



# Análisis de Sentimientos desde la influencia de Twitter en los procesos de Pólizas de seguros del ámbito colombiano

Gustavo Adolfo Martínez Misal

Universidad Jorge Tadeo lozano  
Facultad de ciencias naturales e ingeniería  
Bogotá, Colombia  
2021 - I semestre



# Análisis de Sentimientos desde la influencia de Twitter en los procesos de Pólizas de seguros del ámbito colombiano

Gustavo Adolfo Martínez Misal

Tesis de grado presentada como requisito parcial para optar al título de:  
**Magister en ingeniería y analítica de datos**

Director de proyecto:  
Ph.D., Sebastián Zapata Ramírez

Universidad Jorge Tadeo lozano  
Facultad de ciencias naturales e ingeniería  
Bogotá, Colombia  
2021- I semestre

## RESUMEN

Las redes sociales están jugando un papel importante en la sociedad siendo así un acontecimiento notable de la comunicación hoy en día. Esto permite que los diferentes usuarios puedan publicar una opinión acerca de un determinado tema haciéndolo a través de internet. Esto hace de estas plataformas una fuente para exploración de información que obliga a aprovechar dichos datos para poder interpretarlos con la ayuda de las opiniones realizadas y así con lo anterior llevado a cabo, poder tomar decisiones cruciales para determinar el camino de las aseguradoras en Colombia.

Este proyecto busca desarrollar una visualización de datos que permita analizar el sentimiento (positivo, negativo y neutro) de las opiniones en Twitter, además de presentar el análisis de estos datos (*Tweets*) y el procedimiento con fines de investigación.

Finalmente se proponen técnicas de aprendizaje automático y procesamiento del lenguaje natural aplicando procesamiento basado en texto con análisis de sentimientos.

**Palabras Clave:** Twitter, Análisis de sentimientos, Aprendizaje automático, Aseguradoras.

## ABSTRACT

Social networks are playing an important role in our society, thus being an important communication event today. This allows different users to publish an opinion about a certain topic by doing so through the internet. This makes these platforms a source for information exploration that forces us to take advantage of said data to be able to interpret them with the help of the opinions made and thus with the above carried out, to be able to make crucial decisions to determine the path of insurers in Colombia.

This project seeks to develop a visualization of data that allows analyzing the sentiment (positive, negative and neutral) of the opinions on Twitter, in addition to presenting the analysis of this data (*Tweets*) and the procedure for research purposes.

Finally, machine learning and natural language processing techniques are proposed applying text-based processing with sentiment analysis.

**Keywords:** Twitter, Sentiment Analysis, Machine Learning, Insurers.



# Tabla de contenido

<b>1. Introducción.....</b>	<b>13</b>
<b>2. Marco teórico.....</b>	<b>15</b>
2.1. Twitter .....	15
2.2. <i>Marketing</i> 4.0 .....	16
2.3. API Twitter .....	16
2.4. Conjunto de datos en Twitter .....	17
2.5. Procesamiento del lenguaje natural.....	18
2.5.1. Componentes del PNL.....	19
2.6. Análisis de sentimientos.....	19
2.6.1. Nube de palabras.....	20
2.7. Compañía aseguradora .....	20
2.8. Programación neurolingüística.....	20
2.8.1. Metodología de la PLN .....	21
<b>3. Estado del arte.....</b>	<b>22</b>
<b>4. Planteamiento del problema.....</b>	<b>29</b>
<b>5. Objetivo general y específicos.....</b>	<b>30</b>
5.1. Objetivo general.....	30
5.2. Objetivos específicos .....	30
<b>6. Metodología .....</b>	<b>31</b>
6.1. Comprensión del negocio .....	31
6.2. Comprensión de los datos .....	31
6.3. Preparación de los datos .....	32
6.4. Construcción de modelado .....	32
6.5. Evaluación .....	32

6.6. Implementación.....	32
<b>7. Desarrollo.....</b>	<b>33</b>
7.1. Arquitectura general.....	33
7.2. API key Twitter.....	34
7.3. Requerimientos de extracción.....	34
7.4. Consulta y limpieza de <i>tweets</i> .....	35
7.4.1. Tokenización.....	37
7.4.2. Normalización.....	37
7.4.3. Remover palabras.....	37
7.4.4. Lematización.....	38
7.5. Desarrollo de visualización.....	39
<b>8. Casos de estudio.....</b>	<b>40</b>
8.1. Seguros Sura.....	41
8.2. Seguros del Estado.....	42
8.3. Liberty Seguros.....	43
8.4. Equidad Seguros.....	44
<b>9. Herramientas utilizadas.....</b>	<b>45</b>
9.1. Lenguaje de programación: Python.....	45
9.1.1. Scikit -learn.....	46
9.1.2. NLTK.....	46
9.2. Colaboratory.....	47
9.3. Anaconda.....	48
9.4. MongoDB.....	49
<b>10. Diseño e implantación del sistema.....</b>	<b>50</b>
10.1. Obtención de datos.....	50
10.2. Almacenamiento y procesamiento de datos.....	51
10.3. Datos.....	53



10.4. Código fuente.....	54
<b>11. Casos de análisis .....</b>	<b>55</b>
11.1. Seguros Sura .....	55
11.2. Seguros del Estado.....	59
11.3. Seguros Liberty .....	63
11.4. Seguros Equidad.....	67
<b>12. Cronograma de trabajo .....</b>	<b>71</b>
<b>13. Presupuesto.....</b>	<b>72</b>
<b>14. Conclusiones.....</b>	<b>73</b>
14.1. Limitaciones del modelo.....	73
<b>15. Referencias bibliográficas.....</b>	<b>74</b>

# Índice de figuras

<b>Ilustración 1:</b> Número de usuarios mensuales activos de Twitter en el mundo (Tomado de [1]).....	13
<b>Ilustración 2:</b> API's de Twitter (Tomado de [2]).....	17
<b>Ilustración 3:</b> Ejemplo de publicaciones en Twitter (Fuente: Elaboración propia, 2021).....	18
<b>Ilustración 4:</b> Ejemplo de dataset en Twitter (Fuente: Elaboración propia, 2021). .....	18
<b>Ilustración 5:</b> Ejemplo nube de palabras (Tomado de [3]).....	20
<b>Ilustración 6:</b> Metodología CRISP-DM ( Tomado de [4]).....	31
<b>Ilustración 7:</b> Diseño de arquitectura para generación de análisis de datos obtenidos desde la API de Twitter (Fuente: Elaboración propia, 2020).....	33
<b>Ilustración 8:</b> Llaves de acceso por la API de Twitter (Fuente: Elaboración propia, 2021).....	34
<b>Ilustración 9:</b> Secuencia extracción de datos (Fuente: Elaboración propia, 2021).. .....	35
<b>Ilustración 10:</b> <i>Script</i> de análisis (Fuente: Elaboración propia, 2021).....	35
<b>Ilustración 11:</b> Cuenta de Twitter (Fuente: Elaboración propia, 2021).....	36
<b>Ilustración 12:</b> Proceso de limpieza de datos (Fuente: Elaboración propia, 2021). .....	36
<b>Ilustración 13:</b> Ambiente usuario (Fuente: Elaboración propia, 2021).....	39
<b>Ilustración 14:</b> Cuenta de Twitter, seguros Sura (Tomado de [44]).....	41
<b>Ilustración 15:</b> Cuenta de Twitter, Seguros del Estado (Tomado de [45]).....	42
<b>Ilustración 16:</b> Cuenta de Twitter, Liberty seguros (Tomado de [46]).....	43
<b>Ilustración 17:</b> Cuenta de Twitter, Equidad seguros (Tomado de [47]).....	44
<b>Ilustración 18:</b> Ventajas y desventajas de Python (Tomado de [5]).....	45
<b>Ilustración 19:</b> Interfaz de anaconda (Fuente: Elaboración propia, 2021).....	48
<b>Ilustración 20:</b> Ejemplo de datos, formato JSON en MongoDB (Fuente: Elaboración propia, 2021).....	49
<b>Ilustración 21:</b> Esquema de almacenamiento de <i>tweets</i> (Fuente: Elaboración propia, 2021) .....	50
<b>Ilustración 22:</b> Configuración localhost para MongoDB (Fuente: Elaboración propia, 2021).....	50
<b>Ilustración 23:</b> Referencia de extracción de <i>tweets</i> (Fuente: Elaboración propia, 2021) .....	51
<b>Ilustración 24:</b> Interfaz MongoDB (Fuente: Elaboración propia, 2021).....	52
<b>Ilustración 25:</b> Colecciones en MongoDB ,alojando <i>tweets</i> (Fuente: Elaboración propia, 2021).....	52
<b>Ilustración 26:</b> Exportación de datos en MongoDB (Fuente: Elaboración propia, 2021).....	53

<b>Ilustración 27:</b> <i>Tweets</i> de Sura (Elaboración propia, 2021).....	55
<b>Ilustración 28:</b> Nube de palabras Sura (Fuente: Elaboración propia, 2021).....	55
<b>Ilustración 29:</b> Frecuencia de palabras, Sura (Fuente: Elaboración propia, 2021). .....	57
<b>Ilustración 30:</b> Conteo de palabras, Sura (Fuente: Elaboración propia, 2021). ...	58
<b>Ilustración 31:</b> <i>Tweets</i> de seguros de Estado (Fuente: Elaboración propia, 2021). .....	59
<b>Ilustración 32:</b> Nube de palabras (Fuente: Elaboración propia, 2021).....	60
<b>Ilustración 33:</b> Frecuencia de palabras, Seguros del Estado (Elaboración propia: Python, 2021). ....	61
<b>Ilustración 34:</b> Conteo de palabras, Seguros del Estado (Fuente: Elaboración propia, 2021). ....	62
<b>Ilustración 35:</b> <i>Tweets</i> de Liberty (Fuente: Elaboración propia, 2021).....	63
<b>Ilustración 36:</b> Nube de palabras (Fuente: Elaboración propia, 2021).....	64
<b>Ilustración 37:</b> Frecuencia de palabras , Liberty (Fuente: Elaboración propia, 2021). ....	65
<b>Ilustración 38:</b> Conteo de palabras Liberty (Fuente: Elaboración propia, 2021)..	66
<b>Ilustración 39:</b> <i>Tweets</i> de Equidad Seguros (Fuente: Elaboración propia, 2021). .	67
<b>Ilustración 40:</b> Nube de palabras (Fuente: Elaboración propia, 2021).....	68
<b>Ilustración 41:</b> Frecuencia de palabras, Equidad (Fuente: Elaboración propia, 2021). ....	69
<b>Ilustración 42:</b> Conteo de palabras Equidad (Fuente: Elaboración propia, 2021). .....	70
<b>Ilustración 43:</b> Cronograma del proyecto (Fuente: Elaboración propia, 2020).....	71
<b>Ilustración 44:</b> Presupuesto de proyecto (Fuente: Elaboración propia, 2020). ....	72

## Índice de tablas

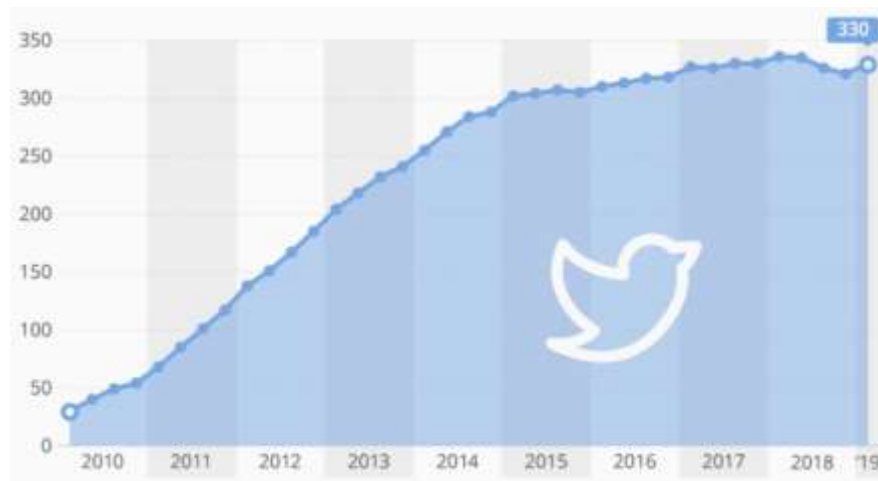
<b>Tabla 1:</b> Características de Twitter (Fuente: Elaboración propia ,Basado en [6])..	16
<b>Tabla 2:</b> Estado del arte (Fuente: Elaboración propia, 2021). .....	25

# 1. Introducción

En una sociedad donde la información llega a generar un gran volumen de datos a nivel global, es importante estar en la capacidad desde almacenar hasta guardar dicha información. Por lo que las aseguradoras deben saber que esta abundancia de información principalmente obtenida en redes sociales puede lograr explicar las percepciones personales de sus clientes y así administrar este tipo de información para intentar ser más competitivas durante los próximos años [7].

Las nuevas tecnologías digitales permiten identificar datos cruciales para entender al cliente, las compañías de seguros podrían entender el comportamiento del cliente en todas sus dimensiones. De hecho, las nuevas tecnologías digitales, permiten acceder a una nueva y valiosa fuente de información. *“Hoy en día, Twitter cuenta con más 340 millones de usuarios, 3,2 millones ubicados en Colombia quienes diariamente almacenan información sobre sus pensamientos, emociones, comportamientos, hábitos y otra información psicológica vital”* [8]. 4.540 millones de personas están ahora en línea, un aumento interanual de 298 millones. Según lo anterior nos acercamos a una penetración de Internet del 60%. Sin embargo, aproximadamente 3.200 millones de personas en todo el mundo permanecen desconectadas.

Con toda esa información se podría incrementar la ejecución de una serie de estudios para permitir a las aseguradoras en Colombia fidelizar a sus clientes y conocer su percepción, además de hacer crecer sus márgenes de ganancia significativamente. Es decir, Twitter está en la capacidad de análisis para toma de decisiones tanto por su cantidad de usuarios como su fuente de datos tal y como se expresa en la ilustración 1.



**Ilustración 1:** Número de usuarios mensuales activos de Twitter en el mundo (Tomado de [1]).

Este crecimiento exponencial de la información ha implicado un gran reto en cuanto a las formas y técnicas que se deben utilizar para almacenar y procesar dicha información de la manera más eficiente, es decir, que implique el menor costo en tiempo y capacidades computacionales posibles [9].

Por lo pronto algunos autores han trabajado sobre aproximaciones frente al análisis de sentimientos [10] ,[11] en Twitter evaluando aseguradoras. A pesar de ello se limitan y poco se ha evaluado sobre la perspectiva de los clientes con las aseguradoras en Colombia. Es así como el debido uso de esos datos tomaría gran importancia en la realización de los planes de éxito para las compañías aseguradoras en Colombia. Por otra parte, el desarrollo de un proceso de más exactitud en cuanto análisis cubriría la necesidad de conocer a los clientes estando alineada con el objetivo de aprovechar los datos de redes sociales, finalmente se resuelve hacer uso de datos estructurados y no estructurados provenientes de la API de Twitter.

El presente proyecto pretende desarrollar un análisis que permita conocer la percepción del sentimiento de los usuarios de cuatro aseguradoras, que posean una cuenta disponible en Twitter. Sin embargo, existe una limitación debido a que cierta cantidad de usuarios no suele usar Twitter, así estén conectados a internet. El análisis abarca el procesamiento de lenguaje natural de las publicaciones escritas en español de la población colombiana. El estudio se limita a realizar una segmentación de las aseguradoras tomando: Sura, Seguros del Estado S.A., Equidad seguros y Liberty.

En este contexto, el siguiente trabajo está organizado de la siguiente manera, una primera parte hablara de la introducción. Donde se proporcionan cuestiones respectivas a Twitter en la actualidad y su impacto como herramienta de análisis para el sector asegurador en Colombia. Otra sección se compone por el marco teórico donde se abarcarán temáticas relacionadas para el desarrollo de este proyecto. También secciones respectivas al estado del arte donde se identifican trabajos con algún grado de similitud a este. Los objetivos son otra de las secciones, adicional de la metodología y el desarrollo que serán vitales para entender el ambiente de desarrollo de este trabajo.

En este trabajo se tocarán temáticas relacionadas con el sector asegurador, incorporando herramientas analíticas para visualización de información. por último se finalizará con los casos de estudio, seguido de los cronogramas, presupuesto y conclusiones que conlleva este trabajo de grado.

