UNIVERSIDAD TÉCNICA DEL NORTE

Facultad de Ingeniería en Ciencias Aplicadas

Carrera de Ingeniería en Sistemas Computacionales

DESARROLLO DE UNA ARQUITECTURA CONCEPTUAL PARA EL ANÁLISIS DE CONTENIDOS EN REDES SOCIALES SOBRE EL TEMA DEL ABORTO USANDO PYTHON

Trabajo de Grado previo a la obtención del título de Ingeniero en Sistemas Computacionales

Autor:

Paolo Roberto Roldán Robles

Director:

Ing. Iván Danilo García Santillán, PhD

Ibarra – Ecuador

2019

UNIVERSIDAD TÉCNICA DEL NORTE BIBLIOTECA UNIVERSITARIA

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

1.- IDENTIFICACIÓN DE LA OBRA:

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a su disposición la siguiente información.

DATOS DEL CONTACTO			
CÉDULA DE IDENTIDAD:	1003017082		
APELLIDOS Y NOMBRES:	Roldán Robles Paolo Roberto		
DIRECCIÓN:	Claudio Manet 2-62 y Salvador Dalí		
EMAIL:	elpao_ro3@yahoo.com		
TELÉFONO FIJO:	062514750	**TELÉFONO MÓVIL:**	0993996406
DATOS DE LA OBRA			
TÍTULO:	"DESARROLLO DE UNA ARQUITECTURA CONCEPTUAL PARA EL ANÁLISIS DE CONTENIDOS EN REDES SOCIALES SOBRE EL TEMA DEL ABORTO USANDO PYTHON"		
AUTOR:	Roldán Robles Paolo Roberto		
FECHA:	27 de febrero del 2019		
PROGRAMA:	[X] Pregrado [ ] Postgrado		
TÍTULO POR EL QUE OPTA:	Ingeniería en Sistemas Computacionales		
ASESOR / DIRECTOR:	Ing. Iván García Santillán. PhD.		

2. CONSTANCIAS

El autor (es) manifiesta (n) que la obra objeto de la presente autorización es original y se desarrolló, sin violar derechos de terceros, por lo tanto, la obra es original y que es (son) el (los) titular (es) de los derechos patrimoniales, por lo que asume (n) la responsabilidad sobre el contenido de la misma y saldrá (n) en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra, a los 27 días del mes de febrero de 2019

EL AUTOR

Paolo Roberto Roldán Robles

UNIVERSIDAD TÉCNICA DEL NORTE

FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS

CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE GRADO A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

Yo, Paolo Roberto Roldán Robles, con cédula de identificación Nro. 100301708-2, manifiesto mi voluntad de ceder a la Universidad Técnica del Norte los derechos patrimoniales consagrados en la ley de propiedad intelectual del Ecuador, artículo 4, 5 y 6, en calidad de autor del trabajo denominado: **"DESARROLLO DE UNA ARQUITECTURA CONCEPTUAL PARA EL ANÁLISIS DE CONTENIDOS EN REDES SOCIALES SOBRE EL TEMA DEL ABORTO, USANDO PYTHON."** que ha sido desarrollado para obtener el título de Ingeniero en Sistemas Computacionales, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En mi condición de autor me reservo los derechos morales de la obra antes citada. En concordancia suscribo este documento en el momento que hago entrega del trabajo final en formato impreso y digital a la biblioteca de la Universidad Técnica del Norte.

Firma:

Nombre: Paolo Roberto Roldán Robles

Cédula: 1003017082

Ibarra, febrero 2019

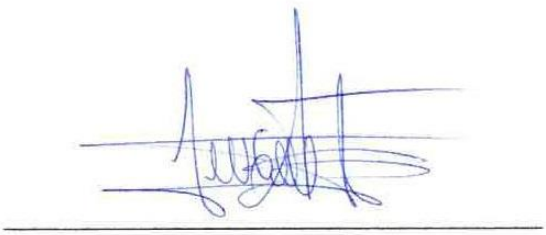
UNIVERSIDAD TÉCNICA DEL NORTE

FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS

CERTIFICACIÓN DIRECTOR DE TRABAJO DE GRADO

Por medio del presente yo Ing. Iván García, PhD. certifico que el Sr. Paolo Roberto Roldán Robles, portador de la cédula de identidad Nro. 100301708-2 ha trabajado en el desarrollo del proyecto de tesis **"DESARROLLO DE UNA ARQUITECTURA CONCEPTUAL PARA EL ANÁLISIS DE CONTENIDOS EN REDES SOCIALES SOBRE EL TEMA DEL ABORTO USANDO PYTHON.",** previo a la obtención del título de Ingeniero en Sistemas Computacionales, lo cual ha realizado en su totalidad con responsabilidad.

Es todo cuanto puedo decir en honor a la verdad.



Ing. Iván García S., PhD

Director de Trabajo de Grado

CERTIFICADO DE FUNCIONAMIENTO

Quito, 15 de febrero de 2019

A quien corresponda.

Por medio de la presente Certifico que el señor PAOLO ROBERTO ROLDAN ROBLES, con Nro. de cédula 100301708-2, ha realizado el "DESARROLLO DE UNA ARQUITECTURA CONCEPTUAL PARA EL ANÁLISIS DE CONTENIDOS EN REDES SOCIALES SOBRE EL TEMA DEL ABORTO, USANDO PYTHON", bajo la dirección de Iván García S. PhD.
La información fue proporcionada en:

- Gráficos estadísticos de porcentajes a favor y en contra del Aborto.
- Mapas de Calor de las localidades donde se destacan las tendencias.
- Nubes de Palabras tanto de frecuencia de hashtags y de usuarios influyentes.
- Línea de tiempo de tweets diarios generados a favor y en contra durante el tiempo de la recolección de la muestra.

La presentación de resultados en forma entendible nos ha permitido tener conocimiento relevante de las posiciones a favor y en contra del aborto en nuestro país Ecuador, lo que ha contribuido para orientar con nuevos argumentos a los adolescentes, jóvenes y familias con las que trabajamos en nuestro Centro y en los distintos lugares que realizamos talleres.

Sin más, me despido agradeciendo su atención y comprensión.

Atentamente,

Marlene Badillo.

G. Marlene Badillo H.
Psicóloga Clínica - Copropietaria Psicointegral.

Dirección: Av. García Moreno y pasaje Morona Santiago (Entrada Llano Grande, Conjunto Aldea Verde), Adm. Oficina 2
Teléfono: 2824-358

Dedicatoria

El presente trabajo de titulación está dedicado principalmente a Dios, por darme la vida y la fuerza para poder terminar la carrera universitaria.
A mi familia, por ser el puntal fundamental por su apoyo constante en todas las etapas de mi vida, y a todos los que han aportado de cualquier forma para que este trabajo sea culminado.

Paolo Roberto Roldán Robles.

Agradecimiento

A mi familia, por su motivación y ayuda total e incondicional y en especial a mi hermana Cristina, a las personas que con cosas sencillas animaron mi caminar en este proceso, a mi Director Ing. Iván García S. PhD, por su disposición al estar presto a apoyarme y guiarme con sus conocimientos, y como siempre a Dios (alejado de Ti nada puedo hacer).

Paolo Roberto Roldán Robles.

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

Resumen

El presente trabajo de titulación DESARROLLO DE UNA ARQUITECTURA CONCEPTUAL PARA EL ANÁLISIS DE CONTENIDOS EN REDES SOCIALES SOBRE EL TEMA DEL ABORTO USANDO PYTHON, pretende presentar el proceso completo de Minería de Datos y lograr un estudio detallado de las opiniones expresadas en redes sociales sobre el tema del aborto, para ello, en primer lugar se presenta un marco teórico sobre el proceso de extracción de conocimiento, para luego pasar a la fase experimental donde en la recolección de datos, se tomó una muestra desde el 16 de agosto hasta el 29 de septiembre de 2018 en Twitter, y las publicaciones de dos páginas de Facebook que fueron seleccionadas porque tenían más seguidores en el lapso de la toma de muestras. A partir de estos datos se llegó a determinar a través del análisis de los resultados, las posiciones **a favor** y **en contra** del aborto en nuestro país Ecuador.

En el desarrollo del presente trabajo, se ha tomado datos de las redes sociales Facebook y principalmente Twitter, además se usó la plataforma de Google maps, los scripts se los realizó en Python usando el Entorno de Desarrollo Integrado (IDE) Spyder (Python 3.6), que es parte de la plataforma Anaconda una de las más utilizadas para programar en este lenguaje.

Se presentan las interpretaciones de los resultados obtenidos, tomando en cuenta algunas directrices planteadas en trabajos anteriores realizados en varios lugares del mundo que abordan el análisis de contenidos y también análisis de sentimientos para extraer información sobre temas que van, desde la popularidad de una institución o marca comercial hasta las reacciones que genera la gente ante un hecho social o político.

Los resultados obtenidos en esta investigación marcan en promedio una posición mayoritaria en contra del aborto más que a favor.

Palabras Clave: Minería de Datos, Análisis de contenidos, Aborto, redes sociales, Twitter, Facebook.

Abstract

The present work to be graduated DEVELOP OF A CONCEPTUAL ARQUITECTURE FOR THE ANALYSIS OF CONTENTS IN SOCIAL NETWORKS ABOUT THE ABORTION USING PYTHON, this presents the complete process of Data Mining in order to achieve a detailed studio of different opinions in Social Networks about the abortion, therefore, in first place, it is gotten the theoretical framework. about the process of knowledge extraction. Then, it goes to the experiment stage where in the data collection since August 16th through September 29th of 2018 It is obtained a sample from Twitter and also from two pages of Facebook publications, those were selected because have more followers during this time than others. From this data it is determined through the analysis of the results for and against positions about the abortion in our country.

During the process of this work, it is obtained some data from Facebook but even more from Twitter, also, there is used the Google Maps Platform. The scripts were made using IDE Spyder (Python 3.6) which is part of Anaconda Platform, this is a more used one.

This work shows the interpretation of the obtained results taking in consideration some established guidelines in other works made around the world which ones talk about analysis of contents and also sentiment analysis to get information about topics that talk about the popularity of an institution or a commercial brand. Also shows the reactions of people at a social or political event.

The results obtained in this investigation set out an average of a predominant position against the abortion than in favors.

Keywords: Data Mining, Analysis of contents, Abortion, Social Networks, Twitter, Facebook.

INTRODUCCIÓN

Antecedentes

En tiempos antiguos, para aceder a la opinión de una persona se debia recurrir a una encuesta e ir a aplicarla en sectores posibles ya que se tenía un sinnúmero de limitaciones, siendo las más relevantes el tiempo a usarse y lo inaccesible de algunos lugares. Desde que se dió el uso masivo de las redes sociales, emitir opiniones de diversos tópicos es muy comun, la información recabada o emitida puede llegar a los extremos de ser por demás trivial o muy importante.

De manera general, la información emitida en estos medios es pública, esa carácteristica hace que se la pueda usar para hacer un sinnúmero de análisis de acuerdo a las conclusiones a las que se quiera llegar, al analizar la misma. A tal efecto, las mismas plataformas sociales ponen a disposición de los desarrolladores sus Aplication Programing Interface (API), que pueden ser usadas por aplicaciones desarrolladas en diferentes lenguajes de programación para conectarse directamente con los servicios disponibles y de esta manera, recolectar la información necesaria para la consecución de sus objetivos particulares.

Twitter, traducido al español como “gorjear” o “trinar”, es sin duda una de las redes sociales más populares actualmente, estableció una nueva forma de comunicación, en la que lo público y lo privado se fusionan. Se la podría definir como un servicio de microblogging por el tamaño de los contenidos que se pueden publicar, concretamente 140 caracteres como mensajes cortos que se denominan tweets y aparecen en la página principal de la cuenta del usuario, o lo que se conoce como timeline. Todos los mensajes publicados en Twitter son de acceso público de manera predeterminada, sin embargo, es posible gestionar la configuración de privacidad para protegerlos.

Proporciona un servicio absolutamente gratuito y sin publicidad que es muy sencillo de utilizar, En Ecuador su uso es aún algo limitado, ya que, a la mayor parte de la población se la puede familiarizar más cuando le hablas de su perfil de Facebook, ahí se establecería al tamaño de la publicación, como una primera diferencia entre estas dos redes, además, de que en Twitter los usuarios pueden seguir a otros usuarios, lo que no necesariamente los convierte en sus amigos, esta relación bidireccional se da cuando el usuario también es seguido por el usuario al que sigue, por tanto, éstos términos están claramente diferenciados. Se ha decidido tomar estas dos redes sociales bajo el análisis de que Facebook al ser la red más usada en Ecuador, marcará la posición clara de un sector de la sociedad más amplio y Twitter, porque, a pesar de no ser la red favorita de los ecuatorianos, proporcionará también información importante de los usuarios que por lo

general son líderes políticos o personas de influencia social o académica. El tema del aborto es relevante ya que una vez más, a polarizado a la sociedad entre los que están a favor y en contra.

Situación Actual.

Se han desarrollado muchos trabajos en base a datos proporcionados por redes sociales, como por ejemplo en procesos electorales, el enfoque primordial ha sido la predicción de resultados en dichos procesos, un ejemplo de ello es el trabajo de (Roldán, 2017); y de reacciones a procesos ya realizados, como el trabajo de (Niklander, 2017), donde analizan ámbitos políticos en Venezuela. Otra área que cubre la extracción de conocimientos en redes sociales, y más actual, es el analisis de las reacciones que pueden causar temas sociales que van del extremo positivo del día de los enamorados hasta el negativo como la guerra en Siria (Perikos & Hatzilygeroudis, 2018).

Hay mucho más que analizar en una publicaciónde Facebook o en un tweet y en el perfil del usuario que lo publica, elementos tales como hashtags, menciones, emoticones vídeos, url's, y en el caso de los usuarios, número de seguidores entre otros, los cuáles en el enfoque principal de este trabajo, darán una idea mucho más clara de que piensa el ecuatoriano en cuanto al tema del aborto. Actualmente, la organización PsicoIntegral no ha incursionado en el uso de tecnologías y arquitecturas conceptuales para sus actividades, pero, consideran muy importante el no estancarse e innovar lo que redundará en resultados positivos en sus expectativas futuras, de allí que es muy importante este análisis de contenidos, ya que el uso de la información recabada acercará a las propietarias mucho más a la realidad actual en el pensamiento general de los ecuatorianos sobre el tema del aborto, tomando opiniones de personas de distintas edades y estratos sociales.

Prospectiva

Mediante la aplicación de los scripts a desarrollar, se pretende establecer cuáles son las opiniones de los usuarios de redes sociales con respecto al tema del aborto en el Ecuador, cabe destacar, que la arquitectura conceptual a desarrollar puede emplearse para analizar las opiniones de éste o cualquier otro tema de interés; bastaría establecer los criterios de búsqueda de contenidos en las distintas redes sociales.

El desarrollo de este tipo de arquitecturas, acercará a la realidad actual en la que no basta con recoger opiniones de los ciudadanos comunes, conocido como encuestas a boca de calle, sino, que es muy importante tomar en cuenta las opiniones y todo lo que se dice en redes sociales, ya que, estas son cada vez más utilizadas por un gran número de usuarios para manifestar su criterio con mayor libertad. Por lo anteriormente expuesto, se espera que los

datos recolectados y analizados correspondan a usuarios que representen a todos los estratos sociales y niveles de educación.

Planteamiento del Problema

En Ecuador no se ha registrado aún, o no existe una Arquitectura Conceptual para el análisis de contenidos en redes sociales sobre el tema del aborto. En la Tabla 1, se presentan las causas y efectos basados en la pregunta de investigación ¿Cómo una arquitectura conceptual permite realizar un análisis de opiniones sobre el aborto en las redes sociales Twitter y Facebook?

TABLA 1

El problema, Causas y Efectos

Causas	Problema	Efectos
Inexistencia de análisis	**¿Cómo una arquitectura conceptual permite realizar un análisis de opiniones sobre el aborto en las redes sociales Twitter y Facebook?**	Falta de Información Estadística confiable
Opiniones a boca de calle		Sesgo Tecnológico
Mala manipulación de la información.		Perdida de información.
Pensar que todos los tweet o comentarios públicos en Facebook son irrelevantes.		No usar la información recogida en redes como insumos de importancia social.
Tomar en cuenta la opinión de solo cierto sector de la sociedad.		Sesgo social.
Conocimiento limitado de los beneficios con fines estadísticos de temas relevante de las redes sociales		Subutilización de las redes
Pensar que el tema del aborto no afecta psicológicamente.		Apatía social

Objetivos

Objetivo General

Desarrollar una arquitectura conceptual para el análisis de contenidos en redes sociales sobre el tema del aborto usando Python.

Objetivos Específicos

- Elaborar un marco teórico acerca del análisis de contenidos en redes sociales.
- Desarrollar cada una de las fases del proceso de construcción de aplicaciones de minería de medios sociales.
- Validar los resultados mediante presentación de estadísticas del análisis, nube de palabras y mapas de calor.

Alcance

Se pretende establecer cuál es la posición de la ciudadanía a nivel nacional con respecto al aborto, con base a las opiniones publicadas principalmente en Twitter y aquellas marcadas como públicas en Facebook.

Se representará dichas opiniones, con gráficos estadísticos, y las ubicaciones geográficas en las que se evidencia un mayor apoyo o rechazo hacia el aborto. para esto se plantea:

- Efectuar un proceso de minería en el contenido de cada uno de los tweets y/o publicaciones de la muestra obtenida, recolectar dicha muestra en base a criterios adecuados de filtro y selección.
- Determinar la relevancia y establecer los elementos a utilizar en cada una de las fases del análisis.
- Identificar a los usuarios más relevantes en base a la frecuencia de retweets de sus tweets, menciones en los tweets de otros usuarios
- Analizar la evolución en el tiempo de los tweets y/o publicaciones de apoyo o rechazo.
- Determinar las zonas geográficas en las cuales se concentran los porcentajes de apoyo o rechazo mediante el uso de mapas de calor.

La arquitectura conceptual consta de las siguientes fases enlistadas en la Tabla 2.

TABLA 2

Arquitectura Conceptual

ARQUITECTURA CONCEPTUAL DE ANÁLISIS DE CONTENIDOS DE REDES SOCIALES	Autenticación
	Recolección de datos
	Limpieza y procesamiento de los datos
	Modelado y análisis
	Presentación de resultados

Las herramientas para la gestión tecnológica que se manejará durante el desarrollo del proyecto se describen a continuación:

- Lenguaje de programación Python, ya que es uno de los lenguajes más utilizados en el desarrollo de aplicaciones de ciencia de datos; el cuál combina en gran forma su potencia con una sintaxis muy clara lo que es atractivo para los programadores principiantes o personas que han dejado de programar por algún tiempo.
- API´s de Twitter y Facebook por ser las redes sociales más populares.
- Google maps para generar las ubicaciones que serán visualizadas en los mapas de calor.

CAPITULO 1

Marco Teórico

1 CAPÍTULO 1 -- MARCO TEORICO

1.1 Conceptos Básicos

1.1.1 Minería de Datos

Si se sabe, que un dato es la representación mediante algún símbolo (número, letra, etc.), de un atributo de una determinada entidad, y que, minería es el arte muy antiguo de extraer los minerales que se han acumulado en el suelo, se puede sacar la siguiente definición de Minería de datos: Disciplina de la informática que estudia el análisis de grandes cantidades de datos con el objetivo de obtener conocimiento a partir de ellos (Lara, 2014) o como el proceso computacional de análisis de grandes cantidades de datos para extraer patrones útiles e información (Gullo, 2015).

La Minería de datos es la etapa más importante, de un proceso más general conocido como Proceso de Descubrimiento del Conocimiento, en inglés, Knowledge Discovery in Databases (KDD), que abarca desde la obtención y comprensión de los datos hasta la obtención de conocimiento a partir de ellos, algunos autores discrepan en el número de fases del KDD, pero en general, la mayoría lo engloban en cinco fases aunque difieren en los nombres de éstas o en su forma de agruparlas El conocimiento extraído del proceso KDD ha de poseer las siguientes características (Lara, 2014):

- No Trivial: que tenga importancia para algo,
- Implícito: que se encuentre oculto en los datos,
- Previamente desconocido: que no haya sido descubierto antes,
- Útil: debe servir para algo.

El proceso KDD consta de las siguientes fases:

a) Recopilación de Datos: En esta fase, los datos, precedentes de diferentes fuentes, se integran en un mismo repositorio de datos.

b) Selección y Limpieza de datos: Aún no se puede aplicar la minería con los datos recabados, debido a que los mismos, no pueden estar limpios o pueden contener atributos irrelevantes, precisamente en esta fase, la segunda del proceso KDD, se los limpia,

c) Transformación y reducción de los datos consiste en modificar la estructura de los datos con el objetivo de facilitar el análisis de estos, este paso resulta fundamental en el proceso global ya que requiere, un buen conocimiento del problema a resolver y una buena intuición, que marcan en gran parte el éxito o fracaso en la extracción de conocimiento.

d) Data Mining o Minería de datos: en esta fase se aplican técnicas concretas para obtener modelos.

e) Interpretación y evaluación de modelos: Los modelos obtenidos en la fase anterior deben ser evaluados, una vez comprobada la calidad de estos, son interpretados y a partir de ellos se obtiene el conocimiento.

La Figura 1, muestra el proceso desde la extracción de los datos iniciales hasta la generación de conocimiento.

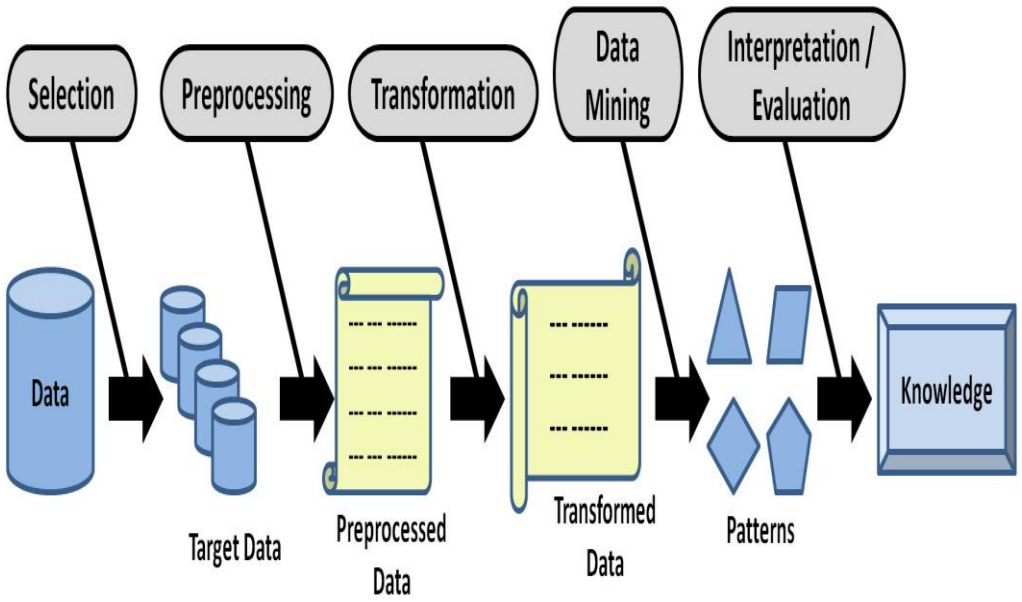


Fig. 1. Proceso KDD (Gullo, 2015)

Se debe tomar en cuenta que si bien es cierto la minería de datos NO es estadística, se apoya en ella.

