python se utiliza. Las pruebas están ubicadas en la carpeta ‘./tests’ y la plantilla utilizada para los archivos de prueba es test_*. Una vez que hayamos Unittest y se encuentra configurado con VSCode, podemos ejecutar la prueba desde la pestaña Prueba o Testing.

Figura 14

Ejemplo en Testing, de VSCode, pasando las pruebas a modo de ejemplo.



O podemos hacer clic en la parte superior de cada prueba o donde ahora deberíamos encontrar la opción para realizar un, Run Test | Debug Test.

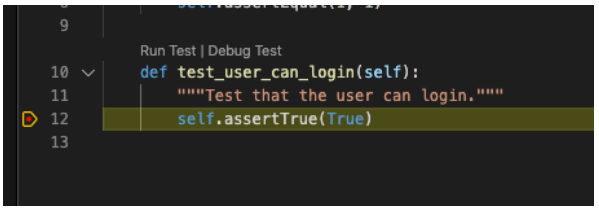
Figura 15

Funciones que servirán a modo de ejemplo fácil, para entender el funcionamiento de Unittest.

Usando ahora Debug Test, es posible fácilmente realizar un Testeo rápido y una depuración (Debugging) en el código, tal como mencione antes, aquí se realizó con el ejemplo del testeo en el código de un script para un micro-sistema de login-user en una Terminal.

Figura 16

En el siguiente código destacamos la línea donde se realizó - Debug Test.



7. Listado De Código, Secuencias De Scripts

A continuación, listare, las secuencias centrales de uno de los scripts o pieza de software, donde también se realiza un análisis psicográfico de sentimientos positivos, neutrales o negativos de una búsqueda y de una cantidad de datos a analizar según como se interactúe con la pieza de software que se ejecuta en una Terminal, y que genera dos gráficos directamente con Python, además de entregar tal como en el script que aquí se mencionó y se demostró en algún minuto, datos determinados y específicos. El gráfico a evaluar es de tipo circular, que determina y muestra datos en porcentajes. Y el siguiente es más bien de uso práctico para mostrar datos analizados en una forma de Nube. Estas operaciones se realizan con la librería de Pythom: Matplotlib, WordCloud, Nltk, entre otras.

En este caso se analizará el concepto de deportes extremos, dentro de Chile –‘Deportes extremos’.

Figura 17 & 18

Nuevamente, se realiza el ejercicio de utilizar las librerías necesarias para nuestro ejercicio.

Figura 18

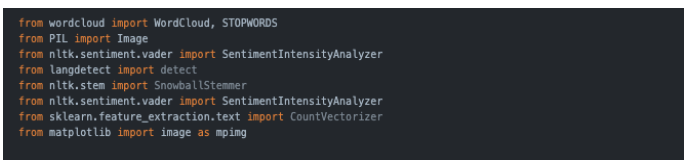
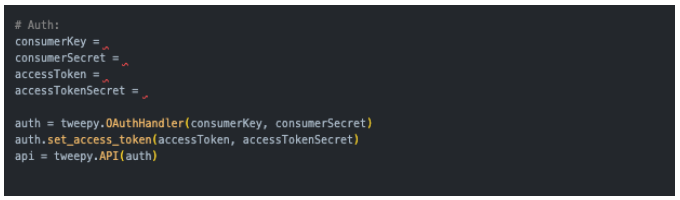


Figura 19

Integración a la API de Twitter.



```
# Auth:
consumerKey = 
consumerSecret = 
accessToken = 
accessTokenSecret = 

auth = tweepy.OAuthHandler(consumerKey, consumerSecret)
auth.set_access_token(accessToken, accessTokenSecret)
api = tweepy.API(auth)
```

Figura 20

Solicitamos las variables necesarias.

```
#Analysis

def percentage(part,whole):
    return 100 * float(part)/float(whole)

keyword = input("INGRESAR UNA KEYWORD OR HASHTAG PARA BUSCAR: ")
noOfTweet = int(input ("INGRESAR UNA CANTIDAD DE TWEETS PARA ANALIZAR: "))


tweets = tweepy.Cursor(api.search, q=keyword).items(noOfTweet)
positive  = 0
negative = 0
neutral = 0
polarity = 0
tweet_list = []
neutral_list = []
negative_list = []
positive_list = []
```

Figura 21 -22 & 23

Utilizando las funciones de librerías de análisis, realizamos comparaciones.

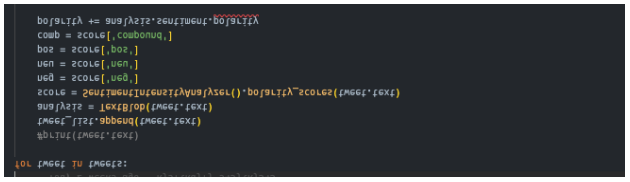
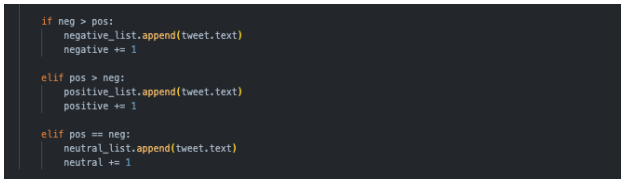


Figura 22



```
    if neg > pos:
        negative_list.append(tweet.text)
        negative += 1

    elif pos > neg:
        positive_list.append(tweet.text)
        positive += 1

    elif pos == neg:
        neutral_list.append(tweet.text)
        neutral += 1
```

Figura 23

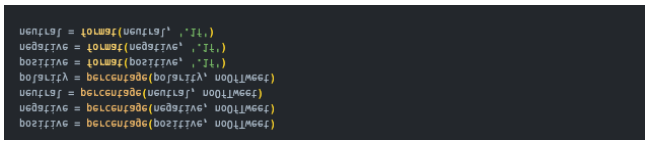


Figura 24

Enviamos a la Terminal, los datos analizados como resultado.

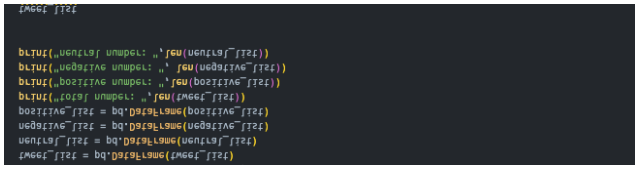
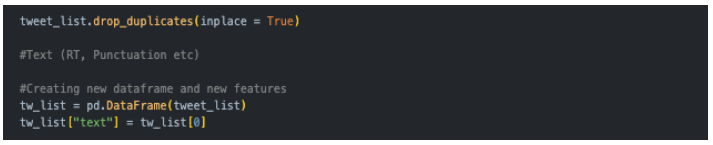


Figura 25

Comenzamos a generar el gráfico circular respectivo y contendrá el análisis en porcentajes.

Figura 26

Verificamos duplicados.



```
tweet_list.drop_duplicates(inplace = True)

#Text (RT, Punctuation etc)

#Creating new dataframe and new features
tw_list = pd.DataFrame(tweet_list)
tw_list["text"] = tw_list[0]
```

Figura 27

Eliminamos cierto tipo de caracteres que puedan ser incómodos frente al análisis.

