UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA
POSTGRADO EN INFORMÁTICA

TESIS MAGISTER SCIENTIARUM

PROGRAMA DE MAESTRÍA EN GERENCIA ESTRATÉGICA EN TECNOLOGÍAS Y SISTEMAS DE INFORMACIÓN – GETSI
VERSIÓN 4 – GESTIÓN 2010-2011

"MODELO DE CUADRO DE MANDO INTEGRAL PARA EL MONITOREO DE SENTIMIENTO DE LA SOCIEDAD BOLIVIANA EN REDES SOCIALES: APLICACIÓN TWITTER"

POSTULANTE: Ing. Freddy Richard Rojas Illatarco
TUTOR: M.Sc. Ing. Cesar Enrique Castellón Escóbar

LA PAZ – BOLIVIA
2019

DEDICATORIA

A la mujer que me enseñó todo lo que sé y a mi compañera de vida.

Este trabajo está dedicado a mi madre y a mi esposa.

RESUMEN

Para conocer la opinión pública general, sobre determinados temas, es necesario llevar a cabo sondeos de opinión.

En el marco de la Ciencia de Datos, el presente trabajo propone un modelo que sirve de alternativa complementaria a los sondeos de opinión empleando la información de Twitter. Para ello, el modelo establece los criterios para el uso de las API de Twitter y conceptos de Teoría de Grafos para encontrar al universo total de usuarios de Twitter en Bolivia. Establece también los criterios para la recolección de Tweets y la aplicación de algoritmos de Aprendizaje Automático orientado al Análisis de Sentimiento sobre ellos.

El Análisis de Sentimiento, también llamado Minería de Opinión, permite determinar si una opinión expresada en un texto escrito en lenguaje natural es positiva o negativa.

El trabajo se completa con una versión demostrativa del modelo donde se determina la opinión pública boliviana respecto de una temática específica. Se hace uso de las capacidades de herramientas Python, Neo4j y PostgreSQL. Los resultados son presentados a través de indicadores clave en un Cuadro de Mando Integral.

Actualmente existen soluciones disponibles en Internet que producen reportes de Análisis de Sentimiento con base en datos de Twitter. A diferencia del modelo propuesto, estas soluciones se basan enteramente en la API de búsqueda de Twitter. Si tres personas juntas emiten diez opiniones, de las cuales seis son positivas y cuatro son negativas, estas soluciones indicarán que el 60% de las opiniones son positivas. El modelo propuesto encuentra a las tres personas que emitieron opiniones y determina cuántos de ellos emitieron opiniones positivas.

SUMMARY

To know the general public opinion, it's necessary to do opinion polls.

This work is a Data Science research and proposes a model that can be used as a complementary alternative to opinion polls using information from Twitter. For this purpose, the model establishes the criteria for the use of Twitter APIs and concepts of Graph Theory to find all Twitter users in Bolivia. The model also establishes the criteria for the gathering of Tweets and the use of Machine Learning algorithms oriented to Sentiment Analysis on the gathered Tweets.

Sentiment Analysis, also called Opinion Mining, allows to determine if an opinion expressed in a text written in natural language is positive or negative.

A demonstrative version of the model completes this work, where Bolivian public opinion is determined regarding a specific theme. The capabilities of Python, Neo4j, and PostgreSQL are used. The results are presented through key indicators in a Dashboard.

Currently, there are some solutions available on the Internet that produce Sentiment Analysis reports based on Twitter data. Unlike the proposed model, these solutions are based on the Twitter Search API. If three people together give ten opinions, where six are positive and four are negative, these solutions will tell that 60% of the opinions are positive. The proposed model on this work finds the three people who gave opinions and finds out how many of them gave positive opinions.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

Los sondeos de opinión son mediciones estadísticas que, basadas en encuestas, se realizan para conocer la opinión pública general sobre ciertos temas. Los resultados de los sondeos permiten conocer a distintos niveles el sentimiento, a favor o en contra, de la sociedad sobre un tema específico.

En el contexto nacional, algunos de los temas polémicos publicados en medios de prensa y redes sociales son: legalidad del aborto, pena de muerte, castración de violadores, alimentos transgénicos, ideología de género y el respeto de los resultados del 21F[1].

En Bolivia, los sondeos de opinión más difundidos son los llevados a cabo por empresas como IPSOS[2] y CIES-MORI[3], y por institutos de investigación de universidades. Debido a los costos que implica llevar a cabo un sondeo de opinión a nivel nacional, éstos no pueden llevarse a cabo con mucha frecuencia.

Una alternativa a los sondeos de opinión, basados en encuestas, es la monitorización de las redes sociales a través del Análisis de Sentimiento de los textos publicados por las personas sobre ciertos temas.

El Análisis de Sentimiento es un área de investigación del Procesamiento del Lenguaje Natural, el cual a su vez es una aplicación de técnicas de Aprendizaje Automático e Inteligencia Artificial. El Análisis de Sentimiento permite determinar si una opinión expresada en lenguaje natural es positiva, negativa o neutra.

---

[1] Referéndum constitucional del 21 de febrero de 2016 donde se consultó al soberano la aprobación o rechazo de un proyecto constitucional para permitir al presidente y vicepresidente postularse nuevamente a las elecciones generales del 2019. El soberano rechazó el proyecto (opción "No") con el 51.3% de los votos.

[2] http://www.ipsos.com.bo/

[3] http://www.ciesmori.com/

Los sondeos de opinión y el Análisis de Sentimiento en las redes sociales se diferencian sustancialmente en que en el primero, el encuestador debe buscar al encuestado y pedir su opinión, y en el segundo, el “encuestador” debe determinar la opinión del “encuestado” a partir de las expresiones que éste hubiera vertido en las redes sociales.

Considerando que en octubre del presente año se llevarán a cabo las elecciones generales para elegir al presidente y al vicepresidente del estado, uno de los temas más polémicos del año será la reelección o no de los actuales mandatarios.

El presente es un trabajo de investigación en el campo de la Ciencia de Datos y propone un modelo para monitorear el sentimiento de la sociedad boliviana respecto de un tema polémico de interés nacional cualquiera. Para probar el modelo se propone determinar la opinión pública boliviana respecto de la potencial reelección de los actuales mandatarios.

CAPITULO I. ASPECTOS GENERALES

1.1 Antecedentes

1.1.1 Sondeos de Opinión en Elecciones Presidenciales

Los sondeos de opinión y las encuestas pueden llegar a fallar, un ejemplo de esto es la victoria de Donald Trump en las elecciones presidenciales de los Estados Unidos el año 2016 la cual no fue prevista por la gran mayoría de las encuestas.

El 8 de noviembre de 2016, Donald Trump ganó las elecciones presidenciales, aunque las 5 encuestadoras más grandes del país del norte habían determinado que en promedio sólo el 41% de la población aprobaba la transición presidencial a favor de Trump[4].

En Bolivia, el 12 de octubre de 2014 se llevaron a cabo las últimas elecciones nacionales. Los resultados publicados por el Órgano Electoral Plurinacional[5] de las elecciones nacionales fueron: MAS-IPSP 61.36%, UD 24.23%, PDC 9.04%, MSM 2.71% y PVB-IEP 2.65%.

Una encuesta fue hecha por IPSOS[6] entre el 8 y el 23 de septiembre de 2014 sobre una muestra de más de 3,000 personas. La encuesta fue publicada por las cadenas de televisión PAT y ATB el 1º de octubre de 2014 con los siguientes resultados: MAS 56.0%, UD 15.0%, PDC 9.0%, MSM 3.0% y PVB 1.0%.

[4] https://fivethirtyeight.com/features/can-you-trust-polling-in-the-age-of-trump/
[5] https://www.oep.org.bo/procesos-electorales-y-consultas/elecciones-generales/elecciones-generales-2014/
[6] http://www.la-razon.com/index.php?_url=/nacional/animal_electoral/Tercera-Ipsos-Doria-Medina-Quiroga_0_2135186542.html

Otra encuesta fue hecha por Equipos MORI[7] entre el 18 y 29 de septiembre de 2014 sobre una muestra de 2,464 personas. La encuesta fue publicada por el periódico El Deber el 2 de octubre de 2014 con los siguientes resultados: MAS 59.0%, UD 18.0%, PDC 9.0%, MSM 3.0% y PVB 2.0%.

En promedio, ambas encuestas estuvieron 3.86% por debajo del resultado final para el ganador y un 7.73% por debajo del resultado final para el segundo.

Las encuestas se hicieron con muestras de alrededor de 3,000 personas. En total 5,319,141 personas emitieron sus votos en las elecciones de octubre de 2014.

1.1.2 El estudio hecho sobre el Referéndum por el Brexit

En noviembre de 2016, se publicaron los resultados de un estudio elaborado por Bruegel[8] sobre el discurso público en Twitter antes del referéndum del 23 de junio de 2016 llevado a cabo en el Reino Unido relativo a la salida del Reino Unido de la Unión Europea – Brexit. Bruegel es un centro de investigación especializado en economía con sede en Bruselas, Bélgica.

El estudio explora la hipótesis de que las redes sociales podrían ser un mejor indicador de las intenciones de voto que las encuestas de opinión tradicionales.

El estudio realizó el análisis de opinión de más de 890,000 tweets publicados desde el 2012 hasta el referéndum. El análisis se hizo en colaboración con el *Dortmund Center for data-based Media Analysis* – DoCMA de la Universidad Técnica de Dortmund, Alemania.

[7] http://www.la-razon.com/nacional/animal_electoral/Encuesta-Equipos-Morales-Doria-Medina_0_2136386441.html
[8] http://bruegel.org/2016/11/tweeting-brexit-narrative-building-and-sentiment-analysis/

El estudio concluye, en que Twitter puede considerarse un tipo novedoso entre los medios de comunicación de élite, que permite a líderes de opinión comunicarse directamente con otros líderes de opinión y una parte del público políticamente activa.

Se afirma que la metodología del estudio captura algo que la teoría de probabilidades y los sondeos de opinión no pudieron revelar, y plantea la interrogante de si será útil en futuras elecciones. Las conclusiones del estudio no fueron acompañadas de los métodos empleados en el estudio para la recolección de datos.

1.1.3 El estado de las Redes Sociales en Bolivia

En junio de 2018, la Agencia de Gobierno Electrónico y Tecnologías de Información y Comunicación – AGETIC publicó el informe del "Estado de las Tecnologías de Información y Comunicación en el Estado Plurinacional de Bolivia"[9].

Los resultados del informe relativos al uso redes sociales tienen base en la "Encuesta Nacional de Opinión sobre TIC"[10] la cual fue hecha entre el 3 y el 18 de diciembre de 2016 sobre una muestra de 5,536 encuestas.

El informe establece que en Bolivia el 94% de la población internauta utiliza Facebook y el 17% utiliza Twitter. Para el efecto, se estableció que un internauta es una persona mayor de 14 años de edad que ha tenido acceso a Internet en los últimos 30 días previos a la encuesta.

El Instituto Nacional de Estadística – INE[11] establece que, con base en el Censo de Población y Vivienda del año 2012, la población proyectada en

[9] https://agetic.gob.bo/pdf/estadotic/AGETIC-Estado-TIC.pdf
[10] https://agetic.gob.bo/pdf/dia_internet_encuesta.pdf
[11] http://datos.ine.gob.bo/binbol/RpWebEngine.exe/Portal?LANG=ESP

Bolivia para el año 2019, entre 15 y 79 años de edad es de 6,980,400 personas. Si el 94% de la población internauta utiliza Facebook y el 17% utiliza Twitter, entonces esas poblaciones serían de aproximadamente 6.5 millones de personas y 1.1 millones de personas respectivamente.

Si bien existen fuentes en Internet que venden información de uso en redes sociales, como **SocialBakers.com**[12] y **Statista.com**[13] cuyos propósitos son mercadeo para grandes empresas y PYME, a la fecha no se cuenta con información actualizada sobre el uso de Twitter en Bolivia de fuentes locales.

1.1.4 El estudio más reciente hecho en Bolivia

En noviembre de 2018, se publicó en el portal **BrujulaDigital.net**[14] el estudio "El oficialismo está perdiendo la batalla de las redes sociales" elaborado por Raúl Peñaranda U. y Milton Condori.

Para el estudio se tomaron como objetos de evaluación las cuentas en Twitter de 5 representantes del gobierno y 5 cuentas en Twitter de representantes de la oposición.

Se tomó como muestra los primeros 30 Tweets publicados desde el 1º de septiembre de 2018 por cada cuenta, es decir, se analizaron los comentarios recibidos de un total 300 Tweets.

Los resultados del estudio son una serie de porcentajes, por cuenta de Twitter, de comentarios favorables, desfavorables y neutros recibidos. Por ejemplo: los representantes del gobierno tuvieron en conjunto un 72.9% de comentarios negativos, 21.9% de comentarios positivos y 5.2% de reacciones neutras.

[12] https://www.socialbakers.com/statistics/twitter/profiles/bolivia/society/
[13] https://www.statista.com/statistics/303931/twitter-users-latin-american-countries/
[14] http://www.brujuladigital.net/politica/estudio-el-oficialismo-esta-perdiendo-la-batalla-de-las-redes-sociales

Este estudio aplica el Análisis de Sentimiento de los comentarios aunque de forma artesanal y sujeta al criterio subjetivo de la persona que hizo el análisis.

1.2 Planteamiento del Problema

La recolección de información por encuestas para sondeos de opinión o estudios de mercado implican un alto costo por las tareas complejas que implica: determinación de una muestra representativa, evaluación de accesibilidad a localidades urbanas y rurales, ejecución y supervisión del trabajo en campo, digitación y procesamiento de datos.

Por ejemplo, el 15 de noviembre de 2018, el Ministerio de Comunicación realizó la contratación de la empresa CIES Internacional S.R.L. para la realización de una Encuesta de Opinión Pública sobre el "Alcance de los Logros de Gestión de Acuerdo a la Agenda 2025". La contratación de la consultoría está publicada en el SICOES[15] y fue hecha por Bs. 49.700,00 con un plazo de 25 días calendario.

Este tipo de estudios son llevados a cabo con métodos probados para asegurar su imparcialidad con márgenes de error aceptables. Sin embargo, los resultados sólo reflejan la percepción de la sociedad en un determinado periodo de tiempo. Se tienen que repetir los estudios para determinar la evolución de dicha percepción. Además, estos estudios no miden el grado de honestidad de los encuestados. Los encuestados pueden decir o no la verdad en el momento en que son cuestionados por un encuestador.

Si bien existen empresas especializadas y centros de investigación que han incursionado en la explotación de la información de redes sociales, ninguna ha publicado los métodos que siguieron para determinar las muestras. En el mejor

[15] https://www.sicoes.gob.bo CUCE: 18-0087-00-904990-0-E

de los casos, se han publicado las conclusiones de los estudios y algunas estrategias para analizar mensajes individuales basadas en Análisis de Sentimiento.

Este escenario plantea preguntas que el presente trabajo responde: ¿Es posible aumentar el tamaño de las muestras? ¿Es posible ampliar el periodo de tiempo de un estudio? ¿Es posible aumentar el grado de honestidad de los encuestados? ¿Cuál sería la solución tecnológica parametrizable que permita implementar un modelo de recolección de información confiable de tendencias de opinión de la sociedad boliviana respecto de una temática específica?

1.3 Objetivos General y Específicos de la Investigación

1.3.1 Objetivo general

Desarrollar un Modelo de Cuadro de Mando Integral para el monitoreo de sentimiento de la sociedad boliviana, respecto de una temática específica, con base en las opiniones expresadas por las personas en la red social Twitter.

1.3.2 Objetivos específicos

Los objetivos específicos son los siguientes:

a) Investigar las capacidades de interacción e integración a través de las Interfaces de Programación de Aplicaciones de Twitter.

b) Seleccionar los recursos tecnológicos de Ciencia de Datos adecuados para aplicar tareas de Análisis de Sentimiento.

c) Diseñar el Modelo de Cuadro de Mando Integral para la explotación de la base de datos de Twitter aplicando Teoría de Grafos y Minería de Opinión con un enfoque de Análisis de Negocios.

d) Implementar una versión demostrativa operativa del Modelo propuesto.

e) Evaluar los resultados de la versión demostrativa y establecer conclusiones.

1.4 Planteamiento de la Hipótesis

La información que proporciona Twitter, a través de sus API, es suficiente para construir una muestra válida para una medición estadística a nivel nacional.

1.4.1 Operacionalización de Variables

El presente trabajo es propositivo con un enfoque cuantitativo por cuanto se fundamenta en una necesidad, realiza una propuesta, es secuencial y probatorio.

Según su alcance, es de tipo descriptivo por cuando describe la naturaleza de un segmento de la población: los usuarios bolivianos de Twitter.