El Data Mining establece tipos de datos según lo que representan y la forma en la que han de ser tratados en función de eso:

- Magnitudes → Cuantitativos
- Categorías → Cualitativos

Dentro de los datos Cuantitativos hay la siguiente clasificación:

• Discretos: Aquellos que pueden tomar un número limitado de valores diferentes. Por ejemplo, el número de estudiantes de una clase.

• Continuos: Aquellos para los que se cumple que, para cualquier par de valores, siempre se puede encontrar un valor intermedio. Por ejemplo, el peso o altura de una persona.

A los Cualitativos se subdivide en:

• Nominales: Aquellos para los cuales existe una asignación puramente arbitraria de números o símbolos para cada una de las categorías. Por ejemplo, el color de una prenda de ropa.

• Ordinales: Aquellos para los cuales existe una relación de orden entre las categorías. Por ejemplo, el número de cita de cada paciente en la consulta de un médico.

Existen también Otros datos no univaluados, tales como:

• Series temporales**,**

• **Documentos,**

• **Datos espaciales,**

•**Datos multimedia: sonidos, imágenes, vídeos, entre otros,**

• **Datos procedentes de la Web.**

Aplicando estos conceptos, y tomando en cuenta que las redes sociales permiten que personas influencien a otras personas, el poder medir esa influencia es unos de los objetivos principales de la minería de datos usando las API´s de las principales de ellas, tales como Facebook o Twitter. Dentro de la metodología empleada en el trabajo de (Roldán, 2017), se propone una Arquitectura para el proceso de construcción de aplicaciones de minería de medios sociales, las fases de dicha arquitectura se muestran en la Figura 2.

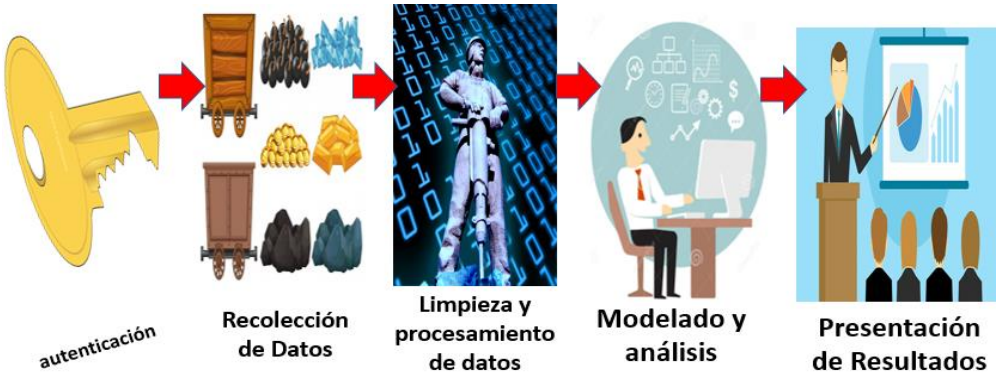


Fig. 2. Fases ilustradas de la Arquitectura Conceptual

La arquitectura conceptual tiene como objetivo el extraer el conocimiento de medios sociales, y es una variante contextualizada del proceso KDD, las diferencias que son más de forma y de adaptación hacia lo específico, y las similitudes que son las de fondo, se podrían detallar de la siguiente manera:

- En la arquitectura, se toma como la fase1 a la autenticación a las redes sociales elegidas,
- La fase 2 de la arquitectura conceptual es la recolección de datos, que es la fase 1 del KDD.
- Para el KDD la fase 2 es el preprocesamiento, y la fase 3 la transformación de datos, estas fases independientes en el KDD se unen en la fase 3 de la arquitectura que es llamada: limpieza y procesamiento de datos
- Tanto para la arquitectura, como para el KDD, la fase 4 busca aplicar técnicas para conseguir modelos, es llamada directamente Modelado y Análisis en la arquitectura, y data mining en el KDD
- Como fase final de la arquitectura está la presentación de los resultados, análoga a la fase 5 del proceso KDD que es la de Análisis y Evaluación, solamente diferenciada, en que el análisis se hizo en la fase 4 de la arquitectura.

1.1.1.1 Relación de la Minería de Datos con otras áreas

- Estadística. Muchas de las técnicas que se aplican en la minería de datos son o tienen su raíz en la estadística, de alguna manera, se podría decir que la estadística es la “madre” de la minería de datos, dado que muchos de los conceptos y técnicas de la estadística, se aplican en minería de datos.
- Bases de datos. El proceso de KDD parte de datos, que, habitualmente se encuentran almacenados en bases de datos. Como se ha comentado anteriormente, dichos datos son preparados para su posterior análisis;
- Visualización. El objetivo final de la minería de datos es obtener conocimiento que sea útil, para lograrlo, es un requisito fundamental, que ese conocimiento pueda ser visualizado por los expertos de cada dominio, de ahí la importancia de las técnicas de visualización (diagramas, gráficos, resúmenes, etc.) en el campo de la minería de datos;

- Aprendizaje Automático. Éste se encuentra profundamente ligado con la minería de datos, ya que ambos, de alguna manera, persiguen la obtención de modelos por medio de mecanismos automáticos;
- Otras. Además de las anteriores, la minería de datos también está relacionada con otras áreas como, por ejemplo: Los sistemas de apoyo a la decisión, La recuperación de información, El tratamiento y procesamiento de señales, entre otras.

La Figura 3, muestra una vista más amplia de cómo se relaciona la minería de datos con otras disciplinas.

Fig. 3. *Relación Minería de Datos y otras disciplinas,* (Sánchez, 2011)

1.1.2 Análisis de Contenidos.

Su primera definición hoy clásica, fue enunciada en 1952, se la dio en el marco de las investigaciones de comunicación de la posguerra por Bernard R. Berelson y se la enuncia como: técnica de investigación que pretende ser objetiva, sistemática y cuantitativa en el estudio del contenido manifiesto de la comunicación. (Comunicólogos, 2016).

El análisis de contenido alude al conjunto de procedimientos interpretativos de información, su principal objetivo es decodificar los mensajes plasmados en las diferentes fuentes.

El análisis de sentimiento o minería de opiniones es una variante del análisis de contenidos, que consiste, en el uso de tecnologías de procesamiento del lenguaje natural, analítica de textos y lingüística computacional para identificar y extraer información subjetiva de contenido de diversos tipos. (Borja, 2016).

1.1.2.1 Estado del Arte

Entre toda la literatura encontrada se nota que hay algunos enfoques para realizar la minería de datos, en este documento se realizó la siguiente clasificación de acuerdo con los criterios de recolección y limpieza de datos:

a) Análisis de textos:

Mohamed Yassine y Hazem Hajj, en su artículo (Yassine & Hajj, 2010), presentan un nuevo framework (que incluye el desarrollo de léxicos especiales), para caracterizar las interacciones emocionales en las redes sociales, y luego, usar estas características para distinguir a los amigos de los conocidos, el propósito, no es identificar emociones específicas sino más bien decir si el texto contiene emociones o no, en otras palabras, si el texto es subjetivo u objetivo, conociendo, que la decisión de transmitir emociones en el texto está influenciado a la fuerza de la relación entre el emisor y el receptor, para este propósito, se realizan técnicas de minería de texto y el análisis de los contenidos recopilados. Se divide al análisis de sentimientos en tres categorías: la primera tiene como objetivo extraer la valencia del texto, indicando si el texto tiene emociones positivas o negativas asociadas con él. La segunda identifica si el texto es subjetivo u objetivo (llamado también factual), la meta es encontrar si el texto es emocionalmente rico o no. La tercera reconoce no solo la emoción sino también su fuerza o excitación. El Marco propuesto para el análisis de texto consiste en 7 partes como se muestra en la Figura 4.

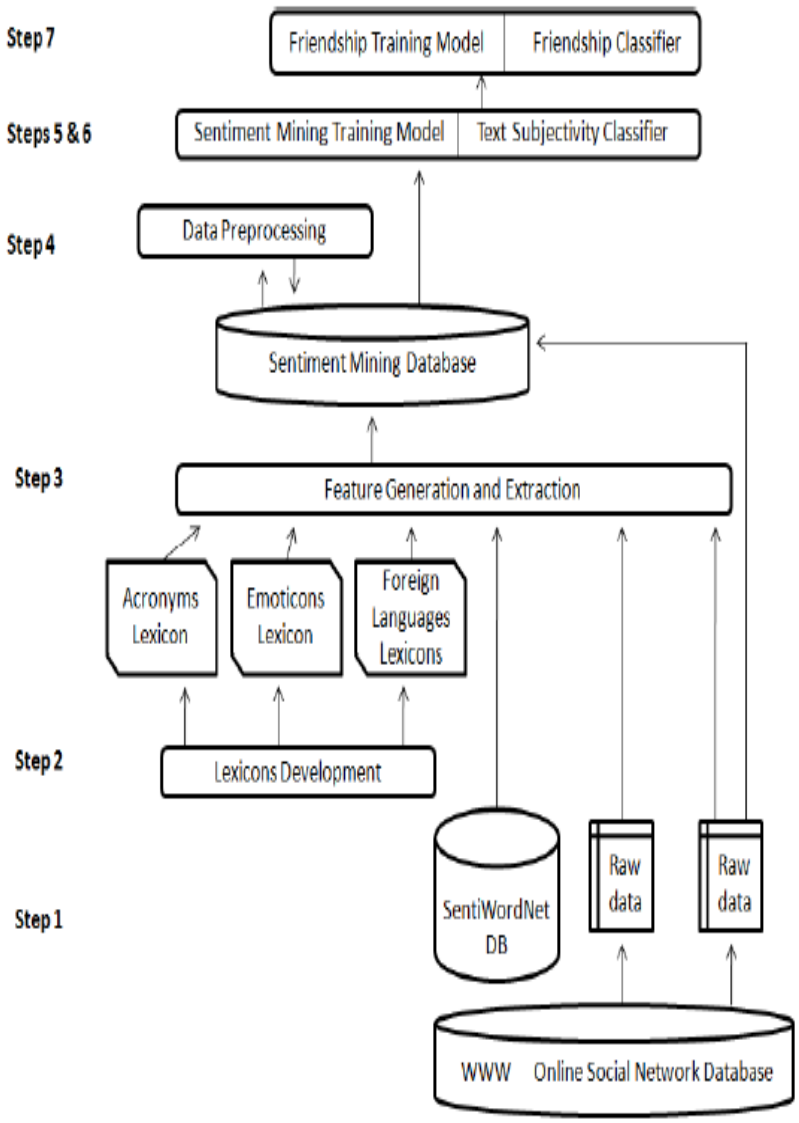


Fig. 4. Fases Framework Propuesto en, (Yassine & Hajj, 2010)

Como resultado, se logró crear los nuevos lexicones incluyendo emoticones, acrónimos sociales, expresiones árabes traducidas al inglés, etc. Usando algoritmos de máquinas de vectores de soporte o Support Vector Machines en inglés (SVM), para definir si un par de usuarios eran conocidos o amigos obteniendo una precisión del 87%. Proponen como trabajo futuro probar primero si la oración está considerando estructuras y claves de sintaxis para mejorar los resultados y por otro lado extraer el contexto del comentario y usar el conocimiento del mundo real para evaluar la emoción del comentario.

En su artículo (Valerio, Herrera-Murillo, Villanueva-Puente, Herrera-Murillo, & Rodríguez-Martínez, 2015) indican que hay varios compromisos de los estudiantes: a) con sus campos de estudio, b) con su universidad en lo afectivo en cuanto a sus actividades como comunidad, c) con la “marca” de su universidad en cuanto a sus servicios, y d) el compromiso digital que

es en el que enfatizó este estudio, específicamente en Facebook, de acuerdo a sus formatos de publicación como imágenes, texto plano, videos y enlaces haciéndose las siguientes preguntas:1. ¿Los formatos utilizados para las diferentes publicaciones en las páginas de los fanáticos de Facebook de las universidades estudiadas tienen un impacto digital? 2. ¿Qué tan estrechamente se relaciona la frecuencia con la que se usa cada formato de publicación con el compromiso digital? Se seleccionó a 28 universidades de México, obteniendo la muestra, la información relacionada con el tipo de publicación y la cantidad de Me gusta, comentarios y acciones fueron coleccionadas usando Facebook Query Language (FQL), un lenguaje de consulta que habilita la concentración de los datos públicos de los usuarios de Facebook. FQL acelera la recopilación de datos y ofrece la posibilidad de objetivamente clasificar datos según el contenido dentro de los parámetros definidos. Como resultados se ve, para la primera pregunta, que el formato de una publicación es factor estadísticamente relevante, la respuesta para la segunda pregunta en términos de frecuencia de uso, se encontró que los enlaces, eran el formato de publicación de uso común en algunas páginas de fans de la universidad, éstas representan el 50,4% del total, seguidos por imágenes (33.7%), texto plano (15.6%) y, finalmente, videos (0.30%). Aunque los enlaces fueron el formato más utilizado, también fueron el formato con el nivel más bajo de aceptación. Las imágenes tuvieron el mayor nivel de aceptación, los videos, que solo se usaron marginalmente, tuvieron una tasa de aceptación promedio en comparación con los enlaces. Concluyeron, que las universidades usan sus páginas de Facebook principalmente para fomentar el compromiso social y de marca y, solo en menor medida, para promover el compromiso académico, por lo tanto, se deben identificar escenarios de observación para este uso.

El artículo (Duwairi & AlFaqeeh, 2015), describe RUM, una herramienta de extracción de datos que permite a los investigadores guardar y analizar fácilmente varios tipos de contenido y estructuras disponibles en las páginas de Facebook respetando la privacidad, ya que solamente se recopilan publicaciones y comentarios públicos sin requerir habilidades de programación sustanciales por parte de los usuarios. RUM es fácil de configurar y usar, dando opciones flexibles a los usuarios para especificar el tipo y cantidad de contenido y estructura que desean recuperar; una característica de RUM es que no hay necesidad de conocer ninguna información previa sobre los nodos visitados desde el lado de los usuarios, tal información puede ser fácilmente extraíble del perfil de las páginas de Facebook. El Extractor de RUM proporciona datos "en bruto" para las páginas. Se ejecuta como una aplicación web y no requiere ser integrado dentro de software como NodeXL, que brinda a los comercializadores controlados por datos, acceso a potentes funciones de análisis de redes sociales que incluyen identificación de personas influyentes, evaluación de marca, escucha social, análisis de contenido, ideación, recolección de clientes potenciales, análisis social y de

campañas de competidores, automatización, etiquetado en blanco y mucho más, por lo tanto, es rápido y utilizable con los datos de Facebook. (NODEXL, 2018) El sistema está configurado para almacenar datos de usuario en servidores RUM durante una hora. Para demostrar cómo se puede usar RUM, se recopiló datos de dos sitios populares de noticias árabes (Aljazeera y Alarabyia). Los datos fueron recolectados de sus respectivas páginas en Facebook entre enero y junio de 2014, los campos que se compararon fueron: el tipo de publicación, la fecha, numero de likes de la página, comentarios totales, comentarios con y sin respuestas, likes en publicaciones, publicaciones compartidas y un consolidado (número total de me gusta, comentarios y comparte). Los datos recopilados muestran que Alarabyia realiza más publicaciones que Aljazeera, también que los lectores de la página de Alarabyia son más entusiastas en interactuar con la página escribiendo comentarios en las publicaciones y compartiendo las publicaciones entre sus redes.

Según el trabajo (Inbal Yahav, Shehory, & Schwartz, 2015), el análisis de los sentimientos y la minería de opiniones representan un gran espacio problemático, a menudo definido de manera ligeramente diferente, que cubre, por ejemplo: opinión extracción, estudio de subjetividad, análisis de emociones y más; el objetivo final del análisis de sentimiento es descubrir lo que la gente piensa o siente hacia un tema determinado; el principal aporte de este artículo está en revelar el discurso del comentario, su parcialidad y discutir sus implicaciones para el preprocesamiento de texto y evaluación del modelo. La segunda contribución es una corrección estadística propuesta al sesgo, entregando modificado el Term Frequency – Inverse Document Frequency (TF-IDF), ya que el tradicional, genera pesos sesgados cuando se lo usa en los textos cortos en el análisis de sentimientos; la estadística de la corrección se mantiene simple por razones prácticas, y las correcciones alternativas se discuten. Como parte del trabajo se examinó la correlación del discurso y su sesgo hacia tf-idf, en dos páginas de seguidores de Facebook con una gran actividad de usuario. La primera, es la página de fans del programa de televisión CommunityTV (www.facebook.com/communitytv), que tiene más de 1.7 millones de fanáticos y sus publicaciones son en su mayoría promocionales; La segunda, es la página de noticias SPORT1 News (www.facebook.com/SPORT1News), que mantiene casi un millón de fanáticos y su enfoque son las noticias deportivas. Para cada página se recopiló un conjunto reciente de 1000 documentos (publicaciones) y sus comentarios, el promedio (mediana) de comentarios por publicación es de aproximadamente 172 en CommunityTV y 96 en SPORT1News, los datos fueron compilados a partir de siete páginas de fans de Facebook seleccionados de diferentes dominios, incluidas noticias, finanzas, política, deporte, compras y entretenimiento lo que permitió diversidad en tema y tamaño y facilitó una aplicabilidad más amplia de los resultados, el ajuste presentado en este documento es aplicable a cualquier

nivel de correlación (incluso en un escenario sin correlación), ya que el ajuste del coeficiente es proporcional al sesgo observado. A pesar de las limitaciones del enfoque, es evidente que el sesgo corrección del tipo que se propuso, puede mejorar significativamente la precisión de la clasificación de comentarios y el tiempo de procesamiento.

El artículo (Kaynar, Görmez, Arslan, & Demirkoparan, 2017) busca hacer una comparativa entre los distintos métodos de selección de atributos en el análisis de contenidos de las redes sociales utilizando algunas estadísticas tales como Chi-cuadrado, Ganancia de información, Radio de ganancia, Coeficiente de Gini, OneR, ReliefF, que son métricas utilizadas para evaluar el rendimiento de cada método, se tomó en cuenta: la tasa (%), precisión (%) y sensibilidad (%), se lo aplicó en los conjuntos de datos de los comentarios de una película, y en general, el coeficiente Gini dio los mejores resultados y ReliefF los peores.

b) Análisis de textos y perfiles de usuarios que postean.

El artículo (Yue, Chen, Li, Zuo, & Yin, 2018), se centra en presentar los métodos típicos para el análisis de sentimientos desde tres perspectivas diferentes orientadas a las tareas, granularidad y a la metodología, y a la vez propone nuevas perspectivas múltiples, desentrañando series de trabajos, herramientas de organización y conjuntos de datos de referencia, utilizados en diversos trabajos de investigación, así como, sus limitaciones, y se analizan las perspectivas esenciales que se avecinan para el análisis del sentimiento. Dentro de la perspectiva orientada a las tareas se detalla a) clasificación de polaridad, b) el nivel de valencia o excitación a una escala específica, c) análisis de información espacial, información temporal, entidades o persona que publica más allá de la polaridad, d) la identificación de subjetividad / objetividad y e) el análisis de sentimiento basado en características; en la perspectiva orientada a la granularidad; se presenta documentación de análisis de sentimiento a nivel de: a) documento, b) oración y c) palabra y en la perspectiva metodológica se analiza el aprendizaje supervisado, semi-supervisado y no supervisado; se da una breve descripción, sus limitaciones y se proporciona el sitio web de las herramientas y diccionarios de minería de datos: General Inquired (GI), SentiWordNet, OpinionFinder, National Taiwan University Sentiment Dictionary (NTUSD), Bing Liu's Opinion Lexicon, SentiStrength, WordNet-Affect, Affective Norms for English Words (ANEW), GPOMS, LIWC, Financial Sentiment Dictionary (FSD), Lexicoder Sentiment Dictionary (LSD), DICTION, TAS/C, LingPipe y Apache OpenNLP. Como resultados se comprobó que, hasta el momento, la mayoría de las investigaciones de análisis de sentimiento, se basan en el procesamiento del lenguaje natural y la lingüística computacional las cuáles se centran en el contenido de texto, mientras que las personas aprovechan cada vez más videos, imágenes y audios para expresar sus opiniones en las plataformas de redes sociales. Se logró además, presentar una visión general del análisis de sentimiento multimodal (MSA) encontrando oportunidades significativas para

futuras investigaciones en el campo multidisciplinario de la fusión multimodal. Finalmente, se estableció una terminología común a través de varias investigaciones, lo que permitió que las personas de diferentes conocimientos básicos pudieran comprenderla fácilmente, y sentó las bases para la investigación avanzada en análisis de sentimientos.

La Figura 5, muestra en resumen todas las perspectivas de investigación para el análisis de sentimientos y minería de opinión.

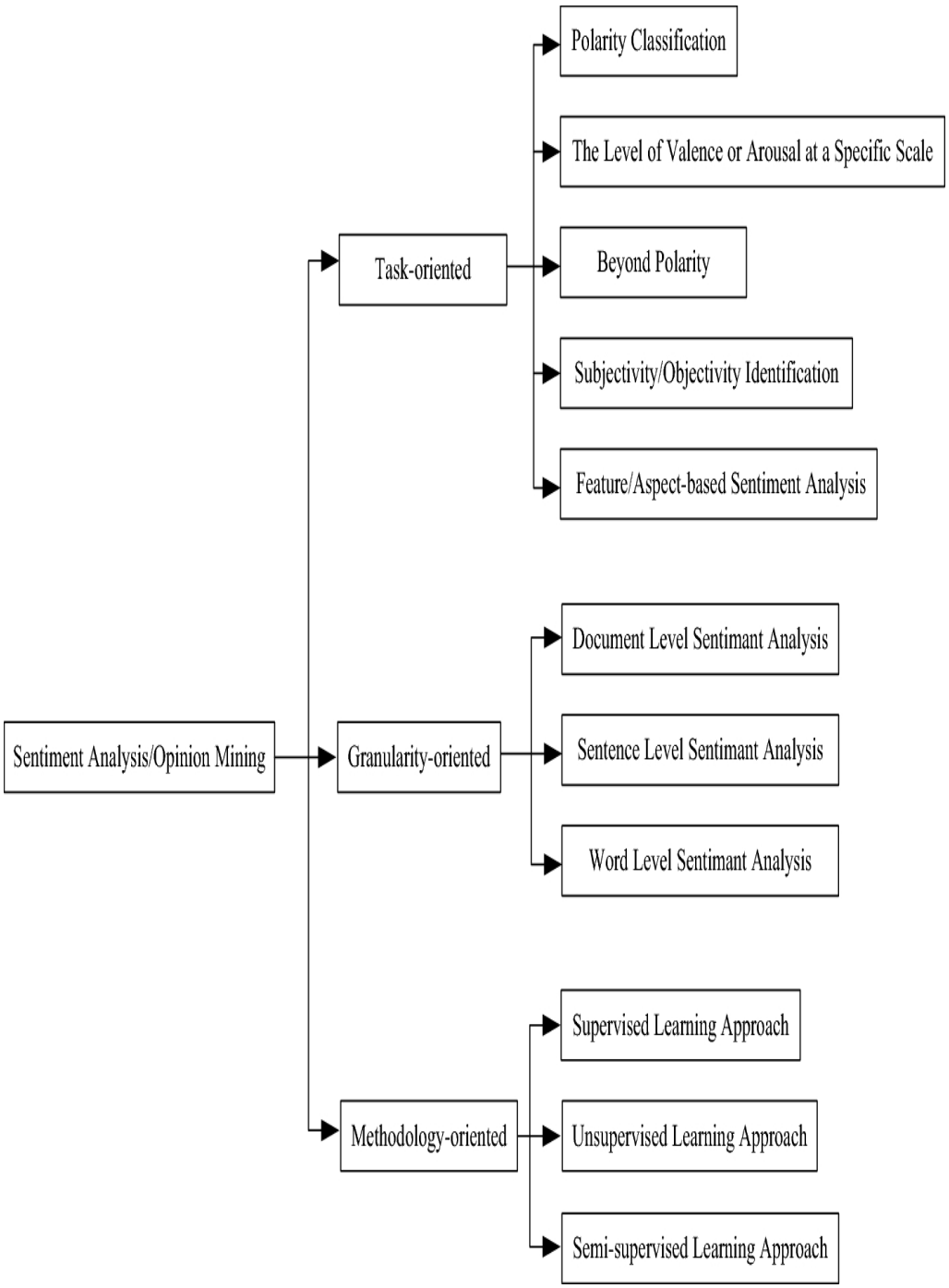


Fig. 5. El análisis del sentimiento visto desde múltiples perspectivas (Yue, Chen, Li, Zuo, & Yin, 2018)

c) Análisis de textos, Imágenes y contenido multimedia:

En su artículo, (Baecchi, Uricchio, Bertini, & Bimbo, 2015) investigan el uso de un enfoque de aprendizaje de características multimodales basados en el modelo: The Continuous Bag-of-Word CBOW y redes neuronales como Denoising Autoencoders (DA) , un tipo de red neuronal entrenada para codificar la entrada en alguna representación (generalmente de menor dimensión) para que la entrada se pueda reconstruir a partir de esa representación, para abordar el análisis de contenidos, como mensajes cortos de Twitter, que están compuestos por un breve texto y posiblemente una imagen, se propone él método CBOW-LR que es la extensión del CBOW con un muestreo negativo y que representa y clasifica el contenido o sentimiento al mismo tiempo , al añadir la posibilidad de analizar imágenes asociadas al tweet agregando a la red neuronal un DA de una sola capa, hasta llegar al método CBOW-DA-LR. Los resultados fueron superiores a otros métodos en 5 experimentos realizados, sobre todo en lo que compete a este trabajo de titulación, en la representación de la polaridad.