Figura 28

Comenzamos a analizar las polaridades del análisis psicográfico.

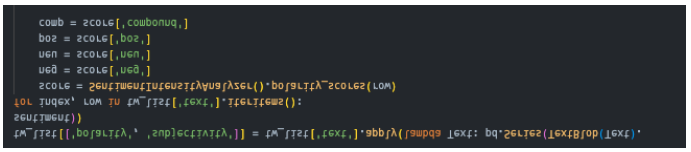


Figura 29

Se realizan comparaciones frente a si tal resultado es mayor o menor, como ejemplo.

```
    if neg > pos:
        tw_list.loc[index, 'sentiment'] = "negative"
    elif pos > neg:
        tw_list.loc[index, 'sentiment'] = "positive"
    else:
        tw_list.loc[index, 'sentiment'] = "neutral"
    tw_list.loc[index, 'neg'] = neg
    tw_list.loc[index, 'neu'] = neu
    tw_list.loc[index, 'pos'] = pos
    tw_list.loc[index, 'compound'] = comp

tw_list.head(10)
```

Figura 30

Contabilizamos, y generamos totales y porcentajes.

```
#Data frames (positivos, negativos and neutrales)

tw_list_negative = tw_list[tw_list["sentiment"]=="negative"]
tw_list_positive = tw_list[tw_list["sentiment"]=="positive"]
tw_list_neutral = tw_list[tw_list["sentiment"]=="neutral"]


#Función: count_values_in single columns

def count_values_in_column(data,feature):
    total=data.loc[:,feature].value_counts(dropna=False)
    percentage=round(data.loc[:,feature].value_counts(dropna=False,normalize=True)*100,2)
    return pd.concat([total,percentage],axis=1,keys=['Total','Percentage'])
```

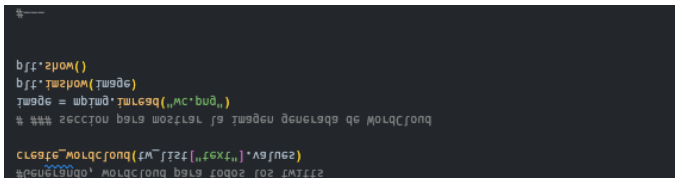
Figura 31

Utilizando el análisis anterior, creamos una 'Nube' gráfica, de las palabras más utilizadas.

```
#Count_values - para analysis -
count_values_in_column(tw_list,"sentiment")
# Wordcloud
def create_wordcloud(text):
    mask = np.array(Image.open("cloud.png"))
    stopwords = set(STOPWORDS)
    wc = WordCloud(background_color="white",
                   mask = mask,
                   max_words=3000,
                   stopwords=stopwords,
                   repeat=True)
    wc.generate(str(text))
    wc.to_file("wc.png")
    print("Word Cloud generada exitosamente!")
    path="wc.png"
```

Figura 32

Sobreescribimos la imagen sobre wc.png y podemos generar una nueva imagen y mostrarla.



7.1 Ejecución De Piezas De Software (Scripts)

La ejecución del script, es directamente usando VSCode.

Figura 33

Utilizando VSCode realizamos la ejecución del código, en la Terminal.

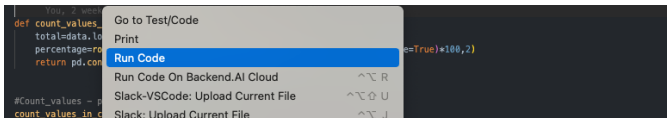


Figura 34

Obtenemos resultados según variables entregadas al microsistema, en la misma Terminal.

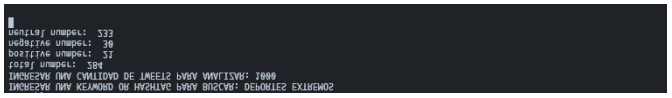


Figura 35

La siguiente figura muestra el tipo de gráfico circular que se puede generar, con el método anterior.

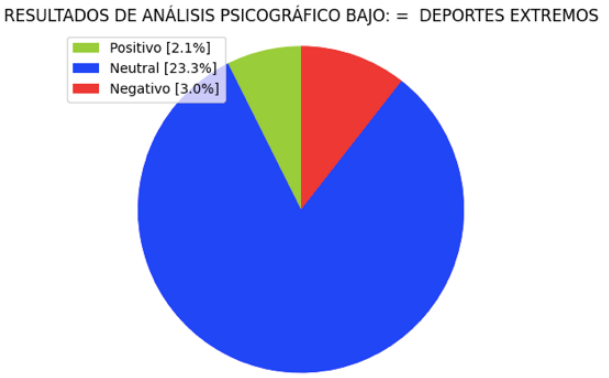


Figura 36

Nos entrega a través de la Terminal un mensaje de que se ha generado exitosamente una imagen de tipo 'Nube' con Python.

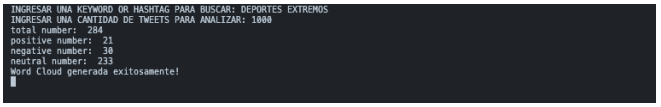
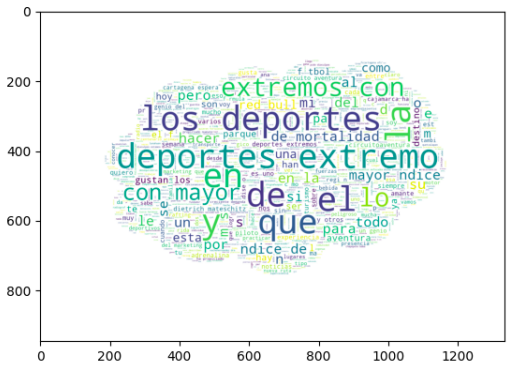


Figura 37

El output es el siguiente, como gráfico:



Nota. Es un gráfico generado de manera automática utilizando Python. Quisiera mencionar que con NodeJS[7], es posible realizar un sistema mixto, con Python y NodeJS para desplegar este tipo de contenidos, con una estructura en otro sistema, que mantenga estos datos y los despliegue en línea.

8. Pruebas De Sistema, Explicación, Criterios De Salida Y Resultados

[7] Node.js es un entorno en tiempo de ejecución multiplataforma, de código abierto, para la capa del servidor (pero no limitándose a ello) basado en el lenguaje de programación JavaScript, asíncrono, con E/S de datos en una arquitectura orientada a eventos y basado en el motor V8 de Google. Referencia: https://bit.ly/3g2Jpg8 (Wikipedia).

En un nivel alto, se necesitan pruebas de software para detectar errores en el software y para probar que el software cumple con los requisitos para obtener los resultados necesarios hacia el cliente.
Esto ayuda al equipo de desarrollo a corregir errores para ofrecer un producto de buena calidad.
Hay varios puntos en el proceso de desarrollo donde el error humano puede llevar a que el software no cumpla con los requisitos necesarios. Fundamentalmente se generarán pruebas de rendimiento, funciones, de estrés. Verificaremos que los datos que entreguen los gráficos, y outputs sean los correctos, en lo que se refiere al algoritmo utilizado.