El método de investigación es inductivo por cuando se parte de observaciones particulares, los usuarios de Twitter, para llegar a conclusiones generales, la población boliviana.

A continuación, se describe la operacionalización de las variables del trabajo.

Operacionalización de variables.

Fuente: Elaboración propia.

VARIABLE	INDICADOR	VALOR	TIPO DE VARIABLE
Sentimiento de opiniones expresadas en Tweets	Polaridad de una opinión	Positiva = 1 Negativa = -1 Neutra = 0	Nominal traducida a Discreta
Sentimiento de cuentas de usuario de Twitter	Promedio de las polaridades de las opiniones	Intervalo cerrado entre -1 y 1	Discreta traducida a Continua
	Cuenta de usuario de Twitter	Identificador de cuenta	
	Localización	Departamento Bolivia	
	Periodo de tiempo	Intervalo cerrado de tiempo	

De acuerdo a su dimensionalidad, la primera variable es unidimensional y la segunda es multidimensional. Siendo que las variables son lógicas, también son categóricas y, por lo tanto, subjetivas.

No hay relación de dependencia entre las variables debido a que el trabajo de investigación es descriptivo. La primera variable es de caracterización y la segunda es la variable de interés.

CAPITULO II. MARCO TEÓRICO

En el presente capítulo se describen los conceptos que son empleados en la definición del modelo propuesto y en la aplicación del modelo. Todos ellos fueron extraídos de la Bibliografía detallada al final del presente documento y de las fuentes indicadas en notas al pie de cada página.

A continuación, el esquema que describe la relación que guardan los conceptos abordados organizados por tipo de marco al que pertenecen.

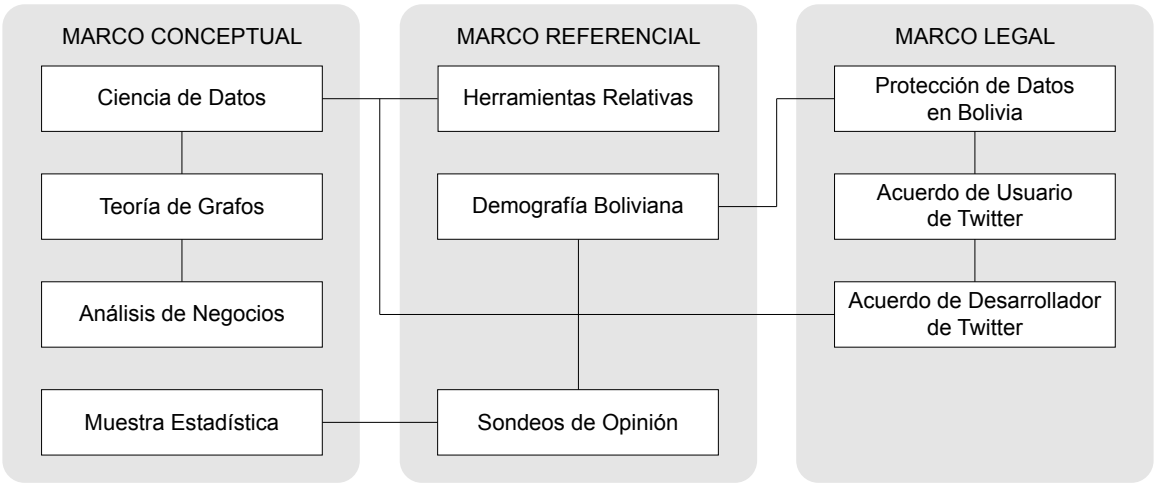

Figura 1: Esquema general de conceptos asociados al modelo.

Fuente: Elaboración propia.

2.1 Marco Conceptual

2.1.1 Ciencia de Datos

La Ciencia de Datos, también conocida como Descubrimiento de Conocimiento, es un campo interdisciplinario que combina métodos de la Estadística, la Minería de Datos, el Aprendizaje Automático y la Analítica Predictiva, para extraer valor de los datos. Estas técnicas se basan en encontrar patrones útiles, conexiones y relaciones dentro de datos estructurados y no estructurados.

El proceso de la Ciencia de Datos es un conjunto de pasos que es independiente de los problemas, algoritmos y herramientas empleadas. Para generar conocimiento, a partir de un tema de análisis, generalmente se aplica un algoritmo de aprendizaje a través de una herramienta de software para Ciencia de Datos como Python, RapidMiner, Lenguaje R o Excel, por citar algunos[16].

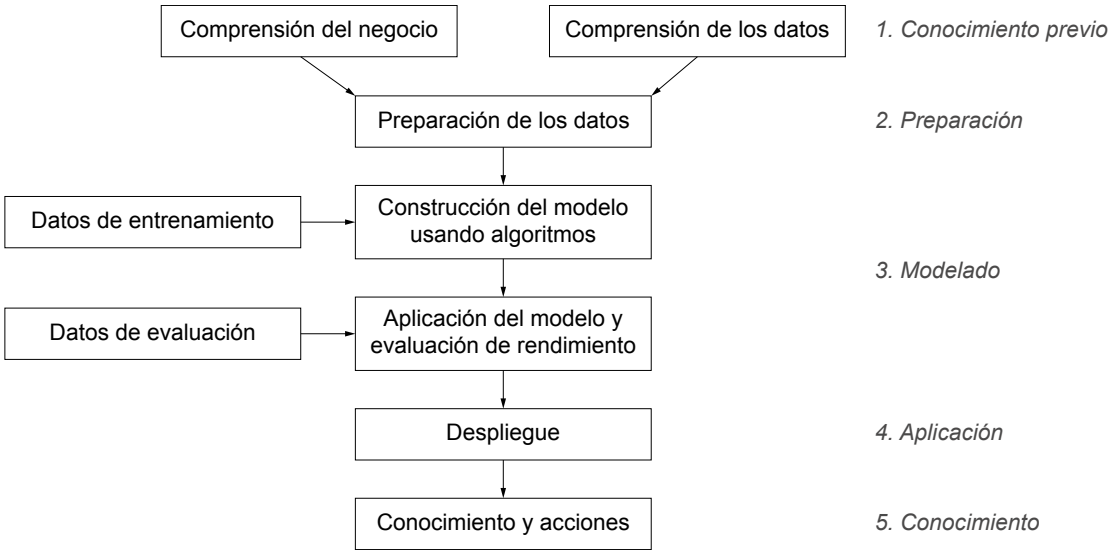

Figura 2: Proceso de la Ciencia de Datos.

Fuente: Ciencia de Datos, Conceptos y Práctica (Kotu, 2019).

El proceso de la Ciencia de Datos comienza con datos, que pueden ser desde una simple colección de observaciones numéricas hasta una matriz compleja de millones de observaciones. Para obtener resultados significativos de un conjunto de datos, antes de que algoritmos de aprendizaje puedan procesarlos, se requiere un gran esfuerzo para recolectarlos, prepararlos, limpiarlos y estandarizarlos.

La Ciencia de Datos, la Inteligencia Artificial y el Aprendizaje Automático están relacionados entre sí. Aunque a menudo se usan indistintamente y se

[16] https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

combinan entre sí, estos tres campos son distintos según el contexto donde se los aplique.

El Aprendizaje Automático, considerado un subcampo de la Inteligencia Artificial, tiene una amplia variedad de técnicas orientadas a la Lingüística y al Procesamiento del Lenguaje Natural.

2.1.1.1 Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural, *Natural Language Processing* o NLP en inglés, cubre cualquier tipo de manipulación informática del lenguaje natural. Se entiende por lenguaje natural al que utilizamos para la comunicación cotidiana con personas en distintos contextos multilingüe.

Por un lado, el Procesamiento del Lenguaje Natural puede ser tan simple como contar frecuencias de palabras para encontrar patrones en escrituras. Por otro lado, el Procesamiento del Lenguaje Natural implica comprender las expresiones humanas complejas, hasta el punto de poder darles respuestas útiles.

Las tecnologías basadas en Procesamiento del Lenguaje Natural se están generalizando cada vez más. Por ejemplo, los teléfonos inteligentes admiten texto predictivo, reconocimiento de escritura manuscrita y generación de voz, los motores de búsqueda entregan resultados a partir de texto no estructurado, la traducción automática nos permite leer textos escritos en distintos idiomas.

En décadas pasadas, solo los expertos con formación en Lingüística, Matemáticas y el Aprendizaje Automático podían participar en el Procesamiento del Lenguaje Natural. Hoy, podemos usar bibliotecas Procesamiento del Lenguaje Natural ya escritas que simplifican el procesamiento previo de textos.

2.1.1.2 Análisis de Sentimiento

Desde inicios del año 2000, el Análisis de Sentimiento, también llamado Minería de Opinión, ha sido una de las áreas de investigación más activas en el Procesamiento del Lenguaje Natural, *Natural Language Processing* o NLP en inglés.

El Análisis de Sentimiento implica el uso de la Minería de Datos, el Aprendizaje Automático y la Inteligencia Artificial para analizar textos en busca de sentimientos e información subjetiva.

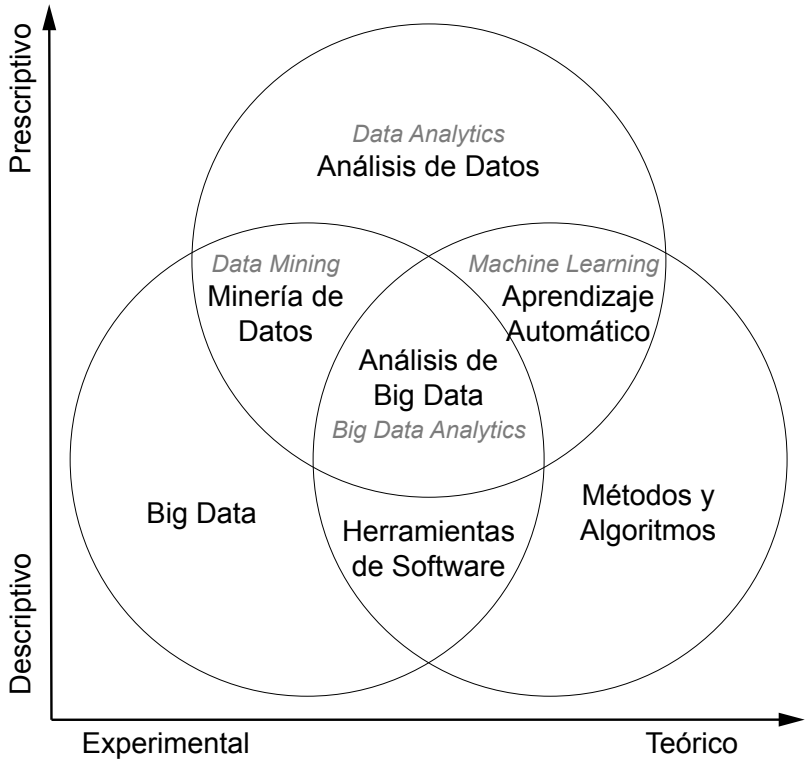

Figura 3: Campos de la Ciencia de Datos.

Fuente: Artículo de Igor Savinkin en Scraping.pro[17]

El objetivo del Análisis de Sentimiento es definir herramientas automáticas capaces de extraer información subjetiva de textos escritos en lenguaje natural, como opiniones y sentimientos, a fin de crear conocimiento

[17] http://scraping.pro/data-analytics-data-analysis-data-mining-data-science-machine-learning-big-data/

estructurado y procesable para ser utilizado por un sistema de soporte a la toma decisiones o por una persona que toma decisiones. Si bien puede existir una confusión entre sentimiento y opinión, se dice que **una opinión implica un sentimiento positivo, un sentimiento negativo o ningún sentimiento**.

A continuación, se describen las características principales que constituyen el Análisis de Sentimiento:

1. **Clasificación de Sentimiento**. Generalmente el primer paso en el Análisis de Sentimiento consiste en distinguir entre oraciones subjetivas y objetivas. Si una oración se clasifica como objetiva, no se requieren otras tareas. Si la oración se clasifica como subjetiva, es necesario estimar su polaridad: positiva, negativa o neutral.

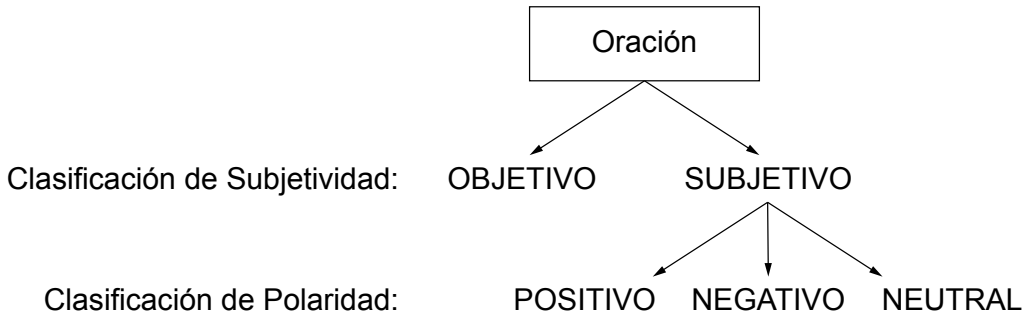

Figura 4: Flujo del Análisis de Sentimiento.

Fuente: Análisis de Sentimiento en Redes Sociales (Pozzi, 2017).

2. **Niveles de Análisis**. La primera opción cuando se aplica el Análisis de Sentimiento es definir qué significa el texto del objeto analizado. En general, el Análisis de Sentimiento en las redes sociales puede hacerse en tres niveles:
    a. Nivel de **Mensaje**. El objetivo es clasificar la polaridad de un mensaje completo.
    b. Nivel de **Oración**. El objetivo es determinar la polaridad de cada oración contenida en el mensaje completo.

c. Nivel de **Entidad** y **Aspecto**. Se realiza un análisis más detallado que a nivel de mensaje y oración. Se trabaja sobre la suposición de que una opinión consiste en un sentimiento y un objetivo de opinión.

3. **Opinión Regular y Opinión Comparativa**. Una opinión puede asignarse a uno de los siguientes grupos: opinión regular u opinión comparativa.
    a. Una **opinión regular** es una opinión estándar que puede ser directa o indirecta. La opinión directa es una expresada directamente sobre una entidad. La opinión indirecta es una expresada indirectamente sobre una entidad sobre base de sus efectos sobre otras entidades.
    b. Una **opinión comparativa** expresa una relación de similitudes o diferencias entre dos o más entidades, o una preferencia del dueño de la opinión en función de algunos aspectos compartidos de las entidades. Una opinión comparativa generalmente se expresa con el uso de la forma comparativa o superlativa de un adjetivo o adverbio.

4. **Opinión Explícita y Opinión Implícita**. Entre los distintos sentidos que puede tener una opinión, se debe distinguir entre opiniones explícitas y opiniones implícitas:
    a. Una **opinión explícita** es una declaración subjetiva que da una opinión regular o comparativa.
    b. Una **opinión implícita** es una declaración objetiva que implica una opinión regular o comparativa que generalmente expresa un hecho deseable o indeseable.

5. **El Papel de la Semántica**. La semántica del lenguaje utilizado en las redes sociales es fundamental para analizar con precisión las

expresiones de los usuarios. Debe tenerse en cuenta el contexto de una expresión para analizar adecuadamente el sentimiento implícito. Una oración analizada tal como está puede parecer positiva o negativa, pero si se analiza adecuadamente desde un punto de vista semántico el resultado puede ser lo opuesto.

6. **Figuras Retóricas**. Una figura retórica es cualquier desviación ingeniosa del modo ordinario de hablar o escribir. Las figuras más problemáticas del habla en el Procesamiento del Lenguaje Natural son la ironía y el sarcasmo. Si bien la ironía a menudo se usa para enfatizar los sucesos que se desvían de lo esperado, el sarcasmo se usa comúnmente para transmitir críticas implícitas con una víctima en particular como su objetivo.

   En muchos casos, las expresiones irónicas y sarcásticas son difíciles de reconocer para los humanos, son mucho más difíciles para las máquinas. En el contexto del Análisis de Sentimiento, donde la ironía y el sarcasmo generalmente se consideran sinónimos, cuando una oración irónica o sarcástica se clasifica como positiva o negativa, probablemente significa lo contrario.

7. **Relaciones en Redes Sociales**. El Análisis de Sentimiento en las redes sociales generalmente se basa en el supuesto de que los textos proporcionados por los usuarios son independientes y distribuidos idénticamente. Un enfoque tratado en el análisis del contenido de las redes sociales está relacionado con el principio de homofilia, la tendencia de las personas por la atracción a sus homónimos. En este contexto, las relaciones de "amistad" pueden usarse para inferir que los usuarios conectados pueden tener opiniones similares.