(Gullo, 2015) Aporta que, debido a la gran disponibilidad de datos de gráficos, la minería de gráficos se ha convertido en un destacado subcampo de minería de datos, y se apela a ella cada vez más continuamente. Las tareas de minería de gráficos prominentes incluyen: la agrupación gráfica, la búsqueda de gráficos, la extracción de subgráfico denso, la clasificación de gráficos, la minería de patrones gráficos, la coincidencia de gráficos, la consulta de gráficos, y la maximización de influencia.

(Purnomo, Sumpeno, Setiawan, & Diana Purwitasaria, 2017), exploran a profundidad algunos pasos para clasificar fraudes escritos como artículos de noticias en forma de alt-facts (difundir intencionalmente información falsa), especialmente, sobre cuestiones médicas en Indonesia; las publicaciones puede estar además de texto, como imagen o video de tal manera que la manipulación de contenido no textual se convierte en un problema porque generalmente los fraudes basados en texto se propagan a través de las redes sociales como Twitter o Facebook y el análisis para reconocerlas no se limitan al texto falso, sino también , al por quién son publicadas, en qué formato se presentan, y su contexto. El experimento fue mostrar cómo se implementó la clasificación de postura, los datos de los archivos médicos se extrajeron de una página web (snopes.com/category/facts/medical/), y se concluyó que el filtrado de la información para evitar engaños de salud en las redes sociales puede asociarse también con la investigación biomédica.

Los autores (Li, Fan, Jiang, Lei, & Liu, 2018), plantearon el análisis audiovisual y otros patrones de la inteligencia artificial, basado en dos enfoques: La representación de nivel medio, y la representación de estudio profundo el cuál se subdivide en: modo de extremo a

extremo, modo de tubería, entre otros, para hacer el análisis de sentimientos de una imagen posteada. Una de las conclusiones de este artículo es que se lograron resultados principalmente ocupando la técnica del estudio profundo.

d) Análisis de (#) Hashtags específicos

El trabajo (Niklander, 2017), indica que el análisis de contenido es una metodología cuantitativa para analizar el contenido de la información que se utiliza para descubrir los significados ocultos de los mensajes, y de quién los produce, es además, una técnica de investigación temática que permite formular inferencias basadas en los datos recopilados, y proporcionar nuevos conocimientos, se la aplicó para clasificar los mensajes comunicados a través de Twitter bajo el hashtag #Maduro, posteriores a las elecciones a la asamblea constituyente en Venezuela realizadas el 30 de julio de 2017, por la relevancia política para América Latina y el mundo entero. La Asamblea Nacional Constituyente de Venezuela, que se estableció al menos por un período de dos años comenzó a trabajar el 4 de agosto de 2017, por esta razón, tomaron ese día para el obtener la muestra, luego de analizar 4540 tweets se vio claramente dos grupos, uno minoritario a favor del presidente reelecto, y otro en contra, que usa principalmente la palabra “dictador” y la frase “fraude electoral” para referirse a la persona de Maduro y los comicios respectivamente.

En su trabajo (Perikos & Hatzilygeroudis, 2018), presentan un marco genérico para el análisis de grandes datos sociales, para el reconocimiento y la representación del estado de ánimo público, que consta de dos partes: a) un esquema clasificador de conjuntos que combina una herramienta basada en el conocimiento, independiente del dominio genérico, con métodos de aprendizaje automático y b) un gráfico emocional que visualiza las emociones o el estado de ánimo en el tema, basados en el análisis del contenido emocional individual de cada usuario. Se realizó el caso de estudio en Twitter analizando el contenido emocional de las publicaciones sobre las áreas temáticas específicas de la guerra civil siria (#syriancivilwar) y el día de San Valentín (#valentinesday) indexando y analizando un número total de 23.352 tweets, los resultados en el gráfico de emociones enmarcado en escala emocional de Ekman (Ekman, 1999) "ira", "Disgusto", "miedo", "felicidad", "tristeza", "sorpresa” se muestran en la Figura 6.

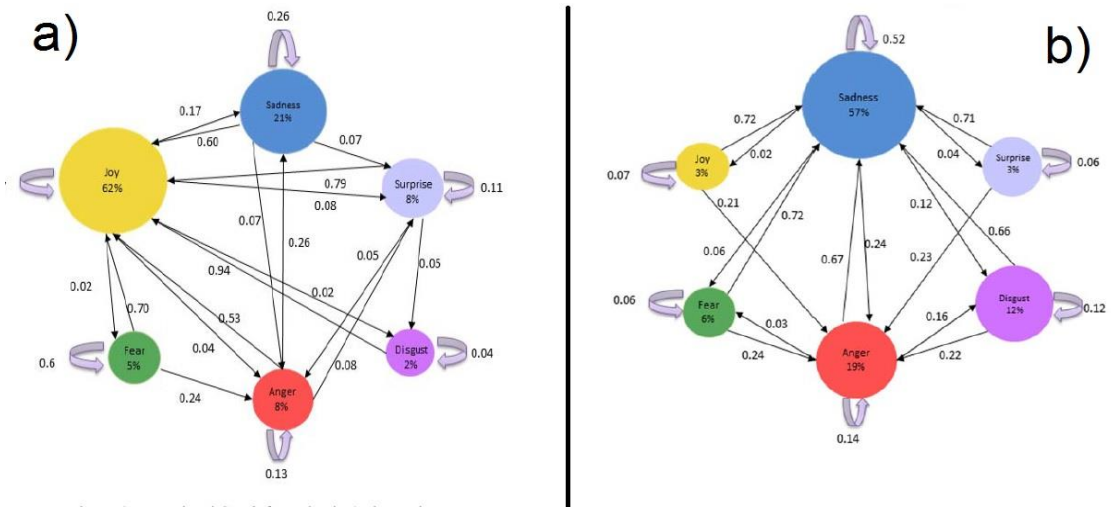


Fig. 6. *Gráfico emocional a) para el tema de la Guerra de Siria y b) para el tema del día de san Valentín,* (Perikos & Hatzilygeroudis, 2018)

e) Análisis de perfiles de usuarios o páginas de redes sociales.

El trabajo de (Yang, Tian, Li, Ma, & Zhang, 2017), está principalmente dirigido a: a)la identificación de líderes de opinión en redes sociales con mayor precisión para la información sensible a un tema específico, b) un análisis adicional para la difusión de la información real y c) el descubrimiento de información comúnmente es sensible a dicho tema; el experimento consistió en analizar un conjunto de datos recopilados de Sina microblog (la red social más importante de China), de noviembre de 2015 a enero de 2016, tomando: a) la información básica de los usuarios: ID de usuario, la cantidad de los microblogs que compartió, la cantidad total de seguidores. b) Información de los microblogs: el contenido, el número de reenvío el número de likes y, el número de comentarios. c) La relación entre los nodos de usuario: el ID de usuario que presta atención al nodo de otro usuario, se obtuvo 38,225 nodos de usuario y la relación entre los nodos de usuario que está compuesta por 57.351 bordes dirigidos, y la distribución de seguidores y se pudo encontrar las personas de mayor influencia enmarcadas en el principio Pareto (La regla 80/20) ampliamente usado en diversos campos, que dice, que la parte pequeña (20%) es vital y la mayoría (80%) es trivial en cualquier cosa, los autores concluyen que su método supera en rendimiento a otros, pero, es costoso debido a la alta complejidad computacional y la sobrecarga temporal de la centralidad de intermediación.

En el documento (Johnsen & Franke, 2017), presentaron un estudio para evaluar el rendimiento de las medidas de centralización basadas en gráficos, para identificar individuos importantes dentro de una red criminal. Estas medidas se han utilizado previamente en redes sociales generales pequeñas y estructuradas, pero ahora se las prueba en un nuevo conjunto de datos que es más grande, poco estructurado que asemeja

una red dentro de los foros de ciberdelincuencia. Se han propuesto métodos de Análisis de Redes Sociales (SNA) para la aplicación de la identificación de individuos centrales dentro de redes criminales. Los datos del análisis provienen de un foro en línea (accesible desde la red transparente) para distribuir software crackeado y negociar credenciales robadas, en un archivo de 9.45 GB, que se filtró el 12.05.2016 con detalles sobre 599 085 cuentas de usuario, (incluidas 800593 privadas) y 3495596 mensajes públicos; se generaron dos gráficos resultantes que fueron exportados luego en un Graph Exchange XML Forumat (GEXF), para facilitar los análisis posteriores. Se combinaron las tablas de los datos públicos y privados para construir el gráfico de comunicación pública. La extracción de datos y el análisis se realizaron en Ubuntu 15.10 con scripts de Python, además usaron los paquetes Networkx y MySQLdb; el resultado de este trabajo muestra que las medidas basadas en gráficos bien establecidas tienen debilidades cuando se aplican a conjuntos de datos grandes y no estructurados.

El objetivo del artículo (Mata-Gómez, Gilete-Tejero, Rico-Cotelo, Royano-Sánchez, & Ortega-Martínez, 2018), fue analizar la situación actual en España del uso de redes sociales en Neurocirugía, para lo cual se tomaron datos recopilados entre febrero y marzo de 2017, en Facebook (número de me gusta de la página), YouTube (número de suscriptores de los canales) y Twitter (número de seguidores y publicaciones), de los servicios y unidades de Neurocirugía, sociedades científicas, publicaciones y asociaciones relacionadas con Neurocirugía; dentro del conocimiento adquirido mediante la investigación, evidencian que muchos usuarios de redes no tienen una base sólida de conocimientos y su elección no depende del prestigio o calidad de la información publicada sino de la popularidad del perfil de redes sociales, esa popularidad puede conseguirse a través de realizar más publicaciones, independientemente de la veracidad de la información científica contenida, por esta razón en la práctica las páginas más populares no son necesariamente las más veraces. La Figura 7, muestra los modelos ideal y real para conseguir popularidad en redes sociales.

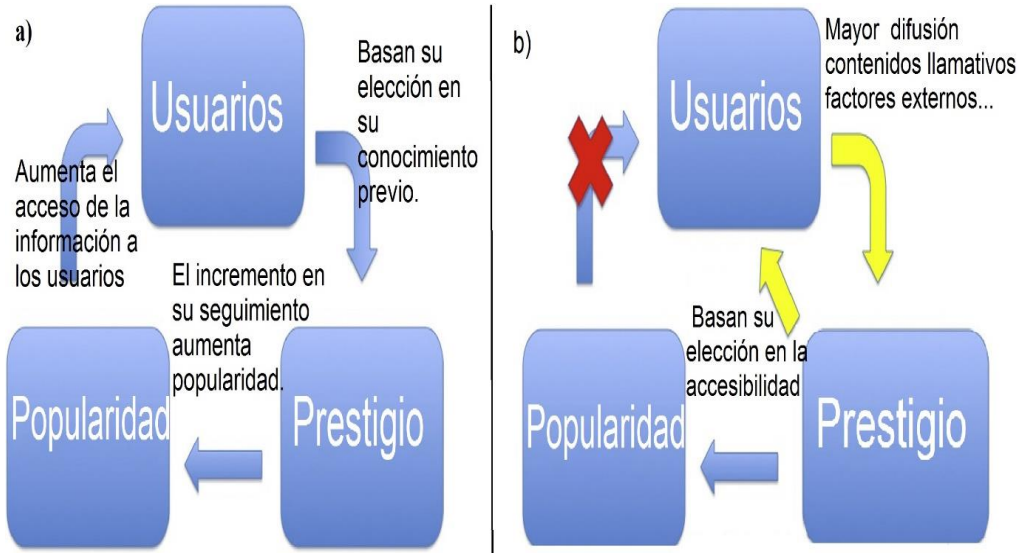


Fig. 7. Modelos de posicionamiento a) Modelo ideal en redes sociales. b) Modelo real en redes sociales. (Mata-Gómez, Gilete-Tejero, Rico-Cotelo, Royano-Sánchez, & Ortega-Martínez, 2018)

El resultado del estudio fue que solo 5 servicios de Neurocirugía de la red pública tienen presencia en redes sociales y son menos populares que las unidades privadas, el resto de los entes analizados tienen una presencia marginal.

1.1.3 Clasificadores

Arboles de Decisión: Proveen de una herramienta de clasificación muy potente, su uso en el manejo de datos les hace ganar en popularidad dadas las posibilidades que brinda y la facilidad con que son comprendidos sus resultados por cualquier usuario. El árbol en sí mismo, al ser obtenido, determina una regla de decisión. (Bouza & Santiago, 2014). Esta técnica permite:

- Segmentación: establecer que grupos son importantes para clasificar un cierto ítem.
- Clasificación: asignar ítems a uno de los grupos en que está particionada una población.
- Predicción: establecer reglas para hacer predicciones de ciertos eventos.
- Reducción de la dimensión de los datos: Identificar que datos son los importantes para hacer modelos de un fenómeno.
- Identificación-interrelación: identificar que variables y relaciones son importantes para ciertos grupos identificados a partir de analizar los datos.
- Recodificación: discretizar variables o establecer criterios cualitativos perdiendo la menor cantidad posible de información relevante

Naive Bayes: Es uno de los clasificadores más utilizados por su simplicidad y rapidez, se trata de una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados. Constituye una técnica supervisada porque necesita tener ejemplos clasificados para que funcione. Está basada en el Teorema de Bayes planteado en 1763, también conocido como teorema de la probabilidad condicionada. (CHiristianCH, 2013).

Además de los clasificadores detallados, existen otros como las Maquinas de Vectores de Soporte y varios tipos de árboles como el J48 y RandomTree que también son usados en diferentes investigaciones como las de (Yassine & Hajj, 2010).

1.1.4 Python

Python es un lenguaje de programación creado por Guido van Rossum a principios de los años 90, su nombre está inspirado en el grupo de cómicos ingleses “Monty Python”, es un lenguaje similar a Perl, pero, con una sintaxis muy limpia y que favorece un código legible.

Entre sus principales características están: Ser un lenguaje interpretado o de Script, Tipado Dinámico, Fuertemente Tipado, Multiplataforma (Windows, Linux, MAC) y orientado a objetos, es uno de los lenguajes más utilizados en el desarrollo de aplicaciones de ciencia de datos porque combina en gran forma su potencia, con una sintaxis muy clara lo que es atractivo para los programadores principiantes o personas que han dejado de programar por algún tiempo.

Sus librerías nos permiten una gran variedad de funcionalidades generales y específicas para: carga y visualización de datos, estadísticas, procesamiento de lenguaje natural y de imágenes, entre otras; y una gran comunidad de usuarios dedicada a fomentar el uso del lenguaje, ayudando a los desarrolladores con cualquier tipo de dudas.

Sus bibliotecas de manejo de tareas relacionadas con grandes volúmenes de datos son: numpy y pandas, éstas incluyen muchas de las capacidades del software R y MATLAB, pero son más intuitivas. Estas características hacen de Python un lenguaje superior o a la par de los lenguajes disponibles, y es por esto, que actualmente se lo elija con más frecuencia para el desarrollo de un gran número de aplicaciones para análisis de datos.

Las siguientes, son estructuras de datos de Python que se usan en este trabajo:

- Listas: son una de las estructuras más básicas de todas, pero también de las más usadas. Se las define como un grupo o conjunto de valores que tienen algo en común que los relaciona. Estos valores se colocan entre corchetes separados por comas, el siguiente es un ejemplo de una lista que contiene los hashtags utilizados por un usuario en sus tweets:

['abortolegal', 'abortoporviolacion'].

En el ejemplo tenemos una lista de Strings y que se basan en el tema propuesto para este trabajo de titulación, pero es posible crear listas de cualquier tipo de datos de Python.

- Diccionarios (Kazil & Jarmul, 2016): son una estructura más compleja que las listas, están conformados por parejas "clave: valor", donde las claves son únicas y pueden ser de cualquier tipo inmutable en Python, esto es que no se pueden modificar. Los valores pueden ser de cualquier tipo, incluso otros diccionarios. Para poder acceder a éstos utilizaremos sus respectivas claves. Los elementos de un diccionario en Python se colocan entre llaves y separados entre sí por comas, el siguiente es un ejemplo de un diccionario cuyas claves son el screen_name de un usuario Twitter, y los hashtags utilizados en sus tweets sus correspondientes valores:

  {dayumaEc:['AbortoPorViolacion','Ecuador'],vikypita:['noalaborto',
  'salvemoslasdosvidas', 'sialavida']}

  A continuación, de las librerías con las que cuenta Python en su "ecosistema científico" (McKinney, 2013), se detallan las que fueron usadas mayormente en el desarrollo de este trabajo:

- **NumPy**: abreviatura de Numerical Python es una librería para la computación científica. Proporciona funcionalidades para trabajar con grandes matrices N-dimensionales, por lo que es muy útil a la hora de trabajar con grandes volúmenes de datos en formato matriz. Incorpora todas las operaciones básicas para operar con cualquier dato numérico, así como también operaciones más complejas como la transformada de Fourier y de algebra lineal. Proporciona también herramientas para la integración de conexiones a C, C ++ y Fortran.

  Su funcionalidad con respecto al análisis de datos es el proporcionar un contenedor primario para pasar los datos entre algoritmos, haciendo más eficiente su almacenamiento y manipulación.

- **Pandas**: proporciona funciones que aceleran y facilitan el trabajo con estructuras de datos. Combina las funciones de alto rendimiento de NumPy con las capacidades de manipulación de datos de hojas de cálculo y bases de datos relacionales. Las estructuras que utilizaremos son los DataFrame, similares a las tablas de bases de datos.

- **MatPlotLib** (Matplotlib: Python Plotting, 2012): es una librería de Python usada para representar gráficas en 2D. Hace uso, entre otros, de Numpy para proporcionar un

correcto funcionamiento con matrices grandes. Es posible crear gráficas simples utilizando pocos comandos, en algunos casos incluso sólo uno.

- **Tweepy** (Roesslein, 2009): es una librería de código abierto que le permite a Python comunicarse con la plataforma de Twitter y utilizar su API. La comunicación entre Tweepy y Twitter se hace mediante el método de autenticación OAuth, al que es necesario pasarle 4 tokens que son proporcionadas por Twitter. La llamada a la API devuelve la información solicitada, en base a los criterios de selección indicados, simplificando de forma significativa la conexión y búsquedas en Twitter.

  La Figura 8, muestra la instalación de tweepy usando el comando pip install, desde Anaconda Prompt (el proceso es igual para el resto de las librerías).

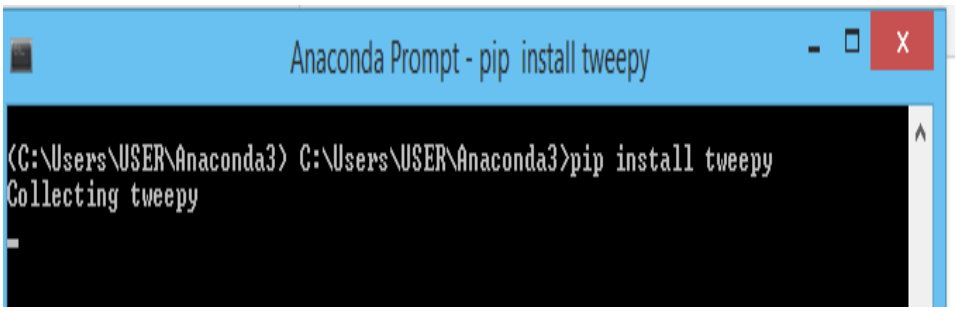


Fig. 8. *Captura de pantalla de la instalación de Tweepy*

- **Folium**: es una librería que permite generar mapas interactivos haciendo uso de los datos procesados en Python (Bonzanini, 2016).
- **Geopy**: es un cliente Python que permite localizar las coordenadas de direcciones, ciudades, países y puntos de referencia en todo el mundo mediante geo codificadores de terceros y otras fuentes de datos.

1.1.5 Twitter

Twitter es una red social que permite a sus usuarios publicar y leer contenidos en forma de mensajes cortos denominados tweets. Los conceptos básicos, características y aspectos técnicos de esta red se detallan a continuación:

- **Tweet**: es un mensaje corto de una longitud máxima de 140 caracteres que puede contener letras, números, signos, emoticones, enlaces, hashtags y menciones, que los usuarios publican en su línea de tiempo (timeline) y es compartido automáticamente con todos sus seguidores. Puede incluirse también contenido multimedia.

- **Retweet**: es un tweet publicado por un usuario en su línea de publicaciones pero que ha sido generado por otro usuario, éste aparecerá en el correspondiente timeline señalando siempre que se trata de una publicación de otro usuario.
- **Timeline**: es la página principal donde se muestran los tweets y retweets publicados por un usuario, las respuestas a mensajes directos entre otros. Es un resumen de la actividad del usuario, a la que únicamente tienen acceso sus seguidores y amigos.
- **Hashtag**: es una palabra o frase que comienza con el símbolo hash (#). Se usan para organizar, clasificar o agrupar las publicaciones de acuerdo su tema o contenido. Para obtener una lista de los mensajes de todos los usuarios que lo han utilizado, basta con dar un clic sobre el mismo.
- **Followers**: denominados seguidores en español son usuarios que se suscriben al contenido publicado por otro usuario; por lo tanto, cuando el usuario al que siguen publica un tweet, éste aparece automáticamente en sus correspondientes timelines y podrá ser comentado, retuiteado, o marcado como favorito por cada uno de los seguidores.
- **Friends**: a diferencia de otras redes sociales, la relación entre usuarios de Twitter no es simétrica, por lo tanto, las conexiones no necesariamente son mutuas: un usuario puede seguir a otro, pero serán “amigos” sólo si se siguen mutuamente.