## 2. Marco teórico

Con el objetivo de entender globalmente toda la terminología en la cual se realiza el siguiente trabajo, es necesario comprender algunos conceptos. Es así, que se estudiarán algunas definiciones que conllevan a comprender las tecnologías de Twitter y herramientas para realizar los análisis de sentimiento entre otras terminologías usadas para el desarrollo del trabajo.

### 2.1. Twitter

*“Se entiende por Twitter como aquella plataforma de uso online que sirve comúnmente para establecer diferentes estados de usuarios, para poner información o para hacer comentarios sobre diferentes tipos de eventos de una persona en sólo 140 caracteres. Twitter es hoy en día una de las plataformas de comunicación online más populares y utilizadas debido a su facilidad de uso, a su rápido acceso y a la simplicidad de su sistema de registro y utilización” [10].*

Twitter también puede ser definida como una red social de similar tipo que Facebook ya que permite que las personas hablen sobre sus diferentes actividades diarias y que otros puedan verlo y hasta conocerlo en el mismo momento en que se realiza la publicación.

Twitter cuenta con características que la identifican y permiten también describir más de sus funcionalidades como se muestra en la tabla 1.

Características	Descripción
<b>Inmediatez</b>	Twitter se caracteriza por la inmediatez de los contenidos, ya que le da importancia a identificar cuáles son las tendencias del momento.
<b>Sencillez</b>	Twitter desde sus inicios se ha caracterizado por su sencillez, haciéndola muy intuitiva y fácil de manejar.
<b>Red de información</b>	Mientras que otras redes sociales están más orientadas hacia un determinado formato o a un tipo de contenido, en Twitter la protagonista es la información, en todos sus formatos.
<b>Ágil, con contenidos directos y concisos</b>	Gracias a su limitación de espacio (280 caracteres), Twitter es una red social muy ágil, en la que de un simple vistazo se puede ver un contenido sin tener que leer grandes párrafos.

<b>Global</b>	Otra de las grandes ventajas de Twitter es que permite acceder a contenido publicado en todo el mundo. Aunque esto se pueda ver también en otras redes, al ser una red basada en la información, se favorece mucho más este aspecto.
---------------	--

**Tabla 1:** Características de Twitter (Fuente: Elaboración propia ,Basado en [6]).

## 2.2. Marketing 4.0

La economía digital requiere un nuevo enfoque de mercadotecnia para guiar a los especialistas en *marketing* a la hora de anticipar y aprovechar las tecnologías disruptivas por ende surge el *Marketing 4.0*. Un enfoque de *marketing* que combina interacciones en línea y fuera de línea entre empresas y clientes [11].

En la economía digital, la interacción digital llega a quedar corta, de hecho, las tecnologías digitales cada vez más migran a estar en línea, además las marcas cada vez más son flexibles y adaptables debido a las tendencias tecnológicas que surgen en estos tiempos.

## 2.3. API Twitter

En el caso de Twitter, sus API's permiten acceder a leer y escribir datos de Twitter, es decir, a través de ellas se pueden crear *tweets* nuevos y leer el perfil de los usuarios y el dato de sus seguidores (entre otros datos de cada perfil). Ya que se identifican las distintas aplicaciones de Twitter y los usuarios que se registran usando la autenticación y autorización *Open Authorization* (OAUTH).

Las respuestas de la API *rest* de Twitter están en formato *JavaScript Object Notation* (JSON) [12]. Para el ejercicio se usará la API pública de Twitter que cuenta con una API de streaming es decir proporciona un acceso a un alto volumen de *tweets* para llevar a cabo el análisis de sentimientos con dichos datos.

### 2.3.1. Las API's de Twitter

Twitter ofrece tres API's en función a diferentes necesidades, *Streaming API*, *Rest API* y *Search API*. Revisar ilustración 2.

- **Streaming API:** Esta API proporciona un subconjunto de *tweets* prácticamente en tiempo real, mediante una conexión permanente por el usuario con los diferentes



servidores de Twitter y realizando una petición http se recibe un flujo de *tweets* de manera continua, en formato JSON todo el tiempo.

- **Search API:** Esta API permite obtener los *tweets* que se ajustan a una *query* solicitada con una antigüedad de 7 días. Se pueden aplicar filtros por cliente utilizado, lenguaje y localización. Además, esta API también muestra la información más particular del *tweet* que se esté consultando.
- **Rest API:** Es más orientada a los desarrolladores, permite el acceso al núcleo de los datos de Twitter [13]. Además, permite realizar mediante la API las operaciones que se realizan vía web soportando varios formatos como XML, JSON, RSS, ATOM.

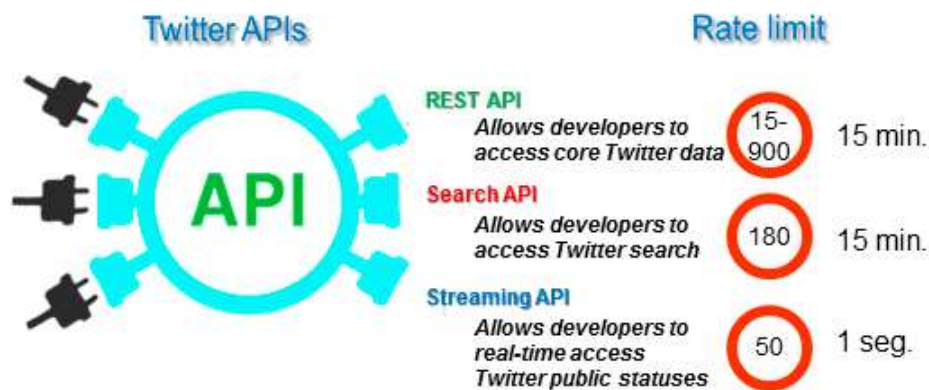


Ilustración 2: API's de Twitter (Tomado de [2]).

## 2.4. Conjunto de datos en Twitter

Los medios sociales son una fuente para poder expresar opinión, estos medios sociales se volvieron una fuente de fácil uso y accesible para poder expresar opinión casi de manera anónima y con una vasta cantidad de información. Surge una extensa data sin analizar con una rica fuente de información para ser explorada y para ser analizada por especialistas en el campo [14]. Específicamente Twitter como medio público de opinión, en el cual se encuentra una gran cantidad de *Tweets* para analizar. Revisar ilustración 3.



**Ilustración 3:** Ejemplo de publicaciones en Twitter (Fuente: Elaboración propia, 2021).

En este trabajo de maestría una vez obtenidos los datos para el proceso de limpieza se cuenta con una forma de etiquetado ya que es necesario que el algoritmo de aprendizaje automático pueda aprender y ajustar sus parámetros para luego poder analizar automáticamente los datos obtenidos por Twitter con sus respectivas variables. Revisar ilustración 4.

	created_at String	id Int64	id_str String	full_text String
1	"Wed Feb 10 23:00:21 +0000 2021"	1359638365478547458	"1359638365478547458"	"Conoce sobre el proceso de reu
2	"Wed Feb 10 22:03:05 +0000 2021"	1359623951798910977	"1359623951798910977"	"RT @77marcong: @SegurosSURA_MX
3	"Wed Feb 10 21:48:45 +0000 2021"	1359620343984439305	"1359620343984439305"	"@aimeooow @GRUPOSURA @Profeco
4	"Wed Feb 10 21:39:28 +0000 2021"	1359618008369487872	"1359618008369487872"	"@Andres210574 @CondusefMX Buen
5	"Wed Feb 10 21:38:43 +0000 2021"	1359617820091375618	"1359617820091375618"	"@soycarmenur Saludos Carmen, u
6	"Wed Feb 10 20:23:02 +0000 2021"	1359598773517905921	"1359598773517905921"	"Ya puedes comprar tu seguro so

**Ilustración 4:** Ejemplo de dataset en Twitter (Fuente: Elaboración propia, 2021)

## 2.5. Procesamiento del lenguaje natural

El procesamiento del lenguaje natural (PLN) es el campo de conocimiento de la inteligencia artificial (IA) que se ocupa de investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino [15].

La lengua humana poco a poco se ve ser tratada de usar por los ordenadores. Lógicamente, existen limitaciones de interés práctico.

Es decir, el PLN se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio del lenguaje natural. El PLN no trata de la comunicación por medio de lenguas naturales de una forma abstracta, sino de diseñar mecanismos para comunicarse que sean eficaces computacionalmente y sobre todo que se puedan realizar por medio de programas que ejecuten o simulen la comunicación.

### 2.5.1. Componentes del PLN

Hay algunos de los componentes del PLN, no todos los análisis que se describen se aplican en cualquier tarea del PLN, sino que depende del objetivo de la aplicación en este caso orientado al proyecto de tesis hablaríamos de:

- **Análisis morfológico o léxico:** Consiste en el análisis interno de las palabras que forman alternativas para extraer lemas, rasgos flexivos, unidades léxicas compuestas.
- **Análisis sintáctico:** Consiste en el análisis de la estructura de las oraciones de acuerdo con el modelo gramatical empleado.
- **Análisis semántico:** Proporciona la interpretación de las oraciones.
- **Análisis pragmático:** Incorpora el análisis del contexto de uso a la interpretación final [16]. Es decir, trata el lenguaje figurado (metáfora e ironía) necesario para entender un texto especializado.

### 2.6. Análisis de sentimientos

El análisis de sentimiento o mejor conocido como minería de opinión se podría definir como el proceso de determinar el tono emocional que hay detrás de una serie de palabras [17]. Este tipo de análisis es comúnmente realizado con la información generada a partir de las redes sociales debido a que la vasta información recopilada que la mayoría de las veces corresponde a opiniones. Esto genera un gran reto que implica almacenar y procesar el gran volumen de información generado, en este caso de Twitter.

El análisis de sentimientos normalmente se usa para intentar comprender las opiniones, actitudes y hasta emociones expresadas en una opinión vía online.



La PNL es a la vez un arte y una ciencia. Se podría definir como un modelo de comunicación que explica el comportamiento humano, puesto que muestra cómo y lo que se hace, es decir, qué estrategias internas se deben seguir.

### **2.8.1. Metodología de la PNL**

La metodología de la PNL es modelar. El modelado consiste en encontrar los componentes esenciales de la conducta que se intenta reproducir [21], para conseguir un resultado equivalente.

El modelado es el proceso que permite recrear comportamientos exitosos. Es un proceso que consta normalmente de dos fases:

- **Fase 1:** Consiste en estudiar detenidamente las actitudes y comportamientos del sujeto a modelar, para averiguar cómo hace y lo que hace de forma óptima.
- **Fase 2:** Consiste en transmitir de forma clara y comprensible las conclusiones extraídas de dicha observación, de modo que otras personas que no hayan participado en la observación sean capaces a partir del modelo creado, de reproducir el comportamiento original que se desea aprender, y obtener unos resultados similares de eficacia.

### 3. Estado del arte

En la actualidad, muchas aseguradoras centran sus esfuerzos en su infraestructura física y servicios prestados, para así satisfacer las demandas y brindar una buena valoración. Por lo tanto, las aseguradoras deberían adoptar soluciones tecnológicas que puedan facilitar el manejo del gran volumen de información en las redes sociales para así poder comprenderlo en su propio beneficio.

En el proceso de investigación del siguiente trabajo se encontraron artículos nacionales e internacionales relacionados con elaboración de análisis de sentimientos mediante el API de Twitter, se identifica la importancia para diferentes campos investigativos. Se puede observar resumidamente a continuación en la tabla 2.

Artículo	Fuente de datos	Herramienta usada	Tema	Ciudad	Año de publicación	Objetivo
[22]	Twitter	Análisis de sentimientos a través de un software	Marcas de empresas en general	Madrid	2013	Necesidad que tienen las empresas en saber, la opinión.
[23]	Twitter	Máquinas de Soporte Vectorial	Opinión en general	Cusco	2020	Implementar una arquitectura para el análisis masivo de datos en Twitter para la identificación de opinión.
[24]	Twitter	Análisis de sentimientos en forma de visualización	Comentarios que estén relacionados con las tic's	Popayán	2019	Aplicar técnicas de análisis de información de la red social Twitter, para identificar tendencias, necesidades, comportamientos en el sector de T.I.
[25]	Twitter	Modelo de aprendizaje	Tweets en particular	Bogotá	2020	Diseñar un modelo de aprendizaje automático que utilice la información

		automático				disponible en Twitter de los seguros de vida .
[26]	Facebook y Twitter	Conocer la percepción de imagen digital	Trabajo cualitativo	Quito	2015	Conocer la percepción de las cuentas en Facebook y Twitter de una ecuatoriana Aseguradora.
[27]	Twitter	Desarrollar un prototipo de una aplicación Web	<i>Tweets</i> en particular	Bogotá	2020	Identificar la polaridad y la actitud de un grupo de usuarios frente a un tema específico.
[28]	Twitter	Conocer la percepción de usuarios	<i>Tweets</i> de opiniones	Jordán	2014	Analizar comentarios o <i>Tweets</i> como positivos, negativos o neutrales sentimientos.
[29]	Twitter	Análisis de sentimientos en forma de visualización	<i>Tweets</i> de opiniones	Burlington	2019	Identificar los sentimientos que los consumidores tienen sobre el seguro médico analizando Twitter.
[30]	Twitter	Modelar con base en percepción de los usuarios	<i>Tweets</i> de opiniones	South África	2020	Modelar la oficina nacional de salud de Sudáfrica basado en usuarios de Twitter y sus interacciones en Twitter.
[31]	Twitter	Probar algoritmo Twitter	<i>Tweets</i> de opiniones	Viña del Mar	2016	Analizar, el desempeño del algoritmo de Adaboost concurrente, en la clasificación de textos de Twitter.
[32]	Twitter	Estudio de la polaridad	<i>Tweets</i> de opiniones	Valencia	2015	Desarrollar una aplicación web basada en Django. agrupando diversas herramientas de clasificación que generen

						estadísticas de polaridad a partir de un conjunto de tuits.
[33]	Twitter	Explorar modelos del aprendizaje profundo aplicado al modelado	<i>Tweets</i> de opiniones	Zaragoza	2018	Analizar el uso de modelos de redes convolucionales (CNN), <i>Long short Term Memory</i> (LSTM), LSTM bidireccionales (BI-LSTM) y una aproximación híbrida entre CNN y LSTM para su uso en el análisis de sentimiento en Twitter.
[34]	Twitter	Clasificación de textos usando métodos de aprendizaje automático	<i>Tweets</i> de opiniones	Puebla	2014	Evaluar el impacto en el uso de diversas características morfológicas sobre la tarea de la detección de carga emocional en <i>Tweets</i> .
[35]	Twitter	Clasificadores de sentimientos que determinan polaridad	<i>Tweets</i> de opiniones	Cauca	2019	Se propone una revisión sistemática de clasificadores de polaridad en análisis de sentimientos, basados en los lineamientos de Kitchenham y el método de bola de nieve.
[36]	Twitter	Minería de opinión para decidir si un texto expresa una opinión	<i>Tweets</i> de opiniones	Málaga	2017	Desarrollar una aplicación web que se conecte con los servicios de la red social Twitter para descargar comentarios en base a criterios de búsqueda.



[37]	Twitter	Técnica basada en minería de datos con el fin de realizar análisis de sentimientos	<i>Tweets</i> de opiniones	Bogotá	2020	Implementar un modelo de minería de datos para clasificar los <i>tweets</i> en base a los sentimientos de los usuarios para conocer su posición con respecto a la JEP, aplicando algoritmos de aprendizaje de máquina al conjunto de datos.
[38]	Twitter	Recursos léxicos de emociones para analizar la polaridad de un conjunto de datos extraídos de Twitter	<i>Tweets</i> de opiniones	Valencia	2017	Análisis del Procesamiento de Lenguaje Natural para determinar la polaridad de un <i>tweet</i> cuando en él aparece lenguaje figurado

**Tabla 2:** Estado del arte (Fuente: Elaboración propia, 2021).

Las redes sociales en los países occidentales se calcula que más del 30% de la población se comunica habitualmente por dichos medios. Según el artículo en el caso de España las redes sociales más populares son Facebook y Twitter. En el artículo [22] se ofrece una funcionalidad de análisis empresarial, se puede contemplar el uso de opiniones que se lleva a cabo en un *software* para el proyecto, que tiene en cuenta un análisis en este caso para las empresas en general y así facilitar la polaridad del contenido.

