



## **UNIVERSIDAD DE GUAYAQUIL**

**FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS  
CARRERA DE INGENIERÍA EN SISTEMAS  
COMPUTACIONALES**

**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN  
DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO  
EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL  
ENTORNO DE TRABAJO JUPYTER NOTEBOOK  
Y EL LENGUAJE DE PROGRAMACIÓN  
PYTHON.**

### **PROYECTO DE TITULACIÓN**

Previa a la obtención del Título de:

**INGENIERO EN SISTEMAS COMPUTACIONALES**

AUTOR:

**Edinson Andrés Jiménez Cárdenas**

TUTOR:

**Ing. Jorge Avilés Monroy, M.Sc.**

**GUAYAQUIL – ECUADOR**

**2019**

 Presidencia de la República del Ecuador	 <b>Plan Nacional</b> de Ciencia, Tecnología, Innovación y Saberes	 <b>SENECYT</b> <small>Beca para la formación Superior,          Ciencia, Tecnología e Innovación</small>
<b>REPOSITORIO NACIONAL EN CIENCIAS Y TECNOLOGÍA</b>		
<b>FICHA DE REGISTRO DE TESIS</b>		
<p><b>TÍTULO:</b> "ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON"</p>		
<p><b>AUTOR:</b>           Jiménez Cárdenas Edinson Andrés       </p>	<p><b>TUTOR:</b> Ing. Jorge Avilés Monroy, M.Sc.</p>	
	<p><b>REVISORES:</b> Ing. Segundo Delgado Menoscal, PhD.</p>	
<p><b>INSTITUCIÓN:</b> Universidad de Guayaquil</p>	<p><b>FACULTAD:</b> Ciencias Matemáticas y Físicas</p>	
<p><b>CARRERA:</b> Ingeniería en Sistemas Computacionales</p>		
<p><b>FECHA DE PUBLICACIÓN:</b></p>	<p><b>Nº DE PAGINAS:</b> 158</p>	
<p><b>ÁREA DE LA TEMÁTICA:</b> Desarrollo local y emprendimiento socio económico sostenible y sustentable.</p>		
<p><b>PALABRAS CLAVE:</b> Análisis de datos, red social Twitter, algoritmos de aprendizaje automático, procesamiento del lenguaje natural, análisis de sentimientos.</p>		
<p><b>RESUMEN:</b> El presente proyecto fue enfocado en la extracción y análisis de los datos que genera la red social Twitter para contribuir a la toma de decisiones, mediante el uso de análisis de sentimiento en español y algoritmos de aprendizaje automático, generando información útil de apoyo en la toma de decisiones a los emprendedores de la ciudad de Guayaquil con base a la problemática planteada al inicio de sus actividades, o al momento de re-direccionar su negocio, la cual consiste en identificar si existe una audiencia para un determinado producto o servicio y examinar qué tipo de sectores e industrias acaparan la mayor cantidad de internautas, la cual ayudaría en el proceso de toma de decisiones en los emprendimientos suscitados en esta ciudad. Una de las metodologías utilizadas en el presente proyecto fue la investigación diagnostica, entre las principales herramientas que posee esta metodología para la recolección de datos y su posterior análisis es la encuesta. Entre los algoritmos a utilizar destacan, algoritmo de clasificación Bayesiano y el uso de una red neuronal para clasificar los tweets por sectores e industrias. El desarrollo de este proyecto fue realizado en Python, en el entorno de desarrollo de Jupyter el cual provee facilidades de uso e implementación de librerías y algoritmos de aprendizaje automático.</p>		
<p><b>Nº DE REGISTRO (en la base de datos):</b></p>	<p><b>No DE CLASIFICACIÓN:</b></p>	
<p><b>DIRECCIÓN URL(tesis en la web):</b></p>		
<p><b>ADJUNTO PDF</b></p>	<p>SI (X)</p>	<p>NO ( )</p>
<p><b>CONTACTO AUTOR:</b></p>	<p><b>Teléfono:</b> 0979589563</p>	<p><b>E-Mail:</b>  <a href="mailto:edinson.jimenezc@ug.edu.ec">edinson.jimenezc@ug.edu.ec</a>  <a href="mailto:jimenezcardenas95@gmail.com">jimenezcardenas95@gmail.com</a> </p>
<p><b>CONTACTO DE LA INSTITUCIÓN:</b> Universidad de Guayaquil</p>	<p><b>Nombre:</b> Ab. Juan Chávez Atocha</p>	
	<p><b>Teléfono:</b> 2307729</p>	

## **APROBACIÓN DEL TUTOR**

En mi calidad de Tutor del trabajo de titulación, “**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON**” Elaborado por el Sr. EDINSON ANDRÉS JIMÉNEZ CÁRDENAS, **Alumno no titulado** de la Carrera de Ingeniería en Sistemas Computacionales, Facultad de Ciencias Matemáticas y Físicas de la Universidad de Guayaquil, previo a la obtención del Título de Ingeniero en Sistemas Computacionales, me permito declarar que luego de haber orientado, estudiado y revisado, la Apruebo en todas sus partes.

**Atentamente**

---

**Ing. Jorge Avilés Monroy, M.Sc.**

**TUTOR**

## **DEDICATORIA**

Se la dedico de manera especial a Dios, a mi familia dado que de ellos aprendí la perseverancia, y han sido el pilar fundamental, en toda mi educación, tanto académica, como en la vida. Este trabajo ha sido posible gracias a ellos.

También dedico este trabajo a mis compañeros de clases, a esos maestros que nos brindaron sus conocimientos en las diferentes asignaturas.

Y sobre todo a mi madre la Sra. Nancy Cárdenas Leyton, ella ha sido la razón por la cual estoy culminando mi carrera y nadie se merece más mi esfuerzo que ella.

Jiménez Cárdenas Edinson Andrés.

## **AGRADECIMIENTO**

Agradezco ante todo a Dios, por permitirme culminar mi carrera universitaria, a mis padres que desde el primer día que ingresé a estudiar Ingeniería en Sistemas Computacionales han estado apoyándome en cada momento de mi vida.

A mis amigos que de una manera u otra aportaron muchísimo durante mi estancia en la universidad, agradezco en especial a la Dra. Sandra Regalado, y a los que aportaron con su granito de arena durante todo mi proceso de titulación.

Jiménez Cárdenas Edinson Andrés.

## **TRIBUNAL PROYECTO DE TITULACIÓN**

---

**Ing. Fausto Cabrera Montes, M.Sc.  
DECANO DE LA FACULTAD  
CIENCIAS MATEMÁTICAS Y  
FÍSICAS**

---

**Ing. Gary Reyes Zambrano, Mgs.  
DIRECTOR DE LA CARRERA DE  
INGENIERÍA EN SISTEMAS  
COMPUTACIONALES**

---

**Ing. Jorge Avilés Monroy, M.Sc.  
PROFESOR TUTOR DEL  
PROYECTO  
DE TITULACIÓN**

---

**Ing. Segundo Delgado Menocal, PhD.  
PROFESOR REVISOR DEL  
PROYECTO  
DE TITULACIÓN**

---

**Ab. Juan Chávez Atocha, Esp.  
SECRETARIO**

## **DECLARACIÓN EXPRESA**

“La responsabilidad del contenido de este Proyecto de Titulación, me corresponden exclusivamente; y el patrimonio intelectual de la misma a la UNIVERSIDAD DE GUAYAQUIL”

---

JIMÉNEZ CÁRDENAS EDINSON ANDRÉS

C.C. 0929377950



UNIVERSIDAD DE GUAYAQUIL  
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

**CARRERA DE INGENIERÍA EN SISTEMAS  
COMPUTACIONALES**

**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE  
PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA  
CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE  
TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE  
DE PROGRAMACIÓN PYTHON**

Proyecto de Titulación que se presenta como requisito para optar por el título  
de INGENIERO EN SISTEMAS COMPUTACIONALES

**Autor:** JIMÉNEZ CÁRDENAS EDINSON ANDRÉS

C.C. 0929377950

**Tutor:** Ing. JORGE AVILÉS MONROY, M.SC.

Guayaquil, septiembre de 2019

## **CERTIFICADO DE ACEPTACIÓN DEL TUTOR**

En mi calidad de Tutor del proyecto de titulación, nombrado por el Consejo Directivo de la Facultad de Ciencias Matemáticas y Físicas de la Universidad de Guayaquil.

### **CERTIFICO:**

Que he analizado el Proyecto de Titulación presentado por el estudiante **JIMÉNEZ CÁRDENAS EDINSON ANDRÉS**, como requisito previo para optar por el título de Ingeniero en Sistemas Computacionales cuyo problema es:

**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON**

Considero aprobado el trabajo en su totalidad.

Presentado por:

**JIMÉNEZ CÁRDENAS EDINSON ANDRÉS**

**C.C. 0929377950**

**Tutor: Ing. JORGE AVILÉS MONROY, M.SC.**

Guayaquil, septiembre de 2019



**UNIVERSIDAD DE GUAYAQUIL**  
**FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS**  
**CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES**

**Autorización para Publicación de Proyecto de Titulación en Formato Digital**

**1. Identificación del Proyecto de Titulación**

<b>Nombre Alumno:</b> Edinson Andrés Jiménez Cárdenas	
<b>Dirección:</b> Duran, Coop. 5 de junio Mz. E1 Sl. 11.	
<b>Teléfono:</b> 0979589563	<b>E-mail:</b> edinson.jimenezc@ug.edu.ec

<b>Facultad:</b> Ciencias Matemáticas y Físicas
<b>Carrera:</b> Ingeniería en Sistemas Computacionales
<b>Proyecto de titulación al que opta:</b> Ingeniero en Sistemas Computacionales
<b>Profesor guía:</b> Ing. Jorge Avilés Monroy, M.Sc.

**Título del Proyecto de titulación:** Análisis de la red social Twitter para la identificación de patrones que generan oportunidades de negocio en la ciudad de Guayaquil utilizando el entorno de trabajo Jupyter notebook y el lenguaje de programación Python

**Tema del Proyecto de Titulación:** Análisis de tweets públicos emitidos en la ciudad de Guayaquil, aplicando algoritmos de aprendizaje automático para obtener información que brinde apoyo en la toma de decisiones de emprendedores y dueños de negocios.

**2. Autorización de Publicación de Versión Electrónica del Proyecto de Titulación**

A través de este medio autorizo a la Biblioteca de la Universidad de Guayaquil y a la Facultad de Ciencias Matemáticas y Físicas a publicar la versión electrónica de este Proyecto de titulación.

**Publicación electrónica:**

Inmediata	<input checked="" type="checkbox"/>	Después de 1 año	<input type="checkbox"/>
-----------	-------------------------------------	------------------	--------------------------

Firma Alumno:

\_\_\_\_\_  
Edinson Andrés Jiménez Cárdenas

C.I. 0929377950

**3. Forma de envío:**

El texto del proyecto de titulación debe ser enviado en formato Word, como archivo .Doc. O .RTF y .Puf para PC. Las imágenes que la acompañen pueden ser: .gif, .jpg o .TIFF.

DVDROM

CDROM

X

## ÍNDICE GENERAL

APROBACIÓN DEL TUTOR .....	III
DEDICATORIA.....	IV
AGRADECIMIENTO.....	V
ÍNDICE GENERAL.....	XI
ABREVIATURAS .....	XIV
ÍNDICE DE CUADROS .....	XVI
ÍNDICE DE GRÁFICOS .....	XVII
RESUMEN.....	XXII
ABSTRACT .....	XXIII
INTRODUCCIÓN .....	1
<b>CAPÍTULO I - EL PROBLEMA .....</b>	<b>4</b>
Ubicación del problema en un contexto .....	4
Situación Conflicto Nudos Críticos.....	5
Causas y consecuencias del problema.....	6
Delimitación del problema .....	7
Formulación del Problema .....	7
Evaluación del problema.....	7
Objetivos .....	9
Alcances del problema .....	10
Justificación e importancia.....	11
Metodología del proyecto.....	14
<b>CAPÍTULO II - MARCO TEÓRICO .....</b>	<b>17</b>
Fundamentación teórica .....	20
Las redes sociales en Ecuador .....	20
Red Social Twitter.....	22
Información que se puede extraer de Twitter.....	23
API Twitter.....	24
Filtrar tweets en tiempo real .....	25
POST estados / filtro .....	27
Datos geolocalizados.....	28

Localizaciones .....	29
Formato de salida JSON.....	31
Limitaciones de Twitter en su API.....	32
Procesamiento del Lenguaje Natural (NLP) .....	32
Análisis de sentimientos.....	33
Minería de datos (Data Mining) .....	36
Extracción de datos .....	38
Inteligencia artificial (IA).....	39
Aprendizaje supervisado .....	43
Aprendizaje no supervisado .....	48
Bolsa de palabras (Bag of Words).....	49
Open Source .....	50
Entorno de desarrollo: Jupyter Notebook.....	51
Librería NumPy.....	55
Twython .....	55
Tweepy .....	56
Librería Pandas.....	57
DataFrame .....	57
Visualización de información para la ayuda a la toma de decisiones .....	58
<b>FUNDAMENTACIÓN LEGAL .....</b>	<b>59</b>
<b>LEY DE PROPIEDAD INTELECTUAL .....</b>	<b>59</b>
Objeto del derecho de autor .....	60
LOTAIP (Ley Orgánica de Transparencia y Acceso a la Información Pública). .	64
Ley sobre el acuerdo de software libre en el Ecuador.....	65
Preguntas científicas a contestarse .....	65
Definiciones conceptuales.....	66
<b>CAPÍTULO III - PROPUESTA TECNOLÓGICA .....</b>	<b>69</b>
Metodología de investigación .....	75
Población y muestra .....	77
Encuesta N. 1.....	80
Encuesta N. 2.....	90

Metodología CRISP-DM.....	101
Descripción de fases de CRISP-DM .....	103
Aplicación de la metodología propuesta .....	106
Fase I. Comprensión del negocio .....	106
Fase II. Comprensión de los Datos.....	107
Fase III. Preparación de los datos.....	116
Fase IV. Modelado .....	122
Construcción del modelo.....	126
Fase V. Evaluación.....	128
Presentación de resultados .....	134
Entregables del proyecto .....	145
CRITERIOS DE VALIDACIÓN DE LA PROPUESTA .....	146
Procesamiento y análisis .....	147
Criterios generales.....	147
Criterios de validación de notebooks .....	148
<b>CAPÍTULO IV - CRITERIOS DE ACEPTACIÓN DEL PRODUCTO ....</b>	<b>150</b>
Privacidad de los datos .....	156
Conclusiones .....	158
Recomendaciones.....	160
<b>BIBLIOGRAFÍA .....</b>	<b>161</b>
ANEXO 1. CRONOGRAMA DE ACTIVIDADES.....	172
ANEXO 2. ENCUESTAS.....	173
ANEXO 3. CRITERIOS DE VALIDACIÓN DE LA PROPUESTA .....	179
ANEXO 4. CRITERIOS DE ACEPTACIÓN DEL PRODUCTO .....	185
ANEXO 5. MANUAL DE USUARIO .....	1
ANEXO 6. MANUAL TÉCNICO .....	1

## ABREVIATURAS

API	Application Programming Interface (Interfaz de programación de aplicaciones)
AS	Análisis de Sentimientos
CSV	Archivo de valores separado por coma
DML	Data Manipulation Language (Lenguaje de Manipulación De Datos)
I.A.	Inteligencia Artificial
ING	Ingeniero
JSON	Notación de Objetos de JavaScript
M.Sc	Master
ML	Machine Learning
NB	Naïve Bayes
NLP	Natural language Processing
NN	Neural Network
PLN	Procesamiento de Lenguaje Natural
PYMES	Pequeñas y medianas empresas
SEPLN	Spanish Society for Natural Language Processing
SVM	Support Vector Machines
TASS	Taller de Análisis Semántico en la SEPLN
TI	Tecnologías de la Información
WWW	World Wide Web (red mundial)

## **SIMBOLOGÍA**

N	Tamaño de la población
n	Tamaño de la muestra
E	Margen de error
K	Nivel de confianza
P	Probabilidad de éxito
Q	probabilidad de fracaso

## ÍNDICE DE CUADROS

<b>CUADRO N. 1 CAUSAS Y CONSECUENCIAS DEL PROBLEMA .....</b>	6
<b>CUADRO N. 2 DELIMITACIÓN DEL PROBLEMA.....</b>	7
<b>CUADRO N. 3 OPCIONES PARA TRANSMITIR TWEETS EN TIEMPO REAL.....</b>	27
<b>CUADRO N. 4 PARÁMETROS PARA FILTRAR TWEETS.....</b>	28
<b>CUADRO N. 5 EJEMPLO DE CAJA DELIMITADORA .....</b>	30
<b>CUADRO N. 6 LIBRERÍAS DE PYTHON CON SUS FUNCIONALIDADES .....</b>	54
<b>CUADRO N. 7 HARDWARE UTILIZADO EN EL DESARROLLO .....</b>	72
<b>CUADRO N. 8 SOFTWARE UTILIZADO EN EL DESARROLLO .....</b>	73
<b>CUADRO N. 9 PRESUPUESTO Y FINANCIAMIENTO.....</b>	74
<b>CUADRO N. 10 POBLACIÓN Y MUESTRA DE LA ENCUESTA N. 1.....</b>	78
<b>CUADRO N. 11 POBLACIÓN Y MUESTRA DE LA ENCUESTA N. 2.....</b>	79
<b>CUADRO N. 12 COMPONENTES Y FUNCIONALIDAD DE NOTEBOOKS DEL PROYECTO..</b>	130
<b>CUADRO N. 13 PERFIL DE LOS EXPERTOS.....</b>	146
<b>CUADRO N. 14 CRITERIOS DE VALIDACIÓN.....</b>	147
<b>CUADRO N. 15 MATRIZ DE RESULTADOS DE CRITERIOS DE ACEPTACIÓN .....</b>	151
<b>CUADRO N. 16 CRITERIOS DE ACEPTACIÓN .....</b>	152
<b>CUADRO N. 17 VALORES DE CALIFICACIÓN DE CRITERIOS DE APROBACIÓN .....</b>	154
<b>CUADRO N. 18 RESULTADOS DE CALIFICACIÓN DE CRITERIOS EXPERTO N. 1 .....</b>	154
<b>CUADRO N. 19 RESULTADOS DE CALIFICACIÓN DE CRITERIOS EXPERTO N. 2 .....</b>	155
<b>CUADRO N. 20 RESULTADOS DE CALIFICACIÓN DE CRITERIOS EXPERTO N. 3 .....</b>	155

## ÍNDICE DE GRÁFICOS

	Pág.
<b>GRÁFICO N. 1</b> PERFIL USUARIOS DE TWITTER POR EDADES EN ECUADOR .....	21
<b>GRÁFICO N. 2</b> AUDIENCIA DE REDES EN ECUADOR. ....	22
<b>GRÁFICO N. 3</b> FORMATO JSON DEVUELTO COMO PETICIÓN AL API DE TWITTER.....	31
<b>GRÁFICO N. 4</b> TÉCNICAS DE APRENDIZAJE AUTOMÁTICO.....	42
<b>GRÁFICO N. 5</b> FUNCIONAMIENTO DE SVM. ....	45
<b>GRÁFICO N. 6</b> RED NEURONAL ARTIFICIAL MULTICAPA. ....	47
<b>GRÁFICO N. 7</b> REPRESENTACIÓN GRÁFICA DE UN ANÁLISIS DE SENTIMIENTO .....	58
<b>GRÁFICO N. 8</b> PREGUNTA 1. ¿POSEE CUENTA EN LA RED SOCIAL TWITTER? .....	81
<b>GRÁFICO N. 9</b> PREGUNTA 2. ¿CON QUÉ FRECUENCIA EMITE TWEETS EN SU CUENTA DE TWITTER?.....	82
<b>GRÁFICO N. 10</b> PREGUNTA 3. ¿SE CONSIDERA UN USUARIO ACTIVO EN LA RED SOCIAL TWITTER?.....	83
<b>GRÁFICO N. 11</b> PREGUNTA 4. ¿CREE QUE ACTUALMENTE PUEDE EXPRESAR LIBREMENTE SU OPINIÓN SOBRE CUALQUIER TEMA EN LA RED SOCIAL TWITTER? 84	
<b>GRÁFICO N. 12</b> PREGUNTA 5. ¿ALGUNA VEZ HA EXPRESADO EN LA RED SOCIAL TWITTER LA NECESIDAD O DESEO SOBRE UN BIEN O SERVICIO?.....	85
<b>GRÁFICO N. 13</b> PREGUNTA 6. ¿CON QUÉ FRECUENCIA HA EXPRESADO SU NECESIDAD O DESEO SOBRE UN BIEN O SERVICIO EN LA RED SOCIAL TWITTER? .....	86

<b>GRÁFICO N. 14</b> PREGUNTA 7. ¿CUÁNDO USTED ADQUIERE, CONSUME O UTILIZA ALGÚN BIEN, PRODUCTO O SERVICIO LO HA PUBLICADO EN LA RED SOCIAL TWITTER? ....	87
<b>GRÁFICO N. 15</b> PREGUNTA 8. ¿CON QUÉ FRECUENCIA HA EXPRESADO USTED SU OPINIÓN SOBRE UN PRODUCTO O SERVICIO EN LA RED SOCIAL TWITTER?.....	88
<b>GRÁFICO N. 16</b> PREGUNTA 9. ¿SABÍA USTED QUE SUS PUBLICACIONES EN LA RED SOCIAL TWITTER PUEDEN AYUDAR EN LA TOMA DE DECISIONES A NIVEL EMPRESARIAL Y DE EMPRENDIMIENTO?.....	89
<b>GRÁFICO N. 17</b> PREGUNTA 10. ¿ESTÁ USTED DE ACUERDO QUE SUS TWEETS PÚBLICOS (CON CARACTERÍSTICAS ESPECÍFICAS) SEAN ANALIZADOS PARA BRINDAR APOYO EN LA TOMA DE DECISIONES EN LOS EMPRENDIMIENTOS DE LA CIUDAD DE GUAYAQUIL? .....	90
<b>GRÁFICO N. 18</b> PREGUNTA 1. SELECCIONE SU SITUACIÓN ACTUAL.....	91
<b>GRÁFICO N. 19</b> PREGUNTA 2. UBIQUE EL SECTOR DE SU NEGOCIO O EMPRENDIMIENTO EN LA SIGUIENTE LISTA. .....	92
<b>GRÁFICO N. 20</b> PREGUNTA 3. ¿SABÍA USTED QUE ECUADOR SE POSICIONA COMO LÍDER EN LA REGIÓN EN EL ÍNDICE DE ACTIVIDAD EMPRENDEDORA TEMPRANA? .....	93
<b>GRÁFICO N. 21</b> PREGUNTA 4. ¿EN SU EMPRENDIMIENTO, UTILIZA LAS REDES SOCIALES PARA PROMOCIONAR SU SERVICIO O PRODUCTO Y MANTENER CONTACTO CON SUS CLIENTES? .....	94
<b>GRÁFICO N. 22</b> PREGUNTA 5. ¿EN SU EMPRENDIMIENTO O NEGOCIO ESTABLECIDO, HACE USO DE LA RED SOCIAL TWITTER?.....	95

<b>GRÁFICO N. 23</b> PREGUNTA 6. EN CASO DE USAR TWITTER, INDIQUE LA FRECUENCIA CON LA QUE USTED LEE COMENTARIOS DE SUS CLIENTES.....	96
<b>GRÁFICO N. 24</b> PREGUNTA 7. CONSIDERA USTED QUE EL TIEMPO QUE EMPLEA PARA LEER LAS OPINIONES EN TWITTER DE SUS POTENCIALES CLIENTES ES SUFFICIENTE PARA DETERMINAR GUSTOS Y NECESIDADES.....	97
<b>GRÁFICO N. 25</b> PREGUNTA 8. ¿SI EXISTIERA UNA PLATAFORMA WEB QUE MEDIANTE SU USO LE PERMITA VISUALIZAR QUE SECTORES/INDUSTRIAS TIENEN MAYOR ACTIVIDAD CON LOS USUARIOS DE LA RED SOCIAL TWITTER, USTED LA USARÍA?.....	98
<b>GRÁFICO N. 26</b> PREGUNTA 9. SI EXISTE UNA PLATAFORMA WEB QUE MEDIANTE SU USO LE PERMITA VISUALIZAR INFORMACIÓN QUE BRINDE APOYO EN LA TOMA DECISIONES AL MOMENTO DE EMPRENDER, ¿ESTARÍA DISPUESTO A USARLA Y CON QUÉ FRECUENCIA?.....	99
<b>GRÁFICO N. 27</b> PREGUNTA 10. ¿QUÉ TIPO DE INFORMACIÓN CREE USTED QUE ES DETERMINANTE PARA EL APOYO EN LA TOMA DE DECISIONES AL MOMENTO DE EMPRENDER?.....	100
<b>GRÁFICO N. 28</b> SECUENCIA DEL PROCESO CRISP-DM.....	102
<b>GRÁFICO N. 29</b> FASES DE LA METODOLOGÍA CRISP-DM .....	103
<b>GRÁFICO N. 30</b> LLAVES Y TOKEN'S ESCRITOS EN EL NOTEBOOK PARA ACCEDER AL API .....	109
<b>GRÁFICO N. 31</b> OBJETO JSON OBTENIDO DE TWYTHON DE APROXIMADAMENTE LINEAS.....	110
<b>GRÁFICO N. 32</b> TWEETS EXTRAÍDOS MEDIANTE EL USO DE TWYTHON.....	111

<b>GRÁFICO N. 33</b> OBJETO JSON OBTENIDO DE TWEETPY CONSTA APROXIMADAMENTE DE 606 LÍNEAS.....	112
<b>GRÁFICO N. 34</b> VISUALIZACIÓN DE ARCHIVO CSV QUE ALMACENA LOS TWEETS.....	114
<b>GRÁFICO N. 35</b> ALMACENAMIENTO DE TWEETS EN FIREBASE. .....	115
<b>GRÁFICO N. 36</b> DATASET PARA EL TASS EMITIDO POR LA SEPLN. ....	118
<b>GRÁFICO N. 37</b> TRAINING Y TEST USANDO EL DATASET DE LA SEPLN.....	119
<b>GRÁFICO N. 38</b> CONJUNTO DE CONOCIMIENTO SUPERVISADO.....	120
<b>GRÁFICO N. 39</b> PROCESO DE ANÁLISIS DE SENTIMIENTO.....	125
<b>GRÁFICO N. 40</b> PARTE DEL CÓDIGO DEL ALGORITMO DE CLASIFICACIÓN.....	127
<b>GRÁFICO N. 41</b> ENTRENAMIENTO DE LA RED NEURONAL .....	128
<b>GRÁFICO N. 42</b> NIVEL DE PREDICCIÓN DEL ALGORITMO DE CLASIFICACIÓN. ....	129
<b>GRÁFICO N. 43</b> ARQUITECTURA NOTEBOOK EXTRACCIÓN DE TWEETS.....	132
<b>GRÁFICO N. 44</b> ARQUITECTURA NOTEBOOK DE BÚSQUEDA DE PRODUCTO O SERVICIO. .....	133
<b>GRÁFICO N. 45</b> ARQUITECTURA NOTEBOOK DE CLASIFICACIÓN POR SECTORES E INDUSTRIAS.....	134
<b>GRÁFICO N. 46</b> CIENCIA DE DATOS EN LA TOMA DE DECISIONES .....	135
<b>GRÁFICO N. 47</b> FLUJO DE ACTIVIDADES MÓDULO DE CLASIFICACIÓN DE TWEETS POR SU POLARIDAD .....	136
<b>GRÁFICO N. 48</b> BÚSQUEDA EN EL DATASET DE VALIDACIÓN .....	137
<b>GRÁFICO N. 49</b> TENDENCIAS RELACIONADAS AL PRODUCTO O SERVICIO.....	137
<b>GRÁFICO N. 50</b> NUBE DE PALABRAS CON MAYOR ÍNDICE DE FRECUENCIA. ....	138

<b>GRÁFICO N. 51</b> VALORACIÓN POR SENTIMIENTOS DE UN PRODUCTO O SERVICIO CON BASE EN EL ANÁLISIS DE OPINIONES DEL DATASET .....	139
<b>GRÁFICO N. 52</b> GRÁFICO DE BARRAS POR SECTORES .....	141
<b>GRÁFICO N. 53</b> VISUALIZACIÓN DE POLARIDAD EN TWEETS .....	142
<b>GRÁFICO N. 54</b> VISUALIZACIÓN GENERAL EN LÍNEA DE TIEMPO .....	143
<b>GRÁFICO N. 55</b> ÍNDICE DE INTERACCIÓN DE TWEETS CON UN SECTOR EN LÍNEA DE TIEMPO .....	144
<b>GRÁFICO N. 56</b> LECTURA DE ARCHIVO PDF GENERADO .....	145
<b>GRÁFICO N. 57</b> PROMEDIO DE CRITERIO GENERAL.....	148
<b>GRÁFICO N. 58</b> PROMEDIO DE CRITERIO POR NOTEBOOKS.....	149



**UNIVERSIDAD DE GUAYAQUIL**  
**FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS**  
**CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES**

**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE  
PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA  
CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE  
TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE  
DE PROGRAMACIÓN PYTHON.**

**Autor:** Edinson Andres Jiménez Cárdenas

**Tutor:** Ing. Jorge Avilés Monroy, M.Sc.

**Resumen**

El presente proyecto fue enfocado en la extracción y análisis de los datos que genera la red social Twitter para contribuir a la toma de decisiones, mediante el uso de análisis de sentimiento en español y algoritmos de aprendizaje automático, generando información útil de apoyo en la toma de decisiones a los emprendedores de la ciudad de Guayaquil con base a la problemática planteada al inicio de sus actividades, o al momento de re-direccionar su negocio, la cual consiste en identificar si existe una audiencia para un determinado producto o servicio y examinar que tipo de sectores e industrias acaparan la mayor cantidad de internautas, la cual ayudaría en el proceso de toma de decisiones en los emprendimientos suscitados en esta ciudad. Una de las metodologías utilizadas en el presente proyecto fue la investigación diagnostica, entre las principales herramientas que posee esta metodología para la recolección de datos y su posterior análisis es la encuesta. Entre los algoritmos a utilizar destacan, algoritmo de clasificación Bayesiano y el uso de una red neuronal para clasificar los tweets por sectores e industrias. El desarrollo de este proyecto fue realizado en Python, en el entorno de desarrollo de Jupyter el cual provee facilidades de uso e implementación de librerías y algoritmos de aprendizaje automático.

**Palabras clave:** Análisis de datos, red social Twitter, algoritmos de aprendizaje automático, procesamiento del lenguaje natural, análisis de sentimientos.



**UNIVERSIDAD DE GUAYAQUIL**  
**FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS**  
**CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES**

ANALYSIS OF THE SOCIAL NETWORK TWITTER FOR THE IDENTIFICATION OF  
PATTERNS THAT GENERATE BUSINESS OPPORTUNITIES IN THE CITY OF  
GUAYAQUIL USING THE WORK ENVIRONMENT JUPYTER  
NOTEBOOK AND THE PROGRAMMING  
LANGUAGE PYTHON.

**Author:** Edinson Andres Jiménez Cárdenas

**Advisor:** Ing. Jorge Avilés Monroy, M.Sc.

**Abstract**

The present project was focused on the extraction and analysis of the data generated by the social network Twitter to contribute to decision making, through the use of sentiment analysis in Spanish and automatic learning algorithms, generating useful information to support decision making to entrepreneurs in the city of Guayaquil based on the issues raised at the beginning of their activities, or at the moment of redirecting your business, which consists of identifying if there is an audience for a certain product or service and examining what type of sectors and industries capture the largest number of Internet users, which would help in the decision-making process in the ventures raised in this city. One of the methodologies used in this project was diagnostic research, among the main tools that this methodology has for data collection and subsequent analysis is the survey. Among the algorithms to use stand out, Bayesian classification algorithm and the use of a neural network to classify tweets by sectors and industries. The development of this project was carried out in Python, in Jupyter's development environment which provides ease of use and implementation of libraries and automatic learning algorithms.

**Keywords:** Data analysis, social network Twitter, automatic learning algorithms, natural language processing, sentiment analysis

## **INTRODUCCIÓN**

El presente proyecto de titulación tiene como objetivo desarrollar una serie de notebooks en el entorno de ciencia de datos más popular hoy en día, el cual es Jupyter, entre los procesos a desarrollar destaca la extracción en tiempo real de tweets emitidos en la ciudad de Guayaquil. Con la finalidad de aplicar algoritmos de aprendizaje automático para el procesamiento del lenguaje natural en el idioma español. Permitiendo generar información que sirva de apoyo en la toma de decisiones de los emprendedores o dueños de negocios.

¿Será que existe un gran índice de actividad en este sector o industria? Es una de las preguntas que la mayoría de los emprendedores se realizan al momento de emprender. En una era en que la información no estructurada es el nuevo petróleo. Nacen nuevos retos y oportunidades en la cual se pueden aplicar las tecnologías en auge. En el presente caso el uso de algoritmos de aprendizaje automático, que generen información basado en un análisis de datos públicos emitidos en la presente ciudad. Brindando apoyo en la toma de decisiones de los emprendedores interesados en hacer uso de estas tecnologías.

¿Qué herramientas existen para obtener un análisis de las opiniones emitidas en redes sociales? Es otra gran interrogante generada por el usuario final del presente proyecto. Se debe dejar en claro la amplia existencia de herramientas para monitorización de redes sociales y medir la presencia e impacto de una marca, el enfoque propuesto en este proyecto no se limita a cuentas específicas. Abarca un análisis de las opiniones públicas emitidas en la red social Twitter por ciudadanos guayaquileños.

Para llegar a un nivel óptimo de predicción estos algoritmos fueron entrenados con datos extraídos en la presente ciudad, para una mayor precisión en la comprensión semántica y clasificación realizada por los algoritmos utilizados.

Cabe recalcar que, basado solo en datos semánticos no se debe tomar decisiones. Por eso el presente proyecto es delimitado como apoyo en la toma de decisiones, mostrando información e insight que deben ser corroborados por los emprendedores para una toma correcta de decisiones basándose en su juicio final.

**En el capítulo I,** se describe la problemática existente y detectada para los fines académicos del presente proyecto. Problemas como la falta de empleo tanto en el sector público y privado, problemas relacionados con una especie de recesión económica que enfrenta el país que ha provocado que un gran número de personas decidan emprender. Las causas y consecuencias de no conocer sobre el uso de herramientas de ciencia de datos y la ventaja que puede generar en quienes hagan uso de ella, se plantean las delimitaciones, los objetivos generales y específicos. La correcta justificación e importancia y alcance del presente proyecto de titulación.

**En el capítulo II,** se describen los antecedentes del uso de machine Learning en la toma de decisiones, se definen los conceptos tecnológicos a utilizar a lo largo del proyecto y se define el marco legal en el que se encasilla el presente proyecto.

**En el capítulo III,** se presenta la propuesta tecnológica, la factibilidad del desarrollo del proyecto y se aplican cada una de las fases de la metodología CRISP-DM utilizada para el desarrollo del proyecto.

**En el capítulo IV,** se presentan los criterios de aceptación del presente proyecto, las conclusiones obtenidas durante el transcurso del desarrollo de los notebooks del presente proyecto y las respectivas recomendaciones para futuras versiones y nuevos enfoques basados en la experimentación realizada este trabajo de titulación.

# **CAPÍTULO I**

## **EL PROBLEMA**

### **PLANTEAMIENTO DEL PROBLEMA**

#### **Ubicación del problema en un contexto**

Actualmente Ecuador es uno de los países con mayor tasa de emprendimientos en Latinoamérica, 1 de cada 3 ecuatorianos, es emprendedor (más que en el resto de países de América Latina); sin embargo, el 90% de los emprendimientos en Ecuador, no llega a los tres años. (Flores, 2018)

En la economía de la ciudad de Guayaquil la mayoría de las personas que emprenden e inician sus actividades comerciales por lo general lo hacen sin realizar un estudio de mercado debido al costo que este representa y en gran medida se limitan a seguir a los demás o en lo que tienen más experiencia. Pero lo antes expresado no garantiza el éxito debido a que no se está escuchando y analizando lo que requiere el mercado en estos momentos o en el corto plazo.

No saber reconocer las tendencias comerciales o que sectores e industrias son las que presentan un mayor flujo de consumo en base al análisis de opiniones, ha limitado a los emprendedores a empezar en una línea comercial poco rentable o a seguir en una línea de negocio la cual ya no tiene mucho potencial de crecimiento.

## **Situación Conflicto Nudos Críticos**

Las personas que emprenden en la ciudad de Guayaquil, en su gran mayoría lo hacen sin realizar un análisis sobre la aceptación de un determinado producto o servicio, y no pueden determinar el nivel de interactividad que tienen los usuarios con un sector e industria. La ausencia de estos factores hace del emprendimiento una actividad con un alto índice de riesgo.

Los emprendedores necesitan información útil en el proceso de toma de decisiones para poder elegir (en caso de requerirlo) en qué sector existe un alto índice de interactividad por parte de los usuarios en una red social. Para decidir emprender en un sector con un determinado producto o servicio con gran aceptación por parte de los usuarios,

El principal problema que se plantea en este proyecto, es el poco uso de estas herramientas que permiten el análisis de datos generados en la red social Twitter y el desconocimiento de la utilización de las mismas, impidiendo obtener información útil en la toma de decisiones al momento de emprender o iniciar un negocio, mediante el análisis, procesamiento y clasificación de tweets. Generando información que permita a los emprendedores tener un panorama acerca de que sectores e industrias tienen mayor interactividad en base a la detección de patrones de consumo.

Por eso es de suma importancia realizar este proyecto para brindar apoyo en la toma de decisiones a los emprendedores en la selección de un sector comercial, que potencie sus negocios o emprendimientos suscitados en esta ciudad, presentando información relevante que ayude a fomentar el emprendimiento.

## Causas y consecuencias del problema

**Cuadro N. 1** Causas y consecuencias del problema

CAUSAS	CONSECUENCIAS
Poco conocimiento acerca de la obtención de información relevante a través de datos públicos.	Pérdida de información muy útil que brinde apoyo en la toma de decisiones.
Noción de una alta inversión para realizar un análisis comercial en los datos.	Los emprendedores pueden llegar a seleccionar un sector o industria en decadencia, en la cual no hay mucho índice de interactividad.
Industrias y sectores con bajo índice de consumo.	Perdida del capital invertido al empezar en un sector o industria en declive.
Librerías de procesamiento del lenguaje natural con funcionalidades disminuidas para el idioma español.	Dificulta la creación de proyectos que permitan el análisis de contenidos emitidos en el idioma español.
Inexistencia de proyectos que permitan a los emprendedores tener información de consumo por sectores e industrias.	Falta de interés en los posibles emprendedores por incertidumbre y nula información en la toma de decisiones.

**Nota:** Factores de incidencia en el emprendimiento.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Formación Gerencial, 2018).

## **Delimitación del problema**

**Cuadro N. 2 Delimitación del problema**

<b>CAMPO:</b>	DESARROLLO LOCAL Y EMPRENDIMIENTO SOCIO-ECONÓMICO SUSTENTABLE
<b>ÁREA:</b>	TECNOLOGÍAS DE LA INFORMACIÓN
<b>ASPECTO:</b>	ANÁLISIS DE DATOS CUANTITATIVOS
<b>TEMA:</b>	ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON.

---

**Nota:** Aspectos principales de respuesta a la problemática planteada.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

## **Formulación del Problema**

¿La utilización de algoritmos de aprendizaje automático aportará información a los emprendedores en la elección de un sector o industria, basado en la búsqueda de patrones de consumo existentes en los tweets emitidos en la ciudad de Guayaquil?

## **Evaluación del problema**

Ausencia de información comercial mediante un análisis social, que permita a los emprendedores tener una idea de la actividad existente en un sector/industria y en base a ello generar sus propias decisiones.

**Delimitado:** Este proyecto está orientado al análisis de los datos generados en la ciudad de Guayaquil en la red social Twitter para identificar patrones en los datos extraídos que den indicio de interactividad de los usuarios con sectores e industrias.

**Claro:** Presentar información gráfica que permita visualizar en qué sector o industria se está desarrollando mayor actividad en base a las opiniones emitidas por usuarios. Permitiendo obtener información útil que podría interesar e influir en los emprendimientos llevados a cabo en la ciudad de Guayaquil, todo esto a través de la utilización de técnicas de procesamiento del lenguaje natural.

**Evidente:** La información resultante será de gran aporte para los emprendedores que tienen o no experiencia en el mundo de los negocios, puesto que el resultado de este proyecto es proporcionar información de qué sectores e industrias son las que generan mayor interacción con los usuarios. A demás, mediante el uso del análisis de sentimiento en español poder medir la aceptación o rechazo hacia un producto o servicio, en base a la cantidad de datos extraídos.

**Concreto:** Con la utilización del entorno de desarrollo Jupyter Notebook podemos obtener resultados concretos sobre qué tendencias han sido utilizadas en un comentario que contiene un producto o servicio, gracias al análisis y extracción de hashtags existentes en los tweets, y así permitir a los emprendedores poder crear publicaciones o campañas más efectivas que les permitan conectar con una mayor audiencia.

**Factible:** Con el resultado final de este proyecto muchos emprendedores podrán tener un soporte de información que les brinde apoyo en la toma de decisiones sobre qué sectores e industrias tienen mayor interactividad con los usuarios de la red social Twitter en la ciudad de Guayaquil, lo cual será de gran ayuda para las decisiones de los emprendedores de esta ciudad, mediante el uso confiable de algoritmos de aprendizaje automático, como lo es el algoritmo de clasificación bayesiano.

**Relevante:** La importancia de este trabajo de titulación radica en proporcionar información relevante a los posibles emprendedores de la ciudad de Guayaquil, debido a que los datos analizados son generados por sus ciudadanos.

**Contextual:** El desarrollo de este proyecto establecerá relación entre la economía, ciencias matemáticas e inteligencia artificial, brindando apoyo en la toma de decisiones a los actuales y futuros emprendedores de la ciudad de Guayaquil, intentando que su probabilidad de éxito sea mayor.

**Identifica los productos esperados:** El resultante final de este proyecto son cuatro cuadernos Jupyter creados en la plataforma web Jupyter Notebook, con un código fuente escrito en el lenguaje de programación Python que contiene la aplicación de algoritmos de aprendizaje automático para generar información relevante acerca de la audiencia existente de un determinado producto o servicio en una ciudad especificada por coordenadas geográficas. Y una clasificación general por sectores e industrias de todas las opiniones extraídas. Con la finalidad de brindar apoyo en la toma de decisiones con base en el análisis de los datos extraídos.

## OBJETIVOS

### OBJETIVO GENERAL

Analizar los datos extraídos de la red social Twitter mediante el uso de algoritmos de aprendizaje automático en la plataforma Jupyter Notebook, para la obtención de información relacionada con la actividad comercial por sectores.

## **OBJETIVOS ESPECÍFICOS**

- Extraer datos públicos de la red social Twitter mediante el uso de su API en la plataforma Jupyter Notebook, para conformar un dataset de tweets generados en la ciudad de Guayaquil durante un periodo de cuatro meses, desde el 1 de mayo al 30 de agosto.
- Seleccionar los algoritmos de aprendizaje automático para la clasificación de tweets que contengan indicios comerciales.
- Aplicar técnicas de Machine Learning y utilizar librerías para la generación de gráficos en el entorno de trabajo Jupyter Notebook, para la visualización de gráficos e información específica.

## **ALCANCES DEL PROBLEMA**

- Seleccionar una herramienta que sea de fácil comprensión, para futuros usos del presente proyecto por parte de los emprendedores.
- Utilizar el lenguaje de marcado ligero Markdown para una mejor comprensión del flujo de los procedimientos aplicados en la plataforma Jupyter Notebook.
- Extraer datos de la red social Twitter mediante el uso de su API.
- Aplicar técnicas de Machine Learning para la clasificación y el reconocimiento de patrones en el entorno de trabajo Jupyter Notebook.
- Crear archivos en formato .csv, para almacenar los tweets y toda información relevante, permitiendo consultar información histórica para que los resultados sean más eficientes.
- Visualizar la información obtenida en cada algoritmo, utilizando la librería Matplotlib en el lenguaje de programación Python.
- Generar resultados en pantalla y descargables en formato PDF que presentarán los datos en gráficos estadísticos.

Las delimitaciones del proyecto serán las siguientes:

1. Los notebooks creados para el presente proyecto solo serán de carácter informativo y servirá como medio de visualización del flujo de trabajo del código fuente para los usuarios, a excepción del notebook que permite la búsqueda de un producto o servicio en el dataset para obtener información del mismo.
2. El proyecto no considera la fase de implementación, los procesos a nivel de prototipo se evaluarán utilizando el juicio de experto.
3. Notebook de análisis de sentimiento a un producto o servicio.  
Se realiza la carga del modelo de análisis de sentimiento en español, previamente desarrollado y entrenado para la clasificación de tweets. Incluye proceso de extracción de los tweets almacenados en Firebase, proceso de data cleaning y data wrangling. Ingreso del nombre de un producto o servicio para su búsqueda en el dataset. Clasificación de los sentimientos encontrados en los tweets referente a la búsqueda emitida. Presentación en forma gráfica de los resultados.
4. Notebook de clasificación de tweets por sectores.  
Se realiza la carga la red neuronal de clasificación de tweets por sectores, previamente desarrollada y entrenada para la clasificación de tweets por patrones de consumo que denoten pertenecía a un sector o industria. Incluye proceso de extracción de los tweets almacenados en Firebase, proceso de data cleaning, data wrangling y filtrado masivo de datos. Presentación en forma gráfica de los resultados.

## **JUSTIFICACIÓN E IMPORTANCIA**

Según (GEM Ecuador, 2014, pág. 23) afirma que: “Ecuador es considerado uno de los países más emprendedores de América Latina, pero la falta de la

calidad en los emprendimientos, poca retroalimentación de parte de los emprendedores y las estrategias no adecuadas dan como resultado que los negocios no sean viables y rentables, aspecto que genera un reto para el país lo cual conlleva a buscar correctas estrategias que produzcan un emprendimiento de alta calidad”.

En 2017 Ecuador mantiene la TEA (Tasa de Actividad Emprendedora) más alta entre los países de América Latina y el Caribe, siempre por encima de la media regional y de las economías de eficiencia. Sin embargo, la TEA Ecuador ha venido declinando gradualmente de 36% en 2013 hasta 29.6% en 2017. (Laso et al, GLOBAL ENTREPRENEURSHIP MONITOR ECUADOR 2017, 2017, pág. 25)

En la investigación desarrollada por (Cabrera & Reyes, 2017, pág. 12) nos da un indicio del uso de herramientas y diferentes usos de la información generada en redes sociales:

“Muchas de las PYMES en Guayaquil, no cuentan, desconocen o se niegan a la utilización de herramientas que les permitan conocer los gustos y preferencias de los usuarios, mucha de esa importante información es expresada mediante las redes sociales”.

Las redes sociales son medios valiosos de comunicación, en las que se genera información de diferente índole, como información personal, gustos, pensamientos y sentimientos de sus usuarios, etc. Esto genera un amplio bagaje de información que puede ser utilizada para realizar diferentes estudios.

La red social Twitter, es lo más parecido a un medio de comunicación dado que gran parte de la información generada en su red social es pública. Debido a que esta red social pone a disposición los datos públicos de sus usuarios,

con ciertas restricciones a través de su API, la realización de este proyecto se ha inclinado por su uso, en comparación con otras redes sociales, las cuales poseen niveles de privacidad más elevados, lo cual limitaría en gran medida la puesta en marcha del presente proyecto. Además que su limitación a 280 caracteres en cada publicación la hace perfecta para el análisis de datos que se desea realizar.

Por lo mencionado previamente, en el presente proyecto se utilizará herramientas tecnológicas para desarrollar un código fuente que brinde apoyo en la toma de decisión a los emprendedores, en la difícil tarea de determinar si existe una audiencia para el producto o servicio que se pretende poner a disposición. Todo esto previo o posterior al haber iniciado sus actividades.

A partir de la extracción de datos públicos generados en la red social Twitter mediante su API y el entorno de trabajo interactivo Jupyter Notebook, en conjunto con librerías para el tratamiento de datos, se procederá a la búsqueda de un producto o servicio en el conjunto de datos extraídos, para su posterior análisis de datos que permita determinar mediante el análisis de sentimiento la existencia de una audiencia, y sobre todo que sea positiva para el sector comercial.