1.1.6 Facebook

Creada por Mark Zuckerberg, Facebook, al igual que Twitter, es una red social explicita que tiene millones de usuarios alrededor del mundo y es la más popular en nuestro país Ecuador (Ekos, 2018).

Fue fundada en 2004, sin embargo, sus versiones en idiomas como portugués y español comenzaron a desarrollarse a partir de 2007.

(Perez-Porto & Gardey, 2013) refiere lo siguiente. El funcionamiento de Facebook es similar al de cualquier otra red social, aunque esta oración deberíamos formularla al revés, ya que es esta la red social que marca los antecedentes y las condiciones que deben cumplir las demás esto se da como queda dicho por su popularidad.

La Figura 9, muestra la pantalla de bienvenida de Facebook para desarrolladores, en la cual la red social permite crear aplicaciones a las personas que tengan una cuenta para poder obtener información en diferentes áreas estadísticas y de marketing entre otras.

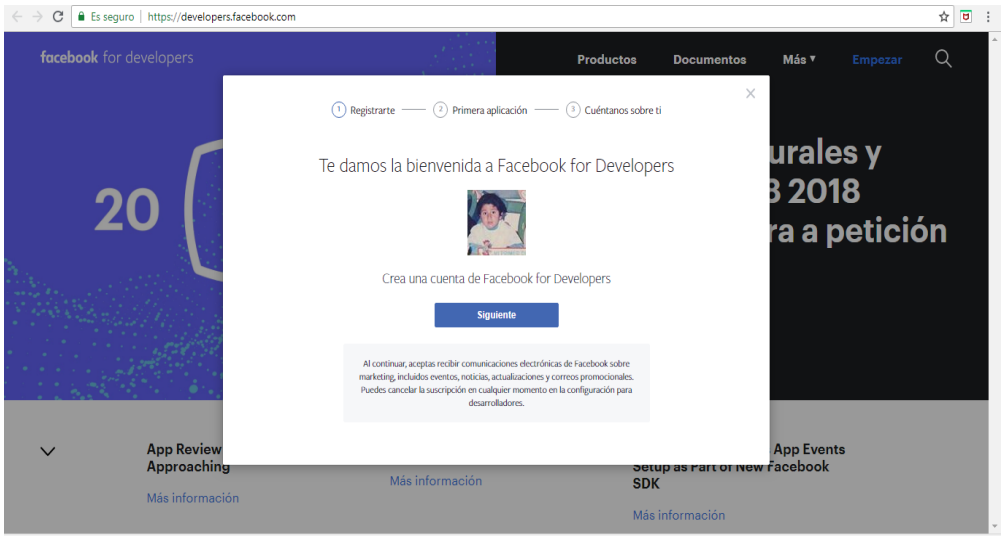


Fig. 9. Creando cuenta de desarrollador en Facebook

1.1.7 API´s

Una API, abreviatura de su nombre en inglés “Application Programming Interface”, es un conjunto de procedimientos, funciones y protocolos, que describen el comportamiento de un componente de software, que podrá ser usado por otras aplicaciones de forma segura y sin preocuparse por el funcionamiento interno de este componente, sino únicamente, por cómo usarlo (Bonzanini, 2016). Los desarrolladores pueden hacer uso de las funcionalidades de las APIs y así evitar realizar el proceso de desarrollo desde cero, sirven para establecer comunicaciones con bases de datos, sistemas operativos, servidores, etc. Las redes sociales también emplean APIs para que las aplicaciones se comuniquen con sus servidores y puedan acceder y gestionar los contenidos disponibles. Algunas de ellas son:

- Facebook GraphAPI: Es la principal forma de ingresar datos en la plataforma de Facebook y extraerlos de esta. Se trata de una API basada en HTTP, las aplicaciones pueden usarla de manera programática para consultar datos, publicar nuevas historias, administrar anuncios, subir fotos y llevar a cabo una amplia gama de otras tareas.
- API de Twitter: Ofrece acceso amplio a los datos de Twitter que los demás usuarios han decidido compartir con el mundo o, administrar su propia información que no es pública (como los Mensajes directos) para que los desarrolladores autorizados, puedan crear su propio software vinculado o aplicaciones,

1.1.8 JSON

(JavaScript Object Notation - Notación de Objetos de JavaScript), la definición aportada por (Nacarro-Arango, 2017) dice que. Es un formato ligero de intercambio de datos, basado en un subconjunto del Lenguaje de Programación JavaScript, leerlo y escribirlo es simple para humanos, mientras que para las máquinas es simple interpretarlo y generarlo. JSON es un

formato de texto que es completamente independiente del lenguaje, pero, hace uso de convenciones que son ampliamente conocidas por los programadores de la familia de lenguajes C, C++, C#, Java, JavaScript, Perl, Python, entre otros, esto hace que JSON sea un lenguaje ideal para el intercambio de datos, JSON 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.

Éstas son estructuras universales, y prácticamente 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. JSON es uno de los formatos más usados actualmente, esto se debe a que es fácil de leer y analizar, de hecho, los registros del archivo JSON adoptan la forma de un diccionario Python, lo que facilita el acceso y manipulación contenido. La Figura 10, muestra una representación del objeto tweet en un archivo JSON:

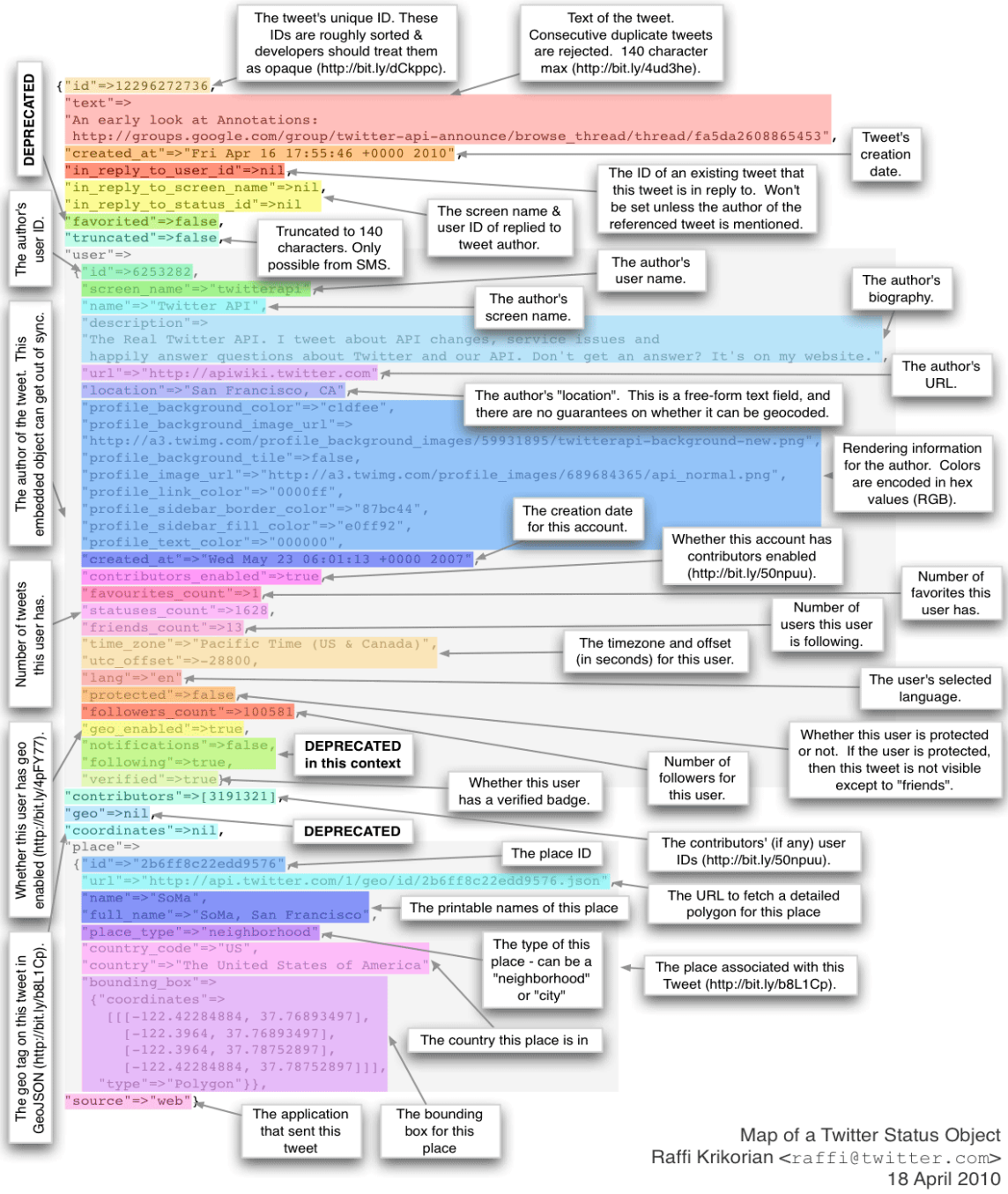


Fig. 10. *Objeto Tweet,* (Migurski, 2012)

Es necesario familiarizarse con los atributos de los objetos: **tweet**, y **user**, porque van a ser utilizados en los diferentes scripts que se implementarán para el desarrollo de la Arquitectura Conceptual, La Tabla 3, muestra la descripción de cada atributo del objeto **tweet**.

TABLA 3

Elementos importantes del objeto tweet, (Roldán, 2017)

Atributo	Descripción
created_at	Fecha de publicación del tweet
Entities	Es un diccionario de URLs, hashtags, y menciones contenidas en el tweet
favourite_count	Número de veces en que el tweet ha sido calificado como favorito.
Geo	Coordenadas de localización
Id	Identificador único del tweet
Lang	Código string que indica el lenguaje en que ha sido escrito el tweet
retweet_count	Número de veces que el tweet ha sido retwiteado
Retweet	El tweet publicado es un retweet
Source	Dispositivo utilizado para publicar el tweet
Text	Es el contenido del tweet
Truncated	Indica si el contenido del tweet ha sido truncado por exceder los 140 caracteres
User	Contiene la información del autor del tweet

A continuación, se muestra cada atributo del objeto **user** con su descripción correspondiente en la Tabla 4.

TABLA 4

Elementos importantes del objeto user, (Roldán, 2017)

Atributo	Descripción
created_at	Fecha de creación de la cuenta del usuario
followers_count	Número de seguidores del usuario
frieds_count	Número de amigos del usuario
geo_enable	En el caso de que este campo sea igual a True, será posible establecer la ubicación del usuario, cada vez que publica un *tweet*
id	Identificador único del usuario
location	Localización del usuario que está asociada con su perfil
name	Nombre propio del usuario
screen_name	Nombre de usuario

1.1.9 Google Maps

Google maps es un servidor de aplicaciones de mapas en la Web que pertenece a Alphabet Inc. (Page, 2018). Ofrece imágenes de mapas desplazables, así como fotos satelitales del mundo entero e incluso la ruta entre diferentes ubicaciones. Desde el 6 de octubre del 2005, Google Maps es parte de Google Local, es similar a Google Earth, una aplicación que ofrece vistas del Globo terráqueo impactantes, pero que no es fácil de integrar a páginas Web. Google Maps fue anunciado por primera vez el 8 de febrero del 2005 y estuvo en su fase beta 6 meses, para conseguir una buena sincronía entre las acciones del usuario y la respuesta de la aplicación Google utilizó Ajax. Fue ya en junio del 2005 Google cuando lanzó su API de Google Maps, haciendo oficialmente modificable casi cualquier aspecto de la interfaz original.

Las fuentes de los mapas disponibles son principalmente: satélites y aviones, aunque también se vale de mapas digitalizados de compañías como TeleAtlas y EarthSat.

Actualmente Google maps, crea experiencias simples y personalizadas para acercar el mundo real a usuarios a través de mapas estáticos y dinámicos, imágenes de Street View y vistas en 360°.

Entre sus APIs más importantes está API Geocoding que convierte las direcciones en coordenadas geográficas y viceversa.

Esta API será usada en este trabajo para realizar los mapas de calor aplicadas en el análisis y modelada y mostrados en la fase de presentación de resultados.

CAPITULO 2

Desarrollo de la Arquitectura Conceptual

2 CAPÍTULO 2 -- DESARROLLO DE LA ARQUITECTURA CONCEPTUAL

La arquitectura conceptual propuesta que consta de cinco fases:

- Autenticación
- Recolección de Datos.
- Limpieza y procesamiento de Datos.
- Modelado y Análisis.
- Presentación de Resultados

Se detalla a continuación, en cada una de las fases, se muestra lo que se hizo experimentalmente, con el objetivo final de indicar cuál es la posición de los ecuatorianos que usan redes sociales en cuanto al tema del Aborto.

2.1 Autenticación

La fase inicial del proceso marcó el camino para obtener los datos, en ella se indica, la forma en que se debe autenticar en Twitter, Facebook y Google maps. Los pasos para autenticarse en Facebook fueron:

1. Tener una cuenta en Facebook,
2. Ir a la página de desarrolladores (https://developers.facebook.com/),
3. Crear una aplicación,
4. Aceptar los términos y condiciones.
5. Esperar que la aplicación pase el proceso de revisión.

Cabe destacar especialmente en Facebook, que una de las complicaciones de este trabajo fue el cambio de las políticas de permisos y a pesar de que se creó aplicaciones como desarrolladores, no se pudo tener acceso para “minar” los datos de cuentas o páginas al no ser administradores).

Aunque el proceso de autenticación es similar en todas las redes, igual se detallan los pasos para la autenticación en Twitter que son:

- Tener una cuenta en Twitter,

- Ir a la página de desarrolladores (https://developer.twitter.com),
- Llenar el formulario que se nos presenta,
- Aceptar los términos legales.
- Al finalizar aparece una pantalla como se muestra en la Figura 11, que indica, que la aplicación entra a revisión (las condiciones para otorgar permisos tuvieron cambios, pero, si se pudo tener acceso a los datos de esta red social).

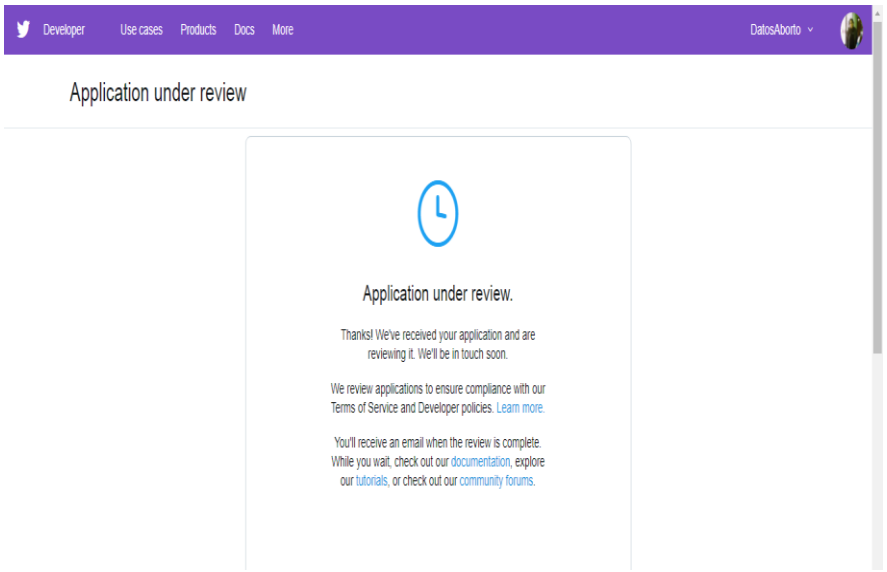


Fig. 11 Fin del proceso de autenticación, (la aplicación queda bajo revisión en Twitter)

El objetivo de este proceso es obtener las claves API key y API secret, y la opción de generación de los tokens: Access Token y Access Token Secrets, todos estos parámetros necesarios para la fase 2 de la Arquitectura Conceptual de Recolección de datos en los scripts de Python.

Para la autenticación en Google maps, se notará que difiere en pequeños detalles con las redes sociales anteriormente detalladas. los pasos que se deben seguir son los siguientes:

1 Tener una cuenta en Google,

2 Ir a la página de desarrolladores (https://cloud.google.com/maps-platform),

3 Habilitar Google Maps Platform y elegir el producto Maps,

4 Seleccionar un proyecto o crear uno nuevo,

5 Completar los datos de facturación (hay una versión de prueba gratuita) y aceptar los términos y condiciones.

De esta forma se obtiene los permisos necesarios para usar las aplicaciones con un número limitado de llamadas diarias. La Figura 12, muestra una captura de pantalla del paso 2 del proceso de autenticación en Google maps.

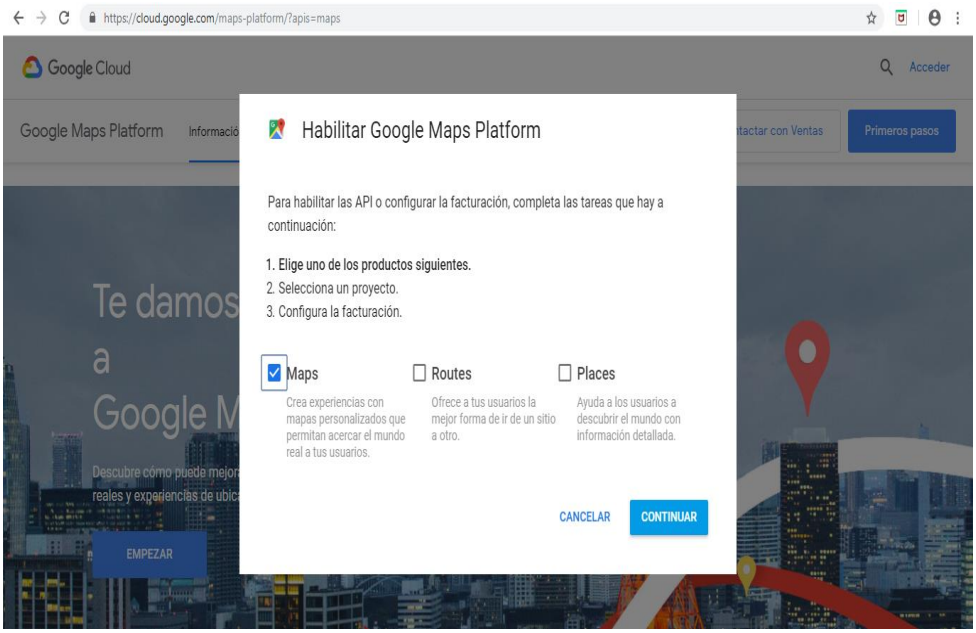


Fig. 12. Configuración y gestión de cuenta paso en el proceso de autenticación

2.2 Recolección de datos

La recolección de datos se realizó desde el 16 de agosto hasta el 29 de septiembre de 2018 y como resultado se obtuvo un archivo JSON de 1’721.287 KB de tamaño, con 344149 registros (tweets); para alcanzar este archivo, se usó un script que pretendió entre otras cosas, limitar un poco el tamaño total de la muestra bajo el criterio de tomar solo los tweets que se enviaban desde el Ecuador con hashtags y cuentas de usuario específicas especializadas en el tema del Aborto, tomadas de una búsqueda avanzada previa, desde la interfaz de Twitter como indica la Figura 13.

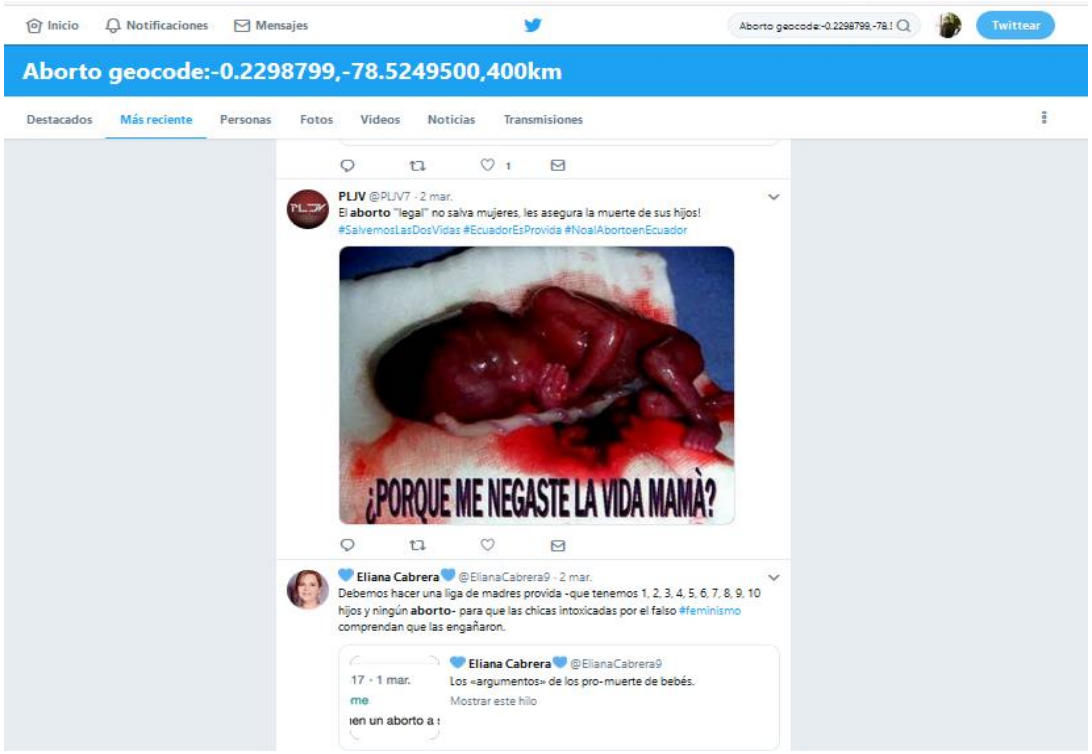


Fig. 13. Búsqueda avanzada de sobre un tema en Twitter

La Figura 14, muestra una captura de pantalla donde se puede notar además de lo expuesto, datos sobre el tamaño y número de registros y el último tweet recolectado con su respectiva fecha del día 29 de septiembre de 2018.