La siguiente figura ilustra las tareas más comunes del Análisis de Sentimiento, entre las que se encuentra la Clasificación de Polaridad que

es la que permite determinar la polaridad de una opinión: positiva, negativa o neutra.

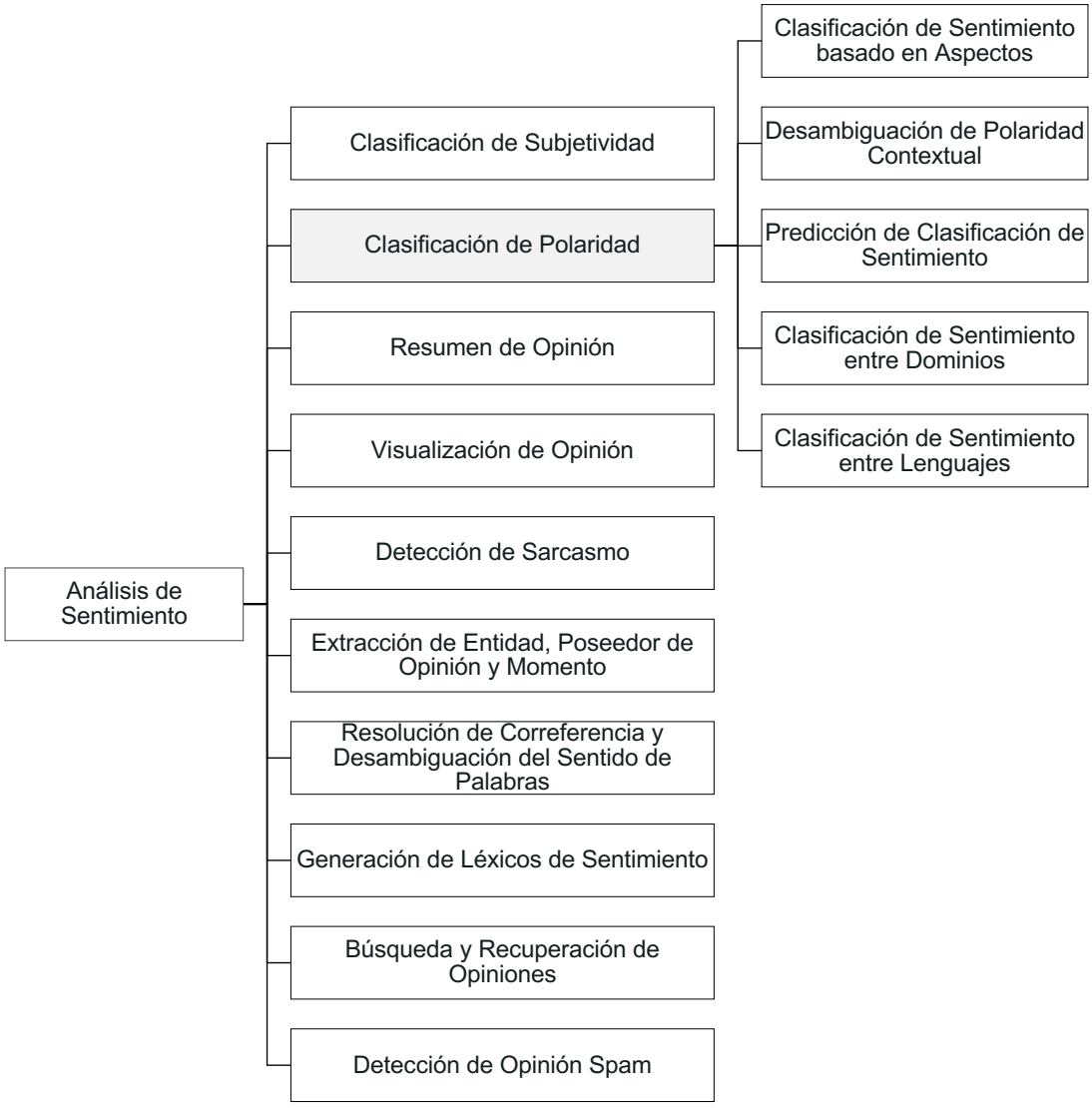

Figura 5: Tareas del Análisis de Sentimiento.

Fuente: Análisis de Sentimiento en Redes Sociales (Pozzi, 2017).

Las estrategias disponibles para el Análisis de Sentimiento ya no son tan efectivas para extraer opiniones en los nuevos entornos. El análisis de las redes sociales, además de los problemas del análisis tradicional y del procesamiento del lenguaje natural, introduce otras complejidades: extranjerismos, mensajes cortos, expresiones regionales, ruido,

abreviaturas no estándares, emoticones y memes, entre otras. En otras palabras, el lenguaje que usamos en los teléfonos inteligentes, las tabletas y las computadoras está evolucionando con la tecnología.

Teniendo en cuenta esta evolución del lenguaje, no todos los sistemas y herramientas de Análisis de Sentimiento disponibles pueden ser aplicados en cualquier contexto. La mayoría de ellos tienen grandes avances, pero en el idioma inglés norteamericano. Desafortunadamente, el castellano boliviano no está en el *roadmap* de estas soluciones.

Sin embargo, existen alternativas para ejecutar la Clasificación de Polaridad a partir de herramientas de Ciencia de Datos como Python.

Python[18] encabeza el último ranking internacional de **KDNuggets.com**[19] de herramientas para Análisis, Ciencia de Datos y Aprendizaje Automático. Entre las bibliotecas de Procesamiento del Lenguaje Natural más populares basados en Python[20] está NTLK.

2.1.1.3 NLTK

El Kit de Herramientas para Procesamiento del Lenguaje Natural, *Natural Language Toolkit* o NLTK en inglés, es una plataforma para crear programas de Python para trabajar con datos del lenguaje humano que actualmente sirve de base para muchos proyectos de investigación.

NLTK se creó originalmente en 2001 como parte de un curso de lingüística computacional en el Departamento de Informática y Ciencias de la Información de la Universidad de Pensilvania. La plataforma proporciona múltiples interfaces y de bibliotecas de procesamiento de texto, recursos

[18] https://www.python.org
[19] https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html
[20] https://www.kdnuggets.com/2018/07/comparison-top-6-python-nlp-libraries.html

léxicos y recursos de corpus de textos los cuales son conjuntos de textos o de datos destinados a la investigación científica.

A continuación, la lista de los módulos más importantes de NLTK:

Tareas de procesamiento del lenguaje y módulos NLTK.
Fuente: Procesamiento del Lenguaje Natural con Python (Bird, 2009).

TAREA DE PROCESAMIENTO	MÓDULOS NLTK	FUNCIONALIDAD
Acceso a corpus	nltk.corpus	Interfaces estandarizadas para corpus y léxicos.
Procesamiento de cadenas	nltk.tokenize nltk.stem	Tokenizadores, tokenizadores de oraciones, stemmers
Descubrimiento de colocación	nltk.collocations	Prueba T, Chi-cuadrado, información mutua puntual.
Etiquetado gramatical	nltk.tag	N-gramos, Backoff, Brill, HMM, TnT.
Clasificación	nltk.classify nltk.cluster	Árboles de decisión, entropía máxima, algoritmos Naive Bayes, EM, K-formas.
Fragmentado	nltk.chunk	Expresiones regulares, N-gramos, entidades con nombre.
Análisis	nltk.parse	Gráficos, basados en características, unificación, probabilístico, dependencia.
Interpretación semántica	nltk.sem nltk.inference	Cálculo Lambda, lógica de primer orden, comprobación de modelos.
Métricas de evaluación	nltk.metrics	Precisión, recuerdo, coeficientes de acuerdo.
Probabilidad y estimación	nltk.probability	Distribuciones de frecuencia, distribuciones de probabilidad suavizadas.

TAREA DE PROCESAMIENTO	MÓDULOS NLTK	FUNCIONALIDAD
Aplicaciones	nltk.app nltk.chat	Coordinador gráfico, analizadores, navegador de WordNet, Chatbots.
Trabajo de campo lingüístico	nltk.toolbox	Manipular datos en formato SIL Toolbox.

Aunque el kit de herramientas proporciona una amplia gama de funciones, el kit no lo hace todo, es una caja de herramientas y continúa evolucionando con el aporte de una gran cantidad de contribuyentes en el campo del Procesamiento del Lenguaje Natural.

Si bien el kit de herramientas es lo suficientemente eficiente como para ejecutar tareas significativas, no está optimizado entornos en producción. Tales optimizaciones a menudo implican algoritmos más complejos o implementaciones en lenguajes de programación de nivel inferior como C o C++. Esto haría que el software de sus módulos fuera menos legibles y más difíciles de instalar. De acuerdo a la filosofía de NLTK, las implementaciones claras son preferibles a las más eficientes pero indescifrables.

2.1.1.4 Clasificadores Naive Bayes

Los clasificadores Naive Bayes son algoritmos de clasificación de Aprendizaje Automático, *Machine Learning* en inglés, que tienen base en el Teorema de Bayes que expresa la probabilidad condicional de un evento aleatorio A dado otro evento aleatorio B.

En los clasificadores Naive Bayes, cada característica es importante en la determinación de qué etiqueta debe asignarse a un valor de entrada dado. Para elegir una etiqueta para un valor de entrada, el clasificador Naive

Bayes comienza calculando la probabilidad previa de cada etiqueta, que se determina verificando la frecuencia de cada etiqueta en el conjunto de entrenamiento. La contribución de cada característica se combina con esta probabilidad previa, para llegar a una estimación de probabilidad para cada etiqueta. La etiqueta cuya estimación de probabilidad es la más alta se asigna al valor de entrada.

Por ejemplo, dadas dos características, la probabilidad posterior de que se de una característica A dado que se dió una característica B será:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Donde:

P(B|A) es la probabilidad de que se de B dado A

P(A) es la probabilidad de A

P(B) es la probabilidad de B

A continuación, se presenta una ilustración abstracta del procedimiento utilizado por el clasificador Naive Bayes para elegir el tema de un documento.

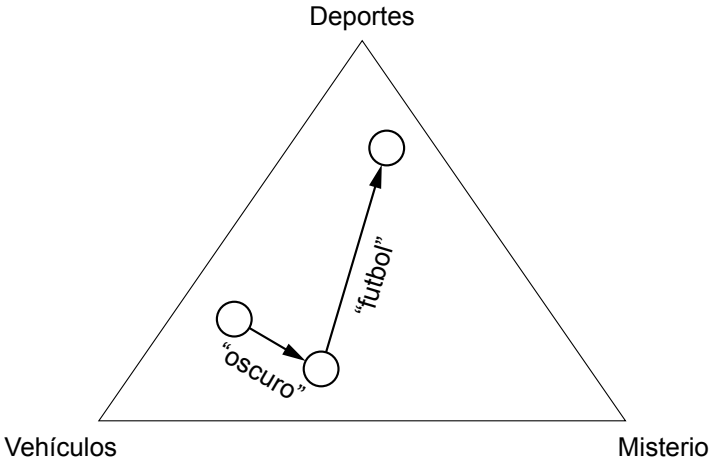

Figura 6: Ilustración del procedimiento de un clasificador Naive Bayes.

Fuente: Procesamiento del Lenguaje Natural con Python (Bird, 2009).

En los datos de entrenamiento, la mayoría de los documentos son de vehículos, por lo que el clasificador comienza en un punto más cercano a la etiqueta "Vehículos". Pero luego considera el efecto de cada característica. En el ejemplo, el documento de entrada contiene la palabra “oscuro”, que es un indicador débil de documentos de “Misterio”, pero también contiene la palabra “fútbol”, que es un indicador fuerte para documentos de “Deportes”. Después de que cada característica hace su contribución, el clasificador verifica a qué etiqueta está más cerca y asigna esa etiqueta a la entrada.

Este clasificador permite predecir clases de una manera fácil y rápida cuando se trata de problemas de clasificación binarios y multiclase. Sin embargo, cuando el conjunto de datos de prueba tiene una característica que no ha sido observada en el conjunto de entrenamiento, el clasificador asignará una probabilidad de cero con lo cual no podrá realizar predicciones.

TextBlob[21] es una biblioteca de Python para procesar datos de textos que emplea NLTK. Proporciona una API simple para tareas comunes de Procesamiento del Lenguaje Natural como el etiquetado gramatical, extracción de frases nominales, análisis de sentimiento, clasificación y traducción.

Los clasificadores Naive Bayes están implementados en el módulo *nltk.classify* de NLTK y es accesible desde el módulo *textblob.classifiers* de TextBlob.

[21] https://textblob.readthedocs.io/en/dev/index.html

2.1.2 Teoría de Grafos

En 1735 el matemático, físico y filósofo suizo-ruso Leonhard Paul Euler publicó el artículo "*Solutio problematis ad geometriam situs pertinentis*", en español "La solución de un problema sobre la geometría de la posición". En este artículo Euler planteó la solución al problema de Los Siete Puentes de Königsberg, una ciudad en la antigua Alemania donde siete puentes unían cuatro regiones de la ciudad.

El problema consistió en encontrar un recorrido para cruzar a pie toda la ciudad, pasando sólo una vez por cada uno de los puentes y regresando al mismo punto de partida.

Euler propuso un modelo matemático para resolver el problema donde nombró a cada región como un punto y a cada puente como una línea que une dos regiones.

Este hito marcó el inicio de la Teoría de Grafos.

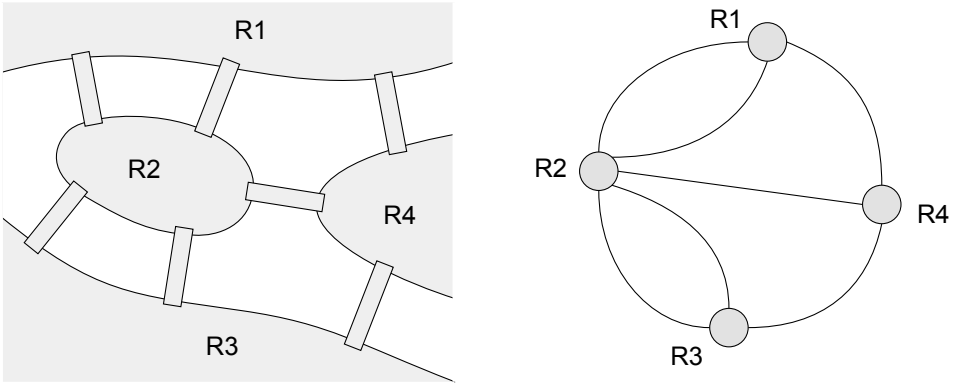

Figura 7: Representación de los Siete Puentes de Königsberg.
Fuente: Adaptación de Wikipedia[22].

Por definición un grafo es una colección de puntos llamados vértices V, unidos por líneas llamadas aristas A, donde cada arista conecta dos vértices. En el caso del problema, el conjunto de vértices está formado por los puntos

[22] https://es.wikipedia.org/wiki/Problema_de_los_puentes_de_K%C3%B6nigsberg

R1, R2, R3 y R4, y las aristas {R1, R2}, {R1, R2}, {R1, R4}, {R2, R3}, {R2, R3}, {R3, R4} y {R2, R4}.

Las aristas pueden ser representadas gráficamente como líneas rectas o líneas curvas. Las aristas que conecta a un vértice con él mismo son llamadas lazos. Las aristas que conectan los mismos vértices más de una vez, son denominadas aristas múltiples.

En el grafo del problema no existen lazos, pero sí existen aristas múltiples que une el vértice R1 con R2, y el vértice R2 con R3.

Cuando las aristas de un grafo tienen dirección, los grafos son denominados dirigidos y la notación usada para las aristas es $\overrightarrow{\{R1, R2\}}$ o (R1, R2).

Entre las definiciones básicas de grafos se encuentran las siguientes:

a) Dos grafos son isomorfos si existe una correspondencia uno a uno entre todos los vértices y todas las aristas del grafo.
b) El grado de un vértice V es el número de aristas que lo conectan.
c) Dos vértices A, B son adyacentes si existe al menos una arista que los conecte.
d) El subgrafo de un grafo es un subconjunto de vértices y aristas del grafo inicial.
e) Una trayectoria es una sucesión de vértices donde cada vértice es adyacente al siguiente y todas las aristas en la sucesión son distintas.
f) Un circuito es una trayectoria que comienza y termina en el mismo vértice.
g) Un grafo es conexo si cualquier par de vértices se pueden conectar por una trayectoria.
h) Un grafo no conexo es denominado disconexo, este está formado por segmentos subgrafos conexos.

Volviendo al problema de Los Siete Puentes de Königsberg, Euler observó que para encontrar una trayectoria que pase por cada arista una sola vez, es necesario que cada vértice, a la que se llega por una arista, debe tener otra arista distinta para dejarla y continuar con el recorrido.