Además, los resultados permitirán visualizar cuales son las tendencias que tienen una correlación con los tweets emitidos que poseen una valoración positiva. Generando para los emprendedores que utilicen el código fuente, información que aporte valor al momento de la toma de decisiones en los emprendimientos, en vías de minimizar el riesgo o fracaso al iniciar un negocio, sea este pequeño o mediano.

Además, reconocer las oportunidades que el mercado ofrece puede ser muy complejo si se desea hacer sin el uso de la tecnología. Pero con la realización

de este proyecto esa barrera que tienen los emprendimientos y ventaja que poseen las grandes empresas se ve reducida, otorgando una igualdad de competencia a todos los que deseen formar parte del sector comercial de la ciudad de Guayaquil.

## **METODOLOGÍA DEL PROYECTO**

### **1. Metodología de desarrollo**

Para el presente proyecto de titulación se utilizará la metodología investigación descriptiva y la metodología CRISP-DM, las cuales se considera que tienen un mayor lineamiento con los fines del presente proyecto.

**Investigación descriptiva:** Es aquella que se detiene en la conceptualización del fenómeno y a la exaltación de sus características más prominentes. La investigación descriptiva se refiere al diseño de la investigación, creación de preguntas y análisis de datos que se llevarán a cabo sobre el tema, sin centrarse en las razones por las que se produce un determinado fenómeno. Para hacer frente al problema, la herramienta a utilizar es la encuesta, permitiendo plantear diferentes preguntas, con el objetivo de identificar datos relevantes del problema y plantear una solución a esta.

**CRISP-DM (Cross Industry Standard Process for Data Mining):** proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software. El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas.

## **2. Supuestos y restricciones**

### **Supuestos**

Dentro de los supuestos están todos los factores que proporcionarán que el proyecto se logre concretar en las fechas y plazos establecidos.

- La programación del código se realizó en el lenguaje Python, en la plataforma Jupyter Notebook.
- Uso de las librerías: pandas, numpy, matplotlib.
- Utilización de archivos en formato csv.
- Uso de algoritmos que en base al aprendizaje permitan detectar patrones comerciales y ayudar en la toma decisiones.
- El resultado que proporcionarán estos algoritmos en la plataforma Jupyter Notebook están orientados a mejorar la toma de decisiones al momento de emprender en el sector comercial.

### **Restricciones**

Entre las posibles limitaciones para el desarrollo del proyecto podemos encontrar:

- Equipos con baja capacidad en memoria RAM lo cual ralentiza la ejecución de los diversos procesos en la plataforma.
- Futuros cambios en las políticas de seguridad de la red social Twitter que impida la extracción de los datos. Debido a que a partir del mes de julio del presente año se debe tener una cuenta de desarrollador validada para crear aplicaciones en la API de Twitter.
- Errores en la conexión de Internet.
- Límite de peticiones a la API; extracción de 180 tweets cada 15 minutos. En caso de anomalías la red social bloqueará la aplicación para precautelar la seguridad de los datos de sus usuarios.

### **3. Plan de calidad (Pruebas a realizar)**

Previo a la presentación final del presente proyecto se realizaron pruebas para validar el correcto funcionamiento de los algoritmos de clasificación y reconocimiento de patrones con el uso de datos públicos extraídos de la red social Twitter. Haciendo especial énfasis en el método de validación cruzada, el cual nos permite subdividir el conjunto de datos en  $n$  subconjuntos de igual dimensión, los cuales se utilizan como conjunto de entrenamiento y el restante en conjunto de prueba. Estos test se realizaron para observar el comportamiento y la funcionalidad de los algoritmos al ser sometidos a datos reales.

## **CAPÍTULO II**

### **MARCO TEÓRICO**

#### **ANTECEDENTES DEL ESTUDIO**

Según (Bertuzzi & Suarez, 2016, pág. 12) mencionan que:

“Históricamente, el manejo de la información ha sido una de las principales causas de la innovación en las TIC, principalmente por la importancia del conocimiento que se puede extraer. Hace unos años no se pensaba en conocimiento oculto dentro de la información. Sin embargo, con el arribo de las nuevas tecnologías y el desarrollo del área de informática, un nuevo sector dentro de la ciencia de la información surgió. Dicho sector denominado Minería de datos (Data Mining en inglés) ha sido el precursor de un nuevo interés por parte del área de inteligencia de negocios, relacionado con la extracción de información a partir de grandes volúmenes de datos para distintos fines”.

A nivel mundial el uso de las redes sociales se ha convertido en una actividad, en el caso de algunos usuarios casi diaria, generando grandes cantidades de datos que se encuentran alojados en data centers, esperando ser solicitados mediante peticiones realizadas por usuarios de la misma red social o por aplicaciones propias de la red social que procesan los datos de cada usuario para ofrecerle una mejor experiencia de uso.

Uno de los servicios utilizados para comunicarse por Internet es micro blog, siendo uno de sus propulsores Twitter, el cual permite a sus usuarios enviar texto plano de corta longitud, con un máximo de 280 caracteres, con el tema que el estime. Las temáticas pueden ir desde conversaciones, noticias, retweets o spam, propagandas, avisos, advertencias y otros más. Esta suerte de publicación y seguidor es lo que se denomina hoy en día una comunidad, en donde existen intercambios de opiniones respecto al tema en cuestión (Alfaro Arancibia, 2016).

Estas opiniones albergan una amplia información respecto a temáticas abordadas por sus creadores, las que podrían contemplar la sensación de bienestar respecto a un tema o el enojo por algún hecho pasado. Pero para obtener esta información se deben analizar estos datos, debido a que al ser opiniones generadas globalmente, la obtención manual de la ventaja que entregan las opiniones se hace imposible.

Los datos de Twitter son la fuente más completa de conversaciones públicas y en vivo en todo el mundo. La API Rest, Streaming y Enterprise permiten el análisis programático de los tweets hasta el primer tweet en 2006. Ya sea que esté creando una solución para marcas, una para su propio equipo o realizando una investigación, sus puntos finales permiten una amplia visión del público y los movimientos del mercado, tendencias emergentes, temas clave, noticias de última hora y mucho más (Twitter I. , 2019).

De hecho, (Asensio, 2015, pág. 10) indica: “En muchos campos, el estudio de redes sociales como herramientas de obtención de datos, ha supuesto un gran avance. Twitter ha abierto nuevas oportunidades de investigación y de negocio. Uno de los temas más interesantes es el análisis de sentimiento y de opinión, donde se obtienen si los textos de los tweets contienen un sentimiento positivo, neutro o negativo”.

La red social Twitter además de permitir la publicación de contenido y la comunicación entre sus usuarios, genera grandes cantidades de datos que pocos emprendedores o dueños de negocios saben utilizar para beneficio de su emprendimiento o negocio. Por lo tanto se considera que la red social Twitter ha abierto nuevas oportunidades de investigación y de negocio. Y con negocio hacemos hincapié en la utilidad de sus datos públicos como ayuda en la toma de decisiones, en base a los datos comerciales generados por los usuarios y existencia de audiencia para un determinado producto o servicio, en el presente caso para la ciudad de Guayaquil.

Según (Torres E. , 2017, pág. 10) manifiesta que:

“La necesidad de datos previo a la toma de decisiones es un factor considerado desde decisiones ancestrales, la diferencia radica en que en la actualidad existe gran cantidad de información, almacenada en diferentes formatos y en diversas fuentes”.

Para (Acosta & Cruz, 2017, pág. 33), hace referencia a las oportunidades de negocio, en su investigación de los factores de éxito para emprendimientos: “Una de las fuentes de desarrollo humano es aprovechar las oportunidades. En este contexto, oportunidad de negocio es el conjunto de circunstancias propicias asociadas a una idea de negocio que contenga un mercado provisor y posea clientes. Por ende la función principal de un emprendedor es detectar, reconocer y generar oportunidades y ejecutarlas”.

Las definiciones citadas previamente son la base y fundamento que dan iniciativa al planteamiento del presente trabajo de investigación, el cual hace uso de herramientas; todas opensource, que permitan extraer datos públicos que los usuarios de la red social Twitter generan al expresar sus pensamientos, sentimientos y emociones. Posteriormente estos datos serán evaluados y procesados para obtener información comercial de gran

relevancia y utilidad para muchos emprendedores que están empezando sus negocios, o para aquellos que ya han iniciado y quieren darle un giro comercial a su negocio.

En la actualidad existen varias herramientas que hacen del análisis de datos algo menos complejo de lo que se cree. La complejidad es el motivo por el cual varios emprendedores no hacen uso e integración de estas herramientas, por eso se ha seleccionado un lenguaje de programación ameno y con una amplia comunidad de programadores, la cual brinda respaldo cuando se tenga uno o varios errores en el código, dicho lenguaje de programación es python, el cual tiene a disposición una gran variedad de librerías para el análisis de datos, algoritmos de aprendizaje automático y generación de gráficos estadísticos lo que facilita la comprensión de los resultados obtenidos.

## FUNDAMENTACIÓN TEÓRICA

### **Las redes sociales en Ecuador**

Según (Formación Gerencial, 2018) menciona que: “Actualmente Ecuador cuenta con más de 13,6 millones de usuarios conectados y con acceso a Internet (octubre 2018), siendo el principal destino de los mismos Facebook, Youtube y Google como buscador, tres sitios que durante los últimos años se han disputan los primeros lugares entre los sitios más visitados y seguidos por diferentes categorías de plataformas de contenido, servicios e interacción, mostrando cada vez mayor nivel de consumo, creación y participación con marcas.”

Se debe tener en cuenta que la cantidad de usuarios antes mencionada, está basada en el crecimiento promedio de usuarios proyectado, que oscila en (1,5%) en base a cifras proporcionada por Arcotel, en el cual se registran un promedio de 9 millones de conexiones “móviles” y 2 millones de conexiones

tradicionales, dado que no se considera múltiples usuarios en una misma conexión, por lo que junto a las cifras de fuentes oficiales de redes sociales, se mantiene la proyección de 13,6 millones de usuarios de Internet promedio expresada en este estudio.

**Gráfico N. 1** Perfil usuarios de Twitter por edades en Ecuador



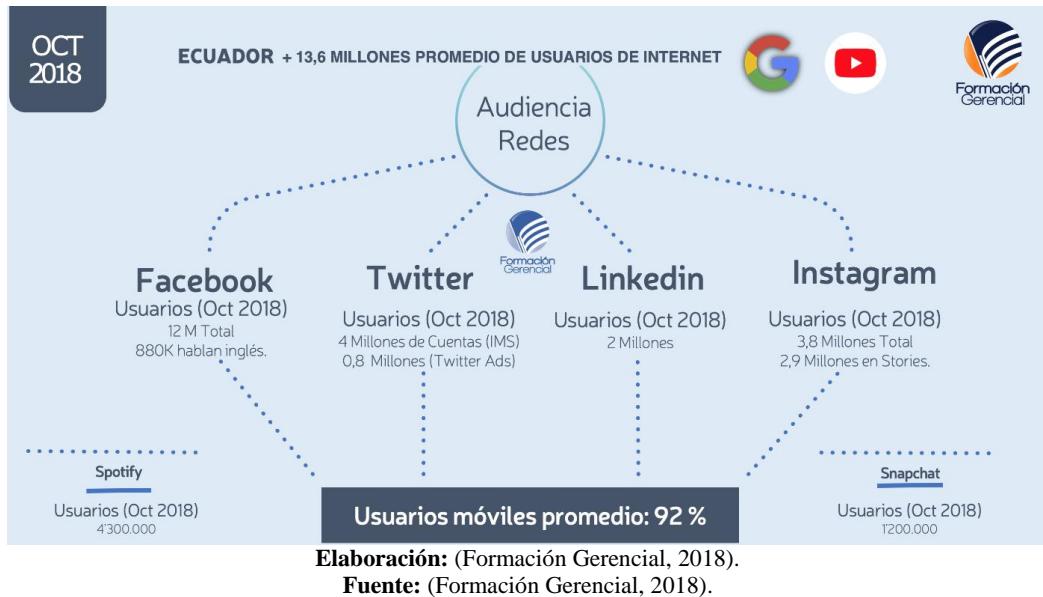
**Elaboración:** (Formación Gerencial, 2018).

**Fuente:** (Formación Gerencial, 2018).

Para el presente proyecto se excluye el perfil de usuarios, por motivos de no incurrir en inconvenientes en el ámbito legal.

Dado la relevancia e importancia en la vida cotidiana que han adquirido nuevas plataformas y redes sociales ya consolidadas, se puede llegar al pensamiento que la red social Twitter ha ido en declive y ya no posee una gran cantidad de adeptos, pero como se refleja en el gráfico N. 2, esta red social sigue tan vigente como las demás mencionadas.

**Gráfico N. 2** Audiencia de redes en Ecuador.



### Red Social Twitter

Twitter fue fundada en marzo de 2006 por los estudiantes de la Universidad de Cornell, en Nueva York. Jack Dorsey, Biz Stone, Evan Williams y Noah Glass quienes fundaron la compañía Obvious que posteriormente pasó a ser Twitter Inc. (Castillo, 2017, pág. 5). Esta compañía fue creada poco después de Facebook, en el año 2006. La idea surgió cuando Dorsey vio por primera vez la mensajería instantánea en marcha y se preguntó si el rendimiento del software del usuario podría ser compartido entre amigos fácilmente.

Twitter es una aplicación web gratuita de microblogging que reúne las ventajas de un blog, redes sociales y la mensajería instantánea. Es una forma de comunicación que permite al usuario estar en contacto en tiempo real con personas de su interés a través de mensajes, publicaciones de tipo textos y contenido, de no más de 280 caracteres los cuales se denominan tweets. Su funcionamiento es similar a cualquier otra red social, en Twitter los usuarios envían y reciben tweets vía web, smartphone, mensajería instantánea o a

través de correo electrónico e incluso desde aplicaciones de terceros (Fernández, 2018).

## Términos de Twitter

A continuación se presenta un breve glosario de algunas de los términos más utilizadas dentro de la red social Twitter (2018):

- **Tweet:** Mensaje o publicación en Twitter que puede contener fotos, GIF, videos y texto.
- **Hashtag:** Se cataloga como un hashtag a cualquier palabra o frase precedida directamente por el símbolo #.
- **Following:** Todos las cuentas o usuarios que un determinado usuario sigue, con el objetivo de informarse sobre sus publicaciones.
- **Timeline:** Lista de publicaciones enviadas por las cuentas que cada usuario sigue, las cuales aparecerán ordenadas de forma cronológica.
- **Retwittear:** La acción de compartir el tweet de otra cuenta con todos tus seguidores citando el autor.

## Información que se puede extraer de Twitter

Las API's provistas por Twitter permite extraer diversos datos de cada publicación, a continuación se hace una pequeña descripción de los datos que se extraen de cada tweet.

- **¿What? (¿Qué?):** El contenido en sí de cada publicación. Puede contener además de texto, imágenes, videos, links o emoticonos. Pero para nuestros fines, solo se hará énfasis en el texto.
- **¿Who? (¿Quién?):** Usuario o cuenta que ha escrito el mensaje o bien a retuiteado una publicación. Esta información contiene nombre completo y lenguaje.

- **¿When? (¿Cuándo?):** Fecha y hora de la publicación.
- **¿Where? (¿Dónde?):** Esta información no aparece en todas las publicaciones, es opcional. Dado a que la red social Twitter tiene opciones de privacidad, cada usuario decide si desea mostrar su ubicación o no. En caso de mostrar la ubicación podemos obtener las coordenadas geográficas desde dónde se ha publicado el tweet.

Para el presente proyecto, el cual está delimitado a la ciudad de Guayaquil es de vital importancia que el principal filtro a la hora de seleccionar los tweets es que contengan coordenadas geográficas de esta ciudad.

### **API Twitter**

Una API (Application Programming Interface) está conformada por un conjunto de funciones y procedimientos que cumplen una o varias funciones con la finalidad de ser utilizadas por otro software. Su principal uso está en implementar funciones que engloba un servicio sin la necesidad de programar de nuevo (San Martín Duchen, 2017, pág. 9).

Entre los grandes motivos de la expansión de la red social Twitter consta la existencia de API's gratuitas que proporciona la empresa, las cuales han propiciado la creación de software de terceros, debido a que por medio de la API se pueden conectar y manejar datos de la aplicación (Twitter, Centro de ayuda, 2018). Se debe tener en cuenta que el API de Twitter posee limitantes en cuanto al acceso a la aplicación, con un total de 150 a 350 solicitudes por hora dependiendo si se tiene registrado la aplicación en el apartado de desarrolladores de Twitter. Esta red social utiliza en su API el protocolo abierto OAuth, el cual permite acceso seguro a las API's.

Twitter ofrece tres APIs: Streaming API, REST API y SEARCH API cada una aplicable a diferentes necesidades (Twitter developer documentation, 2018).

1. **Streaming API:** proporciona subset de tweets en casi tiempo real estableciendo una conexión permanente por usuario con los servidores de Twitter y mediante una petición http se recibe un flujo continuo de tweets en formato json. Esta petición de tweets puede ser una muestra aleatoria o se pueden filtrar por usuarios o palabras claves.
2. **Search API:** Suministra tweets con información más limitada a diferencia de las otras API y con una profundidad en el tiempo de 7 días que se ajustan a la solicitud realizada. También es posible filtrar por, cliente, lenguaje y palabras específicas; los datos se obtienen en formato json.
3. **Rest API:** Es una API web que funciona por HTTP a la cual accedemos a partir de URLs que devuelven contenido en formato JSON, XML, HTML, etc. A diferencia de Search API, no hay limitación temporal, pero sí una limitación del número de resultados devueltos establecido en 3.200 tweets.

### Filtrar tweets en tiempo real

El proceso de filtrar tweets en tiempo real, devuelve tweets públicos que coinciden con uno o más parámetros declarados en el filtro. También permite especificar varios parámetros, lo que permite a la mayoría de los clientes usar una sola conexión a la API de transmisión. Se admiten solicitudes GET y POST, pero las solicitudes GET no deberán contener demasiados parámetros, debido a que la solicitud será rechazada por una longitud excesiva de URL. En aquellos casos la documentación oficial del API de Twitter estipula utilizar una solicitud POST para evitar URL largas (Twitter developer documentation, 2018).

Los campos de seguimiento, seguimiento y ubicaciones deben considerarse como combinados con un operador OR. *track = foo & follow = 1234* devuelve tweets que coinciden con "foo" o creado por el usuario 1234 (Twitter I. , 2019). Debido a que el presente proyecto hace uso de cuadros de ubicación con coordenadas de la ciudad de Guayaquil, no se procederá a usar el campo track para especificar una o varias palabras de búsqueda, evitando obtener redundancia en la búsqueda de tweets y porque la documentación Oficial del API de Twitter lo recomienda.

El nivel de acceso predeterminado permite hasta 400 palabras claves de pista, 5,000 usuarios de seguimiento y 25 cuadros de ubicación de 0.1-360 grados. Para un acceso elevado a la API de transmisión, se deberá utilizar la API de transmisión empresarial PowerTrack (Developer, 2019).

La plataforma API de Twitter ofrece dos opciones para transmitir tweets en tiempo real. Cada opción permite un número variable de filtros y capacidades de filtrado; a continuación se presenta un resumen con detalles:

*Cuadro N. 3 Opciones para transmitir tweets en tiempo real.*

API	Categoría	Número de filtros	Operadores de filtrado	Gestión de reglas
<b>Statuses/filter</b>	Estándar	400 palabras clave, 5,000 usuarios y 25 cajas de ubicación.	<b>Operadores estándar</b>	Una regla de filtro en una conexión permitida, desconexión requerida para ajustar la regla.
<b>PowerTrack</b>	Empresa	Hasta 250,000 filtros por flujo, hasta 2,048 caracteres cada uno.	<b>Operadores premium</b>	Miles de reglas en una sola conexión, no se necesita desconexión para agregar / eliminar reglas usando las API de reglas.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Developer, 2019).

Para evitar incurrir en gastos, el API a utilizar es la denominada API statuses/filter (estados/filtro). Posee características suficientes para el correcto despliegue del proyecto, incluyendo el uso de cajas de ubicación o delimitadoras.

### **POST estados / filtro**

Las solicitudes post devuelven un formato Json de respuesta, el cual dependiendo de la librería que se utilice puede ser limitado, o con todos los campos. Para su funcionamiento se requiere autenticación, y la tasa de respuesta es limitada (Twitter, 2018).

Los parámetros que se pueden utilizar en la solicitud estados / filtro son:

**Cuadro N. 4** Parámetros para filtrar tweets.

Nombre	Necesario	Descripción
<b>Follow</b>	Opcional	Una lista separada por comas de identificadores de usuario, indicando los usuarios que deben devolver los estados de la secuencia.
<b>Track</b>	Opcional	Palabras clave para seguir. Las frases de palabras clave se especifican mediante una lista separada por comas.
<b>Locations</b>	Opcional	Especifica un conjunto de cuadros delimitadores para realizar el seguimiento.
<b>Delimited</b>	Opcional	Especifica si los mensajes deben ser delimitados por longitud.
<b>Stall_warnings</b>	Opcional	Especifica si se deben entregar las advertencias de bloqueo.

Elaboración: Jiménez Cárdenas Edinson Andrés.

Fuente: (Developer, 2019).

## Datos geolocalizados

Según (Asensio, 2015, pág. 20), se dispone de tres maneras diferentes para clasificar los tweets geográficamente, denotando la siguiente clasificación:

1. **Lugar del mensaje:** tweets que están marcados con la localización exacta. Puede ser la localización exacta o el 'Twitter Place'. Localización exacta con coordenadas latitud/longitud: Por ejemplo: -85.7629, 38.2267. El Twitter Place es un metadato que marca un lugar, Por ejemplo: "Louisville Central", y cuatro pares de coordenadas latitud/longitud para definir el área indicada.
2. **Lugar del perfil:** Uno de los datos que puede indicar el usuario es la localidad donde vive, esta aparece en su perfil de manera pública.

3. **Localización mencionada en el mensaje:** Mención de un lugar en el texto del tweet. Por ejemplo: “La lluvia en Sevilla es una maravilla”.

Dada su exactitud, obtener los tweets del primer caso será prioritario, sin embargo sólo un 2% de los tweets publicados globalmente poseen este dato. Dado estos sucesos la extracción de datos se basará en la función de caja delimitadora, la cual solo extrae tweets que sean emitidos entre las coordenadas indicadas y concuerden con la etiqueta ‘Twitter Place’. Se descarta rotundamente almacenar datos sobre la localización mencionada en el texto, pues no es un indicador específico de que el usuario escribe desde ese lugar.

## Localizaciones

Una lista de pares de longitud y latitud separados por comas que especifican un conjunto de cuadros delimitadores para filtrar los tweets. Sólo se incluirán los tweets geolocalizados que se encuentren dentro de las casillas de delimitación solicitadas; a diferencia de la API de búsqueda, el campo de ubicación del usuario no se utiliza para filtrar los tweets.

Cada cuadro delimitador debe especificarse como un par de pares de longitudes y latitudes, siendo la esquina suroeste del cuadro delimitador la primera. Por ejemplo:

*Cuadro N. 5 Ejemplo de caja delimitadora*

<b>Valor del parámetro</b>	<b>Pistas tweets de ciudades</b>
-122.75,36.8, -121.75,37.8	San Francisco
-74,40, -73,41	Nueva York
-122.75,36.8, -121.75,37.8, - 74,40, -73,41	San Francisco O Nueva York

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Developer, 2019).

Las cajas delimitadoras no actúan como filtros para otros parámetros de filtro. Por ejemplo, (**track=Twitter & locations = -122.75, 36.8, -121.75, 37.8**). Esto coincidiría con cualquier Tweet que contenga el término Twitter (incluso Tweets no geográficos) O que provengan del área de San Francisco. Según la documentación oficial de Twitter (Twitter developer documentation, 2018), la API de transmisión utiliza la siguiente heurística para determinar si un Tweet determinado se encuentra dentro de una caja delimitadora:

- Si el campo de coordenadas está lleno, los valores se probarán en el cuadro delimitador. Tenga en cuenta que este campo utiliza el orden geoJSON (longitud, latitud).
- Si las coordenadas están vacías pero el lugar está poblado, la región definida en el lugar se comprueba para detectar la intersección con el cuadro delimitador de ubicaciones. Cualquier superposición coincidirá.
- Si ninguna de las reglas enumeradas anteriormente coincide, el Tweet no coincide con la consulta de ubicación. Tenga en cuenta que el campo geo está en desuso e ignorado por la API de transmisión.

## Formato de salida JSON

El formato para representar la información en Twitter es JSON (JavaScript Object Notation), es un formato sencillo orientado para el intercambio de datos. Una de las ventajas de JSON como formato de intercambio de datos es su simplicidad para escribir un analizador sintáctico, también llamado parser de JSON (Asensio, 2015, pág. 16).

Según la página oficial de JSON (JSON, s.f.), su formato está constituido por dos estructuras:

- Una colección de pares de nombre/valor. En varios lenguajes esto es conocido como un objeto, registro, estructura, diccionario, tabla hash, lista de claves o un arreglo asociativo.
- Una lista ordenada de valores. En la mayoría de los lenguajes, esto se implementa como arreglos, vectores, listas o secuencias.

Estas son estructuras universales; virtualmente todos los lenguajes de programación las soportan de una forma u otra. Es razonable que un formato de intercambio de datos que es independiente del lenguaje de programación se base en estas estructuras.

*Gráfico N. 3 Formato JSON devuelto como petición al API de Twitter.*

```
1 Status(_api=<tweepy.api.API object at 0x0000029C42135DD8>,
2     _json={'created_at': 'Thu Feb 07 22:12:16 +0000 2019',
3             'id': 1093633571506606080,
4             'id_str': '1093633571506606080',
5             'text': '#listas60traVez . De esto se trata. Hay que defender todo lo que se ha hecho!',
6             'display_text_range': [0, 140],
7             'source': '<a href="https://www.hootsuite.com" rel="nofollow">Hootsuite Inc.</a>',
8             'truncated': True,
9             'in_reply_to_status_id': None,
10            'in_reply_to_status_id_str': None,
11            'in_reply_to_user_id': None,
12            'in_reply_to_user_id_str': None,
13            'in_reply_to_screen_name': None,
14            'user': {'id': 588928263,
15                  'id_str': '588928263',
16                  'name': 'Gregorio Álvarez®',
17                  'location': 'Guayaquil Ecuador',
18                  (...)}
```

**Nota:** En el gráfico se puede observar una pequeña parte del archivo JSON, el cual es más extenso y completo si se utiliza la librería Tweepy.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Desde la versión 1.1 del API de Twitter, se requiere autentificación OAuth si se desea realizar una aplicación propia.

### **Limitaciones de Twitter en su API**

Al momento de trabajar con la API de Twitter es necesario tener en cuenta ciertas limitaciones que implica trabajar con ella. Los límites de frecuencia de la interfaz solo permiten realizar hasta 450 peticiones cada 15 minutos. También, se debe tener en cuenta que Twitter filtra gran cantidad de los tweets publicados para que los resultados en las búsquedas sean de mayor calidad, por lo que no está disponible la totalidad de publicaciones que se realizaron sino aquellas que Twitter considera que son más relevantes para el usuario (Martín, 2016, pág. 18).

### **Procesamiento del Lenguaje Natural (NLP)**

El Procesamiento del Lenguaje Natural (NLP por sus siglas en inglés) es el campo de estudio que se enfoca en la comprensión mediante ordenador del lenguaje humano. Abarca parte de la ciencia de datos, Inteligencia Artificial (aprendizaje automático) y la lingüística (González D. , 2017, pág. 35).

En NLP las computadoras analizan el leguaje humano, lo interpretan y dan significado para que pueda ser utilizado de manera práctica. Usando NLP se pueden realizar tareas como resumen automático de textos, traducción de idiomas, extracción de relaciones, análisis de sentimiento, reconocimiento del habla y clasificación de artículos por temáticas (machine learning, 2018).

Se debe recalcar que este campo está muy avanzado en el idioma inglés, en el cual existe una amplia variedad de librerías disponibles para su rápida aplicación a los datos, en el idioma español, las funciones de las librerías de

NLP son reducidas y en algunos casos solo están disponible para el idioma inglés.

### **Análisis de sentimientos**

El análisis de sentimientos trata de juzgar el sentimiento detrás de un escrito. El proceso implica tomar un texto, en nuestro caso un tweet, y analizar la emoción que un usuario está expresando. En el nivel más básico, una herramienta de análisis de opiniones clasificará fragmentos de texto como positivos, negativos o neutrales (Gómez E. , 2018).

Adicional al análisis de temáticas es beneficioso, el poder determinar en qué sentido se realizó cada publicación. Es decir, la posibilidad de determinar si dicha publicación dijo algo positivo, negativo o neutro acerca de un tema particular. Esto aumentaría en gran medida la información que pueda aportar la aplicación ya que no solo presenta tendencias temáticas sino que le adicionará que sentimiento fue preponderante al nombrar dicho tema (Bertuzzi & Suarez, 2016).

La clasificación de la polaridad del sentimiento se trata de determinar si un tweet contiene opinión o no y, de ser así, si esta es positiva, negativa o neutra. El análisis del sentimiento es un proceso en el que se utilizan algoritmos para determinar las emociones positivas o negativas que tienen las personas de una red respecto a un tema, producto, noticia, etc. (Rosas, 2019).

Adicionalmente, esta técnica puede ser aplicada gracias a las capacidades y posibilidades de conexión con datos a analizar de medios sociales como Twitter.

Sin embargo, la expresión humana rara vez es tan sencilla de comprender para un algoritmo. Debido a que cuando hablamos se transmite una amplia gama de emociones que a veces requieren un contexto para ser comprendido

en su totalidad. Todo esto puede suceder dentro de una sola oración (Pérez, 2019).

El análisis supervisado de sentimientos puede además ser ejecutado utilizando recursos libremente disponibles como Python (versiones 2.7 y 3.7) y la interfaz de programa de aplicación (API) de Twitter (REST y Streaming), (Arcila & et al, 2017):

- **API Rest:** permite descargar y filtrar el histórico de mensajes de los últimos 7 días, con lo cual se pueden recolectar mensajes políticos para poderlos clasificar manualmente y que alimenten el modelo.
- **API Streaming:** se puede realizar la conexión al flujo constante de Twitter en tiempo real (limitado al 1% de todos los mensajes producidos en ese momento). Todos los mensajes de Twitter se obtienen de forma semi-estructurada en formato JSON, lo que permite ejecutar filtros sobre las consultas, por ejemplo, de fechas, idiomas, lugares geográficos o etiquetas incluidas en el texto a analizar del mensaje.

Según la investigación de (Pérez, 2019), establece que un algoritmo de análisis de sentimientos basado en datos de entrenamiento de alta calidad debería poder clasificar todo tipo de textos. Esto lo hace comparando partes del texto con ejemplos de sus datos de entrenamiento y su experiencia previa con casos similares.

Existen diferentes maneras en que esta tarea se lleva a cabo dentro de un sistema de análisis de sentimientos. A continuación se describen tres formas:

- **Enfoques basados en reglas:** están basados en reglas definidas manualmente en un script que incorpora técnicas de PNL, como la derivación o tokenización.

- **Los enfoques automáticos:** están basados en técnicas de aprendizaje automático y enmarcan la tarea como un problema de clasificación que debe resolverse mediante redes neuronales, regresión logística u otros modelos estadísticos.
- **Sistemas híbridos:** combinan elementos de ambos enfoques.

Estos tipos de algoritmos tienen sus propias ventajas y desventajas, todos tienen una variedad de aplicaciones potenciales que permiten obtener información que brinda apoyo en la toma de decisiones.

En el presente proyecto se realizó un análisis sobre la polaridad de los tweets. El análisis en español es realmente complejo hoy en día, pues no existen más que conjuntos de palabras clasificadas según el sentimiento o conjuntos de tweets clasificados específicamente para el ámbito político.

Tras un exhaustivo análisis de las librerías disponibles para el lenguaje de programación Python para el análisis de sentimientos, no se ha encontrado ninguna que tenga una función de análisis en de tweets en el idioma español.

Conocer la percepción sobre productos, servicios, eventos o personalidades relevantes, así como monitorizar su reputación online son algunos de los objetivos que las compañías se han marcado a corto plazo. Uno de los primeros problemas a los que se enfrentan estas empresas es discriminar los mensajes pertenecientes a su ámbito de negocio en un medio tan ruidoso como Twitter, donde es posible encontrar opiniones sobre prácticamente cualquier tema. Respecto a esto, las funcionalidades de búsqueda de Twitter se limitan a sencillas funciones como búsqueda por palabras clave, capacidades de búsqueda por idioma o recuperación de los tweets de un determinado autor (Vilares, 2014, pág. 7).

La realización del presente trabajo aborda la extracción de tweets geolocalizados y técnicas que permitan extraer tweets sobre temas de interés para los emprendedores de la ciudad de Guayaquil, que les brinde apoyo en la toma de decisiones e información en relación a la existencia de una audiencia o mercado para un determinado producto o servicio.

Su importancia está en que nuestra percepción de la realidad, y así también las decisiones que tomamos, es condicionada en cierta forma por cómo otras personas ven y perciben el mundo. Es por esto que desde un punto de vista de utilidad, se desea conocer la opinión de otras personas sobre cualquier tema de interés, ya que tienen diversas aplicaciones como recomendar productos y servicios, determinar a qué candidato político se votara en las próximas elecciones o incluso medir la opinión pública ante la medida tomada por una empresa o el gobierno (Martín, 2016, pág. 27).

### **Minería de datos (Data Mining)**

El Data Mining se define como un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos (Sinnexus, 2017).

Según Ribas (2018) establece que:

“La principal finalidad de la minería de datos es explorar, mediante la utilización de distintas técnicas y tecnologías, bases de datos enormes de manera automática con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos que se han ido recopilando con el tiempo. Estos patrones pueden encontrarse utilizando

estadísticas o algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales”.

Dado a que el mayor volumen de datos que circula en internet se genera en las redes sociales, es fundamental para los emprendedores prestar atención a los medios sociales donde se interrelacionan y comparten contenido la mayor parte del mundo, sobre todo un medio social que sea muy utilizado en su localidad, porque es ahí donde se puede encontrar datos no estructurados que al ser procesados se conviertan en una valiosa fuente de información.

En la investigación realizada por (Molina, 2016) se afirma que existen cuatro tipos de tareas que normalmente se involucran en la minería de datos:

- **Clasificación:** la tarea de generalizar una estructura familiar para utilizarla en los nuevos datos.
- **Agrupamiento:** la tarea de encontrar grupos y estructuras en los datos que son de alguna manera u otra lo mismo, sin necesidad de utilizar las estructuras observadas en los datos.
- **Aprendizaje de reglas de asociación:** busca relaciones entre las variables.
- **Regresión:** su objetivo es encontrar una función que modele los datos con el menor error.

Entre las principales características del presente trabajo se destaca el uso de técnicas de minería de datos para optimizar los procesos de buscar patrones de consumo y tendencias a través del análisis de datos, datos que serán extraídos de la red social Twitter. Todo esto en combinación con algoritmos de aprendizaje automático que generen información en relación a la búsqueda de un producto o servicio y mostrar tendencias relacionadas con estos datos. Y mediante el uso de una red neuronal clasificar las opiniones que tengan una connotación con un sector o industria, permitiendo a los emprendedores visualizar que sectores e industrias son las que contienen mayor interacción por parte de los usuarios.

## **Extracción de datos**

La extracción de datos se estipula como un proceso, el cual nos permite obtener datos de uno o varios temas específicos, los cuales pasaran por varios procesos antes de ser considerados información (Astera, 2019).

En el ámbito de los negocios, se emplea la extracción de datos para obtener información que permita a la directiva o a distintos departamentos tomar decisiones acerca del rumbo de una organización. Todo esto en vías del crecimiento empresarial. Un ejemplo muy común es el llevado a cabo en los departamentos de marketing, en el cual se extraen los datos de clientes basado en sus gustos para ofrecerle productos de símil características, o en compras realizadas por otros clientes con unas características similares, lo cual se denomina sistemas de recomendación, algo muy utilizado en las plataformas de compra online (Recalde, 2018).

Para el desarrollo del presente proyecto se ha utilizado la API de la red social Twitter, la cual nos provee acceso a los datos públicos generados por sus usuarios, con unas limitantes las cuales se consideran adecuadas para proteger la reputación de sus servicios web (Developer, 2019).

Entre las acciones a seguir nos encontramos con realizar peticiones a los servidores de la red social Twitter por medio de su API, extraer datos que cumplan con las coordenadas de una caja delimitadora que contiene la ubicación de la ciudad de Guayaquil y el modelo de “Bolsa de palabras”, que permite extraer los tweets que contengan palabras relacionadas al ámbito comercial, para posteriormente clasificarlos por distintos sectores e industrias. Todo esto en vías de dotar de información a los emprendedores de esta ciudad.

## **Inteligencia artificial (IA)**

El término inteligencia artificial representa un conjunto de disciplinas de software, lógica, informática y filosofía que están destinadas a hacer que las PC realicen funciones que se pensaba que eran exclusivamente humanas, como percibir el significado en el lenguaje escrito o hablado, aprender, reconocer expresiones faciales, etc. El campo de la inteligencia artificial tiene una larga historia tras de sí, con muchos avances anteriores, como el reconocimiento de caracteres ópticos, que en la actualidad se consideran como algo cotidiano (Packard, 2018).

Según (Torra, 2015), J. McCarthy define el problema de la inteligencia artificial como el de construir una máquina que se comporte de tal manera que si el mismo comportamiento lo realizara un ser humano, este sería llamado inteligente.

La IA en función de sus objetivos finales de investigación, se puede clasificar en:

- **Inteligencia artificial débil:** Se considera que los ordenadores únicamente pueden simular que razonan, y únicamente pueden actuar de forma inteligente.
- **Inteligencia artificial fuerte:** Se considera que un ordenador puede tener una mente y unos estados mentales, y que, por lo tanto, un día será posible construir uno con todas las capacidades de la mente humana. Este ordenador será capaz de razonar, imaginar, etc.

En la actualidad, en el ámbito de la extracción y análisis de grandes volúmenes de datos se utiliza la Inteligencia Artificial (IA), cuyo objetivo en los ámbitos descritos anteriormente es el tratamiento de datos de forma masiva y automática, que pudieran contener un alto grado de complejidad (Rouse, 2017).

Los datos generados por los usuarios en internet mediante sus interacciones y comportamiento, pueden convertirse en una fuente de información, si se analizan sus quejas, sentimientos, preferencias. Con la finalidad de crear servicios personalizados o tomar decisiones empresariales, mismas que están basados en datos obtenidos mediante el uso de IA (Costa, 2015).

Con las herramientas y recursos tecnológicos disponibles, y cada vez más asequible para aplicar la inteligencia artificial y la ciencia de datos a las decisiones empresariales, se pretende ayudar a los emprendedores, en la gran tarea de identificar qué sectores e industrias son las que poseen mayor interactividad y gran connotación positiva en los comentarios, el cual demuestre un índice de estabilidad, basado en el análisis de tweets pertenecientes a un sector e industria, lo cual será posible gracias a diferentes algoritmos de aprendizaje automático (Machine Learning).

### **Aprendizaje automático (Machine learning)**

El aprendizaje automático es un tipo de inteligencia artificial (AI) que proporciona a las computadoras la capacidad de aprender, sin ser programadas explícitamente. El aprendizaje automático se centra en el desarrollo de programas informáticos que pueden cambiar cuando se exponen a nuevos datos (Rouse, TechTarget, 2017).

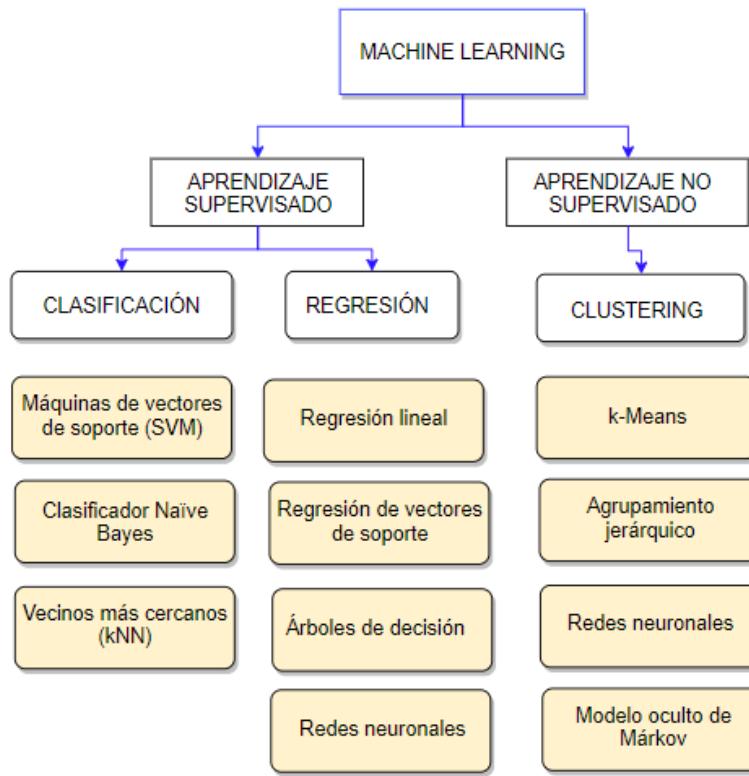
El proceso de aprendizaje automático tiene una similitud con el de minería de datos, dado a que en ambos procesos se busca encontrar patrones entre los datos. Pero, para una mayor comprensión, es fundamental describir que la minería de datos permite extraer los datos para la comprensión humana, y el aprendizaje automático usa los datos para detectar patrones en los datos, lo cual le permite ajustar las acciones del programa o algoritmo en consecuencia a los resultados obtenidos, tratando de minimizar el error y aumentar su efectividad en cada iteración (machine learning, 2018).

Los algoritmos de aprendizaje automático se dividen en:

- **Algoritmos supervisados:** Se utilizan cuando se posee un dataset que contiene información histórica relacionada con el tema a predecir o clasificar, esto permite al algoritmo aprender en la fase de entrenamiento con datos de entrada y salida conocidos, y luego obtener mejores resultados en la fase de pruebas, dado a que ya tuvo interactividad con datos similares y así predecir salidas futuras.
- **Algoritmos no supervisados:** Se diferencia de los algoritmos supervisados en que no se cuenta con información histórica, esto lleva a que el algoritmo descubra de forma autónoma en los datos de entrada: características, correlaciones, categorías, patrones ocultos o estructuras intrínsecas. Por lo tanto requiere menos tiempo de entrenamiento.

Dentro del aprendizaje automático existen varios algoritmos que sirven para cubrir diferentes tipos de aplicaciones, las cuales se verán aclaradas en el gráfico N. 4.

**Gráfico N. 4** Técnicas de aprendizaje automático.



**Nota:** En el gráfico podemos observar cómo se descompone el aprendizaje automático, hasta llegar a sus diferentes algoritmos.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (MathWorks I. , 2018).

El proceso de clasificación depende particularmente de nuestros datos, no se puede predefinir una forma de clasificación genérica qué sea la más óptima para la totalidad de los casos. Por lo tanto dependiendo de nuestro dataset y tipo de clasificación, se deberá utilizar el más apropiado

El aprendizaje automático es actualmente una de las técnicas de análisis de datos más importantes de la inteligencia artificial, la cual enseña a los ordenadores a realizar tareas o procesos que suelen parecer natural para las personas: como lo es aprender en base a la experiencia. Los algoritmos mejoran su rendimiento a medida que aumenta el número de muestras

disponibles para el aprendizaje, lo que se denomina forma o estrategia adaptativa (Sas, 2019).

### **Aprendizaje supervisado**

El aprendizaje supervisado se realiza mediante la recepción por parte del algoritmo de un grupo de ejemplos clasificados llamado corpus que contiene datos de entrenamiento, que consisten en pares de objetos de entrada (típicamente vectores) y los datos de salida deseados. Las salidas de la función corresponden a las predicciones realizadas con el conocimiento de los datos de entrenamiento (Rivera & Villavicencio, 2017).

Este proceso implica entrena al algoritmo otorgándole las preguntas, denominadas características, y las respuestas, denominadas etiquetas. Esto se realiza con la finalidad de que el algoritmo las combine y pueda hacer predicciones (Zambrano, 2018).

El objetivo de este tipo de aprendizaje es encontrar el valor de la función para cualquier entrada válida luego de entrenarse con los datos de ejemplo. Los usos comunes del aprendizaje supervisado están asociados a los problemas de regresión y clasificación; un ejemplo de esto es su aplicación para la segmentación de mercado.

### **Clasificador de Naive Bayes**

Naive Bayes es un método de clasificación supervisado y generativo, se basa en el teorema de Bayes y en la premisa de independencia de los atributos dada una clase. Esta premisa es conocida como “Naive Assumption” y se le llama “Naive” (o ingenua) considerando que en la práctica los atributos raramente son independientes, lo cual en la mayoría de los casos, no afecta los buenos resultados del método (Dubiau, 2014, pág. 10).

Se basan en la teoría probabilística, en especial en el teorema de Bayes, el cual permite estimar la probabilidad de un suceso a partir de la probabilidad de que ocurra otro suceso, del cual depende el primero. Este algoritmo estima la probabilidad de que un documento pertenezca a una categoría.

Cuando el corpus de entrenamiento es pequeño, pueden producirse errores al estimar probabilidades. Por ejemplo, cuando un determinado término no aparece en la etapa de entrenamiento pero aparece en la etapa de pruebas. Esto implica la necesidad de aplicar técnicas de suavizado, a fin de evitar distorsiones en la obtención de las probabilidades.

Según (Dubiau, 2014, pág. 10) con dichas probabilidades obtenidas en el entrenamiento, se puede estimar la probabilidad de que un nuevo documento, dado que contiene un conjunto determinado de términos, pertenezca a cada una de las categorías. La más probable, obviamente, es a la que será asignado. El clasificador Bayes Ingenuo combina el modelo de características independientes con una regla de decisión. El clasificador Bayes (la función Classify) se define como:

$$C_{NB} = \arg \max_i P(C_i) \prod_{k=1}^n P(f_k | C_i)$$

### **Máquina de vectores de soporte (SVM)**

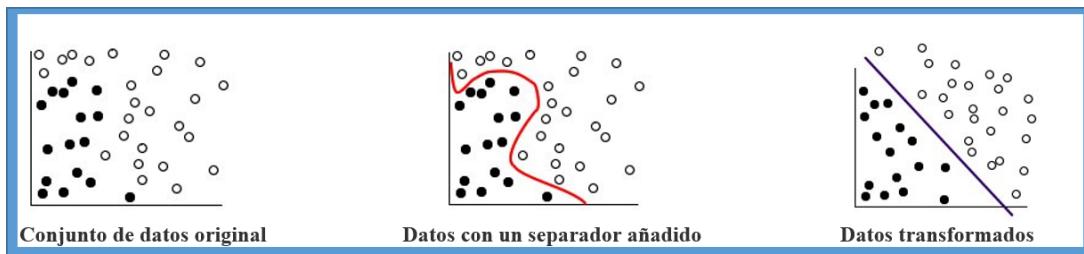
Una máquina de vectores de soporte (SVM) es un algoritmo de aprendizaje supervisado que se puede emplear para clasificación binaria o regresión. Las máquinas de vectores de soporte son muy populares en aplicaciones como el procesamiento del lenguaje natural, el habla, el reconocimiento de imágenes y la visión artificial (MathWorks, 2019).

El objetivo del algoritmo de máquina de vectores de soporte es encontrar un hiperplano en un espacio N-dimensional (N - el número de características) que clasifica claramente los puntos de datos.

SVM funciona correlacionando datos a un espacio de características de grandes dimensiones de forma que los puntos de datos se puedan categorizar, incluso si los datos no se pueden separar linealmente de otro modo. Se detecta un separador entre las categorías y los datos se transforman de forma que el separador se puede extraer como un hiperplano. Tras ello, las características de los nuevos datos se pueden utilizar para predecir el grupo al que pertenece el nuevo registro (IBM, 2019).

Por ejemplo, en el gráfico N. 5 la primera representación los puntos de datos corresponden a dos categorías diferentes. En la segunda representación las dos categorías se pueden separar con una curva. Tras la transformación, el límite entre las dos categorías se puede definir por un hiperplano, como se muestra en la última representación de la presente figura.

*Gráfico N. 5 Funcionamiento de SVM.*



Elaboración: (IBM, 2019).  
Fuente: (IBM, 2019).