Se realizarán pruebas dinámicas, debido a que las piezas y/o pieza de software que se utilizara con el objetivo de entregar análisis de datos, es un tipo de software donde se verificará el comportamiento de entrada y de salida son correctos, obteniendo los datos esperados. Las funciones de la o las piezas de software son operativas. Se realizarán pruebas de unidad para cada pieza o piezas de software que entregara datos específicos. Realizando luego pruebas de sistema, chequeando su carga, operatividad, seguridad, estabilidad, etc. Todo esto debido a que fundamentalmente son scripts, o piezas de software/código que se ejecutarán en una Terminal.

Para el proyecto, se utilizará como se mencionó, un tipo de prueba dinámica y en este caso, de cada una, ya que el objetivo principal de las piezas de software son los outputs o salidas de lo que se procesa (En este caso datos externos), cada pieza de software se consideraría como independiente.

Los resultados obtenidos, durante la prueba son precisos ya que son extraídos de medios específicos, para el análisis de estudio de tipo sintiente y/o psicográfico, por lo tanto, los resultados u outputs o salidas son al mismo tiempo precisos. Sin considerar además que, si existiera un error en el código fuente, no se lograría ejecutar en absoluto el script en la Terminal.

Implementar un servicio de prueba desde cero es una tarea compleja que requiere mucho tiempo.
En proyectos contrastantes, vemos que se han dado pasos pequeños, pero eficientes e incansables hacia el servicio de control de calidad en integración continua. Pasos como la contratación de personas especializadas en la materia, la implementación de herramientas como como Testlink, la gestión de pruebas, SonarQube para evaluar la calidad del código, Jenkins para la integración continúa o Selenium para las pruebas. En el futuro del Testing aparecen horizontes como el Big Data Testing, por lo que el futuro del Testing está garantizado.

8.1 Pruebas De Big Data

En las pruebas de Big Data, los ingenieros de control de calidad verifican el procesamiento exitoso de terabytes de datos utilizando uno y otros componentes de soporte. Esto requiere un alto nivel de habilidades de prueba, el procesamiento es muy rápido y puede ser de los siguientes tipos:

- Bach
- Tiempo real
- Interactivo

Las pruebas de Big Data se pueden dividir en tres.

Paso 1: Validación de la etapa de la etapa de prueba de Big Data, también conocida como etapa previa a Hadoop, implica la validación de los datos de varias fuentes, como fuentes de datos relacionales, blogs, redes sociales, etc., deben validarse para garantizar que los datos correctos están en el sistema. Los datos de origen se comparan con los datos ingresados en el sistema Hadoop para garantizar una coincidencia. Se verifica que los datos correctos estén extraídos y en la ubicación correcta.
Paso 2 - Validación de "MapReduce": en este paso, verifica la validación de la lógica comercial en cada uno y luego los valida después de que se ejecute en múltiples nodos, asegurando que:

- El proceso de MapReduce funciona bien.
- Las reglas de agregación o segregación de datos están en los datos.
- Se pueden generar, “Key Value pairs”.
- Los datos se validan después del proceso MapReduce.

Paso 3 - Fase de validación de los resultados: tercera y última etapa de las pruebas Big Data Testing, es el proceso de validación de los resultados.
Los archivos de datos de salida se generan y se mueven a un Data Warehouse empresarial o cualquier otro sistema basado en estos requisitos.
Las actividades en esta tercera etapa, incluye las siguientes:

- Se comprueba que las reglas de transformación se aplican correctamente.
- Se verifica la integridad de los datos y la descarga exitosa al sistema de destino.
- Se comprueba si hay daños en los datos comparando los datos de destino con los datos del sistema de archivos HDFS.

8.1.1 Hadoop

Hadoop maneja volúmenes de datos muy grandes y una gran cantidad de recursos. Por lo tanto, las pruebas de arquitectura son cruciales para garantizar el éxito de un proyecto como el que se presenta, de Big Data. Un sistema mal diseñado o inadecuado puede resultar en un rendimiento deficiente y el sistema puede no cumplir con los requisitos.

- Los servicios de prueba deben estar ejecutándose en un Hadoop.
- Las pruebas de rendimiento incluyen pruebas para la finalización del trabajo, el uso de la memoria, el rendimiento de datos y métricas del sistema similares.

9. Pruebas Y Análisis De Resultados Ii

Para entender lo que detallare a continuación, y que directamente se refiere a las pruebas y análisis en su parte dos de la explicación, es necesario que pueda dar a entender ciertos modelos de clasificación.

9.1 Clasificadores & Ejemplos De Modelos

Naive Bayes[8], es un modelo simple que considero importante mencionar, entre otros que existen y se pueden utilizar para los fines de análisis y clasificación de datos; este se puede utilizar para la clasificación de datos. En el modelo, la clase ĉ se asigna a un tweet t, tal como se ve en la siguiente fórmula:

Figura 38

Modelo de clasificación de texto, de Naive Bayes.

$$\hat{c} = \underset{c}{argmax}\ \mathrm{P}(\mathrm{c}|\mathrm{t})$$

$$\mathrm{P}(\mathrm{c}|\mathrm{t}) \propto \mathrm{P}(\mathrm{c}) \prod_{i=1}^{n} \mathrm{P}(f_i|\mathrm{c})$$

Nota. Adaptado de "twitter_sentimental_analysis_5.jpg" [Imagen], por pantechsolutions.net - Twitter Sentiment Analysis using Machine Learning on Python. https://bit.ly/3hDjMmy

En la fórmula anterior, fi representa la i-ésima característica del total de n características. P(c) y P(f i |c) pueden obtenerse mediante estimaciones de máxima verosimilitud.

9.1.1 Otros Modelos

Entropía Máxima
El modelo Clasificador de Máxima Entropía se basa en el Principio de Máxima Entropía. La idea principal detrás de esto es elegir el modelo probabilístico más uniforme que maximice la entropía, con restricciones dadas. A diferencia de Naive Bayes, no asume que las características son condicionalmente independientes entre sí. Por lo tanto, podemos agregar funciones como bigramas sin preocuparnos por la superposición de funciones.

Figura 39

Modelo / Ecuación de Máxima entropía. En un problema de clasificación binaria como el que estamos abordando, es lo mismo que usar Regresión Logística para encontrar una distribución sobre las clases. El modelo está representado por la siguiente fórmula:

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

Nota. Adaptado de "Twitter_Sentimental_Analysis_6.jpg" [Imagen], por pantechsolutions.net - Twitter Sentiment Analysis using Machine Learning on Python. https://bit.ly/3O3Asj1

Random Forest. (Utilizado en lo que se presenta)
Random Forest es un algoritmo de aprendizaje conjunto para clasificación y regresión de datos. Random Forest genera una multitud de clasificaciones de árboles de decisión en función de la decisión agregada de esos árboles. Para un conjunto de tweets x 1 , x 2 , . . . x n y sus respectivas etiquetas de opinión y 1 , y 2 , . . . n 'bagging', selecciona repetidamente una muestra aleatoria (X b , Y b) con reemplazo. Cada árbol de clasificación f b se entrena usando una muestra aleatoria diferente (X b , Y b) donde b varía de 1 . . . Finalmente, se toma un voto mayoritario de las predicciones de estos árboles B.