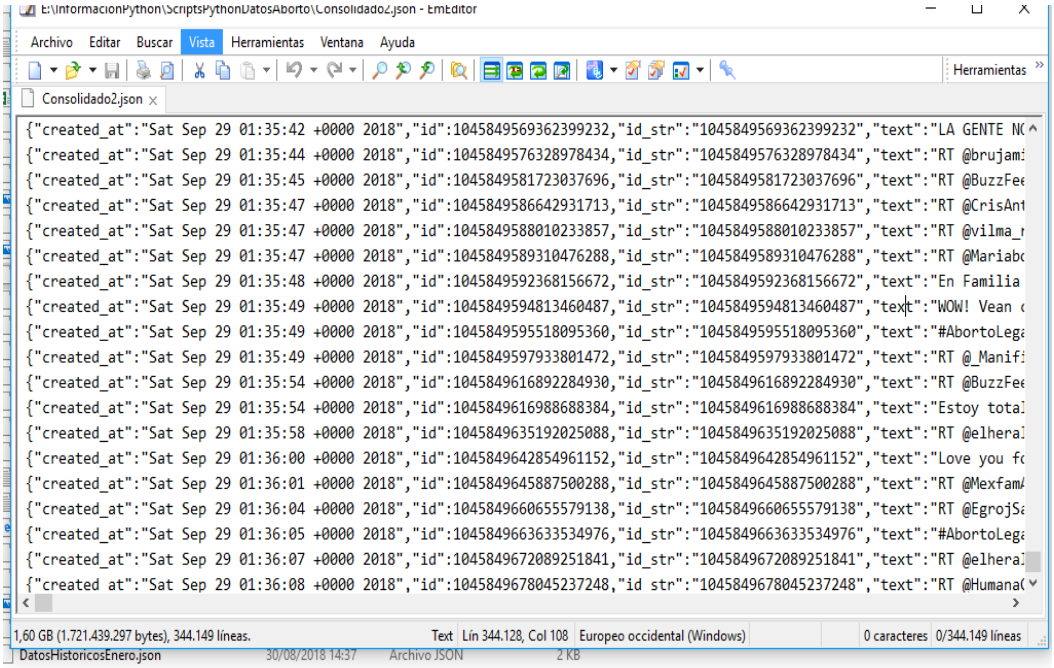


Fig. 14. Captura de pantalla del fragmento final de archivo recolectado

El script de recolección de la muestra estuvo en funcionamiento durante el período señalado en el párrafo anterior, se describe a continuación el funcionamiento del script que se utilizó para la recolección de tweets:

1. Autenticar la aplicación en la plataforma Twitter.
2. Colocar las claves de acceso.
3. Realizar la solicitud de descarga de tweets, incluyendo los criterios de filtrado de la muestra.
4. Generar o abrir el archivo de recolección.
5. Almacenar los datos en el archivo indicado.

La Figura 15, muestra la cantidad de tweets que se obtuvieron cada día mientras duró la fase de recolección de datos.

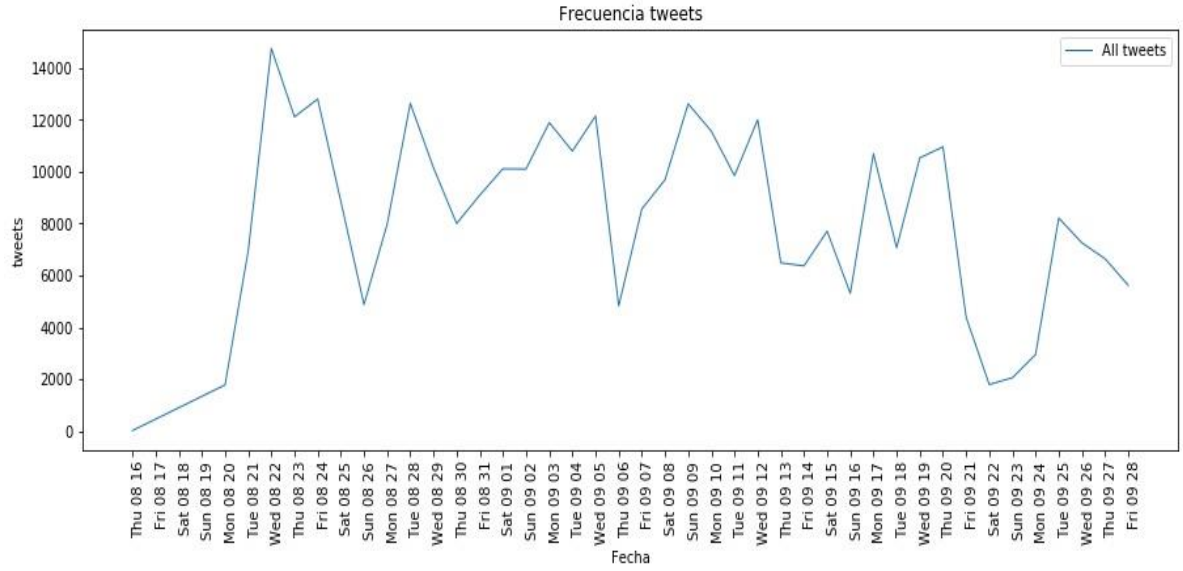


Fig. 15. Tweets recolectados por día durante el tiempo de muestreo

Dentro de la obtención de los datos “minados”, como se expuso anteriormente, se intentó filtrar que la muestra contenga los tweets limitados a Ecuador que es la intención del análisis de este trabajo, es decir, de alguna forma se aspiró tener una muestra limpia más que nada por el tamaño de los archivos que se obtendrían, para esto se usó dentro de la librería stream como filtro location obtenido de (BoundingBox, 2017) como indica (Sogo, 2016) dentro de su trabajo, la Figura 16, muestra el rectángulo formado para enmarcar el mapa de todo el Ecuador.

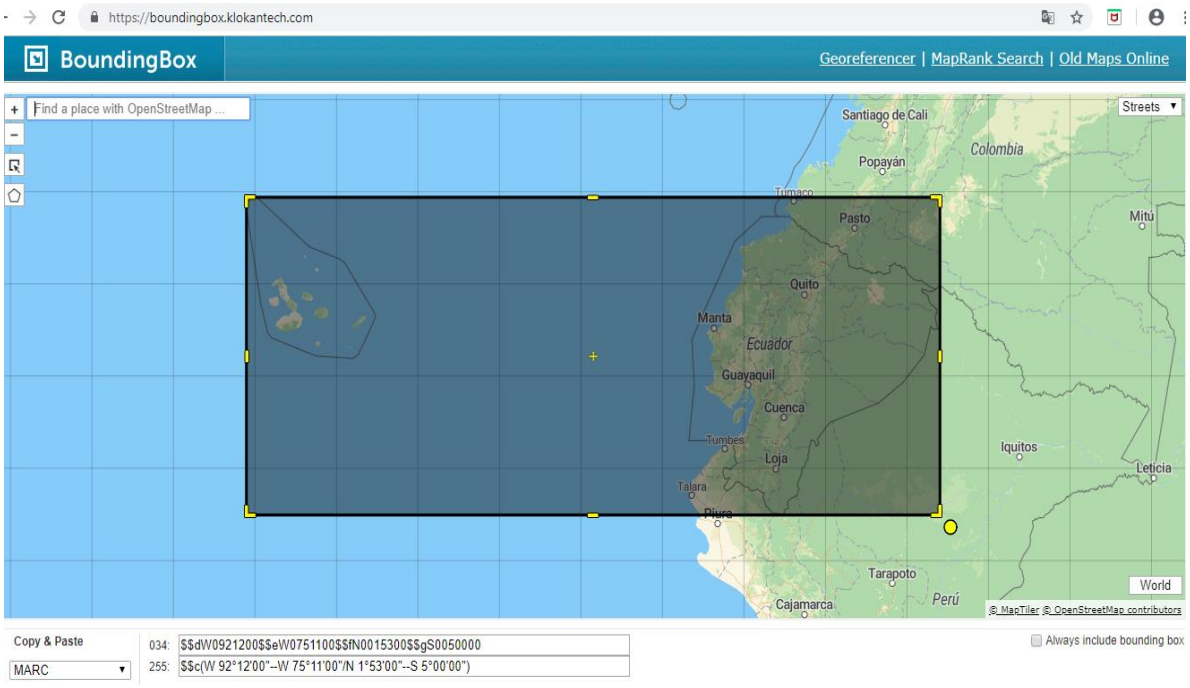


Fig. 16. Rectángulo del Mapa de Ecuador en BoundingBox

Se señaló que el presente trabajo, sería desarrollado principalmente en Twitter, es decir, que nuestra mayor “mina” sería esta red social, sin embargo, Facebook sería la red social que se lo complementaría, por esta razón, se debió analizar diferentes opciones para obtener datos, GraphAPI de Facebook permite guardar información de las páginas que tienen como administrador al dueño de la cuenta, por lo cual una opción era construir una nueva página asociada a una cuenta activa, que se especialice en el tema del Aborto y promocionarla para que la gente interactúe, pero fue descartada ya que por la premura del tiempo de muestreo no se iba a conseguir resultados confiables, la segunda opción consistía en pedir autorización a paginas ya especializadas en el tema y con algún tiempo de estar en Facebook para que se conceda el rol de administradores durante un tiempo idóneo para obtener la información pero no se encontró respuesta favorable, por lo que se optó por el uso de herramientas gratuitas. Para recolectar datos de Facebook hay diferente software como ParseHub que cuenta con su propia API y permite recopilar información pública de Facebook y guardarla en un archivo Excel o JSON o visualizarla dentro de su propia interfaz (parsehub, 2018), por lo cual, se instaló su aplicación de escritorio, para obtener los datos de cada una de las páginas se hizo lo siguiente:

- Ingresar a ParseHub,
- Dar clic en New Project,
- Ingresar la dirección de la página y dar clic en el botón **start Project on this url,**
- Elegir los elementos de interés,
- Damos clic en el botón **get data.**

La Figura 17, muestra una captura de pantalla del penúltimo paso en la página AbortoLibreEc, con algunos elementos elegidos para extraer.

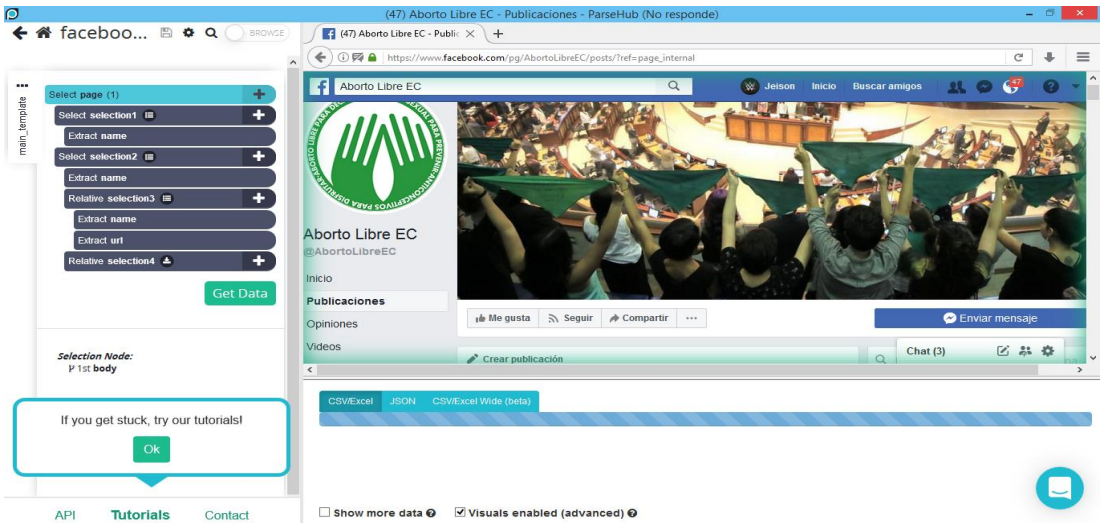


Fig. 17. Recolección de datos de “Aborto Legal Ec” desde ParseHub

2.3 Limpieza y procesamiento de datos

Previo a realizar la limpieza y el procesamiento de los datos, se determinaron los elementos del objeto **tweet** y **user** a usarse a lo largo del desarrollo del análisis de la muestra recolectada con base en los objetivos del presente trabajo, las cuales están detalladas en la Tabla 3 y la Tabla 4 del mismo y que son: entities, user, text, created_at.

2.3.1 Limpieza

En esta tercera fase de la arquitectura conceptual, se descartó de la muestra todos los tweets que no contuvieran información relevante para el análisis, para esto, se hizo correr el script hashtag_frecuency.py cuyas instrucciones se detallan en el siguiente algoritmo:

1. Importar el archivo JSON
2. Para cada línea del archivo hacer:

Extraer el elemento hashtag de la variable entities del objeto tweet

Si el elemento hashtag no está en el diccionario hashtag:

Guardar el elemento en el diccionario de hashtag e inicializar su frecuencia en 0.

Si no

Incrementar en uno la frecuencia.

Se eliminó del archivo obtenido, los hashtags que no tenían relación con el tema del aborto como: saludos, nombres propios y menciones a clubes deportivos o eventos sociales, a continuación, se muestra como se hizo la limpieza:

salvemoslas2vidas 12480

abortolegalya 9467

sialavida 5270

28s 4102

noalaborto 3306

seraley 3164

aborto 2638

dejalonacerrd 2412

~~**ecuador 2106**~~

provida 1958

~~**guayaquil 1349**~~

abortolegal 1154

olaceleste 1145

conmishijosnotemetas 1078

abortolibre 1074

~~**quito 1045**~~

En algunas fases del análisis de los tweets, se realizó acciones adicionales de limpieza, que se listan a continuación:

- Para la generación de los mapas de calor, se descartó los tweets que no contenían el dato de localización y aquellos que no estaban en Ecuador.
- Para la extracción de los usuarios más influyentes, se descartó los que no contenían menciones a otros usuarios, porque, se tomó en cuenta que si no se menciona a otra cuenta es porque el usuario no está interesado en ejercer influencia en otro.
- Para la extracción de hashtags, se descartó los tweets que no contenían hashtags.

2.3.2 Procesamiento de Datos

Terminando la limpieza, se debió procesar la información que se tuvo como resultado, bajo dos categorías: **a favor (Aborto+)** y en **contra (Aborto-)** del aborto, esta fase fue clave, ya que, desde aquí prácticamente se comenzó a buscar el objetivo planteado que es conocer los porcentajes de aceptación o rechazo al aborto en Ecuador. Se quiso hacer un filtro usando solo los hashtags que dentro de la muestra tengan al menos el 10 % de cantidad de menciones respecto al de mayor frecuencia, sin embargo, al final se ocupó absolutamente todos los hashtags (desde los más mencionados hasta los que se mencionaban una sola vez), que tenían que ver con el tema encontrándose un total de 487, de éstos 199 en la categoría a favor y 288 en contra que se los guardó en un archivo llamado Hashtags_ValoracionTotal.txt, a continuación, se detalla la etapa de procesamiento de datos en fragmentos de dicho archivo, este trabajo se hizo de forma manual para evitar perdida o mal uso de la información,

salvemoslas2vidas	A**borto-**
abortolegalya	A**borto+**
sialavida	A**borto-**
28s	A**borto+**
Noalaborto	A**borto-**
Seraley	A**borto+**
…..	
…..	
…..	
yoayudoanacer	**Aborto-**
yoescojo	**Aborto+**
abortosí	**Aborto+**
ecuadorgritaabortolibre	**Aborto+**
findeabortos	**Aborto-**
abortosincondiciones	**Aborto+**

Se realizó la investigación adecuada del origen de cada hashtag y como se lo utilizaba, por esta misma razón, se creyó que hacerlo de forma automática, implicaba tener demasiadas complicaciones por no existir reglas específicas para la creación de hashtags, ya que, algunos no contienen solamente palabras correctas, es decir. que estén dentro de los idiomas sino también: palabras inventadas, mezclas de palabras, palabras unidas con conectores distintos (ya que no pueden estar separadas), palabras con números (como las abreviaturas de fechas en alusión a eventos cercanos o recordatorios importantes de los colectivos a favor y en contra del aborto).

En el caso de Facebook la fase de limpieza y procesamiento de datos no se aplicó ya que al tomar la muestra se lo hizo de páginas claramente identificadas a favor o en contra de modo que se las consideran limpias y procesadas.

2.4 Modelado y Análisis.

Este modelo, consta de dos categorías, a favor y en contra, los registros pasaron por dos algoritmos de clasificadores: Arboles de decisión y Naive Bayes, éstos ya han sido usados en varios estudios similares incluido el de uno reciente de las causas de deserción de los estudiantes de la Universidad Técnica del Norte abordado en (Vila, y otros, 2019), para determinar el análisis de sentimiento ahora también del mensaje corto o texto llamado también tweet en sí.

2.4.1Análisis de Frecuencia de hashtags

Para obtener la frecuencia de hashtags dentro del archivo de datos recolectados, se usó el script frecuencia_hashtag.py, el mismo que dio como resultado el hashtag y el número de menciones, se obtuvo el top 5 de menciones, de los hashtags en el Ecuador sobre el tema del Aborto, dentro del script, se usó la librería **matplotlib.pyplot** para realizar la gráfica de este top que se mostrará más adelante, en la fase de presentación de resultados, siguiendo el orden de la Arquitectura Conceptual. La información se la consiguió simplemente tomando las 5 primeras filas del archivo hashtag_frecuency.txt, luego que sus datos fueron limpiados y procesados.

2.4.2 Análisis de Menciones a Usuarios

Las personas o instituciones más relevantes, o dicho de otra forma, las más mencionados dentro de la muestra obtenida, están relacionados directamente a sus cuentas de usuarios creados, las cuales son añadidas dentro del texto de los mensajes de otras personas, también para este análisis, se tomó en cuenta las personas que más mensajes aportaron a la muestra y las personas cuyos mensajes fueron “retuiteados”, es decir, fueron compartidos por otras cuentas en la línea de tiempo de la recolección de la muestra.

Los resultados de estas menciones aparecen en la siguiente fase de la arquitectura conceptual en la presentación de resultados en una Nube de palabras que muestran cuentas de usuarios relevantes.

2.4.3 Análisis de porcentajes de rechazo y apoyo

Específicamente en este apartado, se usó los algoritmos de los clasificadores, para realizar principalmente análisis de contenidos, pero también, se llegó, hasta el análisis de sentimientos de los textos.

Para la fase de entrenamiento de los dos algoritmos, se usó la misma información: un número de diez (10) hashtags a favor y la misma cantidad en contra, que representan, el 30% de los más relevantes de acuerdo a la frecuencia de aparición dentro del archivo hashtags_frecuency.txt, archivo que se generó al correr el script frecuencia_hashtag.py, además, para elegirlos se procuró que sean los más idóneos para entrenar al clasificador por ser acordes al tema, la siguiente decisión en esta parte del proceso, se hizo al explorar los textos del archivo de muestra ubicadas dentro del atributo **text** del objeto **tweet**, en esta parte del entrenamiento se debía poner los textos más comunes en favor y en contra del aborto, los textos podían ser frases cortas o palabras claves, al final de definió usar siete (7) textos a favor y siete en contra.

La Tabla 5, muestra el listado de hashtags a favor y en contra del aborto que se utilizó en los algoritmos y que fueron tomados del atributo **entities** del objeto **tweet**.

TABLA 5

Hashtags usados en la implementación de los algoritmos para los clasificadores

Hashtags en contra	Hashtags a favor
salvemoslas2vidas	Abortolegalya
sialavida	28s
noalaborto	seraley
provida	abortolegal
olaceleste	abortolibre
porlas2vidas	abortolibreec
salvemoslasdosvidas	hablemosdeaborto
accionprovida	yodecido
nadiemenos	8a
mentiraverde	Gritoglobal

La Tabla 6, muestra las frases a favor y en contra usadas en la fase de entrenamiento de los algoritmos.

TABLA 6

Frases cortas usadas en la implementación de los algoritmos para los clasificadores

Frases cortas en contra	**Frases cortas a favor**
no al aborto	en mi cuerpo decido yo
pro vida	decido abortar
cuidar la vida	aborto legal
las dos vidas	aborto sin riesgos
aborto es asesinato	aborto seguro
provida	aborto libre
abortoNo	abortoXviolacion

En el presente trabajo, se llegó dentro de las características de los arboles de decisión, a la reducción y clasificación de los datos, los hashtags y los textos fueron los argumentos de entrada, y las posiciones a favor y en contra los de salida, para esto, dentro de Python se usó la **librería** tree de **sklern** con el método **DecisionTreeClassifier()** instanciada de la siguiente forma:

clasificador = tree.DecisionTreeClassifier(),

Para la fase de entrenamiento, en la lista de entrada se debió ingresar dos parámetros que como se dijo, fueron los hashtags y textos pero dados valores numéricos así que, se asignó para el primer parámetro a los hashtags en contra, los valores del 1 al 10 y a los a favor los del 11 al 20, para el segundo parámetro, las frases en contra tomaron los valores del 1 al 7 y a las a favor los del 8 al 14, la matriz de aprendizaje usó todas las combinaciones de esta forma:

X = [

[1,1], [1,2], [1,3], [1,4], [1,5], [1,6], [1,7],

[2,1], [2,2], [2,3], [2,4], [2,5], [2,6], [2,7],

…

…

[1,8], [1,9], [1,10], [1,11], [1,12], [1,13], [1,14],

[2,8], [2,9], [2,10], [2,11], [2,12], [2,13], [2,14],

….

…..

[11,8], [11,9], [11,10], [11,11], [11,12], [11,13], [11,14],

[12,8], [12,9], [12,10], [12,11], [12,12], [12,13], [12,14],

…

….

[11,1], [11,2], [11,3], [11,4], [11,5], [11,6], [11,7],

[12,1], [12,2], [12,3], [11,4], [12,5], [12,6], [12,7],

…

…

[25,1], [25,2], [25,3], [25,4], [25,5], [25,6], [25,7],

[25,8], [25,9], [25,10], [25,11], [25,12], [25,13], [25,7],

[1,25], [2,25], [3,25], [4,25], [5,25], [6,25], [7,25],

[8,25], [9,25], [10,25],

[11,25], [12,25], [13,25], [14,25], [15,25], [16,25], [17,25],

[18,25], [19,25], [20,25],

[25,25]

]

Se dejaron los valores de 25 para ambos parámetros, por si llegaban a existir todavía hashtags o textos que no tenían que ver con el tema del Aborto, éstos fueron catalogados como neutros y se omitieron en los análisis.

El análisis de sentimientos, si se lo pudo lograr usando este clasificador, porque permitió manejar, ya no solo el hashtag sino el contenido del tweet y combinarlos, se usó las siguientes condiciones para obtener los resultados en la matriz de salida:

- Si el hashtag es en contra y la frase es en contra se contabilizó un tweet claramente marcado con tendencia en contra.
- Si el hashtag es en contra y la frase a favor se tomó como prioritaria la posición del texto y se añadió como tweet a favor del aborto.
- Si el hashtag es en contra y el texto neutro se lo tomó como un tweet en la posición en contra.
- Si el texto es en contra del aborto y no tiene hashtags o si los tiene estos son neutros se lo añadió al tweet como en contra.

- Si el hashtag es a favor y la frase es a favor se contabilizó como un tweet claramente marcado con tendencia a favor.
- Si el hashtag es a favor y la frase en contra se tomó como prioritaria la posición del texto y se añadió como tweet en contra del aborto.
- Si el hashtag es a favor y el texto es neutro se lo tomó como tweet a favor del aborto.
- Si el texto es a favor del aborto y no tiene hashtags o si los tiene estos son neutros se lo añadió al tweet como a favor.
- Al tweet con texto y hashtags neutros, se lo ignoró, es decir, éste no sumó a ninguna de las dos posiciones.

Se puede notar, que el texto tuvo mayor peso y la razón es que se pueden ocupar los hashtags definidos con alguna tendencia, pero para mostrar rechazo a la misma como, por ejemplo:

El aborto es un asesinato dejemos de apoyar el #abortolegalya.

En mi cuerpo decido yo así no lo entiendan los #provida.

Se indicó, que la matriz de salida se llenaría prácticamente con los valores de la ponderación asignada a favor o en contra, a la combinación entre el hashtag y el texto dados en la matriz de entrada (la opción AbortoNeutro solamente para la opción de entrada [25,25]), dicha matriz tomó la siguiente forma:

Y = ['Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-','Aborto-',

'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-','Aborto-',

…

…

'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+','Aborto+',

'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+','Aborto+',

…

…

'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+',

'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+',

…

…

'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-','Aborto-',

'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-','Aborto-',

…

…

'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+',

'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-', 'Aborto-','Aborto-',

'Aborto-', 'Aborto-', 'Aborto-',

'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+', 'Aborto+',

'Aborto+', 'Aborto+', 'Aborto+',

'AbortoNeutro']

En el caso del Naive Bayes, se tomó como referencia el script antispam.py presentado en el Libro (García Serrano, 2016, págs. 205-207).