Además, para que un vértice sea el punto final o inicial de un circuito, el vértice debe tener una arista adicional que llegue al vértice, es decir, el vértice debe ser de grado impar.

Estas observaciones llevaron a las definiciones de Trayectoria de Euler y Circuito de Euler:

a) Una Trayectoria de Euler es una trayectoria que recorre todas las aristas de un grafo conexo. Si un grafo tiene más de dos vértices de grado impar, entonces no puede tener una trayectoria de Euler.
b) Un Circuito de Euler es un circuito que recorre todas las aristas de un grafo conexo. Si un grafo tiene algún vértice de grado impar, entonces no puede tener un circuito de Euler.

El grafo del problema tiene cuatro vértices, todos los vértices tienen grado impar. Por lo tanto, el grafo no puede tener una Trayectoria de Euler y tampoco puede tener un Circuito de Euler, en otras palabras, no es posible visitar las cuatro regiones de la ciudad cruzando cada puente una sola vez volviendo al mismo punto de partida.

2.1.2.1 Bases de Datos de Grafos

Un grafo también puede ser descrito como un conjunto de nodos y las relaciones que los conectan. Los grafos pueden representar entidades como nodos y las formas en que esas entidades se relacionan. Esta estructura de uso general permite modelar un sin fin de escenarios. Uno de estos escenarios es la representación de las redes sociales de los usuarios de Twitter.

Por ejemplo, los datos de Twitter pueden ser representados fácilmente como un grafo.

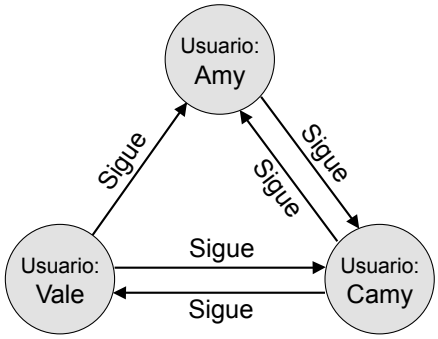

Figura 8: Grafo simple de una red social.

Fuente: Adaptación de la documentación de Neo4j.

En la figura vemos el grafo de una red simple de usuarios de Twitter. Cada nodo representa a un usuario, lo cual indica su papel en la red social. Estos nodos se conectan a través de las relaciones que los usuarios tienen entre ellos, a saber: Amy sigue a Camy, Camy sigue a Amy, Vale y Camy se siguen mutuamente, pero, aunque Vale sigue a Amy, Amy no corresponde. El grafo puede ser ampliado con los Tweets que cada usuario publica en la red social.

Naturalmente, el grafo real de la red social de Twitter, si se toman en cuenta sólo a los Usuarios y no los Tweets, es millones de veces más grande que el del ejemplo, pero funciona bajo los mismos principios.

Una de las principales tecnologías utilizadas para la persistencia de grafos transaccionales, generalmente accesibles desde una aplicación en tiempo real, son los gestores de bases de datos de grafos.

Un gestor de bases de datos de grafos es un sistema de gestión de bases de datos con los métodos estándar de alta, baja, modificación y consulta para modelos de datos orientados a grafos. Estas soluciones usualmente están pensadas para ofrecer rendimiento transaccional, integridad transaccional y disponibilidad operativa.

Existen dos propiedades principales de los gestores de bases de datos de grafos:

a) **Tipo de almacenamiento**. Algunos gestores de bases de datos de grafos utilizan almacenamiento nativo de grafos que está optimizado y diseñado para almacenar y administrar grafos. Sin embargo, no todos los gestores utilizan almacenamiento nativo de grafos. Algunos serializan los datos de los grafos en una base de datos relacional, una base de datos orientada a objetos o algún otro almacén de datos de uso general.

b) **Tipo de procesamiento**. El procesamiento de gráficos nativos, también conocido como adyacencia sin índice, es el medio más eficiente de procesar datos en un grafo porque los nodos conectados se apuntan físicamente entre sí en la base de datos. Los motores de procesamiento de grafos no nativos utilizan otros medios para procesar operaciones de altas, bajas, modificaciones y consultas que no están optimizadas para manejar datos conectados.

2.1.2.2 Neo4j

Neo4j[23] es un gestor de base de datos de grafos que contiene un motor de persistencia embebido e implementado en Java, basado en disco, completamente transaccional y que almacena estructuras de datos de grafos de forma nativa. Neo4j es el gestor más nativo en lo que respecta tanto al almacenamiento como al procesamiento de grafos.

[23] https://neo4j.com

Neo4j encabeza el ranking internacional de **DB-Engines.com**[24] de sistemas de administración de bases de datos orientados a grafos según su popularidad.

El modelo de datos de las bases de datos de grafos de Neo4j se denomina "modelo de grafo de propiedades etiquetadas" y está formado por:

a) Nodos: que son las entidades en el grafo.

b) Etiquetas: cada nodo puede tener una o más etiquetas que especifiquen el tipo de nodo.

c) Relaciones: que son las conexiones entre dos nodos y tienen una sola dirección y tipo.

d) Propiedades: las propiedades son datos organizados en modo clave-valor y pueden ser almacenados tanto en los nodos como en las relaciones.

El modelo del grafo a implementarse depende de la aplicación que se tenga en mente, es decir, hay un sin fin de grafos implementables. Sin embargo, el gestor de base de datos es la única herramienta necesaria para el análisis de los grafos a partir de algoritmos.

Por ejemplo, la documentación de Neo4j contiene casos de aplicación entre los cuales está el grafo *Open Movie* cuyo grafo simplificado es:

[24] https://db-engines.com/en/ranking/graph+dbms

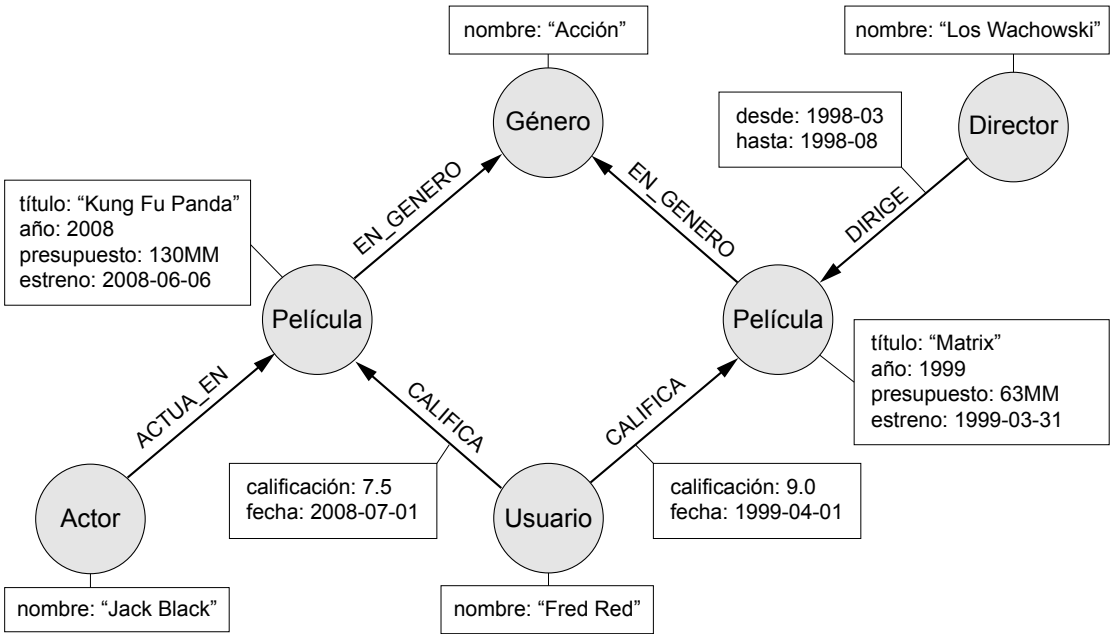

Figura 9: Ejemplo de modelo de datos soportado por Neo4j.

Fuente: Adaptación de la documentación de Neo4j.

En la figura los nodos tienen las etiquetas: Género, Película, Director, Actor y Usuario; el nodo etiquetado como “Película” tiene las propiedades: título, año, presupuesto y fecha de estreno; y la relación etiquetada como “CALIFICA” tiene las propiedades: calificación dada y la fecha en que se la dio.

Así como los gestores de bases de datos relacionales tienen al lenguaje estructurado de consultas SQL para manipular datos, Neo4j tiene a Cypher.

Cypher[25] es un lenguaje de consulta de grafos declarativo que permite consultas y actualizaciones expresivas y eficientes de grafos. Está diseñado para ser simple y adecuado tanto para desarrolladores como para analistas. Las consultas sobre grafos altamente complejos se pueden

[25] https://neo4j.com/docs/cypher-manual/3.5

expresar fácilmente, lo cual permite concentrarse en el tema de análisis, en lugar del acceso a los datos.

Continuando con el ejemplo del grafo *Open Movie*, si se quisiera hacer la consulta ¿Cuántas calificaciones tiene cada película de Matrix?, entonces la comando Cypher que responde a la pregunta sería la siguiente:

Ejemplo de consulta Cypher.

Fuente: Adaptación de la documentación de Neo4j.

	Comando Cypher
1	`MATCH (p:Pelicula)<-[:CALIFICA]-(u:Usuario)`
2	`WHERE p.titulo CONTAINS "Matrix"`
3	`WITH p.titulo AS pelicula, COUNT(*) AS calificaciones`
4	`RETURN pelicula, calificaciones`
5	`ORDER BY calificaciones DESC`
6	`LIMIT 5;`

Donde las filas:

1. Busca un patrón de relación existente en el grafo.
2. Filtra los patrones solo para aquellas que coincidan con la condición: la propiedad título, de los nodos etiquetados como “Pelicula”, contiene el texto “Matrix”.
3. Cuenta la cantidad de patrones que coincidan por cada película.
4. Especifica los datos que serán devueltos por la consulta: película y cantidad de calificaciones.
5. Ordena los resultados por la cantidad de calificaciones en orden descendente.
6. Retorna sólo los primeros cinco resultados.

2.1.2.3 Algoritmos de Detección de Comunidades

Los algoritmos de grafos se utilizan para calcular métricas de grafos, nodos o relaciones. Estos algoritmos pueden proporcionar información relevante sobre entidades como centralidades y clasificación, o

información inherente a la estructura de los grafos como las comunidades: detección de comunidades, división de grafos y agrupaciones.

Muchos algoritmos de gráficos son procedimientos iterativos que frecuentemente atraviesan el grafo utilizando trayectorias aleatorias, búsquedas de amplitud o de profundidad o coincidencia de patrones. Debido al crecimiento exponencial de las posibles trayectorias con una distancia creciente, muchos de los procedimientos también tienen una alta complejidad algorítmica. Neo4j ha implementado estos algoritmos y sus optimizaciones.

Los algoritmos de detección de comunidades evalúan cómo se agrupan o dividen grupo de nodos, así como su tendencia a fortalecerse o separarse.

Los algoritmos para detección de comunidades implementados en Neo4j son:

a) **Algoritmo de Louvain**. El método de Louvain es un algoritmo para detectar comunidades en redes propuesto en 2008 por la Universidad de Louvain.

   El algoritmo maximiza un puntaje de modularidad para cada comunidad, donde la modularidad es una cuantificación de la asignación de los nodos a las comunidades evaluando cuán densamente están conectados los nodos dentro de una comunidad comparado con como estarían conectados en una red aleatoria.

   El algoritmo de Louvain es uno de los algoritmos basados en modularidad más rápidos y funciona bien con grafos grandes. También revela una jerarquía de comunidades a diferentes escalas, lo que puede ser útil para comprender el funcionamiento global de una red.

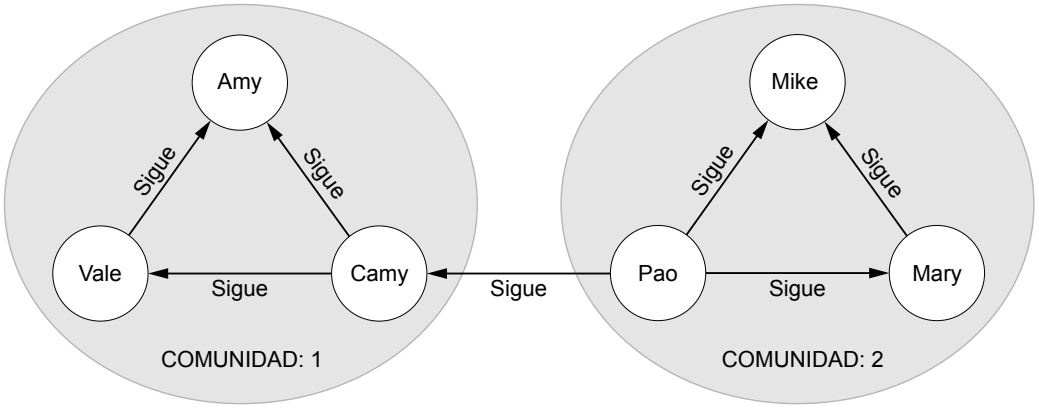

Figura 10: Ejemplo de la aplicación del Algoritmo de Louvain.

Fuente: Adaptación de la documentación de Neo4j.

b) **Algoritmo de Propagación de Etiquetas**. El algoritmo de propagación de etiquetas es un algoritmo rápido para encontrar comunidades en un grafo. Este algoritmo fue propuesto por Raghavan[26] en 2007.

   Detecta las comunidades utilizando solo la estructura de la red como guía, y no requiere una función objetivo predefinida o información previa sobre las comunidades. Una característica importante de este algoritmo es que a los nodos se les pueden asignar etiquetas preliminares para reducir el rango de soluciones generadas. Esto significa que puede usarse como una forma semi-supervisada de encontrar comunidades donde seleccionamos manualmente algunas comunidades iniciales.

[26] https://arxiv.org/pdf/0709.2938.pdf

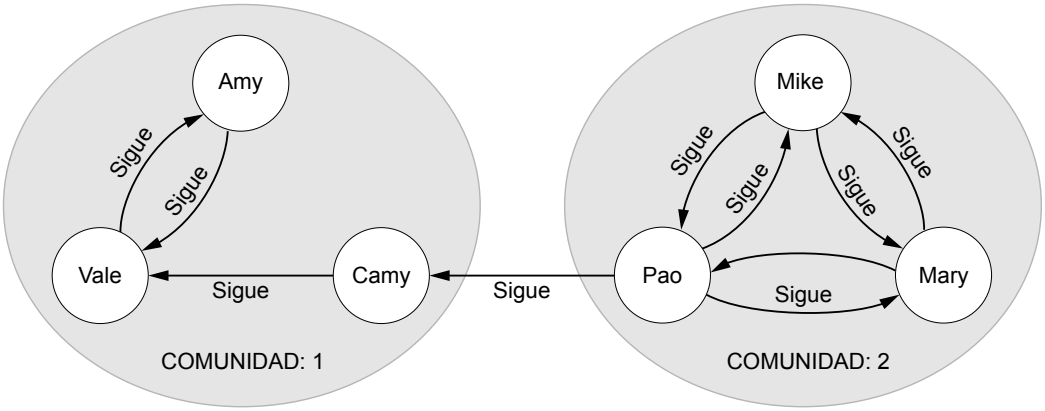

Figura 11: Ejemplo de la aplicación del Algoritmo de Propagación de Etiquetas.

Fuente: Adaptación de la documentación de Neo4j.

c) **Algoritmo de Componentes Conectados**. Este algoritmo busca conjuntos de nodos conectados en un grafo no dirigido donde cada nodo es accesible desde cualquier otro nodo en el mismo conjunto. Se diferencia del algoritmo Componentes Fuertemente Conectados porque solo necesita una ruta para existir entre pares de nodos en una dirección, mientras que el algoritmo de Componentes Conectados necesita una ruta para existir en ambas direcciones. Ambos algoritmos a menudo son usados al inicio de un análisis para comprender la estructura de un grafo.

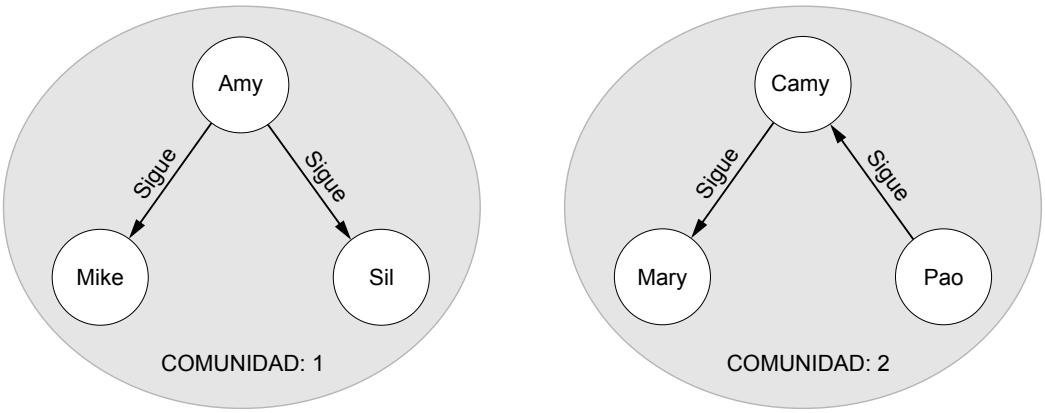

Figura 12: Ejemplo de la aplicación del Algoritmo de Componentes Conectados.