9.1.2 Ecuación Predictiva

Figura 40

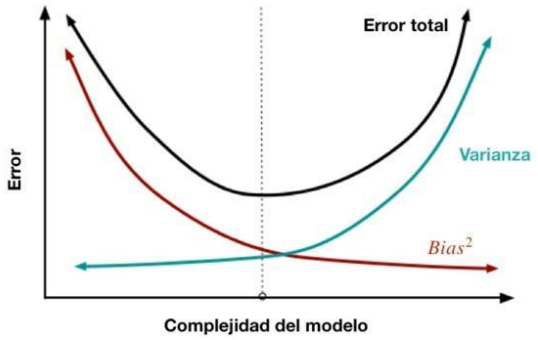
Ecuación del modelo predictivo. Es el algoritmo de entrenamiento para 'bosques o arboles' aleatorios que aplica la técnica general de agregación de arranque o bootstrap aggregating, or bagging, para aprendizaje automático, o machine learning como de tipo árbol.

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

Nota. Adaptado de imagen de Wikipedia, que no es posible de obtener como -nombre-, [Imagen]. Desde https://bit.ly/3tpg6qX & https://bit.ly/3hGNasd

Figura 41

Complejidad del modelo predictivo. Un equilibrio óptimo de bias y varianza nunca sobreequiparía o no sería adecuado para el modelo. Por lo tanto, comprender el bias y la varianza es fundamental para comprender el comportamiento de los modelos de predicción.



Nota. Adaptado de *"46138669365_b98531b89d_b.jpg" [Imagen] por https://flickr.com en https://bit.ly/3O0Pm9Q. Sesgo y Varianza en Machine Learning. https://bit.ly/3UVfRzY*

Además, se puede hacer una estimación de la incertidumbre de la predicción como la desviación estándar de las predicciones de todos los árboles de regresión individuales en x':

[8] Explicación sobre el modelo de Naive Bayes aquí: https://bit.ly/3UU6eRO

Figura 42

Modelo Predictivo, técnica general de agregación de arranque o bootstrap aggregating, or bagging, para aprendizaje automático, o machine learning como de tipo árbol.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

Nota. Adaptado de Adaptado de imagen de Wikipedia, que no es posible de obtener como -nombre-, [Imagen]. Desde Wikipedia en https://bit.ly/3g26QGk.

Es bueno mencionar que existen otros modelos también, como el SVM, o super vector machines. El XGBoost, Xgboost es una forma de algoritmo de gradiente incremental, que produce un modelo de predicción que es un conjunto de árboles de decisión de predicción más débil.

9.1.3 Redes Neuronales

MLP o perceptrón multicapa es una clase de redes neuronales avanzada, que tiene al menos tres capas de neuronas. Cada neurona usa una función de activación no lineal y aprende con supervisión usando un algoritmo de retropropagación. Funciona bien en problemas de clasificación complejos, como el análisis de sentimientos mediante el aprendizaje de modelos no lineales.

9.1.4 Análisis

En este caso, a través de la siguiente pieza de software, y a modo de ejemplo, se utilizará como concepto para analizar, los tweets de 'sentimiento público' relativos a '6 aerolíneas estadounidenses'.
Clasifiqué los tweets en sus categorías, es decir, positivo, neutral y negativo utilizando técnicas de aprendizaje automático en Python. Tal como en un ejemplo anterior, pero bajo otro concepto, y en esta ocasión con la intención de análisis, utilizando Python como herramienta directa para crear aprendizaje de máquina, o machine learning usando diferentes librerías disponibles en el entorno.

Importación de bibliotecas
Para ejecutar los scripts de Python, son necesarias algunas bibliotecas. Tal como se nota a continuación.

Figura 43

Importación de librerías.

```
import numpy as np
import pandas as pd
import nltk
import re

import matplotlib.pyplot as plt
import seaborn as sns
```

9.1.5 Importación Del Conjunto De Datos

El conjunto de datos que se utilizará para entrenar el algoritmo de machine learning, se utilizará un archivo disponible en formato *.CVS. Este archivo contiene un conjunto de datos, como el tweet del usuario, la identificación del tweet, el nombre de la aerolínea con la que se relaciona el texto del tweet, el número de conteo, etc. Se puede usar el método *read_csv()* de la biblioteca Pandas para importar el conjunto de datos a la pieza de software, que realizará el análisis, tal como se muestra en el siguiente script:

Figura 44

Extracción de datos desde archivo *.CVS.

```
df = pd.read_csv("Tweets.csv", encoding="utf-8")
df.head()
```

Figura 45

Producción**.** La siguiente imagen muestra las primeras cinco filas del conjunto de datos.

airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	name	retweet_count	text	tweet_created
neutral	1.0000	NaN	NaN	Virgin America	cairdin	0	@VirginAmerica What @dhepburn said.	2015-02-24 11:35:52 -0800
positive	0.3486	NaN	0.0000	Virgin America	jnardino	0	@VirginAmerica plus you've added commercials t...	2015-02-24 11:15:59 -0800
neutral	0.6837	NaN	NaN	Virgin America	yvonnalynn	0	@VirginAmerica I didn't today... Must mean I n...	2015-02-24 11:15:48 -0800
negative	1.0000	Bad Flight	0.7033	Virgin America	jnardino	0	@VirginAmerica it's really aggressive to blast...	2015-02-24 11:15:36 -0800
negative	1.0000	Can't Tell	1.0000	Virgin America	jnardino	0	@VirginAmerica and it's a really big bad thing...	2015-02-24 11:14:45 -0800

Visualización de datos
A continuación, realizaremos la visualización de datos. Primero, trazamos la distribución de tweets positivos, negativos y neutrales en nuestro conjunto de datos usando un gráfico circular.

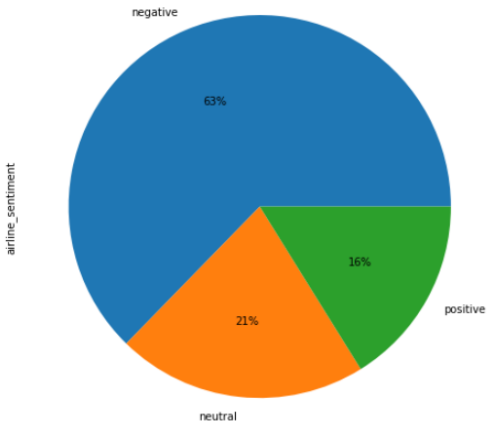
Figura 46

Género en Python una figura circular como gráfico.

```
df.airline_sentiment.value_counts().plot(kind='pie', autopct='%1.0f%%')
fig = plt.gcf()
fig.set_size_inches(8,8)
```

Figura 47

Producción. Output de los resultados a modo de gráfico circular.