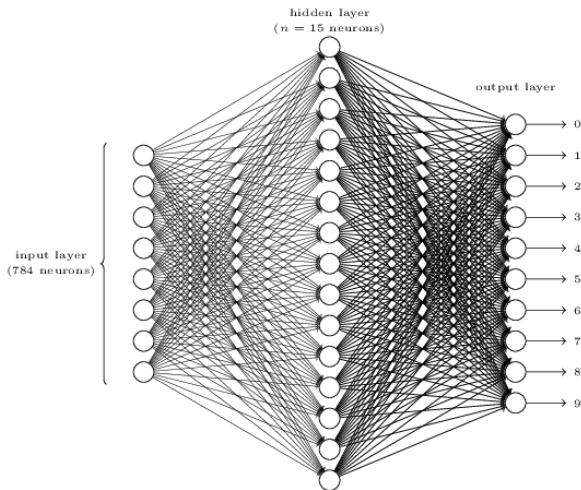
## **Redes neuronales artificiales**

Las redes neuronales artificiales (RNA) han constituido en los últimos tiempos un foco de investigación importante y con una actividad intensa, siendo un paradigma de aprendizaje computacional muy extendido en la resolución de problemas de diversas áreas de la Ingeniería y la Ciencia. Debido a sus excelentes capacidades de ajuste, las RNA se aplican de manera exitosa en distintos ámbitos científicos, sociales y tecnológicos: manufacturación, biología, finanzas, previsión del tiempo, análisis de tendencias y patrones, etc.

Entre las propiedades más destacables, la capacidad de generalización confiere a estos modelos una amplia aplicabilidad en tareas de clasificación y aproximación, entendiendo capacidad de generalización como la propiedad de la RNA para computar correctamente ejemplos de un conjunto de datos que no le han sido mostrados previamente, después de una fase de entrenamiento.

Sin embargo, existen algunas propiedades importantes de los datos que deben tener en cuenta cuando se desea aplicar algunos de estos algoritmos para predicción, y que influyen notablemente en la capacidad de generalización del modelo. Dos propiedades importantes son la calidad de los datos de entrenamiento y su complejidad. La complejidad de un conjunto de datos se puede cuantificar de muchas maneras, fundamentalmente dará una idea del grado de facilidad con el que un conjunto de datos puede ser aprendido y, en caso de las RNA, de la arquitectura y topología de la misma (González & et al, 2015).

**Gráfico N. 6** Red neuronal artificial multicapa.



**Elaboración:** (Machine learnings, 2017).

**Fuente:** (Machine learnings, 2017).

### **Patrones de consumo**

Los patrones de consumo describen la manera en que una población consume cualquier bien o servicio. Jean Baudrillard le da un mayor sentido a esto planteando que el consumo es un modo activo de relacionarse (no solo con los objetos, sino con la comunidad y con el mundo), un modo de actividad sistemática y de respuesta global en el cual se funda todo nuestro sistema cultural (Huamán, 2017).

El presente proyecto aplica la identificación de patrones de consumo para productos o servicio relacionados con sectores e industrias con actividad en la ciudad de Guayaquil, solo deberá configurarse y entrenarse correctamente la red neuronal con una base de conocimiento específico como por ejemplo un conjunto de oraciones o proposiciones que hablen sobre un solo bien/servicio en particular.

## **Aprendizaje no supervisado**

El aprendizaje no supervisado encuentra patrones ocultos o estructuras intrínsecas en los datos. Se emplea para inferir información a partir de conjuntos de datos que constan de datos de entrada sin respuestas etiquetadas (MathWorks I. , 2018).

El clustering en la actualidad es la técnica de aprendizaje no supervisado más común. Se utiliza para el análisis de datos exploratorio, con el objetivo de encontrar patrones o agrupaciones ocultos en los datos. Entre las varias aplicaciones del análisis de clusters se encuentra el análisis de secuencias genéticas, la investigación de mercados y el reconocimiento de objetos.

Algunos de los algoritmos utilizados habitualmente para realizar el clustering son: k-means y k-medoids, clustering jerárquico, modelos de mezclas gaussianas, modelos de Markov ocultos, mapas autoorganizados, clustering difuso de c-means y clustering sustractivo (Sancho, 2018).

Pero con base en los ejemplos expuesto por desarrolladores de Twitter se descarta el uso de algoritmos no supervisados para la clasificación de tweets que se realizara en el presente proyecto.

## **Clustering**

La técnica clustering o algoritmo de agrupamiento es un procedimiento mediante el cual una serie de características como datos de entrada son agrupados en varios clusters o grupos de acuerdo con un criterio. Esta técnica permite admitir o descartar características dependiendo de los resultados que obtienen los clusters a los que pertenecen durante el entrenamiento de aprendizaje (Cartagena, 2017).

Existen dos tipos de clustering o agrupamiento:

- Agrupamiento de partición dura.
- Agrupamiento de partición suave.

En el caso de agrupamiento de **partición dura** cada característica o dato de entrada pertenece exclusivamente a un cluster. Además, los clusters deben cubrir totalmente el conjunto de datos, es decir cada dato tiene que pertenecer a alguno de los clusters.

En el agrupamiento de **partición suave** a cada característica o dato de entrada se le asigna un valor de pertenencia dentro de cada cluster por lo que un dato puede pertenecer parcialmente a más de un cluster. Esto es así porque no siempre es fácil clasificar un dato de entrada en uno de los clusters, dado que dicho dato puede contener características pertenecientes a clusters distintos y encontrarse lo suficientemente cerca de dos clusters en el entrenamiento de aprendizaje (Reyes & Crespo, 2018).

Esta técnica es muy utilizada en minería de datos con el fin de encontrar similitudes entre datos complejos de entre una gran cantidad de datos. Se trata de una técnica **no supervisada**.

### **Bolsa de palabras (Bag of Words)**

El modelo de bolsa de palabras es una técnica simple para extraer características de documentos. En particular este modelo convierte documentos de texto en vectores, cada documento es convertido en un vector que representa la frecuencia de todas las distintas palabras que están presentes en el espacio del vector del documento (Aquino, 2013).

Para el presente proyecto, se creará un listado de palabras sin repeticiones, para generar un conjunto de datos. A partir de este conjunto se extraen las características para la etapa de clasificación del tweet en alguna categoría determinada y se filtra para retener sólo los términos etiquetados.

## **Open Source**

El término open source o código abierto se refiere a algo que las personas pueden modificar y compartir porque su diseño es de acceso público.

El término se originó en el contexto del desarrollo de software para designar un enfoque específico para crear programas de computadora. Sin embargo, hoy en día, "código abierto" designa un conjunto más amplio de valores, lo que se denomina "la forma de código abierto". Los proyectos, productos o iniciativas de código abierto abarcan y celebran los principios de intercambio abierto, participación colaborativa, creación rápida de prototipos, transparencia, meritocracia y desarrollo orientado a la comunidad (Villaverde E., 2019).

El software de código abierto es un software con código fuente que cualquiera puede inspeccionar, modificar y mejorar. El Código fuente es la parte del software que la mayoría de los usuarios de computadoras nunca ven; es el código que los programadores de computadoras pueden manipular para cambiar la forma en que funciona una pieza de software, un programa o aplicación. Los programadores que tienen acceso al código fuente de un programa de computadora pueden mejorar ese programa agregándole características o arreglando partes que no siempre funcionan correctamente (Villaverde, 2018).

Las licencias de código abierto afectan la forma en que las personas pueden usar, estudiar, modificar y distribuir software. En general, las licencias de código abierto otorgan a los usuarios de computadoras permiso para usar

software de código abierto para cualquier propósito que deseen. Algunas licencias de código abierto, lo que algunas personas llaman licencias de "copyleft", estipulan que cualquier persona que libere un programa de código abierto modificado también debe liberar el código fuente de ese programa junto con él. Además, algunas licencias de código abierto estipulan que cualquier persona que altere y comparta un programa con otros también debe compartir el código fuente de ese programa sin cobrar una tarifa de licencia por ello (GNU, 2019).

Por diseño, las licencias de software de código abierto promueven la colaboración y el intercambio, ya que permiten que otras personas realicen modificaciones en el código fuente e incorporen esos cambios en sus propios proyectos. Animán a los programadores de computadoras a acceder, ver y modificar el software de fuente abierta cuando lo deseen, siempre y cuando permitan que otros hagan lo mismo cuando comparten su trabajo (opensource, 2018).

### **Entorno de desarrollo: Jupyter Notebook**

El cuaderno Jupyter es una aplicación web de código abierto que le permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo. Los usos incluyen: limpieza y transformación de datos, simulación numérica, modelado estadístico, visualización de datos, aprendizaje automático y mucho más (Jupyter, 2018).

Jupyter Notebook es una plataforma web, opensource, donde se realizan las pruebas y desarrollo de proyectos de ciencia de datos, esta plataforma consiste en una aplicación web que ayuda a la creación y la portabilidad de documentos que contienen códigos, ecuaciones matemáticas, imágenes. Entre sus grandes cualidades, cuenta con procesos como la transformación y

visualización de datos, modelado de aprendizajes, simulación en el aspecto numérico, modelado estadístico, entre otras más (Ionos, 2019).

Esta plataforma cuenta con soporte para más de 40 lenguajes de programación, entre los cual se incluye Python, el cual es el lenguaje y la versión la versión que se va a utilizar en el presente proyecto. Esta aplicación también permite la exportación de código por medio de código HTML, también de imágenes estadísticas, clúster, imágenes en 2D y visualización de puntos en un mapa por medio de la geolocalización (Bbva, 2015).

### **Firebase Realtime**

Firebase es una plataforma de Google, conocida por su servicio de base de datos en la nube. Este servicio permite conectar aplicaciones web y mobile con esta base de datos y actualizarse en tiempo real bidireccionalmente (Cascante, 2018).

Firebase Realtime Database es una base de datos NoSQL y alojada en la nube. Los datos se almacenan en formato JSON y se sincronizan en tiempo real con cada cliente conectado. Por lo tanto tiene diferentes optimizaciones y funcionalidades en comparación con una base de datos relacional. Realtime Database está diseñada para permitir operaciones que se puedan ejecutar rápidamente. Permitiendo obtener una excelente experiencia de tiempo real que sirve a millones de usuarios sin afectar la capacidad de respuesta (Developers, 2019).

Realtime Database proporciona un lenguaje flexible de reglas basadas en expresiones, llamadas reglas de seguridad de Firebase Realtime Database, para definir cómo se deben estructurar los datos y en qué momento se puede leer o escribir en la base de datos. La comunicación se efectúa de forma

bidireccional, es decir que cuando se modifican datos en la Firebase estos cambios se van reflejados directamente en la plataforma.

## **Python**

Es un lenguaje de desarrollo de programación de código abierto y uno de los más populares que existe en la comunidad de científicos de datos. Es usado desde los más básicos como “scripts”, hasta servidores web de alta prestaciones, es compatible con la licencia publica general de GNU, la cual es una licencia orientada al derecho de autor y usada en el mundo del software libre y del código abierto, posee programación con orientación a objetos, programación funcional e imperativa. Fue desarrollado en finales de los años ochenta y comienzo de los noventas por Guido Van Rossum (Patricia, 2014).

Su utilización en este proyecto se debe a que es sencillo, flexible y open source, además tiene disponible una diversidad de librerías útiles de acuerdo a las distintas necesidades del desarrollador o de lo que se quiera mostrar de acuerdo al área de funcionalidad. A continuación, se mencionaran algunas librerías utilizadas.

*Cuadro N. 6* Librerías de Python con sus funcionalidades

LIBRERÍA	FUNCIONALIDAD
<b>Tweepy</b>	Permite acceder a la API de Twitter desde Python.
<b>Seaborn</b>	Biblioteca de visualización de datos de Python basada en matplotlib. Proporciona una interfaz de alto nivel para dibujar gráficos estadísticos atractivos e informativos.
<b>NLTK</b>	Conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, derivación, etiquetado, análisis y razonamiento semántico, envoltorios para bibliotecas NLP de potencia industrial.
<b>Matplotlib</b>	Librería destinada a la representación gráfica a partir de una serie de datos.
<b>Pandas</b>	Librería destinada al análisis de datos, lo que nos proporcionará una visión de la información recopilada estructurada según lo que le pidamos. Ofrece distintas estructuras, desde series, dataframes, panel, panel4d y panelIND.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (PyData, 2018).

## **Librería NumPy**

NumPy es un paquete fundamental para la computación científica con Python. Contiene entre otras cosas (NumFOCUS, 2019):

- Un poderoso objeto de matriz N-dimensional.
- Funciones sofisticadas (difusión).
- Herramientas para la integración de código C / C ++ y Fortran.
- Álgebra lineal útil, transformada de Fourier y capacidades de números aleatorios.

Además de su uso científico, NumPy también se puede usar como un eficiente contenedor multidimensional de datos genéricos. Se pueden definir tipos de datos arbitrarios. Esto permite que NumPy se integre a la perfección con una amplia variedad de bases de datos (NumPy, 2018).

Dentro del presente proyecto la librería NumPy juega un rol muy importante en el almacenamiento de datos que corresponden a valores numéricos o a los cuales se les ha aplicado una función, como en el caso de funciones de limpieza de datos, que permiten eliminar cuentas de usuarios, URLs, espacios en blanco, etc. (Unipython, 2017).

De esta forma NumPy permite almacenar estos datos en un arreglo sin perderlos. Un uso bastante eficiente en la codificación de proyectos de ciencia de datos.

## **Twython**

Twython es un envoltorio de Python puro mantenido activamente para la API de Twitter. Admite las API de Twitter normales y de transmisión. Twython es una de las principales bibliotecas de Python que proporciona una forma fácil (y actualizada) de acceder a los datos de Twitter. Mantenido activamente y con

soporte para Python 2.6+ y Python 3. Ha sido probado por compañías, instituciones educativas e individuos por igual, su última versión 3.7 fue lanzada el 7 de mayo de 2018, puesto que no lanzaba una nueva versión desde el 2014 (McGrath, 2014).

## Tweepy

Tweepy es una librería de código abierto disponible que permite comunicar Python con la plataforma Twitter. Las funciones definidas por Tweepy simplifican sobremanera la conexión y búsquedas con Twitter (Selva, 2015).

Sin embargo (Asensio, 2015, pág. 19) menciona que:

“Tweepy es probablemente la librería open source más conocida para acceder a la API desde Python, provee acceso a los métodos de la API de Twitter. Se encuentra en GitHub y tiene una documentación muy completa y bastantes ejemplos”.

La presente librería es un "contenedor" escrito en Python que facilita el trabajo con la API de Twitter. Tweepy admite el acceso a Twitter a través de la autenticación básica y el método OAuth. Twitter ha dejado hace varios años de aceptar la autenticación básica, por lo que OAuth es ahora la única forma de usar la API de Twitter.

Tweepy facilita el uso de la API Streaming de Twitter al manejar la autenticación, la conexión, la creación y la destrucción de la sesión, la lectura de los mensajes entrantes y los mensajes de enrutamiento parcial (Sarmiento & Silva, 2017, pág. 40).

## **Librería Pandas**

Pandas es una librería de análisis de datos de Python, de código abierto con licencia BSD que proporciona estructuras de datos de alto rendimiento y fáciles de usar y herramientas de análisis de datos para el lenguaje de programación Python (NumFOCUS, PyData, 2018).

La presente biblioteca ayuda al lenguaje de programación Python en la tarea del análisis y el modelado de datos, dado a que las tareas mencionadas previamente no son su fortaleza, como si lo es la recopilación y preparación de datos. Pandas ayuda a llenar este vacío, permitiendo realizar todo el flujo de trabajo de análisis de datos en Python sin tener que cambiar a un lenguaje de programación más específico como R (Willems, 2016).

Si se combina el excelente kit de herramientas de Jupyter Notebook y otras bibliotecas, el entorno para realizar análisis de datos en python es excelente en rendimiento, productividad y capacidad de colaboración.

## **DataFrame**

El uso de dataframe permite procesar los resultados obtenidos y en cuestión de segundos; dependiendo de la cantidad de datos, visualizarlos de manera ordenada (Aguiar, 2017).

Cada resultado obtenido se guardará con el mismo identificador de la red social. Este procesado se divide en varios campos, entre los que destacan los siguientes:

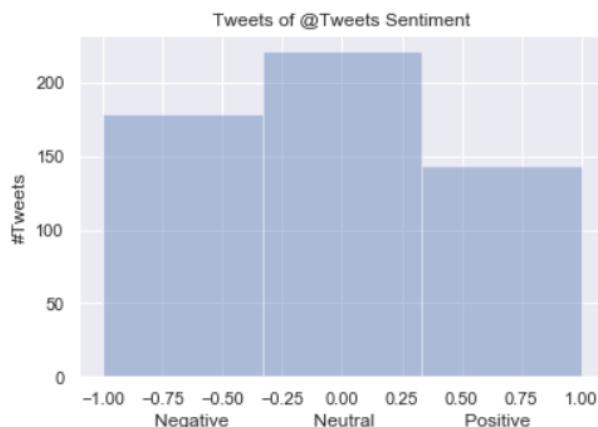
- **Tweet Data:** contiene texto del tweet, fecha y hora de creación, contador de retweets, contador de favoritos y la fuente original del mensaje.
- **At create Data:** contiene la fecha en que fue emitido el tweet, etc.

## Visualización de información para la ayuda a la toma de decisiones

Debido a los grandes volúmenes de datos que se manejan, encontrar y monitorear la opinión de muchos usuarios es una tarea difícil ya que identificar información relevante y extraerla de forma resumida es un procedimiento costoso para ser realizado manualmente. Por lo tanto se debe utilizar sistemas de análisis de sentimientos automáticos que nos permitan sintetizar la información de forma automática (Martín, 2016, pág. 14).

Una forma común es realizar un análisis exploratorio de los datos mediante visualizaciones. En el gráfico N. 7, se visualiza el potencial que otorga el uso de librerías para la creación de gráficos, permitiendo una mayor comprensión de los resultados.

**Gráfico N. 7** Representación gráfica de un análisis de sentimiento



**Nota:** Se puede observar la representación en gráfico de barra de la ejecución de un análisis de sentimientos, permitiendo visualizar la cantidad de tweets y el índice de polaridad.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Las visualizaciones permiten resumir grandes volúmenes de datos en representaciones gráficas. Posteriormente, un emprendedor pueda interpretarlos rápidamente y hacer mejores conclusiones. Luego, puede tomar una decisión basada en la información recolectada.

**FUNDAMENTACIÓN LEGAL**  
**LEY DE PROPIEDAD INTELECTUAL**  
**TITULO I**  
**DE LOS DERECHOS DE AUTOR Y DERECHOS CONEXOS**  
**CAPITULO I**  
**DEL DERECHO DE AUTOR SECCION I**  
**PRECEPTOS GENERALES**

**Art. 4.** Se reconocen y garantizan los derechos de los autores y los derechos de los demás titulares sobre sus obras.

**Art. 5.** El derecho de autor nace y se protege por el solo hecho de la creación de la obra, independientemente de su mérito, destino o modo de expresión. Se protegen todas las obras, interpretaciones, ejecuciones, producciones o emisión radiofónica cualquiera sea el país de origen de la obra, la nacionalidad o el domicilio del autor o titular. Esta protección también se reconoce cualquiera que sea el lugar de publicación o divulgación.

El reconocimiento de los derechos de autor y de los derechos conexos no está sometido a registro, depósito, ni al cumplimiento de formalidad alguna.

El derecho conexo nace de la necesidad de asegurar la protección de los derechos de los artistas, intérpretes o ejecutantes y de los productores de fonogramas.

**Art. 6.** El derecho de autor es independiente, compatible y acumulable con:

- La propiedad y otros derechos que tengan por objeto la cosa material a la que esté incorporada la obra;
- Los derechos de propiedad industrial que puedan existir sobre la obra; y,
- Los otros derechos de propiedad intelectual reconocidos por la ley.

También se tendrá en cuenta artículos referenciados al derecho de autor la cual se citará:

## **Sección II**

### **Objeto del derecho de autor**

**Art. 8.** La protección del derecho de autor recae sobre todas las obras del ingenio, en el ámbito literario o artístico, cualquiera que sea su género, forma de expresión, mérito o finalidad. Los derechos reconocidos por el presente Título son independientes de la propiedad del objeto material en el cual está incorporada la obra y su goce o ejercicio no están supeditados al requisito del registro o al cumplimiento de cualquier otra formalidad.

Las obras protegidas comprenden, entre otras, las siguientes:

a) Libros, folletos, impresos, epistolarios, artículos, novelas, cuentos, poemas, crónicas, críticas, ensayos, misivas, guiones para teatro, cinematografía, televisión, conferencias, discursos, lecciones, sermones, alegatos en derecho, memorias y otras obras de similar naturaleza, expresadas en cualquier forma;

- b) Colecciones de obras, tales como antologías o compilaciones y bases de datos de toda clase, que por la selección o disposición de las materias constituyan creaciones intelectuales, sin perjuicio de los derechos de autor que subsistan sobre los materiales o datos;
- c) Obras dramáticas y dramático musicales, las coreografías, las pantomimas y, en general las obras teatrales;
- d) Composiciones musicales con o sin letra;
- e) Obras cinematográficas y cualesquiera otras obras audiovisuales;
- f) Las esculturas y las obras de pintura, dibujo, grabado, litografía y las historietas gráficas, tebeos, comics, así como sus ensayos o bocetos y las demás obras plásticas;
- g) Proyectos, planos, maquetas y diseños de obras arquitectónicas y de ingeniería;
- h) Ilustraciones, gráficos, mapas y diseños relativos a la geografía, la topografía, y en general a la ciencia;
- i) Obras fotográficas y las expresadas por procedimientos análogos a la fotografía;
- j) Obras de arte aplicada, aunque su valor artístico no pueda ser disociado del carácter industrial de los objetos a los cuales estén incorporadas;
- k) Programas de ordenador; y,

I) Adaptaciones, traducciones, arreglos, revisiones, actualizaciones y anotaciones; compendios, resúmenes y extractos; y, otras transformaciones de una obra, realizadas con expresa autorización de los autores de las obras originales, y sin perjuicio de sus derechos.

Sin perjuicio de los derechos de propiedad industrial, los títulos de programas y noticieros radiales o televisados, de diarios, revistas y otras publicaciones periódicas, quedan protegidos durante un año después de la salida del último número o de la comunicación pública del último programa, salvo que se trate de publicaciones o producciones anuales, en cuyo caso el plazo de protección se extenderá a tres años.

### **Sección III**

#### **Titulares de los derechos**

**Art. 11.** Únicamente la persona natural puede ser autor. Las personas jurídicas pueden ser titulares de derechos de autor, de conformidad con el presente Libro.

**Art. 12.** Se presume autor o titular de una obra, salvo prueba en contrario, a la persona cuyo nombre, seudónimo, iniciales, sigla o cualquier otro signo que lo identifique aparezca indicado en la obra.

## **Constitución De La República Del Ecuador**

### **Sección primera**

#### **Educación**

**Art. 350.** El sistema de educación superior tiene como finalidad la formación académica y profesional con visión científica y humanista, la investigación científica y tecnológica, la innovación, promoción, desarrollo y difusión de los saberes y las culturas, la construcción de situaciones para los problemas del país, en relación con los objetivos del régimen de desarrollo.

**Art. 355.** El estado reconocerá a las universidades y escuelas politécnicas autonomía académica, administrativa, financiera y orgánica acorde con los objetivos del régimen de desarrollo y los principios establecidos en la Constitución.

**Art. 385.** El sistema nacional de ciencia, tecnología, innovación y saberes ancestrales, en el marco del respeto al ambiente, la naturaleza, la vida, las culturas y la soberanía, tendrá como finalidad:

1. Generar, adaptar y difundir conocimientos científicos y tecnológicos.
2. Recuperar, fortalecer y potenciar los saberes ancestrales.
3. Desarrollar tecnologías e innovaciones que impulsen la producción nacional, eleven la eficiencia y productividad, mejoren la calidad de vida y contribuyan a la realización del buen vivir.

**Art. 386.** El sistema comprenderá programas, políticas, recursos, acciones, e incorporará a instituciones del Estado, universidades y escuelas politécnicas, institutos o jurídicas, en tanto realizan actividades de investigación, desarrollo tecnológico, innovación y aquellas ligadas a los saberes ancestrales.

**Art. 136.- Trabajos realizados por investigadores y expertos extranjeros.-**

El reporte final de los proyectos de investigación deberán ser entregados por los centros de educación superior, en copia electrónica a la Secretaría Nacional de Educación Superior Ciencia, Tecnología e Innovación. Esta información será parte del Sistema Nacional de Información de la Educación Superior de investigación pública y particulares, empresas públicas y privadas, organismos no gubernamentales y personas naturales

**LOTAIP (Ley Orgánica de Transparencia y Acceso a la Información Pública).**

**Art. 6.- Información confidencial.-** Se considera información confidencial aquella información pública personal, que no está sujeta al principio de publicidad y comprende aquella derivada de sus derechos personalísimos y fundamentales, especialmente aquellos señalados en los artículos 23 y 24 de la Constitución Política de la República. El uso ilegal que se haga de la información personal o su divulgación, dará lugar a las acciones legales pertinentes. No podrá invocarse reserva, cuando se trate de investigaciones

que realicen las autoridades, públicas competentes, sobre violaciones a derechos de las personas que se encuentren establecidos en la Constitución Política de la República, en las declaraciones, pactos, convenios, instrumentos internacionales y el ordenamiento jurídico interno. Se excepciona el procedimiento establecido en las indagaciones previas.

### **Ley sobre el acuerdo de software libre en el Ecuador**

**Art. 2.-** Se entiende por Software Libre, a los programas de computación que se pueden utilizar y distribuir sin restricción alguna, que permitan su acceso a los códigos fuentes y que sus aplicaciones puedan ser mejoradas.

Estos programas de computación tienen las siguientes libertades:

- a) Utilización del programa con cualquier propósito de uso común
- b) Distribución de copias sin restricción
- c) Estudio y modificación del programa (Requisito: código fuente disponible)
- d) Publicación del programa mejorado (Requisito: código fuente disponible).

### **Preguntas científicas a contestarse**

1. ¿Cuáles serán los beneficios que obtendrán los emprendedores al utilizar la plataforma Jupyter Notebook y un código fuente para la extracción de datos públicos de la Red Social Twitter?
2. ¿De qué manera se emplean los algoritmos de aprendizaje automático para brindar apoyo en la toma de decisiones a los emprendedores en la ciudad de Guayaquil?

## **Definiciones conceptuales**

**Algoritmo:** se define como una secuencia de pasos con una etapa inicial y con una etapa final que da como resultado una ejecución de tareas, cada paso se expresa de forma específica (Pamies, 2017).

**Análisis de sentimientos:** Es una parte del procesamiento del lenguaje natural (PLN) y está estrechamente relacionado con la minería de opinión y el análisis de subjetividad, por ende estudia los campos subjetivos, un significado más profundo puede demostrarse en las expresiones lingüísticas de los estados particulares en un texto relacionados con las palabras sueltas, frases u oraciones (Gómez & Gallego, 2018).

**Aprendizaje automático:** También denominado aprendizaje de máquinas es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan (Digital, 2016).

**Audiencia:** Número de individuos destinatarios que están expuestos a un tipo particular de publicidad o a algún medio de comunicación y que, en ocasiones, interactúa con ellos. Las audiencias suelen dividirse según diversas variables, como la edad o el sexo, para determinar los contenidos que se les ofrecen (Marketing, 2016).

**Corpus:** El corpus es un conjunto de documentos (conjunto de datos de entrenamiento) que se utiliza inicialmente como entrada para entrenar un algoritmo clasificador de sentimientos, proporciona una base para cualquier prueba investigativa, tanto textual como oral, para el presente caso solo se tendrá en cuenta la parte escrita ya que se trabajará con tweets escritos por usuarios de las redes sociales Twitter (Gómez & Gallego, 2018).

El corpus de datos en el análisis de sentimientos puede estar compuesto por una recopilación de textos, materiales lingüísticos, semánticos; según el tema del dominio a tratar. Para este caso en particular el estudio se centró en la polaridad transmitida por los usuarios en la red social Twitter.

**Emprendedor:** persona que diseña, lanza y pone en funcionamiento un nuevo negocio. El emprendimiento es la capacidad y el deseo de desarrollar, organizar y manejar un negocio junto con sus respectivos riesgos, y con el fin de obtener una ganancia (Latam, 2019).

**Emprendimiento:** Efecto de emprender, llevar adelante una obra o un negocio. El emprendimiento suele ser un proyecto que se desarrolla con esfuerzo y haciendo frente a diversas dificultades, con la resolución de llegar a un determinado punto (Latam, 2019).

**Gráficos estadísticos:** Son muestras gráfica para describir numéricamente un conjunto de datos relacionados para proporcionar un análisis final (Ingenio, 2017).

**Industria:** Forma específica de actividades comerciales dentro de una región específica. Una industria es una división de la economía, formada por un gran número de empresas comerciales, que tienen actividades relacionadas (Sawakinome, 2018).

**Notebook:** Es un fichero generado por Jupyter Notebook o Jupyter Lab que se puede editar desde un navegador web, permitiendo mezclar la ejecución de código Python con anotaciones (Torres J. , Deep learning: Introducción práctica con Keras, 2018, pág. 66).

**Oportunidad de negocio:** Ocación u oportunidad para comenzar una idea empresarial, adentrarse en un nuevo sector laboral o el lanzamiento de un

nuevo producto en el mercado. Aprovechar una necesidad de los consumidores, satisfacer una demanda o presentar un servicio o artículo nuevo en el mercado que destaque por su potencial innovador (García I. , 2017).

**Sector:** Representa un grupo de industrias que tienen atributos comunes. Los diferentes sectores tienen características específicas, lo que significa que las industrias del sector operan en una línea específica de productos o servicios. Un sector incluye una división de economía donde una gran cantidad de industrias tienen actividades relacionadas (Sawakinome, 2018).

**Stopword:** Son palabras que no agregan ninguna información al texto desde un punto de vista semántico (Inboundcycle, 2016).

**Tendencia:** Inclinación o disposición natural que una persona tiene hacia una cosa determinada. En un sentido general, es un patrón de comportamiento de los elementos de un entorno particular durante un período (Consuegra, 2014).

**Trending topic (hashtags #):** Es una de las palabras o frases más repetidas en un momento concreto en Twitter. Los diez más relevantes se muestran en la página de inicio, pudiendo el usuario escoger el ámbito geográfico que prefiera, mundial o localizado, o personalizadas, en función además de a quién sigue el propio usuario (García V. , 2018, pág. 4).

## **CAPÍTULO III**

### **PROPUESTA TECNOLÓGICA**

El presente proyecto, propone la implementación de algoritmos de aprendizaje supervisado para el análisis de tweets en la plataforma web Jupyter Notebook, ampliamente usada en proyectos de ciencia de datos, en la cual se incorpora el uso de análisis de sentimiento en español para determinar si existe una audiencia positiva o negativa en base a la búsqueda de un determinado producto o servicio que los emprendedores deseen ofertar. Como parte general del proyecto se incorpora el uso de una red neuronal para clasificar los tweets por sus respectivos sectores e industrias, mediante la búsqueda de patrones de consumo que permitan generar insight a los futuros emprendedores, teniendo como finalidad convertirse en un medio que permita brindar apoyo en la toma de decisiones de los emprendimientos suscitados en la ciudad de Guayaquil.

En relación con la implementación de los algoritmos de aprendizaje supervisado, se hará uso de herramientas de software libre; el lenguaje de programación utilizado es Python, además del uso de lenguaje de marcado Markdown para una mayor comprensión de la estructura del notebook, y librerías específicas para la ciencia de datos con Python; para la manipulación y el acceso a la información se hará uso de archivos CSV, la cual provee de mayor flexibilidad y rapidez al momento de utilizar el notebook en otro computador o en la nube a través de Google Colaboratory, debido a que levantar una base de datos para un usuario final que no posee conocimientos en informática puede resultar un proceso laborioso que conlleva tiempo de aprendizaje.

Es de alta probabilidad que los emprendedores o dueños de negocios no usen estas herramientas porque desconocen el valor de la información que se puede obtener mediante el uso de la herramienta Jupyter Notebook, la cual aportaría una ventaja competitiva con el uso de la tecnología en la toma de decisiones, generando confianza gracias al análisis de datos reales.

- **Análisis de factibilidad**

En base a las investigaciones realizadas en diferentes revistas especializadas en tecnología como lo es Tendencias 21 y revistas de emprendimiento como la revista Líderes, se conoce que en Ecuador no se presentan indicios al momento de emprender del uso de plataformas opensource para el análisis de datos que permita determinar mediante el análisis de sentimientos cómo se encuentra posicionado un determinado producto o servicio, reflejando una polaridad que denote si está en auge o en decadencia por medio de las opiniones de los usuarios en Twitter. Y lo más importante que mediante el uso del procesamiento del lenguaje natural y algoritmos de aprendizaje automático se proceda a clasificar las opiniones vertidas en esta red social, para determinar en qué sector o industria existe una mayor interactividad por parte de los usuarios, y cuál es la polaridad dominante. Facilitando la toma de decisiones a los emprendedores de la ciudad de Guayaquil.

La factibilidad de este proyecto se basa en brindar apoyo a los emprendedores al momento de tomar decisiones, brindando una mejor experiencia al momento de disipar dudas sobre cómo está posicionado en la opinión del usuario un producto o servicio. Esto se realiza a través del análisis y procesamiento de datos, convirtiendo todo esto en información dentro de la plataforma web Jupyter Notebook.

## **- Factibilidad operacional**

El presente proyecto tiene como usuario final a los emprendedores y dueños de negocio que estén interesados en el análisis de las opiniones efectuadas en las redes sociales, en este caso en particular la red social Twitter. Para evaluar que tan posicionado entre los usuarios se encuentra su actual o futuro producto o servicio. Y visualizar que sectores son los que poseen mayor interactividad en esta red mediante el conjunto de tweets analizados.

La plataforma web Jupyter Notebook tiene una interfaz amigable y de fácil adiestramiento. Y gracias al uso del lenguaje de marcado Markdown se puede realizar una breve explicación en texto de cómo usar el notebook. Y para una mayor usabilidad se ha adoptado el uso de archivos CSV, que permiten una rápida implementación y uso del notebook en cualquier computador que tenga previamente instalado la plataforma Jupyter.

Actualmente elegir un sector en el que se desea emprender requiere de una correcta decisión, y contar con una herramienta que mediante la incorporación de diferentes algoritmos brinde ayuda a estos usuarios en la toma de decisiones resulta importante al momento de realizar un emprendimiento o efectuar un cambio en vuestro negocio.

El desarrollo de este proyecto de ciencia de datos es factible operacionalmente, debido a que la plataforma web seleccionada emplea una interfaz fácil y amigable para los usuarios y estos no necesitan tener un amplio conocimiento sobre entornos de desarrollo web para interactuar y hacer uso de ella, permitiéndoles acceder a la misma desde cualquier computadora que tenga instalado Jupyter y acceso a internet para poder extraer tweets que posteriormente serán analizados con toda facilidad y rapidez, en el menor tiempo posible.

### - Factibilidad técnica

El uso de la plataforma web Jupyter Notebook se efectúa desde cualquier computador, indistintamente del sistema operativo en el cual se instalará la aplicación: ya sea Windows, MacOs o Linux.

El API Streaming de Twitter mediante la cual se realizaron las peticiones de los tweets, permite el acceso a tweets públicos mediante el uso de credenciales, entregadas por la empresa dueña de la red social, en este caso Twitter Inc. Siempre y cuando se respete los límites de extracción en su versión Standard (Twitter, 2018).

Al utilizar el lenguaje de programación Python se cuenta con una gran variedad de librerías, necesarias para la realización de proyectos de ciencia de datos. Las cuales proveen de documentación y comunidades en línea para consultar su implementación y solucionar errores que suelen surgir. Además las herramientas de desarrollo software utilizadas son de libre acceso, es por esto que técnicamente el desarrollo de este proyecto en la plataforma Jupyter Notebook es factible. A continuación, se presenta el cuadro N. 7, donde se indica el hardware y software utilizado para el desarrollo y pruebas del proyecto.

*Cuadro N. 7 Hardware utilizado en el desarrollo*

Hardware	Características
Laptop DELL Inspirion 5000	Windows 10 Professional Procesador Intel Core i5 – 5200 2.20 Ghz. Almacenamiento 1000 Gb RAM 4 Gb.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.  
**Fuente:** Datos de la investigación.

**Cuadro N. 8** Software utilizado en el desarrollo

Software	Versión
Anaconda Navigator	1.9.7
Jupyter Notebook	5.7.8
Python	3.7.3
NLTK	3.4.4
Firebase Admin Python SDK	2.17.0

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

Con base al estudio y los resultados obtenidos en las encuestas realizadas (Ver gráfico N. 8, pág. 81) se ha determinado que la mayor parte de la población hace uso de la red social Twitter en su vida cotidiana por lo tanto es apto para la realización del presente proyecto y en el cuadro N. 8 se mencionan las herramientas utilizadas para el presente desarrollo del proyecto.

### - Factibilidad Legal

El presente proyecto no vulnera ninguna de las reglamentaciones consideradas para su desarrollo como: la privacidad de la información de los tweets extraídos mediante el API dado a que estos son públicos y no se almacena información que permita identificar al emisor de un tweets como lo es el *nombre* o su *screen\_name* (identificador o el alias con el que un usuario se puede identificar en Twitter). La información mencionada en el capítulo II, se refiere a los artículos que menciona el uso de hardware libre y software libre, de manera más específica en el Decreto Ejecutivo 1014 en sus artículos 1 y 2. Además que su desarrollo hace uso de tecnologías open source y no incurre en ninguna infracción legal que imposibilite la ejecución del proyecto.

## - Factibilidad Económica

El proyecto propuesto es sin fines de lucros y está enfocado en servir de apoyo en la toma de decisiones de personas que desean emprender o ya poseen un negocio, por lo que las herramientas utilizadas para su desarrollo son de Software libre y no tienen costo alguno.

El periodo de tiempo comprendido para el presente proyecto es de tres meses, desde el 10 de junio hasta el 30 de agosto.

**Cuadro N. 9** Presupuesto y financiamiento

Rubros	Valor mensual	Meses	Subtotal
Recursos Humanos	\$300	3	\$900
Recursos de Software	\$0	3	\$0
Conectividad e Internet	\$20	3	\$60
Servicios básicos	\$25	3	\$75
Gastos varios	\$25	3	\$75
<b>TOTAL:</b>			<b>\$1110</b>

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

La factibilidad económica entre recursos humanos, transporte, servicios de internet y gastos varios, da un total de \$1110 que serán invertidos en el proyecto.

La inversión de este proyecto es totalmente financiada por el propio estudiante. Dentro de estos gastos también están considerado el transporte, viáticos, los gastos varios que serían los costos de impresiones, copias, anillados, empastados, que son cubiertos por el estudiante realizador del proyecto.

## • Etapas de la metodología del proyecto

En este proyecto se emplea la metodología de investigación diagnostica, para recabar información útil en el desarrollo de los notebooks del proyecto, por lo

que se ha utilizado como técnica de recolección de datos la encuesta, la cual será aplicada utilizando como instrumento un cuestionario electrónico a la muestra correspondiente con el fin de capturar características necesarias que involucra el proyecto.

## **Metodología de investigación**

### **Metodología de Investigación descriptiva**

La presente metodología de investigación consiste en describir características o funciones del mercado. La investigación descriptiva supone que el investigador tiene conocimiento previo acerca de la situación del problema a estudiar y se caracteriza por la formulación previa de hipótesis. Por tal motivo, la información requerida debe definirse con precisión. Este tipo de investigación es usada por diferentes motivos, entre los que se destaca:

- Describe características de grupos relacionados como pueden ser consumidores, vendedores organizaciones o sectores del mercado
- Se utiliza para calcular un porcentaje de una población específica que tiene patrones de conducta.

Entre de las principales herramientas que esta metodología pone a disposición para la recolección de datos y posterior análisis, es la encuesta. Dando uso de ella para elaborar un cuestionario con múltiples opciones, el cual fue elaborado en línea a través de los servicios de Google, como lo es Google Forms. El formato de las preguntas se encuentra en el Anexo 2.

## **Procesamiento y análisis**

Antes de comenzar con el desarrollo del presente proyecto es importante realizar una encuesta a los interesados, fundamentalmente a las personas que usan Twitter y las personas que desean emprender algún negocio con el fin

de presentar un proyecto que sea útil. Para este proyecto se empleó la encuesta como herramienta para recolectar datos para ver qué tan alto es el uso de Twitter y también para determinar si las personas están interesadas en un proyecto que presente información útil en el apoyo de toma de decisiones de algún tipo de emprendimiento.

En el presente proyecto se realizaron dos encuestas:

1. Encuesta N. 1 para usuarios de la red social Twitter en la ciudad de Guayaquil
2. Encuesta N. 2 para emprendedores y dueños de negocios de la ciudad de Guayaquil.

### **Recolección de información**

Para la recolección de información en el presente proyecto se utilizó la plataforma Google Forms, esta es una herramienta online gratuita que nos permite crear formularios de preguntas y posteriormente enviarla a las personas que van a ser encuestadas, estas encuestas pueden ser compartidas por diferentes medios.

### **Herramientas de análisis**

Para el análisis de los datos se dio uso del software estadístico IBM® SPSS Statistics, el cual proporciona distintas técnicas, incluyendo el análisis ad-hoc, pruebas de hipótesis e informes, facilitando la gestión de datos, seleccionar y realizar análisis y compartir los resultados. Combinado con la exportación de las respuestas registradas en las encuestas creadas en Google Forms .Se obtiene un análisis eficiente de los datos obtenidos en las encuestas.

## **Población y muestra**

Según el Instituto Nacional de Estadísticas y Censos (INEC), en Ecuador hay 17,2 millones de habitantes, de ellos 13,8 millones están conectados a Internet y 12 millones son usuarios de redes sociales.

Con base en el estado digital de Ecuador emitido en febrero del 2019 por Formación Gerencial, se conoce que la red social Twitter tiene 4 millones de perfiles registrados en Ecuador, de los cuales solo 800.000 constan como cuentas activas. Además, gracias a Formación Gerencial se conoce que la red social más utilizada en Ecuador es Facebook, con un total del 23% de usuarios en Ecuador registrados en la ciudad de Guayaquil (Del Alcazar, 2019).

Para Twitter no se tiene registro del porcentaje de usuarios en la ciudad de Guayaquil, pero se utilizará el porcentaje registrado de usuarios de Facebook para determinar el tamaño de la población, así que de las 800.000 cuentas activas de la red social Twitter se tomará el 23%. Obteniendo un total de 184.000 cuentas para la ciudad de Guayaquil.

Para realizar esta encuesta se realizó un tipo de muestreo probabilístico, cuya fórmula es la siguiente.

$$n = \frac{P \times Q \times N}{(N - 1)E^2 / k^2 + P \times Q}$$

Reemplazando valores obtenemos el tamaño de la muestra:

**Cuadro N. 10** Población y muestra de la encuesta N. 1.

Variable	Descripción	Valor
<b>P</b>	Probabilidad de éxito. Este dato es generalmente desconocido y se suele suponer que p=q=0.5	0.5
<b>Q</b>	Probabilidad de fracaso. Este dato es generalmente desconocido y se suele suponer que p=q=0.5	0.5
<b>N</b>	Tamaño de la población	184000
<b>E</b>	Error de estimación	0.05
<b>K</b>	Constante depende del nivel de confianza que asignemos	2%

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

$$n = \frac{0.5 \times 0.5 \times 184000}{(184000 - 1)x(0.05)^2 / (2)^2 + 0.5 \times 0.5}$$

$$n = \frac{46000}{(183999)x(0.0025) / 4 + 0.25}$$

$$n = \frac{46000}{115.25}$$

$$n = 399,13 \quad \textbf{n} = \mathbf{399}$$

Luego de haber aplicado la fórmula y proceder a los cálculos para obtener la muestra se obtiene un total de 399 habitantes para la primera encuesta. Se logró encuestar en su totalidad a usuarios de la red social Twitter en la ciudad de Guayaquil.

Para la encuesta N. 2 se usará la primicia de (Flores, 2018) que 1 de cada 3 ecuatorianos, es emprendedor. Para obtener el tamaño de la población de dueños de negocios y emprendedores con una cuenta activa en la red social Twitter. De un total de 184.000 se tomará el 33.33%, obteniendo un total de 61.327 usuarios; no es una cifra corroborada debido a que no existe un estudio que indique el número exacto.

Reemplazando valores obtenemos el tamaño de la muestra para la segunda encuesta:

**Cuadro N. 11** Población y muestra de la encuesta N. 2.

Variable	Descripción	Valor
P	Probabilidad de éxito.	0.5
Q	Probabilidad de fracaso.	0.5
N	Tamaño de la población	61327
E	Error de estimación	0.05
K	Constante depende del nivel de confianza que asignemos	2%

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

$$n = \frac{0.5 \times 0.5 \times 61327}{(61327 - 1)x(0.05)^2 / (2)^2 + 0.5 \times 0.5}$$

$$n = \frac{15331.75}{(61326)x(0.0025) / 4 + 0.25}$$

$$n = \frac{15331.75}{38.58}$$

$$n = 397,41 \quad \mathbf{n = 397}$$

Luego de haber aplicado la fórmula y proceder a los cálculos para obtener la muestra para la segunda encuesta se obtuvo como resultado un total de 397 habitantes para esta encuesta. Se logró encuestar en su totalidad a usuarios de la red social Twitter que están emprendiendo o son dueño de negocios en la ciudad de Guayaquil.

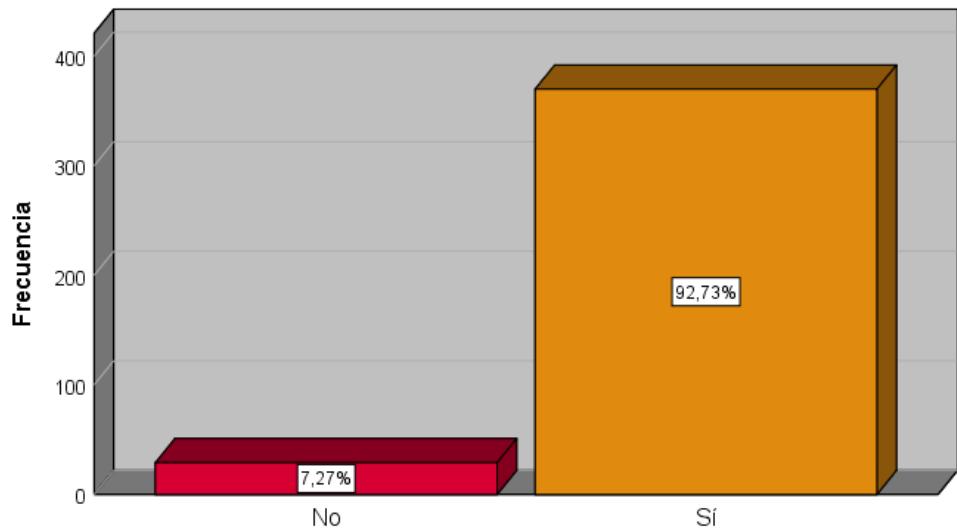
### **Encuesta N. 1**

Esta encuesta está dirigida para los usuarios de Twitter para identificar qué tan a menudo utilizan esta red social, frecuencia de emisión de tweets, corroborar si emiten tweets con valor comercial y útil en el apoyo en la toma de decisiones, y consultar si estarían de acuerdo de que sus tweets públicos sean analizados. Para esta encuesta se seleccionaron 399 personas, correspondiente al tamaño de la muestra para la encuesta N. 1.

Pregunta 1:

1. ¿Posee cuenta en la red social Twitter?

**Gráfico N. 8** Pregunta 1. ¿Posee cuenta en la red social Twitter?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

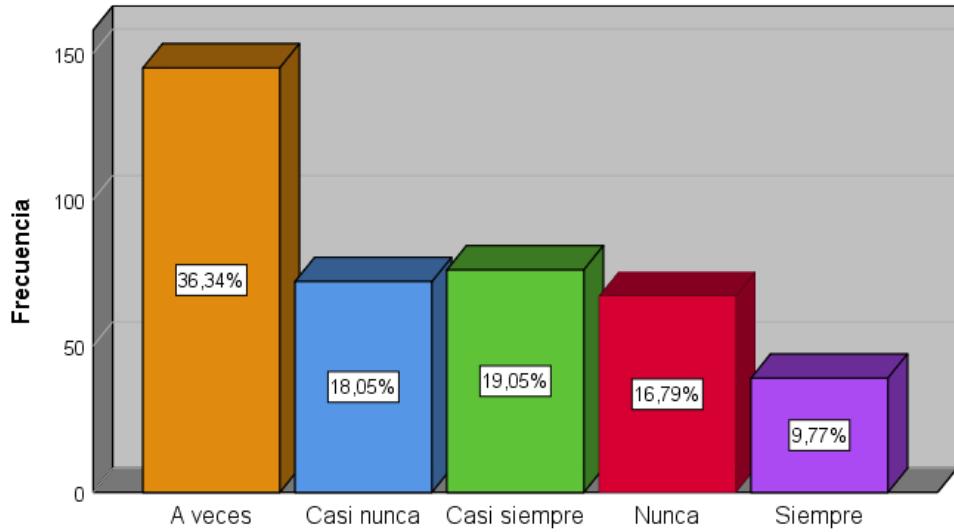
**Fuente:** Datos de la investigación.