Con base en el estudio del artículo [23] se recomienda el desarrollo de una tecnología con fines de ayudar a la exploración y percepción libre de opinión con ayuda de máquinas de soporte vectorial, es decir en este caso las opiniones representan el análisis. Sin embargo, carece de un tema específico ya que se analizan *tweets* en general, se puede identificar como utilizan procesos sistematizados para dar con el resultado.

Con el uso de algoritmos se da entrenamiento manual, por medio de aprendizaje automático, así el modelo aprende y finalmente es capaz de predecir tendencias. Para el caso del artículo [24] en cuanto a la visualización de tendencias en este caso tendencias

laborales en tecnología (TI) se intenta poder entender y analizar información, sin embargo se logra proyectar en gráficos tipo histograma y nubes de palabras.

El siguiente estudio tienen como objetivo desarrollar un modelo de clasificación de *tweets* utilizando aprendizaje automático supervisado con redes neuronales artificiales, que se segmentan en categorías de riesgo a los propietarios de una cuenta de Twitter, el artículo [25]. Sin duda alguna es el trabajo que posee mayor similitud al tratarse de seguros, sin embargo, toma la categoría de seguros de vida y para ello analiza variables en base a *tweets* relacionados en práctica de deportes, niveles de estrés, consumo de tabaco y alimentos para poder así obtener el respectivo análisis y su visualización.

A través de la recolección de información a una muestra representativa de usuarios de las redes con encuestas personales en línea, y un complemento utilizando la metodología cualitativa, que consistió en la realización de entrevistas personales. El artículo [26] logra estudiar la percepción de compañías de seguros y con base en lo anterior analiza la imagen digital a través de los usuarios.

Por otra parte, se intenta diseñar y desarrollar un prototipo de una aplicación web que permita exploración y generación de herramientas de análisis escalable de los datos generados en la red social Twitter [27].

Para el artículo [28] se propone explorar una amplia plataforma llena de pensamientos, emociones, reseñas y comentarios. Esto utilizado en muchos aspectos para conocer la opinión, que a su vez permite generar mayor escalabilidad para generar procesos que requieran mayor nivel de procesamiento a futuro.

El análisis de sentimientos estableció qué emociones específicas estaban asociadas con el seguro y proveedores médicos, identificando emociones. El artículo [29] es otro trabajo que utilizó la herramienta analítica para intentar predecir emociones a pesar de las circunstancias para la recolección de datos.

La investigación en el artículo [30] concluye la importancia del análisis para describir la distribución de grados de las relaciones entre los usuarios de Twitter.

En el proyecto del artículo [31] se pretende generar una herramienta que habilite la extracción de opiniones sobre la red social Twitter, aplicadas por ejemplo en las opiniones acerca de los candidatos en las elecciones presidenciales e industria del *retail*, para luego realizar un análisis con el algoritmo de Adaboost.

Otros proyectos han desarrollado un sistema de resumen automático que se basa en la extracción de los *tweet* más representativos con ayuda de sistemas basados en el análisis semántico latente el cual cuenta la popularidad de un *tweet* [32]. Es por ello que este proyecto abarca desde la elaboración de un corpus de *tweets* puntuados manualmente por relevancia, hasta el estudio de las distintas herramientas de Twitter que hacen que un *tweet* se pueda considerar popular o relevante.

En este tipo de artículo se tiene como objetivo la aplicación de técnicas basadas en redes neuronales profundas para dar con una clasificación de la polaridad respecto a los *tweets* [33]. Los resultados que se obtienen de la aplicación de algoritmos de CNN, LSTM y un algoritmo híbrido es combinar ambos algoritmos y así demostrar una mayor métrica de exactitud para analizar los *tweets*.

En el artículo se presentan una serie de experimentos dirigidos al análisis de sentimientos sobre textos escritos en Twitter [34]. En particular, se analizan distintas características morfológicas para la representación de textos con el objetivo de proporcionar el óptimo rendimiento para detectar la carga emocional contenida en los *Tweets*.

La investigación de este artículo se centra en conocer la estructura de los clasificadores y saber qué tipos de algoritmos se utilizan en el proceso de clasificación de *tweets* [35]. Concluyendo usar técnicas de preprocesamiento de datos para normalizar los mensajes antes del proceso de clasificación y así mostrar los resultados de polaridad más precisos con respecto a otros clasificadores de análisis de sentimientos.

Integrar comentarios de redes sociales es el objetivo de este proyecto con ayuda de herramientas de análisis de sentimientos para analizar los comentarios de los usuarios, de manera que se pueda obtener información que refleje opiniones generales sobre algunos temas [36]. Por lo tanto, en este proyecto se desarrolla una aplicación web en la que un usuario puede realizar una búsqueda y análisis de comentarios para su posterior consulta y extracción de estadísticas.

Para este artículo se diseñó una estrategia de minería de datos para analizar los sentimientos de los *tweets* respecto a la jurisdicción especial para la paz (JEP), llevando a cabo una investigación con la cual se busca ordenar y ejecutar una serie de pasos para realizar este proceso de manera óptima y sostenible [37]. Se logra implementar un modelo de minería de datos para clasificar los correspondientes *tweets* en base a los sentimientos de los usuarios para conocer su posición con respecto a la JEP y sus procesos.

En este artículo se lleva a cabo una serie de experimentos para evaluar cómo afectan diferentes recursos léxicos de emociones en el lenguaje figurado y en el lenguaje literal [38]. El artículo apunta a la inclusión de *tweets* para ayudar a clasificar correctamente la polaridad, por ello puede ser de gran importancia desarrollar técnicas capaces de representar la información de manera que sea posible clasificar el sentimiento.

En todos los proyectos expuestos anteriormente se identifica la forma en que con diferentes herramientas analíticas se busca generar un mejor entendimiento de la opinión a través de los *tweets* registrados por los usuarios para diferentes objetivos. Hoy por hoy existe una necesidad de llevar a cabo análisis con la red social Twitter, es así que este proyecto de tesis busca generar una visualización de información, que permita la elaboración de diferentes tipos de análisis generados de Twitter sobre el ámbito asegurador en Colombia.

## 4. Planteamiento del problema

La percepción de los clientes frente a las pólizas de seguros poco se ha estudiado, pese a la situación que actualmente afronta el mundo con la pandemia. En Colombia se cree que suele existir el desconocimiento, los costos y la desconfianza en las aseguradoras siendo las principales razones por las que solo 30 % de las personas y los hogares tienen seguros voluntarios en Colombia y siendo esta la información limitada con la que las aseguradoras normalmente toman sus decisiones [39].

Basado en los estudios anteriores, es necesario hoy por hoy que las compañías aseguradoras se cuestionen del porque un seguro nunca es una primera opción en la mayoría de los casos además de la deserción de sus clientes. La falta de cultura de seguros es histórica en Colombia y en países de Latinoamérica, ya que no es un tema visto como positivo para la comunidad.

Para conocer las opiniones implica el desarrollo de una estrategia enfocada en conocer e identificar las percepciones en cuanto a los sentimientos de los clientes que poseen pólizas de seguros. Dicha estrategia debe estar determinada con la intención de obtener información sobre pensamientos, comportamientos, emociones e incluso información psicológicamente relevante para conocer la percepción con ayuda de *tweets*. Esto finalmente para que el análisis de sentimientos en el desarrollo del siguiente trabajo tenga relevancia a futuro e incluso funcione para toma de decisiones estratégicas en el sector asegurador en Colombia.

La selección de las 4 aseguradoras para el desarrollo de este trabajo, se basó tomando las aseguradoras primeramente que contaran con presencia en el territorio colombiano, que estuvieran en la categoría de seguros generales y además que dispusieran de la red social de Twitter.

## 5. Objetivo general y específicos

### 5.1. Objetivo general

Diseñar y desarrollar una visualización de análisis de sentimientos con el API de Twitter, que permita determinar exploración, visualización y generación de los datos con la finalidad de innovar en las empresas aseguradoras en Colombia mediante comparación a través de la percepción de los *tweets* de sus clientes.

### 5.2. Objetivos específicos

- Realización y configuración de *script* para obtención de datos desde el API de Twitter.
- Identificar los datos incompletos, inexactos, o no pertinentes para el análisis de sentimientos (limpieza de los datos).
- Implementar unas visualizaciones de *tweets* con lenguaje natural con ayuda de herramientas de análisis de sentimientos.
- Determinar la relación entre los comentarios (*tweets*) de los usuarios considerados como clientes por las aseguradoras: Sura, Seguros del Estado S.A., Equidad seguros y Liberty para el primer trimestre del 2021.

## 6. Metodología

La metodología que se utilizará para el siguiente trabajo de maestría será *cross industry standard process for data mining* (CRISP-DM), se contempla en la ilustración 6 ,”se trata de un modelo estándar abierto que proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos” [40]. A continuación, se describe brevemente cada una de las fases asociadas con este trabajo de maestría.

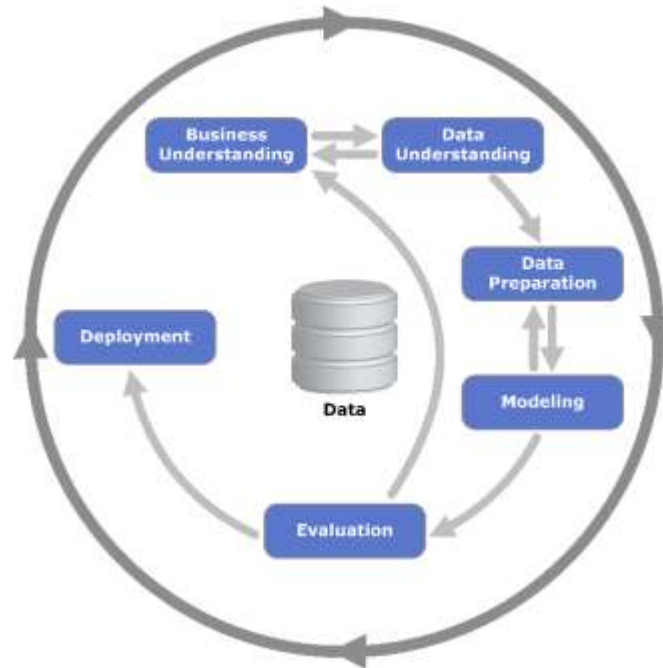


Ilustración 6: Metodología CRISP-DM ( Tomado de [4]).

### 6.1. Comprensión del negocio

En la primera etapa implica conocer y entender los propósitos y requerimientos del sector asegurador en Colombia. Con el objetivo de comprender las oportunidades desde la perspectiva de los clientes además de tener como herramienta diferentes estudios e investigaciones como referencia para este proyecto. Por ello la importancia de intentar extraer la mayor cantidad de opinión respecto a los *tweets*.

### 6.2. Comprensión de los datos

En cuanto a la etapa de comprensión de los datos se parte de que cada mensaje o *tweet* no solo se entiende como un texto sino como un objeto complejo, definido por varios atributos en una estructura JavaScript *Object Notation* (JSON).

### **6.3. preparación de los datos**

En cuanto a la preparación de los datos se procede con herramientas de PNL para así identificar cuáles son los datos incompletos o por consiguientes repetitivos y así tomar diferentes medidas ya sea para eliminar, completar o integrar la información que compete a las aseguradoras en cuestión.

### **6.4. Construcción de modelado**

Para el propósito de la etapa de modelado, se analiza la polaridad de todos los *tweets* por cliente para determinar su relevancia en cuanto a opinión crítica. Un propósito fundamental es poder caracterizar un *tweet* a partir de su relevancia evaluando el servicio de las aseguradoras y así poder escoger los *tweets* más adecuados para llevar a cabo su procesamiento. Para ello, es necesario una aplicación que interactúe con el API de Twitter en este caso Python.

### **6.5. Evaluación**

Una vez obtenido los resultados, se visualizará la información con la ayuda de herramientas de análisis de sentimientos y se comparan las aseguradoras para su posterior análisis.

### **6.6. Implementación**

Ya para la fase final se procederá a la estandarización del proceso de visualización de la información recolectada. Al mismo tiempo se examinará para poner a disposición a las áreas estratégicas de las aseguradoras en cuestión principalmente para la ciudad de Bogotá.



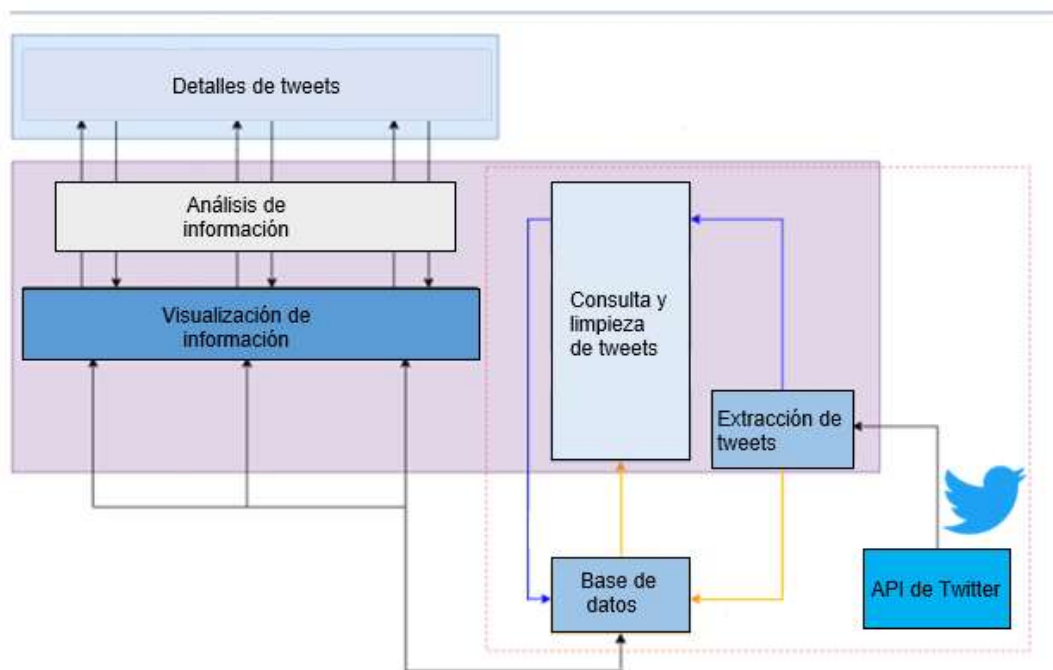
## 7. Desarrollo

Para el desarrollo de este trabajo de tesis se realiza una visualización de datos que permita a los diferentes usuarios o empresas realizar análisis de *tweets* sobre las aseguradoras en cuestión que quieran consultar.

### 7.1. Arquitectura general

La arquitectura utilizada para este proyecto presentada en la ilustración 7, cuenta con dos funcionalidades relevantes. Un proceso de extracción de *tweets* que se basa en conectarse a la API de Twitter que enviará información a la base de datos sobre los *tweets* en tiempo previsto por el cronograma del proyecto, cada vez que se reciba un *tweet* se realizará un análisis de este para quedarse con la información la cual se guardará en una base de datos posteriormente.

En la base de datos se obtienen los *tweets* para entrenar el modelo, una vez el modelo se encuentre en un estado óptimo. El segundo proceso importante será la construcción de la visualización para ser consumida por el cliente, dicha información es procesada y enviada al cliente para su visualización y así poder tomar decisiones estratégicas en cuanto al sector asegurador.



**Ilustración 7:** Diseño de arquitectura para generación de análisis de datos obtenidos desde la API de Twitter  
(Fuente: Elaboración propia, 2020).

## 7.2. API key Twitter

Para obtener los *tweets* se llevó a cabo la creación de una cuenta con permisos de desarrollador en Twitter. Los permisos de desarrollador sobre la cuenta ofrecen unas *Keys* mostradas en la ilustración 8. Con las que haciendo uso de la librería *tweepy* de Python se pueden hacer consultas de los *tweets* asociados.

```
consumer_key = '6vgrPlQ[redacted]zuxV'  
consumer_secret = 'ffAJF3S6fCimOR7rR[redacted]5RJozkhekg3Sc'  
access_token = '940788[redacted]-aKtE7Zkhh3sspyQQ5LCS8mle18F5TX7'  
access_token_secret = 'T2SP0aIxAh20l7Jse[redacted]PXIQUS'
```

**Ilustración 8:** Llaves de acceso por la API de Twitter (Fuente: Elaboración propia, 2021).

la primera funcionalidad se basa en conectarse a la API de Twitter mediante una *key* de esa forma la API enviará información de los datos sobre los *tweets*. Cada vez que se reciba un evento que contenga un *tweet* se deberá realizar un análisis de este para quedarse con la información relevante la cual se guardará en la base de datos de MongoDB posteriormente.