Ya en la práctica, se debió instalar **TextBlob** y dentro de **textblob.classifiers** se importó **NaiveBayesClassifier,** Los datos para el entrenamiento ya no fueron numéricos como el árbol de decisiones, por eso, fue necesario darle las claves de aprendizaje al clasificador utilizando una matriz que recibió los datos, cada uno de ellos con dos parámetros, el primero fue el hashtag o frase y el segundo la polaridad, en este parámetro a la posición **a favor** se la nombró **pos** y a la posición **en contra neg**, quedando de la siguiente manera:

claves_aprendizaje = [

('salvemoslas2vidas', 'neg'),

('provida','neg'),

('noalaborto','neg'),

('sialavida', 'neg'),

("salvemoslasdosvidas", 'neg'),

('olaceleste', 'neg'),

('accionprovida ', 'neg'),

('porlas2vidas', 'neg'),

('nadiemenos', 'neg'),

('mentiraverde', 'neg'),

('abortolegalya', 'pos'),

('28s', 'pos'),

('hablemosdeaborto', 'pos'),

('abortolibre.', 'pos'),

('seraley', 'pos'),

('abortolegal.', 'pos')

('abortolibreec.', 'pos'),

('yodecido', 'pos'),

('8a.', 'pos'),

('gritoglobal', 'pos'),

('en mi cuerpo decido yo', 'pos'),

('decido abortar', 'pos'),

('aborto legal', 'pos'),

('aborto sin riesgos', 'pos'),

('aborto seguro', 'pos'),

('aborto libre', 'pos'),

('abortoXviolación', 'pos'),

('no al aborto', 'neg'),

('cuidar la vida', 'neg'),

('pro vida', 'neg'),

('las dos vidas', 'neg'),

('aborto es asesinato', 'neg'),

('provida', 'neg'),

('abortoNo', 'neg')]

Esta matriz fue el parámetro del clasificador para la fase de entrenamiento, quedando detallado así dentro del script:

clasificador = NaiveBayesClassifier(claves_aprendizaje)

Como se puede evidenciar, al implementar este algoritmo dentro de la clase de Python no se pudo ser tan especifico en las reglas como se hizo en el otro clasificador, en la siguiente

fase de la arquitectura se comprobará si esto influenció o no en sus valores de resultados para algunas métricas específicas de cada clasificador requeridas para llenar la Tabla 7, que se decidió evaluar en base al trabajo (Vila, y otros, 2019), el cuál usó Weka, cuya definición según (Córdoba-Fallas, 2011). Es una herramienta de tipo software para el aprendizaje automático y minería de datos, que contiene una colección de algoritmos para realizar análisis de datos y modelado predictivo, clasificaciones, entre otros, además tiene herramientas para la visualización de estos datos, y provee una interfaz gráfica que unifica las herramientas para que estén a una mejor disposición.

TABLA 7

Modelo para comparar métricas de evaluación específicas para los clasificadores (weighted average)

Clasificador	TP Rate	FP Rate	Precisión	Recall	F1 score	ROC Área
Clasificador 1						
Clasificador 2						

Con el propósito de conseguir la información requerida, se usó los metodos que ofrece **sklearn** en Python como se detalla en los fragmentos de código para:

- la matriz de confusión:

  ```
  from sklearn import metrics
  print(metrics.confusion_matrix(Y_true,Y_predict))
  ```

- la Precisión:

  ```
  from sklearn.metrics import accuracy_score
  accuracy_score(Y_true, Y_predict)
  print ('Acurracy      '+ str(accuracy_score(Y_true, Y_predict))).
  ```

- El F1_score (también conocido como F1 Meassure):

  ```
  from sklearn.metrics import f1_score
  F1Meassure = f1_score(Y_true, Y_predict, average='weighted')
  print (' valor F1Meassure      '+ str(F1Meassure))
  ```

- El Recall:

  ```
  from sklearn.metrics import recall_score
  Recall=recall_score(Y_true, Y_predict, average='weighted')
  ```

print (' valor Recall '+ str(Recall)).

- Roc_score:

  from sklearn.metrics import roc_auc_score

  Roc_Area = roc_auc_score(Y_true, Y_predict)

  print (' Roc_Area '+ str(Roc_Area))

La Figura 18, muestra una captura de pantalla mientras se implementaba el Script dentro de Spyder IDE de Python, para evaluar a los clasificadores utilizados.

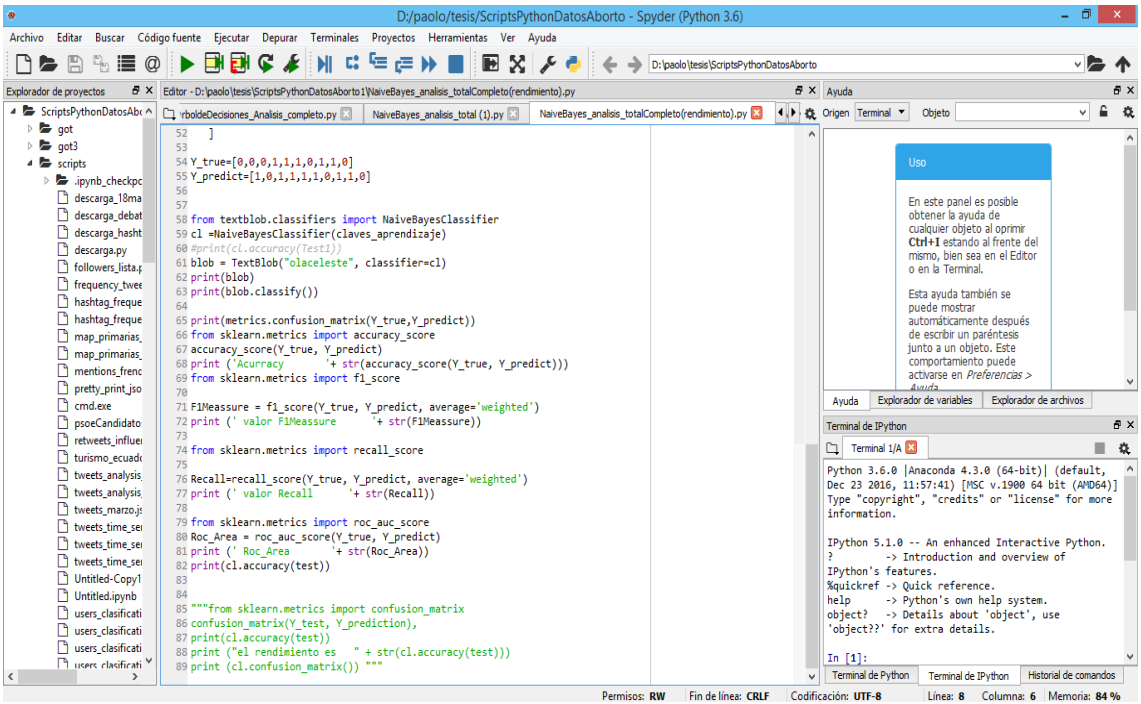


Fig. 18. Evaluación de clasificadores usando

Spyder IDE para Python 3,6 (Captura de Pantalla)

2.4.4 Análisis de Localización.

Para obtener las localizaciones de los datos obtenidos y que pasaron los filtros anteriores se tomó en cuenta:

- Los tweets que tenían activado al atributo geo_enable del objeto user, (los tweets que lo tenían deshabilitado tomaron el valor de Missing).
- Dentro de las localizaciones existieron valores de sitios irreales, los cuales, no pudieron ser ubicados dentro del mapa.

Considerando el tamaño de la muestra, se optó por el uso del servicio Nominatim de OpenStreetMap, ya que el mismo ofrece las mismas funcionalidades de las APIS de Google Maps, pero de forma gratuita y por tanto no fue necesario proporcionar datos de facturación.

En la clase get_user_location del script de análisis de la muestra, se realizó la llamada a Nominatim y se obtuvo las coordenadas correspondientes a las localizaciones donde se generaron los tweets, finalmente, pasaron el proceso de conversión a coordenadas para ser incluidas en la gráfica del mapa dentro del archivo HTML, donde se guardaron claramente diferenciados los sitios de donde salieron los tweets originales.

Para lograrlo, se ocupó los diccionarios de datos de Python para guardar tanto el nombre del usuario como su localización, para esto, se usó también la información de los atributos **screen_name, geo_enable** y **location** recolectada en el objeto **user,** el seudocódigo del script es el siguiente:

```
1  Importar el archivo JSON.
2  Para cada línea del archivo hacer:
       Extraer el screen_name del objeto user:
          Si el screen_name no está en diccionario users_location:
             Si el elemento location es diferente de None:
                Guardar screen_name y location en el diccionario
                users_location
             Fin Si
```

El diccionario se guardó de la siguiente forma:

users_location = {EmmaGabriela:'Manta,Ecuador', TrendsEcuador:'Ecuador', La_caritosh: 'Ecuador', barrabasec: 'Ibarra, Ecuador',} Domenikam: 'Guayaquil, Ecuador', yosoykiwi: 'Ecuador'}

Para el caso de Facebook, en esta fase se decidió, que el modelo sea simplemente comparativo en base al análisis de contenidos de las dos páginas elegidas, se tomó en cuenta el número de interacciones que generaron y que serán mostrados en la fase final del proceso de arquitectura.

2.5 Presentación de Resultados.

El objetivo principal de este trabajo de grado consistió en presentar los porcentajes a favor y en contra del aborto, la fase final de la arquitectura conceptual permitirá mostrar de formas distintas los resultados de los scripts usados desde la fase 3 donde se comenzaron a filtrar los datos y la fase 4, en donde estos datos filtrados pasaron por análisis determinados y se transformaron en conocimiento; este conocimiento ahora dice algo, es decir, es útil y trascendente cumpliendo enunciados de la minería de datos. A continuación, se detalla el seudocódigo del script usado para este propósito:

1. Importar el archivo JSON
2. Extraer el contenido del archivo
   Para cada línea del archivo hacer:
   Pasar por el clasificador
   Extraer la polaridad del tweet
3. Ubicar al tweet en el grupo correspondiente
4. Para cada grupo calcular el porcentaje de tweets

El análisis para la red social Facebook, se hizo de forma complementaria considerando la tabla comparativa de páginas de fans de acuerdo a las cantidades de “me gusta”, “compartir”, “comentarios” de los estudios de (Inbal Yahav, Shehory, & Schwartz, 2015) y (Duwairi & AlFaqeeh, 2015), añadiéndole las filas “seguidores” , “me entristece”, “me encanta”, “me enoja” y “me divierte”, como se muestra en la Tabla 8, se analizaron las publicaciones de las páginas: “Aborto Libre Ec” y “Pro Vida Ecuador”, elegidas porque en la búsqueda inicial en las fechas que se tomaron las muestras fueron los sitios con mayor cantidad de seguidores a favor y en contra del aborto respectivamente.

TABLA 8

Modelo pata comparativa de interacciones entre páginas de Facebook

ESTADISTICAS EN PUBLICACIONES			
LAPSO DE LA MUESTRA			
ESTADISTICAS	DESCRIPCION	PAGINA 1	PAGINA 2
Seguidores	Número de seguidores de las páginas		
TYPE_POST	se refiere a los tipos de publicaciones existentes en la pagina		
POST_PUBLISHED	Número de publicaciones en el intervalo de muestreo		
LIKES	Número de reacciones me gusta en publicaciones		
ME ENCANTA	Número de reacciones me encanta en publicaciones		
ME ENOJA	Número de reacciones me enoja en publicaciones		
ME DIVIERTE	Número de reacciones me divierte en publicaciones		
ME ENTRISTECE	Número de reacciones me entristece en publicaciones		
COMENTARIOS	Número total de comentarios incluidos respuestas		
COMENT_BASE	Número total de solo los comentarios iniciales		
REPLICAS	Respuestas a comentarios		
COMENTS LIKE	Reacciones a comentarios		
COMPARTIDOS	Número de veces que las publicaciones fueron compartidas		
CONSOLIDADO	Suma total de las estadísticas		

Los resultados obtenidos en la fase 5 de la arquitectura y su análisis se mostrarán en el Capítulo 3 de esta investigación.

CAPÍTULO 3

Validación de Resultados

3. CAPÍTULO 3 -- VALIDACIÓN DE RESULTADOS

3.1 Pruebas de Funcionamiento y Análisis e Interpretación de Resultados.

El resultado que se logró al completar todas las fases de la arquitectura conceptual debería generar información que pueda ser medida o comparada, pero, sobre todo, útil dentro de algún ámbito, bajo esa premisa, en esta investigación uno de los resultados, es el uso de los hashtags los cuales son presentados en la Figura 19, que deja visualizar el top 5 de los más utilizados dentro de la muestra obtenida en la fase de Recolección de datos.

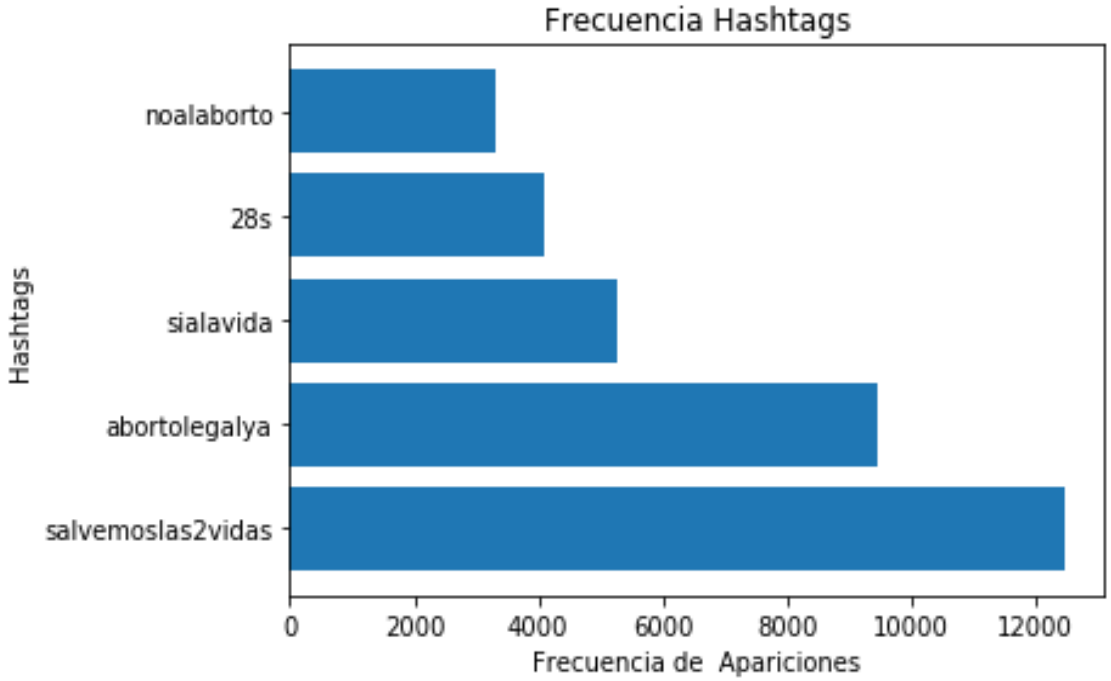


Fig. 19. Top 5 de apariciones de hashtags.

Dentro del cuadro de frecuencias encontramos que: Los puestos 1, 3 y 5 (#salvemoslas2vidas, #sialavida y #noalaborto), corresponden a hashtags identificados claramente **en contra** del aborto mientras que los puestos 2 y 4 (#abortolegalya y #28s), son hashtag **a favor** del aborto.

Se aseveró que el puesto número 4 es a favor tras indagar que hashtag #28s, fue creado en alusión al 28 de septiembre, un día emblemático para la causa que defiende el aborto, a

partir del V Encuentro Feminista Latinoamericano y del Caribe de 1990 realizado en Argentina, en este encuentro se propone enaltecer esta fecha, porque el 28 de septiembre de 1871 se promulgó en Brasil la ley de libertad de vientres donde se declaraba libres a las y los hijos que nacieron de esclavas.

La Figura 20, deja ver la frecuencia de hashtags en formato de nube de palabras, la cual marca con el tamaño superior al hashtag más usado y de allí en más, reducido el tamaño gradualmente al resto de hashtags hasta llegar al último.

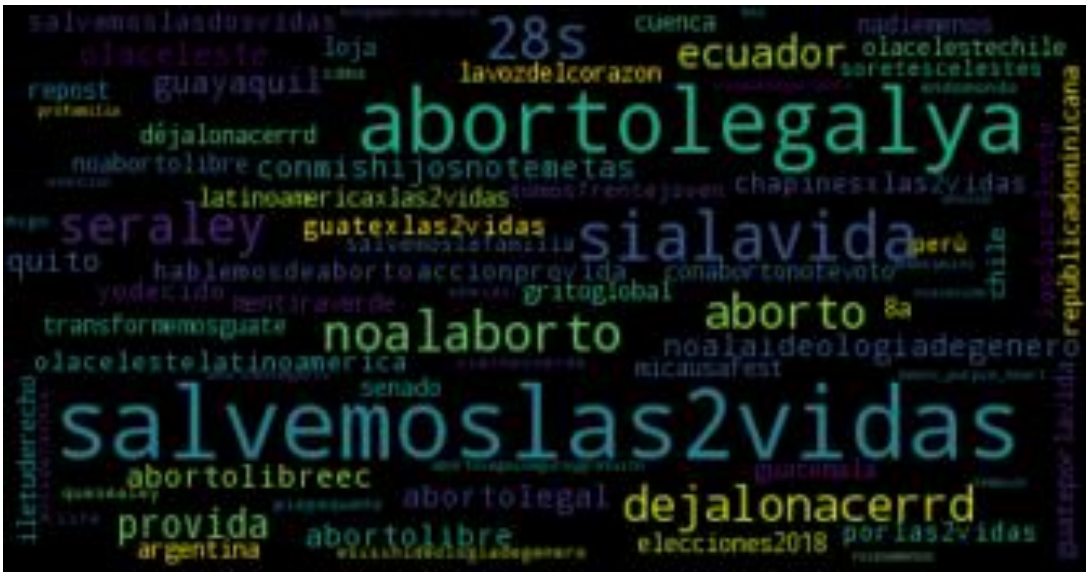


Fig. 20. Nube de Palabras de los hashtags más usados

Obviamente, como se había demostrado el hashtag #Salvemoslas2vidas es él más usado pero independientemente de este dato, que ya lo obtuvo en el top 5 expuesto anteriormente, esta nube de palabras permite ver aún un buen número de hashtags que son irrelevantes en el objetivo final de este estudio pero que aparecieron incluidos en la fase 2 de recolección de datos pero que más tarde fueron eliminados en la fase 3 de limpieza y procesamiento de datos evidenciando el proceso de Arquitectura Conceptual.

Se generó, además, gráficos estadísticos que fueron obtenidos usando dos algoritmos diferentes, generados desde Python al usar de **pyplot** la librería **matplotlib**. La Figura 21, muestra los resultados para las dos categorías del algoritmo de árboles de decisión en forma de pastel e indica el 59,3% de la posición en contra y el 40,7% de la posición a favor.

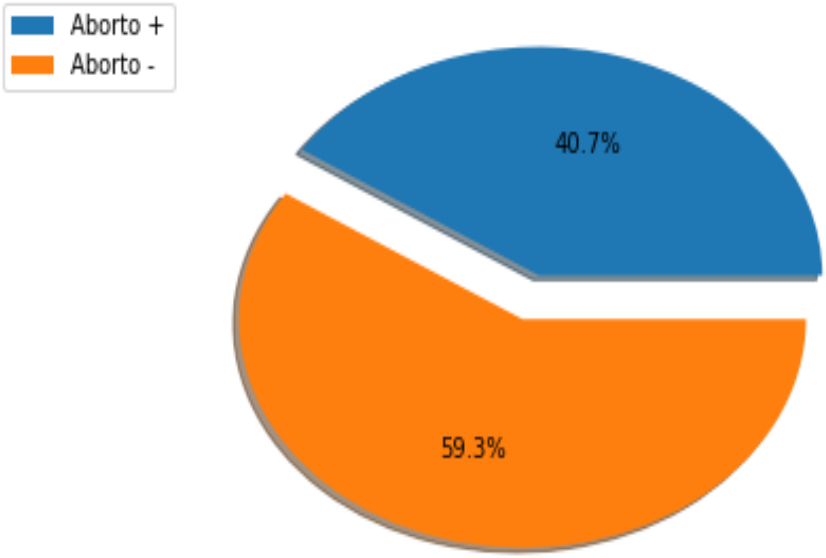


Fig. 21. Gráfica Pastel porcentajes de tendencias A favor y en contra del Aborto basado en el análisis de contenidos con el algoritmo árboles de decisión

Se obtiene los siguientes resultados presentados en la Figura 22, para el análisis ahora con el algoritmo de Naive Bayes e indica el 55,4% de la posición en contra y el 44,6% de la posición a favor.

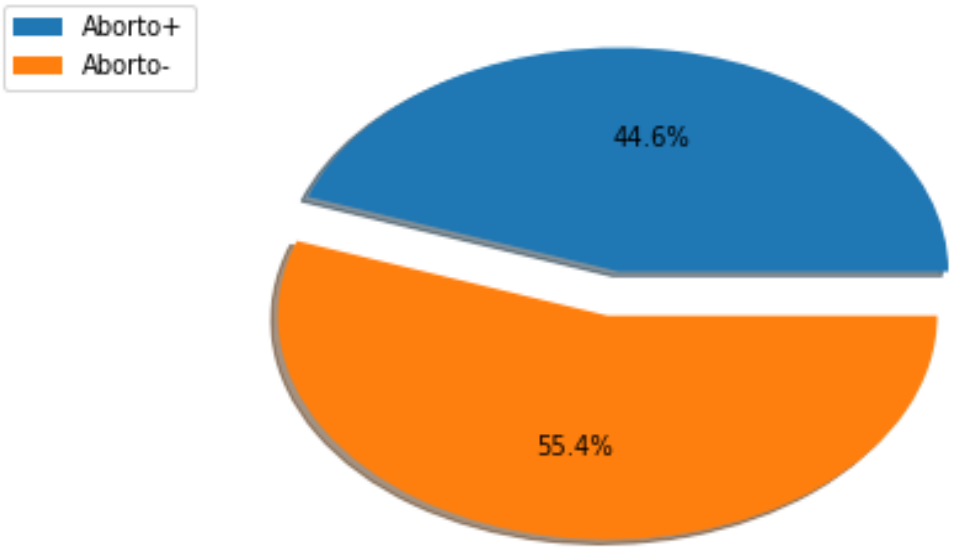


Fig. 22. Gráfica pastel porcentajes de tendencias a favor y en contra del Aborto basado en el análisis de contenidos con el algoritmo Naive Bayes

Luego de evaluar los clasificadores se procede a presentarlos de forma adecuada, esta información aparece en la Tabla 9, de la cuál, permite comparar los dos clasificadores, y por

los valores obtenidos para Naive Bayes, comprobar que si existió influencia de lo reducido de las claves de aprendizaje que se le pudo ingresar al algoritmo dentro de la fase de entrenamiento.