**Análisis:** podemos observar que el 92.73% de los encuestados, si poseen una cuenta en la red social Twitter, mientras que el 7.27% no posee una cuenta en esta red social. Estos resultados son positivos para el presente proyecto debido a que se requiere una gran cantidad de datos para ser analizadas.

Pregunta 2:

2. ¿Con qué frecuencia emite tweets en su cuenta de Twitter?

**Gráfico N. 9** Pregunta 2. ¿Con qué frecuencia emite tweets en su cuenta de Twitter?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

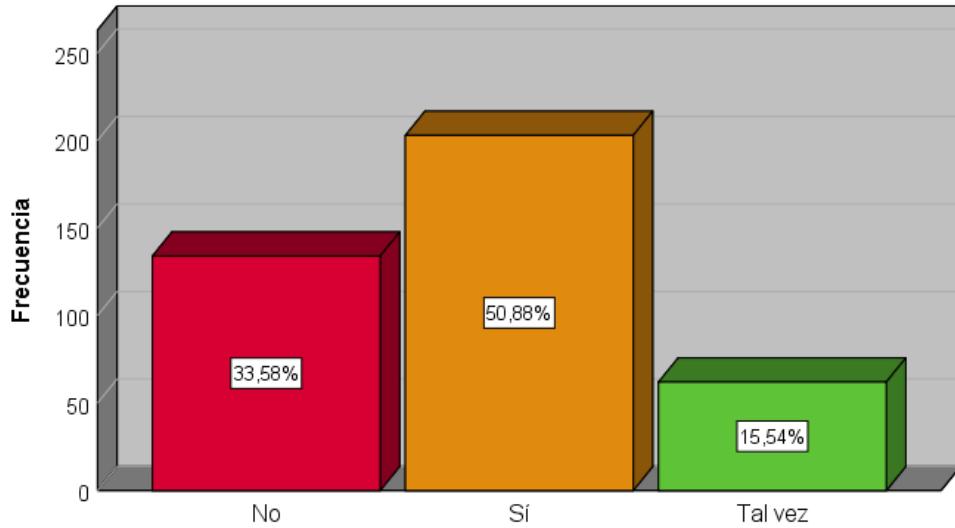
**Fuente:** Datos de la investigación.

**Análisis:** según el análisis de los datos podemos observar que el 36.34% de los encuestados a veces emiten tweets, para el 18.05% casi nunca emiten tweets, para el 19.05% casi siempre publican tweets y para el 16.79% nunca publican tweets, mientras que solo un 9.77% siempre emiten tweets. Para interés del proyecto se procede a sumar los valores de siempre, casi siempre y a veces, el cual nos da un total del 65.16% de frecuencia de emisión de tweets, no es un valor recurrente, pero nos da un indicio positivo.

Pregunta 3:

3. ¿Se considera un usuario activo en la Red Social Twitter?

**Gráfico N. 10** Pregunta 3. ¿Se considera un usuario activo en la Red Social Twitter?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

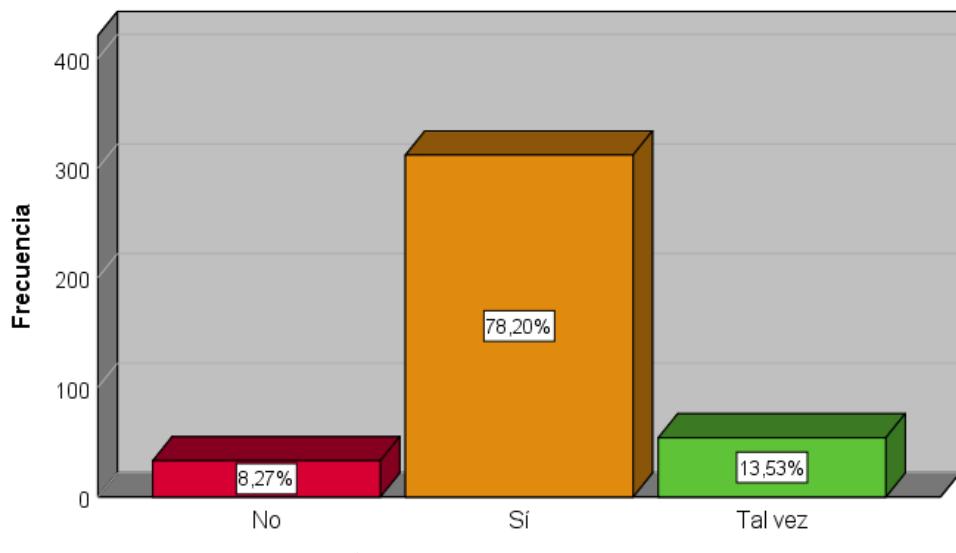
**Fuente:** Datos de la investigación.

**Análisis:** como podemos observar solo el 50.88% de los encuestados se consideran usuarios activos en la red social Twitter, mientras que el 33.58% de los encuestados se definen como no activos y el 15.54% de los encuestados tal vez se consideran usuarios activos, debido a que se conectan con semanas de separación de actividad en la red social.

Pregunta 4:

4. ¿Cree que actualmente puede expresar libremente su opinión sobre cualquier tema en la red social Twitter?

**Gráfico N. 11** Pregunta 4. ¿Cree que actualmente puede expresar libremente su opinión sobre cualquier tema en la red social Twitter?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

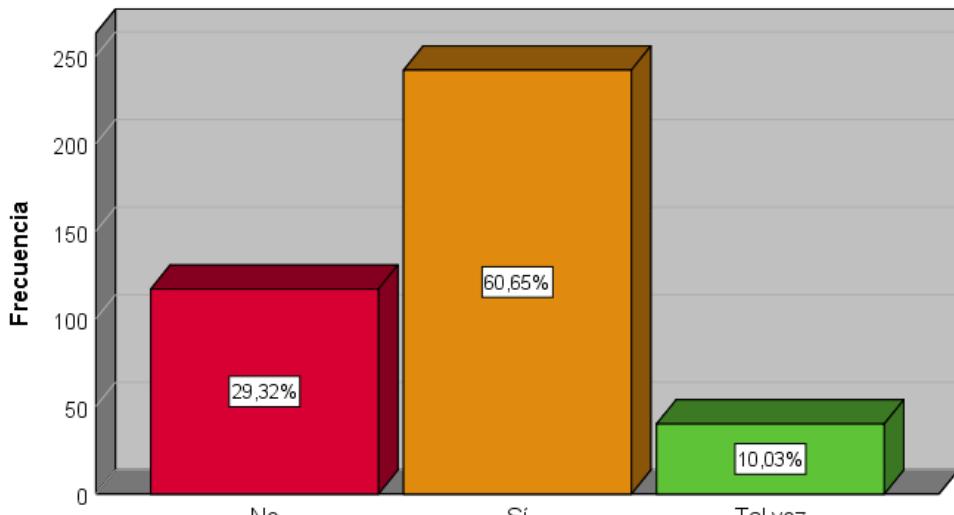
**Fuente:** Datos de la investigación.

**Análisis:** para el 78.20% de los individuos encuestados creen que actualmente pueden expresar libremente su opinión sobre cualquier cuestión en la red social Twitter, mientras que el 8.27% creen no sentirse libres al emitir un comentario, y para 13,53% de los encuestados creen que tal vez; debido a que la mayoría de opiniones que se realizan en Twitter pueden ser reportadas.

#### Pregunta 5:

5. ¿Alguna vez ha expresado en la red social Twitter la necesidad o deseo sobre un bien o servicio?

**Gráfico N. 12** Pregunta 5. ¿Alguna vez ha expresado en la red social Twitter la necesidad o deseo sobre un bien o servicio?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

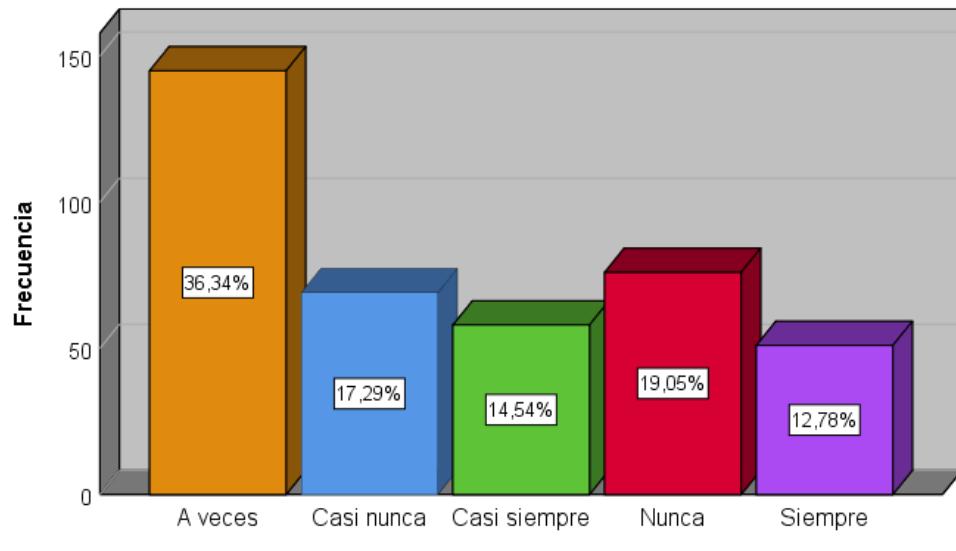
**Fuente:** Datos de la investigación.

**Análisis:** según los datos analizados podemos observar que el 60.65% de los individuos encuestados si ha expresado la necesidad o deseo sobre un bien o un servicio, mientras que el 29.32% no lo han hecho, y el 10.03% tal vez creen que han expresado la necesidad o deseo sobre un bien o un servicio en sus tweets emitidos.

Pregunta 6:

6. ¿Con qué frecuencia ha expresado su necesidad o deseo sobre un bien o servicio en la red social Twitter?

**Gráfico N. 13** Pregunta 6. ¿Con qué frecuencia ha expresado su necesidad o deseo sobre un bien o servicio en la red social Twitter?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

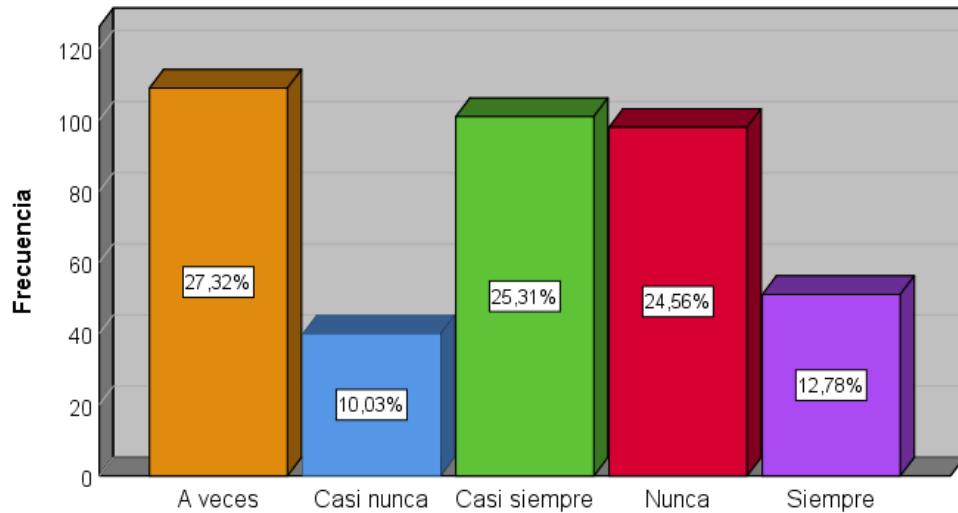
**Fuente:** Datos de la investigación.

**Análisis:** según los datos analizados podemos ver que el 36.34% de los encuestados a veces han expresado su necesidad o deseo sobre un bien o un servicio, para el 17.29% casi nunca se han expresado, para el 14.54% casi siempre se han expresado, el 19.05% nunca expresan en este medio la necesidad de un bien o un servicio, mientras que el 12.78% siempre ha manifestado su deseo sobre un bien o un servicio en la red social Twitter. Para interés del proyecto se procede a sumar los valores de siempre, casi siempre y a veces, el cual nos da un total del 63.66% de frecuencia de emisión de necesidad o deseo sobre un producto o servicio, se deja en claro que no es un valor recurrente.

Pregunta 7:

7. ¿Cuándo usted adquiere, consume o utiliza algún bien, producto o servicio lo ha publicado en la red social Twitter?

**Gráfico N. 14** Pregunta 7. ¿Cuándo usted adquiere, consume o utiliza algún bien, producto o servicio lo ha publicado en la red social Twitter?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

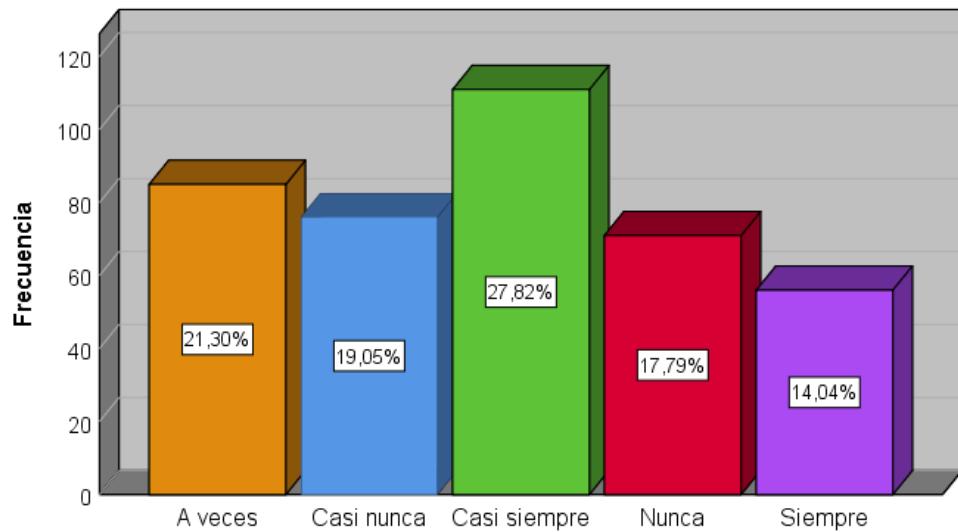
**Fuente:** Datos de la investigación.

**Análisis:** con estos resultados podemos apreciar que gran parte de las personas encuestadas que han consumido o utilizan un producto o servicio lo han publicado en la red social, con un total de 25.31% para la opción casi siempre, mientras que un 27.32% a veces lo ha publicado, para el 12.78% siempre han emitido un tweets respecto al producto o servicio que han adquirido, mientras que un 24.56% deja en claro que nunca han emitido un tweets de esta índole y un 10.03% casi nunca lo hacen.

Pregunta 8:

8. ¿Con qué frecuencia ha expresado usted su opinión sobre un producto o servicio en la red social Twitter?

**Gráfico N. 15** Pregunta 8. ¿Con qué frecuencia ha expresado usted su opinión sobre un producto o servicio en la red social Twitter?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

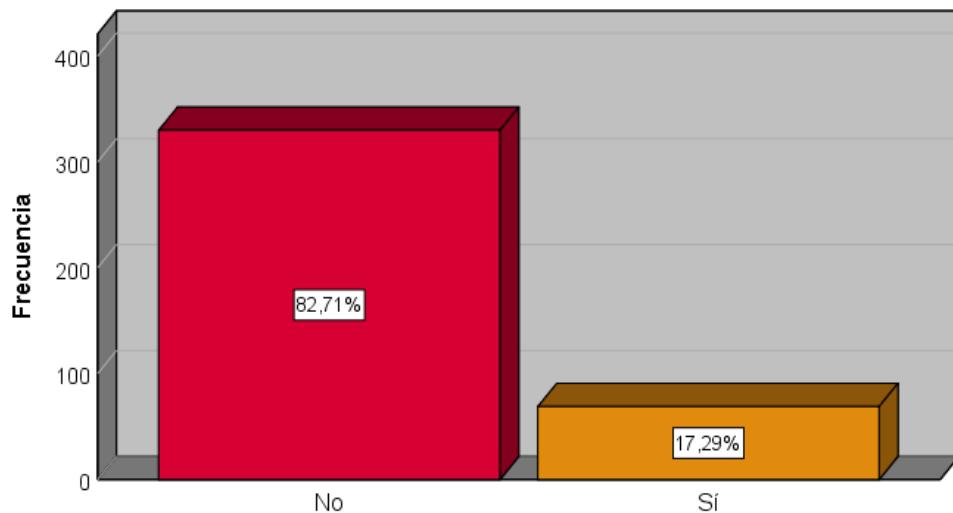
**Fuente:** Datos de la investigación.

**Análisis:** con estos resultados podemos apreciar que la mayoría de los encuestados casi siempre han expresado su opinión sobre un producto o servicio en la red social Twitter que son un 27.82%, para el 21.30% a veces han expresado su opinión, para 19.05% casi nunca han expresado su opinión de un servicio o producto en la red social, para el 17.79% nunca han expresado su opinión de un servicio mientras que solo un 14.04% siempre han expresado su opinión por un producto o servicio adquirido.

Pregunta 9:

9. ¿Sabía usted que sus publicaciones en la red social Twitter pueden ayudar en la toma de decisiones a nivel empresarial y de emprendimiento?

**Gráfico N. 16** Pregunta 9. ¿Sabía usted que sus publicaciones en la red social Twitter pueden ayudar en la toma de decisiones a nivel empresarial y de emprendimiento?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

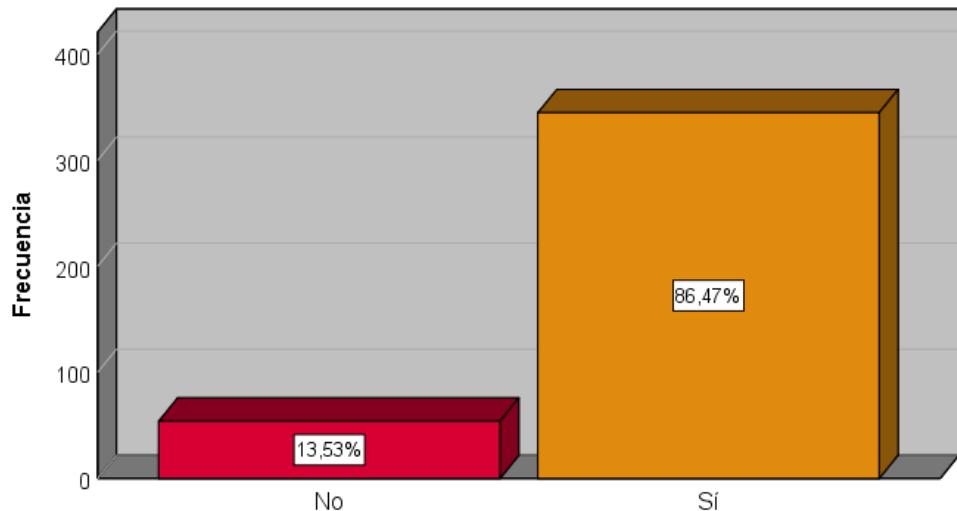
**Fuente:** Datos de la investigación.

**Análisis:** podemos observar que el 82.71% de los encuestados no sabían que sus publicaciones en la red social Twitter puede ayudar en la toma de decisiones a nivel empresarial y emprendimiento, mientras que el 17.29% si sabían que las publicaciones pueden ayudar en la toma de decisiones, mediante el análisis de datos.

#### Pregunta 10:

10. ¿Está usted de acuerdo que sus tweets públicos (con características específicas) sean analizados para brindar apoyo en la toma de decisiones en los emprendimientos de la ciudad de Guayaquil?

**Gráfico N. 17** Pregunta 10. ¿Está usted de acuerdo que sus tweets públicos (con características específicas) sean analizados para brindar apoyo en la toma de decisiones en los emprendimientos de la ciudad de Guayaquil?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

**Análisis:** podemos observar que el 86.47% de los encuestados si aceptan que sus tweets sean analizados para brindar apoyo en la toma de decisiones en los emprendimientos de la ciudad de Guayaquil, mientras que el 13.53% no está de acuerdo, debido a que creen que se irrumpen con la privacidad de los datos emitidos en esta red social.

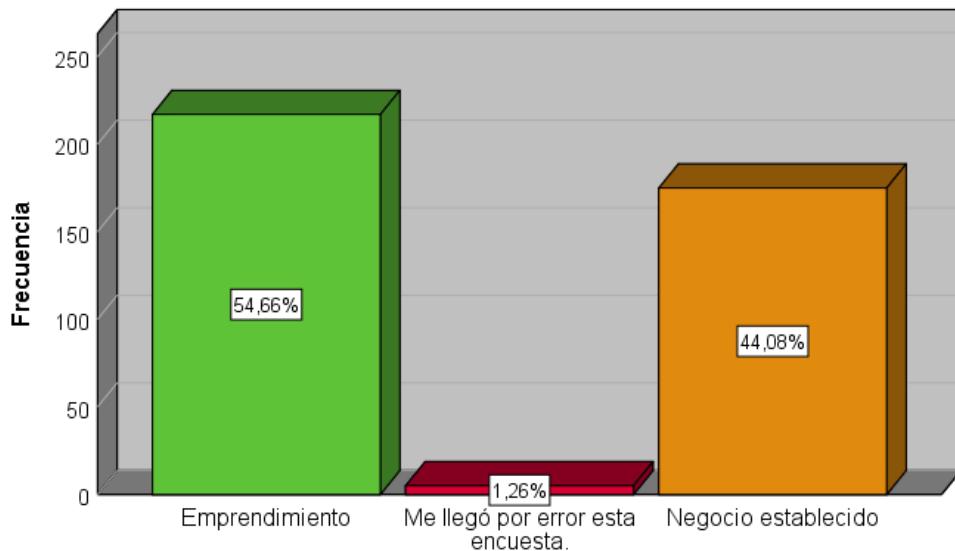
## Encuesta N. 2

Esta encuesta está dirigida a las personas que van a utilizar el notebook en la ciudad de Guayaquil para conocer su opinión sobre este proyecto, dedicado al análisis de los tweets generando información que brinde apoyo en la toma de decisiones al momento de emprender o redireccionar un negocio ya establecido, para esta encuesta se tomó el valor total de la muestra calculada para la encuesta N. 2 con un total de 397 personas encuestadas.

Pregunta 1:

1. Seleccione su situación actual:

**Gráfico N. 18** Pregunta 1. Seleccione su situación actual.



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

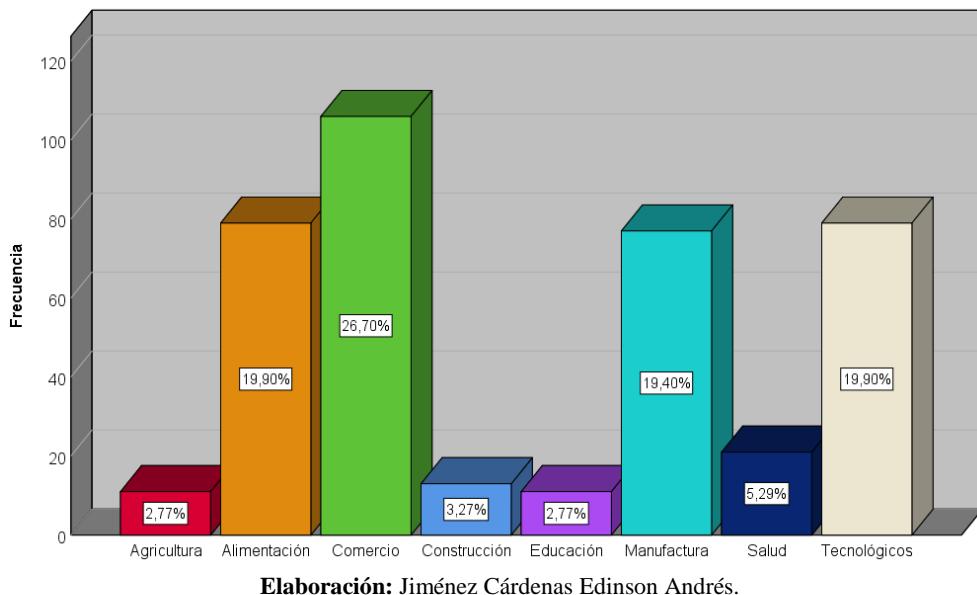
**Fuente:** Datos de la investigación.

**Análisis:** Con estos resultados podemos apreciar que el 54.66% de las personas interesadas son personas que han realizado un emprendimiento, esto es bueno para los fines del proyecto ya que el análisis de datos del presente proyecto es para brindar información acerca de los sectores e industrias con las que más interactúan los usuarios de la red social Twitter en la presente ciudad. Con un 1.26% de personas que les llegó por error esta encuesta y un 44.08% de los encuestados son personas dueñas de negocios.

Pregunta 2:

2. Ubique el sector de su negocio o emprendimiento en la siguiente lista:

**Gráfico N. 19** Pregunta 2. Ubique el sector de su negocio o emprendimiento en la siguiente lista.



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

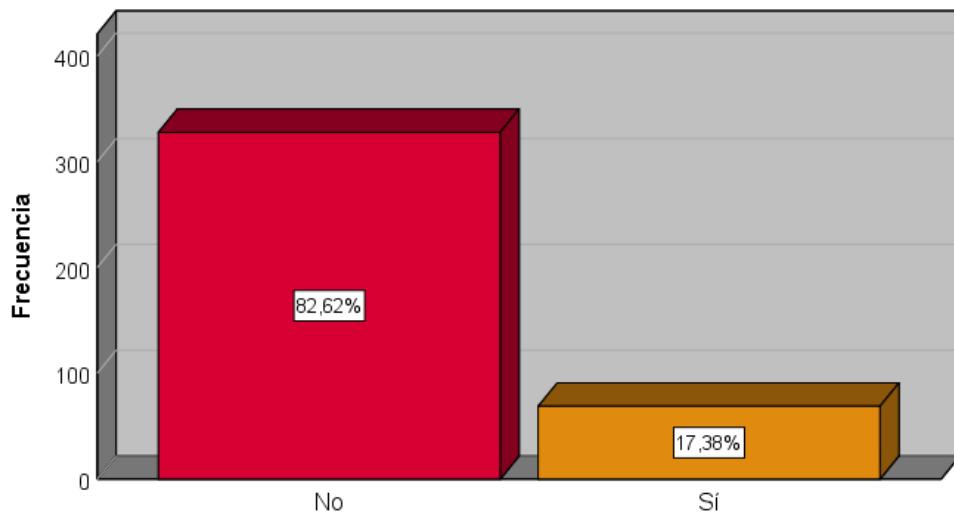
**Fuente:** Datos de la investigación.

**Análisis:** con estos resultados podemos observar que 26.70% de las personas encuestadas se dedican al comercio siendo la cifra más alta de los encuestados seguido de la industria alimenticia que tiene un 19,90% y del sector tecnológico con un 19,90% mientras que 2.77% es para el sector agrícola, un 3.27% es para la industria de la construcción y el 2.77% es para la educación y el 19.40% es para la industria de manufactura y al final tenemos 5.29% para el sector salud. Podemos ver que los negocios con mayor auge en la ciudad de Guayaquil son los relacionados con el comercio, la alimentación y la tecnología, siendo mayor las personas que se dedican al sector comercial.

Pregunta 3:

3. ¿Sabía usted que Ecuador se posiciona como líder en la región en el índice de actividad emprendedora temprana?

**Gráfico N. 20** Pregunta 3. ¿Sabía usted que Ecuador se posiciona como líder en la región en el índice de actividad emprendedora temprana?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

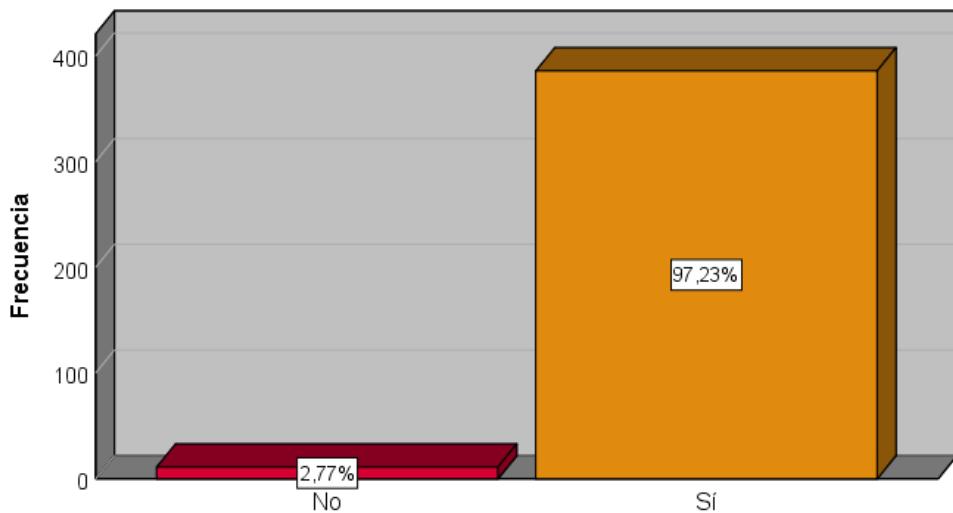
**Fuente:** Datos de la investigación.

**Análisis:** según el análisis de los datos se puede observar que el 82.62% de las personas encuestadas no sabían que Ecuador es el país líder en emprendimiento según la revista Líderes, mientras que el 17.38% de personas sí sabían que Ecuador es líder en actividad emprendedora temprana.

Pregunta 4:

4. ¿En su emprendimiento, utiliza las redes sociales para promocionar su servicio o producto y mantener contacto con sus clientes?

**Gráfico N. 21** Pregunta 4. ¿En su emprendimiento, utiliza las redes sociales para promocionar su servicio o producto y mantener contacto con sus clientes?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

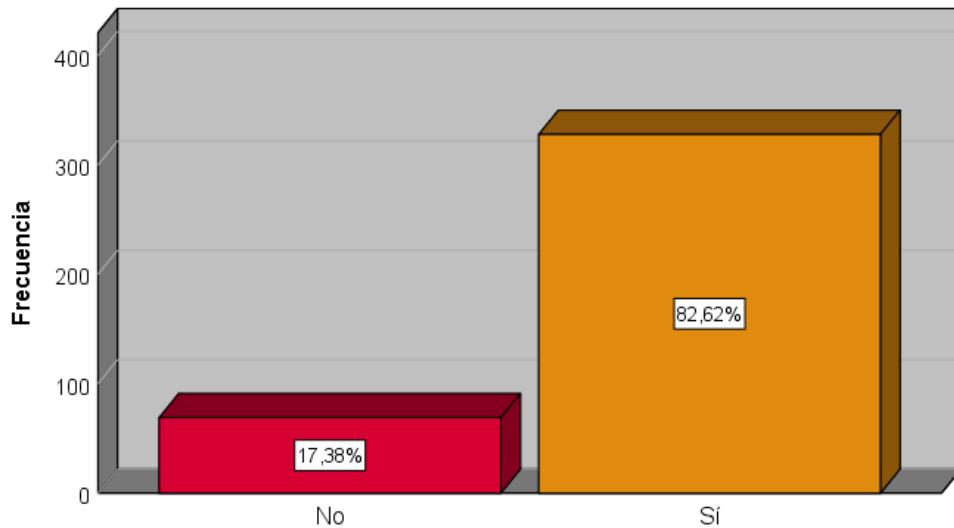
**Fuente:** Datos de la investigación.

**Análisis:** según los datos analizados podemos ver que el 97.23% de las personas encuestadas que realizan algún emprendimiento utilizan las redes sociales para promocionar sus servicio o producto mientras que el 2.77% no utilizan las redes sociales para promocionar sus negocios siendo esto una desventaja para aumentar su número de clientes e ingresos de dinero porque según ARCONEL la ciudad de Guayaquil ha aumento en el uso de las telecomunicaciones y las redes sociales.

Pregunta 5:

5. ¿En su emprendimiento o negocio establecido, hace uso de la red social Twitter?

**Gráfico N. 22** Pregunta 5. ¿En su emprendimiento o negocio establecido, hace uso de la red social Twitter?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

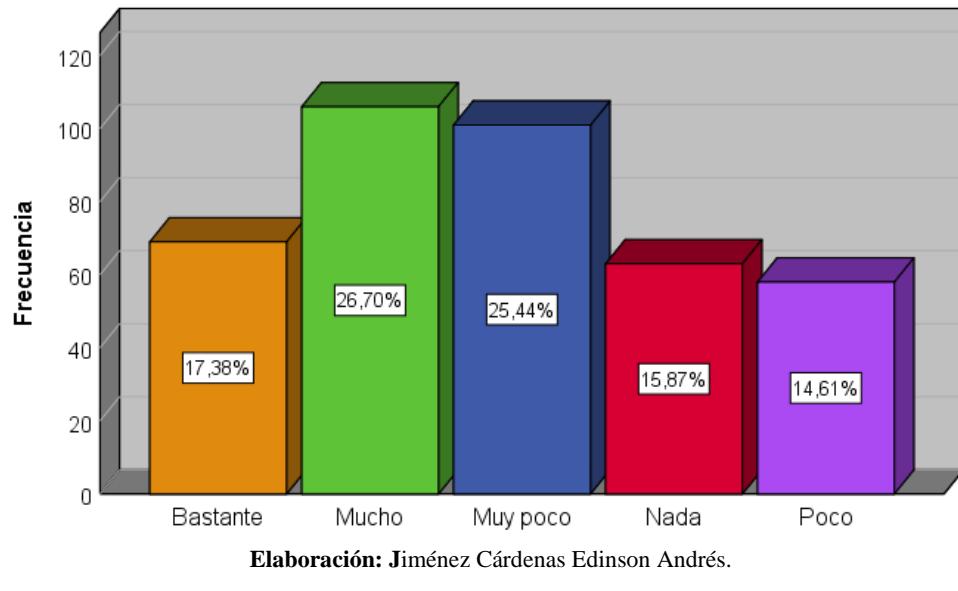
**Fuente:** Datos de la investigación.

**Análisis:** Se visualiza que el 82.62% de las personas utilizan Twitter para promocionar su negocio ya que aumenta más la posibilidad de ser conocidos y que no se fijan en una sola red social mientras que el 17.38% no utiliza esta red en su negocio.

#### Pregunta 6:

6. En caso de usar Twitter, indique la frecuencia con la que usted lee comentarios de sus clientes.

**Gráfico N. 23** Pregunta 6. En caso de usar Twitter, indique la frecuencia con la que usted lee comentarios de sus clientes.

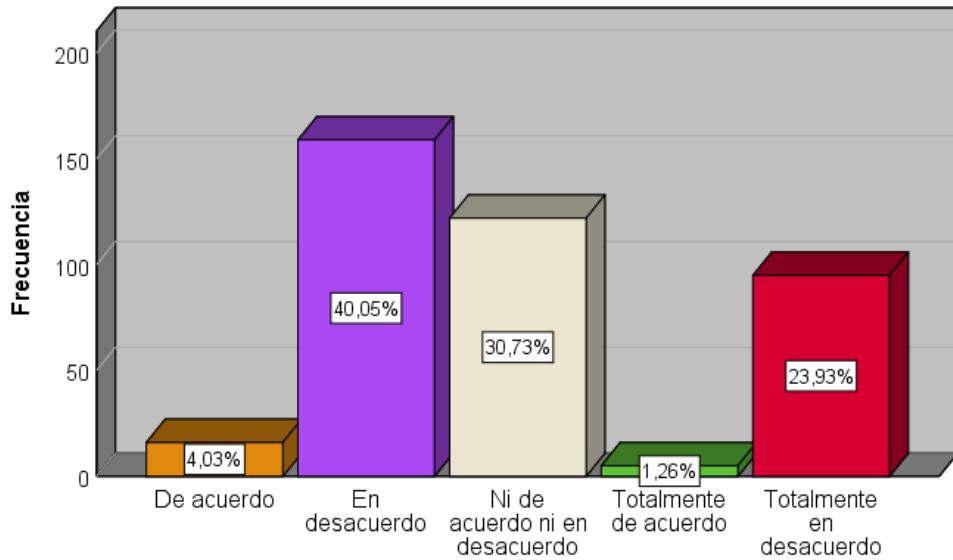


**Análisis:** se puede observar que el 25.44% de los individuos encuestados leen muy poco los comentarios emitidos en Twitter, pero el 26.70% si proceden a leer mucho los comentarios, el 14.61% los lee poco, mientras que el 15.87% no utiliza nada esta red, mientras que el 17.38% lee bastante los comentarios recibidos en esta red, con estos resultados podemos ver que los usuarios que usan la red social Twitter, tienden a no leer los comentarios que les emiten sus clientes por sus servicios o productos.

**Pregunta 7:**

7. Considera usted que el tiempo que emplea para leer las opiniones en Twitter de sus potenciales clientes es suficiente para determinar gustos y necesidades.

**Gráfico N. 24** Pregunta 7. Considera usted que el tiempo que emplea para leer las opiniones en Twitter de sus potenciales clientes es suficiente para determinar gustos y necesidades.



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

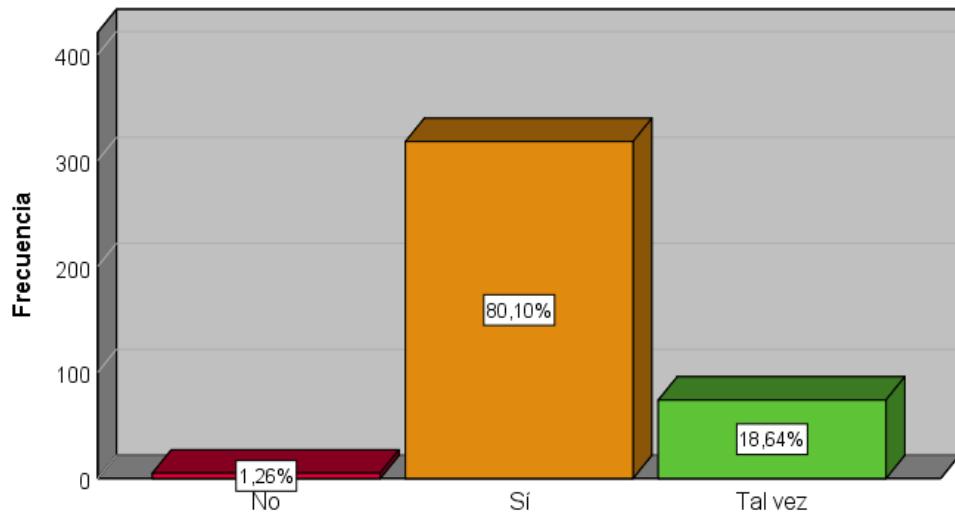
**Fuente:** Datos de la investigación.

**Análisis:** podemos observar que para el 40.05% de los encuestados, están en desacuerdo que el tiempo que emplean en leer las opiniones de los potenciales clientes no es suficiente para determinar gusto y necesidades, para el 30.73% no están de acuerdo ni en desacuerdo, para el 23.93% están total en desacuerdo, para el 1.26% están de acuerdo y para el 4.03% están totalmente de acuerdo; como se puede ver con estos resultados que los encuestados creen que con leer las opiniones de sus posibles clientes no es suficiente para determinar gusto y necesidades.

Pregunta 8:

8. ¿Si existiera una plataforma web que mediante su uso le permita visualizar que sectores/industrias tienen mayor actividad con los usuarios de la red social Twitter, usted la usaría?

**Gráfico N. 25** Pregunta 8. ¿Si existiera una plataforma web que mediante su uso le permita visualizar que sectores/industrias tienen mayor actividad con los usuarios de la red social Twitter, usted la usaría?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

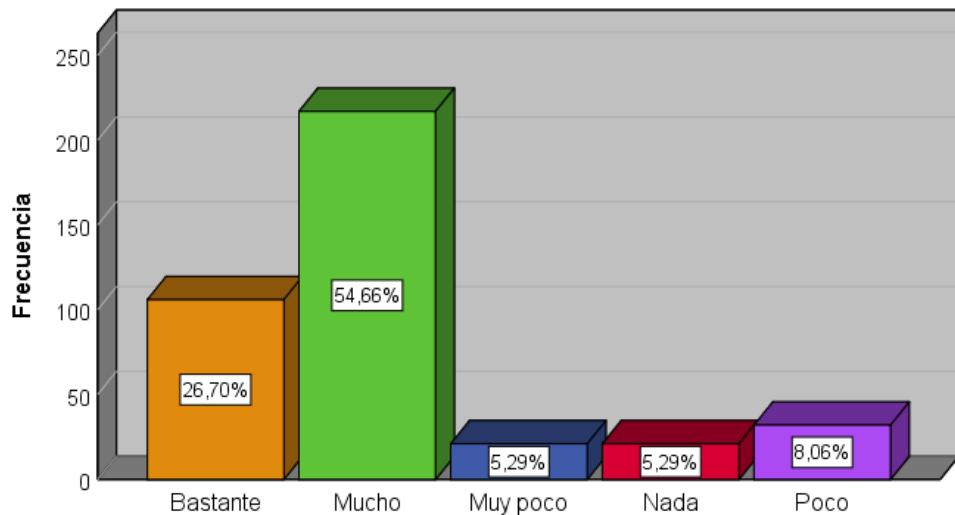
**Fuente:** Datos de la investigación.

**Análisis:** podemos observar que el 80.10% de las personas encuestadas sí utilizarían una plataforma web de ciencia de datos que les permita visualizar que sectores e industrias tienen mayor actividad con los usuarios de la red social Twitter, para el 18.64% tal vez la utilizarían, mientras que el 1.26% no la usarían. Estos resultados son positivos debido que al visualizar que sectores o industrias tienen mayor actividad se puede observar cuando hay un alto número de clientes en un sector y tomar mejores decisiones de negocios, todo esto en base a la opinión de los usuarios.

**Pregunta 9:**

9. Si existe una plataforma web que mediante su uso le permita visualizar información que brinde apoyo en la toma decisiones al momento de emprender, ¿estaría dispuesto a usarla y con qué frecuencia?

**Gráfico N. 26** Pregunta 9. Si existe una plataforma web que mediante su uso le permita visualizar información que brinde apoyo en la toma decisiones al momento de emprender, ¿estaría dispuesto a usarla y con qué frecuencia?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

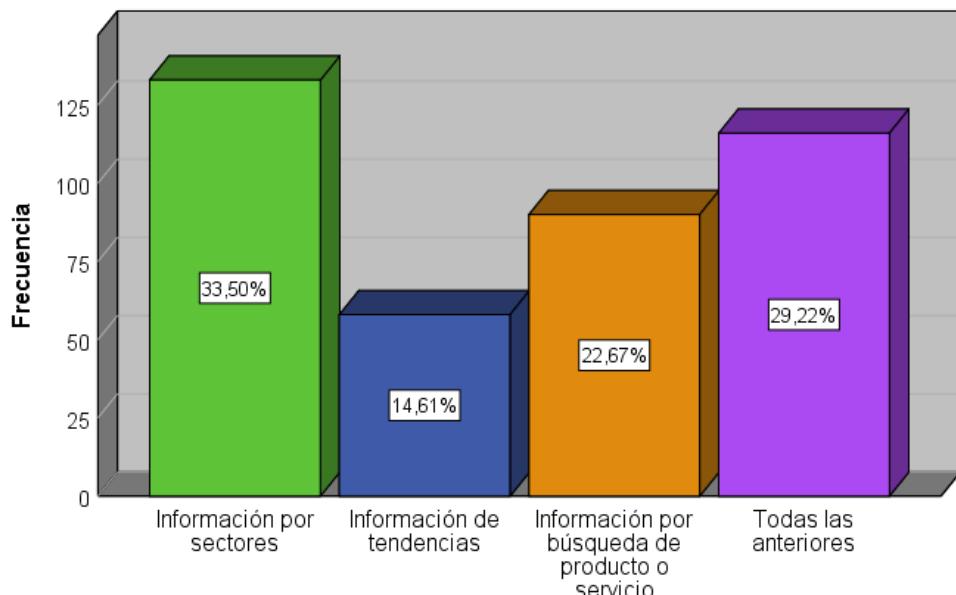
**Fuente:** Datos de la investigación.

**Análisis:** se visualiza que el 54.66% de los encuestados utilizarían mucho una plataforma web que le brinde apoyo en la toma de decisiones al momento de emprender, mientras que el 26.70% la utilizarían bastante, y para el 8.06% la utilizarían poco y el 5.29% la utilizarían muy poco, mientras que el otro 5.29% no la utilizarían. Estos resultados son positivos para los fines del presente proyecto, debido a que las personas si lo utilizarían al momento de realizar un emprendimiento, teniendo información visual al momento de tomar una decisión, o al momento de elegir un sector o industria para emprender.

Pregunta 10:

10. ¿Qué tipo de información cree usted que es determinante para el apoyo en la toma de decisiones al momento de emprender?

**Gráfico N. 27** Pregunta 10. ¿Qué tipo de información cree usted que es determinante para el apoyo en la toma de decisiones al momento de emprender?



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

**Análisis:** según el análisis de los datos podemos observar que el 33.50% de los encuestados creen que obtener información por sectores es determinante para el apoyo en la toma de decisiones en negocios y emprendimientos, para 14.61% es la información de tendencias, para el 22.67% el factor determinante es información por búsqueda de producto o servicio, mientras que 29.22% cree que todos los factores antes nombrados son determinantes para el apoyo en la toma de decisiones.

## **Metodología de desarrollo**

En la investigación realizada por (Galán, 2015, págs. 49-107) se demuestra la eficacia de la metodología CRISP-DM en proyecto de minería de datos en el entorno universitario. Por ende en el presente proyecto se emplea la metodología que se aplica en proyectos de ciencia y minería de datos, debido a que se hace uso de análisis de sentimientos en español y categorización por sectores de tweets públicos emitidos en la ciudad de Guayaquil utilizando el lenguaje de programación Python, ante lo expuesto anteriormente se hará uso de la metodología CRISP-DM.

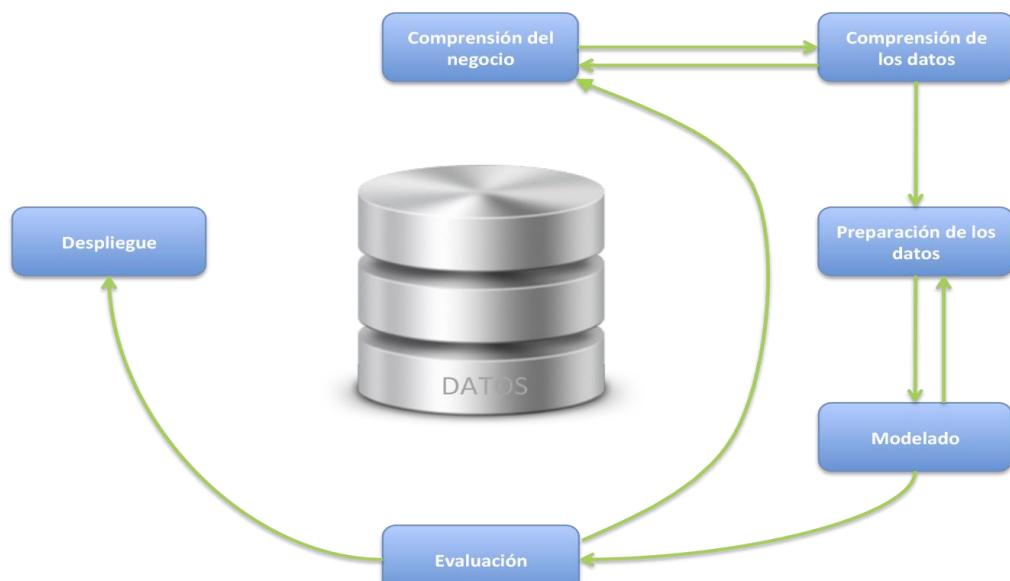
La metodología CRISP-DM, está dividida en seis fases que abarcan las actividades empleadas en la construcción de proyectos de data Mining; la fase de implementación no aplica al desarrollo del presente proyecto por tratarse de un proyecto académico. Los procesos empleados para el desarrollo de cada fase se detallaran más adelante conforme se vaya explicando cada fase de esta metodología aplicada a este proyecto.