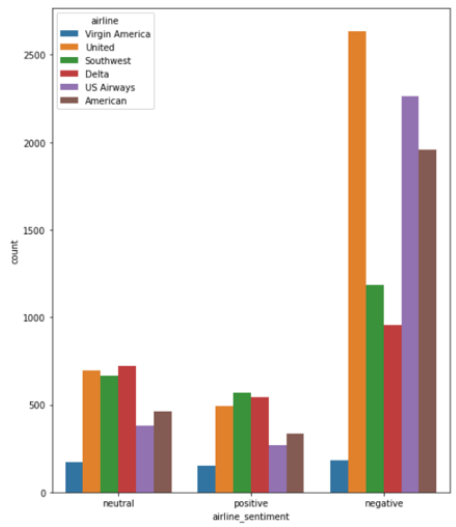
El resultado muestra que el 63 % de los tuits generales son negativos, mientras que el 21 % y el 16 % son respectivamente neutrales y positivos.

Figura 48

Se traza un diagrama de barras que muestre el recuento de tweets negativos, positivos y neutrales para las 6 aerolíneas como se muestra a continuación:

Figura 49

Producción. Genero el gráfico de columnas, en Python, para segmentar el sentiment análisis.



El gráfico anterior muestra que United Airlines tiene los tweets más negativos y neutrales, mientras que Airlines tiene los tweets más positivos. Virgin America tiene la cantidad más pequeña de tweets positivos y neutrales. Sin embargo, la razón podría ser que la participación general de tweets de Virgin America es más baja que la de las aerolíneas.

9.1.6 Preprocesamiento De Datos

Ahora nosotros necesitamos eliminar los números y ciertos caracteres de los tweets. Definiremos una función llamada *text_preprocess()* que acepta cadenas de texto y elimina todo el texto excepto los alfabetos. Los espacios simples y dobles creados como resultado de la eliminación de dígitos y los caracteres especiales se eliminan posteriormente.
Se ejecuta el siguiente script para definir la función *text_preproces()*. La primera línea de la función elimina números y caracteres especiales. La segunda línea de la función elimina todos los únicos generados y este representa también un resultado de la eliminación de caracteres especiales. Finalmente, la tercera línea de la función *text_preprocess()* elimina los espacios vacíos dobles y los reemplaza con un solo espacio.

Figura 50

Genero una función para procesar los caracteres especiales y números.

```
def text_preprocess(sen):

    sen = re.sub('[^a-zA-Z]', ' ', sen)

    sen = re.sub(r"\s+[a-zA-Z]\s+", ' ', sen)

    sen = re.sub(r'\s+', ' ', sen)

    return sen
```

Figura 51

Antes de que podamos de esta forma, 'limpiar' los tweets, se debe dividir los datos en características y etiquetas:

Figura 52

A continuación, ejecutamos un bucle *foreach()* que pasa iterativamente tweets de la lista de tweets X al método *text_preprocess()* que limpia el texto del tweet. El siguiente script es el que realiza la operación:

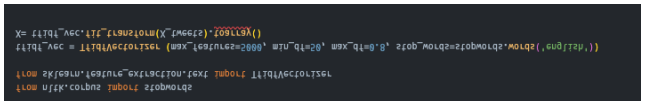
```
X_tweets = []
messages = list(X)
for mes in messages:
    X_tweets.append(text_preprocess(mes))
```

9.1.7 Conversión De Texto A Números

Dado que los algoritmos de aprendizaje automático se basan en matemáticas y las matemáticas funcionan con números, es necesario convertir los tweets de texto en forma numérica.

Figura 53

Aunque hay varias formas de hacerlo, utilizaré en este caso la clase *TfidfVectorizer del módulo sklearn.feature_extraction.text.* Para hacerlo, se puede usar el método *fit_transform()* de la clase *TfidifVectorizer* que se muestra en el siguiente script:



El atributo *max_features* se utiliza para especificar el número de palabras que aparecen con más frecuencia para convertir, que es 5000 en este caso. El atributo *min_df* especifica el número mínimo de documentos en los que debe aparecer una palabra (50). Finalmente, *max_df* especifica la proporción máxima de documentos en los que debe aparecer una palabra, que es del 80% en el script anterior. También eliminamos palabras vacías como an, is, are, we, at, ya que no brindan mucha información para la clasificación.

9.1.8 División De Datos En Conjuntos De Entrenamiento Y Prueba

Los algoritmos de aprendizaje automático o machine learning, se entrenan en conjuntos de entrenamiento y se evalúan en conjuntos de prueba.

Figura 54

Para dividir los datos en conjuntos de entrenamiento y prueba, se puede usar el método *train_test_split()* del módulo *sklearn.model_selection* tal como se muestra a continuación en este script:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
```

9.1.9 Entrenamiento De Algoritmos De Aprendizaje Automático

Aunque se puede usar cualquier algoritmo de clasificación de la lista de clasificadores aquí. Se utilizará el clasificador Random Forest, ya que es el más robusto. Para usar el clasificador Random Forest en este caso, se puede usar la clase *RandomForestClassifier de sklearn.ensemble* los scripts entregados como ejemplo. Para entrenar la clase RandomForestClassifier en el conjunto de entrenamiento, se debe pasar las funciones de entrenamiento (X_train) y las etiquetas de entrenamiento (y_train) al método *fit() de la clase RandomForestClassifier* .

Figura 55

Una vez que se entrena el modelo, se puede hacer predicciones pasando las características de prueba *(X_test)* al método *predict() de RandomForestClassifier.*clase. Se debe ejecutar el siguiente script para entrenar el clasificador *Random Forest* y hacer predicciones.