Fuente: Adaptación de la documentación de Neo4j.

2.1.3 Análisis de Negocios

El Análisis de Negocios, o *Business Analytics* en inglés, es el uso de datos, tecnología de la información, análisis estadístico, métodos cuantitativos, modelos matemáticos y modelos basados en computadora, orientado a tomadores de decisión para comprender el comportamiento de las empresas y para tomar mejores decisiones basadas en hechos.

El Análisis de Negocios es un proceso de transformación de datos que puede hacerse con varias herramientas, desde Microsoft Excel, pasando por software comercial especializado, herramientas de Aprendizaje Automático, hasta suites de Inteligencia de Negocios.

La librería Python de Aprendizaje Automático **scikit-learn**[27] contiene un conjunto amplio de herramientas simples y eficientes para Minería de Datos, Análisis de Datos y Análisis de Negocios. Esta librería está construida sobre las librerías NumPy, SciPy y matplotlib. Para la construcción del Cuadro de Mando Integral se cuenta con la librería gráfica **Plotly**[28].

El Análisis de Negocios comienza con la recopilación, organización y manipulación de datos y está soportado por tres componentes principales:

a) **Análisis Descriptivo**. El Análisis Descriptivo es el tipo de análisis más utilizado y mejor entendido. La mayoría de las empresas comienzan con Análisis Descriptivo para comprender el comportamiento pasado y actual de la empresa y tomar decisiones informadas en consecuencia. El Análisis Descriptivo resume los datos en gráficos e informes significativos, por ejemplo, ventas, ingresos, costos y presupuestos. Este proceso permite a los tomadores de decisión obtener informes para, por ejemplo, revisar el rendimiento

[27] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
[28] https://plot.ly/python/

del negocio, para encontrar problemas, áreas de oportunidad o identificar patrones y tendencias. La pregunta típica que responde el Análisis Descriptivo es ¿Qué pasó?

b) **Análisis Predictivo**. El Análisis Predictivo busca predecir el futuro examinando datos históricos, detectando patrones o relaciones entre los datos y luego extrapolando estos hacia adelante en el tiempo. El Análisis Predictivo puede predecir el riesgo y encontrar relaciones en los datos que podrían no ser tan evidentes con técnicas tradicionales. El uso de técnicas avanzadas, permite la detección de patrones ocultos en grandes cantidades de datos segmentados y agrupados en conjuntos homogéneos para predecir su comportamiento particular y detectar tendencias. Se ha usado Análisis Predictivo para alertar a un cliente de tarjeta de crédito de un posible cargo fraudulento. La pregunta típica que responde el Análisis predictivo es ¿Qué pasará?

   Las funciones matemáticas utilizadas en los modelos analíticos predictivos incluyen las siguientes funciones: lineal, logarítmica, polinómica, de potencia y exponencial.

c) **Análisis Prescriptivo**. El Análisis Prescriptivo utiliza los conceptos de optimización para identificar las mejores alternativas para minimizar consecuencias o maximizar resultados. Muchos problemas en empresas implican demasiadas opciones o alternativas para que un tomador de decisión las considere de manera efectiva y oportuna. Este tipo de análisis es usado en muchas áreas de negocios, incluidas las operaciones, el marketing y las finanzas. Por ejemplo, para identificar la cantidad óptima de efectivo que debe almacenar un determinado cajero automático. Las técnicas matemáticas y estadísticas del Análisis Predictivo pueden ser combinar con la

optimización para tomar decisiones que tengan en cuenta la incertidumbre en los datos. La pregunta típica que responde el Análisis Prescriptivo es ¿Qué debemos hacer?

2.1.4 Muestra Estadística

La Estadística es una herramienta de investigación usualmente asociada a la recolección de grandes cantidades de datos y su presentación en tablas y gráficos, cálculo de totales, promedios y porcentajes que permite, además, diseñar experimentos y muestras.

El concepto de Estadística y su aplicación maneja los conceptos de variable, valor y dato. Una variable es una característica de una población que puede adoptar una serie de valores, y dato es el valor que adopta una variable.

El proceso de construcción de cada variable con sus respectivas categorías para una investigación se denomina operacionalización de variables. Este proceso reduce progresivamente un concepto relativamente abstracto, la idea inicial de la investigación, a un número concreto de conceptos menos abstractos y más limitados en su alcance, hasta alcanzar las variables referentes de la realidad que se quiere analizar.

Una vez definidas las variables, se debe definir a quiénes se van a estudiar. Al conjunto de todos los individuos que son del interés de la investigación se denomina universo o población.

Cuando la población es muy grande, resulta imposible estudiarla en su totalidad, razón por la cual se hace necesario tomar muestras del total de la población. Una muestra es una porción representativa de individuos que permitirá generalizar los resultados de una investigación al total de la población.

Luego de haber definido la población objetivo, se debe establecer los pasos a seguir para obtener una muestra. El diseño de muestras se hace siguiendo un plan de muestro, cuyos tres primeros pasos son:

1. **Definición de la población, unidades de muestreo, extensión del muestreo y tiempo de recolección de información**. El proceso inicia con la identificación del marco de muestreo, es decir, la definición del conjunto de todos los individuos que pueden salir seleccionados durante el proceso de selección de la muestra. Esto implica, definir cada una de las unidades de muestreo. Una unidad de muestreo es un elemento o un conjunto de elementos de una población que pueden salir seleccionados durante el proceso de muestreo.

   Es fundamental establecer la extensión del muestreo, es decir, se debe establecer el lugar donde se llevará a cabo el muestreo. Para concluir, es necesario especificar cuándo se recolectará la información, es decir, el periodo de tiempo en el cual se hará el trabajo de recolección de información.

2. **Identificación del marco de muestreo**. El marco de muestreo es una lista de todas las unidades de observación o individuos que pueden salir seleccionados durante el proceso de selección de la muestra. Dicho marco debe asemejarse lo más posible a la población objetivo para evitar dejar fuera elementos de la población con determinadas características particulares.

3. **Determinación del tamaño de la muestra**. La definición del tamaño de la muestra es decidir a cuántas unidades de observación o individuos serán consultados en la investigación. Para definir el tamaño de la muestra se debe realizar un cálculo a partir de un estadístico muestral.

Los pasos finales del plan de muestreo son: elección del método de selección de muestras, y definición del procedimiento de estimación y errores de muestreo.

2.1.4.1 Definición de la Muestra

Un **estimador** es un valor que permite estimar un parámetro desconocido de la población. La media de una muestra será el mejor estimador de la media μ de la población.

Para definir el tamaño de la muestra es necesario establecer el nivel de precisión deseado para la definición. El nivel de precisión está dado por: el **nivel de confianza** y el **error muestral**.

El **nivel de confianza** está relacionado a los intervalos de confianza. Los intervalos de confianza determinan cuán precisa será la media seleccionada en función al tamaño de la muestra. Si la muestra es pequeña, existe una alta probabilidad de que el valor de la media de la muestra no concuerde con el valor de la media de la población. El nivel de confianza se traduce en un **grado de confianza** que es una medida probabilística de certeza de que el intervalo de confianza contiene el parámetro de la población. El complemento del grado de confianza se simboliza con la letra griega alfa minúscula.

Los valores más utilizados del grado de confianza son:

a) 90%, con α = 100% - 90% = 10% = 0.10

b) 95%, con α = 100% - 95% = 5% = 0.05

c) 99%, con α = 100% - 99% = 1% = 0.01

El grado de confianza del 95%, el cual es el más usado en trabajos de investigación, indica que existe una probabilidad del 95% de que los

intervalos contengan al parámetro de la población. Complementariamente, existe una probabilidad del 5% de que los intervalos no contengan el parámetro de la población.

El **valor Z** es un indicador asociado al grado de confianza y representa a las desviaciones estándar asociadas a la distribución normal estándar.

$$Z = \frac{X - \mu}{\sigma}$$

Donde:

X es el valor a evaluar.

μ es la media de la distribución original.

σ es la desviación estándar de la misma distribución.

Tabla 1: Valor Z en la distribución normal estándar.
Fuente: Encuesta y Estadística (Blanco, 2011).

GRADO DE CONFIANZA	α	Z
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

La fórmula para determinar el tamaño de la muestra es la siguiente:

$$n = \frac{Z^2\ p\ (1 - p)}{e^2}$$

Donde:

Z es el valor asociado al grado de confianza.

p es la probabilidad de éxito en la muestra.

e es el error muestral permitido.

La diferencia entre la media obtenida de una muestra y la media de la población se denomina **error muestral** y se define como la máxima

diferencia probable, con una probabilidad del grado de confianza, entre la media de una muestra y la media de la población.

La fórmula para determinar el error muestral es la siguiente:

$$\mathrm{e} = Z\sqrt{\frac{p\ (1-p)}{n}}$$

Si se desconoce la probabilidad de éxito *p*, entonces se puede proporcionar un valor de manera convencional. Siendo que el tamaño de la muestra debe ser el más grande posible, valor del numerador de la fracción dentro de la raíz debe ser el más grande posible.

Está comprobado que el producto de la probabilidad de éxito por su complemento logra su valor máximo cuando p es del 50%. Se dice que esta es la forma más conservadora de determinar el tamaño de una muestra.

Tabla 2: Distribución de la probabilidad de éxito p.
Fuente: Encuesta y Estadística (Blanco, 2011).

p	(1 – p)	p (1 – p)
0.5	0.5	0.25
0.4	0.6	0.24
0.3	0.7	0.21
0.1	0.9	0.09
0.01	0.99	0.0099

Estadísticamente, las muestras de mayor tamaño llevan consigo mayores probabilidades de que los resultados obtenidos, a partir de ellas, concuerden con las características estudiadas de la población. En otras palabras, para un margen de error pequeño y un nivel de confianza alto, el tamaño de muestra debe ser grande.

2.2 Marco Referencial

2.2.1 Soluciones Relativas

Existen soluciones de software orientada a la Inteligencia de Negocios que brindan funcionalidades de Análisis de Sentimiento en Redes Sociales, a continuación, se describen algunas de ellas:

1. **Spagobi**[29] es una suite de Inteligencia de Negocios de código abierto que ofrece una amplia gama de herramientas analíticas para realizar informes, análisis multidimensionales, gestión de indicadores clave, visualización, cuadros de mandos, consultas e informes ad-hoc, inteligencia de ubicación y análisis de redes. Desde la versión 6.0, esta suite es conocida como **Knowage**[30] e incluye una edición empresarial licenciada.

2. **Power BI**[31] es un servicio licenciado de Análisis de Negocios que proporciona información detallada orientada a la toma de decisiones. Proporciona servicios para transformación de datos en imágenes, explora y analiza datos visualmente en la nube, presta recursos en para colaboración y personalización de informes interactivos.

3. **Pentaho**[32] agrupa soluciones licenciadas para Integración de Datos y Análisis de Negocios. Ofrece una arquitectura multiusuario que permite aplicaciones integradas de alta escalabilidad a través de interfaces web que permiten ejecutar procesos en la nube, análisis ad-hoc, visualizaciones interactivas, mapas en línea, aprendizaje automático y procesamiento de Big Data con Apache Spark.

[29] https://www.spagobi.org
[30] https://www.knowage-suite.com
[31] https://powerbi.microsoft.com
[32] https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform.html

4. **Tableau**[33] ofrece una serie de soluciones licenciadas para la visualización de datos interactivos orientados a la Inteligencia de Negocios. Las soluciones van desde opciones para escritorio y servidores, hasta integración con servicios en la nube con herramientas colaborativas, análisis de datos, descubrimiento de contenido, gobierno, preparación de datos, acceso y desarrollo. Ofrece también una versión gratuita para escritorio con limitadas funcionalidades.

Estas soluciones proporcionan herramientas para monitorear redes sociales, entre ellas Twitter. El usuario final puede monitorear los Tweets correspondientes a palabras clave o cuentas de usuario particulares y aplicar sobre ellos Análisis de Sentimiento.

Sin embargo, ninguna de ellas identifica o dimensiona muestras de poblaciones a nivel de un país y tampoco asegura la aplicación de Minería de Opinión con cuerpos léxicos del idioma castellano.

Además, con excepción de Spagobi, las técnicas y algoritmos de Análisis de Sentimiento de estas soluciones están guardadas detrás de servicios en la Web.

2.2.1.1 Soluciones OSINT

La recopilación, transformación y correlación de datos de distintas fuentes, utilizando herramientas interactivas de Minería de Datos, es conocida como Inteligencia de Código Abierto, *Open Source Intelligence* – OSINT en inglés.

[33] https://www.tableau.com

Las soluciones OSINT explotan los datos públicos disponibles en Internet y están orientadas a resolver aspectos de Seguridad de Tecnologías de la Información, Análisis Forense y Análisis de Relaciones.

Entre las soluciones OSINT más conocidas están **Maltego**[34], **Shodan**[35], **TheHarvester**[36], **Metagoofil**[37] y **Recon-NG**[38] por citar algunos[39].

Estas soluciones, dado su propósito, no están pensadas para realizar tareas de Análisis de Sentimiento.

2.2.2 Demografía Boliviana

Las proyecciones demográficas son estimaciones de la población futura, a corto y medio plazo. Estas proyecciones se hacen con base en el estudio de fenómenos demográficos y utilizando distintos los indicadores demográficos. Las cifras de población por edad y por sexo se proyectan bajo diversas hipótesis denominadas escenarios.

En Bolivia, el Instituto Nacional de Estadísticas – INE realiza proyecciones demográficas con base en los resultados obtenidos del último Censo de Población y Vivienda llevado a cabo el año 2012[40]. Una de esas proyecciones es la de población de ambos sexos según edad por departamento desde la gestión 2012 hasta la gestión 2020.

De acuerdo a las proyecciones de población del INE, la población entre 18 y 79 años de edad proyectada por departamento para los años 2014, 2017 y 2019 son:

---

[34] https://www.maltego.com
[35] https://www.shodan.io
[36] https://github.com/laramies/theHarvester
[37] http://www.osintux.org/documentacion/metagoofil
[38] http://www.osintux.org/documentacion/recon-ng
[39] https://securitytrails.com/blog/top-20-intel-tools
[40] https://www.ine.gob.bo/index.php/demografia/introduccion-2

Tabla 3: Proyecciones de población entre 18 y 79 años.
Fuente: Instituto Nacional de Estadísticas.

DEPARTAMENTO	PROYECCIÓN PARA EL 2014	PROYECCIÓN PARA EL 2017	PROYECCIÓN PARA EL 2019
BENI	243,483	260,709	272,590
CHUQUISACA	340,222	355,895	367,737
COCHABAMBA	1,101,750	1,171,113	1,219,728
LA PAZ	1,727,202	1,791,603	1,837,840
ORURO	307,012	323,362	335,493
PANDO	68,783	79,858	87,494
POTOSÍ	469,024	488,796	504,814
SANTA CRUZ	1,719,005	1,886,011	1,998,289
TARIJA	319,390	341,707	356,415
TOTAL BOLIVIA	**6,295,871**	**6,699,054**	**6,980,400**

En Bolivia, el voto es obligatorio para los ciudadanos a partir de los 18 años de edad. Aunque el Tribunal Supremo Electoral – TSE ha anunciado de que el voto no es obligatorio para los ciudadanos mayores de 70 años de edad, la población habilitada para emitir su voto incluye a todos los ciudadanos mayores de 18 años.

El Órgano Electoral Plurinacional – OEP tiene publicados los resultados de las Elecciones Generales de 2014[41] y las Elecciones Judiciales de 2017[42]. Las estadísticas del padrón electoral a nivel nacional para las elecciones nacionales 2019[43] fueron publicadas el 12 de septiembre de 2019.

[41] http://atlaselectoral.oep.org.bo/#/sub_proceso/17/1/1/graficos
[42] https://www.oep.org.bo/elecciones-judiciales-2017/
[43] https://www.oep.org.bo/elecciones-generales-2019/

Las poblaciones de ciudadanos inscritos y habilitados para emitir su voto por departamento para las elecciones nacionales llevadas a cabo el año 2014, para las elecciones judiciales llevadas a cabo el año 2017 y para las elecciones nacionales que se llevarán a cabo el presente año son:

Tabla 4: Población habilitada para votar en 2014, 2017 y 2019.