## **Metodología CRISP-DM**

CRISP-DM (Cross Industry Standard Process for Data Mining), es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema (Galán, 2015, pág. 21). Para implementar una tecnología en un negocio se requiere una metodología. Estos métodos suelen venir de experiencias propias y también de los procedimientos estándar más conocidos. En el caso de los proyectos que hacen uso de técnicas de minería de datos una de las metodologías que ha tenido más apoyo de las empresas privadas y organismos públicos es la metodología CRISP-DM.

Según (Galán, 2015) esta metodología incluye un modelo y una guía, estructurada en seis fases, algunas bidireccionales, indicando que de una fase en concreto se puede volver a una fase anterior para poder revisarla, por lo que la sucesión de fases no tiene porqué ser ordenada desde la primera hasta la última. En el gráfico N. 28, se visualiza las fases en las que está dividida CRISP-DM y las secuencia a seguir.

**Gráfico N. 28** Secuencia del proceso CRISP-DM

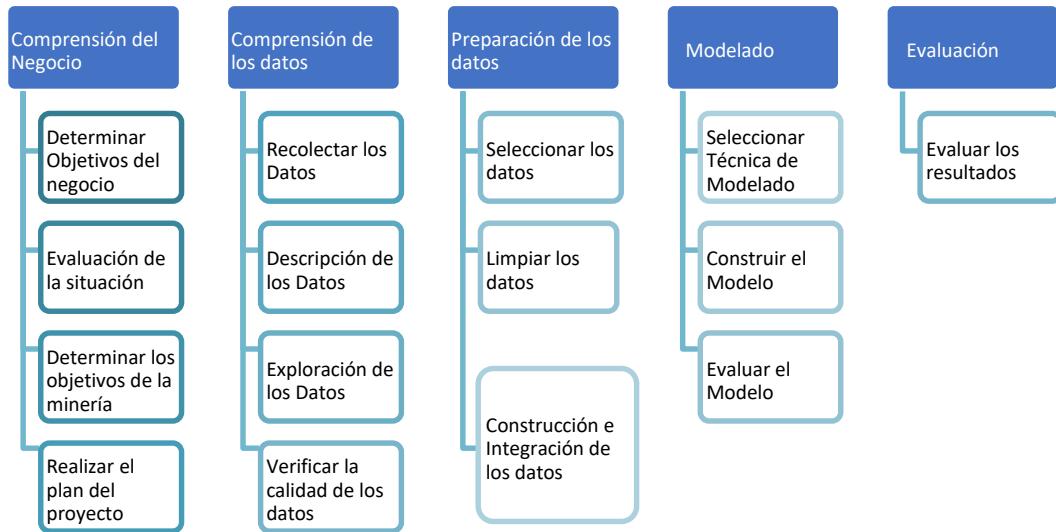


**Elaboración:** (Calvo, 2016).

**Fuente:** (Calvo, 2016).

La metodología CRISP-DM consta de 6 fases y cada una de ellas establecen tareas, la última fase de implementación no se hará uso en el presente proyecto, las fases a utilizar se describen a continuación:

**Gráfico N. 29** Fases de la metodología CRISP-DM



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario, 2015).

## Descripción de fases de CRISP-DM

### Comprensión del negocio.

La presente fase es probablemente la más importante y agrupa las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Si no se logra la comprensión de estos objetivos, ningún algoritmo por muy sofisticado que sea permitirá obtener los resultados esperados (Galán, 2015, pág. 22). Entender de manera completa el problema que se pretende resolver permite recolectar los datos correctos e interpretar correctamente los resultados.

### **Comprensión de los datos.**

Aquí se realiza una recolección inicial de los datos relacionados con el problema, posteriormente se realizar un análisis de los datos con la finalidad de detectar problemas de calidad e identificar relaciones interesantes entre los mismos que permitan definir las primeras hipótesis y generar conocimiento.

### **Preparación de los datos.**

En esta fase se efectúa una selección de los datos más relevantes para la creación de los dataset (conjunto de datos) a partir de los datos recopilados al inicio, a partir de los cuales se deberá realizar tareas de limpieza y transformación de tablas, registros y atributos para tener información de calidad y poder adaptarla a las técnicas de minería de datos que se utilizaran posteriormente, entre las que destaca la visualización de datos, búsqueda de patrones, etc.

La preparación de los datos incluye tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, data clean (limpieza de datos), generación de variables adicionales en el procesamiento de los datos, integración de diferentes orígenes de datos y cambios de formato.

### **Modelado**

En esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Según (Galán, 2015), las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada para el problema.
- Disponer de los datos adecuados.

- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Se procede a utilizar el conjunto de datos determinado, para procesarlo con la ayuda de una herramienta de minería de datos que implemente las técnicas necesarias para la construcción del modelo.

## **Evaluación**

En la presente fase se debe evaluar y comprobar la eficacia del modelo generado, verificando si los resultados que devuelve son los correctos, si está realizando las predicciones necesarias. Revisar los pasos realizados para la construcción del modelo y verificar que estos sean apropiados en función de los objetivos del negocio.

En su trabajo (Galán, 2015), indica que es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se pueda haber cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo.

## **Implantación o despliegue.**

Una vez que el modelo ha sido construido y validado, se procede a transformar el conocimiento obtenido en acciones dentro del proceso de negocio, esto puede hacerse por ejemplo cuando el analista recomienda acciones basadas en la observación del modelo y sus resultados. Generalmente un proyecto de minería de datos no concluye con la implantación del modelo, debido a que se debe documentar y presentar los resultados de manera comprensible para el

usuario con el objetivo de lograr un incremento del conocimiento. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados (Galán, 2015, pág. 30).

## **Aplicación de la metodología propuesta**

### **Fase I. Comprensión del negocio**

A continuación se detallan cada una de las tareas de las que consta esta primera fase en el proceso de la minería de datos, su finalidad es determinar los objetivos y requisitos del proyecto desde una perspectiva de negocio, para más adelante poder convertirlos en objetivos desde el punto de vista técnico y en un plan de proyecto.

### **Objetivos del negocio**

El objetivo del presente proyecto en su orientación de apoyo en la toma de decisiones y generar insight de negocio es permitir la búsqueda de un producto o servicio en el dataset de tweets almacenados para obtener información de que tendencias están relacionadas a esta búsqueda, ejecutar un análisis de sentimientos de las opiniones emitidas sobre esta búsqueda, permitiendo visualizar la apreciación o rechazo de un producto o servicio con base en la polaridad encontrada en estas opiniones. Además se clasificará los tweets por sectores e industrias dependiendo de su pertenencia o no, para una visualización general de interacción y polaridad encontrada que permita detectar cuáles son los sectores e industrias con mayor índice de interactividad e índice de consumo relacionados con un producto o servicio perteneciente a una industria o sector.

De esto último también se puede inferir que es de utilidad para la toma de decisiones, al poder determinar si un producto o servicio es aceptado o rechazado por la ciudadanía en base a sus opiniones emitidas.

Para el presente proyecto se han definido los siguientes objetivos:

- Búsqueda de patrones comerciales en los tweets para clasificarlos por sector e industria.
- Corpus para clasificación de sentimientos en español.
- Modelo para la clasificación de sentimientos en español.
- Análisis de sentimientos para determinar polaridad sobre un producto o servicio.
- Corpus de patrones comerciales en español.
- Red neuronal para la clasificación de tweets por sectores e industrias.

### **Criterios de éxito del negocio**

Desde el punto de vista del negocio se establece como criterio de éxito la correcta clasificación de sentimientos de los tweets y clasificación por sector o industria. Generando información que brinde apoyo en la toma de decisiones a los emprendedores y dueños de negocios de la ciudad de Guayaquil.

### **Fase II. Comprensión de los Datos**

En esta segunda fase de la metodología CRISP-DM se realiza la recolección inicial de los datos para poder establecer un primer contacto con el problema, familiarizarse con los datos y averiguar su calidad, así como identificar las relaciones más evidentes para formular las primeras hipótesis.

#### **Recolectar los datos iniciales**

El presente proyecto hace uso de tweets públicos generados en la ciudad de Guayaquil. Para su extracción de la red social Twitter se hizo uso del API Streaming, proporcionada por la empresa Twitter Inc.

Entre las ventajas que determinaron el uso de la red social Twitter para el presente proyecto, es que la mayoría de sus cuentas son públicas, esto permite extraer una gran y diversa cantidad de datos para analizar.

Para una mayor comprensión de este proceso se detallará de forma exhaustiva, paso a paso el proceso de obtención de los tweets públicos.

### **Registrar la cuenta de desarrollador para Twitter**

Para poder extraer tweets de la red social Twitter se debe registrar una cuenta Twitter de uso común en su sección de desarrollador, lo cual permite a la empresa Twitter asegurarse quién es usted y que va a realizar con la aplicación. Véase el *anexo 5*, en el manual de usuario se detalla este procedimiento.

Una vez concluido este proceso se tiene habilitado el acceso a las diferentes API's de Twitter, en nuestro caso específico al API Streaming.

Finalmente para conectarse desde Python a Twitter, se prepara la conexión con Twitter véase el gráfico N. 30 Utilizando la aplicación creada en la plataforma de desarrollador de Twitter, procedemos a usar el paquete Twython. Este paquete sirve para poder trabajar con todos los datos públicos que están en la API de la red social Twitter.

**Gráfico N. 30** Llaves y token's escritos en el notebook para acceder al API.

```
#Se define el token de la aplicacion
CONSUMER_KEY = "*****" # <---- Add your API Key
CONSUMER_SECRET = "*****" # <---- Add your API Secret
#Se define el acceso al usuario

ACCESS_KEY = "*****" # <---- Add your access token
ACCESS_SECRET = "*****" # <----Add your access token secret

stream = MiStream(CONSUMER_KEY,CONSUMER_SECRET,ACCESS_KEY,ACCESS_SECRET)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

## Accediendo a los datos

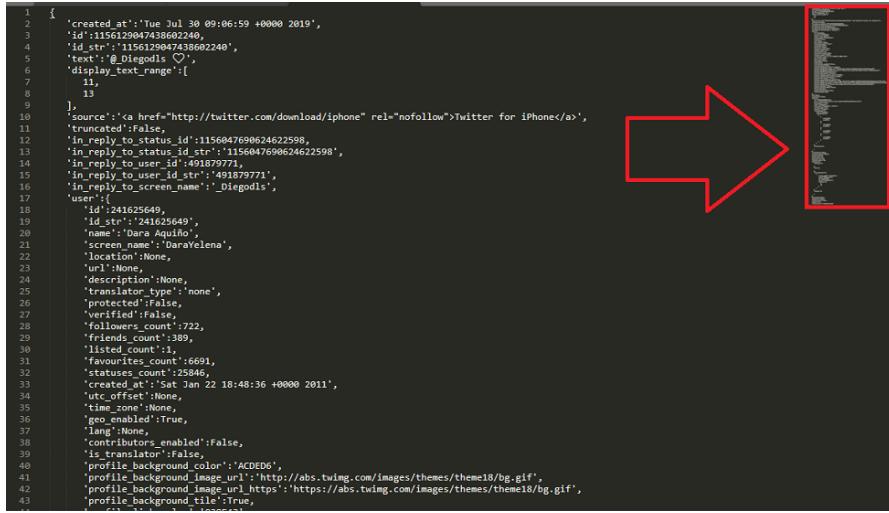
Para el proceso de accediendo a los datos se debe seleccionar una librería que permita la conexión a las API's de Twitter, haga uso de autenticación mediante el método OAuth, y soporte el uso de la API Streaming con el uso de geolocalización para la extracción de tweets de una localidad específica en base a coordenadas, en el presente será la ciudad de Guayaquil.

Con base en lo antes mencionado para el desarrollo de este proyecto, en primera instancia se hizo uso de la librería Twython, por la simplicidad de su objeto Json como respuesta.

## Librería Twython

Twython proporciona una forma fácil de acceder a los datos de Twitter. En el gráfico N. 31 se visualiza una parte de la estructura del objeto json obtenido como respuesta por cada tweet extraído por esta librería, el cual está compuesto por 129 líneas, 26 atributos que contienen una estructura de pares (clave, valor).

**Gráfico N. 31** Objeto Json obtenido de Twython de aproximadamente 129 lineas.



```
1  {
2      'created_at': 'Tue Jul 30 09:06:59 +0000 2019',
3      'id': 1156129047438602240,
4      'id_str': '1156129047438602240',
5      'text': '@_Diegolis <3>',
6      'display_text_range': [
7          11,
8          13
9      ],
10     'source': '<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>',
11     'truncated': False,
12     'in_reply_to_status_id': 115604769062462598,
13     'in_reply_to_status_id_str': '115604769062462598',
14     'in_reply_to_user_id': 491879771,
15     'in_reply_to_user_id_str': '491879771',
16     'in_reply_to_screen_name': '_Diegolis',
17     'user': {
18         'id': 1241625649,
19         'id_str': '1241625649',
20         'name': 'Dara Aquino',
21         'screen_name': 'DaraHelena',
22         'location': None,
23         'url': None,
24         'description': None,
25         'translator_type': 'none',
26         'protected': False,
27         'verified': False,
28         'followers_count': 722,
29         'friends_count': 389,
30         'listed_count': 1,
31         'favourites_count': 66801,
32         'statuses_count': 25846,
33         'created_at': 'Sat Jan 22 18:48:36 +0000 2011',
34         'utc_offset': None,
35         'time_zone': None,
36         'geo_enabled': True,
37         'lang': None,
38         'contributors_enabled': False,
39         'is_translator': False,
40         'profile_use_background_image': True,
41         'profile_background_image_url': 'https://abs.twimg.com/images/themes/theme18/bg.gif',
42         'profile_background_image_url_https': 'https://abs.twimg.com/images/themes/theme18/bg.gif',
43         'profile_background_tile': True,
44     }
45 }
```

Elaboración: Jiménez Cárdenas Edinson Andrés.

Fuente: Datos de la investigación.

En el transcurso de las pruebas y familiarización con el API Streaming, se comprobó que los tweets extraídos mediante el uso de esta biblioteca, no contenían el total de caracteres que a partir del 2017 permite la red social, los cuales son 280 al día de hoy. La presente biblioteca solo permitía obtener 140 caracteres, los cuales eran los establecidos por la red social hasta mediados del 2017. En el gráfico N. 32 se puede visualizar como los tweets que exceden los 140 caracteres quedan inconclusos mediante el uso de esta librería y registra el exceso de caracteres mediante la incorporación de puntos suspensivos, que denotan el faltante de caracteres. Esto se debe a que en el formato Json de la librería twython no se implementaron las actualizaciones que la red social Twitter implementó en su objeto Json como respuesta a cada petición que se le realiza.

### *Gráfico N. 32 Tweets extraídos mediante el uso de Twython.*

```
10
11 [],@paulmartinez29 @wilfridolaz Con este gobierno traidor que pisotea la constitución y no respeta la independencia de...
https://t.co/98NYZ9yYJt,jualomal2,Guayaquil
12
13 [],Sabe que es un drone? Cuanto es la autonomía? Suena bonito decirlo pero suena más lógico tener cámaras fijas no? Si...
https://t.co/IWU7nBSA5P,wgcv,Guayaquil-Ecuador
14
15 ['Sánchezjoselo'],@washingtonpost @dialogodeport @miguelycaza @CarlosVictorM ¿cuál es la nueva cuenta de #Sánchezjoselo por favor?. Te...
https://t.co/D0FYDeutES,thefloodgate,Guayaquil-Ecuador
16
17 [],"Primero me aseguraba traer un Reemplazo para ambos dinneno y nahuelpan, pero no...",DavoBravo14,Guayaquil
18
19 ["'OsbaldoLastra', 'EfectoBSC']","Yo sé que duele, yo sé que lloras...versión 2019 cantada por #OsbaldoLastra #EfectoBSC",DanielCalle76,
20
21 [],"Ja,ja,ja,ja, esto sí que causa risa y al mismo tiempo coraje ,decir que presentan los primeros resultados de "luch...
https://t.co/FLUtByj74j",perjudicado2010,
22
23 [],La fiscal subrogante dice que Topic no ha llegado a la fiscalía que es la "dueña"del caso y que esto no impide los...
https://t.co/amPq2v0JaK,perjudicado2010,
24
25 [],@Ecuulenin @jorgebg87 @JohnnyXavierSal si señor...estos son los fanáticos que hacen que la gente no vaya a los esta...
https://t.co/mo25CPMvdz,edudoradocolom,
26
27 [],"@barceblaccio Bien sacado la puta, eso es todo ojalá y la gente ya se sacuda y empieze a coger sin miedo a estos p...
https://t.co/01XXV3edUW",edyrivera74,"Guayaquil, Ecuador"
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

Para corregir el problema expuesto anteriormente, se realizó una investigación en proyectos de minería de datos usando el API de Twitter, y se precede a determinar que la biblioteca o librería idónea para acceder al API Streaming es Tweepy.

### **Uso de la librería Tweepy**

Tweepy devuelve un objeto json como respuesta mucho más amplio que el empleado por librería Twython, el cual se puede visualizar una pequeña parte en el gráfico N. 33 en el recuadro rojo, presentando una estructura mucho más extensa y detallada, permitiendo dar solución a la problemática de la extracción incompleta de tweets, proporcionando un atributo denominado *extended\_tweet* que almacena la versión completa del tweet en caso de que este exceda los 140 caracteres.

**Gráfico N. 33** Objeto Json obtenido de Tweepy consta aproximadamente de 606 líneas

```

1 Status._api<--tweepy.api.API object at 0x7fdf2284363b>
2     _json={
3         'created_at': 'Sun Jul 28 04:14:03 +0000 2019',
4         'id': 1155330551862678625,
5         'id_str': '1155330551862678625',
6         'text': 'Guardo cada detalle de las armas, pueden regalarme un chicle y lo guardaré .En la escuela se reian de mí por ser al... https://t.co/vnTOxvNCo',
7         'display_text_range':
8             [
9                 0,
10                140
11             ],
12         'source': '<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>',
13         'truncated': True,
14         'in_reply_to_status_id': None,
15         'in_reply_to_status_id_str': None,
16         'in_reply_to_user_id': None,
17         'in_reply_to_user_id_str': None,
18         'in_reply_to_screen_name': None,
19         'user': {
20             'id': 109399934916403715,
21             'id_str': '109399934916403715',
22             'name': 'Nuruka_QT',
23             'screen_name': 'Nuruka_QT',
24             'location': 'Guayaquil, Ecuador',
25             'url': 'https://www.instagram.com/nurukquinonez23_04/?hl=es-la',
26             'description': '♡ KIM NANIJOON \n♡ KIM SEOKJIN\n♡ MIN YOUNG\n♡ JUNG HOSEOK\n♡ PARK JIMIN\n♡ KIM TAEHYUNG\n♡ JEON JUNGKOOK\n♡\n        ♡ BTS♡',
27             'profile_image_url_type': 'normal',
28             'protected': False,
29             'verified': False,
30             'followers_count': 453,
31             'friends_count': 1127,
32             'listed_count': 0,
33             'statuses_count': 3712,
34             'created_at': 'Fri Feb 08 22:25:45 +0000 2019',
35             'utc_offset': None,
36             'time_zone': None,
37             'geo_enabled': True,
38             'lang': None,
39             'contributors_enabled': False,
40             'is_translator': False,
41             'profile_background_color': '000000',
42         }
43     }
44 
```

Elaboración: Jiménez Cárdenas Edinson Andrés.

Fuente: Datos de la investigación.

Adicionalmente su estructura json al ser muy bien dotada de datos referente al tweet extraído, permite mediante la incorporación de validaciones, determinar si el objeto extraído pertenece a una respuesta, un retweet, o sencillamente un tweet. Según (Twitter I., 2019), nos indica que la carga útil JSON de un solo Tweet puede tener más de 100 pares clave-valor.

## Exploración de datos

A continuación se listan los datos seleccionados de cada objeto json obtenido:

- **id\_str:** representa la cadena del identificador único para este tweet. La documentación de Twitter recomienda que las implementaciones deben usar este campo en lugar del entero grande que existe en el campo id. Debido a que el campo id contiene un número que es superior a 53 bits y algunos lenguajes de programación pueden tener dificultades/defectos de silencio al interpretarlo. Usar un entero de 64

bits como el existente en id\_str para almacenarlo como identificador es seguro.

- **created\_at:** Hora UTC cuando se creó este tweet.
- **location:** cuando está presente, indica que el tweet está asociado a un lugar (pero no necesariamente donde fue emitido).
- **text:** El texto real UTF-8 de la actualización de estado.
- **full\_text:** almacena todos los caracteres del tweet con un máximo de 280, debido a que si es retweet se trunca en 140 caracteres en el campo text.
- **in\_reply\_to\_status\_id:** Si el Tweet representado es una respuesta, este campo contendrá la representación entera del ID de autor del tweet original. Esto no siempre será necesariamente el usuario mencionado directamente en el tweet. Este campo no se almacena, solo se usa para validación del tipo de tweet.
- **is\_quote\_status:** Indica si se trata de un tweet citado.

El notebook de captura se construyó en Jupyter utilizando la librería Tweepy y el API Streaming de Twitter, la información es almacenada en archivos CSV para su posterior procesamiento, en este tipo de archivo se almacenaban 4 atributos de cada tweet (véase el gráfico N. 34). La ventana de captura se estableció desde las 9h00 a 13h00 y luego a las 18h00 a 23h00, tiempo según el cual se presenta mayor interactividad de los usuarios de Twitter (Formación Gerencial, 2018), el cual se considera un factor importante para la capturar de tweets en la ciudad de Guayaquil.

**Gráfico N. 34** Visualización de archivo CSV que almacena los tweets.

date	location	tweetID	tweetText	typeTweet
2019-08-26T08:09:00	GUAYAQUIL ECUADOR	1165898929394855937	@Joserandez1 Pon el Link para ir directamente...	respuesta
2019-08-26T08:09:10	NaN	1165898970759086086		retweeted
2019-08-26T08:09:33	Guayaquil - Ecuador	1165899066401681409	ATM señalizando Guayaquil #atminnova #roadmark...	tweet
2019-08-26T08:09:36	Ecuador	1165899078204612610	La mitad de una mentira no es la verdad... ❤\n...	tweet
2019-09-05T14:35:47	Guayaquil, Ecuador	1169620143179214849	@AparicioCC Eso pasa cuando pones a alguien q ...	respuesta

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

Los flujos de información extraídos son de datos públicos que fluyen a través de Twitter. Una vez que se establece una conexión a un punto final de transmisión, el API Streaming entrega un feed de Tweets, sin tener que preocuparse por los límites de las tasas como en el API REST. De acuerdo a la documentación oficial de la red social, los usuarios que realizan excesivos intentos de conexión o solicitud (sean estas exitosas o no) corren el riesgo de que su aplicación sea bloqueada automáticamente.

La recolección de datos realizada inicialmente se efectuaba en archivos CSV, lo cual no otorgaba un grado de automatización del proceso en la parte de la extracción de tweets, por lo tanto al presente proyecto se le incorpora una mejora sustancial, realizando el almacenamiento de los tweets extraídos; en una base de datos no relacional como Firebase (Google Developers, 2019).

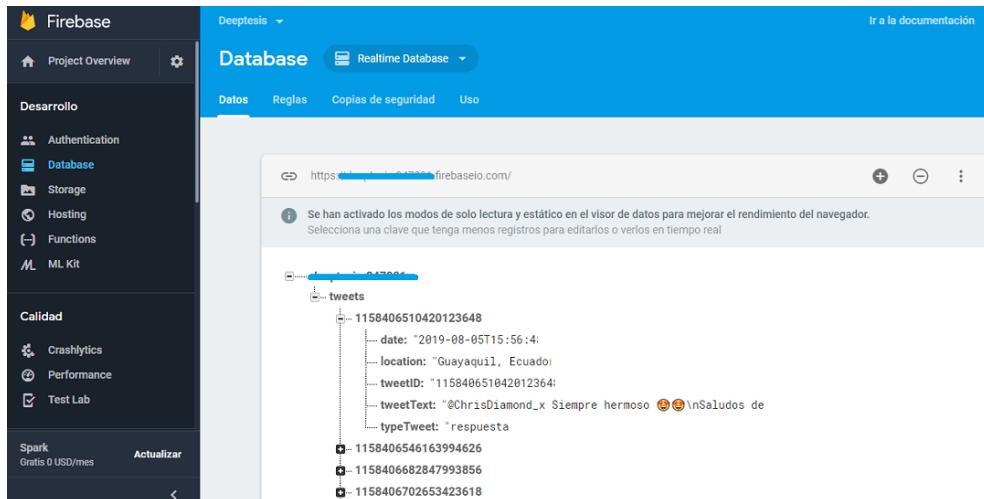
Firebase Realtime otorga almacenamiento y posterior consulta en tiempo real de los datos. Permitiendo el acceso a un conjunto de datos uniforme para cada usuario del notebook (archivo con código Python ejecutable en el entorno de desarrollo Jupyter). Erradicando el proceso manual de extracción de tweets por cuenta propia de cada usuario del notebook, centrando la atención del usuario solo en el uso del notebook que realiza el análisis de sentimiento en

español, y obtener la valoración emocional de los usuarios a un producto o servicio.

Debido a que ciertos campos almacenados al inicio del desarrollo de este proyecto no contribuían en gran medida fueron eliminados y reemplazados por campos que aportan información de mayor relevancia, y cuando se proceda al apartado de análisis de los datos puede otorgar mejores métricas que contribuyan al objetivo de este proyecto que es brindar apoyo en la toma de decisiones a los emprendedores y dueño de negocios de la ciudad de Guayaquil.

Se puede visualizar en el gráfico N. 35 que la estructura de almacenamiento de cada tweet en la base de datos Firebase, al ser NO-SQL; almacena los datos extraídos en formato JSON.

*Gráfico N. 35 Almacenamiento de tweets en Firebase.*



**Elaboración:** (Google Developers, 2019).

**Fuente:** (Google Developers, 2019).

Se almacena cada tweet extraído con un número de identificación único, el cual es provisto por la propia red social al asignarle un id único a cada tweet

que se emite en su plataforma. Cada tweet almacenado cuenta con cinco campos específicos, los cuales serán detallados a continuación:

- **date:** fecha de emisión del tweet.
- **location:** ciudad desde la cual fue emitido el tweet.
- **tweetID:** identificador único para cada tweet emitido en Twitter, consta de 19 dígitos.
- **tweetText:** tweet recuperado en su versión completa.
- **typeTweet:** campo que mediante validaciones determina si es tweet, retweet o respuesta.

### **Fase III. Preparación de los datos**

En base a la investigación realizada por (Sarmiento & Silva, Repositorio Institucional Universidad Distrital , 2017), la captura de datos realizada en la fase previa, tiene el propósito de brindar un acercamiento a la estructura de los flujos de información provenientes de la red social Twitter, para determinar de forma adecuada las técnicas de pre procesamiento y el modelo de datos que mejor se ajuste a los objetivos del caso de estudio. La fase de preparación de los datos tuvo en cuenta dos actividades principales, las cuales se describen a continuación:

#### **Estructuración de los datos**

La construcción de un modelo de análisis de sentimientos en español, orientado a determinar la polaridad de la opinión que los usuarios tienen sobre un determinado producto o servicio, requiere de un dataset (conjunto de datos) de conocimiento supervisado. Estos dataset en su mayoría contienen información clasificada de forma manual, que permita enseñar al modelo la forma correcta de evaluar los datos que se le suministren.

Conociendo todos estos por menores, se deja en claro que la captura final de los tweets públicos emitidos en la red social Twitter con geolocalización en la ciudad de Guayaquil, fue realizada por el lapso de cuatro meses, desde inicios del mes de mayo a finales del mes de agosto; permitiendo obtener un total de 65.000 tweets. Los cuales servirán para la parte de validación, en la cual se realiza una búsqueda específica de un producto o servicio por parte del usuario final del notebook. Obteniendo como resultado la polaridad expresada (positiva o negativa) mediante tweets.

Previamente para el apartado de entrenamiento del modelo de clasificación de sentimientos en español, se hizo uso de un dataset muy popular para análisis de sentimientos en español, el cual es de propiedad de la SEPLN, y su principal uso se da en el Taller de análisis de sentimientos en la SEPLN (TASS), que se lleva a cabo cada año.

Según (SEPLN, 2019), su corpus contiene más de 70 000 tweets, escritos en español por casi 200 personalidades y celebridades conocidas del mundo de la política, la economía, la comunicación, los medios de comunicación y la cultura, entre noviembre de 2011 y marzo de 2012. El conjunto de datos TASS es un corpus de textos (principalmente tweets) en el idioma español etiquetados para tareas relacionadas con el análisis de sentimientos. Se divide en varios subconjuntos creados para las diversas tareas propuestas en las diferentes ediciones a lo largo de los años.

Aunque el contexto de extracción tiene un sesgo enfocado en España, la nacionalidad diversa de los autores, incluyendo personas de España, México, Colombia, Puerto Rico, Estados Unidos y muchos otros países, hace que el corpus alcance una cobertura global en el mundo de habla hispana y sea

considerado ampliamente en proyectos que involucren análisis de sentimientos y clasificación de documentos en español.

El corpus del TASS descargado cuenta con más 70,000 documentos (textos cortos, comentarios o tweets), de los cuales 7,219 tienen el propósito de entrenar al sistema y se encuentra previamente clasificados en base a su sentimiento, los más de 62,798 restantes para probar la exactitud del algoritmo de clasificación de sentimientos (Librado, 2017, pág. 62).

Con base en lo antes expuesto se corrobora que el corpus está dividido en un conjunto de training (10%), que se proporciona a los participantes del TASS para facilitar el entrenamiento de los modelos, y un conjunto de test (90%) que sirve para la evaluación competitiva de los resultados de los diferentes experimentos llevados a cabo por los participantes.

**Gráfico N. 36** Dataset para el TASS emitido por la SEPLN.



**Dataset Download**

**General Corpus (2012)**

- Train set (tagged with entities, 5-level global and aspect-based sentiment and topics)
- Train set (tagged with entities, 3-level global and aspect-based sentiment and topics)
- Test set
- Test 1k subset
- Test set (tagged with 5-level global sentiment and topics)
- Test set (tagged with 3-level global sentiment and topics)
- Test 1k subset (tagged with 5-level global sentiment and topics)
- Test 1k subset (tagged with 3-level global sentiment and topics)
- Test Gold standard (for 5-level global sentiment)
- Test Gold standard (for 3-level global sentiment)
- Test 1k subset Gold standard (for 5-level global sentiment)
- Test Gold standard (for 3-level global sentiment)
- Test 1k subset Gold standard (for 3-level global sentiment)
- Test Gold standard (for topics, TASS 2012-2014, Task 2)
- User information (manually tagged with political orientation, TASS 2013, Task 3)

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

Este corpus al ser procesado en el apartado de entrenamiento con sus más de 7000 tweets clasificados nos otorga un accuracy (precisión) de 76% como se puede visualizar en el gráfico N. 37. Pero al ser evaluado para clasificar la polaridad existente en cada tweet emitido en la ciudad de Guayaquil demostró

no ser tan efectivo, otorgando un accuracy del 59%, que para fines del presente proyecto no se considera un valor óptimo.

**Gráfico N. 37** Training y test usando el dataset de la SEPLN.

```
57658/57658 [=====] - 31s 534us/step - loss: 0.6906 - acc: 0.7407
Epoch 27/30
57658/57658 [=====] - 29s 510us/step - loss: 0.6719 - acc: 0.7490
Epoch 28/30
57658/57658 [=====] - 29s 507us/step - loss: 0.6533 - acc: 0.7547
Epoch 29/30
57658/57658 [=====] - 29s 503us/step - loss: 0.6364 - acc: 0.7623
Epoch 30/30
57658/57658 [=====] - 29s 500us/step - loss: 0.6140 - acc: 0.7697

Ahora evaluamos con los datos de test
```

In [74]: loss, precision\_test = modelo\_sentimiento.evaluate(X\_test, y\_test)  
precision\_test

Out [74]: 0.5927159209446571



**Elaboración:** Jiménez Cárdenas Edinson Andrés.  
**Fuente:** Datos de la investigación.

Debido a que los términos expuestos en la licencia de uso del dataset TASS no se ajustan a los fines del proyecto, en la cual se menciona que el Licenciatante (SEPLN) otorga al usuario final el derecho de usar el Conjunto de datos, para su propio uso interno y no comercial y solo con fines de investigación científica. En conjunto con los resultados obtenidos en las pruebas de clasificación de sentimientos, se procede a descartar el uso de este dataset.

Por consiguiente se optó por la creación de un dataset de entrenamiento propio para el presente proyecto, con datos que se ajusten al léxico empleado en la presente ciudad. Para la creación del corpus de entrenamiento, se tomó un total de 9000 tweets extraídos en la etapa inicial del proyecto, cuando los tweets eran almacenados en archivos CSV, los cuales fueron clasificados a mano por el autor del presente proyecto; asignando una valoración positiva o negativa en base al sentimiento encontrado en cada tweet. Formando un

dataset con 3000 tweets con su respectiva polaridad (sentimiento). Todo esto, con la finalidad de otorgar una mayor eficacia al algoritmo de clasificación de sentimientos en español, y debido a que el dataset utilizado al inicio del proyecto, perteneciente a la Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN), no se ajustaba al léxico empleado en los tweets de validación, y el algoritmo no predecía con eficacia los sentimientos expresados en los tweets emitidos en la ciudad de Guayaquil.

Este conjunto supervisado fue almacenado en un archivo .csv, que posteriormente fue cargado al modelo, a continuación se puede visualizar en el gráfico N. 38 una pequeña parte del corpus.

**Gráfico N. 38** Conjunto de conocimiento supervisado.

68	negativo Se retrasa el Pleno AsambleaEC
69	negativo Se demoran en la entrega del pedido
70	positivo Tuiteando desde Bucanero
71	positivo Muchas gracias por todas las felicitaciones del fondo del alma Gracias Obrigado
72	positivo Marcelino Iglesias en el congrespsc en la política como en la vida hace falta mucho trabajo y un
73	positivo En el Comite Regional de la Prefectura trabajando por un futuro mejor
74	positivo organizaciones agrarias aplauden recuperar el M de Agricultura para dar mayor atencion al sector
75	positivo Viteri le ha propuesto a Lenin pacto ante Asamblea Por lo demás reedicion del debate electoral
76	negativo No imposible la wifi va y viene esto es real siempre
77	negativo Aqui frente al Hospital con un dolor de pies importante
78	positivo Espero que podamos arreglar definitivamente este asunto
79	positivo Vale ya casi casi veo que cada minuto somos mas
80	positivo que buen servicio MARAVILLOSO GRACIAS
81	positivo Tenemos enormes fortalezas y debemos ser conscientes de ellas investidura
82	negativo Malas noticias ADN cierra Un abrazo para los compañeros de la redaccion y mucha suerte
83	negativo La nostalgia es como una droga Te impide ver las cosas tal como son The Walking Dead
84	negativo Todo va a acabar fatal porque la gente no creo que siente ni a cenar
85	positivo me encantan los jugos naturales de la san martin
86	positivo Duran asegura que de ninguna manera el Govern subira los precios
87	positivo Preparada para la cena con los compischapa y pintura en minutos hoy he batido todos los records
88	positivo Que feliz me hace ver que less gusta tanto el nuevo video Esta hecho para ustedes Los quierooooo
89	negativo Los dos han puesto de su parte pero no pudieron lograrlo
90	negativo A quien no le guste que no lo vea

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

La clasificación manual del corpus que contiene: el sentimiento y el tweet, fue un proceso complejo que involucró tiempo, conocimiento y búsqueda de cuáles son las palabras con las que los usuarios describen sus sentimientos, deseos y necesidades, con la finalidad de dar un criterio acertado de la semántica del tweet.

## **Formateo de los datos**

Los procesos de pre-procesamiento de datos son tan importantes como la misma extracción y almacenamiento de datos, estos procesos involucran una gran cantidad de recursos para llevar a cabo esta actividad, por lo tanto se emplearon diferentes técnicas y herramientas para limpiar y formatear la información del conjunto de conocimiento supervisado, y los nuevos tweets que serán evaluados por el modelo de clasificación de sentimientos en español.

Debido a la utilización de textos de Twitter, la tarea se hace más compleja ya que no siempre se trata con textos gramaticales tanto a nivel léxico como sintáctico, es decir, hay ciertas características peculiares de Twitter como palabras alargadas, eliminaciones de caracteres, emoticonos, urls, hashtags, etc., que dificultan el tratamiento lingüístico de textos (González J. , 2017, pág. 2).

Para llevar a cabo este proceso se emplearon diferentes funciones y el uso del módulo *Re* en el lenguaje Python, permitiendo aplicar varias técnicas de data cleaning (limpieza de datos) a partir de un corpus de conocimiento. A continuación se mencionan los diferentes procesos de limpieza aplicados en este proyecto:

- Extraer hashtags
- Remover los saltos de líneas
- Eliminar tildes
- Remover URL's
- Remover usuarios
- Remover números
- Remover caracteres especiales
- Eliminación de letras repetidas

Según (Conde, 2018), la clave en estos desarrollos, donde el algoritmo a aplicar esta probado en su potencia y capacidad, está en la previa, el llamado pre-procesamiento de datos que “limpia” el input que se va analizar y se aclaran las decisiones aplicadas sobre el set de datos.

#### **Fase IV. Modelado**

Twitter se ha convertido en una fuente que alberga una gran cantidad de datos de forma distribuida, por ello, proporciona un amplio abanico de posibilidades para realizar investigaciones en campos de PLN como sentiment analysis. Sentiment analysis o opinion mining, es un área de investigación dentro de PLN cuyo objetivo es identificar la emoción o polaridad subyacente en un determinado documento, frase o aspecto; empleando algoritmos de machine learning (González J. , Aprendizaje profundo para el procesamiento del lenguaje natural, 2017, pág. 53).

En concreto, Machine Learning se define como el subcampo de la inteligencia artificial que proporciona a los ordenadores la capacidad de aprender sin ser explícitamente programados, es decir, sin que necesiten que el programador indique las reglas que debe seguir para lograr su tarea sino que las hace automáticamente.

Generalizando, se puede decir que Machine Learning consiste en desarrollar para cada problema un “algoritmo” de predicción para un caso de uso particular. Estos algoritmos aprenden de los datos con el fin de encontrar patrones o tendencias para comprender qué nos dicen los datos y de esta manera construir un modelo para predecir y clasificar los elementos (Torres J. , 2018).

## Análisis de sentimientos

El análisis de sentimientos o emociones es algo con lo que sueñan muchos profesionales de marketing, y es que evaluar la opinión pública sobre un evento o producto a través del análisis de datos, en una escala que ningún humano podría alcanzar; le da a su equipo la capacidad de averiguar lo que realmente piensa la gente.

Actualmente la mayor barrera para la adopción de herramientas de análisis de sentimientos es la falta general de conocimiento sobre qué hacen estas herramientas, cómo funcionan y su posible efecto (Pérez, 2019).

El análisis de sentimiento en inteligencia artificial se puede realizar mediante el uso de sistemas basados en redes neuronales o sistemas de clasificación, como puede ser en el presente caso el uso de un clasificador bayesiano (Vilanova, 2019).

- **Datos clasificados.** El primer paso es disponer de lo que denominamos dataset (conjunto de datos) con información de tweets sobre la temática que estamos analizando clasificados o puntuados según su sentimiento. Es decir, partimos de ejemplos para lo que denominamos train (o entrenamiento) de nuestro algoritmo de inteligencia artificial.
- **Datos pre-procesados.** Es ideal en problemas donde en realidad se evalúa el lenguaje natural (los tweets) eliminar información que no es relevante, típicamente puntos, comas, puntos y comas, paréntesis, pronombres y artículos, etc. para que el sistema reciba como entrada verbos, adjetivos, adverbios, etc. Información específica. Adicionalmente otras técnicas se realizaron para preparar la información antes de enviarla a nuestro algoritmo de clasificación de sentimientos y la red neuronal, como vectorizaciones y normalizaciones.

- En la fase de train o entrenamiento el sistema aprende de los tweets que le enviamos, los cuales son positivos o negativos.
- A partir de aquí, con ajustes al dataset y algoritmo, conseguimos lo que se denomina modelo que probaremos con otro conjunto de validación de tweets.
- Una vez se ha llegado mediante iteraciones al sistema idóneo, se puede utilizar nuestro modelo óptimo (nunca perfecto) para poder enviar los futuros tweets y analizar su sentimiento, lo que denominamos predicciones.
- Estos análisis serán almacenados en archivos CSV para poder extraer conclusiones.
- Adicionalmente los nuevos tweets clasificados y revisados se podrán añadir al dataset de train (entrenamiento) para mejorar nuestro modelo basado en inteligencia artificial.

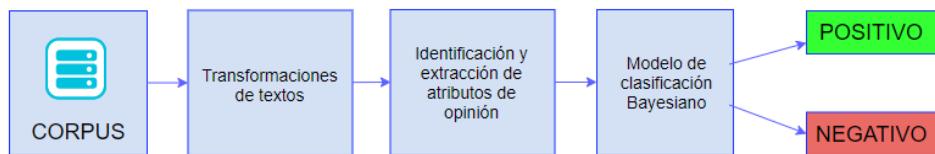
El análisis en español es realmente complejo, pues hoy en día no existen más que conjuntos de palabras clasificadas según el sentimiento o conjuntos de tweets clasificados específicamente para el ámbito político.

Tras un exhaustivo análisis de las librerías disponibles para el lenguaje de programación Python para el análisis de sentimientos, no se ha encontrado ninguna que tenga una función de análisis en de tweets en el idioma español.

Por obvias razones en primera instancia del proyecto se optó por utilizar librerías de traducción de idiomas, para traducir los tweets al idioma inglés y proceder a asignarle un valor relacionado a su polaridad. Pero debido al alto consumo de tiempo para realizar este procedimiento se procedió a reestructurar este proceso del proyecto. Para el presente proyecto que involucra un análisis de sentimiento orientado al ámbito comercial, se procede a la creación de un corpus desde cero (véase en la Fase III de la metodología).

El objetivo del análisis de datos realizado en este proyecto es determinar el sentimiento positivo o negativo de tweets referente a un producto o servicio ingresado por el usuario final, por lo tanto se seleccionó el algoritmo de clasificación bayesiano Bernoulli porque se enfoca en procesos de clasificación, partiendo de un dataset supervisado para entrenar el modelo y de esta forma predecir comportamientos de nuevos datos, ajustándose a los objetivos propuestos en esta investigación (Sarmiento & Silva, Repositorio Institucional Universidad Distrital , 2017).

**Gráfico N. 39** Proceso de análisis de sentimiento



Elaboración: Jiménez Cárdenas Edinson Andrés.

Fuente: (Dubiau, 2014).

### Selección de la técnica de modelado

Según (Rivera & Villavicencio, Pontificia Universidad Católica de Valparaíso, 2017), el modelo de Naive-Bayes (Bayes ingenuo) es una de las formas más tradicionales y sencillas de las redes bayesianas, cuya investigación surge en la década de 1950; este modelo ha presentado éxito considerable en su aplicación en temas relacionados a los problemas de clasificación. Este modelo es llamado ingenuo debido a que asume que las palabras son condicionalmente independientes entre sí dada una cierta clase. Esto resulta ser falso dentro del ámbito de la clasificación de textos, ya que la independencia condicional no se puede aplicar a la aparición de palabras en los documentos. El clasificador utiliza el conjunto de datos de entrenamiento para estimar la probabilidad de pertenencia a una clase dada la frecuencia de palabras del documento de cada instancia del conjunto de entrenamiento. Esta

probabilidad se calcula mediante el teorema de Bayes con una leve variación, dada su naturaleza ingenua y asumiendo la condicionalidad independiente entre palabras.

En el presente proyecto se utilizará el modelo de clasificación BernoulliNB, el cual es un clasificador Bayes ingenuo para modelos multivariantes. Al igual que MultinomialNB, este clasificador es adecuado para datos discretos. La diferencia es que mientras MultinomialNB trabaja con conteos de ocurrencias, BernoulliNB está diseñado para características binarias/booleanas.

## **Construcción del modelo**

### **Construcción del modelo de clasificación de sentimientos**

Una vez que se tiene definida las características, se procede con la construcción de un clasificador para tratar de predecir la categoría de una publicación. Se empleará el uso de un clasificador ingenuo de Bayes, específicamente BernoulliNB, que gracias al uso de la librería scikit-learn, proporciona una buena línea de base para la ejecución de este proyecto. Aunque la librería incluye varias variantes de este clasificador como es la variante multinomial. Para los presentes fines se usará BernoulliNB.

La construcción del modelo de clasificación de sentimientos en español se hizo utilizando Pipeline, para facilitar el trabajo del clasificador vectorizador y transformador. Scikit-learn proporciona una clase Pipeline que se comporta como un clasificador compuesto, tal como se puede observar en el gráfico N. 40.

### **Gráfico N. 40** Parte del código del algoritmo de clasificación

```
# Crea el pipeline junto con el clasificador
text_clf = Pipeline([('vect', CountVectorizer()),
                     ('tfidf', TfidfTransformer(use_idf=False)),
                     ('clf', BernoulliNB()),
                     ])

from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score

def evaluar_modelo():
    print('***** EVALUAR CORPUS *****')
    polaridad, texto = cargar_corpus('corpus_gye/corpus_ok2.csv')
    # preparing data for split validation. 60% training, 40% test
    X_train, X_test, y_train, y_test = train_test_split(texto, polaridad, test_size=0.1, random_state=10)

    classifier = text_clf.fit(X_train,y_train)
    print("Score:", classifier.score(X_test,y_test))
    predicted = classifier.predict(X_test)
    print (classification_report(y_test,predicted))
    print ("Nivel de predicción {:.2%}".format(accuracy_score(y_test,predicted)))
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Se espera que este método sea llamado de forma consecutiva en diferentes partes del dataset para implementar el aprendizaje fuera del núcleo o en línea. Esto es especialmente útil cuando todo el conjunto de datos es demasiado grande para caber en la memoria a la vez.

### **Construcción de la red neuronal para clasificar por sectores/industrias**

Se empieza por tener una capa de neuronas de entrada donde alimenta sus vectores de características y los valores fluyen hacia una capa oculta. En cada conexión, está alimentando el valor hacia adelante, mientras que el valor se multiplica por un peso y se agrega un sesgo al valor. Esto sucede en cada conexión y al final llega a una capa de salida con uno o más nodos de salida (Código fuente, 2018).

### **Gráfico N. 41** Entrenamiento de la red neuronal

```
X = np.array(training)
y = np.array(output)

start_time = time.time()
#base 20/ ultima 80
train(X, y, hidden_neurons=10, alpha=0.1, epochs=100000, dropout=False, dropout_percent=0.2)

elapsed_time = time.time() - start_time
print ("processing time:", elapsed_time, "seconds")
```

Training with 10 neurons, alpha:0.1, dropout:False  
Input matrix: 193x603 Output matrix: 1x8  
delta after 10000 iterations:0.0020727063880344116  
delta after 20000 iterations:0.0014273074951094615  
delta after 30000 iterations:0.0011510241283886476  
delta after 40000 iterations:0.0009889917611635207  
delta after 50000 iterations:0.0008795690597870302  
delta after 60000 iterations:0.0007994103183379701  
delta after 70000 iterations:0.0007374801172901102  
delta after 80000 iterations:0.0006878006379426522  
delta after 90000 iterations:0.0006468173451536323  
delta after 100000 iterations:0.0006122676261427977  
saved synapses to: synapses.json  
processing time: 95.19737124443054 seconds

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

El archivo synapse.json contiene todos los pesos sinápticos, esto es nuestro modelo. Posteriormente se requiere llamar a la función classify () para realizar la clasificación de los tweets por sectores una vez que se han calculado los pesos de sinapsis.

### **Fase V. Evaluación**

El modelo toma como factor clave el sentimiento del tweet presente en el conjunto de datos y arroja un nivel de predicción del 77,69% luego de varios entrenamientos (Ver gráfico N. 42), con un dataset supervisado que posee inicialmente 1500 tweets. La estructura de este modelo se considera robusta, de tal forma que permite clasificar textos en un grado significativo de precisión.

**Gráfico N. 42** Nivel de predicción del algoritmo de clasificación.