```
from sklearn.ensemble import RandomForestClassifier

rf_clf = RandomForestClassifier(n_estimators=250, random_state=0)
rf_clf.fit(X_train, y_train)
y_pred = rf_clf.predict(X_test)
```

Evaluación De Los Algoritmos

Se puede utilizar Accuracy, F1, Recall y Confusion Matrix como métricas para evaluar el rendimiento de un algoritmo de clasificación.

Figura 56

Para hacerlo en Python, se puede usar el módulo *sklear.metrics* para encontrar los valores de estas métricas como se muestra en el siguiente script:

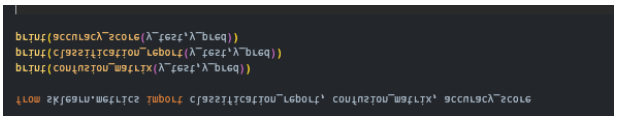
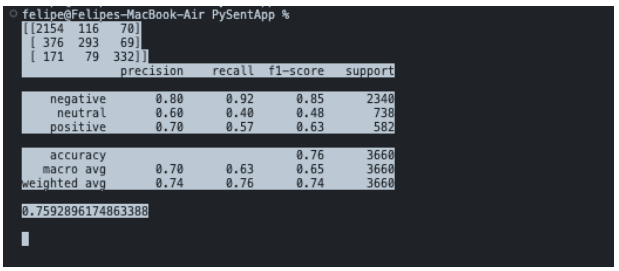


Figura 57

Este es el output en la terminal de los resultados a nivel general, basándonos en el algoritmo propuesto:



El resultado muestra que el algoritmo y esta pieza de software, puede clasificar de manera eficiente y precisa, un tweet como 'positivo, negativo o neutral' con una total precisión del 75,92 %.

10. Conclusiones Y Recomendaciones

10.1 Tema De Estudio, Importancia

Las nuevas tendencias de marketing para empresas, marcan el uso de las redes sociales como medio de comunicación de acciones principalmente por su gran potencial para establecer relaciones con los clientes a través de ellas.
Las redes sociales también facilitan los esfuerzos de comunicación corporativa para transmitir los valores de marca, atraer nuevos consumidores, difundir información de marketing u obtener resultados medibles de clientes insatisfechos.
Según Pak y Paroubek (2010), las redes sociales han sido

una herramienta útil para simplificar la comunicación con los clientes.

Es importante conocer el propósito de cada conjunto social, el formato y el público objetivo y estar en las redes sociales donde el público y el target de estos esperan encontrar la marca.

10.1.1 Objetivo General De La Investigación

Esta investigación tiene como objetivo medir el sentimiento de usuarios en la red social Twitter en relación a cómo en base a ciertos conceptos, se pueden entender un modelo psicográfico y sus correspondientes modelos predictivos.

Para hacer esto, la investigación se basa en datos de los perfiles de Twitter de usuarios, organizaciones, empresas etc., las cuales fueron obtenidas utilizando la API de Twitter.
Una vez que se descargan los tweets, se aplica un algoritmo desarrollado que funciona con machine learning, utilizando Python, para dividir la muestra (n=X tweets) en sentimientos negativos, neutrales y positivos.

10.1.2 Puntos Principales De La Investigación

Análisis de Sentimiento con Machine Learning
El análisis de sentimientos se define como el proceso de formación de opiniones en términos de calificaciones, actitudes y emociones sobre un tema en particular. (Fiorini and Lipsky, 2012).

El análisis de sentimientos generalmente sirve para dos propósitos, primero, expresiones de sentimiento y definición de la orientación del sentimiento por parte de los individuos. (Honeycutt and Herring ,2009; Saura, 2018). El análisis de sentimiento permite detectar la expresión positiva, negativa y neutra sobre un tema específico, un producto o servicio, entidad, persona física, etc., de un elemento textual. (Boyd, 2017; Chunga, 2017).

Debes (2017), indica que el análisis de sentimiento puede referirse a enfoques y estar basado en características y etiquetas automáticas, conversaciones, o puede ser el caso de etiquetas comunes en una temática o eventos concretos de uso, en emoticon o en recursos como léxicos de sentimiento que se con tweets positivos, neutrales o negativos. Los léxicos fundamentales etiquetan palabras recogidas en una dimensión necesaria de manera semántica, llamada "sentimiento", "valencia" u "orientación semántica" (Saura, Palos-Sanchez / Cerdá, 2017).

Los algoritmos desarrollados en Python para realizar análisis de sentimiento tienen poder predictivo.
La predicción viene determinada por machine learning. El aprendizaje automático es una forma de inteligencia artificial que entrena a la máquina virtual a través de la exploración de datos para automatizar el proceso de análisis de datos, entre otras características.

10.1.3 Objetivos, Logros Y Formatos

Luego del desarrollo del proceso metodológico que incluye el análisis y extracción de datos, se procedió al análisis de sentimiento psicográfico en esta investigación, como resultado del análisis, se logró obtener el promedio, mediante la aplicación del algoritmo con aprendizaje automático.

En la tabla u *output en la terminal, de las pruebas y resultados (Parte II),* entregaron los resultados del proceso de análisis de tweets procesados sobre el sentimiento de una cantidad determinada de usuarios respecto a 6 aerolíneas de Estados Unidos, además del uso de las interacciones por parte de las empresas y las interacciones y comentarios realizados por los usuarios sobre las mismas, la categorización realizada según sentimiento y la veracidad promedio obtenida como resultado del algoritmo machine learning. Tal como se especificó anteriormente, se logró clasificar de manera eficiente y precisa, un tweet como 'positivo, negativo o neutral' con una total precisión del 75,92 %.

10.1.4 Viabilidad Y Potencial De La Investigación Y El Proyecto

El motor principal de la aplicación utilizado es el aprendizaje automático, con uso repetido y entrenamiento del promedio de los aumentos de resultados promedio.
Como se ha demostrado, Twitter se configura como la red social óptima para los usuarios que pueden expresar sus sentimientos, opiniones y comentarios de forma específica, en tiempo real y en todo el mundo.
Twitter se ha utilizado como objeto de estudio en múltiples periodos durante la última década.
En esta investigación se utilizó Twitter para jerarquizar el sentimiento de las ofertas, y por tanto la calidad, de las empresas que componen la muestra cuando publican en dicha red social.
Los resultados de la investigación pueden ser utilizados por las empresas para mejorar el desarrollo de sus estrategias en las redes sociales y, más concretamente, en torno al Twitter social.

Además, los resultados de búsqueda identifican comunicaciones y ofertas a través de Twitter para otros relacionados con los datos analizados, en términos de tono, y qué pueden aprovechar las empresas en las redes sociales.
Esta investigación proporciona datos verificados de acciones en Twitter que se pueden usar para futuras estrategias de marketing y servirán como una fuente de investigación sobre el tema especificado.

10.1.5 Recomendaciones Para Continuar La Investigación

- ¿Por qué debe hacerse?

Sin menospreciar el valor de los KPI cuantitativos, el sentimiento e interés por establecer con mayor precisión la audiencia a la que se dirige una marca, es importante, ya que permite analizar la respuesta emocional de los usuarios que han interactuado con él. En el entorno digital actual, las personas tienen más herramientas para expresar su opinión, compartir sus dudas y expresar sus inquietudes, ya sean positivas o negativas.
Es por ello que el análisis de sentimiento se postula como un activo que permite obtener información sobre el tono en el que hablan los usuarios en redes sociales, o foros de una determinada marca, algo que permite a la gente saber lo que le gusta y lo que no.

- ¿Qué cosa se debe hacer?

El análisis de sentimiento es un proceso en el que se utiliza