TABLA 9

Resultados de las Métricas de evaluación específicas para los clasificadores (weighted average)

Clasificador	TP Rate	FP Rate	Precisión	Recall	F1 score	ROC Área
Naive Bayes	1	0,4	0,791	0,8	0,791	0,81
Arboles de Decisión.	1	0,021	0.979	0.989	0.989	0,989

Se obtuvo también el rendimiento de los clasificadores que asignó para los árboles de decisión 98,93% y para Naive Bayes es 80%, con dicha información se completó el primer análisis de arquitectura conceptual en la fase 5 presentando la Tabla 10, que muestra los resultados obtenidos desde Python con el uso del algoritmo Naive Bayes y Arboles de Decisión para las posiciones a favor y en contra del aborto, que en resumen, para el árbol de decisión los resultados arrojó 59,3% en contra y 40,7% a favor y para Naive Bayes 55,4% en contra y 44,6% a favor, se nota que difieren en algunos puntos porcentuales, pero dan el mismo resultado general, el cual es que, la posición en contra del aborto es mayor a la posición a favor.

TABLA 10

Porcentajes de resultados favor y en contra del Aborto, resultantes de los clasificadores y totales promediados

Algoritmo	Porcentaje A favor	Porcentaje en contra
Naive Bayes	44,6	55,4
Arboles de Decisión	40,7	59,3
Resultados Finales		
Promedio	42,65	57,35

Los resultados del promedio de las tendencias de la Tabla 10, permiten mostrar en forma de pastel Figura 23.

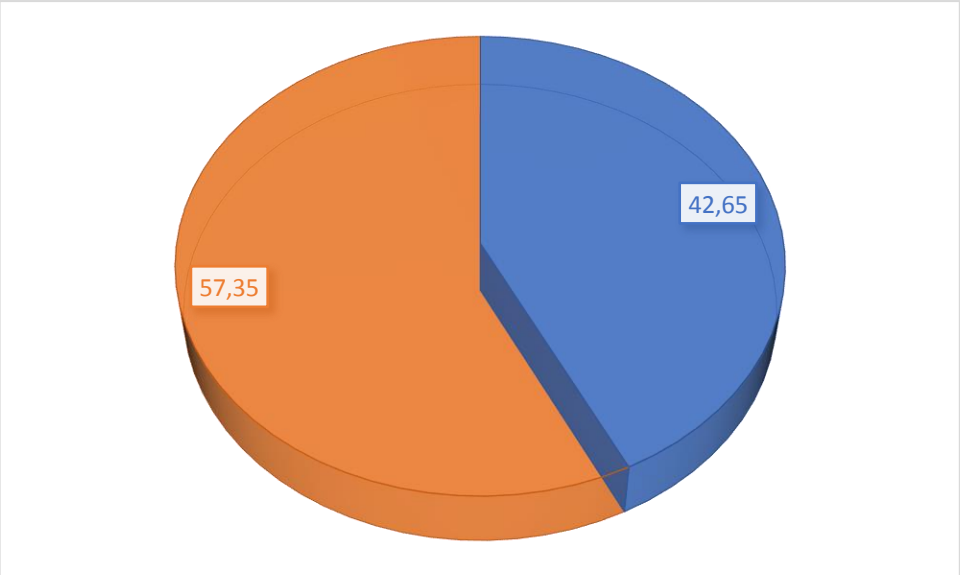


Fig. 23. Promedio de resultados de clasificadores sobre las tendencias

A favor y en contra del Aborto

Otro gráfico estadístico interesante es el que se presenta en la Figura 24, y es la evolución en el tiempo de los tweets a favor y en contra del aborto que está relacionada con el alcance al cuál se quería llegar en esta investigación.

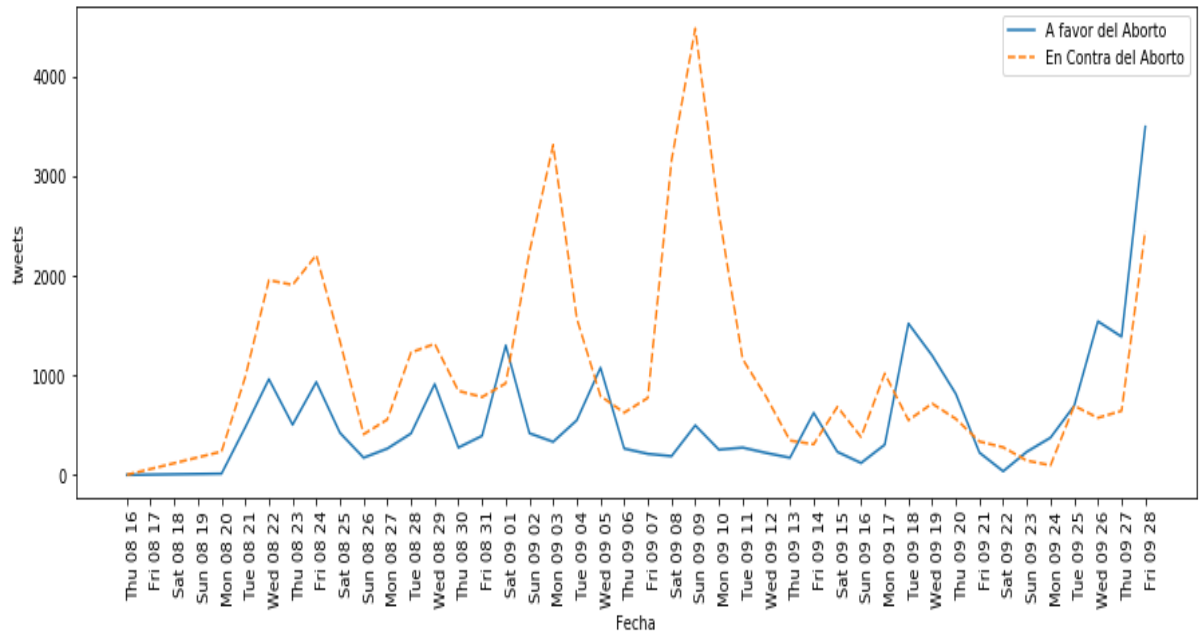


Fig. 24. Línea de tiempo en base a la frecuencia de Tweets a favor y en contra del Aborto

Se evidencia, que los tweets en contra son menos estables que los a favor y sus picos notables están el 24 de agosto, 03 de septiembre y el más alto el 09 de septiembre, la posición a favor empieza a dispararse al final de la muestra justamente al acercarse al 28 de septiembre y con tendencia a crecer.

Se puede notar, que desde que inició la muestra la posición en contra siempre supera a la a favor salvo excepciones como la que se evidencia a partir del 17 de septiembre.

Sobre el pico más alto de la posición en contra el día 09 de septiembre de 2018, se puede detallar más y explicarlo como una respuesta masiva en redes luego de que prácticamente había pasado un mes del día 08 de agosto de 2018, en el cual, se negó la legalidad del aborto en el senado de Argentina tema que repercutió en Latinoamérica en general y que generó opiniones también en el Ecuador que fueron recabadas en nuestra muestra. Con el pico en tweets a favor del aborto que está justamente en los últimos días de la toma de la muestra, se concluye que, es por lo que representa la fecha 28 de septiembre para las personas que apoyan el aborto.

Se destaca que en el top 5 de frecuencias precisamente el hashtag #28s que alude a la fecha conmemorativa ocupa la cuarta ubicación, evidenciando lo dicho anteriormente.

Los resultados de las ubicaciones se presentan en los Mapas de calor, la Figura 25, muestra el primer mapa del análisis del Aborto General, es decir, de todo el archivo JSON para el siguiente análisis se ocupó un archivo de extensión txt, el mismo que ya tuvo los filtros en base a la ponderación de los hashtags, dentro del mapa, el color rojo representa la clasificación **en contra** y el color azul **a favor**.

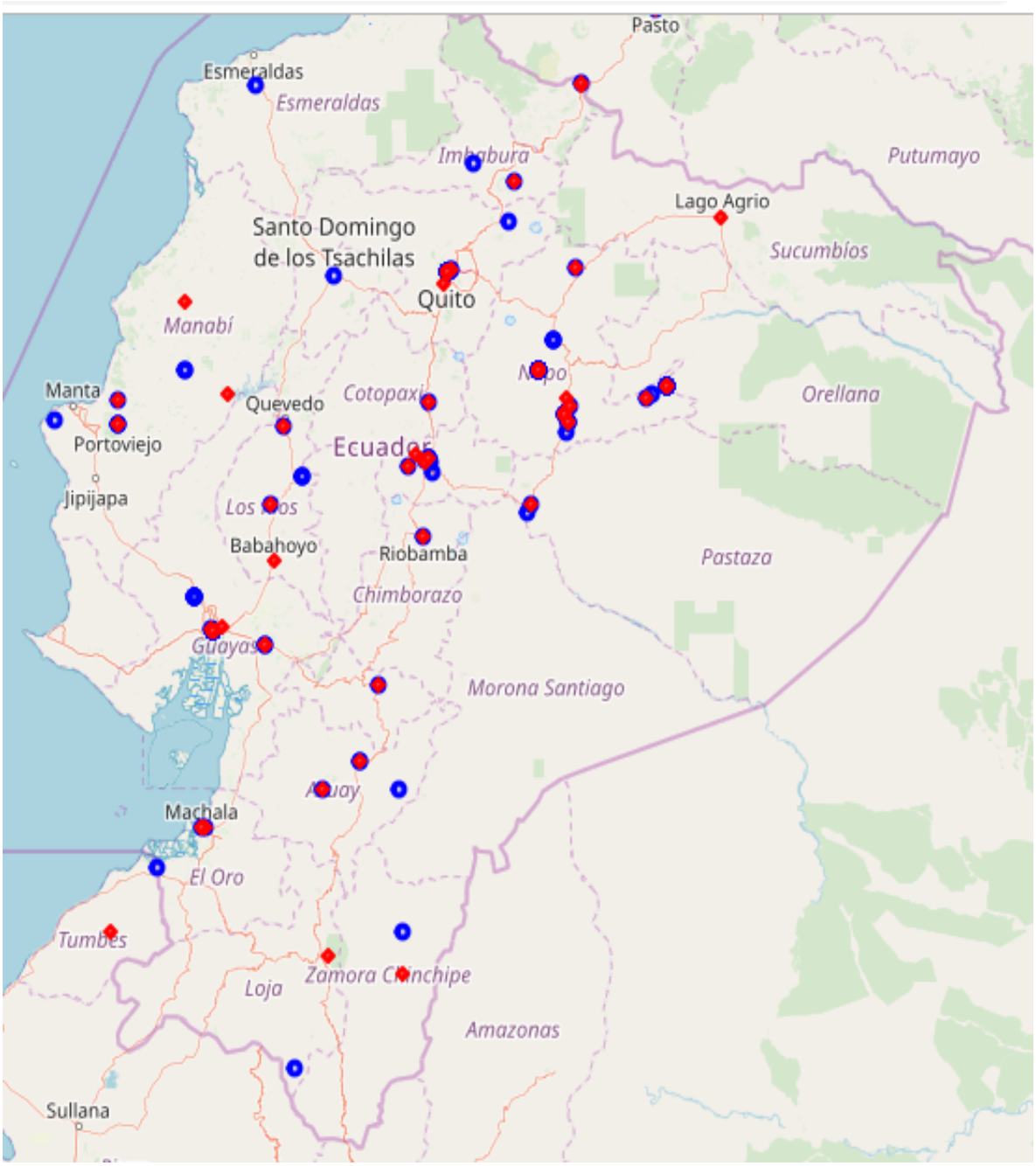


Fig. 25. Mapa de calor de comentarios a favor y en contra del aborto

La Figura 26, muestra El Mapa de Calor del análisis del tema del Aborto filtrado por el uso de Hashtags y es el resultado de las localizaciones luego que del archivo se usaran sólo los comentarios que contengan los hashtags con valoración y ponderación.

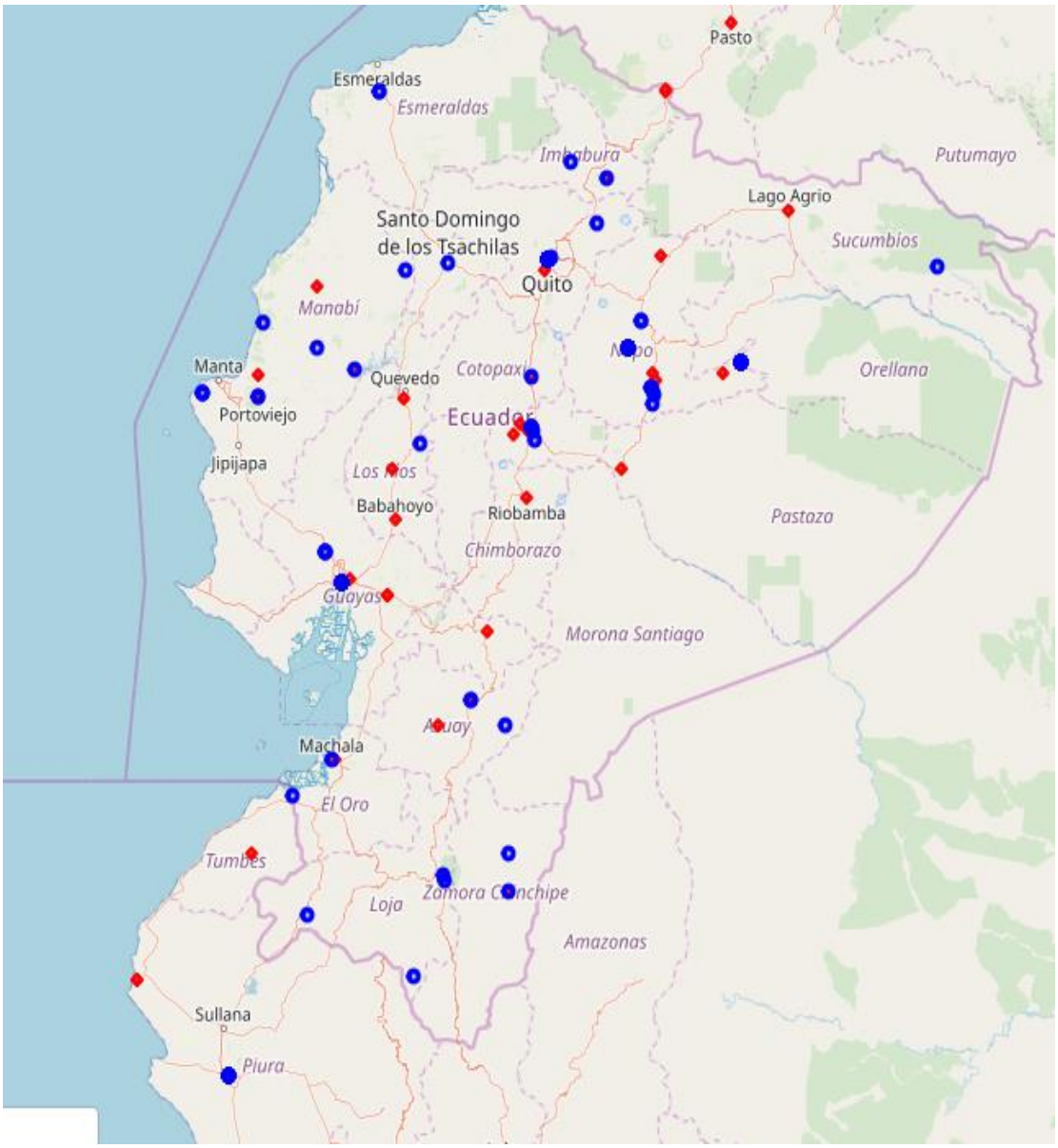


Fig. 26. Mapas de Calor bajo el filtrado de hashtags

Se puede observar, que las figuras son similares, sin embargo, la lectura que se puede dar es que la mayoría de los usuarios de la muestra que colocaron una ubicación existente y de forma correcta, también emitieron su tweet sobre la temática del aborto y lo hicieron señalando un hashtag definido a favor o en contra.

La Figura 27, es la Nube de palabras sobre las cuentas de usuario que realizaron más publicaciones, también los que recibieron más “retweets”, y/o fueron más mencionadas desde otras cuentas, este resultado permitió ver la influencia de estos usuarios dentro de los datos tomados en la muestra.

Fig. 27. Nube de Palabras de usuarios influyentes

Se recalca, que estos no son los hashtags, sino los usuarios tomados del atributo **name** del objeto **user** de Python; la cuenta Salvemoslas2vidas con tendencia **en contra** del aborto ocupa el primer lugar seguida de la cuenta abortolegalya en el segundo casillero y con tendencia **a favor**, porlavida2014 ocupa la tercera ubicación, sialavida la cuarta éstas dos **en contra** y 28s que es una cuenta **a favor** ocupa la quinta casilla cerrando este top 5 de usuarios o cuentas incluyentes,

Muy visibles en la nube de palabras se encuentran las cuentas de MashiRafael y Lenin ex y actual presidente del Ecuador respectivamente y esto se debe no necesariamente a que los mencionados políticos generen tweets respecto al tema, sino, que: esta nube sale del archivo inicial de recolección de datos que como quedó señalado desde esta fase 2 marcaba como territorio de alcance de la investigación al Ecuador, y que muchos usuarios hacen menciones en sus tweets a personas importantes para ganar cierta notoriedad en lo que publican,

Para la red social Facebook, se presentan los resultados visibles en la Tabla 11, obtenidos de las páginas: “Aborto Libre Ec” y “Pro Vida Ecuador”.

TABLA 11

Comparativa de interacciones entre las Páginas de Facebook “Aborto Libre Ec” y “Provida Ecuador”

ESTADISTICAS EN PUBLICACIONES DESDE 16 AGOSTO HASTA 29 SEP 2018			
ESTADISTICAS	DESCRIPCION	ABORTO LIBRE EC	PROVIDA ECUADOR
Seguidores	Número de seguidores de las páginas	2760	4200
TYPE_POST	se refiere a los tipos de publicaciones existentes en la pagina	Todos	Todos
POST_PUBLISHED	Número de publicaciones en el intervalo de muestreo	13	18
LIKES	Número de reacciones me gusta en publicaciones	489	47
ME ENCANTA	Número de reacciones me encanta en publicaciones	332	15
ME ENOJA	Número de reacciones me enoja en publicaciones	0	4
ME DIVIERTE	Número de reacciones me divierte en publicaciones	12	2
ME ENTRISTECE	Número de reacciones me entristece en publicaciones	0	5
COMENTARIOS	Número total de comentarios incluidos respuestas	19	9
COMENT_BASE	Número total de solo los comentarios iniciales	16	9
REPLICAS	Respuestas a comentarios	3	0
COMENTS LIKE	Reacciones a comentarios	0	0
COMPARTIDOS	Número de veces que las publicaciones fueron compartidas	392	26
CONSOLIDADO	Suma total de las estadísticas	**1283**	**117**

Se analiza de la información obtenida de la página “Pro Vida Ecuador”, que, a pesar de tener mayor cantidad de seguidores, posee menos información relevante al tema y por ende menos interacciones; además, que los seguidores de la página “Aborto Libre Ec” son mucho más entusiastas e interactúan mucho más, es decir, se encontró mayor difusión y apoyo a la página a favor del aborto, que es un dato interesante puesto que hasta ese momento, todos los resultados obtenidos eran mayores para la posición en contra del aborto.

Considerando el trabajo de (Mata-Gómez, Gilete-Tejero, Rico-Cotelo, Royano-Sánchez, & Ortega-Martínez, 2018), en primer lugar al tomar en cuenta el número de seguidores de cada

página de Facebook, se coincide con ellos en que la cantidad de seguidores no es directamente proporcional a la calidad de los contenidos; por otro lado se notó que, es importante también reconocer que las páginas especializadas en un tema dentro de Facebook, deben tomar en cuenta que es importante subir periódicamente contenidos preocupándose: a) por la calidad tanto en la forma y el fondo de éstos y b) por su veracidad.

Luego de obtener la Tabla comparativa de Facebook, se pudo encontrar una coincidencia con los resultados obtenidos en Twitter en cuanto a los porcentajes de cada tendencia, basándose solamente en el número de seguidores de las dos páginas, la discrepancia, se dio por el número de interacciones que generaron sus contenidos que son muy bajas para la página elegida en esta red social en contra del aborto, con un amplio margen de diferencia con las de la página a favor, esto enriquece este trabajo y muestra que hay un completo campo de análisis abierto para investigaciones mucho más profundas sobre el tema tanto en criterios dentro de una sola red social comparada con otra, como en un conjunto total de datos mezclados.

Finalmente, se tomó en cuenta un valor, que si bien cierto no está estandarizado, permitió realizar un análisis posterior, para comprobar si los datos recabados en la muestra de Twitter tienen una cierta relación solamente con él número de "me gusta" de las páginas de Facebook "Pro Vida Ecuador" y "Aborto Legal Ec" presentados en la Tabla 12**, (l**as capturas de pantalla que muestran: portada, perfil, numero de likes y seguidores se muestran posteriormente en el Anexo A y Anexo B respectivamente, a la fecha del 24 de enero de 2019 (aproximadamente 4 meses luego de que se tomaron los últimos datos de la muestra principal).

TABLA 12

Número de "me gusta" de páginas de Facebook analizadas en esta investigación

Página	**Tendencia respecto al aborto**	**Nro. De "me gusta" a la página (al 24-01-19)**
Pro Vida Ecuador	En contra	4819
Aborto Libre Ec	A favor	3604

El número total de "me gusta" de ambas páginas fue de 8423, en un porcentaje a favor del aborto 42,79% y en contra 57,21%, valor bastante parecido al del resultado conseguido en el promedio de Twitter que arrojó 57,35% en contra y 42,65% a favor, mostrados en la presentación de resultados específicamente en la Figura 23.

La lectura que se obtiene de esta prueba es que, se puede calificar los resultados de la fase 5 del proceso de arquitectura conceptual propuesto como confiables, ya que coinciden

en los resultados generales obtenidos, en que la posición en contra del aborto supera a la posición a favor, los valores difieren solamente en 0,14 puntos porcentuales con los obtenidos bajo el criterio de “me gusta” de páginas de Facebook.

3.2 Discusión

Dentro de su caso de estudio el articulo (Duwairi & AlFaqeeh, 2015), presenta una tabla comparativa la cuál se usó para el estudio de los resultados de Facebook añadiendole las filas “seguidores”,“me encanta”, “me entristece”, “me enoja”, “me divierte”, se lo creyó conviente, para marcar un precedente para trabajos futuros en los cuales, se pueda analizar por estas reacciones las emociones inherentes a los seres humanos que son analizadas hasta llegar a formas gráficas de respresentarlas por (Perikos & Hatzilygeroudis, 2018).

Sobre cualquier tema técnico o tecnológico, el Aborto como tema elegido para el desarrollo de la Arquitectura Conceptual, es el principal aporte de esta investigación, por ser uno de los más comentados en los tiempos actuales en la sociedad en general así como por los usuarios de Twitter en específico, donde lo que se expresa en Ecuador, sustentado por el 97,6 de precisión del árbol de decisión, representa un 40,7% a favor del aborto y en contra 59, 3%, éste clasificador superó al de Naive Bayes que arrojó un 79,1% de precisión, adentrándose al tema de la evaluación de los clasificadores, (Vila, y otros, 2019) concuerdan con este trabajo en un porcentaje similar para su árbol 96,8%, que coincidencialmente, también aunque por poco margen, superó al 96,7% de precisión del clasificador Naive Bayes.