```
In [11]: evaluar_modelo()
("***** EVALUAR CORPUS *****
Score: 0.7768595041322314
      precision    recall   f1-score   support
negativo       0.83     0.77     0.80      70
positivo       0.71     0.78     0.75      51
accuracy        -         -     0.78     121
macro avg       0.77     0.78     0.77     121
weighted avg    0.78     0.78     0.78     121
Nivel de predicción 77.69%" data-bbox="338 341 531 356" style="border: 2px solid red; padding: 2px; display: inline-block; margin-right: 10px;">
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Se realizaron diversas pruebas para la comparación con el modelo de Maquinas de vectores de vectores el cual denotó un accuracy del 80%, pero en las pruebas las predicciones generadas por este modelo no fueron las esperadas, clasificando correctamente 6 de cada de 10 tweets. Por lo tanto se considera correcto el uso del modelo de clasificación bayesiano el cual nos otorga unas predicciones correctas en 8 de cada 10 tweets.

## Revisión e integración

Una vez construido el modelo de análisis de datos, realizado la validación y correcta predicción del mismo, se procede a construir los notebooks que generan la información respectiva que brinde apoyo en la toma de decisiones al usuario final y permita visualizar los patrones existentes en las opiniones de un producto o servicio.

La delimitación de la experiencia de los usuarios finales en el uso de notebooks, determino que la construcción de los mismos esté enfocada en la practicidad y con solo dar clic en la opción *run* (ejecutar), funcione perfectamente sin realizar cambios en el código, solo ingresar los datos de

interés que el sistema requiere para realizar una búsqueda específica de un producto o servicio de interés para el usuario final.

La investigación realizada en el presente trabajo, permitió identificar la necesidad de construir 4 notebooks con las características descritas en el cuadro N. 12.

**Cuadro N. 12** Componentes y funcionalidad de notebooks del proyecto.

Componente	Ubicación	Funcionalidad
Extracción de tweets	Notebook: <b>Downloading_tweets</b>	Permite conexión con el API Streaming para extraer tweets emitidos en la ciudad de Guayaquil, en tiempo real. Realizando el almacenamiento de los tweets en la base de datos no relacional Firebase.
Construcción y entrenamiento del modelo de clasificación	Notebook: <b>Sentiment_model</b>	Construcción del algoritmo y posterior lectura del dataset supervisado para entrenar el algoritmo de clasificación de sentimientos.
Búsqueda de producto o servicio en tweets y clasificación de sentimientos	Notebook: <b>Analisis_sentimientos</b>	Permite realizar búsquedas en el dataset de tweets sobre un producto o servicio de interés del usuario final. Asignar una clasificación positiva o negativa, teniendo en cuenta la semántica del tweet.
Visualización de tweets por categorizados por sector/industria.	Notebook: <b>clasificacion_sector</b>	Permite visualizar la clasificación de todos los tweets por categorías. Permitiendo tener un análisis general de qué sectores tienen mayor actividad mediante el uso de una red neuronal. Generando información gráfica que brinde apoyo en la toma de decisiones de los emprendedores y dueño de negocios.

**Nota:** Descripción de los notebooks del presente proyecto.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

En lo referente a la identificación de patrones comerciales, el presente proyecto se centra en la búsqueda de patrones de consumo, detectados en los tweets. Se seleccionaron dos procesos a nivel organizacional que permiten determinar la opinión de un producto o servicio en Twitter, los cuales a su vez son soportados por los notebooks creados, los cuales se encargan de presentar la opinión emitida referente al producto o servicio, y categorizar los tweets por sectores/industrias, basándose en el contexto de los mismos y patrones de consumo.

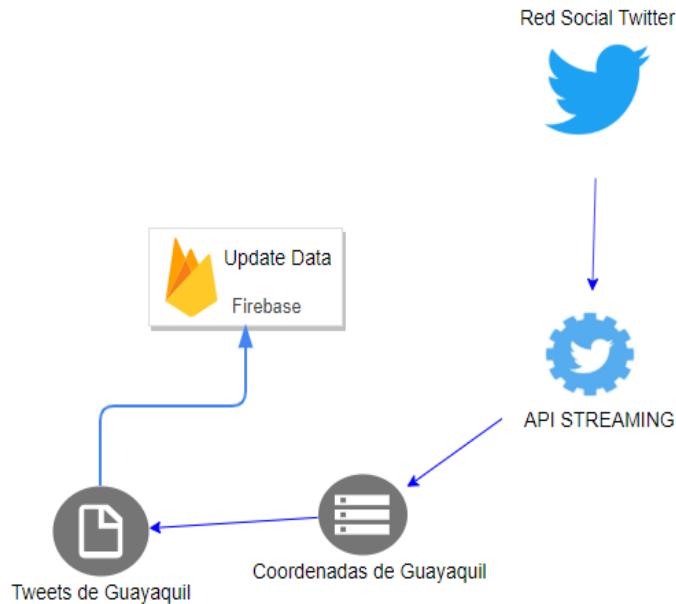
### **Arquitectura de los notebooks**

El proyecto está dividido en tres notebooks principales que hacen uso de los datos de la red social Twitter, el cuarto es el notebook de creación del modelo de clasificación de sentimiento en español.

### **Arquitectura del notebook de extracción de datos de la red social Twitter.**

En el gráfico N. 43 se visualiza la arquitectura implementada en la construcción del primer notebook, el cual es el encargado de realizar la extracción de tweets emitidos en la ciudad de Guayaquil en tiempo real. Almacenando esta información en Firebase y respetando la privacidad de sus usuarios no se almacena el Identificador de usuario o su nombre en la red social.

**Gráfico N. 43** Arquitectura notebook extracción de tweets.



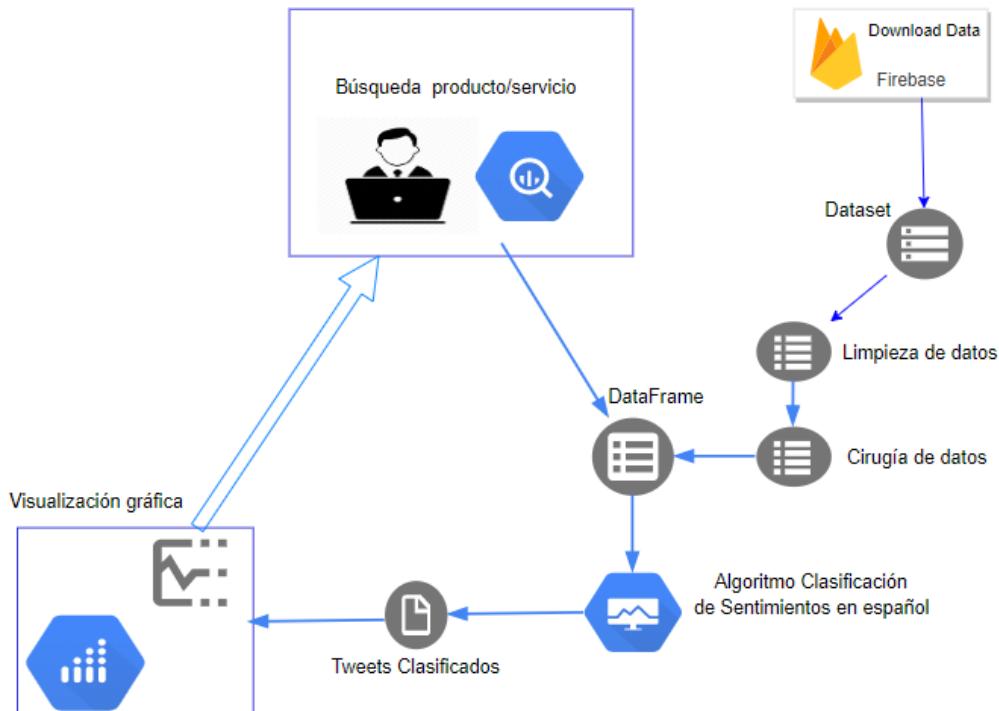
**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

### **Arquitectura del notebook de búsqueda de productos o servicios y aplicación de análisis de sentimientos en español.**

El siguiente notebook hace uso de los datos almacenados en Firebase Realtime, para aplicar las funciones de limpieza y cirugía a los datos, permitiendo tener un dataset listo para aplicar la función de clasificación de sentimientos a cada tweet encontrado en el dataset. Previamente se solicita al usuario que ingrese el nombre de un producto o servicio del cual desee obtener información mediante su búsqueda en el dataset y la polaridad obtenida mediante el análisis de sentimiento.

**Gráfico N. 44** Arquitectura notebook de búsqueda de producto o servicio.



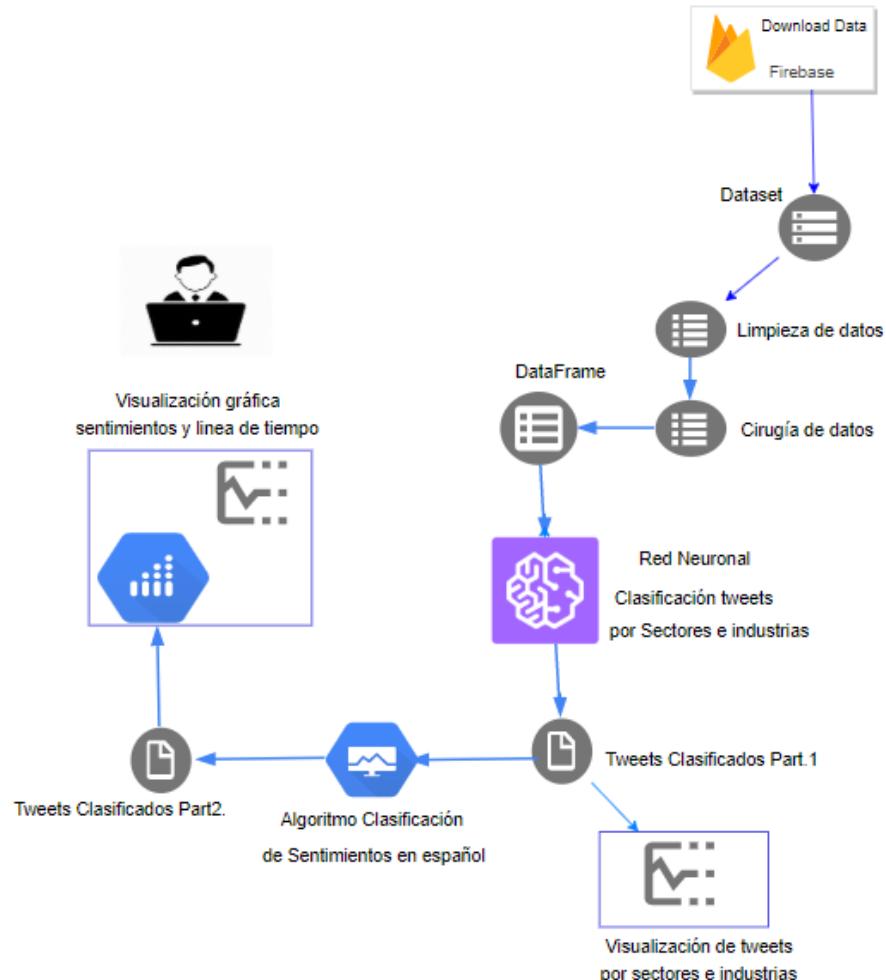
**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

### Arquitectura del notebook de clasificación de tweets por sectores e industrias.

El siguiente notebook hace uso de una arquitectura un poco más compleja, permitiendo la clasificación de los diferentes tweets por sectores, siempre y cuando se encuentren en estos tweets patrones de consumo que haga referencia a un sector o industria. Este objetivo se logra gracias al uso de una red neuronal dotada de patrones que permiten realizar esta clasificación, brindando información como que industrias son las que poseen mayor interactividad con los usuarios, tipo de polaridad encontrada en esta industria, gráficos en línea de tiempo de índices de consumo, etc.

**Gráfico N. 45** Arquitectura notebook de clasificación por sectores e industrias.



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

## Presentación de resultados

A continuación se hace una presentación de los principales gráficos e información resultante de la ejecución de los notebooks, generando información útil en la toma de decisiones a través de sus dos notebooks más importantes para el usuario final, clasificación de sentimientos y el de clasificación de tweets por sectores, basándose en patrones comerciales previamente prescritos en el dataset de entrenamiento.

**Gráfico N. 46** Ciencia de datos en la toma de decisiones



**Elaboración:** (Nube colectiva, 2018).  
**Fuente:** (Nube colectiva, 2018).

### **Módulo de clasificación de sentimientos**

Este módulo cuenta con una de las funciones más importantes del proyecto que corresponde a la clasificación de tweets (positivo o negativo) de acuerdo al modelo de clasificación de sentimientos construido. Los pasos seguidos para alcanzar este objetivo fueron descritos en la fase IV.

Para la construcción de las gráficas del módulo se utilizaron las librerías NumPy y Matplotlib de Python, cuyo uso se enfoca en la construcción de aplicaciones de ciencias matemáticas e ingeniería, porque aprovechan la velocidad de procesamiento generada por el alto nivel de acoplamiento de las librería NumPy para números y la librería Matplotlib para la representación gráfica (Sarmiento & Silva, 2017).

**Gráfico N. 47** Flujo de actividades módulo de clasificación de tweets por su polaridad



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

El notebook a continuación descrito permite extraer todos los datos almacenados en Firebase y luego realizar una búsqueda de productos o servicio en el dataset para identificar las tendencias existentes que tienen relación con el producto o servicio buscado. La visualización de gráficos se hará efectiva siempre y cuando existan tweets con dicha mención para su posterior procesamiento y visualización. A continuación en el gráfico N. 48 se puede visualizar como ejemplo el ingreso de un producto popular en nuestra ciudad para su posterior análisis.

## *Gráfico N. 48* Búsqueda en el dataset de validación

### Paso 6: Búsqueda por productos y servicios

- Por favor no escribir cosas no relacionadas o el Kernel no podrá generar una búsqueda acertada.

```
In [27]: inputSearch = input("Ingrese Su Busqueda: ").lower()
#inputSearch = str(input("Ingrese Su Busqueda: "))
filtro = df[df['tweetFinal'].str.contains(inputSearch, case=False)]
len(filtro)

Ingrese Su Busqueda: cerveza

Out[27]: 81
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

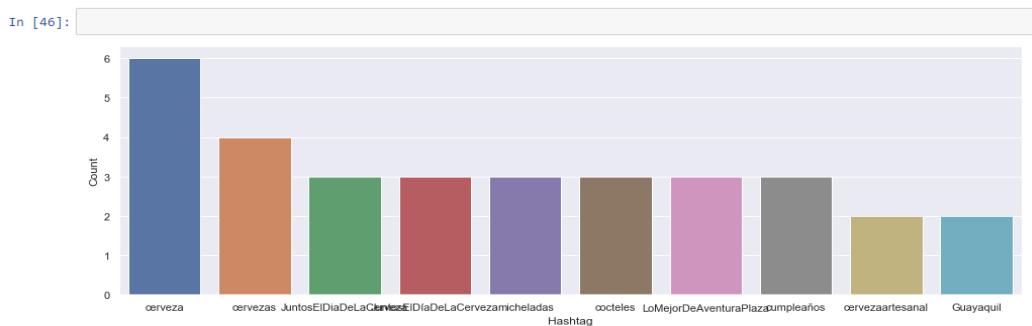
**Fuente:** Jiménez Cárdenas Edinson Andrés.

A continuación se visualiza un gráfico de barras que permite observar las tendencias relacionadas con el producto o servicio buscado.

Esta información es muy importante en caso de tener un negocio con el producto o servicio en cuestión, debido a que permite identificar y sumarse a estas tendencias para tener una mayor interactividad con los usuarios y ganar más seguidores en la red social. También existe la visualización de este mismo gráfico en su connotación negativa que permite evitar sumarse en tendencias que generan repudio u odio en cuanto al servicio o producto en cuestión. En el gráfico N. 49 se puede corroborar lo antes expuesto.

## *Gráfico N. 49* Tendencias relacionadas al producto o servicio

### Paso 8: Visualización de tendencias más frecuentes por sentimientos.



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación

Como siguiente visualización tenemos el gráfico N. 50, el cual genera una nube de palabras que presenta un gráfico que permite identificar que palabras fueron de mayor mención en cada tweets que contenía la presencia del nombre del producto o servicio buscado. Permitiendo obtener información respecto a qué piensa la gente cuando menciona este producto. Proporcionando información para que los emprendedores o dueños de negocios la analicen y puedan detectar productos secundarios ligados a la compra del producto principal. Estableciendo un punto de partida en la toma de decisiones acerca de incorporar nuevos productos a su stock ya existente.

**Gráfico N. 50** Nube de palabras con mayor índice de frecuencia.



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

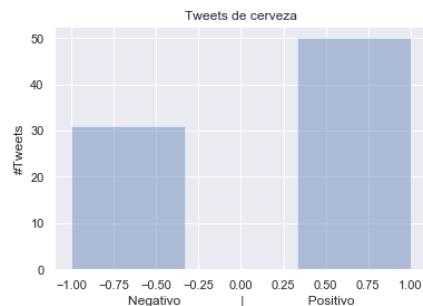
Siguiendo el flujo de ejecución del notebook se genera el gráfico N. 51, el cual es el último gráfico del primer notebook. Obtenemos un gráfico de clasificación final de sentimientos. Que nos permite evidenciar el nivel de aceptación o rechazo sobre un producto o servicio en cuestión, todo esto en base a la opinión solo de los tweets extraídos en esta red social y en la presente ciudad. No se debe aceptar como un resultado definitivo, pero si como un punto de

partida que permita a los emprendedores y dueños de negocios plantear hipótesis, con respecto a su toma de decisiones. El presente gráfico nos permite visualizar la cantidad de tweets analizados para este proceso y su respectiva tendencia en cuanto a la apreciación textual emitida en tweets.

**Gráfico N. 51** Valoración por sentimientos de un producto o servicio con base en el análisis de opiniones del dataset.

**Paso 10: Valoración final de sentimientos de tweets.**