## 7.3. Requerimientos de extracción

Por medio de la ilustración 9, corresponde directamente a diagramas de extracción de datos y secuencias respectivamente donde se definen las funcionalidades que contemplara el desarrollo del proyecto y la manera en que se interactuara con cada uno de los componentes a detalle.



**Ilustración 9:** Secuencia extracción de datos (Fuente: Elaboración propia, 2021).

Tras explicar anteriormente el proceso para conectar Python y Twitter a través del acceso a la API de Twitter, se comienzan a extraer los *tweets* y se proceden a analizar. Para ello se detalla a continuación los paquetes utilizados para la elaboración de un *script* que permite el correcto análisis de la información contemplado en la ilustración 10.

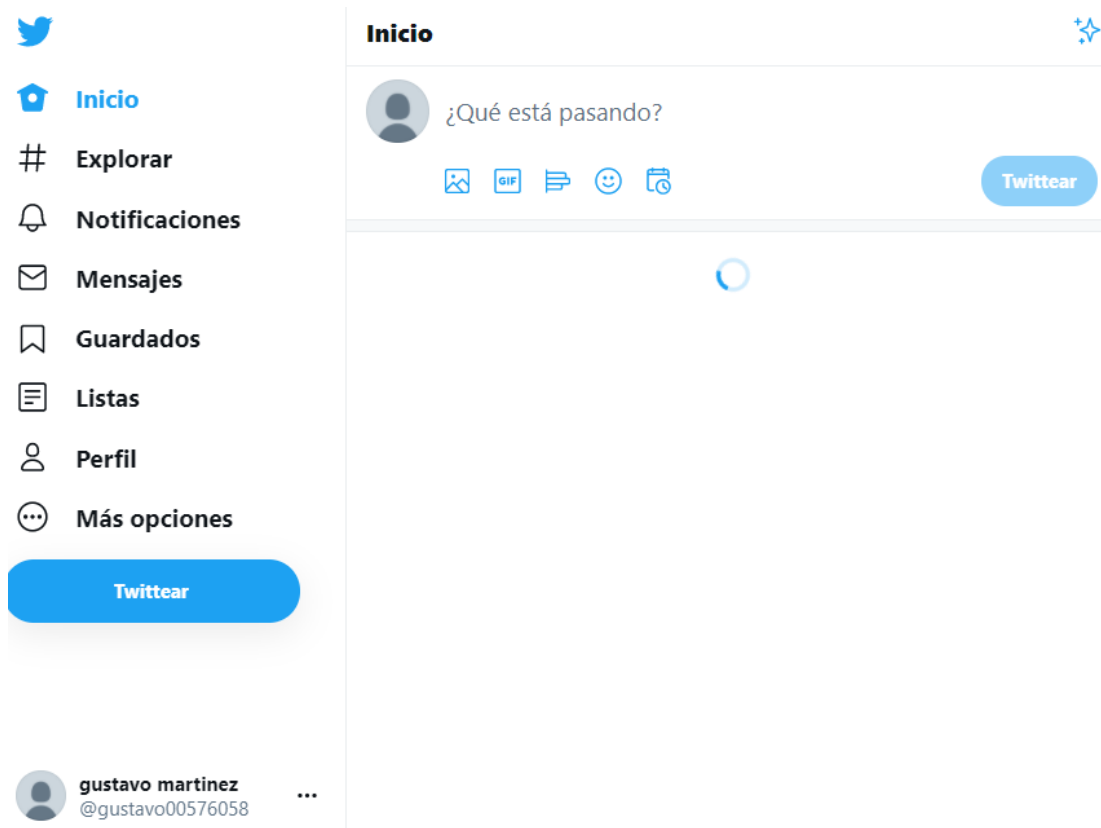
```
import numpy as np
import pandas as pd
from os import path
from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator

import matplotlib.pyplot as plt
% matplotlib inline
```

**Ilustración 10:** *Script* de análisis (Fuente: Elaboración propia, 2021).

## 7.4. Consulta y limpieza de *tweets*

En cuanto a la consulta de los *tweets* es necesario configurar una cuenta de Twitter con permisos de desarrollador. Se creó la cuenta @gustavo00576058 mostrada en la ilustración 11 y a esta cuenta se le solicitó a Twitter permisos de desarrollador para así poder realizar consultas de datos por medio de sus servicios gratuitos prestados.



**Ilustración 11:** Cuenta de Twitter (Fuente: Elaboración propia, 2021).

Para realizar los análisis de los *tweets* se requiere que la información este primeramente extraída para así proceder a realizar procesos de limpieza para obtener los datos que les den más valor a las herramientas de análisis generadas, para los procesos de limpieza de datos resaltados en la ilustración 12.



**Ilustración 12:**Proceso de limpieza de datos (Fuente: Elaboración propia, 2021).

### 7.4.1. Tokenización

Este término consiste en que se dividen las cadenas de texto más largas en piezas más pequeñas o *tokens*. Los trozos de texto más grandes pueden ser convertidos en oraciones, las oraciones pueden ser tokenizadas en palabras. El procesamiento adicional generalmente se realiza después de que una pieza de texto ha sido apropiadamente concatenada [41].

Para el desarrollo de este proyecto de tesis se llevará a cabo:

- Tokenización manual.
- Tokenización y limpieza con *Natural Language Toolkit* (NLTK).

La tokenización suele conocerse como segmentación de texto o análisis léxico. A veces la segmentación se usa para referirse al desglose de un gran trozo de texto en partes más grandes que las palabras. Es decir, se reserva para el proceso de desglose que se produce exclusivamente en el *tweet* extraído.

### 7.4.2. Normalización

Consiste en colocar todas las palabras en las mismas condiciones [27], eliminando puntuaciones, símbolos, números y colocando todas las palabras en mayúsculas o minúsculas.

Respecto a los *tweets* de este trabajo de maestría:

- No parece haber números que requieran manipulación.
- No hay marcadores de sección (por ejemplo, I, II, etc.).
- El uso del guion (-) es para continuar oraciones, según el usuario.
- Hay signos de puntuación como comas, apóstrofes, signos de interrogación.

### 7.4.3. Remover palabras

Respecto al proceso de limpiar texto, muchas veces puede traducir en la necesidad ya sea de dividir un texto en palabras, y hasta llegar a manejar la puntuación en cada *tweet*. Para este proyecto, por consiguiente, es necesario remover las palabras que de alguna manera no agregan valor o que generan ruido a los resultados de los análisis con datos recopilados.

De hecho, hay muchas formas y métodos para remover palabras, pero enfocado en la preparación de textos. Sin embargo, esto depende de cómo se lleve a cabo la función del PNL.

Para este proyecto de maestría lo anterior puede significar:

- Remover los espacios de una cadena de *tweets*.
- Quitar comas e incluso los diferentes signos de puntuación de una cadena de *tweets*.

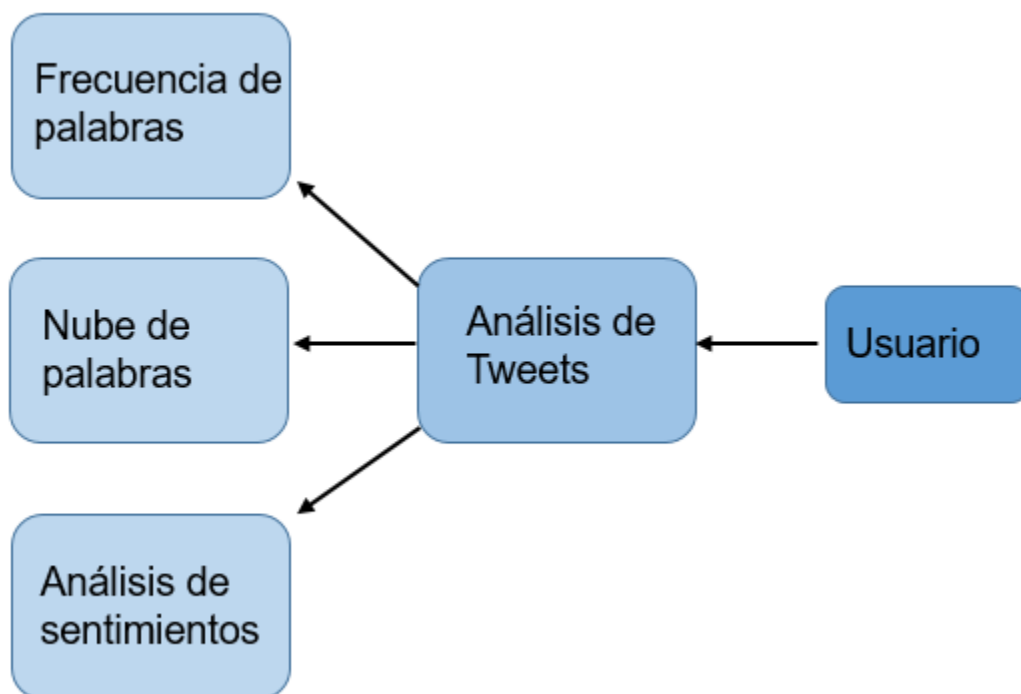
#### **7.4.4. Lematización**

La lematización consiste en un proceso lingüístico , dada una forma flexionada para así hallar el lema correspondiente del texto. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra [42]. Es decir, el lema de una palabra, es la misma palabra que se encontraría como entrada en determinado texto.

La lematización por consiguiente es un proceso clave en muchas tareas prácticas del PLN, suele ser probabilística, lo que hace que en algunos casos se obtengan resultados inesperados.

## 7.5. Desarrollo de visualización

La visualización de datos permite al usuario o entidad aseguradora generar diferentes herramientas de análisis con solo consultar por el nombre de la aseguradora en cuestión. Las herramientas de análisis implementadas para este proyecto se encuentran en la ilustración 13.



**Ilustración 13:** Ambiente usuario (Fuente: Elaboración propia, 2021).

## 8. Casos de estudio

Twitter que está compuesta por una gran cantidad de usuarios que producen información [43]. Sin embargo, para poder desarrollar este proyecto es necesario particionar o agrupar a las aseguradoras en cuestión, con base en 3 características en común:

1. Pertenecer a la categoría de seguros generales, que son aquellos que protegen desde bienes (muebles e inmuebles) contra siniestros, daños, robos, incendios, inundaciones, e incluso desastres naturales.
2. Contar con presencia de sus servicios en el territorio colombiano.
3. Disponer de la red social de Twitter.

A continuación, se hará una breve descripción de los *Tweets* obtenidos mediante las etiquetas:

- @SegurosSURAcól
- @SegurosdelEstado
- @LibertySegCol
- @SegurosEquidad



## 8.1. Seguros Sura

“Es una compañía latinoamericana especializada en la industria de seguros y gestión de tendencias y riesgos, reconocida por su experiencia de más de 75 años en el mercado. El grupo Sura es un grupo empresarial y financiero latinoamericano con foco estratégico en los sectores seguros, pensiones, ahorro, inversión, gestión de activos y banca, en los que cuentan con más de 50 millones de clientes, cuenta con más de 57 mil empleados en 11 países de la región” [44].



Ilustración 14: Cuenta de Twitter, seguros Sura (Tomado de [45]).

Grupo Sura es la aseguradora que cuenta con más seguidores, de hecho, este factor se debe ya que su presencia en el mercado es mayor que las demás y por lo tanto posee un tanto más posicionamiento que su competencia. Sin embargo, cabe aclarar que no será un factor determinante para el análisis ya que las aseguradoras en cuestión cuentan con un número limitado de tweets de acuerdo con este proyecto de grado.

## 8.2. Seguros del Estado

*“Compañía de seguros, donde predomina la constante innovación de productos y servicios, atienden las necesidades de los usuarios, han estado vigentes durante 60 años en el sector asegurador. Se consolidan como una de las principales aseguradoras de Colombia” [46].*

Tienen como objetivo proteger el patrimonio de los colombianos. por medio de productos diseñados acorde a los riesgos que puedan afectarlo, cuentan con el respaldo de los reaseguradores multinacionales de primera línea. Cuentan con amplia cobertura y servicio en todo el territorio nacional, con presencia en 23 ciudades. Y una red comercial de más 4.000 agentes independientes, agencias y corredores.



**Ilustración 15:** Cuenta de Twitter, Seguros del Estado (Tomado de [47]).

### 8.3. Liberty seguros

*“En Liberty seguros creen que el compromiso con los clientes es importante y el impacto en las comunidades. Cada año movilizan en todo el mundo más de 50 millones de dólares de inversión a la comunidad que se ven reflejados en tiempo, dinero y recursos. Están en 30 países del mundo además cuentan con 900 oficinas para facilitarle la vida a los usuarios” [48].*



Ilustración 16: Cuenta de Twitter, Liberty seguros (Tomado de [49]).

## 8.4. Equidad seguros

*“Es una aseguradora cooperativa que brinda servicios de protección a las personas, sus familias, bienes y empresas. Conocida anteriormente como Seguros La Equidad es una empresa colombiana, administradora de seguros y riesgos profesionales. Su sede principal, así como su mercado, se ubican principalmente en Bogotá” [50].*



**Ilustración 17:** Cuenta de Twitter, Equidad seguros (Tomado de [51]).

## 9. Herramientas utilizadas

Se mostrará una descripción de las herramientas tecnológicas y de los lenguajes de programación que han sido utilizados para llevar a cabo este proyecto de grado.

### 9.1. Lenguaje de programación: Python

Python es un lenguaje de *scripting* independiente de plataforma y orientado a objetos, preparado para realizar cualquier tipo de programa, desde aplicaciones Windows a servidores de red o incluso, páginas web [52]. Es un lenguaje interpretado, lo que significa que no se necesita compilar el código fuente para poder ser ejecutado, lo que ofrece ventajas como la rapidez de desarrollo con una menor velocidad. A continuación, en la ilustración 18 se presentan sus ventajas y desventajas.

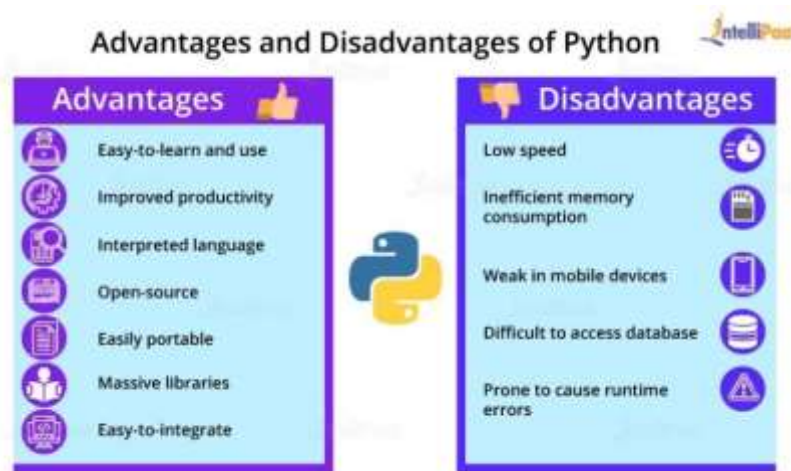


Ilustración 18: Ventajas y desventajas de Python (Tomado de [5]).

Python es el lenguaje es ideal para este proyecto por:

- La cantidad de librerías que contiene, tipos de datos y todas sus funciones incorporadas en el propio lenguaje, que ayudan a realizar diferentes tareas sin necesidad de tener que programarlas desde cero.
- La sencillez y velocidad con la que se crean los programas. Un programa en Python puede tener de 3 a 5 líneas de código.
- La cantidad de plataformas en las que se puede desarrollar, como Unix, Windows, etc.
- Python es gratuito, incluso para propósitos empresariales.

### 9.1.1. Scikit-learn

Scikit-Learn es una librería gratuita para Python. Cuenta con algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad [53]. Además, tiene compatibilidad con otras librerías de Python como NumPy, SciPy y Matplotlib.