Fuente: Órgano Electoral Plurinacional.

DEPARTAMENTO	HABILITADOS 2014	HABILITADOS 2017	HABILITADOS 2019
BENI	223,598	240,509	265,586
CHUQUISACA	323,129	348,421	370,680
COCHABAMBA	1,128,351	1,216,294	1,325,896
LA PAZ	1,678,769	1,792,978	1,910,654
ORURO	293,576	314,747	335,777
PANDO	57,596	67,061	72,580
POTOSÍ	409,144	428,683	452,047
SANTA CRUZ	1,533,638	1,677,634	1,863,543
TARIJA	323,351	352,474	377,600
TOTAL BOLIVIA	**5,971,152**	**6,438,801**	**6,974,363**

El crecimiento de las poblaciones de ciudadanos habilitados para emitir su voto entre las gestiones 2014 y 2017, y entre las gestiones 2017 y 2019 es:

Tabla 5: Crecimiento de la Población Habilitada para Votar de 2014, 2017 y 2019.

Fuente: Elaboración propia.

DEPARTAMENTO	CRECIMIENTO 2014-2017	CRECIMIENTO 2017-2019
BENI	7.56%	10.43%
CHUQUISACA	7.83%	6.39%
COCHABAMBA	7.79%	9.01%
LA PAZ	6.80%	6.56%

DEPARTAMENTO	CRECIMIENTO 2014-2017	CRECIMIENTO 2017-2019
ORURO	7.21%	6.68%
PANDO	16.43%	8.23%
POTOSÍ	4.78%	5.45%
SANTA CRUZ	9.39%	11.08%
TARIJA	9.01%	7.13%
TOTAL BOLIVIA	**7.83%**	**8.32%**

2.2.3 Sondeos de Opinión

Los sondeos de opinión basados en encuestas han sido definidos como un método científico de recolección de datos basado en la teoría estadística y técnicas de muestreo. Este método es generalmente administrado por encuestadores entrenados o distribuidos a una muestra para su autoadministración. La recolección de datos se hace a través de formularios con cuestionarios estandarizados.

En la década de los años treinta, George Horace Gallup dio inicio al uso de encuestas a estudios de mercado y opinión pública. Gallup fue un matemático, estadístico y periodista norteamericano, quien fundó el *American Institute of Public Opinion*, el Instituto de Opinión Pública estadounidense.

En general, los sondeos son más simples y rápidos que las encuestas. Un sondeo consiste normalmente en una sola pregunta con respuesta de opción múltiple. Las encuestas normalmente tienen una cantidad mayor de preguntas que pueden ser abiertas.

En Bolivia, entre las empresas especializadas en sondeos de opinión más conocidas están: IPSOS Bolivia[44], CIES-MORI[45], Mercados y Muestras[46], Tal Cual[47] y Captura Consulting[48].

Considerando el tema de análisis de la versión demostrativa del modelo propuesto, entre abril y septiembre de este año, se hicieron sondeos de intención de voto respecto de los candidatos a la presidencia y la vicepresidencia del estado.

2.2.3.1 Sondeos de Intención de Voto

Entre el 25 de abril y 11 de septiembre de 2019 se publicaron resultados de 10 sondeos de intención de voto en Bolivia. A continuación, el resumen de resultados de dichos sondeos de opinión:

Resumen de sondeos de opinión del 25/04/2019 al 11/09/2019.

Fuente: Prensa nacional.

FECHA	HECHO POR/PARA	RECOLECCIÓN
25/04/2019	IPSOS Bolivia para la Red RTP[49].	Estudio a nivel nacional con recolección de datos hecho del 6 al 22 de abril en las ciudades capitales de departamento incluyendo El Alto.
28/04/2019	Mercados y Muestras para Página Siete[50].	Estudio a nivel nacional con recolección de datos hecho del 13 al 17 de julio en las ciudades capitales de departamento incluyendo El Alto y poblaciones intermedias.

[44] http://www.ipsos.com.bo
[45] https://www.ciesmori.com
[46] http://www.mymbolivia.com
[47] http://www.talcualbolivia.com
[48] https://www.capturaconsulting.com
[49] http://www.rtpbolivia.com.bo/2019/04/25/ipsos-para-rtp-morales-lidera-intencion-de-voto-con-33-mesa-tiene-un-25
[50] https://www.paginasiete.bo/nacional/2019/4/28/evo-obtiene-el-34-mesa-el-28-en-intencion-de-voto-216403.html

FECHA	HECHO POR/PARA	RECOLECCIÓN
19/05/2019	Tal Cual para La Razón[51].	Estudio a nivel nacional con recolección de datos hecho del 25 de abril al 12 de mayo en las ciudades capitales de departamento incluyendo El Alto, 17 ciudades intermedias y 31 poblaciones rurales.
21/07/2019	CIES-MORI para El Deber y la Red Unitel[52].	Estudio a nivel nacional con recolección de datos hecho del 12 al 18 de julio en las ciudades capitales de departamento y poblaciones rurales.
01/08/2019	Captura Consulting para la Red PAT y la revista Poder y Placer[53].	Estudio a nivel nacional con recolección de datos hecho del 15 al 29 de julio en las ciudades capitales del eje troncal del país incluyendo El Alto. La encuesta fue ampliada a las demás ciudades capitales y área rural del país.
04/08/2019	Mercados y Muestras para Página Siete, Los Tiempos y Asuntos Centrales[54].	Estudio a nivel nacional con recolección de datos hecho del 20 al 24 de julio en ciudades capitales de departamento incluyendo El Alto y poblaciones intermedias.
25/08/2019	IPSOS Bolivia para la Red RTP, La Razón, Opinión, La Patria y El Día[55].	Estudio a nivel nacional con recolección de datos hecho del 1 al 13 de agosto en ciudades capitales de departamento incluyendo El Alto y área rural.

[51] http://la-razon.com/nacional/encuesta-primer-urbano-rural-bolivia-ventaja-elecciones_0_3150284943.html
[52] https://www.eldeber.com.bo/bolivia/Oscar-Ortiz-saca-ventaja-a-Evo-y-a-Mesa-en-el-departamento-cruceno-20190722-0010.html
[53] http://redpatdigital.com/node/37177
[54] https://www.lostiempos.com/actualidad/pais/20190804/provincias-mantienen-adhesion-evo-morales
[55] https://www.la-razon.com/nacional/animal_electoral/encuesta-elecciones-bolivia-ciudades-evo-morales-carlos-mesa_0_3209079069.html

FECHA	HECHO POR/PARA	RECOLECCIÓN
01/09/2019	Mercados y Muestras para Página Siete[56].	Estudio a nivel nacional con recolección de datos hecho del 17 al 21 de agosto en ciudades capitales de departamento incluyendo El Alto y área rural.
08/09/2019	CIES-MORI para la Red Unitel[57].	Estudio a nivel nacional con recolección de datos hecho del 20 de agosto al 4 de septiembre.
11/09/2019	"Tu Voto Cuenta" de la UMSA, la Fundación Jubileo y varios medios de comunicación[58].	Estudio a nivel nacional con recolección de datos hecho del 31 de agosto al 2 de septiembre.

El resumen de las fichas técnicas de los sondeos de opinión es:

Tabla 6: Fichas técnicas de sondeos de opinión del 25/04/2019 al 11/09/2019.

Fuente: Prensa nacional.

FECHA	TAMAÑO DE LA MUESTRA	GRADO DE CONFIANZA	ERROR MUESTRAL
25/04/2019	2,000	95%	2.19%
28/04/2019	800	95%	3.47%
19/05/2019	2,250	95%	2.5%
21/07/2019	2,015	95%	2.2%
01/08/2019	900	95%	3.3%
04/08/2019	800	95%	3.47%
25/08/2019	2,000	95%	2.19%
01/09/2019	800	95%	3.47%

[56] https://www.paginasiete.bo/nacional/2019/9/1/si-hoy-fueran-las-elecciones-mesa-gana-en-ciudades-evo-en-provincias-229390.html

[57] https://www.unitel.tv/asi-decidimos-2019/

[58] https://www.bolpress.com/2019/09/11/mesa-acorta-distancia-al-mas-encuesta-perfila-segunda-vuelta/

FECHA	TAMAÑO DE LA MUESTRA	GRADO DE CONFIANZA	ERROR MUESTRAL
08/09/2019	2,221	95%	2.07%
11/09/2019	14,238	N/A	2.9%

A continuación, los detalles de los sondeos de intención de opinión publicados en la prensa nacional donde:

a) La intención de voto a favor de los actuales mandatarios es considerada como una opinión a favor de su reelección.

b) La intención de voto a favor de otros candidatos es considerada como una opinión en contra.

c) La intención de voto por ningún candidato, blanco o nulo, y "no sabe/no responde" es considerada como una opinión neutra.

Tabla 7: Resultados de sondeos de opinión del 25/04/2019 al 11/09/2019.

Fuente: Prensa nacional.

FECHA	A FAVOR	EN CONTRA	NEUTRA
25/04/2019	33%	45%	22%
28/04/2019	34%	42%	24%
19/05/2019	38%	46%	16%
21/07/2019	37%	42%	21%
01/08/2019	39%	37%	24%
04/08/2019	26%	48%	26%
25/08/2019	31%	47%	22%
01/09/2019	28%	49%	23%
08/09/2019	36%	43%	21%
11/09/2019	31%	40%	29%

La difusión del sondeo de opinión del 11/09/2019 fue suspendida por el Tribunal Supremo Electoral – TSE debido a que, según el TSE, no es clara la fuente de financiamiento del estudio entre otras observaciones[59]. Sin embargo, Televisión Universitaria – TVU difundió los resultados, los cuales fueron reproducidos por las redes sociales.

2.3 Marco Legal

2.3.1 Protección de Datos en Bolivia

De acuerdo a la Red Iberoamericana de Protección de Datos[60] existe un carácter consciente de que la protección de datos personales debe ser establecido como Derecho Fundamental de las personas. Esto se vio reflejado en el apoyo político dado en la Declaración Final de la XIII Cumbre de Jefes de Estado y de Gobierno de los países iberoamericanos celebrada en la ciudad de Santa Cruz de la Sierra el 15 de noviembre de 2003.

Sin embargo, a la fecha Bolivia no cuenta con una Ley de Protección de Datos.

Si bien existen leyes que otorgan derechos relativos a la protección de datos a las personas, éstos se aplican cuando las personas se ven afectadas y reclaman.

La Fundación InternetBolivia.org[61] ha establecido, a través de talleres y seminarios, la discusión sobre una norma para el uso y resguardo de datos personales en Bolivia.

[59] https://www.paginasiete.bo/nacional/2019/9/11/tse-frena-difusion-de-encuesta-exige-conocer-quien-la-financio-230583.html
[60] http://www.redipd.es
[61] https://internetbolivia.org

Para la promoción del debate sobre la importancia de una ley de protección de datos personales, la Fundación **InternetBolivia.org**, con la colaboración de **AccessNow.org**, elaboró una guía[62] que describe la problemática de la protección de datos personales en Bolivia y ofrece conceptos, principios y reglas que deberían tenerse en cuenta en su legislación.

Producto de estos esfuerzos, el 10 de mayo de 2019, la Fundación presentó el Anteproyecto de Ley de Protección de Datos Personales[63] a la Asamblea Legislativa Plurinacional. A la fecha, no se ha recibido respuesta por parte de la asamblea.

2.3.1.1 Constitución Política del Estado

La actual Constitución Política del Estado entró en vigencia el 7 de febrero de 2009, fecha en la cual fue publicada en la Gaceta Oficial de Bolivia. Fue aprobada en referéndum del 25 de enero de 2009.

En cuando a los derechos civiles de los ciudadanos, el numeral 2 del Artículo 21 establece que:

> *Las bolivianas y los bolivianos tienen derecho a la privacidad, intimidad, honra, honor, propia imagen y dignidad.*

En el marco de las Garantías Jurisdiccionales y Acciones de Defensa, para la Acción de Protección de Privacidad, el Artículo 130 establece:

I. Toda persona individual o colectiva que crea estar indebida o ilegalmente impedida de conocer, objetar u obtener la eliminación o rectificación de los datos registrados por cualquier medio físico, electrónico, magnético o informático, en archivos o bancos de datos

[62] https://www.accessnow.org/cms/assets/uploads/2019/03/Guia-Basica-Proteccion-de-Datos-Bolivia.pdf
[63] https://misdatos.internetbolivia.org/

públicos o privados, o que afecten a su derecho fundamental a la intimidad y privacidad personal o familiar, o a su propia imagen, honra y reputación, podrá interponer la Acción de Protección de Privacidad.

II. *La Acción de Protección de Privacidad no procederá para levantar el secreto en materia de prensa.*

Complementariamente, el Artículo 131 establece:

I. *La Acción de Protección de Privacidad tendrá lugar de acuerdo con el procedimiento previsto para la acción de Amparo Constitucional.*

II. *Si el tribunal o juez competente declara procedente la acción, ordenará la revelación, eliminación o rectificación de los datos cuyo registro fue impugnado.*

III. *La decisión se elevará, de ocio, en revisión ante el Tribunal Constitucional Plurinacional en el plazo de las veinticuatro horas siguientes a la emisión del fallo, sin que por ello se suspenda su ejecución.*

IV. *La decisión final que conceda la Acción de Protección de Privacidad será ejecutada inmediatamente y sin observación. En caso de resistencia se procederá de acuerdo con lo señalado en la Acción de Libertad. La autoridad judicial que no proceda conforme con lo dispuesto por este artículo quedará sujeta a las sanciones previstas por la ley.*

2.3.1.2 Ley General de Telecomunicaciones, TIC

De acuerdo a los numerales 6 y 17 del Artículo 54 de la Ley Nº 164 del 8 de agosto de 2011, las usuarias o los usuarios de los servicios de

telecomunicaciones y tecnologías de información y comunicación tienen derecho a:

6. *Exigir respeto a la privacidad e inviolabilidad de sus comunicaciones, salvo aquellos casos expresamente señalados por la Constitución Política del Estado y la Ley.*

17. *Recibir protección del proveedor del servicio sobre los datos personales contra la publicidad no autorizada por la usuaria o usuario, en el marco de la Constitución Política del Estado y la presente Ley.*

Respecto de la Inviolabilidad y Secreto de las Comunicaciones, el Artículo 56 establece que:

> *En el marco de lo establecido en la Constitución Política del Estado, los operadores de redes públicas y proveedores de servicios de telecomunicaciones y tecnologías de información y comunicación, deben garantizar la inviolabilidad y secreto de las comunicaciones, al igual que la protección de los datos personales y la intimidad de usuarias o usuarios, salvo los contemplados en guías telefónicas, facturas y otros establecidos por norma.*

El numeral 13 del Artículo 59 establece que entre las Obligaciones de los Operadores y Proveedores está:

> *Brindar protección sobre los datos personales evitando la divulgación no autorizada por las usuarias o usuarios, en el marco de la Constitución Política del Estado y la presente Ley.*

En cuanto al Correo Electrónico Personal, el Artículo 89 establece:

A los efectos de esta Ley el correo electrónico personal se equipara a la correspondencia postal, estando dentro del alcance de la inviolabilidad establecida en la Constitución Política del Estado. La protección del correo electrónico personal abarca su creación, transmisión, recepción y almacenamiento.

2.3.1.3 Ley de Ciudadanía Digital

La Ley Nº 1080 del 11 de julio de 2018, en su artículo 1, establece que la Ley tiene por objeto establecer las condiciones y responsabilidades para el acceso pleno y ejercicio de la ciudadanía digital en el Estado Plurinacional de Bolivia.

De acuerdo al Artículo 4:

I. *La ciudadanía digital consiste en el ejercicio de derechos y deberes a través del uso de tecnologías de información y comunicación en la interacción de las personas con las entidades públicas y privadas que presten servicios públicos delegados por el Estado.*

II. *El uso de los mecanismos de la ciudadanía digital implica que las instituciones mencionadas en el Parágrafo anterior, puedan prescindir de la presencia de la persona interesada y de la presentación de documentación física para la sustanciación del trámite o solicitud.*

En cuanto a la protección de datos personales, el Artículo 12 establece:

I. *Las y los servidores y funcionarios de las instituciones previstas en la presente Ley, utilizarán los datos personales y la información generada en la plataforma de interoperabilidad y ciudadanía digital únicamente para los fines establecidos en normativa vigente.*

II. *El incumplimiento de la anterior previsión, será sujeto a responsabilidad por la función pública; para el caso de instituciones privadas que presten servicios públicos delegados por el Estado, el ente que ejerza supervisión respecto a sus funciones deberá establecer los mecanismos pertinentes a fin de dar cumplimiento a esta norma.*

La Ley establece que la Agencia de Gobierno Electrónico y Tecnologías de Información y Comunicación – AGETIC es la entidad que establecerá y dirigirá los lineamientos y estándares técnicos a ser adoptados para la implementación de ciudadanía digital.