```
In [70]: sentimient = filtrado["prediction"]
user_id = inputSearch
grafico_sentimiento = anl_tweets(sentimient, user_id)
```



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

El diseño de los gráficos presentados en este proyecto se ha realizado gracias a las librerías de visualización de datos estadísticos Seaborn y Matplotlib. Las cuales han permitido la representación gráfica de información generada en este proyecto.

### Módulo de clasificación de tweets por sectores

El presente notebook hace uso de técnicas de procesamiento de lenguaje natural como lo es stemming, vectorización de palabras, bag of words, red neuronal para la clasificación de tweets por sectores, con base en patrones prescritos en el dataset de entrenamiento.

Tabla de índice del notebook N. 4:

1. Cargar las bibliotecas necesarias
2. Red neuronal
3. Cargar datos de Firebase
4. Data cleaning (limpieza de los datos)
5. Aplicación de red neuronal para clasificar tweets por sectores
6. Importar el modelo para clasificación de sentimientos
7. Comparar el sentimiento entre los diferentes sectores
8. Visualizar sentimiento a lo largo del tiempo por las sectores
9. Almacenar información en archivo pdf

A continuación se describen los resultados a obtener en el presente notebook, y una forma de interpretar dichos gráficos. Luego que se ha importado las respectivas librerías, se ha importado el modelo y se ha aplicado el procesamiento a los datos. Se podrá obtener una categorización de tweets por diferentes sectores e industrias véase el grafico N. 52.

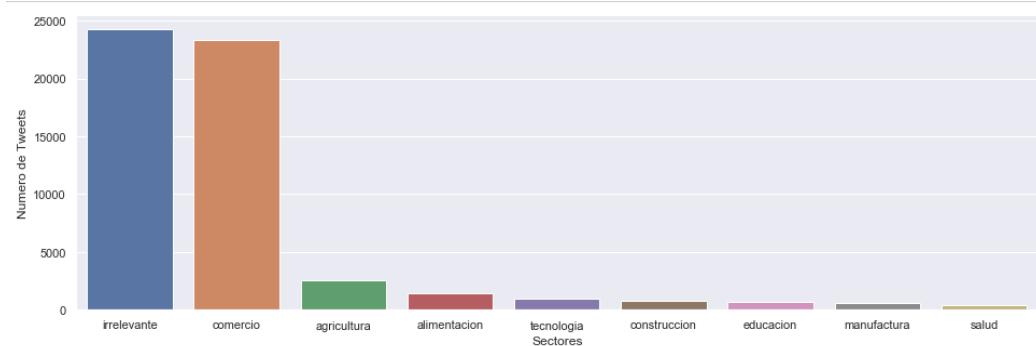
Los sectores e industrias fueron seleccionados con base en la pregunta N. 2 de la encuesta N. 2. Los cuales son los siguientes:

- Industrias manufactureras
- Sector comercial
- Sector agrícola
- Industria de la construcción
- Sector de la salud
- Sector de Tecnologías
- Industria alimentaria
- Sector educativo.

Permitiendo a los usuarios una visualización rápida respecto a qué sector tiene mayor cantidad de actividad en la presente red social. Estableciendo un punto inicial en la toma de decisiones en caso de emprender en un sector específico,

corroborando que en base a un análisis de 65.000 opiniones procesadas, cierta cantidad hizo hincapié de consumo en un producto o servicio de un determinado sector e industria.

**Gráfico N. 52** Gráfico de barras por sectores



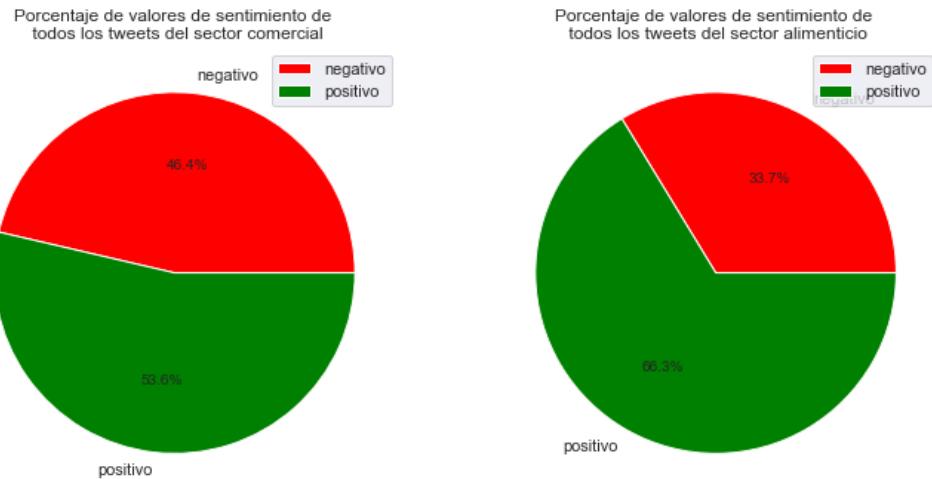
**Nota:** Visualización rápida de interactividad de usuarios por sectores

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

Posteriormente se procede a aplicar el modelo de clasificación de sentimientos desarrollado para el presente proyecto. Para obtener información más detallada acerca de las opiniones categorizadas por sector/industria. Permitiendo obtener información de la polaridad detectada, para identificar cuáles son los sectores que tienen una mayor aceptación de productos o servicio. En cuales existe mayor incidencia negativa con base a comentarios analizados, y cuáles son los sectores donde abunda el positivismo de acuerdo a la opinión receptada y analizada. Los gráficos N. 53 se generaran uno por cada sector, en total 4 gráficos, cada uno contiene información acerca de dos industrias.

**Gráfico N. 53** Visualización de polaridad en tweets



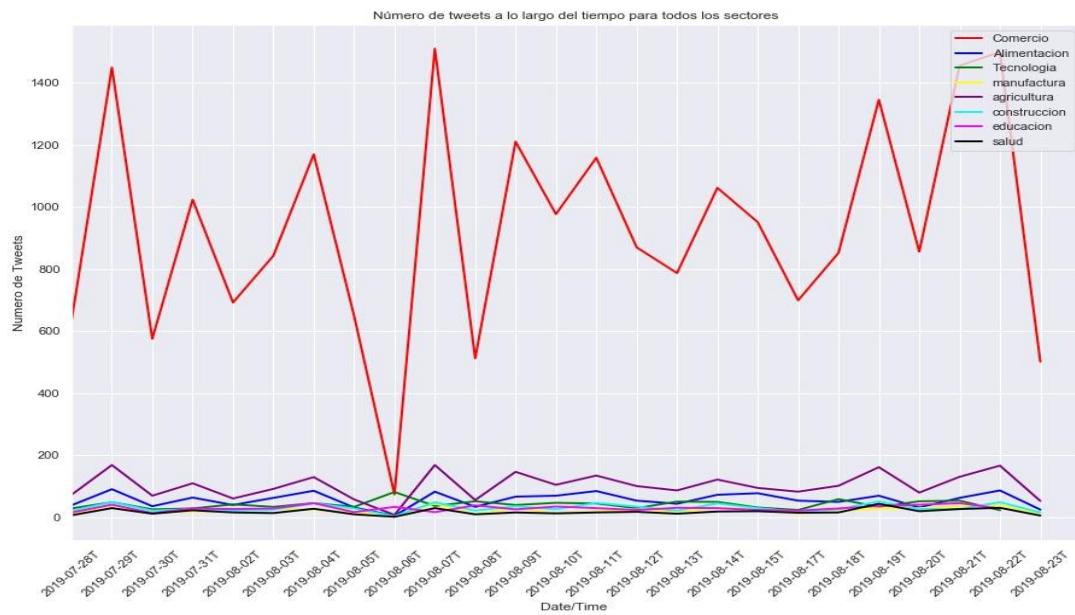
**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

El grafico N. 54, se genera para tener una vista en línea de tiempo que presenta el índice de interactividad de usuarios que detonaron el gusto o la compra de un producto en un sector. Permitiendo a los emprendedores obtener información respecto al alza y baja en el índice de actividad que genera cada sector e industria. Denotando picos de actividad de consumo útiles en la toma de decisiones, que aportan a los emprendedores información a mediano plazo para seleccionar una industria o sector en la cual exista un índice de actividad continuo en base al tiempo. Y no sectores que generan índices de consumo solo por temporadas. Lo cual podría generarles perdidas si emprenden justo en el momento en que ha terminado una alta tasa de actividad de consumo y les espera un periodo de recesión, en caso de existir, hasta una nueva temporada en que se generan estos índices de actividad. A los dueños de negocios les permite en caso de visualizar un alto índice de consumo en su industria dotarse de mercadería para abastecer una demanda en caso de existir. Toda esta información es en base al análisis del lenguaje

natural, y debe ser validada por una investigación propia de cada usuario con base a números que terminen corroborando si la predicción realizada en base a este proyecto es similar a los ingresos que genera cada industria o sector.

**Gráfico N. 54** Visualización general en línea de tiempo



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

Posteriormente se genera el gráfico N. 55, que realiza el análisis presentado anteriormente pero de forma individual para cada sector/industria presente en el desarrollo de este proyecto.

**Gráfico N. 55** Índice de interacción de tweets con un sector en línea de tiempo



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

Para finalizar el flujo del presente notebook, toda la información generada será archivada en un archivo en formato PDF para posterior consulta del usuario final. Dotándole de información basada en el análisis de opiniones registradas en la red social Twitter. Para su corroboración del presente proyecto se presenta el grafico N. 56.

**Gráfico N. 56** Lectura de archivo PDF generado



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

## Entregables del proyecto

Se detallan los siguientes entregables del proyecto que son los siguientes:

- Manual de usuario: Es un documento fundamental que contiene la información para el conocimiento de los módulos y el funcionamiento de la aplicación web, ver **Anexo 5**.
- Manual técnico: Este documento es la guía para usuarios con conocimientos especializados sobre el desarrollo del presente proyecto. Dentro del documento se detalla el procedimiento de instalación de la plataforma Jupyter Notebook. Creación de la base de datos no relacional Firebase Realtime. Conceptos técnicos para comprender el código fuente y el flujo de trabajo realizado. Para mayor detalle se adjunta el manual en el **Anexo 6**.
- Notebooks del proyecto (Código fuente)
- Corpus de entrenamiento con tweets en español categorizados por sentimiento.
- Archivo Json para importación de base de datos.
- Archivo .m para uso del algoritmo de clasificación de sentimiento.

## CRITERIOS DE VALIDACIÓN DE LA PROPUESTA

Para la validación del tema propuesto “ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON”, se realizó un juicio de expertos con tres profesionales del área de la computación, los cuales probaron los notebooks desarrollados en el presente proyecto y evaluaron los mismos bajo los criterios elaborados para este fin. En el **anexo 3** se encuentra las fichas llenadas por los entendidos.

A continuación, se presenta el perfil de los expertos:

*Cuadro N. 13 Perfil de los expertos*

Nº	Experto	Título académico	Cargo	Años de experiencia
1	Ing. José Baque	Ingeniero en Sistemas Computacionales	Desarrollador software en la empresa Cursor.	3
2	Ing. Luis Alcides Mora Camacho	Ingeniero en Sistemas Computacionales	Desarrollador Senior en la empresa Sipecom S.A.	3
3	Lsi. Carlos Víctor Ramírez	Licenciado en Sistema Información	Antiguo Jefe Sistemas en el Hospital León Becerra. Actualmente Gerente en el Grupo Artegrafía AD.	6

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

La valoración en promedio otorgada por los profesionales se muestra en el siguiente cuadro:

**Cuadro N. 14** Criterios de validación

Ítem	Criterios	Nivel de cumplimiento
1	Herramienta de ciencia de datos amigable	95%
2	Información se guarda correctamente	93%
3	Validación en la extracción de datos	91%
4	Tiempo de procesamiento Aceptable	87%
5	Fiabilidad de datos presentados	88%

Ítem	Notebooks	Nivel de cumplimiento
1	Extracción de tweets	98%
2	Creación y entrenamiento del algoritmo de clasificación de sentimientos.	87%
3	Búsqueda y procesamiento de opiniones	93%
4	Clasificación de las opiniones por sectores.	91%
5	Visualización de información en archivo PDF.	100%

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

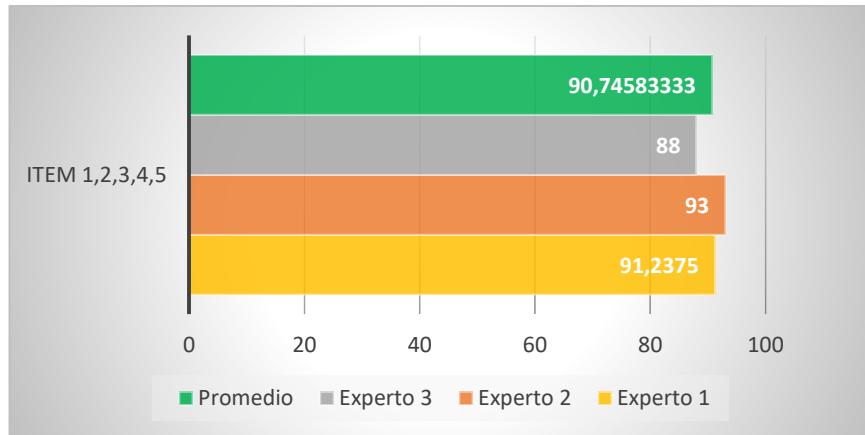
## Procesamiento y análisis

Para una interpretación más detallada se dividió los criterios de validación en:

### Criterios generales

Valoración promediada emitida por los expertos para los criterios generales.

*Gráfico N. 57 Promedio de criterio general*



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

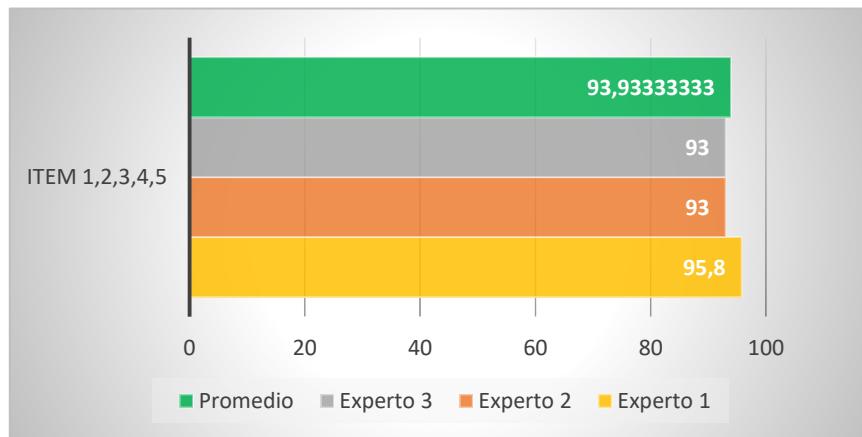
**Análisis:** Según los datos de las entrevistas, los expertos determinaron que el presente proyecto cumple en un 91% con los criterios generales establecidos. Siendo esta una cifra aceptable para un proyecto de titulación.

Entre los puntos en que más hicieron énfasis los expertos fue en el tiempo de procesamiento, esto se debe a la cantidad de tweets procesados en el presente proyecto, cuya cantidad es de 65.000 en un periodo comprendido desde inicios de mayo y finales del mes de agosto, lo cual hace que le tome entre 9 a 12 minutos al algoritmo de clasificación de sentimientos determinar la polaridad de todos los tweets. Una cifra de tiempo aceptable dada la magnitud del dataset. En el resto de campos el proyecto no obtuvo observaciones, fue aprobado en su totalidad.

### **Criterios de validación de notebooks**

Valoración promediada emitida por los expertos para los criterios de validación de notebooks.

*Gráfico N. 58* Promedio de criterio por notebooks



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

**Análisis:** Según los datos de las entrevistas, el promedio de los expertos determina que el proyecto cumple un 93% en que los notebooks realizan el objetivo general y específico del presente proyecto.

Entre los puntos en que más hicieron énfasis los expertos fue en la creación y entrenamiento del algoritmo, según su punto de vista la propuesta de creación de un algoritmo de clasificación de sentimientos en español es muy buena, pero sería mejor llegar al umbral de 5000 tweets en el dataset de entrenamiento, actualmente el dataset posee 3000 tweets de entrenamiento, logrando un accuracy de 78%. Bajo el criterio de los expertos esta cifra es aceptable y viable para un proyecto de esta índole. En el resto de campos de criterios por notebooks el proyecto no obtuvo indicaciones y fue aprobado en su totalidad.

## **CAPÍTULO IV**

### **CRITERIOS DE ACEPTACIÓN DEL PRODUCTO**

A continuación, se detallarán los criterios de aceptación, conclusiones y recomendaciones correspondientes al presente trabajo de titulación, una vez finalizado con el presente proyecto se espera que la calidad del producto sea adecuada para proceder a ser aceptable, por tal motivo se realizó una ficha con actividades para ser revisado y valorado por usuarios.

Este proyecto se desarrolló con el objetivo de poder brindar apoyo en la toma de decisiones a los emprendedores y dueños de negocios, permitiéndoles mediante el uso de notebooks extraer datos públicos de la red social Twitter generados en la presente ciudad y su posterior análisis mediante la aplicación de algoritmos de aprendizaje automático, generando información visual que les permita obtener insight e información útil en la toma de decisiones.

Los criterios de aceptación fueron planteados en base a todos los datos recolectados en las encuestas realizadas, y además criterios que permita verificar el correcto funcionamiento de cada notebook desarrollado.

El número de usuarios a encuestar es el 10% de la muestra obtenida para usuarios que están emprendiendo o son actualmente dueño de negocios, esto con el fin de corroborar lo que los juicios de expertos dieron como resultado. El tamaño de la muestra es de 40 usuarios (10% de 397 usuarios encuestados =  $39.7 = 40$  usuarios) para los presentes fines.

Los criterios de evaluación que se han tomado en cuenta para la aceptación del proyecto por parte de 40 usuarios finales, entre los cuales 18 de ellos son actuales dueños de negocios y 22 usuarios que están emprendiendo en la ciudad de Guayaquil., los cuales pudieron evidenciar el uso de los notebooks

y sus resultados, por consiguiente evaluaron su funcionamiento de acuerdo al documento mencionado en el cuadro N. 15. Previo a cada prueba de validación, se verificó el correcto funcionamiento de los notebooks para garantizar el correcto funcionamiento del código fuente en la plataforma Jupyter.

El proyecto cuenta con la aceptación de los usuarios, quienes fueron los encargados de hacer uso de los notebooks que contienen el código fuente del proyecto, probando las funcionalidades indicadas.

**Cuadro N. 15** Matriz de resultados de criterios de aceptación

RESULTADO DE LOS ASPECTOS DE ACEPTACIÓN									
	P1	P2	P3	P4	P5	P6	P7	P8	P9
Si	40	40	36	28	32	36	32	40	40
No	-	-	4	12	8	4	8	-	-

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

Mediante criterios de evaluación cerrados donde “Si” nos indica que han aprobado el criterio y “No” que han rechazado el criterio planteado. Analizando estos resultados, los cuales indican que el proyecto obtiene un porcentaje de aprobación del 90%, dando un resultado de que cumple con los criterios de evaluación planteados

Dando por terminado el análisis de los resultados obtenidos, se procede a concluir que el proyecto es viable y de acuerdo con el problema existente resulta ser una alternativa de ayuda en la toma de decisiones en los emprendedores y dueños de negocios en la presente ciudad.

Las evidencias de las respuestas emitidas por los usuarios se pueden encontrar en el **Anexo 4**.

A continuación, en el cuadro N. 16 se detalla el porcentaje de aceptación para cada pregunta de la encuesta con los criterios de aceptación y satisfacción emitidos.

**Cuadro N. 16** Criterios de aceptación

ÍTEM	CRITERIOS	SI	NO
1	¿Los datos extraídos pertenecen a la ciudad de Guayaquil? ¿Aceptaría usar estos notebooks para obtener información	100%	0%
2	que le permita conocer el grado de actividad y aceptación de un producto o servicio?	100%	0%
	¿Los notebooks le son útil para obtener información pública		
3	acerca de los comentarios emitidos respecto a un producto o servicio?	90%	10%
4	¿Le resulta ventajoso ver los análisis de sentimientos de los distintos sectores e industrias para una toma de decisiones?	70%	30%
5	¿La plataforma de ciencia de datos utilizada le pareció de fácil uso?	80%	20%
	¿Considera provechoso que este proyecto le permita conocer en qué sector e industria están interactuando los usuarios con su respectiva línea de tiempo, visualizando alza y baja en los índices de interactividad, y así poder sacar sus propias conclusiones que le permitan tomar decisiones?		
6	¿La visualización de los gráficos generados en los análisis realizados le pareció de fácil comprensión luego de la explicación realizada por el presente autor?	90%	10%
7	¿La información obtenida se guarda correctamente en el archivo PDF?	80%	20%
8	¿Considera usted que la información generada le permite tomar mejores decisiones de negocios o al emprender?	100%	0%
9		100%	0%

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Datos de la investigación.

Mediante la demostración del proyecto a usuarios, se logra determinar que los notebooks cumplen con los objetivos propuestos en este proyecto, adicionalmente los emprendedores manifestaron que el uso de algoritmos de aprendizaje automático para el análisis de datos les pareció un campo fascinante.

Denotando un amplio interés por la clasificación de opiniones mediante el análisis de sentimiento y el uso de una red neuronal para detectar patrones de consumo en diferentes sectores e industrias, esto les pareció muy útil, dado que gracias al uso de librerías para la creación de gráficos se presenta información visual, que les permite comprender que sectores tienen mayor actividad en la red social.

Los gráficos en línea de tiempo les permitió observar que días hay una mayor emisión de comentarios respecto al consumo por sectores e industrias y qué polaridades se encontraron en estos tweets, que reflejaban el sentimiento de cada usuario en gráficos de barra o tipo pastel. Permitiendo visualizar el nivel de satisfacción o rechazo.

La búsqueda de producto o servicio en el conjunto de datos les permitió determinar el nivel de aceptación o rechazo existente, en base al análisis de tweets que contienen en su sintaxis el nombre del servicio o producto ingresado previamente en la búsqueda por teclado, permitiendo obtener un análisis rápido si un producto es popular o no, todo esto depende de la existencia del mismo en el conjunto de datos extraídos.

Por lo tanto se concluye que el uso de algoritmos de aprendizaje automático planteado cumple su propósito de generar información útil en el proceso de toma de decisiones en el emprendimiento o en negocios ya establecidos.

## **Informe de aceptación y aprobación para productos de SOFTWARE / HARDWARE**

La validación del producto por parte del experto N. 1 El Ing. José Sánchez. El presente experto realizó la comprobación del correcto funcionamiento de los notebooks, y se realizó la debida calificación mediante el uso de porcentajes, representados en el cuadro N. 17.

**Cuadro N. 17** Valores de calificación de criterios de aprobación

<b>Valoración</b>	<b>Cumplimiento</b>	<b>Rango</b>
<b>5</b>	Excelente	100%
<b>4</b>	Muy bueno	99% - 76%
<b>3</b>	Bueno	75% - 51%
<b>2</b>	Regular	50% - 26%
<b>1</b>	Insuficiente	25% - 0

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

A través del formulario de validación que se encuentra en el **anexo 3** se obtuvieron los siguientes resultados.

**Cuadro N. 18** Resultados de calificación de criterios experto N. 1.

<b>RESULTADOS DE LOS ASPECTOS DE VALIDACIÓN</b>	
<b>Excelente</b>	1 de 10 aspectos
<b>Muy bueno</b>	9 de 10 aspectos
<b>Bueno</b>	-----
<b>Regular</b>	-----
<b>Insuficiente</b>	-----

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Los resultados del experto N. 2. El Ing. Luis Mora Camacho. Realizó la comprobación del correcto funcionamiento de los notebooks, y se realizó la debida calificación mediante el uso de porcentajes, representados en el cuadro N. 19. Del cual se obtuvieron los siguientes resultados.

**Cuadro N. 19** Resultados de calificación de criterios experto N. 2

**RESULTADOS DE LOS ASPECTOS DE VALIDACIÓN**

<b>Excelente</b>	<b>3 de 10 aspectos</b>
<b>Muy bueno</b>	<b>7 de 10 aspectos</b>
<b>Bueno</b>	-----
<b>Regular</b>	-----
<b>Insuficiente</b>	-----

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Los resultados del experto N. 3. El LSi. Carlos Ramírez, realizó la comprobación del correcto funcionamiento de los notebooks, y se realizó la debida calificación mediante el uso de porcentajes, representados en el cuadro N. 20. Del cual se obtuvieron los siguientes resultados.

**Cuadro N. 20** Resultados de calificación de criterios experto N. 3

**RESULTADOS DE LOS ASPECTOS DE VALIDACIÓN**

<b>Excelente</b>	<b>2 de 10 aspectos</b>
<b>Muy bueno</b>	<b>8 de 10 aspectos</b>
<b>Bueno</b>	-----
<b>Regular</b>	-----
<b>Insuficiente</b>	-----

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Mediante los resultados obtenidos en la validación del producto por parte de los expertos, se procede a concluir que el presente proyecto cumple con los objetivos y alcances propuestos al inicio del proyecto. La constancia de la validación de los expertos puede verse en el **anexo 3**.

### **Informe de aseguramiento de la calidad para productos de SOFTWARE/ HARDWARE**

El presente proyecto fue desarrollado utilizando el lenguaje de programación Python, en la plataforma web Jupyter Notebook, el cual ofrece apoyo en la toma de decisiones a los emprendedores y dueño de negocios, permitiéndoles obtener información mediante el análisis de datos públicos extraídos en la red social, que al ser procesados por los notebooks desarrollados en el presente proyecto generan información de productos y servicios mencionados en la red, además de una clasificación de tweets por sectores e industrias.

Para el correcto funcionamiento de los notebooks se recomienda a los usuarios instalar el paquete de anaconda, que incluye Anaconda Navigator el cual trae embebida la plataforma Jupyter empleada en este proyecto.

Además se realizó la validación de juicio de experto, los cuales validaron en base a pruebas y funcionalidad los notebooks del presente proyecto, de la misma forma se midió la aceptación por parte de los usuarios de acuerdo con las funcionalidades establecidas.

### **Privacidad de los datos**

La privacidad de los datos de los usuarios está considerada en el presente proyecto, es por eso que no se extraen datos que permitan identificar al usuario emisor de cada tweet. Evitando así problemas legales con la empresa Twitter.

Una vez realiza la demostración en vivo del correcto funcionamiento del algoritmo de clasificación de sentimientos y la red neuronal artificial para la clasificación por tweets por sectores e industrias, el proyecto valida el almacenamiento tweets con un lapso de un mes, respetando el periodo de almacenamiento establecido en los acuerdos de uso de sus APIs. Al iniciar un nuevo mes no se podrá obtener acceso a los datos almacenados en el mes anterior debido a que habrán sido eliminados.

De este modo se logra el uso adecuado de la información extraída mediante el API Streaming.

## **Conclusiones**

Los notebooks desarrollados para el presente proyecto permiten analizar los datos extraídos de la red social Twitter, mediante el uso de algoritmos de aprendizaje automático, aplicando técnicas de minería de datos, adicionando el uso del procesamiento del lenguaje natural, para una mayor comprensión del idioma español, todo este proceso fue realizado en la plataforma Jupyter Notebook, permitiendo la obtención de información relacionada con la actividad comercial por sectores.

Se extrajeron datos públicos de la red social Twitter, haciendo uso del API Streaming, la cual brinda acceso a los datos en tiempo real y mediante el uso de cajas delimitadoras que permiten incorporar coordenadas, se logró extraer y conformar un dataset de 65.000 tweets para la prueba de los algoritmos creados en el presente proyecto. La extracción de tweets se realizó desde inicios del mes de mayo a finales del mes de agosto, logrando una extracción de datos de cuatro meses.

Se seleccionaron los algoritmos de aprendizaje automático como el clasificador bayesiano ingenuo, que permitió crear el modelo de clasificación de sentimientos; y para la clasificación de tweets que contengan indicios comerciales se desarrolló una red neuronal artificial entrenándola con patrones respectivos a cada sector e industria.

Se realizaron diversas pruebas para determinar el algoritmo adecuado para la clasificación de sentimientos, el modelo de Maquinas de vectores de vectores denotó un accuracy del 80%, pero en las pruebas de validación las predicciones generadas por este modelo no fueron las esperadas, clasificando correctamente 6 de cada de 10 tweets. El modelo de clasificación bayesiano mediante el uso del corpus de entrenamiento obtiene un accuracy del 77.69%

y en las pruebas de validación otorga unas predicciones correctas en 8 de cada 10 tweets. Por lo tanto se considera correcto el uso del modelo de clasificación bayesiano para los presentes fines.

El modelo de red neuronal artificial desarrollado en el presente proyecto alcanza una precisión del 83%, lo cual se considera aceptable para los presentes fines. Otorgando una clasificación correcta de 8 tweets de cada 10 de tweets que poseen connotación comercial o un patrón de consumo. Se puede aumentar la precisión al agregar más datos al conjunto de entrenamiento y realizar más entrenamientos con más épocas y validar si se obtienen mejores resultados.

Se logró aplicar con éxito técnicas de Machine Learning y se utilizaron librerías para la generación de gráficos como Matplotlib y Seaborn en el entorno de trabajo Jupyter Notebook, logrando realizar una diversidad de gráficos que permiten la visualización de información específica y de todos los sectores e industrias.

Adicionalmente el presente proyecto hizo uso de notebooks en plataformas cloud como Microsoft Azure Notebooks y de Google Colaboratory, y dado a su mayor velocidad de procesamiento se concluye ejecutar estos notebooks de preferencia en Google Colaboratory.

## **Recomendaciones**

Es recomendable leer el manual de usuario para conocer los notebooks desarrollados en el presente proyecto.

Aumentar el número de ejemplos en el dataset de entrenamiento del modelo de clasificación de sentimientos en español. Al momento de finalizar el presente proyecto el dataset posee un total de 3000 frases con sus respectivas polaridades que sirven para entrenar el modelo. La finalidad de incorporar más datos de entrenamiento es aumentar la exactitud del modelo, que actualmente posee una eficacia del 86% utilizando el algoritmo de clasificación ingenuo bayesiano.

Se recomienda migrar este proyecto a un servicio web, que haga uso del código fuente escrito en este trabajo de titulación, con el objetivo de dotar al usuario final de una mejor experiencia debido a que si no se lee el manual de usuario se requiere la guía de un usuario con conocimiento técnico en el campo de la ciencia de datos.

Se recomienda realizar la incorporación de más ejemplos y palabras claves en el dataset de entrenamiento de la red neuronal que permite clasificar las opiniones de los usuarios por distintos sectores e industrias. Así aumentar su capacidad de clasificación al tener más datos, para clasificar con mayor eficacia en distintos sectores e industrias.

Se recomienda el uso de la nueva plataforma de ciencia de datos Jupyter Lab. Tomando en cuenta que la nueva generación de Jupyter Notebook es Jupyter Lab es de suma importancia aclarar que los cuadernos Jupyter desarrollados en el presente proyecto podrán seguir siendo utilizados con la misma eficacia en la plataforma Jupyter Lab.

## Bibliografía

- Acosta, M., & Cruz, K. (2017). *Repository Institucional de la Universidad de Guayaquil*. Obtenido de Repository Institucional de la Universidad de Guayaquil: <http://repository.ug.edu.ec/bitstream/redug/20713/1/trabajo%20de%20titulaci%C3%B3n%20%28tesis%29.pdf>
- Aguiar, V. (2017). *Medium*. Obtenido de Medium: <https://medium.com/data-hackers/uma-introdução-simples-ao-pandas-1e15eea37fa1>
- Alfaro Arancibia, R. (2016). *Pontificia Universidad Católica de Valparaíso*. Obtenido de Pontificia Universidad Católica de Valparaíso: [http://opac.pucv.cl/pucv\\_txt/txt-8500/UCD8528\\_01.pdf](http://opac.pucv.cl/pucv_txt/txt-8500/UCD8528_01.pdf)
- Anaconda, I. (2019). *Anaconda*. Obtenido de Anaconda: <https://www.anaconda.com/>
- Aquino, G. (2013). Caracterización de documentos utilizando técnicas de minería de textos. *SEDICI*. Obtenido de <http://sedici.unlp.edu.ar/handle/10915/63166>
- Arcila, C., & et al. (2017). ANÁLISIS SUPERVISADO DE SENTIMIENTOS POLÍTICOS EN ESPAÑOL. *Repository Español de Ciencia y Tecnología*, 977. Obtenido de <https://recyt.fecyt.es/index.php/EPI/article/viewFile/epi.2017.sep.18/36488>
- Asensio, E. (2015). *RiuNet Repository Institucional de la Universitat Politècnica de València*. Obtenido de RiuNet Repository Institucional de la Universitat Politècnica de València: <https://riunet.upv.es/bitstream/handle/10251/56102/ASENSIO%20-Apliaci%C3%B3n%20de%20t%C3%A9cnicas%20de%20miner%C3%ADA%20de%20datos%20en%20redes%20sociales.pdf>
- Astera. (2019). *Astera*. Obtenido de Astera: <https://www.astera.com/es/soluciones/soluciones-tecnológicas/la-extracción-de-datos/>

- AWS. (2019). *Amazon Web Services*. Obtenido de Amazon Web Services: <https://aws.amazon.com/es/comprehend/>
- Bbva. (08 de 05 de 2015). *Bbva Open 4 U*. Obtenido de Bbva Open 4 U: <https://bbvaopen4u.com/es/actualidad/herramientas-basicas-para-los-desarrolladores-en-python>
- Bertuzzi, L., & Suarez, D. (2016). *RIDAA UNICEN*. Obtenido de RIDAA UNICEN : <http://www.ridaa.unicen.edu.ar/xmlui/bitstream/handle/123456789/643/Tesis%20de%20grado%20Bertuzzi-Suarez.pdf?sequence=1&isAllowed=y>
- Blanco, E. (2016). *UPCommons*. Obtenido de UPCommons: <https://upcommons.upc.edu/bitstream/handle/2117/82434/113257.pdf?sequence=1&isAllowed=y>
- Cabrera, S., & Reyes, Z. (2017). *Repositorio Universidad de Guayaquil*. Obtenido de Repositorio Universidad de Guayaquil: <http://repositorio.ug.edu.ec/bitstream/redug/19514/1/UG-FCMF-B-CISC-PTG-1253.pdf>
- Calvo, D. (26 de 11 de 2016). *Diego Calvo* . Obtenido de Diego Calvo : <http://www.diegocalvo.es/modelo-crisp-drm-data-mining/>
- Cartagena, B. P. (09 de 2017). Obtenido de Predicción de la probabilidad de éxito en la adquisición de clientes
- Cascante, R. (17 de 06 de 2018). *Medium*. Obtenido de Medium: <https://medium.com/@margalida.kaskante/empezando-con-firebase-realtime-database-authentication-a5c54b3b67d6>
- Castillo, J. (2017). Análisis Automático de mensajes y usuarios de Twitter en Chile. *Universidad Católica De Valparaíso*, 5.

Chamorro, V. (09 de 2018). *UNIVERSIDAD COMPLUTENSE DE MADRID*. Obtenido de  
UNIVERSIDAD COMPLUTENSE DE MADRID:  
<https://eprints.ucm.es/49774/1/TFM%20Veronica%20Chamorro%20Alvarado.pdf>

Código fuente. (29 de 10 de 2018). Obtenido de Código fuente:  
<https://www.codigofuente.org/clasificacion-texto-python/>

Conde, M. (2018). *Analítica Sports*. Obtenido de Analítica Sports:  
<https://www.analiticaspports.com/analisis-de-sentimientos-para-entender-el-humor-social-en-el-mundial/>

Consuegra, J. (03 de 02 de 2014). *Analitica web*. Obtenido de Analitica web:  
<https://www.analiticaweb.es/analisis-de-sectores-mediante-tendencias/>

Costa, C. (2015). *Universidad Complutense de Madrid*. Obtenido de Universidad Complutense de Madrid:  
[https://www.ucm.es/data/cont/media/www/pag-73273/TesisDoctoral\\_CarlosCosta-2015.pdf](https://www.ucm.es/data/cont/media/www/pag-73273/TesisDoctoral_CarlosCosta-2015.pdf)

Developers, G. (2019). *Firebase*. Obtenido de Firebase:  
<https://firebase.google.com/docs/database/?hl=es-419>

Digital, I. (2016). *Lujo digital*. Obtenido de Lujo digital: <http://lujodigital.org/en-los-proximos-anos-la-ia-seguira-contribuyendo-al-crecimiento-economico/>

Dubiau, L. (10 de 2014). *Facultad de Ingeniería de la Universidad de Buenos Aires*. Obtenido de Facultad de Ingeniería de la Universidad de Buenos Aires:  
<http://materias.fi.uba.ar/7500/Dubiau.pdf>

Fernández, I. (2018). *UCM*. Obtenido de UCM: <https://www.ucm.es/data/cont/docs/1334-2019-03-27-Guía%20de%20actuación%20def%202019%20WEB.pdf>

GAD Municipal de Guayaquil. (2015). Obtenido de GAD Municipal de Guayaquil:  
<https://guayaquil.gob.ec/Paginas/negocios-guayaquil.aspx>

- Galán, V. (2015). Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario. *Universidad Carlos III de Madrid / Biblioteca*, 21.
- Obtenido de [https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC\\_Victor\\_Galan\\_Cortina.pdf](https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf)
- García, I. (10 de 07 de 2017). *Economia simple*. Obtenido de Economia simple: <https://www.economiasimple.net/glosario/oportunidad-de-negocio>
- García, V. (2018). *Repositori Universitat Jaume I*. Obtenido de Repositori Universitat Jaume I:<http://repositori.uji.es/xmlui/bitstream/handle/10234/180348/Memoria%20TFM%20VictorGarciaPerez%20.pdf?sequence=1&isAllowed=y>
- GNU. (31 de 07 de 2019). *GNU*. Obtenido de GNU: <https://www.gnu.org/philosophy/free-sw.es.html>
- Gómez, E. (03 de 2018). Influencia de redes sociales en el análisis de sentimiento aplicado a la situación política en Ecuador. *SciELO*, 3. Obtenido de [http://scielo.senescyt.gob.ec/scielo.php?script=sci\\_arttext&pid=S1390-65422018000100067](http://scielo.senescyt.gob.ec/scielo.php?script=sci_arttext&pid=S1390-65422018000100067)
- Gómez, J., & Gallego, A. (2018). *DSpace*. Obtenido de DSpace: <http://dspace.tdea.edu.co/bitstream/tda/435/1/DIAGNOSTICO%20DEL%20ESTADO%20DE%20LA%20CUESTION%20DEL%20ETIQUETADO%20LINGUISTICO.pdf>
- González, & et al. (2015). Algoritmos de clasificación y redes neuronales en la observación automatizada de registros. *SciELO*, 35.
- González, D. (2017). *Archivo Digital UPM*. Obtenido de Archivo Digital UPM: [http://oa.upm.es/48921/1/TFG\\_DANIELA\\_GONZALEZ\\_HERRERA.pdf](http://oa.upm.es/48921/1/TFG_DANIELA_GONZALEZ_HERRERA.pdf)

González, J. (2017). Obtenido de  
<https://pdfs.semanticscholar.org/37d4/dd4a6535bef307325164891f16bcdd293721.pdf>

González, J. (2017). Aprendizaje profundo para el procesamiento del lenguaje natural.  
Universitat Politècnica de València, 53. Obtenido de  
<https://riunet.upv.es/handle/10251/86279>

Google Developers. (2019). *Firebase*. Obtenido de Firebase:  
<https://firebase.google.com/docs/database/web/structure-data?hl=es-419>

Huamán, C. (2017). *Universidad Nacional Santiago Antúnez de Mayolo*. Obtenido de  
Universidad Nacional Santiago Antúnez de Mayolo:  
<http://repositorio.unasam.edu.pe/handle/UNASAM/2017>

IBM. (2019). *IBM*. Obtenido de IBM:  
[https://www.ibm.com/support/knowledgecenter/es/SS3RA7\\_sub/modeler\\_mainhelp\\_client\\_ddita/clementine/svm\\_howwork.html](https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/svm_howwork.html)

IIC. (2019). *Instituto de Ingeniería del Conocimiento*. Obtenido de Instituto de Ingeniería del  
Conocimiento: <http://www.iic.uam.es/soluciones/inteligencia-de-cliente/procesamiento-lenguaje-natural/>

Inc., T. (2018). *Twitter*. Obtenido de Twitter: <https://help.twitter.com/es/glossary>

Ingenio, v. (2017). *Ingenio virtual*. Obtenido de Ingenio virtual:  
<https://www.ingeniovirtual.com/tipos-de-graficos-y-diagramas-para-la-visualizacion-de-datos/>

Ionos. (28 de 02 de 2019). *Ionos*. Obtenido de Ionos:  
<https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/jupyter-notebook/>

JSON. (s.f.). *JSON*. Obtenido de JSON: <https://www.json.org/json-es.html>

Jupyter. (2019). *Jupyter*. Obtenido de Jupyter: <https://jupyter.org/>

- Laso et al. (2014). GEM Ecuador. *ESPAE Graduate School of Management*, 23.
- Laso et al. (2017). GLOBAL ENTREPRENEURSHIP MONITOR ECUADOR 2017. *ESPAE Graduate School of Management*, 25-26.
- Latam, S. (29 de 08 de 2019). *Medium*. Obtenido de Medium:  
<https://medium.com/@LatamServices/emprendedores-1012df96857f>
- Leo, B. (2001). Random Forests.
- Librado, H. (01 de 2017). *Gelbukh*. Obtenido de Gelbukh:  
<https://www.gelbukh.com/thesis/Hugo%20Librado%20Jacobo%20-%20MSc.pdf>
- Machine learning, A. (27 de 12 de 2018). *Aprende machine learning*. Obtenido de Aprende machine learning: <https://www.aprendemachinelearning.com/procesamiento-del-lenguaje-natural-nlp/>
- Machine learnings. (26 de 01 de 2017). *Machine learnings*. Obtenido de Machine learnings:  
<https://machinelearnings.co/text-classification-using-neural-networks-f5cd7b8765c6>
- Marketing, d. (2016). *Marketing directo*. Obtenido de Marketing directo:  
<https://www.marketingdirecto.com/diccionario-marketing-publicidad-comunicacion-nuevas-tecnologias/audiencia-2>
- Martín, C. (2016). *Repositorio Digital de la Universidad Nacional de Córdoba*. Obtenido de Repositorio Digital de la Universidad Nacional de Córdoba:  
[https://rdu.unc.edu.ar/bitstream/handle/11086/3751/Becerra%202016\\_analisis-de-sentimiento.pdf?sequence=1](https://rdu.unc.edu.ar/bitstream/handle/11086/3751/Becerra%202016_analisis-de-sentimiento.pdf?sequence=1)
- MathWorks. (2019). *MathWorks*. Obtenido de MathWorks:  
<https://la.mathworks.com/discovery/support-vector-machine.html>
- MathWorks, I. (2018). *Mathworks*. Obtenido de Mathworks:  
<https://la.mathworks.com/discovery/machine-learning.html>
- McGrath, R. (2014). *twython*. Obtenido de twython: <https://twython.readthedocs.io/en/latest/>

- Mertricks. (7 de 8 de 2016). *Mertricks*. Obtenido de Mertricks.
- Molina, J. G. (2016). *REVISTA ONTARE DE LA UNIVERSIDAD EAN*. Obtenido de REVISTA  
ONTARE DE LA UNIVERSIDAD EAN:  
<https://journal.universidadean.edu.co/index.php/Revistao/article/view/1440>
- NumFOCUS. (2018). *PyData*. Obtenido de PyData: <https://pandas.pydata.org/#python-data-analysis-library>
- NumFOCUS. (2019). *Jupyter*. Obtenido de Jupyter : <https://jupyter.org/>
- NumFOCUS. (2019). *Numpy*. Obtenido de Numpy: <https://www.numpy.org/>
- NumPy. (2018). *NumPy*. Obtenido de NumPy: <http://www.numpy.org/>
- Olgún, A. (2015). *ResearchGate* . Obtenido de ResearchGate :  
[https://www.researchgate.net/publication/281979580\\_Que\\_se\\_escribe\\_respecto\\_al\\_marxismo\\_en\\_redes\\_sociales\\_Analisis\\_de\\_patrones\\_de\\_texto\\_a\\_traves\\_de\\_Twitter\\_por\\_medio\\_de\\_Data\\_Mining](https://www.researchgate.net/publication/281979580_Que_se_escribe_respecto_al_marxismo_en_redes_sociales_Analisis_de_patrones_de_texto_a_traves_de_Twitter_por_medio_de_Data_Mining)
- opensource. (2018). Obtenido de opensource: <https://opensource.com/resources/what-open-source>
- Packard, H. (2018). *Hewlett Packard Enterprise*. Obtenido de Hewlett Packard Enterprise:  
<https://www.hpe.com/lamerica/es/what-is/artificial-intelligence.html>
- Pamies, B. (09 de 2017). *Universidad de Alicante*. Obtenido de Universidad de Alicante:  
[https://rua.ua.es/dspace/bitstream/10045/69432/1/Prediccion\\_de\\_la\\_probabilidad\\_de\\_exito\\_en\\_la\\_adqui\\_PAMIES\\_CARTAGENA\\_BENJAMIN.pdf](https://rua.ua.es/dspace/bitstream/10045/69432/1/Prediccion_de_la_probabilidad_de_exito_en_la_adqui_PAMIES_CARTAGENA_BENJAMIN.pdf)
- Patricia, S. (12 de 2014). *Escuela Politécnica Nacional*. Obtenido de Escuela Politécnica Nacional: <https://bibdigital.epn.edu.ec/bitstream/15000/8992/3/CD-6006.pdf>
- Pérez, C. (11 de 01 de 2019). *CustomerTrigger* . Obtenido de CustomerTrigger :  
<https://www.customertrigger.com/analizar-emociones-plus-las-areas-marketing/>

- Recalde, S. (04 de 2018). *Universidad Central del Ecuador*. Obtenido de Universidad Central del Ecuador.
- Reyes, G., & Crespo, C. (2018). Algoritmos de agrupación difusos. *Innovative Space of Scientific Research Journals*, 19. Obtenido de Innovative Space of Scientific Research Journals
- Rivera, M., & Villavicencio, J. (2017). *Pontificia Universidad Católica de Valparaíso*. Obtenido de Pontificia Universidad Católica de Valparaíso: [http://opac.pucv.cl/pucv\\_txt/txt-8000/UCC8097\\_01.pdf](http://opac.pucv.cl/pucv_txt/txt-8000/UCC8097_01.pdf)
- Rivera, M., & Villavicencio, J. (12 de 2017). *Pontificia Universidad Católica de Valparaíso*. Obtenido de Pontificia Universidad Católica de Valparaíso: [http://opac.pucv.cl/pucv\\_txt/txt-8000/UCC8097\\_01.pdf](http://opac.pucv.cl/pucv_txt/txt-8000/UCC8097_01.pdf)
- Rosas, R. (06 de 2019). *Rosana Rosas*. Obtenido de Rosana Rosas: <https://rosanarosas.com/analisis-sentimiento-redes-sociales/>
- Rouse, M. (04 de 2017). *Search data center*. Obtenido de Search data center: <https://searchdatacenter.techtarget.com/es/definicion/Inteligencia-artificial-o-AI>
- Rouse, M. (1 de 2017). *TechTarget*. Obtenido de TechTarget: <https://searchdatacenter.techtarget.com/es/definicion/Aprendizaje-automatico-machine-learning>
- Ruiz, S. (20 de 07 de 2017). *Analitica web*. Obtenido de Analitica web: <https://www.analiticaweb.es/algoritmo-knn-modelado-datos/>
- Sabrino, J. (06 de 2018). *Repositorio Universitat Oberta de Catalunya*. Obtenido de Repositorio Universitat Oberta de Catalunya: <http://openaccess.uoc.edu/webapps/o2/bitstream/10609/81435/6/jsobrinosTFM0618memoria.pdf>

- San Martín Duchen, D. (2017). Proceso y aplicación Web para la gestión de librerías de datos nucleares: Interfaz de usuario y herramientas. *Universidad Politécnica de Madrid*, 9.
- Sancho, F. (26 de 12 de 2018). *Ciencias de la computación*. Obtenido de Ciencias de la computación: <http://www.cs.us.es/~fsancho/?e=77>
- Sarmiento, E., & Silva, D. (2017). *Repositorio Institucional de la Universidad Francisco José de Caldas*. Obtenido de Repositorio Institucional de la Universidad Francisco José de Caldas.
- Sarmiento, E., & Silva, D. (2017). *Repositorio Institucional Universidad Distrital*. Obtenido de Repositorio Institucional Universidad Distrital : <http://repository.udistrital.edu.co/bitstream/11349/5911/1/TesisAnalisisDeFlujosDeInformacionDeLaRedSocialTwitter.pdf>
- Sas. (2019). Sas. Obtenido de Sas: [https://www.sas.com/es\\_mx/insights/analytics/machine-learning.html](https://www.sas.com/es_mx/insights/analytics/machine-learning.html)
- Sawakinome. (2018). *Sawakinome*. Obtenido de Sawakinome: <https://es.sawakinome.com/articles/business/unassigned-85.html>
- Selva, J. (09 de 2015). *Universitat Politècnica de València*. Obtenido de Universitat Politècnica de València: <https://riunet.upv.es/bitstream/handle/10251/55471/SELVA%20-%20Desarrollo%20de%20un%20sistema%20de%20análisis%20de%20sentimiento%20sobre%20Twitter.pdf?sequence=1>
- SEPLN. (2019). *SEPLN*. Obtenido de SEPLN: [http://www.sepln.org/workshops/tass/tass\\_data/download.php](http://www.sepln.org/workshops/tass/tass_data/download.php)
- Sinnexus. (2017). *Sinnexus* . Obtenido de Sinnexus : [https://www.sinnexus.com/business\\_intelligence/datamining.aspx](https://www.sinnexus.com/business_intelligence/datamining.aspx)

Speroni, R. (03 de 2017). *Universidad de la República - Uruguay*. Obtenido de Universidad de la República - Uruguay:  
[https://www.fing.edu.uy/inco/grupos/pln/prygrado/Informe\\_Speroni\\_Steglich.pdf](https://www.fing.edu.uy/inco/grupos/pln/prygrado/Informe_Speroni_Steglich.pdf)

StatCounter. (2017). *StatCounter GlobalStats*. Recuperado el 12 de 2019, de <http://gs.statcounter.com/os-market-share/mobile/ecuador>

Suárez, J. (22 de 06 de 2015). *Ecuavisa*. Obtenido de Ecuavisa:  
<https://www.ecuavisa.com/articulo/televistazo/noticias/114266-guayaquil-cuna-negocios>

Torra, V. (2015). *Instituto de Investigación en Inteligencia Artificial*. Obtenido de Instituto de Investigación en Inteligencia Artificial:  
[http://www.fgcsic.es/lychnos/es\\_ES/articulos/inteligencia\\_artificial](http://www.fgcsic.es/lychnos/es_ES/articulos/inteligencia_artificial)

Torres, E. (2017). *Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto*. Obtenido de Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto:  
[https://alicia.concytec.gob.pe/vufind/Record/UCVV\\_269957dacb9f56e75847b961f240b78c/Description](https://alicia.concytec.gob.pe/vufind/Record/UCVV_269957dacb9f56e75847b961f240b78c/Description)

Torres, J. (2018). *Deep learning: Introducción práctica con Keras*. Barcelona: WATCH THIS SPACE.

Torres, J. (10 de 06 de 2018). *Torres.ai*. Obtenido de Torres.ai: <https://torres.ai/deep-learning-inteligencia-artificial-keras/>

Twitter. (2018). *Centro de ayuda*. Obtenido de Centro de ayuda:  
<https://help.twitter.com/es/rules-and-policies/twitter-api>

Twitter. (2018). *Twitter developer documentation*. Obtenido de Twitter developer documentation: <https://dev.twitter.com/docs>

Twitter, I. (2019). *Developer*. Obtenido de Developer:  
<https://developer.twitter.com/en/docs/tweets/filter-realtime/overview>

Unipython. (2017). *Unipython*. Obtenido de Unipython: <https://unipython.com/numpy-algebra/>

Vilanova, L. (10 de 02 de 2019). *Luis Vilanova*. Obtenido de Luis Vilanova:  
<https://luisvilanova.es/twitter-y-inteligencia-artificial/>

Vilares, D. (2014). *Lengua Y Sociedad de la Información*. Obtenido de Lengua Y Sociedad de la Información: <http://www.grupolys.org/biblioteca/Vil2014a.pdf>

Villaverde. (2018). *Vviza gestion*. Obtenido de Vviza gestion:  
<https://www.vizagestion.com/2018/01/03/que-es-open-source/>

Villaverde, E. (03 de 1 de 2019). *Engagebs*. Obtenido de Engagebs:  
<https://www.engagebs.com/2018/01/03/que-es-open-source/>

Zambrano, J. (30 de 03 de 2018). *Medium*. Obtenido de Medium:  
<https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b>

## ANEXO 1. CRONOGRAMA DE ACTIVIDADES

<b>Cronograma de actividades</b>				
	<b>Actividad</b>	<b>Fecha de inicio</b>	<b>Fecha de fin</b>	<b>#de días</b>
		13/06/2019	30/08/2019	79
Capítulo I	El Problema	13/06/2019	24/06/2019	12
	Establecer la limitación del problema	13/06/2019	17/06/2019	5
	Plantear y definir precisamente el problema	18/06/2019	24/06/2019	7
Capítulo II	Marco teórico	25/06/2019	16/07/2019	22
	Definir las herramientas con las que se van a trabajar para optimizar el proyecto	25/06/2019	29/06/2019	5
	Investigar los antecedentes del caso para establecer un punto de partida	30/06/2019	04/07/2019	5
	Considerar los distintos proyectos guías que se puedan asemejar el proyecto	05/07/2019	12/07/2019	8
	Sintetizar la información obtenida con respecto a la investigación bibliográfica	13/07/2019	16/07/2019	4
Capítulo III	Propuesta tecnológica	17/07/2019	20/08/2019	35
	Establecer los parámetros y realizar un análisis de factibilidad del proyecto	17/07/2019	19/07/2019	3
	Establecer los parámetros y definir qué tan factible operacionalmente es el proyecto	20/07/2019	21/07/2019	2
	Consultar la factibilidad legal que cuenta el proyecto	22/07/2019	24/07/2019	3
	Establecer la metodología del proyecto	25/07/2019	26/07/2019	2
	Establecer algoritmo de clasificación binaria para el análisis de sentimientos.	27/07/2019	28/07/2019	2
	Desarrollar el notebook búsqueda de productos o servicios.	29/07/2019	04/08/2019	7
	Establecer algoritmo de clasificación de sectores/industrias por patrones de consumo	05/08/2019	07/08/2019	3
	Realizar las pruebas a los algoritmos. Y comparar el resultado con las pruebas que esperadas	08/08/2018	10/08/2018	3
	Desarrollar el notebook de clasificación de sectores/industrias.	11/08/2019	20/08/2019	10
Capítulo IV	Resultado	21/08/2019	30/08/2019	10
	Presentar los resultados del proyecto	21/08/2019	24/08/2019	4
	Presentar manual de usuario	25/08/2019	26/08/2019	2
	Establecer las conclusiones del proyecto	27/08/2019	30/08/2019	4

## ANEXO 2. ENCUESTAS

### ENCUESTA N.1

#### ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENEREN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON.

Proyecto de tesis de estudiante de la Universidad de Guayaquil.

Sexo:

- Mujer
- Hombre

¿Posee cuenta en la Red Social Twitter?

- Sí
- No

¿Con que frecuencia emite tweets en su cuenta de Twitter?

- Siempre
- Casi siempre
- A veces
- Casi nunca
- Nunca

¿Se considera un usuario activo en la Red Social Twitter?

- Sí
- No
- Tal vez

¿Cree que actualmente puede expresar libremente su opinión sobre cualquier tema en la red social Twitter? \*

- Sí
- No
- Tal vez

¿Alguna vez ha expresado en la red social Twitter la necesidad o deseo sobre un bien o servicio? \*

- Sí
- No
- Tal vez

¿Con qué frecuencia ha expresado su necesidad o deseo sobre un bien o servicio en la red social Twitter? \*

- Siempre
- Casi siempre
- A veces
- Casi nunca
- Nunca

¿Cuándo usted adquiere, consume o utiliza algún bien, producto o servicio lo ha publicado en la red social Twitter? \*

- Siempre
- Casi siempre
- A veces
- Casi nunca
- Nunca

¿Con qué frecuencia ha expresado usted su opinión sobre un producto o servicio en la red social Twitter? \*

- Siempre
- Casi siempre
- A veces

¿Sabía usted que sus publicaciones en la red social Twitter pueden ayudar en \* la toma de decisiones a nivel empresarial y de emprendimiento?

Sí

No

¿Está usted de acuerdo que sus tweets públicos (con características específicas) sean analizados para brindar apoyo en la toma de decisiones en los emprendimientos de la ciudad de Guayaquil? \*

Sí

No

## ENCUESTA N.2

### ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENEREN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON.

Proyecto de tesis de estudiante de la Universidad de Guayaquil.

Sexo: \*

- Mujer
- Hombre

Seleccione su situación actual: \*

- Emprendimiento
- Negocio establecido
- Me llegó por error esta encuesta

Ubique el sector de su negocio o emprendimiento en la siguiente lista: \*

- Agricultura
- Alimentación
- Manufactura
- Comercio
- Construcción
- Educación
- Salud
- Tecnológicos
- Otros

¿Sabía usted que Ecuador se posiciona como líder en la región en el índice de Actividad Emprendedora Temprana? \*

Sí

No

¿En su emprendimiento, utiliza las redes sociales para promocionar su servicio o producto y mantener contacto con sus clientes? \*

Sí

No

¿En su emprendimiento o negocio establecido, hace uso de la red social Twitter? \*

Sí

No

En caso de usar Twitter, indique la frecuencia con la que usted lee comentarios de sus clientes. \*

Bastante

Mucho

Poco

Muy poco

Nada

Considera usted que el tiempo que emplea para leer las opiniones en Twitter de sus potenciales clientes es suficiente para determinar gustos y necesidades. \*

Totalmente de acuerdo

De acuerdo

Ni de acuerdo ni en desacuerdo

En desacuerdo

Totalmente en desacuerdo

Considera usted que el tiempo que emplea para leer las opiniones en Twitter \* de sus potenciales clientes es suficiente para determinar gustos y necesidades.

- Totalmente de acuerdo
- De acuerdo
- Ni de acuerdo ni en desacuerdo
- En desacuerdo
- Totalmente en desacuerdo

+++  
¿Si existiera una plataforma web que mediante su uso le permita visualizar que sectores/industrias tienen mayor actividad con los usuarios de la red social Twitter, usted la usaría? \*

- Sí
- No
- Tal vez

Si existe una plataforma web que mediante su uso le permita visualizar información que brinde apoyo en la toma decisiones al momento de emprender, ¿estaría dispuesto a usarla y con qué frecuencia? \*

- Bastante
- Mucho
- Poco
- Muy poco
- Nada

¿Qué tipo de información cree usted que es determinante para el apoyo en la toma de decisiones al momento de emprender? \*

- Información por sectores
- Información por tendencias
- Información por búsqueda de producto/servicio
- Todas las anteriores

### ANEXO 3. CRITERIOS DE VALIDACIÓN DE LA PROPUESTA

Guayaquil, 23 agosto del 2019

**EXPERTO N.- 01**

#### **CONSTANCIA DEL JUICIO DE EXPERTO**

Quien suscribe, **JOSÉ LUIS SÁNCHEZ BAQUE** con cedula de identidad N.- **0931147284**, de profesión **INGENIERO EN CIENCIAS COMPUTACIONALES**, ejerciendo actualmente como **DESARROLLADOR DE SOFTWARE** en la empresa CURSOR.

Por medio de la presente hago constar que he realizado la evaluación del proyecto de minería de datos realizado por el Sr. EDINSON ANDRÉS JIMÉNEZ CÁRDENAS, estudiante no titulado de la carrera Ingeniería en Sistemas Computacionales, Facultad de Ciencias Matemáticas y Físicas de la universidad de Guayaquil, el cual se encuentra realizando el proyecto de titulación "**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON.**"

Luego de hacer las pruebas y observaciones pertinentes concluyo que dicho test es adecuado para el proyecto de minería de datos realizado.

**C U R S O R**



—  
RUELA AUTÓDOLA

**ING. JOSÉ LUIS SÁNCHEZ BAQUE**

**C.I. 0931147284**



**UNIVERSIDAD DE GUAYAQUIL**  
**FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS**  
**CARRERA DE INGENIERIA EN SISTEMAS COMPUTACIONALES**

**Criterios de validación de la propuesta**

Este documento tiene el objetivo de validar, a través de juicios de expertos, el proyecto "Análisis de la red social Twitter para la identificación de patrones que generan oportunidades de negocio en la ciudad de Guayaquil utilizando el entorno de trabajo Jupyter Notebook y el lenguaje de programación Python."

Su opinión es muy importante para este proyecto, por favor sea honesto al responder cada uno de los siguientes ítems.

Coloque el número que usted crea conveniente dentro del casillero, de acuerdo al nivel de cumplimiento.

	<b>Cumplimiento</b>	<b>Rango</b>
5	Excelente	100%
4	Muy bueno	99% - 76%
3	Bueno	75% - 51%
2	Regular	50% - 26%
1	Insuficiente	25% - 0

<b>Ítem</b>	<b>Criterios</b>	<b>Nivel de Cumplimiento</b>
1	Herramienta de ciencia de datos amigable	95%
2	Información se guarda correctamente	95%
3	Validación en la extracción de datos	89%
4	Tiempo de procesamiento Aceptable	90%
5	Fiabilidad de datos presentados	

<b>Ítem</b>	<b>Notebooks</b>	<b>Nivel de Cumplimiento</b>
1	Extracción de tweets	95%
2	Creación y entrenamiento del algoritmo de clasificación de sentimientos.	85%
3	Búsqueda y procesamiento de opiniones	100%
4	Clasificación de las opiniones por sectores.	99%
5	Visualización de información en archivo PDF.	100%

Observaciones:

Nombre: José Luis Sánchez Baque  
Firma: \_\_\_\_\_ C.I: 0931147284

**C U R S O R**  


Guayaquil, 26 agosto del 2019

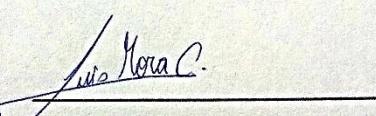
**EXPERTO N.- 02**

**CONSTANCIA DEL JUICIO DE EXPERTO**

Quien suscribe, **LUIS ALCIDES MORA CAMACHO** con cedula de identidad N.- **0952299030**, de profesión **INGENIERO EN SISTEMAS COMPUTACIONALES**, ejerciendo actualmente como **DESARROLLADOR SEMI SENIOR** en la empresa **SIPECOM S.A.**

Por medio de la presente hago constar que he realizado la evaluación del proyecto de minería de datos realizado por el Sr. EDINSON ANDRÉS JIMÉNEZ CÁRDENAS, estudiante no titulado de la carrera Ingeniería en Sistemas Computacionales, Facultad de Ciencias Matemáticas y Físicas de la universidad de Guayaquil, el cual se encuentra realizando el proyecto de titulación "**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON.**"

Luego de hacer las pruebas y observaciones pertinentes concluyo que dicho test es adecuado para el proyecto de minería de datos realizado.

  
\_\_\_\_\_  
**ING. LUIS ALCIDES MORA CAMACHO**  
**C.I. 0952299030**



UNIVERSIDAD DE GUAYAQUIL  
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS  
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

**Criterios de validación de la propuesta**

Este documento tiene el objetivo de validar, a través de juicios de expertos, el proyecto "Análisis de la red social Twitter para la identificación de patrones que generan oportunidades de negocio en la ciudad de Guayaquil utilizando el entorno de trabajo Jupyter Notebook y el lenguaje de programación Python."

Su opinión es muy importante para este proyecto, por favor sea honesto al responder cada uno de los siguientes ítems.

Coloque el número que usted crea conveniente dentro del casillero, de acuerdo al nivel de cumplimiento.

	<b>Cumplimiento</b>	<b>Rango</b>
5	Excelente	100%
4	Muy bueno	99% - 76%
3	Bueno	75% - 51%
2	Regular	50% - 26%
1	Insuficiente	25% - 0

<b>Ítem</b>	<b>Criterios</b>	<b>Nivel de Cumplimiento</b>
1	Herramienta de ciencia de datos amigable	100%.
2	Información se guarda correctamente	95%.
3	Validación en la extracción de datos	95%.
4	Tiempo de procesamiento Aceptable	85%.
5	Fiabilidad de datos presentados	90%.

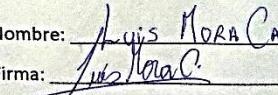
<b>Ítem</b>	<b>Notebooks</b>	<b>Nivel de Cumplimiento</b>
1	Extracción de tweets	100%.
2	Creación y entrenamiento del algoritmo de clasificación de sentimientos.	85%.
3	Búsqueda y procesamiento de opiniones	90%.
4	Clasificación de las opiniones por sectores.	90%.
5	Visualización de información en archivo PDF.	100%.

Observaciones:

---

---

Nombre: Luis Mora Capacho

Firma: 

Guayaquil, 26 agosto del 2019

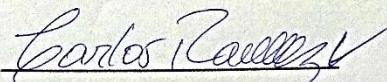
**EXPERTO N.- 03**

**CONSTANCIA DEL JUICIO DE EXPERTO**

Quien suscribe, **CARLOS ENRIQUE RAMÍREZ VÍCTOR** con cedula de identidad N.- 0930066691, de profesión **LICENCIADO EN SISTEMA DE INFORMACIÓN**, ejerciendo actualmente como **GERENTE** en el grupo **ARTEGRAFÍA AD.**

Por medio de la presente hago constar que he realizado la evaluación del proyecto de minería de datos realizado por el Sr. EDINSON ANDRÉS JIMÉNEZ CÁRDENAS, estudiante no titulado de la carrera Ingeniería en Sistemas Computacionales, Facultad de Ciencias Matemáticas y Físicas de la universidad de Guayaquil, el cual se encuentra realizando el proyecto de titulación "**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL ENTORNO DE TRABAJO JUPYTER NOTEBOOK Y EL LENGUAJE DE PROGRAMACIÓN PYTHON.**"

Luego de hacer las pruebas y observaciones pertinentes concluyo que dicho test es adecuado para el proyecto de minería de datos realizado.



LSI. CARLOS ENRIQUE RAMÍREZ VÍCTOR

C.I. 0930066691



UNIVERSIDAD DE GUAYAQUIL  
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS  
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

**Criterios de validación de la propuesta**

Este documento tiene el objetivo de validar, a través de juicios de expertos, el proyecto "Análisis de la red social Twitter para la identificación de patrones que generan oportunidades de negocio en la ciudad de Guayaquil utilizando el entorno de trabajo Jupyter Notebook y el lenguaje de programación Python."

Su opinión es muy importante para este proyecto, por favor sea honesto al responder cada uno de los siguientes ítems.

Coloque el número que usted crea conveniente dentro del casillero, de acuerdo al nivel de cumplimiento.

	<b>Cumplimiento</b>	<b>Rango</b>
5	Excelente	100%
4	Muy bueno	99% - 76%
3	Bueno	75% - 51%
2	Regular	50% - 26%
1	Insuficiente	25% - 0

<b>Ítem</b>	<b>Criterios</b>	<b>Nivel de Cumplimiento</b>
1	Herramienta de ciencia de datos amigable	90%
2	Información se guarda correctamente	90%
3	Validación en la extracción de datos	90%
4	Tiempo de procesamiento aceptable	85%
5	Fiabilidad de datos presentados	85%

<b>Ítem</b>	<b>Notebooks</b>	<b>Nivel de Cumplimiento</b>
1	Extracción de tweets	100%
2	Creación y entrenamiento del algoritmo de clasificación de sentimientos.	90%
3	Búsqueda y procesamiento de opiniones	90%
4	Clasificación de las opiniones por sectores.	85%
5	Visualización de información en archivo PDF.	100%

Observaciones:

---

---

Nombre: Carlos Ramirez

Firma: Carlos Ramirez V

## ANEXO 4. CRITERIOS DE ACEPTACIÓN DEL PRODUCTO



UNIVERSIDAD DE GUAYAQUIL  
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS  
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

### Criterios de aceptación del producto

Esta encuesta está dirigida a usuarios que poseen un negocio en la ciudad de Guayaquil o están emprendiendo, con el objeto de obtener su opinión acerca del presente proyecto.

Su opinión es muy importante para este proyecto, por favor sea honesto al responder cada uno de los siguientes ítems.

Coloque el número que usted crea conveniente dentro del casillero, de acuerdo al nivel de cumplimiento.

ítem	Criterios	SI	NO
1	¿Los datos extraídos pertenecen a la ciudad de Guayaquil?		
2	¿Aceptaría usar estos notebooks para obtener información que le permita conocer el grado de actividad y aceptación de un producto o servicio?		
3	¿Los notebooks le son útil para obtener información pública acerca de los comentarios emitidos respecto a un producto o servicio?		
4	¿Le resulta ventajoso ver los análisis de sentimientos de los distintos sectores e industrias para una toma de decisiones?		
5	¿La plataforma de ciencia de datos utilizada le pareció de fácil uso?		
6	¿Considera provechoso que este proyecto le permita conocer en qué sector e industria están interactuando los usuarios con su respectiva línea de tiempo, visualizando alza y baja en los índices de interactividad, y así poder sacar sus propias conclusiones que le permitan tomar decisiones?		
7	¿La visualización de los gráficos generados en los análisis realizados le pareció de fácil comprensión luego de la explicación realizada por el presente autor?		
8	¿La información obtenida se guarda correctamente en el archivo PDF?		
9	¿Considera usted que la información generada le permite tomar mejores decisiones de negocios o al emprender?		

Nombre: \_\_\_\_\_

Firma: \_\_\_\_\_

**ANEXO 5. MANUAL DE USUARIO**



**UNIVERSIDAD DE GUAYAQUIL**

**FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS  
CARRERA DE INGENIERÍA EN SISTEMAS  
COMPUTACIONALES**

**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN  
DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO  
EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL  
ENTORNO DE TRABAJO JUPYTER NOTEBOOK  
Y EL LENGUAJE DE PROGRAMACIÓN  
PYTHON.**

**MANUAL DE USUARIO**

**AUTOR:**

**Edinson Andrés Jiménez Cárdenas**

**TUTOR:**

**Ing. Jorge Avilés Monroy, M.Sc.**

**GUAYAQUIL – ECUADOR**

**2019**

## Introducción

El propósito del manual de usuario es explicar e informar al usuario las diferentes actividades que se realizan dentro de los notebooks desarrollados en el presente proyecto.

## Objetivos

- Detallar los notebooks del presente proyecto.
- Conocer las funciones que permite realizar cada notebook.
- Guía para los usuarios finales para su manejo dentro de la plataforma Jupyter.

## Registrar la cuenta de desarrollador para Twitter

El primer paso para poder extraer tweets de la red social Twitter es registrar una cuenta Twitter de uso común en su sección de desarrollador, lo cual permite a la empresa Twitter asegurarse quién es usted y que va a realizar con la aplicación.

1. Se ingresa o se crea una cuenta en Twitter:

<https://apps.twitter.com/>

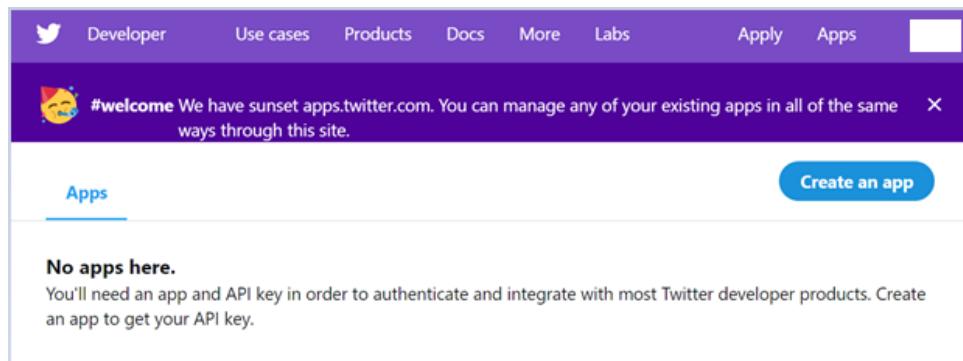
*Gráfico N. 1 Login de la sección desarrollador en Twitter.*



**Elaboración:** Jiménez Cárdenas Edinson Andrés.  
**Fuente:** (Twitter I. , 2019).

- Una vez registrados en la plataforma, se nos da la bienvenida y se procede a crear un proyecto. Al intentar crear un proyecto por primera vez, le informaran que solicite una cuenta desarrollador para poder hacer uso de las API's de Twitter. Todo esto en vías de seguir cumpliendo con la responsabilidad con sus usuarios y proporcionar un lugar que respalte la salud y seguridad de la conversación en Twitter. Evitando así el mal uso de su plataforma, además han introducido algunos requisitos nuevos para los desarrolladores.

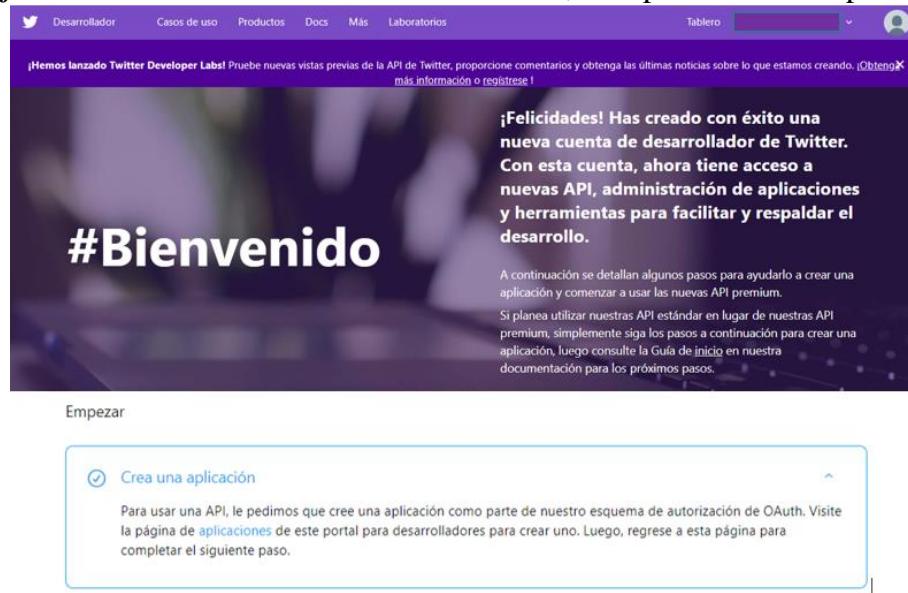
*Gráfico N. 2 Login de la sección desarrollador en Twitter.*



- En el transcurso de la creación de la cuenta desarrollador, la plataforma le ira solicitando ciertos datos de validación como número de teléfono, y que describa cómo utilizará la API de Twitter o los datos de Twitter, se recomienda responder en ingles en los campo que se solicita el ingreso de una respuesta, debido a que se requiere una aprobación de la cuenta que se ha creado para su posterior utilización, si las respuestas son efectuadas en inglés; la aprobación no tomara más de 5 horas, por experiencia propia en el desarrollo de este proyecto. En caso de haber llenado

correctamente le otorgaran acceso a la plataforma de desarrollador.

**Gráfico N. 3** Cuenta de desarrollador habilitada, lista para crear una aplicación.



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Twitter I. , 2019).

4. Debido a que Twitter emplea el método de autenticación Oauth se necesita la clave de acceso a la API, es decir, API Key y API secret. Por otra parte, se requieren las claves del token (Access Token y Access Token Secret) para acceder a la aplicación creada.
5. Posteriormente se tendrá total acceso a la creación de una aplicación que nos permita hacer uso de la API Streaming y la obtención de claves y token's para la extracción de tweets. Se debe tener cuidado con el uso de las mismas y no ser expuestas en público, por cuestiones de privacidad y que nadie haga uso indebido de ellas, ya que puede incurrir en sanciones legales por parte de la empresa Twitter Inc.

**Gráfico N. 4** Llaves y token's otorgados por Twitter para acceder a sus API's.

The screenshot shows the 'Llaves y fichas' (Keys and tokens) section of the Twitter developer dashboard. It includes fields for Consumer API keys (API Key and API Secret), Access Token (Access Token and Access Token Secret), and a 'Regenerado' (Regenerated) button. Below these are sections for 'Token de acceso y secreto de token de acceso' (Access token and access token secret) and a 'Leer y escribir' (Read and write) permission level.

**Claves API de consumidor**

API Key: **4p0Rm1c1ox1ekdravggqNk1c** ( clave API )  
API Secret: **6GPM00...BMEf7U...pQy0...LWTo...j...20...Q2D...p...FM...8C...4** ( clave secreta API )

**Regenerado**

**Token de acceso y secreto de token de acceso**

Access Token: **6GPM00...BMEf7U...pQy0...LWTo...j...20...Q2D...p...FM...8C...4** ( token de acceso )  
Access Token Secret: **oVg...0...97LEWl...Y...p...IVZ...2...I...TTf...f...ON...Q...j...IM...9E...6** ( secreto de token de acceso )

**Leer y escribir** ( nivel de acceso )

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Twitter I. , 2019).

6. Una vez concluido este proceso se tiene habilitado el acceso a las diferentes API's de Twitter, en nuestro caso específico al API Streaming.

Una vez que se poseen las credenciales para dar uso al API Streaming y se ha seguido el manual técnico que explica como instalar el paquete de anaconda 3, que trae instalado Jupyter y también su propia versión de python. Se procede a la explicación de uso de cada notebook.

### **Notebook 1: Downloading\_tweets**

Este notebook permite la extracción de tweets emitidos en la ciudad de Guayaquil y posteriormente realiza el almacenamiento de estos datos en Firebase Realtime.

A continuación se visualiza el inicio del notebook cuando se procede a abrirlo en Jupyter.

**Gráfico N. 5** Notebook de extracción

The screenshot shows a Jupyter Notebook interface. The title bar says 'jupyter 1\_Downloading\_tweets\_FINAL Last Checkpoint: el lunes pasado a las 3:29 (autosaved)'. The menu bar includes File, Edit, View, Insert, Cell, Kernel, Help. The toolbar has icons for New, Open, Save, Run, Cell, Kernel, Help, and Markdown. The status bar shows 'Trusted' and 'Python 3'. The main area contains a text cell with the following content:

```
Almacenar tweets públicos de la ciudad de Guayaquil .  
Firebase Realtime.
```

Below this is another text cell with the heading 'Paso 1: Instalar librerías' and the command 'In [1]: pip install tweepy' followed by its output:

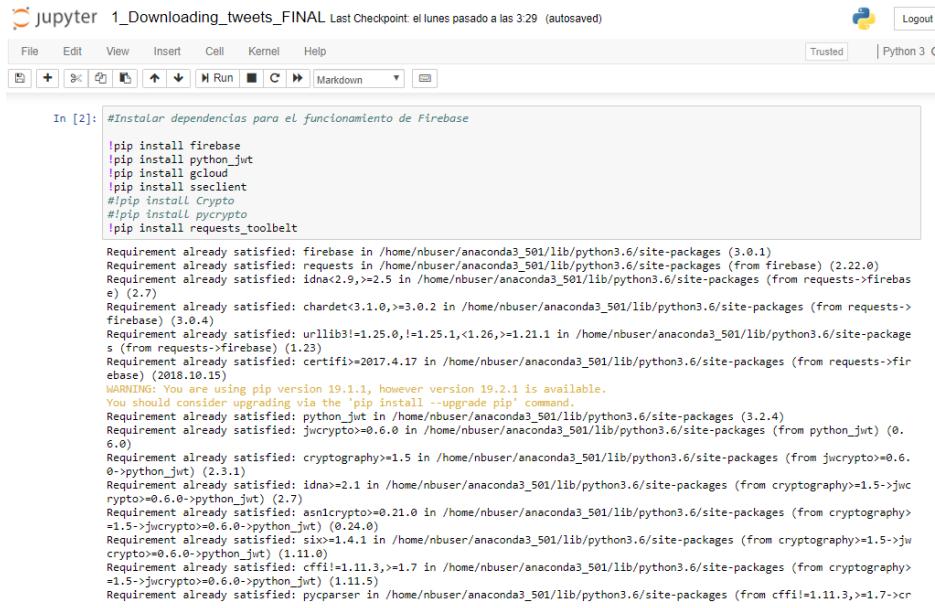
```
In [1]: pip install tweepy  
Collecting tweepy  
  Downloading https://files.pythonhosted.org/packages/36/1b/2bd38043d22ade352fc3d3902cf30ce0e2f4bf285be3b304a2782a767aec/tweepy-3.8.0-py3-none-any.whl  
Requirement already satisfied: six>=1.10.0 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from tweepy) (1.11.0)  
Requirement already satisfied: requests-oauthlib>=0.7.0 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from tweepy) (1.2.0)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.  
**Fuente:** Jiménez Cárdenas Edinson Andrés.