La gran variedad de algoritmos y utilidades de Scikit-learn hace que esta herramienta este en la capacidad de estructurar los sistemas de análisis datos y modelado estadístico. Los algoritmos de Scikit-Learn se combinan y depuran con otras estructuras de datos y aplicaciones externas como Pandas o PyBrain necesarias para el desarrollo de este proyecto.

La funcionalidad que proporciona Scikit-learn incluye:

- Regresión, incluida la regresión lineal y logística.
- Clasificación, incluidos los vecinos K más cercanos.
- Agrupación, incluidos *K-Means*.
- Selección de modelo.

### 9.1.2. NLTK

El kit de herramientas de lenguaje natural, o más comúnmente NLTK, es un conjunto de bibliotecas y programas para PLN simbólico y estadísticos para el lenguaje de programación Python. NLTK incluye demostraciones gráficas y datos de muestra [54]. NLTK está destinado a apoyar la investigación y la enseñanza en PNL o áreas muy relacionadas, que incluyen la lingüística empírica, las ciencias cognitivas, la inteligencia artificial, la recuperación de información, y el aprendizaje de las máquinas.

### 9.1.3. Numpy y pandas

Pandas es un paquete de Python que proporciona estructuras de datos similares a los dataframes de R. Pandas depende de Numpy, la librería que añade un potente tipo matricial a Python [55]. Los principales tipos de datos que pueden representarse con pandas son:

- Datos tabulares con columnas de tipo heterogéneo con etiquetas en columnas y filas.
- Series temporales.

Pandas es ideal para este proyecto ya que proporciona herramientas que permiten:

- Leer y escribir datos en diferentes formatos: CSV, Microsoft Excel, bases SQL.
- Seleccionar y filtrar de manera sencilla tablas de datos en función de posición, valor o etiquetas.
- Fusionar y unir datos.
- Transformar datos aplicando funciones tanto en global como por ventanas.
- Manipulación de series temporales.
- Hacer gráficas.

## 9.2. Colaboratory

Colaboratory o también llamado Colab, permite escribir y ejecutar código de Python en un navegador [56]. Es ideal para este proyecto ya que cuenta con las siguientes particularidades:

- Funciona sin configuración requerida.
- Posee acceso gratuito a GPU.
- Tiene facilidad para compartir código.

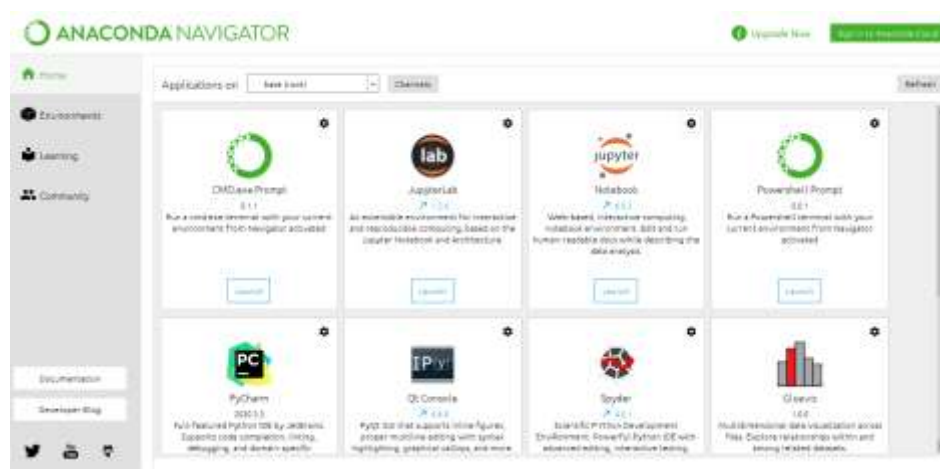
Colab funciona sobre todo en este proyecto ya que, si se quiere mostrar o hacer un modo de presentación de su código en Python se puede ejecutar y mostrar los resultados de cada paso de desarrollo. Esto permite mejorar el código en Python y así tener la facilidad de incorporar todas las librerías para trabajar con aprendizaje automático y también todas las librerías disponibles para Python.



### 9.3. Anaconda

Anaconda es una distribución libre y abierta de los lenguajes Python y R, utilizada en ciencia de datos, y aprendizaje automático [57]. Esto incluye procesamiento de grandes volúmenes de información, análisis predictivo y cómputos científicos.

Está orientado a simplificar el despliegue y administración de los paquetes de software. La distribución de Anaconda es usada por más de 13 millones de usuarios e incluye más de 1.400 paquetes populares de ciencia de datos adecuados para Windows, Linux y MacOS.



**Ilustración 19:** Interfaz de anaconda (Fuente: Elaboración propia, 2021).

Para el desarrollo de este proyecto se destaca por:

- Permite trabajar localmente en el desarrollo de este trabajo.
- Funciona con lenguaje Python.
- Es un software gratuito.



## 9.4. MongoDB

MongoDB es una base de datos orientada a documentos. Esto quiere decir que, en lugar de guardar los datos en registros, guarda los datos en documentos. Estos documentos son almacenados en JSON revisar ilustración 20, que es una representación binaria de JSON.

```
{
  Nombre: "Pedro",
  Apellidos: "Martínez Campo",
  Edad: 22,
  Aficiones: ["fútbol", "tenis", "ciclismo"],
  Amigos: [
    {
      Nombre: "María",
      Edad: 22
    },
    {
      Nombre: "Luis",
      Edad: 28
    }
  ]
}
```

**Ilustración 20:** Ejemplo de datos, formato JSON en MongoDB (Fuente: Elaboración propia, 2021).

### Características:

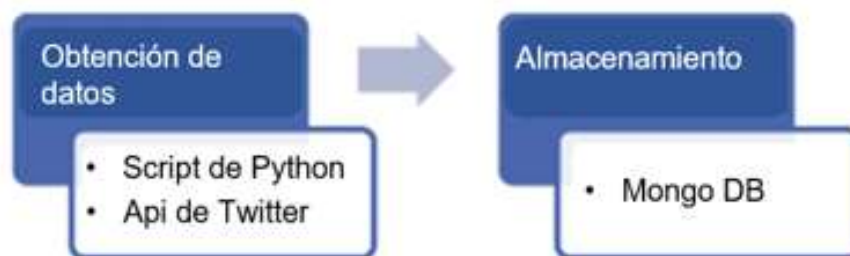
- MongoDB viene con una consola donde se pueden ejecutar los distintos comandos.
- Además de las funciones de MongoDB, se pueden utilizar muchas de las funciones propias de JavaScript.
- Cualquier aplicación que necesite almacenar datos semiestructurados puede usar MongoDB.

Para el desarrollo de este proyecto Mongo DB destaca por:

Una de las diferencias más importantes con respecto a las bases de datos relacionales, es que no es necesario seguir un esquema [58]. Además, los documentos pueden pertenecer a una misma colección, concepto similar a una tabla de una base de datos relacional, lo cual facilita de una u otra forma la extracción de *tweets* desde el API de Twitter.

## 10. Diseño e implantación del sistema

Una vez obtenida la información necesaria, es decir el conjunto de *tweets* de las opiniones de las aseguradoras en cuestión, es necesario una herramienta para procesar y guardar esa información con el fin de analizarla lo mejor posible. Para esto se utiliza la herramienta MongoDB, una novedosa base de datos no SQL que destaca por sus soluciones que demandan un rápido acceso a grandes cantidades de datos. Revisar ilustración 21.



**Ilustración 21:** Esquema de almacenamiento de *tweets* (Fuente: Elaboración propia, 2021)

Respecto a MongoDB como plataforma de administración para alojar los *tweets*, se conecta mediante un usuario con roles definidos específicamente para esta operación. Además de contar con la información del localhost de la maquina local como se muestra en la ilustración 22. Internamente, la base de datos cuenta con cuatro colecciones destinados a los *tweets* guardados.



**Ilustración 22:** Configuración localhost para MongoDB (Fuente: Elaboración propia, 2021).

### 10.1. Obtención de datos

Para este proyecto el análisis será sobre la red social Twitter, así obteniendo un conjunto de *tweets* correspondientes a comentarios u opiniones de las aseguradoras en cuestión. Sobra decir, que en un rango específico para su posterior análisis. Para utilizar la API de

Twitter se hace de manera sencilla ya es necesario tener una cuenta de Twitter creada, la cual se usará para obtener una serie de parámetros necesarios para utilizar la API de Twitter.

La información que se extraerá es de carácter público como son los *tweets* generados por los usuarios de las aseguradoras. Respecto a limitaciones no han supuesto grandes inconvenientes a la hora de desarrollar este proyecto. Ya que a la hora de recuperar los *tweets* estos hacen referencia normalmente a una etiqueta del que se esté twitteando, por lo que, si se quieren *tweets* de algún tema del pasado o del que no haya muchos usuarios hablando, el *script* producirá un error.

Respecto al *script* utilizado mostrado en la ilustración 23. Este obtendrá un conjunto de *tweets* con varias variables que serán extraídas en formato JSON a una carpeta de destino especificada. Como parámetros obligatorios se debe introducir la etiqueta que contendrá todos los *tweets* que serán extraídos.

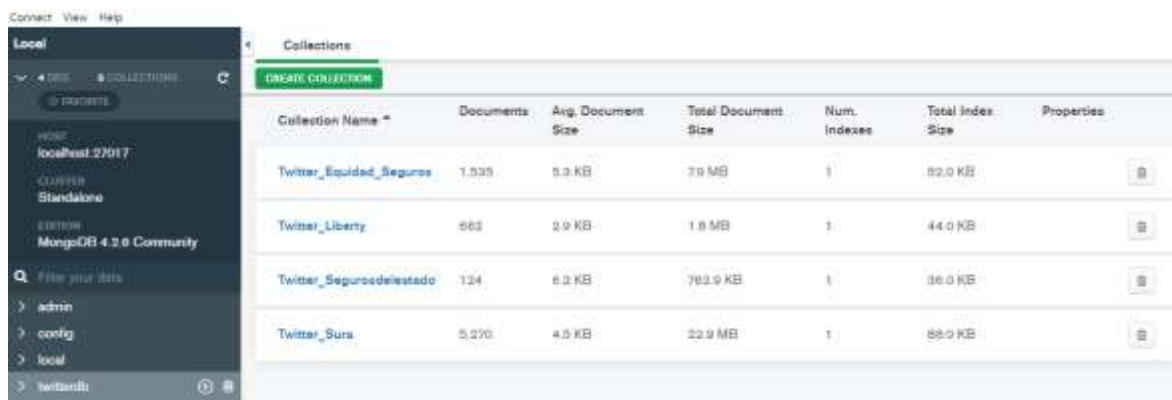
```
#for tweet in tweepy.Cursor(api.search,q='#SegurosEquidad',since='2021-01-01',until='2021-02-25').items(1000):
api = tweepy.API(auth)
for tweet in tweepy.Cursor(api.search, q='@SegurosEquidad', tweet_mode='extended',lang="es",since="2021-01-01").items(1000):
    client = MongoClient('localhost', 27017)
    db = client['twitterdb']
    collection = db['Twitter_Equidad_Seguros']
    tweet_json = json.loads(json.dumps(tweet._json))
    collection.insert(tweet_json)
```

**Ilustración 23:** Referencia de extracción de *tweets* (Fuente: Elaboración propia, 2021).

Como resultado se obtendrá un archivo en formato JSON con el conjunto de *tweets* recuperados para determinada etiqueta alusivo a la aseguradora en cuestión.

## 10.2. Almacenamiento y procesamiento de datos

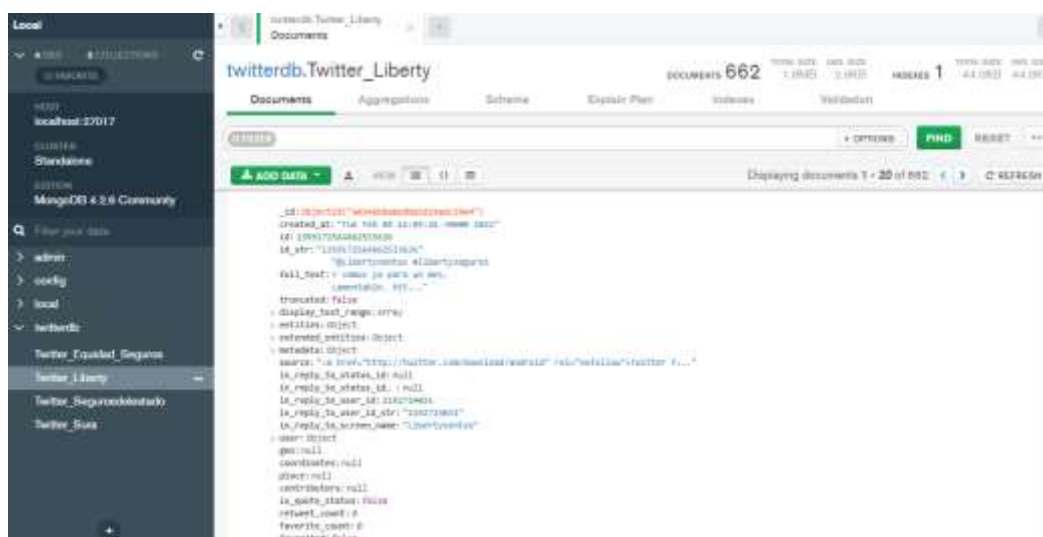
Obtenidos los datos que se van a analizar, se guardaran en una base de datos utilizando la herramienta MongoDB visualizada en la ilustración 24. La cual permite crear colecciones de datos. Se crean colecciones que almacenen todos los *tweets* que se han recuperado de las etiquetas en una base de datos local llamada twitterdb que contendrá colecciones las cuales alojaran los *tweets* dependiendo la aseguradora.



Collection Name	Documents	Avg. Document Size	Total Document Size	Num. Indexes	Total Index Size	Properties
Twitter_Equidad_Seguros	1,535	5.3 KB	7.9 MB	1	82.0 KB	
Twitter_Liberty	662	2.9 KB	1.8 MB	1	44.0 KB	
Twitter_SegurodelEstado	134	6.2 KB	762.9 KB	1	36.0 KB	
Twitter_Sura	5,270	4.5 KB	22.9 MB	1	88.0 KB	

**Ilustración 24:** Interfaz MongoDB (Fuente: Elaboración propia, 2021).

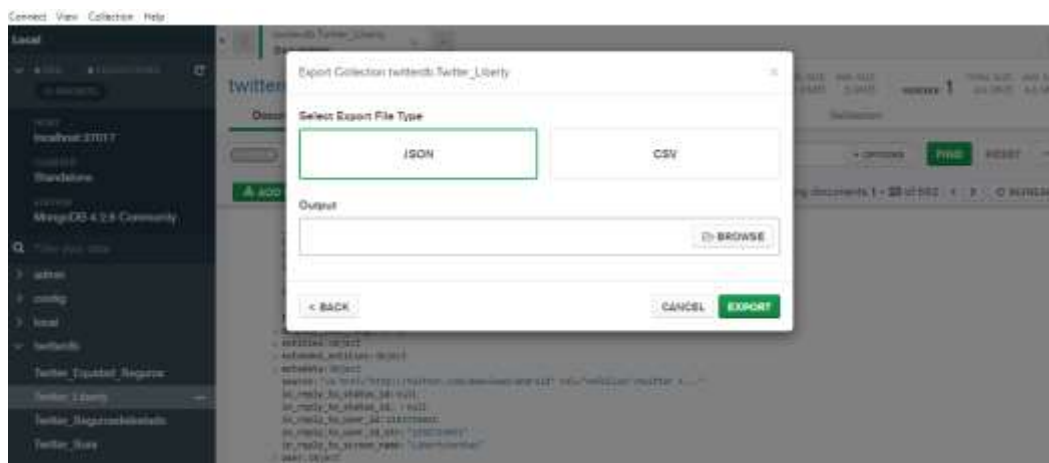
En cada una de las colecciones en la que se van a insertar los documentos, también se crea mediante la interfaz MongoDB, en la pestaña de *collections* pulsando *add collection* y añadiendo sus respectivos nombres para diferenciar las aseguradoras y que correctamente estén alojados sus respectivos *tweets*, se puede apreciar en la ilustración 25.