aprendizaje automático o machine learning para encontrar un punto de una palabra clave o sentimiento, y colocar la información de interés en el proceso. El sentimiento puede definirse como el resultado de "un juicio o un juicio formado sobre alguna cosa, no necesariamente hechos o conocimientos". Pero con el uso del análisis de sentimientos y la ciencia de datos, se entiende la opinión o el juicio.
Es una evaluación subjetiva de algo basada en la experiencia empírica personal.
Se compone de hechos objetivos y en parte por emociones.
Una opinión se puede interpretar como una dimensión en los datos sobre un tema en particular.
Es un conjunto de significantes que, combinados, presentan una visión, es decir, un sentimiento sobre un tema en particular.

Algoritmos de análisis de sentimiento
Hay dos métodos principales de análisis de sentimiento.

- *Enfoque basado en reglas*

El análisis de opiniones basado en reglas se basa en un algoritmo con una descripción claramente definida de una opinión identificada.

- *Incluye identificación de subjetividad, polaridad o tema de opinión.*

El enfoque basado en reglas implica una rutina básica de procesamiento del lenguaje natural.

Así es como funciona:
Hay dos listas de palabras.
Uno de ellas solo incluye puntos positivos, neutros y los otros puntos negativos.
El algoritmo escanea el contenido, encuentra palabras que coinciden con los criterios.
Después de eso, el algoritmo calcula qué tipo de palabras son más frecuentes en el contenido.
Si hay más palabras positivas, se considera que el texto tiene una polaridad positiva.

- ¿A quién beneficiará?

El análisis de sentimiento o análisis psicográfico, se ocupa de la percepción de los usuarios en torno a ideas o productos, a partir de la comprensión del mercado, a través del prisma de los datos de sentimiento que se pueden encontrar ya registrados por marcas y empresas en redes sociales o en su general en Internet.
Hay muchas fuentes de información públicas y privadas de las que puede obtener información sobre la calidad del producto por parte del cliente o la calidad de la relación del usuario o cliente con el producto.
Para nombrar unos pocos:

- Llamadas de atención al cliente y correos electrónicos.
- Reseñas de productos generadas por los usuarios.
- Publicaciones, respuestas o comentarios en redes sociales.
- Foros generales y especiales.
- Registro de las interacciones con los clientes.

El análisis de sentimientos puede ayudar a las empresas a dar sentido y agregar valor a la acumulación de datos no estructurados y transformarlos.
Una visión claramente definida de lo que ciertos segmentos de clientes piensan del producto o de la empresa en general.
Una inmersión profunda en el estado del mercado desde la perspectiva del consumidor.
En cualquier caso, es un factor influyente en la formulación y elaboración de la propuesta de valor para un segmento de audiencia específico. Si bien al principio estas actividades son relativamente fáciles de realizar con soluciones básicas, en algún momento se vuelve lógico usar herramientas más elaboradas y extraer ideas más sofisticadas.

- ¿Quién lo hará?

Son ingenieros, y fundamentalmente, Data Scientists, o empresas/emprendimientos dedicados a estos temas en específico, donde amplían mucho más la mirada a distintas disciplinas.
Un Data Scientist o científico de datos es el profesional dedicado a la recopilación, análisis y a la comprensión de grandes volúmenes de datos y su respectiva extracción.
Son personas que aplican sus conocimientos de estadística y programación para analizar e interpretar los datos a disposición de las empresas y extraer información valiosa.
Las organizaciones tienen una gran cantidad de información que, si se usa correctamente, puede traducirse en beneficios comerciales. En un entorno cada vez más digital, aprovechar la información que las empresas obtienen con su entorno es casi imprescindible. De ahí la creciente necesidad de profesionales que puedan analizar y dar sentido a todos estos datos, para que tengan verdadero valor.

¿Qué hace un Data Scientist en una empresa?
Las funciones de un científico de datos pueden diferir de una organización a otra, pero a grandes rasgos comprenden las siguientes:

- Minería de datos.
- Obtener toda la información que considere útil de varias fuentes.
- El volumen de datos puede diferir.
- Limpieza de datos.
- Eliminar toda la información que no sea es irrelevante preparar los datos para su procesamiento.
- Procesamiento de datos.
- Procesar datos aplicando enfoques estadísticos, análisis, aprendizaje automático, modelos predictivos, etc.
- Visualización de datos.
- Representar los datos de diferentes maneras para hacerlos comprensibles de la manera más precisa.

- ¿Dónde se hará?

Estas implementaciones de tecnología se realizan en sistemas controlados y manejados idealmente por ingenieros en empresas de tecnología, y de análisis de datos. Esta rama, es de constante investigación, por lo que es común y muy natural que estas empresas realicen mucha investigación de manera paralela a la entrega de servicios, desarrollo de productos tecnológicos para sí misma y/o de asesoría.

11. Conclusiones Generales[9]

Sin duda, el planteamiento del problema de una investigación técnica supone un paso importante para su desarrollo, permitiendo establecer quiénes serán estudiados y, por en consecuencia, cuáles serán resueltas.
El planteamiento del problema, así como los objetivos de la investigación, el punto de partida de este tema, actualizando lo realizado en el tema anterior, estableciendo de forma clara y concisa la dirección inicial que se entregó a la investigación.
Esto ocurre ya que la parte de la investigación constituye previamente una reflexión ordenada y coherente que destaca una transición lógica de acciones y objetivos encaminados al problema detectado.
Un producto a nivel de software, hecho a la medida, tiene un impacto positivo y deseable en la organización donde se implementa.
Esto sucede porque fue diseñado y basado en las necesidades particulares de quienes lo necesitan.
Para que este impacto sea verdaderamente positivo y genere beneficios, su desarrollo debe realizarse de manera ordenada, con estándares adecuados y buenas prácticas que garanticen el cumplimiento de los plazos comprometidos, los costos asociados y que funcione para lo que fue creado, cumpliendo con los requisitos del cliente y del proceso.

El ciclo de vida de desarrollo de una aplicación o producto de tecnología de la información no es más que una secuencia estructurada y claramente definida de pasos a considerar al desarrollar un producto tipo software o que es tecnología de la información.

Sin ninguna duda, el desarrollo de un prototipo, capaz del hardware o software de un sistema, es aquel en el que se debe tener en cuenta una gran cantidad de variables. Además, incluye una primera etapa de diseño, una etapa intermedia que incluye el desarrollo y una final donde se evalúan los resultados, esta última etapa en la que se prueba la interrelación de las partes. Por su parte, los diagramas de flujo ofrecen una forma única de visualizar y organizar procesos complejos de una manera fácil, lo que los convierte en una herramienta excelente para mejorar la resolución de problemas, así como también, una forma eficaz de compartir información.
Tan importante como verificar que un usuario puede usar la aplicación, es igualmente importante verificar que el sistema continúa funcionando correctamente cuando se toman acciones inesperadas o se ingresan datos incorrectos.