Con el uso los resultados generados, se puede evidenciar que aparece información adicional importante y coherente con el impacto social que se quiso lograr, tanto en las nubes de palabras de hashtags usados y usuarios influyentes, de causas que se las puede considerar afines a las tendencias, por ejemplo, el hashtag #niunamenos (que promueve mayormente la erradicación del femicidio, y el no maltrato a la mujer en ningún área y bajo ninguna circunstancia), aparece considerablemente en la muestra, esto se entiende porque, al promover que el aborto es un derecho propio de cada mujer está a favor del aborto; del otro lado el hashtag #conmishijosnotemetas y la cuenta del mismo nombre (que apoyan la causa que mayormente rechaza la enseñanza de la ideología de género y demás corrientes por considerarlas atentatorias contra la moral de las personas), también aparece notoriamente en la muestra, se cree porque, considera al aborto un asesinato, es decir, tiene tendencia en contra. Es evidente, que si se quiere hacer estudios concretos sobre el femicidio y la ideología de género basándose en (Niklander, 2017), se debería pensar en el uso de los hashtags #niunamenos y #conmishijosnotemetas respectivamente; para estudios sobre el tema del aborto los hashtags específicos que se sugieren son: #salvemoslasdosvidas y #abortolegalya, que fueron los tuvieron mayor frecuencia de aparición.

Con respecto al artículo (Mata-Gómez, Gilete-Tejero, Rico-Cotelo, Royano-Sánchez, & Ortega-Martínez, 2018), se consiguió establecer puntos de comparación ente las dos páginas de Facebook elegidas, en base al número de me gusta, 4819 para “Pro Vida Ecuador” 3604

para AbortoLibreEc (información tomada el 24 de enero de 2019); es interesante que el número de seguidores es menor respecto al número de “me gusta” de la misma páginas para “Pro Vida Ecuador”4785 (34 menos) y mayor para AbortoLibreEc 3640 (36 más), ésto tiene que ver con el interes que producen los contenidos de cada página, se llegó a concluir en base a esos resultados, que la página más popular, no necesariamente es la que genera contenido más interesante, otra conclusión es que las personas, pueden dar un “me gusta” a una pagina de Facebook, por lo que representa, promueve o defiende, más que por los contenidos que publica.

Se sabe de antemano, que bajo ningún concepto este trabajo pondrá fin a la controversia sobre el aborto y que los resultados obtenidos no serán los definitivos en las tendencias a favor y en contra, sin embargo, existe la seguridad que aportará un punto de referencia confiable.

Una de las limitaciones del presente trabajo, fue el no poder hacer un análisis en los textos publicados en Facebook, por eso se enfatiza que hay que tomar en cuenta que existen cosas que no dependen del investigador, sino, de las políticas de seguridad y privacidad de los diferentes sitios de los cuáles queremos minar los datos, es importante prevenir inconvenientes, sin embargo, lo más seguro es que van a aparecer de cualquier forma a medida que se van desarrollando los trabajos, pero, hay que saber sortearlos adecuadamente.

3.3Análisis de Impacto.

El análisis de impacto de este trabajo, se deberá medir en base a la importancia social del tema que se propuso tratar, por esta razón, se cree que es muy importante que esta información sea compartida a las personas idóneas por ejemplo en el campo psicológico, educativo, político y aún religioso, para que, con los resultados obtenidos y los análisis propuestos cada especialista dentro de sus competencias pueda sacar sus propias conclusiones y pueda aportar en favor de la sociedad en general.

Cabe destacar que, actualmente en la Asamblea Nacional del Ecuador, se discute el despenalizar el aborto por violación para todas las mujeres ya que en el país el aborto no tiene implicaciones legales solamente cuando:

a. Se lo hace con fines terapéuticos, es decir, cuando en una cirugía por salvar la vida de la mujer pudo morir el feto o,
b. después de una violación, una mujer con enfermedades mentales quedo embarazada.

En específico, la Comisión de Justicia de la Asamblea, es la responsable de profundizar y dar una respuesta sobre el tema que centra el debate en los artículos de la Constitución, (Asamblea Nacional Constituyente de Ecuador, 2008) que tienen un cierto nivel de contradicción, puesto que, el artículo 45, habla de los derechos desde la concepción y el artículo 43, menciona la protección prioritaria y cuidado de la salud integral de la mujer y de su vida durante el embarazo, parto y posparto, los textos completos de estos artículos se los puede ver en el Anexo C de esta investigación.

En lo que respecta específicamente a la parte técnica de este trabajo, se confía que motivará a más estudiantes a introducirse en la minería de datos, el análisis de contenidos y demás temas afines, usando otros lenguajes de programación, distintas redes sociales, y nuevas técnicas de minería de datos, ya no solo para el análisis de contenidos y análisis de sentimientos de textos sino hasta llegar el análisis de las imágenes como observamos en los trabajos de (Baecchi, Uricchio, Bertini, & Bimbo, 2015), (Li, Fan, Jiang, Lei, & Liu, 2018) entre otros, para hacer nuevas investigaciones con el objetivo de sacar conocimiento de otro tipo de temas o también para seguir profundizando en el del aborto.

En la parte psicológica, usando el conocimiento que adquirimos en base a la muestra que se tomó, se evidenció principalmente, que nuestra sociedad tiende cada vez más a ser liberal a pesar de que todavía es conservadora. Los resultados muestran, que el criterio de la gente joven está tomando un peso relevante y que su influencia crece, los temas como el Aborto dejaron hace rato de ser un tabú social para ser comentados con libertad, pero siguen siendo materia de análisis (más allá de su despenalización o no), las causas y efectos de esta práctica. Opinar sobre un tema, no necesariamente indica tener conocimiento profundo acerca

del mismo lo que lleva a pensar que nunca serán demasiado los esfuerzos para educarse sobre todo lo que implica el aborto desde las dos perspectivas.

Otro punto de impacto, tiene que ver con que las redes sociales son una herramienta muy poderosa para propagar todo tipo de contenidos, por esto, el uso responsable de las mismas es importante, a pesar de que en este trabajo no se profundiza en buscar los contenidos o porcentajes de información falsa, ya que, el empeño estaba enfocado en buscar posiciones a favor y en contra del aborto, no se puede dejar de mencionar que la información falsa o aún verídica, pero publicada sin ningún tipo de criterios, causa un efecto negativo considerable. Cabe destacar que cuando se realizó este análisis, Ecuador y específicamente la ciudad de Ibarra se encontraban dentro de una coyuntura de tensión, ya que, comentarios, imágenes y videos de violencia viralizados en las redes sociales (principalmente en Facebook), provocaron stress social, que, como punto más álgido, desembocaron en comportamientos nunca antes vistos en esta ciudad como la Xenofobia en contra de ciudadanos venezolanos; todo lo relacionado al tema ha sido publicado bajo los hashtags #ibarraestadeluto y #Diana. Solo días después de estos hechos lamentables, Venezuela vivió momentos cruciales dentro de su situación política actual, puesto, que luego de las marchas de convocadas a propósito de conmemorar el retorno a la democracia en Venezuela en el golpe de Estado del 23 de enero de 1958, el presidente de la Asamblea de este país Juan Guaidó, quien a la vez representa la oposición al régimen, se autoproclamó presidente del país generando reacciones a nivel mundial. el hashtag #GuaidoEsMiPresidente es uno de los que refleja el apoyo a esta iniciativa y que junto a hashtags como #MaduroRenuncia forman parte de la posición contra Nicolás Maduro, quien fue el ganador de las últimas elecciones presidenciales del vecino país, el análisis de esos comicios se lo observa en el trabajo de (Niklander, 2017), en contra parte, los hashtags #VenezuelaConMaduro y #VenezuelaSeRespeta muestran respectivamente, el apoyo a Maduro y la inconformidad por el respaldo que recibió Guaidó desde la comunidad internacional catalogándola como injerencia en la soberanía del Estado.

Los ejemplos citados, son una muestra clara del caudal de datos y opiniones que se generan en las redes sociales diariamente sobre los diversos temas de interés general permanente, o que tienen que ver con la coyuntura, quedando demostrado que aplicar la Arquitectura Conceptual a estos temas y conseguir correctamente la información relevante logra causar un alto impacto dentro de la sociedad.

Conclusiones

En el presente trabajo se abordó el uso de la Minería de Datos para realizar una Arquitectura Conceptual sobre el tema del Aborto, luego de culminar las fases propuestas se concluye que:

1. Se logró obtener el análisis de contenidos al evaluar los hashtags con su polarización en Twitter y el número de seguidores de las páginas elegidas en Facebook, y también en una forma general se consiguió el análisis de sentimientos (esto solamente en Twitter) al usar los clasificadores para definir la polaridad del contenido del texto del tweet, tanto el análisis de contenidos como en el análisis de sentimientos, son muy similares en sus resultados, esto se debe a que el texto de los tweets suele estar muy relacionado con los hashtags utilizados en los mismos, salvo en algunos casos donde se usa el hashtag para dentro del texto mostrar oposición.
2. Dentro de la literatura usada como guía para el desarrollo de este trabajo, se ha podido clasificar en base a los datos recolectados por los distintos investigadores cinco categorías de análisis: a) de textos b) de textos y perfiles que los postean, c) textos, imágenes y contenido multimedia, d) hashtags específicos y e) perfiles de usuarios y páginas de redes sociales. La tercera categoría, es decir, la que tiene que ver con el análisis de textos, imágenes y contenido multimedia es la más compleja, puesto que, no se trata de analizar las reacciones que genera ese contenido (compartir, comentarios de respuesta, me gusta, o cualquier otra reacción) o de indagar en el texto que acompaña a la publicación (como una nota al pie), sino que, busca analizar el sentimiento que genera la imagen, el video u otro contenido multimedia en sí mismo, por esto las técnicas usadas para obtener los resultados son mucho más sofisticadas.
3. Se ha cumplido cada fase de la arquitectura conceptual propuesta, de donde se destaca en la recolección de datos, una muestra idónea por el lapso de tiempo en la que se tomó, ya que, las fechas en las que hubo una mayor cantidad de tweets coinciden con eventos que se dieron a nivel internacional, en las que diferentes colectivos manifestaron su postura frente al aborto, marcando una clara evidencia de la influencia que ejercen determinados usuarios, en las opiniones de otros; y en la presentación de resultados el conocimiento en concreto generado que muestra que en el Ecuador la posición en contra del aborto todavía es superior a la posición a favor.

4. El uso de Python como lenguaje de programación para generar los scripts necesarios al llevar a cabo la construcción de la Arquitectura Conceptual, fue conveniente ya que existen plataformas completas donde se incluyen los IDE y contiene documentación adecuada.

5. Según la información mostrada en los mapas de calor, es en la sierra, con mayor porcentaje en la provincia de Pichincha, en donde hay una mayor actividad en Twitter, aunque es importante señalar que, en la mayoría de los tweets de la muestra, el campo location, no estaba activo, por lo tanto, las ubicaciones presentadas en los mapas de calor no reflejan la cantidad de tweets de la muestra.

6. Se marca un precedente ya que no se ha encontrado datos (oficiales o no oficiales) en Ecuador, de porcentajes de las posiciones a favor y en contra del aborto ya sea de forma general, menos aún, de información de redes sociales de forma específica como la que se logró obtener en la arquitectura conceptual, donde se pudo ampliar el campo de estudio a otros elementos de los tweets de la muestra, dichos elementos permitieron conocer aspectos tan relevantes, como por ejemplo los lugares del país, en los que la despenalización del aborto es un tema de mucho interés.

7. Mediante la obtención de información concreta de los porcentajes de las posiciones a favor y en contra del Aborto sustentada en los resultados obtenidos a partir de los datos recabados (en Twitter en promedio 43,3% a favor y 56,7% en contra de y en Facebook 43,2% a favor y 56,8 % en contra), se puede responder a la pregunta de investigación ¿Cómo una arquitectura conceptual permite realizar un análisis de opiniones sobre el aborto en las redes sociales Twitter y Facebook? realizada al inicio de este documento en el Planteamiento del Problema.

8. Al comparar a los clasificadores por el resultado de su rendimiento se probó, que el árbol de decisión, con un 98,93% superó a Naive Bayes que alcanzó un 80%, lo que le hace ser el clasificador más confiable para este tipo de investigaciones.

Recomendaciones

1. Dar importancia al estudio de la minería de datos al análisis de contenidos y al análisis de sentimientos en redes sociales impartiendo seminarios y talleres enfocados a extraer conocimiento no solo en temas netamente académicos sino del interés común.

2 Como trabajos a futuro, investigaciones del mismo tema, en una muestra tomada en este año 2019 o en los posteriores para poder hacer una comparativa con este trabajo y así determinar si los porcentajes han variado de forma leve o si se marcan nuevas tendencias con claras diferencias a las obtenidas.

3 El alcanzar a topar en este trabajo de manera muy general el análisis de sentimientos, permite indicar que es un área poco explotada y muy interesante para ir más allá del contenido literal de una publicación o de un tweet y poder llegar a profundizar aún en el contenido multimedia siendo una buena opción para seguir investigando, ya que existe documentación adecuada sobre el tema.

4 Tomar en cuenta las nuevas redes sociales como Instagram para realizar investigaciones del tema del Aborto, encontrando diferencias o semejanzas con el presente trabajo e introduciendo nuevos criterios como el género, nivel de educación y edades de las personas que opinan y publican,

5 Que, con el uso de las mismas herramientas de programación u otras más acordes a cada investigador, se pueda realizar más trabajos de análisis de contenidos en temas de relevancia social, como el femicidio o la ideología de género, adentrándose mucho más en el análisis de sentimientos del texto de las publicaciones, además, de tomar en cuenta todas las reacciones que permite Facebook o las particularidades de cada red social.

6 Promover se genere dentro de los perfiles en las diferentes redes sociales de la Universidad Técnica del Norte, el debate de los diferentes temas de interés social y académico para que pueda ser usada por investigadores y contrastada para analizar si coincide o no con el pensamiento del Ecuador en general.

7 Dentro del enfoque social y humano, y bajo el conocimiento de los hechos Xenofóbicos suscitados en Ibarra recientemente, se recomienda de que el hecho de tener discrepancias con otra persona respecto a criterios sobre temas controversiales, no debe ser argumento válido para aislarla, discriminarla, dejar de ser solidarios o peor aún violentarla.

Anexo A: “Pro Vida Ecuador” Página en contra del Aborto Foto de perfil y portada, y número de seguidores, captura tomada el 24-01-19

Anexo B: “Aborto Libre EC”, Página a favor del del Aborto Foto de perfil y portada, y número de seguidores, captura tomada el 24-01-19

Anexo C: Tabla que muestra los artículos de la Constitución sobre el tema del Aborto.

Articulos de la Constitución	
Articulo 45 Usado por la posición en contra del Aborto	Art. 45.- Las niñas, niños y adolescentes gozarán de los derechos comunes del ser humano, además de los específicos de su edad. El Estado reconocerá y garantizará la vida, incluido el cuidado y protección **desde la concepción**. Las niñas, niños y adolescentes tienen derecho a la integridad física y psíquica; a su identidad, nombre y ciudadanía; a la salud integral y nutrición; a la educación y cultura, al deporte y recreación; a la seguridad 35 social; a tener una familia y disfrutar de la convivencia familiar y comunitaria; a la participación social; al respeto de su libertad y dignidad; a ser consultados en los asuntos que les afecten; a educarse de manera prioritaria en su idioma y en los contextos culturales propios de sus pueblos y nacionalidades; y a recibir información acerca de sus progenitores o familiares ausentes, salvo que fuera perjudicial para su bienestar. El Estado garantizará su libertad de expresión y asociación, el funcionamiento libre de los consejos estudiantiles y demás formas asociativas. **(Asamblea Nacional Constituyente de Ecuador, 2008)**
Artículo 43 Usado por la posición a favor del Aborto	El Estado garantizará a las mujeres embarazadas y en periodo de lactancia los derechos a: 1. No ser discriminadas por su embarazo en los ámbitos educativo, social y laboral. 2. La gratuidad de los servicios de salud materna. **3. La protección prioritaria y cuidado de su salud integral y de su vida durante el embarazo, parto y posparto.** 4. Disponer de las facilidades necesarias para su recuperación después del embarazo y durante el periodo de lactancia. **(Asamblea Nacional Constituyente de Ecuador, 2008)**

Bibliografía

Asamblea Nacional Constituyente de Ecuador, d. 2. (2008). Constitución de la República del Ecuador.

Baecchi, C., Uricchio, T., Bertini, M., & Bimbo, A. D. (2015). A multimodal feature learning approach for sentiment analysis of social network multimedia. *Multimed Tools Appl (2016)*, 19. doi:10.1007/s11042-015-2646-x

Bonzanini, M. (2016). *Mastering Social Media Mining with Python.* Birmingham: Packt Publishing Ltd.

Borja, M. (2016). *BrainSINS*. Recuperado el 10 de diciembre de 2018, de BrainSINS: https://www.brainsins.com/es/blog/analisis-del-sentimiento-y-mineria-de-opiniones/99679

BoundingBox. (2017). Obtenido de https://boundingbox.klokantech.com/

Bouza, C., & Santiago, A. (2014). LA MINERÍA DE DATOS: ARBOLES DE DECISION Y SU APLICACION EN ESTUDIOS MEDICOS. En C. Bouza, & A. Santiago, *LA MINERÍA DE DATOS: ARBOLES DE DECISION Y SU APLICACION EN ESTUDIOS MEDICOS* (págs. 64-78).

CHiristianCH. (02 de mayo de 2013). *Clasificador Naïve Bayes* . Obtenido de Clasificador Naïve Bayes : http://naivebayes.blogspot.com

Comunicólogos. (2016). Recuperado el 10 de diciembre de 2018, de Comunicólogos: https://www.comunicologos.com/enciclopedia/t%C3%A9cnicas/an%C3%A1lisis-de-contenido/

Córdoba-Fallas, L. (16 de junio de 2011). *Mineria de Datos*. Obtenido de Mineria de Datos: http://cor-mineriadedatos.blogspot.com/2011/06/weka.html

Duwairi, R. M., & AlFaqeeh, M. (2015). RUM Extractor: A Facebook Extractor for Data Analysis. 5. doi:10.1109/FiCloud.2015.116

Ekman, P. (1999). *Basic emotions. Handbook of cognition and emotion,.*

Ekos, R. (12 de agosto de 2018). *Ekos*. Obtenido de Ekos: http://www.ekosnegocios.com/negocios/verArticuloContenido.aspx?idArt=10877

García Serrano, A. (2016). *Inteligencia Artificial Fundamentos, Prácticas y Aplicaciones* (2da Edición ed.). España: RC Libros.

Gullo, F. (2015). From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. 5. doi:10.1016/j.phpro.2015.02.005

Inbal Yahav, Shehory, O., & Schwartz, a. D. (08 de Agosto de 2015). Comments Mining With TF-IDF: The Inherent Bias and Its Removal. 14. doi:10.1109/TKDE.2018.2840127

Johnsen, J. W., & Franke, K. (2017). Feasibility Study of Social Network Analysis on Loosely. 5. doi:10.1016/j.procs.2017.05.172

Kaynar, O., Görmez, Y., Arslan, H., & Demirkoparan, F. (2017). Duygu Analizinde Öznitelik Seçim Yöntemleri. 5. doi:10.1109/IDAP2017:8090187

Kazil, J., & Jarmul, K. (2016). *Data Wrangling with Python.* O'Reilly Media.

Lara, J. (2014). *Minería de Datos.* Madrid: Edima.

Li, Z., Fan, Y., Jiang, B., Lei, T., & Liu, W. (2018). A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications*, 29. doi:10.1007/s11042-018-6445-z

Mata-Gómez, J., Gilete-Tejero, I., Rico-Cotelo, M., Royano-Sánchez, M., & Ortega-Martínez, M. (2018). Situación actual del uso de redes socialesen Neurocirugía en España. *NEUCIR-313*, 7. doi:10.1016/j.neucir.2018.01.001

Matplotlib: Python Plotting. (2012). Obtenido de Matplotlib: Python 2D plotting: http://matplotlib.org

McKinney, W. (2013). *Python for Data Analysis.* O'Reilly Media, Inc..

Migurski, M. (10 de Noviembre de 2012). *a beginners guide to streamed data from Twitter*. Obtenido de a beginners guide to streamed data from Twitter: http://mike.teczno.com/notes/streaming-data-from-twitter.html

Nacarro-Arango, R. (28 de noviembre de 2017). *DABACODLAB*. Obtenido de DABACODLAB: https://dabacodlabblog.wordpress.com/2017/11/28/que-es-json/

Niklander, S. (2017). Content Analysis on Social Networks: Exploring the #Maduro Hashtag. 5. doi:10.1109/ICCNI.2017.8123803

NODEXL. (2018). Recuperado el 10 de 01 de 2019, de NODEXL: https://nodexl.com/

Page, L. (2018). *Alphabet*. Obtenido de Alphabet: https://abc.xyz

parsehub. (2018). Recuperado el 16 de octubre de 2018, de parsehub: https://www.parsehub.com/features

Perez-Porto, J., & Gardey, A. (2013). *Definicion.de*. Obtenido de Definicion.de: https://definicion.de/facebook/

Perikos, I., & Hatzilygeroudis, I. (2018). A Framework for Analyzing Big Social Data and Modelling Emotions in Social Media. 5. doi:10.1109/BigDataService.2018.00020

Purnomo, M. H., Sumpeno, S., Setiawan, E. I., & Diana Purwitasaria, c. (2017). Biomedical Engineering Research in the Social Network Analysis Era:. 7. doi:10.1016/j.procs.2017.10.049

Roesslein, J. (2009). *Tweepy*. Obtenido de https://tweepy.readthedocs.io/en/v3.5.0/getting_started.html

Roldán, C. (2017). ANÁLISIS DE CONTENIDOS DE TWITTER CASO DE ESTUDIO: ELECCIONES PRIMARIAS DEL PSOE 2017 . Madrid, España.

Sánchez, J. (2 de Abril de 2011). *Breve Paseo por la minería de Datos*. Obtenido de Breve Paseo por la minería de Datos: https://es.slideshare.net/jculacio/brevepaseoporla-mineradedatos

Sierra, J. (s.f.). *Inteligencia Artificial.*

Sogo, J. G. (06 de mayo de 2016). *Lingẅars.* Obtenido de Lingẅars: http://lingwars.github.io/blog/twitter-stream.html

Valerio, G., Herrera-Murillo, D. J., Villanueva-Puente, F., Herrera-Murillo, a., & Rodríguez-Martínez, M. d. (2015). The Relationship between Post Formats and Digital Engagement: A Study of the Facebook Pages of Mexican Universities. *RUSC. Universities and Knowledge Society Journal*, 14. doi:10.7238/rusc.v12i1.1887

Vila, D., Cisneros, S., Granda, P., Ortega, C., Posso-Yepez, M., & García-Santillan, I. (2019). Detection of Desertion Patterns in University. *Springer Nature Switzerland AG 2019*, 10. doi:10.1007/978-3-030-05532-5_31

Yang, L., Tian, Y., Li, J., Ma, J., & Zhang, J. (2017). Identifying opinion leaders in social networks with topic limitation. *Cluster Comput*, 11. doi:10.1007/s10586-017-0732-8

Yassine, M., & Hajj, H. (2010). A Framework for Emotion Mining from Text in Online Social Networks. *2010 IEEE International Conference on Data Mining Workshops*, 7. doi:10.1109/ICDMW.2010.75

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2018). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 47. doi:10.1007/s10115-018-1236-4