Collection Name	Documents	Avg. Document Size	Total Document Size	Num. Indexes	Total Index Size	Properties
Twitter_Equidad_Seguros	1,535	5.3 KB	7.9 MB	1	82.0 KB	
Twitter_Liberty	662	2.9 KB	1.8 MB	1	44.0 KB	
Twitter_SegurodelEstado	134	6.2 KB	762.9 KB	1	36.0 KB	
Twitter_Sura	5,270	4.5 KB	22.9 MB	1	88.0 KB	

**Ilustración 25:** Colecciones en MongoDB ,alojando *tweets* (Fuente: Elaboración propia, 2021).

A continuación, se puede apreciar como es el proceso para exportar una colección, para ello se necesita utilizar la opción de exportar que permite exportar la gran cantidad de datos en un archivo JSON o CSV, contemplado en la ilustración 26.



**Ilustración 26:** Exportación de datos en MongoDB (Fuente: Elaboración propia, 2021).

**Nota:** Para el desarrollo de este trabajo se extraen en formato CSV por temas de análisis de los *tweets*.

### 10.3. Datos

Para este trabajo de tesis se obtuvieron los siguientes datos tomados desde el 01 de enero del 2021 hasta 25 de febrero del 2021 teniendo en cuenta la presentación de este proyecto de tesis. Además, se logran recolectar aproximadamente 15.353 *tweets*. Todos estos *tweets* pertenecen a las debidas aseguradoras en cuestión con su respectivo *hashtag* mostrados a continuación.

- 10.204 *Tweets* @SegurosSURAcól
- 936 *Tweets* @SegurosdelEstado
- 1.271 *Tweets* @LibertySegCol
- 2.942 *Tweets* @SegurosEquidad

## 10.4. Código fuente

En el siguiente link se encuentra el código fuente para el desarrollo de este proyecto de tesis junto con las indicaciones para la instalación de las librerías y su debida configuración.

<https://github.com/gutak12345/Gustavo-Adolfo-Martinez-Misal/blob/master/CODIGO%20TRABAJO%20DE%20MAESTRIA%20-%20SEGUROS>

## 11. Casos de análisis

Se realizan cuatro casos de estudio para analizar los resultados obtenidos (*tweets*) de las aseguradoras en cuestión. En el proceso de implementación de este proyecto de grado fue necesario la construcción del dataset y finalmente la limpieza de los datos. Teniendo en cuenta lo anterior se puede concluir que los resultados que se obtuvieron y en el tiempo en que se llevó a cabo son satisfactorios. Los resultados obtenidos fueron los siguientes.

### 11.1. Seguros Sura

En la ilustración 27 se pueden observar los resultados obtenidos de la consulta de la aseguradora Sura junto con su respectivo *script*, de igual manera se puede ver el enunciado del *tweet* para el posterior análisis de sentimientos de cada uno de los mensajes generados.

```
#Read data from Google Sheet: Asguradoras
df = pd.read_csv("/content/Twitter_Sura.csv")

df.columns
#df.head()
df['full_text']
#print(df.shape)
```

```
0      Conoce sobre el proceso de reutilización de re...
1      RT @77marcong: @SegurosSURA_MX Buen día, llevo...
2      @aimeoooow @GRUPOSURA @Profeco Buenas tardes, A...
3      @Andres210574 @CondusefMX Buenas tardes, André...
4      @soycarmenur Saludos Carmen, un placer ayudart...
...
10199  @SegurosSURAcól Les recomiendo que envíen a qu...
10200  RT @SegurosSURAcól: ¿La vacuna sí me protege p...
10201  RT @SegurosSURAcól: ¿La vacuna sí me protege p...
10202  RT @SegurosSURAcól: ¿La vacuna sí me protege p...
10203  RT @SegurosSURAcól: ¿La vacuna sí me protege p...
Name: full_text, Length: 10204, dtype: object
```

**Ilustración 27:** *Tweets* de Sura (Elaboración propia, 2021).

Por otra parte, para el caso de la aseguradora Sura, el conteo de apariciones de las palabras también nos permite realizar una nube de palabras que es una representación gráfica de las palabras clave más recurrentes mostradas en la ilustración 28.



**Ilustración 28:** Nube de palabras, Sura (Fuente: Elaboración propia, 2021).

En este trabajo de tesis como se ha comentado en el apartado de la metodología, una nube de palabras es una representación visual de la frecuencia de ciertas palabras extraídas de un texto.

En este tipo de gráfico se puede observar como las palabras más usadas o recurrentes tienen un tamaño de letras más grande y colores más llamativos visualmente, que permiten identificar más rápidamente el contenido como se muestra en la ilustración 28. Además, ayudan a visualizar aquellas palabras que poseen cierto peso y con las que posiblemente se extraerán algunas conclusiones respecto a esta aseguradora. Combinada con el análisis de emociones realizado sobre los datos, proporciona algo más intuitivo para conocer lo que los usuarios sienten cuando tuitean sobre el tema de las aseguradoras.

La nube correspondiente a la aseguradora Sura en la ilustración 28, permite apreciar las palabras que fueron clasificadas con emociones pertenecientes al conjunto de *tweets* que contenían su búsqueda. El tamaño de las palabras crece según lo hace su frecuencia en los datos. Se pueden obtener algunas conclusiones observando dichas palabras:

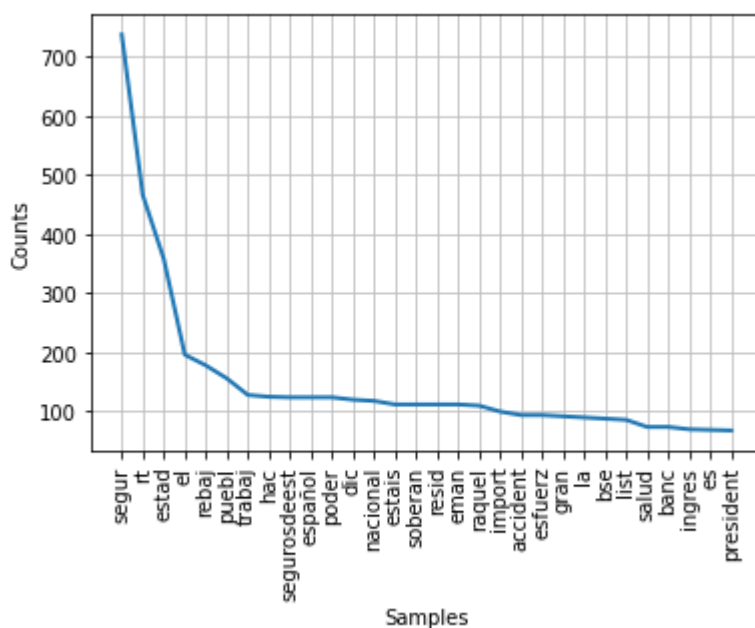


## Categorías de sentimiento

En la categoría negativa se puede observar como las principales palabras relacionadas a «intentando comunicarme», «llevo intentando», «no he podido» o «malas hacer». Podría ser un reflejo de la imagen negativa existente hacia la aseguradora Sura, en la cual varios usuarios indican la mala disponibilidad de la aseguradora.

En la categoría positiva, destacan palabras como «protege para», «he podido», o «eficaces». El uso de estas palabras nos permite identificar que algunos usuarios hablan de forma positiva de la aseguradora Sura, sin embargo es muy poco lo cual demuestra que Sura no tiene buen posicionamiento positivo por lo usuarios y debería mejorar la relacion con sus usuarios.

Por último, en la categoría neutral se puede leer «vacuna», «contra covid» o «verdadero vacunas». El uso de estas palabras dan a demostrar que esta aseguradora esta a la vanguardia en temas de vacunacion contra el covid 19 ya que es lo que mas resalta respecto a lo que tuitean los usuarios.



**Ilustración 29:** Frecuencia de palabras, Sura (Fuente: Elaboración propia, 2021).

En cuanto al proceso de conteo de palabras, se identifican cuáles son las palabras que son más frecuentes y que tiene relación con la aseguradora Sura en cuanto al resultado de *tweets* obtenidos.

Con esta información se genera un diagrama de series que muestra cuales son las palabras que tienen más frecuencia en la información obtenida como se puede ver en la ilustración 29. Además, estos gráficos ayudan a comprender mejor las nubes de palabras

representadas anteriormente con la finalidad de dar una idea de las veces que aparece un término específico en la muestra.

bse:88	
list:86	segur:737
salud:74	rt:466
banc:74	estad:358
ingres:70	el:196
es:69	rebaj:178
president:68	puebl:156
si:68	trabaj:128
encuentr:62	hac:125
fuerz:60	segurosdeest:124
te:55	español:124
invit:55	poder:124
garantiz:50	dic:120
equip:50	nacional:118
quier:49	estais:112
pag:49	soberan:112
consult:48	resid:112
las:46	eman:112
aqu:44	raquel:110
riesg:44	import:100
asum:42	accident:94
aut:40	esfuerz:94
guair:40	gran:92
virtual:38	la:90
en:38	
person:38	
cad:38	

**Ilustración 30:** Conteo de palabras, Sura (Fuente: Elaboración propia, 2021).

Por otra parte, el conteo de apariciones de las palabras mostrado en la ilustración 30, permite realizar una idea de cuáles fueron las palabras clave más recurrentes respecto a la aseguradora Sura.

## 11.2. Seguros del Estado

Se realiza la consulta para la aseguradora Seguros del Estado, se obtuvieron 936 registros para analizar cargados a Python como se observa en la ilustración 31.

```
#Read data from Google Sheet: Asguradoras
df = pd.read_csv("/content/Twitter_Segurosdelestado.csv")

df.columns
#df.head()
df['full_text']
#print(df.shape)

0      RT @SegurosdeEstado: Regresa a clases presenci...
1      RT @SegurosdeEstado: Regresa a clases presenci...
2      RT @SegurosdeEstado: Regresa a clases presenci...
3      RT @SegurosdeEstado: Te invitamos a participar...
4      RT @SegurosdeEstado: Te invitamos a participar...
...
931    2- Muchos jefes de sistemas o encargados de si...
932    @_hugo_drax_ @RafaHistorel @elindepcom Somos d...
933    RT @FrentePreveN: Frente Preventivo Guárico, p...
934    Algunas personas odian estar en el volante por...
935    @AlexiaRivasG1 @feliperaytyson @nayibbukele Pa...
Name: full_text, Length: 936, dtype: object
```

**Ilustración 31:** *Tweets* de Seguros de Estado (Fuente: Fuente: Elaboración propia, 2021).

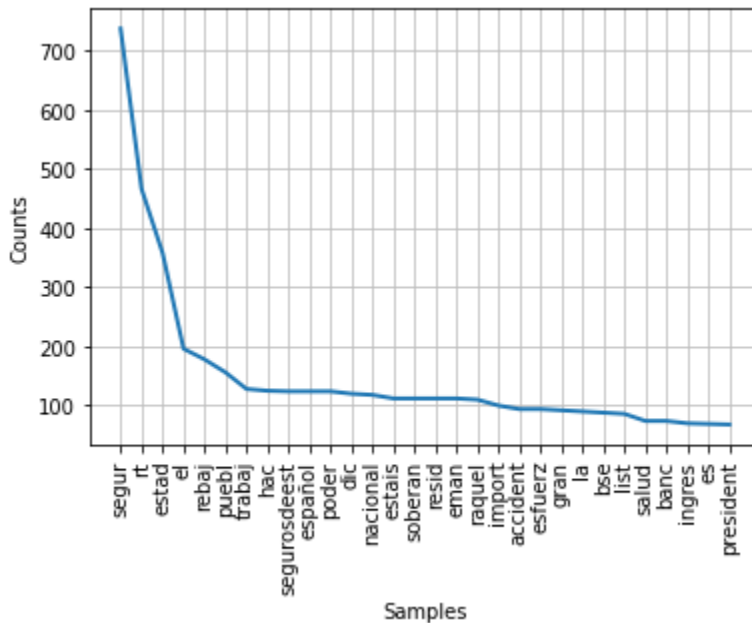
A continuación, se muestra en la ilustración 31 los resultados obtenidos de la consulta de la aseguradora Seguros del Estado, de igual manera se puede ver el enunciado de algunos *tweets* para el posterior análisis de sentimientos de cada uno de los mensajes generados.



**Ilustración 32:** Nube de palabras, Seguros del Estado (Fuente: Elaboración propia, 2021).

La nube correspondiente a la aseguradora Seguros del Estado mostrada en la ilustración 32. Se aprecian las palabras que fueron clasificadas con emociones pertenecientes al conjunto de *tweets* que contenían su búsqueda. El tamaño de las palabras crece según lo hace su frecuencia en los datos, se pueden obtener algunas conclusiones observando dichas palabras.

Una de las limitaciones de la aseguradora seguros del Estado con este trabajo es el uso de los *tweets* por los usuarios basados en lexicones, ya que al otorgar puntuaciones normalmente se obtienen más resultados “neutrales”. Esto ocurre ya que este análisis no es basado en aprendizaje automático del todo (modelos), sino que se respalda con análisis manual (observación) para el análisis de este ejercicio. Esto es debido a la forma de actuar de este proyecto y al hecho de que los *tweets* se limitan a una serie de palabras o frases cualquiera. Por lo tanto, no cuenta con carga sentimental para el análisis, aunque realmente la tenga.



**Ilustración 33:** Frecuencia de palabras, Seguros del Estado (Fuente: Elaboración propia, 2021).

Respecto al conteo de palabras se logran identificar cuáles son las palabras que son más frecuentes y que tiene relación con la aseguradora Seguros de Estado en el resultado de *tweets* obtenidos. Una vez realizado el proceso de limpieza de los *tweets* y con las palabras ya transformadas en *tokens*, se utiliza para este proceso la librería de Python Counter la cual muestra el resultado dependiendo del número de apariciones de cada una de las palabras en los *tweets* obtenidos.

Con esta información se genera un diagrama de series que muestra cuales son las palabras que tienen más frecuencia en la información obtenida como se puede ver en la ilustración 33.

segur:737	banc:74
rt:466	ingres:70
estad:358	es:69
el:196	president:68
rebaj:178	si:68
puebl:156	encuentr:62
trabaj:128	fuerz:60
hac:125	te:55
segurosdeest:124	invit:55
español:124	garantiz:50
poder:124	equip:50
dic:120	quier:49
nacional:118	pag:49
estais:112	consult:48
soberan:112	las:46
resid:112	aqu:44
eman:112	riesg:44
raquel:110	asum:42
import:100	aut:40
accident:94	guair:40
esfuerz:94	virtual:38
gran:92	en:38
la:90	person:38
bse:88	cad:38
list:86	
salud:74	

**Ilustración 34:** Conteo de palabras, Seguros del Estado (Fuente: Elaboración propia, 2021).

Por otra parte, el conteo de apariciones de las palabras mostrado en la ilustración 34, permite realizar una idea de cuáles fueron las palabras clave más recurrentes respecto este caso la aseguradora Seguros del Estado.

### 11.3. Seguros Liberty

Se realiza la consulta de la aseguradora Liberty, se obtuvieron 1.271 registros para analizar.

```
#Read data from Google Sheet: Asguradoras
df = pd.read_csv("Twitter_Liberty.csv")

df.columns
#df.head()
df['full_text']
#print(df.shape)
```

```
0      @LibertyVentas #libertyseguros\nY vamos ya par...
1      Aprovecha ahora la tarifa de nuestro servicio ...
2      #LibertySeguros anunció el lanzamiento de plat...
3      @libertyseguros @LibertySeguros2 Estimados, re...
4      Después de 42 días calendario de mamadera de g...
...
1266   @LibertySegCol Ya solucioné, lo charro es que ...
1267   Liberty Seguros: notifica tu siniestro por int...
1268   @efecastro @SegurosdeEstado @SegurosBolivar @L...
1269   RT @efecastro: Resultados excelentes del secto...
1270   Resultados excelentes del sector asegurador en...
Name: full_text, Length: 1271, dtype: object
```

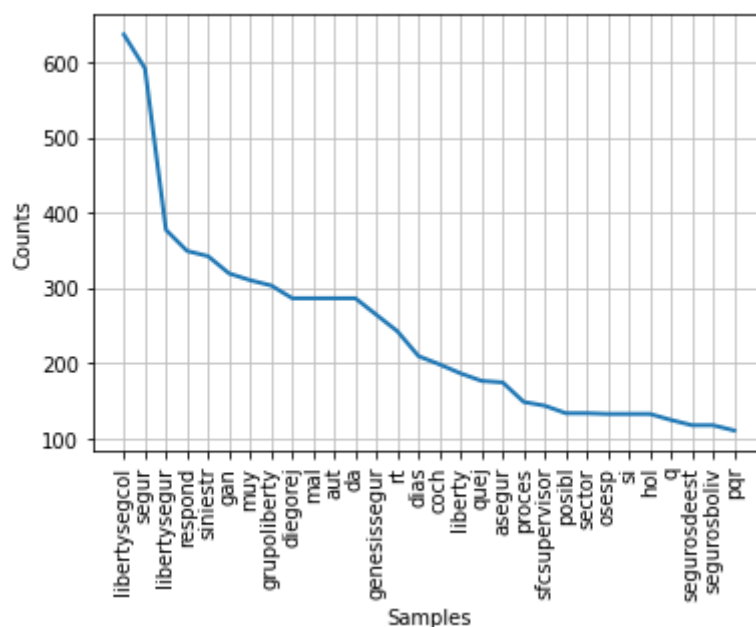
**Ilustración 35:** *Tweets* de Liberty (Fuente: Elaboración propia, 2021)

En la ilustración 35 se pueden observar los resultados obtenidos de la consulta de la aseguradora Liberty, de igual manera se puede ver el enunciado del *tweet* para el posterior análisis de sentimientos de cada uno de los mensajes generados.