Los notebooks están desarrollados para ser ejecutados en cualquier plataforma que permita el uso de los mismos.

Por eso cada paquete instalado en el presente proyecto está escrito en el bloque de código de cada notebook, para que el usuario final no tenga que lidiar con la visualización de errores por parte de librerías no encontradas.

### Gráfico N. 6 Importación de librerías



The screenshot shows a Jupyter Notebook interface with the title "jupyter 1\_Downloading\_tweets\_FINAL Last Checkpoint: el lunes pasado a las 3:29 (autosaved)". The menu bar includes File, Edit, View, Insert, Cell, Kernel, and Help. On the right, there are icons for Python 3, Logout, Trusted, and a refresh button. Below the menu is a toolbar with various icons. The main area is titled "In [2]". The code cell contains the following Python code:

```
#Instalar dependencias para el funcionamiento de Firebase
!pip install firebase
!pip install python_jwt
!pip install gcloud
!pip install googleclient
#!pip install Crypto
#!pip install pycrypto
!pip install requests_toolbelt
```

Output from the command shows various package requirements and their versions, such as:

```
Requirement already satisfied: firebase in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (3.0.1)
Requirement already satisfied: requests in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from firebase) (2.22.0)
Requirement already satisfied: idna<2.9,>=2.5 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from requests->firebase) (2.7)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from requests->firebase) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,<1.26,>=1.21.1 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from requests->firebase) (1.23)
Requirement already satisfied: certifi>=2017.4.17 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from requests->firebase) (2018.10.15)
WARNING: You are using pip version 19.1.1, however version 19.2.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
Requirement already satisfied: python_jwt in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (3.2.4)
Requirement already satisfied: jwcrypto>=0.6.0 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from python_jwt) (0.6.0)
Requirement already satisfied: cryptography>=1.5 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from jwcrypto>=0.6.0->python_jwt) (2.3.1)
Requirement already satisfied: idna>=2.1 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from cryptography>=1.5->jwcrypto>=0.6.0->python_jwt) (2.7)
Requirement already satisfied: asn1crypto>=0.21.0 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from cryptography>=1.5->jwcrypto>=0.6.0->python_jwt) (0.24.0)
Requirement already satisfied: aio>=1.4.1 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from cryptography>=1.5->jwcrypto>=0.6.0->python_jwt) (1.11.0)
Requirement already satisfied: cffi!=1.11.3,>=1.7 in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from cryptography>=1.5->jwcrypto>=0.6.0->python_jwt) (1.11.5)
Requirement already satisfied: pycparser in /home/nbuser/anaconda3_501/lib/python3.6/site-packages (from cffi!=1.11.3,>=1.7->cryptography)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Una vez instaladas las librerías y dependencias, el entorno de trabajo está listo para ejecutar los siguientes bloques de código

En caso de que el usuario final desee modificar el proceso de extracción, en este bloque debe hacer uso de las credenciales y token's obtenidos por Twitter para usar sus API's.

### *Gráfico N. 7 Incorporación de key's y token's*

#### Paso 2: Preparando Tweepy para capturar tweets

```
In [17]: import tweepy
import sys
import csv
import json
import os
import time

from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener

In [18]: #Imports de Firebase
#from Crypto.Publickey import RSA
from firebase import firebase

In [19]: consumer_key = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxx'
consumer_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_token = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_token_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

El siguiente bloque de código es el encargado de la conexión con la base de datos Firebase, si ha creado su base de datos en Firebase Realtime agregue la dirección que le otorgan en la consola de Firebase.

### *Gráfico N. 8 Notebook de extracción*

```
In [20]: #Conexion a La base de datos
FBConn = firebase.FirebaseApplication('https://xxxxxxxx-247901.firebaseio.com/', None)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

En caso de que el usuario lo desee puede expandir la caja delimitadora y ampliar el ratio de filtrado. Posteriormente se debe ejecutar el último bloque de código y usted estará extrayendo tweets de la presente ciudad y almacenándolos en Firebase Realtime.

## Gráfico N. 9 Coordenadas geográficas

### Paso 3: GEOLOCALIZACIÓN (Ciudad de Guayaquil)

```
In [22]: GUAYAQUIL_GEO = [-80.088362, -2.240213, -79.891261, -2.061090]
```

### Paso 4: Sorber del firehouse

```
In [ 1]: #filter_track = ['tesla', 'wall']
file_name = 'tweets_gye.csv'
#twitter_stream_listener(file_name, filter_track, time_limit=36000)
twitter_stream_listener(file_name, GUAYAQUIL_GEO, time_limit=36000)
```

El script capturará todos los tweets que encajan dentro de la caja que configuramos.

Una cosa importante a tener en cuenta es que la API no es 100% exacta en los datos que devuelve. Se puede encontrar varios tweets geocodificados que no pertenecían a la caja especificada.

Como el script tiene que estar ejecutándose para poder capturar todos los tweets, puedes ejecutarlo en una computadora de repuesto si tienes una, o alternativamente puedes considerar servicios en línea como Google Colaboratory o PythonAnywhere, o alquilar su propia máquina en la nube con servicios como Amazon Web Services.

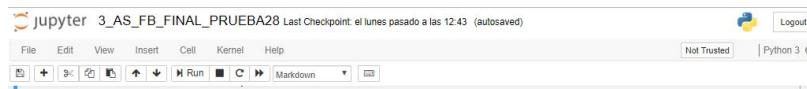
**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

## Notebook 2: Análisis\_sentimientos

Una vez extraído una cantidad considerable de tweets, se procede a hacer uso de los notebooks de análisis de datos. Este notebook incorpora el modelo de análisis de sentimientos en español generado para fines del presente proyecto, en el inicio del notebook obtenemos la siguiente imagen.

### Gráfico N. 10 Notebook de análisis de sentimientos



Análisis de Sentimiento a tweets públicos de la ciudad de Guayaquil .



#### Tabla de contenidos

1. [Instalar dependencias para el funcionamiento de Firebase](#)
2. [Conexión y extracción de tweets de Real Time Firebase](#)
3. [Procesamiento de los Tweets extraídos](#)
4. [Extraer los hashtags de cada tweet](#)
5. [Data Cleaning \(limpieza de los datos\)](#)
6. [Búsqueda por productos y servicios](#)
7. [Importar el modelo de clasificación de sentimientos](#)
8. [Visualización de tendencias más frecuentes por sentimientos](#)
9. [Explorando tweets positivos y negativos](#)
10. [Valoración final de sentimientos de tweets](#)

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

El presente notebook incorpora dos bloques de instalación de librerías que una vez instaladas nos permitirán hacer uso del notebook y sus funcionalidades. Las cuales al ejecutar cada bloque se irán instalando. En el paso 2 se procede a extraer los tweets de Firebase y convertir el objeto json devuelto en un DataFrame para una fácil manipulación de los datos.

### **Gráfico N. 11 Extracción de datos desde Firebase**

**Paso 2: Conexión y extracción de tweets de Real Time Firebase**

```
In [4]: FBConn = firebase.FirebaseApplication('https://deeptesis-247901.firebaseio.com/', None)
#firebase = firebase.FirebaseApplication('https://xxxx.firebaseio.com/', None)
result = FBConn.get('tweets', '')
```

- Se debe eliminar la última declaración de impresión "print(result)", debido a que excede el límite de 10 MB en el cuaderno.

```
In [5]: #Convertir el archivo Json extraido de Firebase en DataFrame
df_fb = pd.DataFrame(result)
#Transponer Dataframe
df2_fb=df_fb.T
```

```
In [12]: #df2_fb.head(3)
```

```
In [7]: #Cantidad de filas del Dataframe
len(df2_fb)
```

```
Out[7]: 54939
```

```
In [8]: #Convertir DataFrame a archivo -----CSV df.to_csv(path='Final_22.csv', encoding ='utf-8')
path = 'corpus_give/'
df2_fb.to_csv(path+"tweets_fb.csv", encoding ='utf-8')
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Posterior a la obtención de los datos se debe aplicar los procesos de limpieza de datos, los cuales corresponden al paso 5, descrito en el notebook.

### **Gráfico N. 12 Proceso de limpieza de datos**

**Paso 5: Data Cleaning (limpieza de los datos)**

```
In [9]: # remove username
df["tweetText"] = df["tweetText"].str.replace(r"(@)[A-Za-z0-9_]+", "")
df["tweetText"] = df["tweetText"].dropna()
```

```
In [10]: # función para limpiar tweets
import re
import itertools

def tweet_clean(x):
    tweet = re.sub('((www\.[^\s]+)|(https?://[^ \s]+))', '', x) # remove URLs
    sinUrl = re.sub('http[s]?://(?:[a-zA-Z][0-9]|[$-_@.&#39;][!%\\(\v,\])|^(?:[0-9a-fA-F][0-9a-fA-F])+)', '', tweet)
    #Eliminar comillas
    comillas = re.sub([''''<>>'], '', sinUrl)
    #Eliminar salto de líneas
    sinSalto = re.sub('\n', ' ', comillas)
    #Separar palabras juntas. Transforma AguanteRiver en Aguante River
    separar = ("".join(re.findall('[A-Z][A-Z]*', sinSalto)))
    #Estandarizar palabras. Frases como "estoy muuuuy feliz" las convertimos en "estoy muy feliz"
    estandarizar = ''.join([s[1:2] for _, s in itertools.groupby(separar)])
    #Convertir a minuscula todos los tweets
    minuscula= estandarizar.lower()
    # retorna tweet Limpio
    return minuscula
```

```
In [11]: df["tweetFinal"] = df["tweetText"].apply(tweet_clean)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Luego de haber ejecutado cada bloque de código, deténgase en el paso 6, el cual le solicitará que ingrese el nombre de un producto o servicio para su posterior análisis en base a sus datos extraídos.

### *Gráfico N. 13* Ingreso de búsqueda por teclado

#### Paso 6: Búsqueda por productos y servicios

- Por favor no escribir cosas no relacionadas o el Kernel no podrá generar una búsqueda acertada.

```
In [28]: inputSearch = input("Ingrese Su Búsqueda: ").lower()
#inputSearch = str(input("Ingrese Su Búsqueda: "))
filtro = df[df['tweetFinal'].str.contains(inputSearch, case=False)]
len(filtro)

Ingres Su Búsqueda: sushi
Out[28]: 17
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Luego se procede a importar el algoritmo clasificador de sentimientos, para darle una polaridad a cada tweet, de acuerdo a su semántica.

### *Gráfico N. 14* Importación de modelo de clasificación de sentimientos

```
In [29]: # Dado un texto predice su sentimiento
def obtener_sentimiento(texto):
    modelo = open("corpus_aye/corpus_prog.m", "rb")
    text_clf2 = pickle.load(modelo)
    modelo.close()

    newTexto = [(texto)]
    sentimiento = text_clf2.predict(newTexto)

    sentimiento = str(sentimiento)
    sentim = re.sub('[\[\]]', '', sentimiento)
    sentim2 = re.sub('^\'', '', sentim)
    return(sentim2)

In [30]: #len(filtro)
filtrado = pd.DataFrame(filtro)

In [31]: #APLICA FUNCIÓN DE OBTENER SENTIMENTO
filtrado["prediction"] = filtrado["tweetText"].apply(obtener_sentimiento)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Se aplica el algoritmo de clasificación de sentimientos a todos los tweets, luego de varios minutos obtendremos un dataset clasificado.

Posteriormente se procede a presentar la información de manera gráfica para una mayor comprensión de la información generada. Entre las gráficas

obtenidas tenemos, gráfico de barras para la visualización de las tendencias relacionadas con el producto o servicio buscado.

**Gráfico N. 15** Generación de gráficos de barras



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

El siguiente gráfico que obtenemos es una nube de palabras más utilizadas cuando se emiten tweets con el nombre del producto o servicio buscado. De este tipo de gráfico se generan dos, uno con las palabras más utilizadas en tweets categorizados como positivos y otro gráfico con las palabras más utilizadas en tweets categorizados como negativos.

**Gráfico N. 16** Nube de palabras más usadas

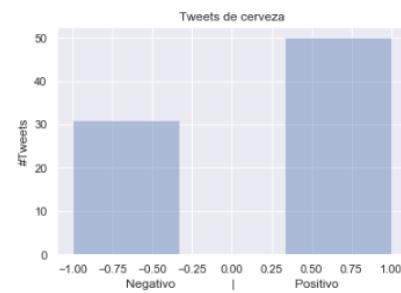


**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Siguiendo el flujo de ejecución de bloques obtenemos una gráfica final de la valoración de sentimientos, que nos da una visualización acerca del sentimiento encontrado sobre un producto o servicio, en base a los datos extraídos previamente

**Gráfico N. 17** Gráfico total de sentimientos



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

### Notebook 3: clasificacion\_sector

El presente Notebook hace uso de una red neuronal para lograr la clasificación de tweets por sectores, basándose en patrones comerciales previamente prescritos por categorías. A continuación se detalla el flujo de trabajo del notebook

Se importan las librerías que hace uso el código escrito en el notebook. Y se procede a ejecutar todos los bloques de código que incluye la red neuronal

**Gráfico N. 18** Red neuronal del proyecto

#### Paso 2: Red Neuronal

```
In [6]: words = []
classes = []
documents = []
ignore_words = ['i?']
# Bucle mediante el cual cada frase va al data training
for pattern in training_data:
    # tokenizar cada palabra de la oración
    w = nltk.word_tokenize(pattern['sentence'])
    # add to our words list
    words.extend(w)
    # agregar documentos en nuestro corpus
    documents.append((w, pattern['class']))
    # agrega a nuestra lista de clases
    if pattern['class'] not in classes:
        classes.append(pattern['class'])

# stemming y minusculas a cada palabra y remover duplicados
words = [stemmer.stem(w.lower()) for w in words if w not in ignore_words]
words = list(set(words))

# remover duplicados
classes = list(set(classes))

print (len(documents), "documents")
print (len(classes), "classes", classes)
#print (len(words), "unique stemmed words", words)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Una vez que se han ejecutado los bloques de código perteneciente a la red neuronal, el modelo está listo para ser aplicado a los tweets y categorizar solo a aquellos tweets en los que detecte un patrón de consumo a un producto perteneciente a un sector e industria. A los tweets que no posean ninguna de las características antes mencionadas se los cataloga como irrelevantes.

### **Gráfico N. 19** Entrenamiento de la red neuronal

```
In [32]: X = np.array(training)
y = np.array(output)

start_time = time.time()
#base 20| ultima 80
train(X, y, hidden_neurons=10, alpha=0.1, epochs=100000, dropout=False, dropout_percent=0.2)

elapsed_time = time.time() - start_time
print ("processing time:", elapsed_time, "seconds")

Training with 10 neurons, alpha:0.1, dropout=False
Input matrix: 193x603 Output matrix: 1x8
delta after 10000 iterations:0.020727063880344116
delta after 20000 iterations:0.0014273074951094615
delta after 30000 iterations:0.001151024128386476
delta after 40000 iterations:0.0009889917611635207
delta after 50000 iterations:0.0008795690597870302
delta after 60000 iterations:0.0007994103183379701
delta after 70000 iterations:0.0007374801172901102
delta after 80000 iterations:0.0006878006379426522
delta after 90000 iterations:0.0006468173451536323
delta after 100000 iterations:0.0006122676261427977
saved synapses to: synapses.json
processing time: 95.19737124443054 seconds
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Luego este notebook hace uso de los pasos mencionados en el notebook anterior, como lo es instalación de librerías para su correcto funcionamiento y extracción de los datos almacenados en Firebase, se aplican los pasos de limpieza y extracción siguiendo el flujo del Notebook anterior.

Una vez llegado al paso 5 se aplica el modelo de clasificación por sectores, lo cual permite clasificar los tweets que presenten patrones de consumo de un producto perteneciente a un sector.

## **Gráfico N. 20** Aplicación de la red neuronal para clasificación

### **Paso 5: Aplicacion de red neuronal para clasificar tweets por sectores**

- Tiempo de espera 4 minutos con 54000 tweets

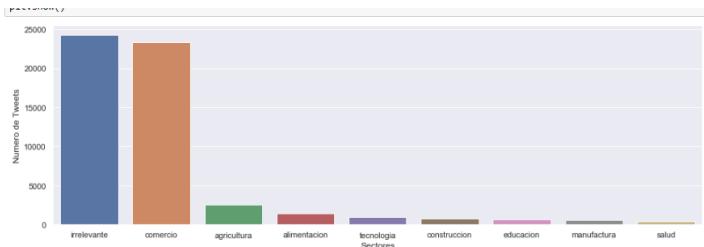
```
In [41]: #Aplica Clasificacion de tweets por sector  
df_copy["sector"] = df_copy["tweetFinal"].apply(classify)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Una vez aplicado el algoritmo clasificador de tweets por sectores e industrias. Se presenta un gráfico de Barras para una visualización rápida de cuáles son los sectores o industrias en los que más se ha detectado índices de consumo.

## **Gráfico N. 21** Grafico de barras de índice de consumo por sectores



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Siguiendo el flujo de trabajo del notebook se procede a hacer uso del modelo de clasificación de sentimientos creado en el notebook anterior. Permitiendo agregar una polaridad a cada tweet perteneciente a un sector o industria.

## **Gráfico N. 22 Importación de modelo de clasificación de sentimientos**

### **Paso 6: Importar el modelo para clasificación de sentimientos**

- Tiempo de espera 7 minutos con 54000 tweets

```
In [46]: # Dado un texto predice su sentimiento
def obtener_sentimiento(texto):
    #modelo = open("/content/corpus_prog.m", "rb") # Google Colab
    modelo = open("corpus_gye/corpus_prog.m", "rb")
    text_cif2 = pickle.load(modelo)
    modelo.close()

    newTexto = [(texto)]
    sentimiento = text_cif2.predict(newTexto)

    sentimiento = str(sentimiento)
    sentimiento = re.sub(r'[\[\]]', '', sentimiento)
    sentimiento2 = re.sub(r'\'' , '', sentimiento)
    return(sentimiento2)

In [47]: df_copy["prediction"] = df_copy["tweetFinal"].apply(obtener_sentimiento)

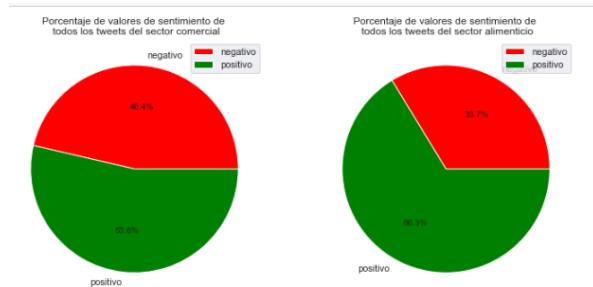
In [48]: df.to_csv(path="Final_clasificacion.csv", encoding ='utf-8')
#df.head(5)
```

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Habiendo obtenido las polaridades de cada tweet, el notebook genera gráficos que permiten visualizar los niveles de polaridad encontrada en comentarios por sector e industria, permitiendo observar los niveles de positividad o negatividad de cada sector e industria. Se generaran 4 de los gráficos mostrados a continuación, los cuales contienen los datos de dos industrias por gráficos.

## **Gráfico N. 23 Grafico de pastel de sentimientos por sectores**



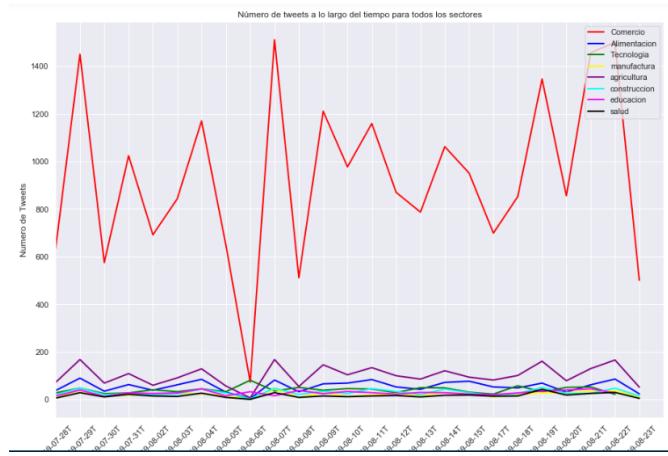
**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

Posteriormente el notebook genera un gráfico en línea de tiempo que permite a los usuarios observar que industrias son las que poseen mayor índice de interactividad y consumo en base a los comentarios obtenidos, suministrando

información relevante que permita a los emprendedores y dueños de negocios generar sus propias conclusiones y tomar en base a su criterio decisiones acertadas.

**Gráfico N. 24** Línea de tiempo de índice de consumo



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

**ANEXO 6. MANUAL TÉCNICO**



**UNIVERSIDAD DE GUAYAQUIL**

**FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS  
CARRERA DE INGENIERÍA EN SISTEMAS  
COMPUTACIONALES**

**ANÁLISIS DE LA RED SOCIAL TWITTER PARA LA IDENTIFICACIÓN  
DE PATRONES QUE GENERAN OPORTUNIDADES DE NEGOCIO  
EN LA CIUDAD DE GUAYAQUIL UTILIZANDO EL  
ENTORNO DE TRABAJO JUPYTER NOTEBOOK  
Y EL LENGUAJE DE PROGRAMACIÓN  
PYTHON.**

**MANUAL TÉCNICO**

**AUTOR:**

**Edinson Andrés Jiménez Cárdenas**

**TUTOR:**

**Ing. Jorge Avilés Monroy, M.Sc.**

**GUAYAQUIL – ECUADOR**

**2019**

## **Herramientas usadas para el desarrollo del proyecto**

Para el desarrollo del proyecto se usan las siguientes herramientas:

- Anaconda Distribution
- Python
- Jupyter Notebook
- Firebase

Todas estas herramientas pueden ser descargadas desde las páginas oficiales.

## **Descarga e instalación de las herramientas utilizadas**

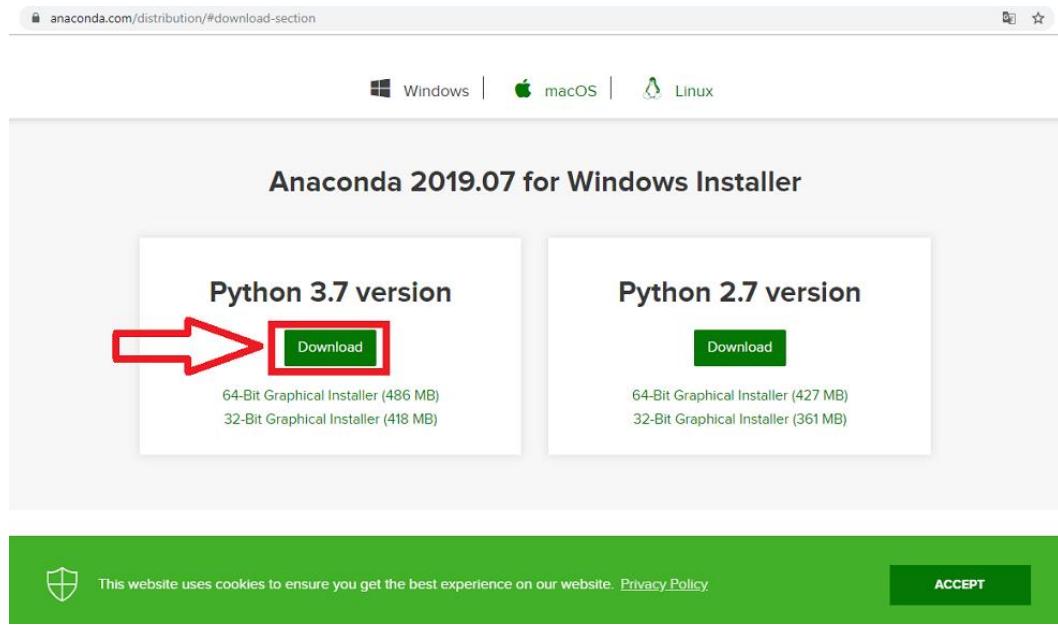
A continuación, se detalla como descargar e instalar cada una de las herramientas utilizadas para el desarrollo del proyecto.

Cabe recalcar que la instalación del paquete Anaconda Distribution ya incorpora la instalación del entorno Jupyter Notebook, y su propia versión de Python. Así que una vez instalada esta distribución solo restaría la creación de la base de datos no relacional Firebase Realtime.

### **Anaconda Distribution 2019.07 (plataforma de ciencia de datos)**

Para descargar la distribución de Anaconda se debe ingresar a la siguiente dirección: [<https://www.anaconda.com/distribution/>] como se observa en el gráfico N. 1, Anaconda permite seleccionar una versión de python para su uso en esta plataforma, se debe dar clic en la opción que se encuentra en el rectángulo de color rojo. Automáticamente se descargará el instalador.

## Gráfico N. 1 Descargar Anaconda Distribution

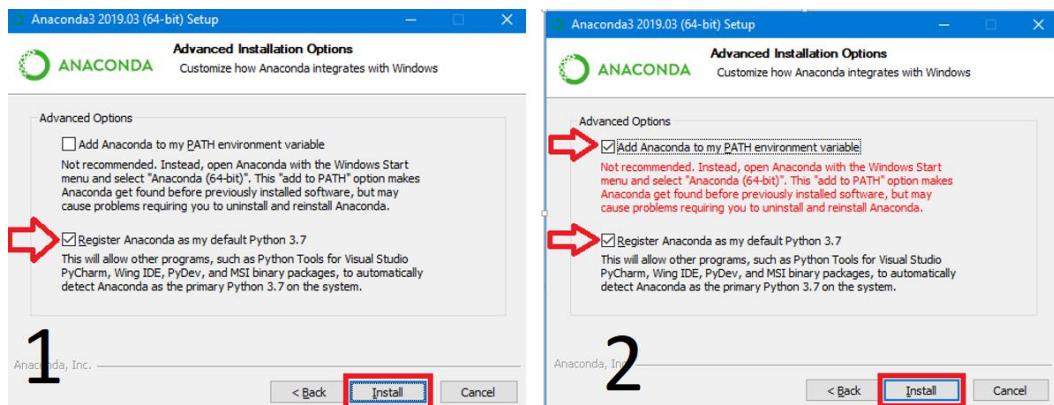


## Instalar Anaconda Distribution

1. Abrir la carpeta donde se encuentra el archivo descargado con nombre *Anaconda3-2019.07-Windows-x86\_64.exe*. Ese es el nombre del archivo en la que se redacta el presente trabajo.
2. Dar doble clic sobre el archivo y clic en la opción **ejecutar**.
3. Posteriormente se abrirá una ventana de instalación, donde deberá dar clic en **Next** (siguiente).
4. En la siguiente ventana lea los términos de licencia y haga clic en la opción **I Agree** (Acepto).

5. El siguiente paso es delimitar la instalación, aparecerá una ventana en la cual se debe escoger entre solo para su usuario o para todos los usuarios. El fabricante recomienda la opción **Just Me** (Solo su usuario), a menos que esté instalando para todos los usuarios (lo cual requiere privilegios de administrador de Windows) y haga clic en **Siguiente**.
6. Seleccione una carpeta de destino mediante el uso del botón **browser**, haga clic en el botón **Siguiente** para instalar Anaconda. Nota: Instale Anaconda en una ruta de directorio que no contenga espacios o caracteres unicode.
7. Elegir si se desea agregar Anaconda a su variable de entorno PATH. Se Recomienda no agregar Anaconda a la variable de entorno PATH, ya que esto puede interferir con otro software. En su lugar, use el software Anaconda, abriendo Anaconda Navigator o el símbolo de Anaconda desde el menú Inicio. El enfoque recomendado es no marcar la primera casilla como se visualiza en la figura N. 2, empleando la opción 1 para su uso. Si desea poder usar Anaconda en su símbolo del sistema (o git bash, cmder, powershell, etc.), utilice el enfoque alternativo y marque las dos casillas como se observa en la opción 2 de la figura N. 2.

**Gráfico N. 2 Opción de instalación avanzada**

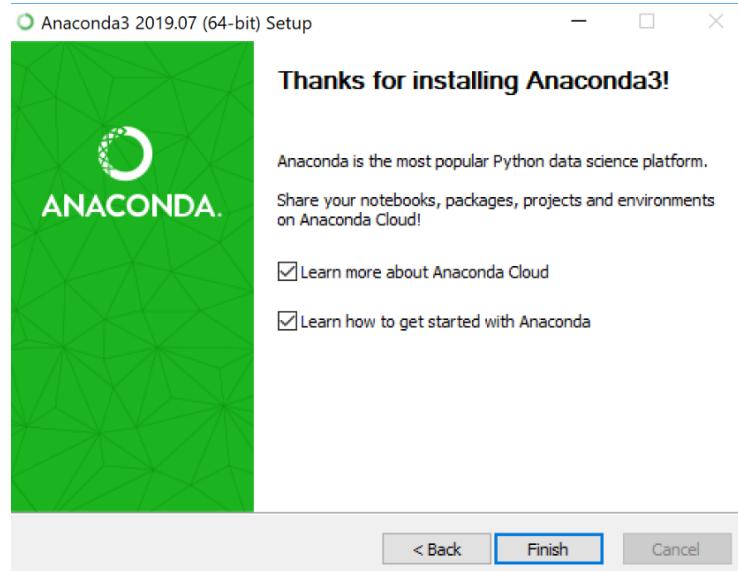


**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Anaconda, 2019).

8. Elija si desea registrar Anaconda como su versión de Python predeterminada. A menos que planee instalar y ejecutar múltiples versiones de Anaconda, o múltiples versiones de Python, acepte el valor predeterminado y deje esta casilla marcada.
9. Haga clic en el botón **Instalar**. Si desea ver los paquetes que Anaconda está instalando, haga clic en **Mostrar detalles**.
10. Opcional: para instalar PyCharm para Anaconda, haga clic en el enlace a [<https://www.anaconda.com/pycharm>]. O para instalar Anaconda sin PyCharm, haga clic en el botón **Siguiente**.
11. Después de una instalación exitosa, verá el cuadro de diálogo "Gracias por instalar Anaconda" véase en el gráfico N. 3. Si desea leer más sobre Anaconda Cloud y cómo comenzar a usar Anaconda, marque las casillas "Aprenda más sobre Anaconda Cloud" y "Aprenda cómo comenzar a usar Anaconda". Haga clic en el botón **Finalizar**.

*Gráfico N. 3 Cuadro de diálogo final de instalación.*

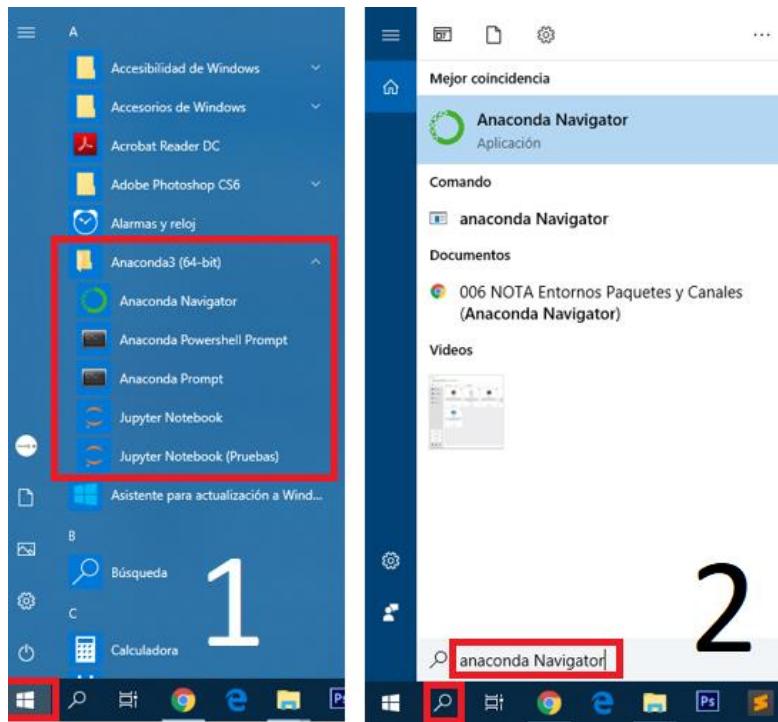


**Elaboración:** (Anaconda, 2019).

**Fuente:** (Anaconda, 2019).

12. Una vez completada la instalación, verifique abriendo Anaconda Navigator, el cual es la interfaz gráfica que incluye Anaconda. Desde el menú Inicio de Windows, seleccione el acceso directo a Anaconda Navigator desde el Agregado recientemente como se muestra en la figura N. 4, opción 1. O emplee la opción 2 usando la opción de búsqueda y escribiendo "Anaconda Navigator". Si se abre Navigator, ha instalado correctamente Anaconda. De lo contrario, verifique que haya completado cada paso anterior.

*Gráfico N. 4 Apertura de Anaconda Navigator.*



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** Jiménez Cárdenas Edinson Andrés.

## Python

Como se detalló previamente Anaconda incorpora su propia versión de Python, una vez instalado Anaconda ya se posee una versión de este lenguaje de programación, la cual usted haya seleccionado, puede ser la versión 2.7 o 3.7. Para el presente proyecto fue utilizada la versión 3.7.

### Jupyter Notebook (Entorno de desarrollo)

Jupyter Notebook es una aplicación web de código abierto que le permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo. Los usos incluyen: limpieza y transformación de datos, simulación numérica, modelado estadístico, visualización de datos y aprendizaje automático (Jupyter, 2019).

## **Instalación de Jupyter Notebook usando Anaconda**

La documentación oficial de Jupyter recomienda la instalación de Python y Jupyter utilizando la distribución Anaconda, que incluye Python, el Notebook Jupyter, y otros paquetes de uso común para la computación científica y la ciencia de datos.

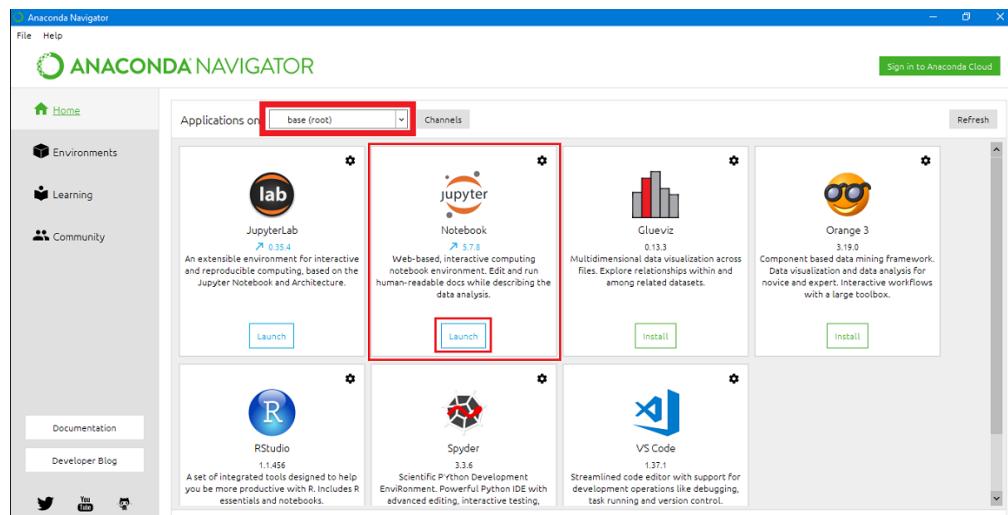
- Primero se descarga Anaconda, cuya instalación se ha detallado previamente. También se recomienda descargar la última versión de Python 3 de Anaconda.
- Segundo, instale la versión de Anaconda que descargó, siguiendo las instrucciones ya especificadas.

Una vez instalado Anaconda Distribution, se obtiene Anaconda Navigator que es la interfaz gráfica de usuario (GUI) de escritorio incluida en la distribución Anaconda que le permite iniciar aplicaciones y administrar fácilmente paquetes, entornos virtuales, sin usar comandos de línea de comandos.

Gracias a la instalación de Anaconda Distribution se obtiene Jupyter Notebook sin necesidad de instalarlo. Para proceder a usar Jupyter notebook se deben seguir los siguientes pasos:

1. Abrir Anaconda Navigator. En el gráfico N. 5, se puede observar la ventana principal de Anaconda Navigator. En el entorno raíz (**root**) viene preinstalado Jupyter Notebook.

**Gráfico N. 5** Anaconda Navigator.

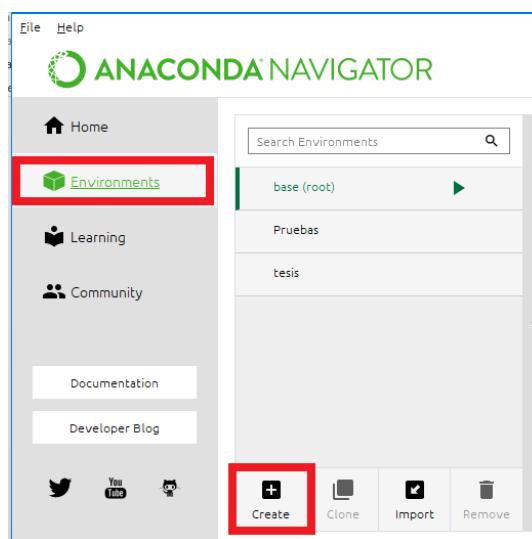


**Elaboración:** (Anaconda, 2019).

**Fuente:** (Anaconda, 2019).

2. Procedemos a crear un entorno de desarrollo dando clic en Environments y clic en **create**, como se detalla en el gráfico N. 6.

**Gráfico N. 6** Apertura de Anaconda Navigator.

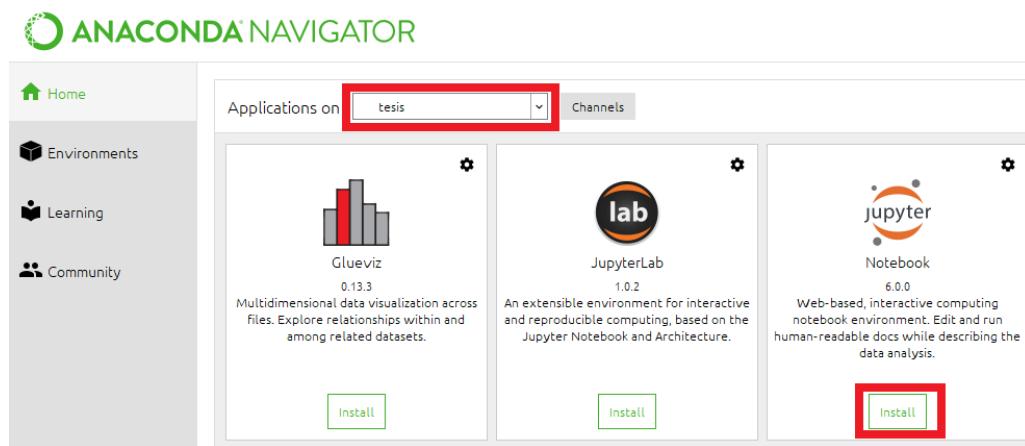


**Elaboración:** (Anaconda, 2019).

**Fuente:** (Anaconda, 2019).

3. Una vez creado el nuevo entorno, procedemos a utilizarlo; seleccionando el nombre de nuestro entorno en la opción **Applications on**. Posteriormente se debe instalar Jupyter Notebook en este nuevo entorno dando clic en la opción **install**, debido a que viene preinstalado en el entorno virtual **root**. Y cada vez que se cree un nuevo entorno virtual se debe instalar en caso de requerir esta aplicación. En el gráfico N. 7 se visualiza este proceso.

*Gráfico N. 7 Selección de entorno virtual e instalación de Jupyter.*

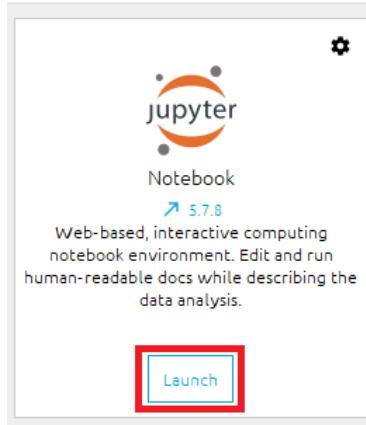


**Elaboración:** (Anaconda, 2019).

**Fuente:** (Anaconda, 2019).

4. Una vez instalado Jupyter Notebook nos aparecerá la opción **Launch** como se puede ver en el gráfico N. 8, debemos darle clic para empezar a utilizar la aplicación.

*Gráfico N. 8* Opción para abrir Jupyter en el navegador.

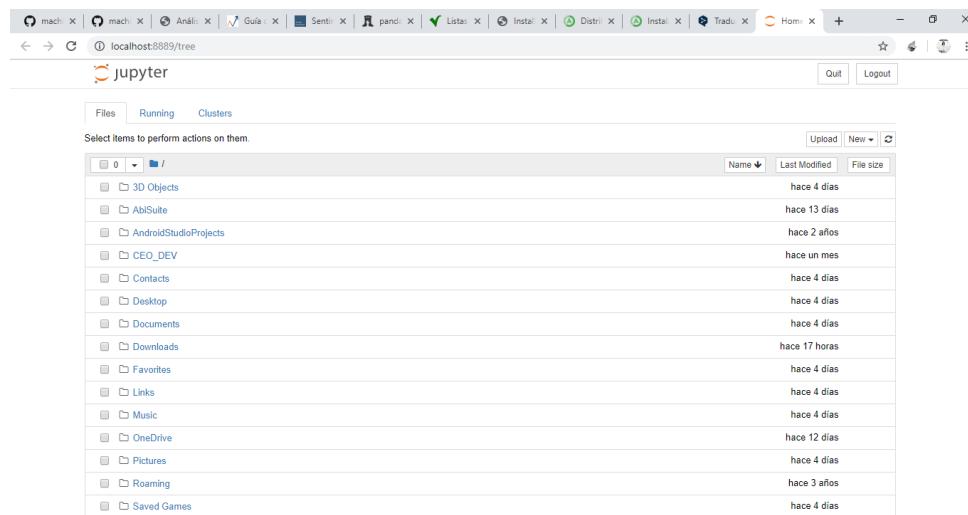


**Elaboración:** (Anaconda, 2019).

**Fuente:** (Anaconda, 2019).

5. La aplicación web se ejecutara en el navegador web que haya definido como predeterminado, presentando las carpetas contenidas en su usuario, posteriormente deberá buscar la carpeta donde haya almacenado los notebooks del presente proyecto y continuar con su uso. Para más detalles del uso de los notebooks desarrollados en este proyecto, revisar el manual de usuario

*Gráfico N. 9* Jupyter en el navegador.



**Elaboración:** (Anaconda, 2019).

**Fuente:** (Jupyter , 2019).

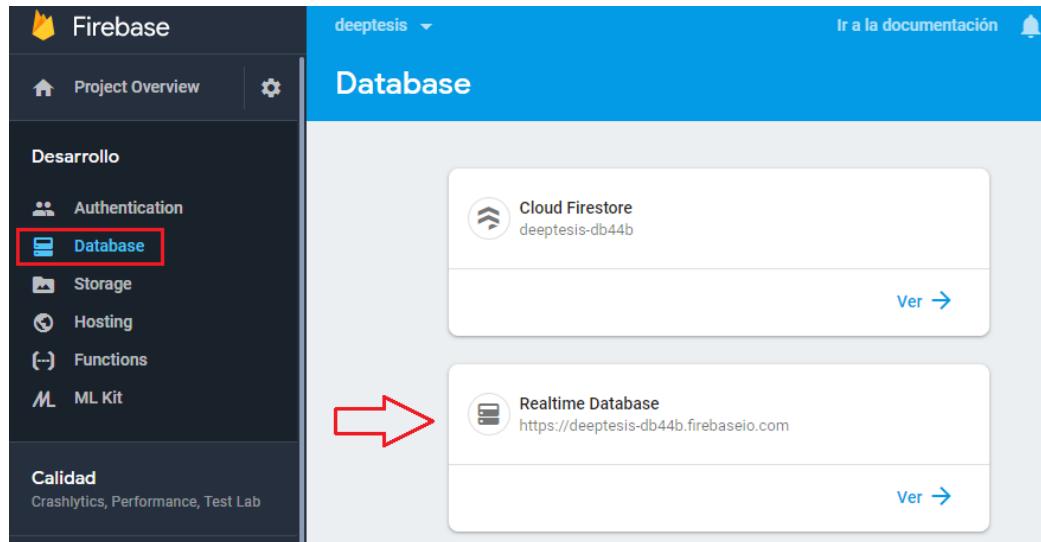
## Firebase

En Firebase se tiene dos tipos de bases de datos: Realtime Database y Cloud Firestore. Para el presente proyecto se hizo uso del primer tipo. Se debe destacar que las dos son bases de datos no relacionales (NoSQL).

Pasos para crear la base de datos Realtime en Firebase:

1. Para registrarse en Firebase lo podemos hacer con nuestra cuenta de Google.
2. Se accede a <https://firebase.google.com/> y se ingresa con la cuenta de Google.
3. Posteriormente se debe ir a la Consola de Firebase.
4. Crear un nuevo proyecto, con su respectivo nombre y región donde se va a almacenar de forma definitiva la base de datos a crear.
5. En el apartado de desarrollo seleccionar Database y seleccionar Realtime Database como se muestra en el gráfico N. 10.

**Gráfico N. 10** Creación de Realtime Database.



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Google Developers, 2019).

6. A continuación se procede a dotar de un nombre al árbol principal que va a albergar todos nuestros registros.
7. A diferencia de una base de datos de SQL, en Firebase no hay tablas ni registros. Cuando se agrega datos al árbol JSON, estos se convierten en un nodo de la estructura JSON existente con una clave asociada. A continuación en el grafico N. 11 se detalla la estructura empleada en el almacenamiento de los tweets.

**Gráfico N. 11** Estructura Json de la base de datos no relacional.

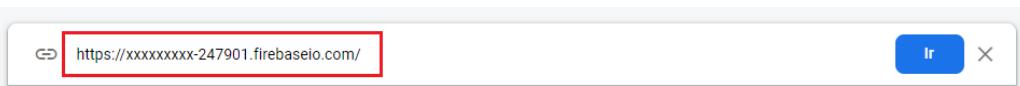
The screenshot shows the Firebase Realtime Database interface. At the top, there's a header with 'Deeptesis' and a dropdown, followed by tabs for 'Realtime Database' and 'Database'. Below that, there are tabs for 'Datos', 'Reglas', 'Copias de seguridad', and 'Uso'. The main area shows a tree view of data under 'tweets'. One node is expanded, showing fields: 'date' (2019-08-23T01:32:1), 'location' (Guayaquil, Ecuador), 'tweetID' (116471192550004326), 'tweetText' (@comandante Que se decida @fabian142), and 'typeTweet' (respuesta). Other nodes in the tree include 1164711893895917569 and 1164712016126328833.

**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Google Developers, 2019).

8. Una vez creada la base de datos se procede a copiar la dirección de la base de datos en el notebook para proceder a almacenar tweets.

**Gráfico N. 12** Estructura Json de la base de datos no relacional.



**Elaboración:** Jiménez Cárdenas Edinson Andrés.

**Fuente:** (Google Developers, 2019).