2.3.1.4 Ley del Código Penal

La Ley Nº 1768 del 10 de marzo de 1997, en su Artículo 363 ter., de la Alteración, acceso y uso indebido de datos informáticos, establece:

> *El que sin estar autorizado se apodere, acceda, utilice, modifique, suprima o inutilice, datos almacenados en una computadora o en cualquier soporte informático, ocasionando perjuicio al titular de la información, será sancionado con prestación de trabajo hasta un año o multa hasta doscientos días.*

2.3.1.5 DS 28168 Acceso a la Información

El Decreto Supremo Nº 28168 del 17 de mayo de 2005 tiene por objeto garantizar el acceso a la información, como derecho fundamental de toda persona y la transparencia en la gestión del Poder Ejecutivo.

En cuanto a la Petición de Habeas Data, el Artículo 19 establece que:

I. *Toda persona, en la vía administrativa, podrá solicitar ante la autoridad encargada de los archivos o registros la actualización,*

complementación, eliminación o rectificación de sus datos registrados por cualquier medio físico, electrónico, magnético o informático, relativos a sus derechos fundamentales a la identidad, intimidad, imagen y privacidad. En la misma vía, podrá solicitar a la autoridad superior competente el acceso a la información en caso de negativa injustificada por la autoridad encargada del registro o archivo público.

II. *La petición de Habeas Data se resolverá en el plazo máximo de cinco (5) días hábiles. En caso de negativa injustificada de acceso a la información, la autoridad jerárquica competente, adicionalmente tendrá un plazo de quince (15) días hábiles para proporcionar la información solicitada.*

III. *La petición de Habeas Data no reemplaza ni sustituye el Recurso Constitucional establecido en el Artículo 23 de la Constitución Política del Estado. El interesado podrá acudir, alternativamente, a la vía administrativa sin que su ejercicio conlleve renuncia o pérdida de la vía judicial. El acceso a la vía judicial no estará condicionado a la previa utilización ni agotamiento de esta vía administrativa.*

2.3.2 Acuerdo de Usuario de Twitter

Twitter es público y los Tweets pueden ser vistos y buscados por cualquier persona en todo el mundo. El usuario puede utilizar Twitter bajo un seudónimo si prefiere no utilizar su nombre.

El Acuerdo de Usuario de Twitter se compone de: Términos de Servicio[64], Política de Privacidad[65] y Reglas de Twitter[66] para los usuarios que residen

[64] https://twitter.com/es/tos
[65] https://twitter.com/es/privacy
[66] https://help.twitter.com/es/rules-and-policies

fuera de los Estados Unidos. La versión actual del acuerdo está vigente desde el 25 de mayo de 2018.

Las Reglas de Twitter tienen como objetivo garantizar que todas las personas puedan participar en la conversación pública de manera libre y segura.

2.3.2.1 Términos de Servicio

El usuario de Twitter proporciona información al hacer uso de los Servicios de Twitter.

El numeral 2 de Privacidad establece que el Usuario de Twitter, mediante el uso de los Servicios de Twitter, consiente la recopilación y uso de la información que proporciona, incluida su transferencia a los Estados Unidos, Irlanda o a otros países para su almacenamiento, procesamiento y uso por parte de Twitter y sus filiales.

La Política de Privacidad describe cómo es manejada la información proporcionada por los usuarios cuando hacen uso de los Servicios de Twitter.

2.3.2.2 Política de Privacidad

La mayor parte de la actividad en Twitter es pública, lo que incluye la información de perfil del usuario, su zona horaria e idioma, la fecha de creación de su cuenta y sus Tweets, así como cierta información de sus Tweets como la fecha, hora y la aplicación y versión de Twitter desde la que tuiteó.

El usuario puede decidir publicar su ubicación en sus Tweets o en su perfil de Twitter. Las listas que crea, la gente a la que sigue y que le sigue, así como los Tweets a los que hace “me gusta” o retuitea. La información

publicada sobre un usuario por otros usuarios que utilicen los servicios de Twitter también puede ser pública. Por ejemplo, etiquetas en fotos o menciones en Tweets.

El usuario es responsable de sus Tweets y otras informaciones que proporcione a través de los servicios de Twitter. Twitter establece que el usuario debe pensar con detenimiento lo que hace público, especialmente si se trata de información sensible. En función a la configuración que haga el usuario, Twitter también proporciona a ciertos terceros datos personales para facilitar la oferta o prestación de los servicios de Twitter.

Twitter conserva los datos de registro de los usuarios durante un máximo de 18 meses. Si se sigue el procedimiento de desactivación de cuentas, el nombre de cuenta, nombre de usuario y perfil público dejan de ser visibles en Twitter, luego de la desactivación se hace la eliminación. En caso de que un usuario desactive su cuenta accidentalmente o de forma improcedente, todavía será posible restaurarla durante los 30 días posteriores a la desactivación.

Twitter, además de proporcionar información pública al mundo directamente a través de sus aplicaciones, también utiliza tecnologías como Interfaces de Programación de Aplicaciones – API y otras integraciones para permitir que esta información pueda ser utilizada por sitios web, aplicaciones y otros. Por lo general, Twitter proporciona este contenido de forma gratuita en cantidades limitadas y cobra tarifas de licencia para el acceso a gran escala. Estos terceros no están asociados con Twitter y sus contenidos pueden no reflejar las actualizaciones que los usuarios hagan en Twitter.

Twitter cuenta con condiciones estándar que rigen cómo pueden utilizarse estos datos y un programa de cumplimiento para implementar dichas

condiciones. Estas condiciones están descritas en el Acuerdo de Desarrollador de Twitter.

2.3.3 Acuerdo de Desarrollador de Twitter

La versión actual del Acuerdo de Desarrollador de Twitter[67], *Developer Agreement and Policy* en inglés, está vigente desde el 25 de mayo de 2018 y sólo cuenta con una versión redactada en inglés.

El acuerdo está compuesto de las siguientes secciones:

I. API de Twitter y contenido de Twitter, con los términos y definiciones usados en el acuerdo.

II. Restricciones en el uso de materiales con licencia, con las limitaciones en cuanto a la aplicación de ingeniería inversa, límites de uso de las API de Twitter, información geográfica, uso de marcas de Twitter y aspectos de seguridad.

III. Modificaciones, con la aclaración de que Twitter puede modificar sus API bajo su criterio.

IV. Propiedad y retroalimentación, con las definiciones de derechos otorgados en el acuerdo.

V. Terminación, con las condiciones para rescisión o suspensión del acuerdo.

VI. Confidencialidad, en cuanto a la información patentada por Twitter.

VII. Otros términos importantes.

[67] https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

El numeral 1 de la sección VII contiene las condiciones relevantes en cuanto a la Protección de los datos del Usuario. Este numeral establece que el Contenido de Twitter y la información derivada del Contenido de Twitter no pueden ser utilizados, mostrados, distribuidos o puestos a disposición de cualquier manera a:

1. Cualquier entidad del sector público, o cualquier otra entidad que brinde servicios a entidades del sector público, con fines de vigilancia, incluidos:

    a. Investigar o rastrear a los usuarios de Twitter o su contenido en Twitter.

    b. Hacer seguimiento, alertar u otro monitoreo de eventos delicados como protestas, manifestaciones o reuniones de organización comunitaria.

2. Cualquier entidad del sector público, o cualquier otra entidad que brinde servicios a entidades del sector público, cuya función o misión principal incluya realizar vigilancia o recopilar inteligencia.

3. Cualquier entidad con el propósito de llevar a cabo o proporcionar vigilancia, análisis o investigación que aísle a un grupo de individuos o cualquier individuo para cualquier propósito ilegal o discriminatorio o de una manera que sea inconsistente con las expectativas razonables de privacidad de los usuarios de Twitter.

4. Cualquier entidad para identificar, segmentar o perfilar a individuos en función de su salud, estado o condición financiera, afiliación política o creencias, origen racial o étnico, afiliación o creencias religiosas o filosóficas, vida sexual u orientación sexual, afiliación comercial, datos relacionados con cualquier comisión presunta o real

de un delito, o cualquier otra categoría sensible de información personal prohibida por ley.

5. Cualquier entidad que se crea razonablemente usará dichos datos para violar la Declaración Universal de Derechos Humanos[68], incluidos, entre otros, los artículos:

   Art. 12. Nadie será objeto de injerencias arbitrarias en su vida privada, su familia, su domicilio o su correspondencia, ni de ataques a su honra o a su reputación. Toda persona tiene derecho a la protección de la ley contra tales injerencias o ataques.

   Art. 18. Toda persona tiene derecho a la libertad de pensamiento, de conciencia y de religión; este derecho incluye la libertad de cambiar de religión o de creencia, así como la libertad de manifestar su religión o su creencia, individual y colectivamente, tanto en público como en privado, por la enseñanza, la práctica, el culto y la observancia.

   Art. 19. Todo individuo tiene derecho a la libertad de opinión y de expresión; este derecho incluye el no ser molestado a causa de sus opiniones, el de investigar y recibir informaciones y opiniones, y el de difundirlas, sin limitación de fronteras, por cualquier medio de expresión.

Si el personal de las fuerzas del orden público solicita información sobre Twitter o sus usuarios para fines de una investigación en curso, el

[68] https://www.un.org/es/universal-declaration-human-rights/index.html

desarrollador de Twitter debe remitirlos a las Guías para la Aplicación de la Ley de Twitter[69].

El acuerdo concluye con la Política del Desarrollador, *Developer Policy* en inglés, el cual está compuesto de 2 secciones: Principios guía y Reglas para servicios o características específicas de Twitter.

[69] https://t.co/le

CAPITULO III. MODELO PROPUESTO

Conforme lo expuesto en el apartado “2.1.1.2 Análisis de Sentimiento”, el Análisis de Sentimiento en las redes sociales enfrenta múltiples problemas adicionales a los del Procesamiento del Lenguaje Natural tradicional: ironía, sarcasmo, extranjerismos, expresiones regionales, ruido, abreviaturas no estándares, emoticones y memes por citar algunos.

Actualmente existen herramientas para Procesamiento del Lenguaje Natural en la Web. Sin embargo, la mayoría de ellas tienen implementaciones de cuerpos léxicos del idioma inglés y las pocas implementaciones del idioma español no incluyen el castellano boliviano.

El modelo propuesto plantea una alternativa para ejecutar la tarea de Clasificación de Polaridad del Análisis de Sentimiento en las opiniones expresadas en el texto de los Tweets publicados por usuarios bolivianos con un enfoque distinto al adoptado por las soluciones descritas en el apartado “2.2.1 Soluciones Relativas”.

3.1 Twitter

En términos de sus autores[70], “Twitter es lo que está pasando en el mundo y los temas sobre los que está hablando la gente.”

Twitter, como la Fuente de Datos del presente modelo, es la red social de micro-blogueo, *microblogging* en inglés, más popular empleado en la actualidad. Su sencillez y rapidez la han convertido en uno de los medios de comunicación más efectivos entre el común de las personas, periodistas, activistas, autoridades, políticos, celebridades, empresas, instituciones públicas y marcas.

[70] https://about.twitter.com/es.html

De acuerdo a **Statista.com**[71], Twitter ocupa la posición número 12 en el ranking de las redes sociales más populares en el mundo a julio del 2019.

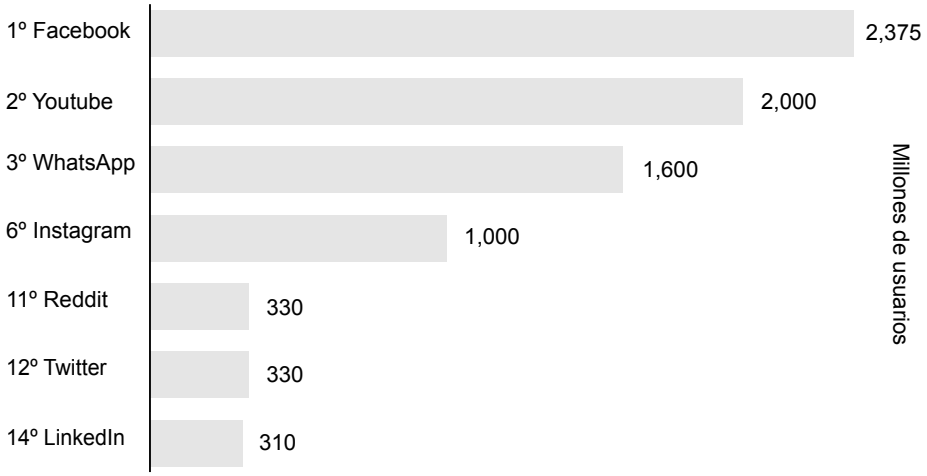

Figura 13: Ranking mundial de redes sociales a julio de 2019.

Fuente: Statista.com

Según la misma fuente, el crecimiento de la red social continuará en ascenso. El pronóstico de crecimiento en Latinoamérica[72] establece que Twitter tendrá 101.7 millones de usuarios el año 2020.

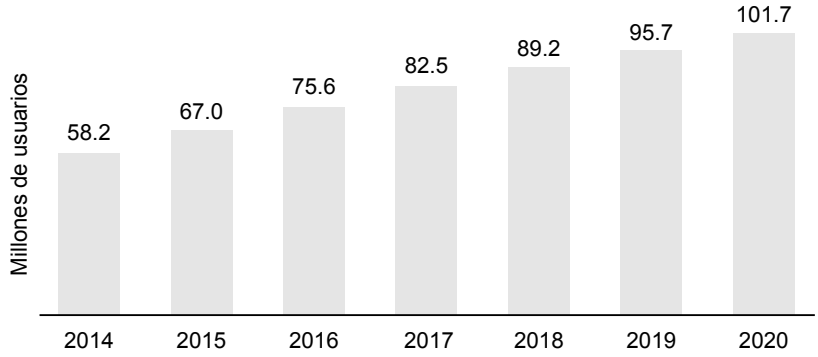

Figura 14: Crecimiento y pronóstico de usuarios de Twitter.

Fuente: Statista.com

De acuerdo a la Agencia de Gobierno Electrónico y Tecnologías de Información y Comunicación – AGETIC[73] el 17% de las personas mayores de 14 años de edad con acceso a Internet utiliza Twitter en Bolivia.

[71] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[72] https://www.statista.com/statistics/303923/twitter-users-latin-america/
[73] https://agetic.gob.bo/pdf/estadotic/AGETIC-Estado-TIC.pdf

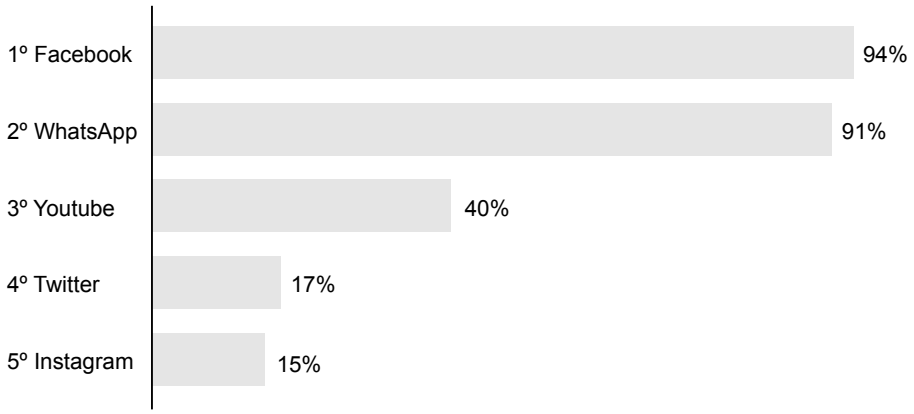

Figura 15: Redes sociales más utilizadas en Bolivia.

Fuente: AGETIC.

El carácter público de la información publicada en Twitter favorece su explotación. Dada la naturaleza del presente trabajo, la mejor opción para la aplicación de Minería de Opinión es la red social Twitter.

3.2 Capas del Modelo

Los componentes del modelo propuesto pueden ser distribuidos en cuatro capas, empezando en el lugar del cual se encuentran los datos que servirán al modelo y terminando en el análisis de resultados obtenidos a partir del procesamiento de los datos recuperados de la fuente de datos.