Por último, en la categoría neutral se puede leer «seguros auto», «los siniestros» o «grupo liberty». El uso de estas palabras nos permite identificar que los usuarios aquí representados solo posicionan la aseguradora ya que cabe recordar que el solo hecho de que se hable a diario de Liberty quiere decir que se están logrando posicionar día a día.



**Ilustración 37:** Frecuencia de palabras, Liberty (Fuente: Elaboración propia, 2021).

Por otro lado, en el proceso de conteo de palabras, básicamente consiste en tratar de identificar cuáles son las palabras que son más frecuentes y que tiene relación con la aseguradora Liberty en el resultado de *tweets* obtenidos. Una vez realizado el proceso de limpieza de los *tweets* y con las palabras ya transformadas en *tokens*, se utiliza para este proceso la librería de Python Counter la cual muestra el resultado dependiendo del número de apariciones de cada una de las palabras en los *tweets* obtenidos.

Con esta información se genera un diagrama de series que muestra cuales son las palabras que tienen más frecuencia en la información obtenida como se puede ver en la ilustración 37.

q:124	libertysegcol:637
segurosdeest:117	segur:592
segurosboliv:117	libertysegur:377
pqr:110	respond:349
sig:105	siniestr:342
tecnolog:104	gan:319
deposit:100	muy:310
poliz:100	grupoliberty:303
radic:99	diegorej:286
mail:99	mal:286
cay:99	aut:286
revis:99	da:286
dm:99	genesissegur:264
cas:99	rt:242
ya:89	dias:209
favor:88	coch:198
ecosistem:83	liberty:186
excelent:81	quej:176
segurosencolombi:81	asegur:174
calendari:77	proces:148
gall:77	sfcsupervisor:143
devolu:77	posibl:133
prim:77	sector:133
corre:77	osesp:132
	si:132
	hol:132

**Ilustración 38:** Conteo de palabras, Liberty (Fuente: Elaboración propia, 2021).

Por otra parte, el conteo de apariciones de las palabras también nos permite saber cuáles fueron las palabras más recurrentes del conjunto de los *tweets* se muestra en la ilustración 38.

## 11.4. Seguros Equidad

Se realiza la consulta de la aseguradora Equidad, se obtuvieron 2.942 registros para analizar.

```
#Read data from Google Sheet: Asguradoras
df = pd.read_csv("/content/Twitter_Equidad_Seguros.csv")

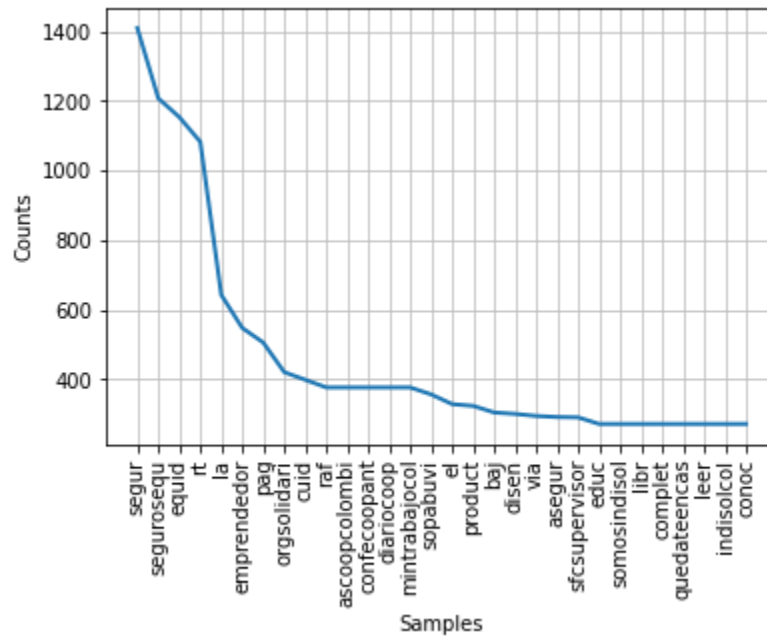
df.columns
#df.head()
df['full_text']
#print(df.shape)

0      RT @SegurosEquidad: Felicitamos a todos los Pe...
1      Que no hay saldo en su cuenta. Llamo a @Seguro...
2      Hola ! @SegurosEquidad hace dos semanas inform...
3      RT @CaquetaResiste: #Educación co #SomosIndiso...
4      RT @CaquetaResiste: #Educación co #SomosIndiso...
...
2937   RT @SegurosEquidad: Equidad Seguros anuncia pó...
2938   RT @SegurosEquidad: Equidad Seguros lanza al m...
2939   RT @sopabuvi: @OrgSolidarias @coogranada @rafa...
2940   RT @CaquetaResiste: #Educación co #SomosIndiso...
2941   RT @sopabuvi: @OrgSolidarias @rafa5023 @Mintra...
Name: full_text, Length: 2942, dtype: object
```

**Ilustración 39:** Tweets de Equidad Seguros (Fuente: Elaboración propia ,2021)

En la ilustración 39 se pueden identificar los resultados obtenidos de la consulta de la aseguradora Equidad, de igual manera se puede ver el enunciado del *tweet* para el posterior análisis de sentimientos de cada uno de los mensajes generados.





**Ilustración 41:** Frecuencia de palabras, Equidad (Fuente: Elaboración propia, 2021).

En el ejercicio de conteo de palabras se obtienen las palabras que son más frecuentes y que tiene relación con la aseguradora Equidad seguros en el resultado de *tweets* obtenidos. Una vez realizado el proceso de limpieza de los *tweets* y con las palabras ya transformadas en *tokens*, se utiliza para este proceso la librería de Python Counter la cual muestra el resultado dependiendo del número de apariciones de cada una de las palabras en los *tweets* obtenidos.

Con esta información se genera un diagrama de series que muestra cuales son las palabras que tienen más frecuencia en la información obtenida como se puede ver en la ilustración 41.

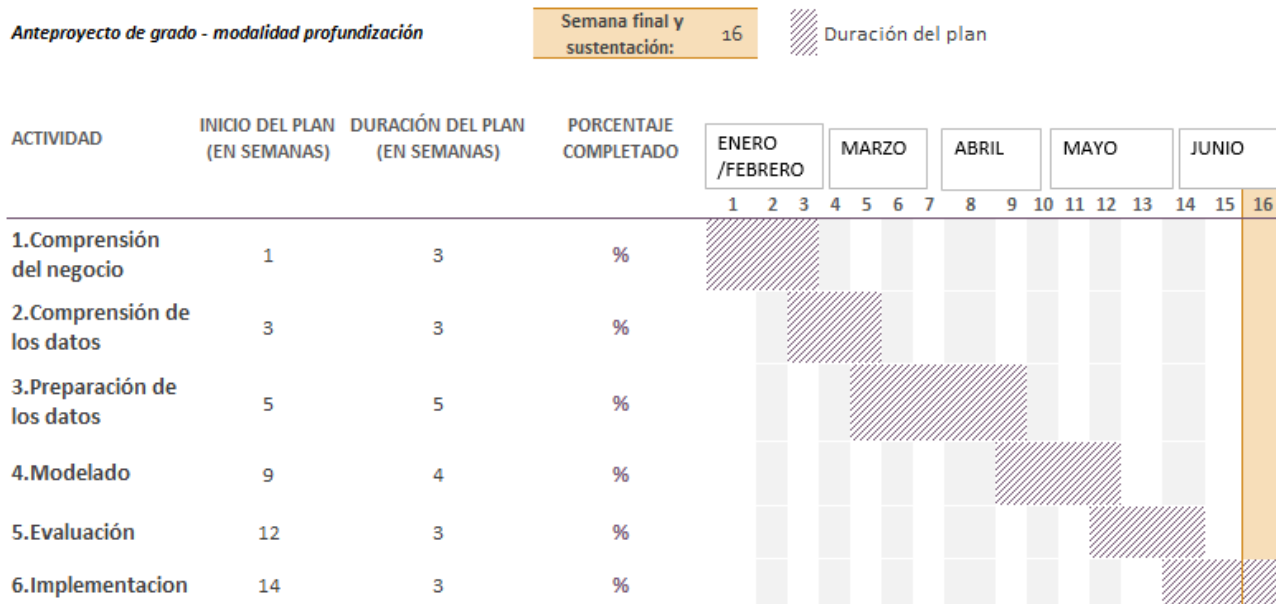
segur:1410	leer:272
segurosequ:1208	indisolcol:272
equid:1154	conoc:272
rt:1083	httpstcozxukmg:252
la:643	coogran:248
emprendedor:548	caquetares:239
pag:506	fech:235
orgsolidari:421	oanalf:226
cuid:399	periodicoecon:226
raf:377	judicial:218
ascoopcolombi:377	ecos:215
confecoopant:377	de:214
diariocoop:377	part:205
mintrabajocol:377	sector:193
sopabuvi:357	solidari:193
el:329	nuev:193
product:324	sol:190
baj:305	club:189
diseñ:301	no:181
via:295	proteg:177
asegur:292	pas:176
sfcsupervisor:291	febrer:174
educ:272	lleg:172
somosindisol:272	
libr:272	
complet:272	
quedateencas:272	

**Ilustración 42:** Conteo de palabras, Equidad (Fuente: Elaboración propia, 2021).

Por otra parte, el conteo de apariciones de las palabras mostrado en la ilustración 42. Permite realizar una idea de cuáles fueron las palabras clave más recurrentes respecto en este caso a la aseguradora Equidad seguros.

## 12. Cronograma de trabajo

En el desarrollo del cronograma de trabajo para el año 2021 se plantean las siguientes actividades que se llevarán a cabo en un rango de tiempo de 16 semanas, iniciando desde el mes de enero y finalizando en el mes de junio aproximadamente.



**Ilustración 43:** Cronograma del proyecto (Fuente: Elaboración propia, 2020).

## 13. Presupuesto

La fuente de financiación de este proyecto proyectado para el año 2021 se apoya en recursos propios y la colaboración de algunas de las aseguradoras en cuestión (Sura y Equidad seguros) para el desarrollo del proyecto.

### Anteproyecto de grado

NOMBRE	Gustavo Adolfo Martinez	MOTIVO	Propuesta anteproyecto
DEPARTAMENTO	Facultad de Ciencias Naturales e Ingeniería	INICIO	1/01/2021
		FINALIZACIÓN	1/06/2021
		PREPARADO POR	Gustavo Adolfo Martinez
		APROBADO POR	

VARIABLE	CANTIDAD DE HORAS GASTADAS*SEMANA	DESCRIPCIÓN	PRECIO
Honorarios del estudiante	8	Correspondientes al total de horas dedicadas	\$6.000.000,00
Honorarios de los tutores	5	Referentes al total de horas dedicadas en el acompañamiento y monitoreo de las actividades	\$2.000.000,00
Costos de la herramienta	7	La suscripción ( el costo no será asumido por tratarse de un uso académico y un acceso gratuito a la Api twitter )	\$700.000,00
TOTAL EN PESOS COLOMBIANOS			\$8.700.000,00

**Ilustración 44:** Presupuesto del proyecto (Fuente: Elaboración propia, 2020).



## 14. Conclusiones

- En primer lugar, se llevó a cabo una revisión de la literatura acerca del procesamiento de texto y el análisis de tweets usando técnicas de aprendizaje automático, con el objetivo de aumentar el conocimiento sobre dicho tema y así explorar ciertos enfoques para llegar a la solución de proyecto de tesis. Los resultados obtenidos se identifican para diferentes aseguradoras donde se puede incorporar el análisis que se generan a través de Twitter y que pueden realizar una mejor caracterización de los usuarios, con esto obtener más información para analizar.
- Con el desarrollo de este trabajo de tesis se desarrolla una visualización de datos que es de fácil uso lo que permite que pueda ser utilizado por cualquier persona o usuario sin necesidad de tener un conocimiento muy detallado en el desarrollo de las herramientas utilizadas para llevar a cabo el proyecto.
- Como partes críticas del proyecto, se podría destacar la obtención de los datos en el API de Twitter. Aparentemente parece un trabajo sencillo por la gran cantidad de información de la que se dispone, de hecho, encontrar los datos concretos para trabajar es un gran desafío, puesto que se tenían que cumplir una serie de requisitos para que el análisis tuviese éxito.
- Para el desarrollo de este trabajo fueron valiosas las herramientas que hacen posible el análisis de datos en las redes sociales, además de como todas las utilidades y aplicaciones que tiene en la industria, sobre todo por la gran cantidad de información que se expone al día en la red de Twitter y de la cual se puede extraer mucha información como patrones de comportamiento, intereses por distintas temáticas, etc. No se puede dejar en alto el impacto del COVID-19, tal como se pudo ver en los diferentes *tweets*, las personas que se han visto afectadas a nivel económico y por lo tanto es una tendencia al alza en cuanto a opiniones de cualquier tema.

### 14.1. Limitaciones del modelo

- Esta investigación proporciona datos contrastados del sector asegurador. El análisis de *tweets* en el sector asegurador podrá ser usado para estrategias de *marketing* en trabajos futuros y podrá servir como futuras investigaciones sobre la temática. Las limitaciones de la investigación son aquellas relativas al tamaño de la muestra y el número de aseguradoras que participan en el estudio.

## 15. Referencias bibliográficas

- [1] Statista, «statista.com,» 25 04 2019. [En línea]. Available: <https://es.statista.com/grafico/17792/usuarios-activos-mensuales-de-twitter/>. [Último acceso: 04 03 2021].
- [2] M. Congosto, «Barriblog,» 09 09 2017. [En línea]. Available: <http://www.barriblog.com/2017/10/lo-siempre-quiso-saber-del-api-twitter-nunca-se-atrevio-preguntar-actualizado-2017/>. [Último acceso: 04 03 2021].
- [3] «AulaPlaneta,» 09 09 2017. [En línea]. Available: <http://www.letoko.com/sitio/elblog/2017/10/09/cinco-herramientas-tic-crear-nubes-palabras/>. [Último acceso: 04 03 2021].
- [4] «healthdataminer,» [En línea]. Available: <https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/>. [Último acceso: 04 03 2021].
- [5] «Todoequipo,» 9 07 2020. [En línea]. Available: <https://todoequipo.com/ventajas-y-desventajas-de-python/>. [Último acceso: 04 03 2021].
- [6] M. Tejero, «Lasocialmedia,» Maria Tejero, 05 12 2020. [En línea]. Available: <https://lasocialmedia.es/caracteristicas-twitter>. [Último acceso: 04 03 2021].
- [7] M. P. Á. Rodríguez, «Universidad Politecnica de València,» Análisis de tweets y su influencia en los seguros de vida en el ámbito colombiano, 2019. [En línea]. Available: <https://riunet.upv.es/bitstream/handle/10251/150168/%C3%81vila%20-%20An%C3%A1lisis%20de%20tweets%20y%20su%20influencia%20en%20los%20seguros%20de%20vida%20en%20el%20%C3%A1mbito%20colombiano..pdf?sequence=1>. [Último acceso: 29 09 2020].
- [8] J. Fernandez, «We Are Social,» Digital 2020 El uso de las redes sociales abarca casi la mitad de la población mundial, 30 01 2020. [En línea]. Available: <https://wearesocial.com/es/blog/2020/01/digital-2020-el-uso-de-las-redes-sociales-abarca-casi-la-mitad-de-la-poblacion-mundial>. [Último acceso: 29 09 2020].
- [9] D. R. Pastor, «Universidad De Barcelona,» Big Data en sectores asegurador y financiero, 2015. [En línea]. Available: [http://diposit.ub.edu/dspace/bitstream/2445/140208/1/TFM-DEAF-189\\_Ramos.pdf](http://diposit.ub.edu/dspace/bitstream/2445/140208/1/TFM-DEAF-189_Ramos.pdf). [Último acceso: 29 09 2020].
- [10] C. Bembibre, «definicionabc,» definicionabc, 11 Septiembre 2010. [En línea]. Available: <https://www.definicionabc.com/tecnologia/twitter.php>. [Último acceso: 11 02 2021].