Por lo tanto, un buen conjunto de pruebas debería llevar la aplicación o el software al límite, sin mencionar que, en el caso de las pruebas automatizadas, estas también son código, por lo que también merecen una consideración detallada.
La investigación es la sistematización lógica de información utilizada para obtener nuevos conocimientos, para descubrir datos relevantes o verdades relacionadas con los hechos que se analizan.
Por otro lado, las normas APA contienen lineamientos que han sido universalmente aceptados a la hora de documentar los distintos pasos realizados durante una investigación.

---

[9] Resumen a modo de comentario general, de todo el material estudiado durante el ramo de Proyecto de Título. IACC (2022).

Referencias.

Balestrini, M. (2006). Cómo se elabora el proyecto de investigación. 7ma edición. Caracas, Venezuela: Consultores Asociados.

Restrepo, M. (2008). Producción de textos educativos. Bogotá, Colombia: Editorial Magisterio.

Fernández, V. (2006). Desarrollo de Sistemas de Información una Metodología Basada en el Modelado. Barcelona, España: Edicions UPC.

Granados, R. (2014). Desarrollo de aplicaciones web en el entorno servidor. IFCD0210. Málaga, España: IC Editorial.

Sommerville, I. (2005). Ingeniería de software. Madrid, España: Pearson Education.

Barranco, J. (2001). Metodología del análisis estructurado de sistemas. 2da edición. Madrid, España: Universidad Pontificia Comillas.

Granollers, T.; Vidal, J. y Cañas, J. (2005). Diseño de sistemas interactivos centrados en el usuario. Barcelona, España: Editorial UOC.

López, A. (2007). Introducción al desarrollo de programas con Java. México D.F., México: Universidad Nacional Autónoma de México.

Mompín, J. (1988). Introducción a la bioingeniería. Barcelona, España: Marcombo.

Álvarez, L. (2009). La materialización de ideas. Realidades, necesidades, oportunidades, encuentros y desencuentros. Tesis de postgrado. Barcelona, España: Universitad Autónoma de Barcelona.

Gómez, S. y Moraleda, E. (2020). Aproximación a la ingeniería del software. Editorial Universitaria Ramón Areces.

Granados, R. (2015). Despliegue y puesta en funcionamiento de componentes software. IFCT0609. IC Editorial.

Serna, E. (2013). Prueba funcional del software: un proceso de verificación constante. Medellin, Colombia: Instituto Tecnológico Metropolitano.

Barranco, J. (2001). Metodología del análisis estructurado de sistemas. Madrid, España: Universidad Pontificia Comillas de Madrid.

Garriga et al. (2010). Introducción al análisis de datos. Madrid, España: Editorial UNED.
PopulationPyramid.ne (2019) Densidad de población por país. https://bit.ly/3sTwwYl

Requena Serra, Bernart (2014). Diagrama de barras. Universo Fórmulas. https://bit.ly/2Q1L7PQ

Rodríguez, E. (2005). Metodología de la investigación. Tabasco, México: Universidad Juárez Autónoma de Tabasco.

Saiz Carvajal, Rosario (2016). Técnicas de análisis de información (resumen). https://bit.ly/3zA53ia

Malhotra, N. (2004). Investigación de mercados: un enfoque aplicado. 4ta edición. México: Pearson Educación.

Muñoz, C. (2018). Metodología de la investigación. México D. F., México: Editorial Progreso S. A.

Pardinas, F. (2005). Metodología y técnicas de investigación en ciencias sociales. 38va edición. México D. F., México: Siglo XXI Editores S. A.

American Psychological Association (2002). Manual de estilo de publicaciones de la American Psychological Association. México: Editorial El Manual Moderno. Normas Apa (2019). Normas Apa 2019 Actualizadas. Https://Normasapa.Com/

Rodríguez, E. (2005). Metodología de la Investigación. México: Publicaciones de la Universidad Juárez Autónoma de Tabasco. SANCHEZ, CARLOS (2014). Normas Apa – 7ma (Séptima) Edición. Https://Normas-Apa.Org/

Otras referencias de uso y de investigación

Big Data con Python – Recolección, almacenamiento y procesos. Por R. Caballero, E. Martin & A. Riesco. (Libro personal).

Programación en C, C++, Java y UML Ingeniería de Software. Por Luis Joyannes e Ignacio Zahonero. (Libro personal).

Estadística descriptiva, probabilidades – Inferencia – Modelos de regresión y métodos no paramétricos. Por Pedro Vergara Vera. (Libro personal).

BOYD, D. (2007). "Social network sites: Definition, history, and scholarship". Journal of
Computer-Mediated Communication.

BULUT, A. (2015). Lean Marketing: Know who not to advertise to! Electronic Commerce
Research and Applications.

CHUNGA, A., ANDREEVA, P., BENYOUCEF, M., DUANE, A., O'REILLY, P. (2017). Managing an
organisation's social media presence: An empirical stages of growth model. International
Journal of Information Management.

DEBES, V., SANDEEP, K. AND VINNETT, G. (2017). Predicting information diffusion
probabilities in social networks: A Bayesian networks-based approach. Journal of Knowledge-
Based Systems.

FAGAN, J. C. (2014). The Suitability of Web Analytics Key Performance Indicators in the
Academic Library Environment. The Journal of Academic Librarianship.

FILE, K. M., AND PRINCE, R. A. (1993). Evaluating the effectiveness of interactive
marketing. Journal of Services Marketing, 7(3), 49-58. doi:10.1108/08876049310044574

FIORINI, P. M., AND LIPSKY, L. R. (2012). Search marketing traffic and performance models.
Computer Standards and Interfaces.

HERRÁEZ, B., BUSTAMANTE, D. Y SAURA, J.R. (2017). Information classification on social
networks. Content analysis of e-commerce companies on Twitter.

HONEYCUTT, C., & HERRING, S. C (2009). Beyond microblogging: Conversation and
collaboration via Twitter. In 42nd Hawaii International Conference on System Sciences.

JÄRVINEN, J., AND KARJALUOTO, H. (2015). The use of Web analytics for digital marketing
performance measurement. Industrial Marketing Management.

LEEFLANG, P, VERHOEF, P., DAHSLTRÖM, P. Y FREUNDT, T. (2014). Challenges and solutions
for marketing in a digital era. European Management Journal.

MATHEWS, S., BIANCHI, C., PERKS, K. J., HEALY, M., AND WICKRAMASEKERA, R. (2016).
Internet marketing capabilities and international market growth. International Business
Review.

MCDERMOTT, J. (2017, December 03). Black Friday stats: The numbers behind the madness.
Retrieved April 18, 2018.

PAK, A., & PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In
Proceedings of LREC, 2010. Valletta, Malta.

PALOS-SANCHEZ, P.; SAURA, J.R. (2018). The Effect of Internet Searches on Afforestation:
The Case of a Green Search Engine.

SAURA, J. R., PALOS-SÁNCHEZ, P., & CERDÁ SUÁREZ, L. M. (2017). Understanding the
Digital Marketing Environment with KPIs and Web Analytics. Future Internet, 9(4), 76.

SAURA, J.R., PALOS-SANCHEZ, P.R. & RIOS MARTIN, M.A. (2018). Attitudes to
environmental factors in the tourism sector expressed in online comments: An exploratory

study. International Journal of Environmental Research and Public Health.

T.T., KUO, S.-C. HUNG, W.-S. LIN, N. PENG, S.-D. LIN AND W.F.LIN (2012). Exploiting latent information to predict diffusions of novel topics on social networks, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics.

Algorithmic future. Big data, sensor data and mobile media, Co-authors: Jose Correa and Charles Thraves.

Sort Algorithms and Data structures, por Nchena Linos.

Data Structures and Algorithms with JavaScript, por Wahyu Prayogo. JS & Data Structures.

Programming Problems: Advanced Algorithms (Volume 2) Paperback – February 27, 2013, por Guido Noto La Diega, PhD.

Superintelligence: Paths, Dangers, Strategies Reprint Edition, por Nick Bostrom (Author), PhD.