- [11] H. K. I. S. Philip Kotler, Marketing 4.0: transforma tu estrategia para atraer al consumidor digital, España: LID Editorial, 2016.
- [12] Banco BBVA, «Banco BBVA,» Cómo usar la API de Twitter en tu negocio, 05 09 2018. [En línea]. Available: <https://bbvaopen4u.com/es/actualidad/como-usar-la-api-de-twitter-en-tu-negocio>. [Último acceso: 29 09 2020].
- [13] P. R. Montero, «Universidad Carlos III de Madrid,» ANÁLISIS DE LA CYBERRIVALIDAD Y HOSTILIDAD EN REDES SOCIALES. CASO DE ESTUDIO TWITTER, 19 Junio 2017. [En línea]. Available: [https://e-archivo.uc3m.es/bitstream/handle/10016/27297/TFG\\_Patricia\\_Rodriguez\\_Montero.pdf?sequence=1](https://e-archivo.uc3m.es/bitstream/handle/10016/27297/TFG_Patricia_Rodriguez_Montero.pdf?sequence=1). [Último acceso: 12 02 2021].
- [14] A. E. O. C. y. A. C. C. Lopez, «ANALISIS MASIVO DE DATOS EN TWITTER PARA IDENTIFICACION DE OPINION,» *Repositorio Unsaac*, 2020.
- [15] A. Moreno, «Instituto De Ingenieria Del Conocimiento,» Procesamiento de lenguaje natural, 17 10 2017. [En línea]. Available: <https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/>. [Último acceso: 29 09 2020].
- [16] A. Moreno, «Instituto de Ingenieria Del Conocimiento,» Modelos para procesamiento, 17 10 2017. [En línea]. Available: <https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/>. [Último acceso: 29 09 2020].
- [17] K. Bannister, «Brand Watch,» Entendiendo el análisis de sentimiento, 10 02 2015. [En línea]. Available: <https://www.brandwatch.com/es/blog/analisis-de-sentimiento/>. [Último acceso: 29 09 2020].
- [18] T. Diaz, «Economiasimple,» 9 09 2018. [En línea]. Available: <https://www.economiasimple.net/glosario/nube-de-palabras>. [Último acceso: 08 03 2021].
- [19] M. Mendoza, «Rankia Seguros,» Qué es una empresa aseguradora?, 23 12 2018. [En línea]. Available: <https://www.rankia.cl/blog/mejores-seguros-chile/4110087-que-empresa-aseguradora>. [Último acceso: 29 09 2020].
- [20] A. V. Steve Bavister, «Psicología Y Mente,» Programación Neurolingüística (PNL), 2014. [En línea]. Available: <https://psicologiymente.com/vida/programacion-neurolinguistica>. [Último acceso: 29 09 2020].
- [21] «divulgaciondinamica,» 5 Abril 2017. [En línea]. Available: <https://www.divulgaciondinamica.es/blog/la-pnl-programacion-neurolinguistica/>. [Último acceso: 11 02 2021].
- [22] L. Gallego Preciado Cobos, «Universidad Carlos III De Madrid,» Análisis de sentimientos acerca de marcas mediante tweets, 19 11 2013. [En línea]. Available:

- [https://e-archivo.uc3m.es/bitstream/handle/10016/18524/PFC\\_Laura\\_Gallego-Preciado\\_Cobos.pdf?sequence=1&isAllowed=y](https://e-archivo.uc3m.es/bitstream/handle/10016/18524/PFC_Laura_Gallego-Preciado_Cobos.pdf?sequence=1&isAllowed=y). [Último acceso: 29 09 2020].
- [23] A. C. Casaverde Lopez, «Universidad Nacional De San Antonio Abad Del Cusco,» Análisis masivo de datos en twitter para identificacion de opinion, 2020. [En línea]. Available: [http://repositorio.unsaac.edu.pe/bitstream/handle/UNSAAC/5252/253T20200108\\_TC.pdf?sequence=1&isAllowed=y](http://repositorio.unsaac.edu.pe/bitstream/handle/UNSAAC/5252/253T20200108_TC.pdf?sequence=1&isAllowed=y). [Último acceso: 29 09 2020].
- [24] L. A. O. Palma, «Universidad Nacional Abierta Y a Distancia,» Aplicación de técnicas de análisis de información de la red social twitter, para la visualización de tendencias y necesidades laborales y de formación en el sector de t.i., 10 02 2019. [En línea]. Available: <https://repository.unad.edu.co/bitstream/handle/10596/28027/%20%09laortizpa.pdf?sequence=5&isAllowed=y>. [Último acceso: 29 09 2020].
- [25] M. P. Ávila Rodríguez, «Universidad Politecnica de Valencia,» Análisis de tweets y su influencia en los seguros de vida en el ámbito colombiano, 20 07 2020. [En línea]. Available: <https://riunet.upv.es/handle/10251/150168>. [Último acceso: 30 09 2020].
- [26] D. A. G. Corrales, «Pontificia Universidad Católica Del Ecuador,» Estudio de percepción de la imagen digital de aseguradora del sur, compañía de seguros y reaseguros, 2015. [En línea]. Available: <http://repositorio.puce.edu.ec/bitstream/handle/22000/9808/disertacion.pdf?sequence=1&isAllowed=y>. [Último acceso: 30 09 2020].
- [27] T. Novoa Triana, «Universidad de Bogotá Jorge Tadeo Lozano,» Prototipo de Sistema de Exploración y Generación de Herramientas de análisis para datos en twitter, 06 2020. [En línea]. Available: <https://expeditiorepositorio.utadeo.edu.co/bitstream/handle/20.500.12010/12532/Trabajo%20de%20grado.pdf?sequence=1&isAllowed=y>. [Último acceso: 30 09 2020].
- [28] R. M. ,. N. S. ,. S. R. R. M. Duwairi, «Jordan University of Science and Technology,» Sentiment Analysis in Arabic Tweets , 26 06 2014. [En línea]. Available: <http://www.just.edu.jo/~rehab/ICICS2014.pdf>. [Último acceso: 30 09 2020].
- [29] A. J. A. Eline M. van den Broek-Altenburg, «University of Vermont,» Using Social Media to Identify Consumers' Sentiments towards Attributes of Health Insurance during Enrollment Season, 13 05 2019. [En línea]. Available: <https://www.mdpi.com/2076-3417/9/10/2035>. [Último acceso: 30 09 2020].
- [30] I. Struweg, «University of Johannesburg,» A Twitter Social Network Analysis: The South African Health Insurance Bill Case, 01 04 2020. [En línea]. Available: [https://link.springer.com/chapter/10.1007/978-3-030-45002-1\\_11](https://link.springer.com/chapter/10.1007/978-3-030-45002-1_11). [Último acceso: 30 09 2020].

- [31] J. A. P. Castro, «Pontificia Universidad Católica De Valparaíso,» Análisis de sentimiento y clasificación de texto mediante Adaboost Concurrente, 2016. [En línea]. Available: [http://opac.pucv.cl/pucv\\_txt/txt-3500/UCD3631\\_01.pdf](http://opac.pucv.cl/pucv_txt/txt-3500/UCD3631_01.pdf). [Último acceso: 30 09 2020].
- [32] J. S. Castelló, «riunet,» Universidad Politècnica de València, 11 09 2015. [En línea]. Available: <https://riunet.upv.es/bitstream/handle/10251/55471/SELVA%20-%20Desarrollo%20de%20un%20sistema%20de%20an%C3%A1lisis%20de%20sentimiento%20sobre%20Twitter.pdf?sequence=1>. [Último acceso: 19 02 2021].
- [33] R. A. R. d. H. Rosa Montañes, «Instituto Tecnológico de Aragón,» Application of a hybrid deep learning model for Sentiment, 08 2018. [En línea]. Available: [https://zaguan.unizar.es/record/75832/files/texto\\_completo.pdf](https://zaguan.unizar.es/record/75832/files/texto_completo.pdf). [Último acceso: 19 02 2021].
- [34] D. P. V. L. Miguel Jasso Hernandez, «Benemerita Universidad Autonoma de Puebla,» Analisis de sentimientos en Twitter: impacto de las características morfológicas, 2014. [En línea]. Available: [https://www.rcs.cic.ipn.mx/2014\\_72/Analisis%20de%20sentimientos%20en%20Twitter\\_%20impacto%20de%20las%20caracteristicas%20morfológicas.pdf](https://www.rcs.cic.ipn.mx/2014_72/Analisis%20de%20sentimientos%20en%20Twitter_%20impacto%20de%20las%20caracteristicas%20morfológicas.pdf). [Último acceso: 19 02 2021].
- [35] J. L. P. R. J. A. T. Jason Paul Anturi Martínez, «Universidad del Cauca,» Clasificadores para el Análisis de Sentimientos en Twitter, 30 06 2016. [En línea]. Available: [https://www.researchgate.net/profile/Carlos\\_Cobos2/publication/337719633\\_Classifiers\\_for\\_Sentiment\\_Analysis\\_on\\_Twitter\\_a\\_review/links/5de6a44692851c83645fbf40/Classifiers-for-Sentiment-Analysis-on-Twitter-a-review.pdf](https://www.researchgate.net/profile/Carlos_Cobos2/publication/337719633_Classifiers_for_Sentiment_Analysis_on_Twitter_a_review/links/5de6a44692851c83645fbf40/Classifiers-for-Sentiment-Analysis-on-Twitter-a-review.pdf). [Último acceso: 19 02 2021].
- [36] D. R. Cortés, «ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA,» Tool for Opinion Mining and Sentiment Analysis on Twitter, 06 2017. [En línea]. Available: <https://riuma.uma.es/xmlui/bitstream/handle/10630/15422/DavidrubiocortesMemoria.pdf?sequence=1>. [Último acceso: 19 02 2021].
- [37] A. F. M. Erika Paola Paez Guarnizo, «IMPLEMENTACIÓN DE UN MODELO DE ANÁLISIS DE SENTIMIENTOS CON RESPECTO A LA JEP BASADO EN MINERÍA DE DATOS EN TWITTER.,» Universidad Católica De Colombia, 2020. [En línea]. Available: <https://repository.ucatolica.edu.co/jspui/bitstream/10983/24981/1/JEPDocumentoFinal.pdf>. [Último acceso: 19 02 2021].
- [38] A. E. Pérez, «El impacto de las emociones en el análisis de la polaridad en textos con lenguaje figurado en Twitter.,» Universidad Politecnica De Valencia, 07 2017. [En línea]. Available:

- <https://riunet.upv.es/bitstream/handle/10251/86127/ESCORTELL%20-%20EI%20impacto%20de%20las%20emociones%20en%20el%20an%C3%A1lisis%20de%20la%20polaridad%20en%20textos%20con%20lenguaje%20fig....pdf?sequence=1>. [Último acceso: 21 02 2021].
- [39] V. C. Bernal, «Los retos del sector asegurador en Colombia,» El Espectador, Bogotá, 2018.
- [40] J. V. Román, «sngular,» 02 08 2016. [En línea]. Available: <https://www.sngular.com/es/data-science-crisp-dm-metodologia/>. [Último acceso: 3 11 2020].
- [41] M. Mayo, «Preprocesamiento de datos de texto: un tutorial en Python,» Datos y ciencia, 3 Mayo 2018. [En línea]. Available: <https://medium.com/datos-y-ciencia/preprocesamiento-de-datos-de-texto-un-tutorial-en-python-5db5620f1767>. [Último acceso: 12 02 2021].
- [42] L. A. U. Fernández, «<https://medium.com/>,» 04 05 2019. [En línea]. Available: <https://medium.com/qu4nt/reducir-el-n%C3%BAmero-de-palabras-de-un-texto-lematizaci%C3%B3n-y-radicalizaci%C3%B3n-stemming-con-python-965bfd0c69fa>. [Último acceso: 02 03 2021].
- [43] I. Ruiz, «<https://webescuela.com/>,» [En línea]. Available: <https://webescuela.com/que-es-twitter-como-funciona/>. [Último acceso: 08 03 2021].
- [44] Sura, «Grupo Sura,» Seguros , Tendencias y Riesgos, 11 06 2020. [En línea]. Available: <https://segurossura.com/acerca-de-suramericana/nuestra-compania/>. [Último acceso: 12 02 2021].
- [45] «Twitter Sura,» [En línea]. Available: [https://twitter.com/segurossuracol?lang=es#:~:text=Seguros%20SURA%20Colombia%20\(%40SegurosSURAcól\)%20%7C%20Twitter](https://twitter.com/segurossuracol?lang=es#:~:text=Seguros%20SURA%20Colombia%20(%40SegurosSURAcól)%20%7C%20Twitter). [Último acceso: 08 03 2021].
- [46] «segurosdelestado,» [En línea]. Available: <https://www.segurosdelestado.com/>. [Último acceso: 09 03 2021].
- [47] «Twitter,» [En línea]. Available: <https://twitter.com/segurosdeestado?lang=es>. [Último acceso: 08 03 2021].
- [48] «Libertycolombia,» [En línea]. Available: <https://www.libertycolombia.com.co/>. [Último acceso: 09 03 2021].
- [49] «Twitter,» [En línea]. Available: [https://twitter.com/libertysegcol?lang=es#:~:text=Liberty%20Seguros%20\(%40LibertySegCol\)%20%7C%20Twitter](https://twitter.com/libertysegcol?lang=es#:~:text=Liberty%20Seguros%20(%40LibertySegCol)%20%7C%20Twitter). [Último acceso: 08 03 2021].

- [50] «Laequidadseguros,» [En línea]. Available: <https://www.laequidadseguros.coop/>. [Último acceso: 09 03 2021].
- [51] «Twitter,» [En línea]. Available: [https://twitter.com/segurosequidad?lang=es#:~:text=La%20Equidad%20Seguros%20\(%40SegurosEquidad\)%20%7C%20Twitter](https://twitter.com/segurosequidad?lang=es#:~:text=La%20Equidad%20Seguros%20(%40SegurosEquidad)%20%7C%20Twitter). [Último acceso: 08 03 2021].
- [52] M. A. Alvarez, «Desarrollo Web,» Desarrollo Web, 19 Noviembre 2003. [En línea]. Available: <https://desarrolloweb.com/articulos/1325.php>. [Último acceso: 12 02 2021].
- [53] «UNIVERSIDAD DE ALCALÁ,» UNIVERSIDAD DE ALCALÁ, 11 Enero 2020. [En línea]. Available: <https://www.master-data-scientist.com/scikit-learn-data-science/>. [Último acceso: 12 02 2021].
- [54] «<http://www.nltk.org/>,» wikipedia, 13 Abril 2020. [En línea]. Available: <http://www.nltk.org/>. [Último acceso: 12 02 2021].
- [55] «Bioinformatics,» Bioinformatics, 2019. [En línea]. Available: <https://bioinf.comav.upv.es/courses/linux/python/pandas.html#:~:text=pandas%20es%20un%20paquete%20de,potente%20tipo%20matricial%20a%20Python.&text=Datos%20tabulares%20con%20columnas%20de,etiquetas%20en%20columnas%20y%20filas..> [Último acceso: 12 02 2021].
- [56] «Google Colab,» [En línea]. Available: <https://colab.research.google.com/notebooks/welcome.ipynb?hl=es-AR>. [Último acceso: 12 02 2021].
- [57] «Anaconda,» [En línea]. Available: <https://www.anaconda.com/>. [Último acceso: 12 02 2021].
- [58] «enbeta,» enbeta, 04 Febrero 2014. [En línea]. Available: <https://www.genbeta.com/desarrollo/mongodb-que-es-como-functiona-y-cuando-podemos-usarlo-o-no>. [Último acceso: 12 02 2021].