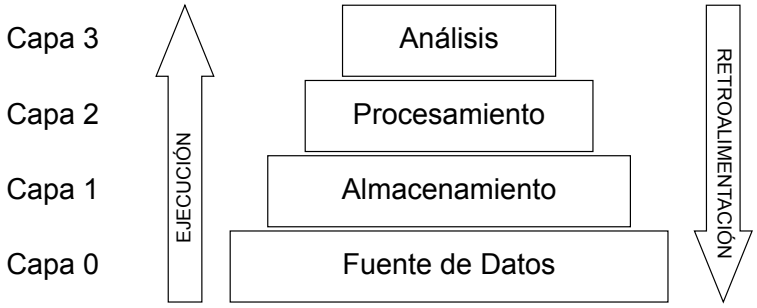

Figura 16: Esquema de las capas del modelo propuesto.

Fuente: Elaboración propia.

1. **Capa de Fuente de Datos**. Se definen los datos que serán extraídos de la Fuente de Datos. Para este propósito se debe analizar la información que es proporcionada por la Fuente de Datos y las

interfaces por la cuales esta información es puesta a disposición de terceros. Es posible que se deban prever estrategias de transformación de los datos extraídos para que puedan ser almacenados en un componente de la Capa de Almacenamiento. En este caso, los datos extraídos dejarán de ser nativos, dependiendo del nivel de transformación pueden ser desde cuasi-nativos hasta pre-procesados.

2. **Capa de Almacenamiento**. Se definen las estructuras de la base de datos que servirán para el almacenamiento de los datos extraídos de la Fuente de Datos.

   La definición de estas estructuras debe hacerse teniendo en mente el volumen de los datos posibles de extraer y la versatilidad de la base de datos para satisfacer los requerimientos de los componentes de la Capa de Procesamiento. Se definen también las estrategias para los procesos de Extracción, Transformación y Carga, *Extract, Transform and Load* – ETL en inglés, que se seguirán para conectar a la Capa de Fuente de Datos.

3. **Capa de Procesamiento**. Se definen las estrategias para el procesamiento de los datos almacenados en cuanto a: selección, depuración, traducción, validación y clasificación.

4. **Capa de Análisis**. Se definen las estrategias para obtener resultados producto del análisis de los datos procesados en función a los objetivos del modelo aplicado. El análisis de datos puede hacerse también sobre los datos almacenados.

Las estrategias definidas en cada capa no son definitivas, los resultados obtenibles están abiertos al conocimiento de quienes ejecutan el modelo. El modelo puede ser ajustado o redefinido producto de la retroalimentación en cada ejecución.

3.3 Interacción e Integración con Twitter

A continuación, se describen las capacidades de interacción e integración a de la fuente de datos que representa Twitter a través de sus API.

Para la interacción e integración con Twitter existen tres familias de API:

1. Las **API estándar** consisten en API REST de consulta y de transmisión/recepción en línea. Estas API son de uso gratuito, con límites de velocidad en su uso.

2. Las **API empresariales** consisten en API con mayores capacidades, filtros, búsqueda histórica, herramientas de análisis de datos más profundos y aplicaciones empresariales adicionales. El uso de estas estas API tiene costo y se paga por suscripción.

3. Las **API especiales** consisten en versiones confiables y asequibles de API empresariales. El costo de estas API tiene costo y se paga por uso.

De acuerdo a Twitter, las API estándar gratuitas son excelentes para comenzar, probar una integración o validar un concepto. **El modelo propuesto plantea el uso de las API estándar.**

La documentación de las API de Twitter[74] está organizada en secciones. Las secciones donde se encuentran los métodos relevantes para el modelo son:

1. En la sección de Cuentas y Usuarios (*Accounts and users*) se encuentran los métodos para Seguimiento, búsqueda y recolección (*Follow, search, and get users*):
    a. Método *GET users/show*
    b. Método *GET followers/list*

[74] https://developer.twitter.com/en/docs/api-reference-index

c. Método *GET friends/list*

2. En la sección de Tweets se encuentran los Objetos relativos a Tweets (*Tweet objects*):
    a. Objeto Usuario (*User object*)
    b. Objeto Tweet (*Tweet object*)

El método *GET users/show* es empleado para obtener los metadatos de cuentas semilla. Las cuentas de seguidores y amigos son obtenidas a través de los métodos *GET followers/list* y *GET friends/list* respectivamente.

El método *GET statuses/user_timeline* es empleado para obtener los Tweets publicados por los usuarios de Twitter.

Los métodos *GET followers/ids* y *GET friends/ids* son descartados del proceso debido a que tienen la misma tasa de transferencia que los métodos *GET followers/list* y *GET friends/list* respectivamente y, tal como se describe en la documentación de las API de Twitter, estos métodos sólo entregan identificadores de usuarios.

Los datos proporcionados por los métodos de las API de Twitter están formados por objetos de datos y sus atributos representados en formato JSON (*JavaScript Object Notation*).

El detalle de los métodos y objetos citados anteriormente se encuentra en el "Anexo I. Relevamiento de las API de Twitter".

Adicionalmente, la documentación de las API de Twitter incluye también información de los métodos de búsqueda de Tweets[75]:

a. Búsqueda especial (*Premium search*)

[75] https://developer.twitter.com/en/docs/tweets/search/overview

b. Búsqueda empresarial (*Enterprise search*)

c. Búsqueda estándar (*Standard search*)

Las soluciones descritas en el apartado “2.2.1 Soluciones Relativas" hacen uso de estos métodos de búsqueda. **El modelo propuesto no usa ninguno de estos métodos, esta es la principal diferencia del modelo con esas soluciones**.

3.3.1 Tasas de Transferencia

El uso de las API estándar tiene una limitación de velocidad que se realiza por usuario, dicho con mayor precisión, por *token* de acceso de usuario. Si un método permite 15 solicitudes por ventana de límite de velocidad, entonces permite 15 solicitudes por ventana por *token* de acceso.

Las tasas de transferencia, se dividen en intervalos de 15 minutos, cada intervalo es una ventana. Todos los métodos requieren autenticación, por lo que no existe invocaciones de métodos sin una autenticación previa.

A continuación, se encuentran los límites por ventana[76], por hora y por día de los métodos descritos anteriormente.

Tabla 8: Tasas de transferencia de los métodos de las API de Twitter.

Fuente: Documentación de API de Twitter.

MÉTODO	SOLICITUDES POR 15 MIN	SOLICITUDES POR HORA	SOLICITUDES POR DÍA
GET users/show	900	3,600	86,400
GET followers/list GET followers/ids	15	60	1,440
GET friends/list GET friends/ids	15	60	1,440

[76] https://developer.twitter.com/en/docs/basics/rate-limits

MÉTODO	SOLICITUDES POR 15 MIN	SOLICITUDES POR HORA	SOLICITUDES POR DÍA
GET statuses/user_timeline	900	3,600	86,400

Estos límites son máximos y referenciales. En aplicaciones reales, las cantidades de solicitudes atendidas por ventana pueden ser menores a las indicadas.

3.4 Selección de Recursos Tecnológicos

Entre los componentes del modelo propuesto se encuentran los recursos tecnológicos requeridos para su aplicación en una versión demostrativa.

Estos recursos son herramientas propias de plataformas para Análisis, Ciencia de Datos y Aprendizaje Automático.

A continuación, se presenta el último *ranking* internacional del portal especializado en Descubrimiento de Conocimiento KDNuggets.com[77] para herramientas para Análisis, Ciencia de Datos y Aprendizaje Automático.

Tabla 9: Ranking de herramientas para Ciencias de Datos del año 2018

Fuente: KDNuggets.com

HERRAMIENTA	PARTICIPACIÓN 2018	CAMBIO 2017 - 2018
Python	65.6%	11%
RapidMiner	52.7%	65%
R	48.5%	-14%
SQL	39.6%	1%
Excel	39.1%	24%

Python[78] es un lenguaje de programación interpretado que se ejecuta en tiempo real. Su curva de aprendizaje es lineal y el hecho de que sea un

[77] https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html
[78] https://www.python.org

lenguaje de alto nivel asegura su comprensión por parte de analistas que no sean desarrolladores. Además, su gran comunidad asegura la disponibilidad de soporte gratuito y una amplia variedad de librerías para implementar procesos para ETL, Procesamiento del Lenguaje Natural y Análisis de Negocios.

RapidMiner[79] es una solución de software desarrollada para análisis de datos mediante el encadenamiento de procesos predefinidos a través de un entorno gráfico. La versión gratuita del cliente permite la obtención de hasta 10,000 filas de datos, con 1 procesador lógico en 30 días de prueba. Los planes de uso licenciado inician en 5,000 dólares anuales. Las limitaciones de la versión gratuita lo descartan como la herramienta para el modelo propuesto.

R[80] es un entorno y lenguaje de programación orientado al análisis de datos y estadísticas para proyectos académicos y de Investigación y Desarrollo. Desafortunadamente, su curva de aprendizaje es más lenta que Python y presenta dificultades sobre todo al inicio de su uso.

Considerando la popularidad de Python, su curva de aprendizaje y el hecho de que cuenta con librerías para Procesamiento del Lenguaje Natural lo convierten en la herramienta adecuada para la aplicación de tareas de Análisis de Sentimiento entre otras.

En consecuencia, siguiendo el esquema de capas descrito en el apartado “3.2 Capas del Modelo”, los recursos tecnológicos intangibles están distribuidos de la siguiente forma:

[79] https://rapidminer.com
[80] https://www.r-project.org

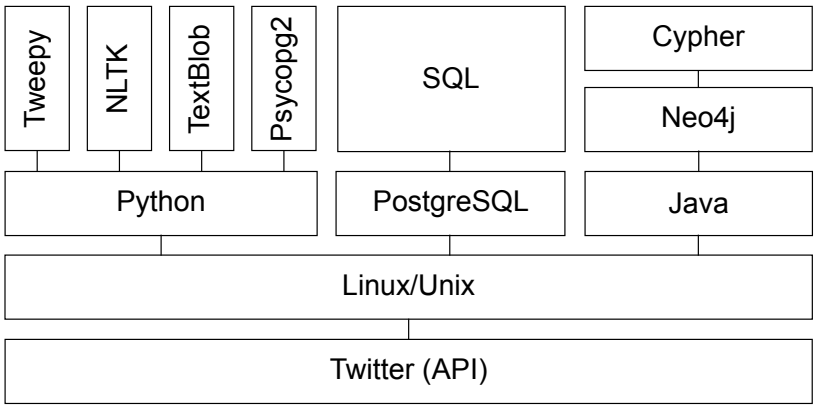

Figura 17: Esquema de recursos tecnológicos sugeridos para el modelo.

Fuente: Elaboración propia.

1. En la **Capa de Fuente de Datos** está Twitter a través de sus API[81]. Conforme se describe en el apartado "2.3 Marco Legal", no existen impedimentos en el uso de la información proporcionada por las API de Twitter fundamentalmente porque sus usuarios consienten la recopilación y uso de la información que proporcionan al hacer uso de los Servicios de Twitter.

2. En la **Capa de Almacenamiento** están:

    a. El gestor de bases de datos PostgreSQL.

    b. El intérprete del lenguaje Python[82], la librería Tweepy[83] para acceder a las API de Twitter y la librería Psycopg2[84] para conexión con bases de datos PostgreSQL.

    PostgreSQL[85] ocupa el cuarto lugar del *ranking* internacional de DB-Engines.com[86] de sistemas de administración de bases de datos relacionales. PostgreSQL está después de Oracle, MySQL

[81] https://developer.twitter.com/en/docs/api-reference-index
[82] https://www.python.org/
[83] https://www.tweepy.org/
[84] https://pypi.org/project/psycopg2/
[85] https://www.postgresql.org/
[86] https://db-engines.com/en/ranking/relational+dbms

y SQL Server en el *ranking*. Es el sistema de código abierto más usado en entornos MacOS.

Neo4j[87] encabeza el *ranking* internacional de DB-Engines.com[88] de sistemas de administración de bases de datos de grafos según su popularidad.

3. En la **Capa de Procesamiento** están:
    a. El intérprete del lenguaje Python, la librería Psycopg2 para conexión con bases de datos PostgreSQL, la librería NLTK[89] del Kit de Herramientas para Procesamiento del Lenguaje Natural, y la librería TextBlob[90] para Procesamiento del Lenguaje Natural.
    b. El gestor de bases de datos relacionales PostgreSQL y el lenguaje de consultas SQL[91].
    c. El gestor de bases de datos de grafos Neo4j, el cual está basado en Java, y el lenguaje de consulta de grafos Cypher[92].
4. En la **Capa de Análisis** está el gestor de bases de datos relacionales PostgreSQL y el lenguaje de consultas SQL.

Adicionalmente, como herramienta para desarrollo Python y uso de sus módulos se tiene al IDE PyCharm[93], y como cliente SQL y herramienta de administración de bases de datos se tiene a la aplicación DBeaver[94].

[87] https://neo4j.com/
[88] https://db-engines.com/en/ranking/graph+dbms
[89] https://www.nltk.org/
[90] https://textblob.readthedocs.io/en/dev/
[91] https://www.postgresql.org/docs/current/sql.html
[92] https://neo4j.com/docs/cypher-manual/current/
[93] https://www.jetbrains.com/pycharm/
[94] https://dbeaver.io/

La disponibilidad de los recursos tecnológicos descritos asegura la factibilidad técnica del modelo propuesto para una versión demostrativa.

3.4.1 Procesos ETL

El concepto de Extracción, Transformación y Carga de datos, o ETL por sus siglas en inglés, ganó popularidad en la década de años 70 cuando las organizaciones empezaron a usar múltiples repositorios de datos para almacenar diferentes tipos de información[95].

ETL hace referencia al proceso, no a las soluciones de software que implementan el proceso.

a) Para la Extracción, Python cuenta con la librería Tweepy[96] para acceder a la API de Twitter.

b) Para la Transformación, Python cuenta con 47 módulos para ejecutar transformaciones de textos multipropósito[97].

c) Para la Carga de datos, Python cuenta con la librería Psycopg2[98] para conexiones con bases de datos PostgreSQL.

La documentación de la librería Tweepy[99] y de la librería Psycopg2[100] contiene las rutinas recomendadas para acceder a la API de Twitter y para conectar a bases de datos PostgreSQL respectivamente.

A continuación, el emparejamiento de los métodos de las API de Twitter con los métodos de la librería Tweepy.

---

[95] https://www.sas.com/en_us/insights/data-management/what-is-etl.html
[96] https://www.tweepy.org
[97] https://docs.python.org/3/whatsnew/3.8.html#improved-modules
[98] https://pypi.org/project/psycopg2
[99] https://tweepy.readthedocs.io/en/latest/getting_started.html
[100] http://initd.org/psycopg/docs/usage.html

Emparejamiento de métodos de API de Twitter con Tweepy.

Fuente: Elaboración propia.

MÉTODO API DE TWITTER	MÉTODO DE TWEEPY
GET users/show	API.get_user
GET followers/list	API.followers
GET friends/list	API.friends
GET statuses/user_timeline	API.user_timeline

El “Capítulo IV. Aplicación del Modelo” describe de forma detallada el uso de las librerías Tweepy y Psycopg2 en *scripts* Python.

3.5 Diseño del Modelo

3.5.1 Tema de Análisis

El Tema de Análisis es la razón de ser de la aplicación del modelo y sintetiza las preguntas de las que se necesitan respuestas.

En el marco de la aplicación del modelo, se requiere la definición de palabras clave, o *keywords*, las cuales serán utilizadas como puntos de referencia para encontrar las opiniones relativas al Tema de Análisis.

Las palabras clave son los Hashtags que encabezan listas de Temas de Tendencia de Twitter, *Trending Topic* en inglés. La definición de las palabras clave depende del conocimiento de quienes ejecutan el modelo en cuanto a la red social Twitter en el contexto nacional y su comprensión del Tema de Análisis.

3.5.2 Datos Fuente

Los datos fuente son los requeridos para el modelo y surgen del relevamiento descrito en el “Anexo I. Relevamiento de las API de Twitter”:

1. Del **Objeto Usuario** (*user-object*) los datos que deben ser recolectados son:

Datos recolectados del objeto Usuario a través de las API de Twitter.

Fuente: Elaboración propia.

DATO	TIPO	DESCRIPCIÓN
id_str	String	La representación de cadena del identificador único del usuario.
screen_name	String	El nombre de pantalla, el identificador o el alias con el que el usuario se identifica.
created_at	String	La fecha y hora UTC en que se creó la cuenta de usuario en Twitter.
location	String	La ubicación definida por el usuario para el perfil de esta cuenta. No necesariamente es una ubicación, ni es analizable por máquina.
followers_count	Int	El número de seguidores que el usuario tiene actualmente.
friends_count	Int	El número de usuarios que el usuario está siguiendo, también conocidos como sus "amigos".
protected	Boolean	Indicador que señala si el usuario ha elegido proteger sus Tweets.

2. Del **Objeto Tweet** (*tweet-object*) los datos que deben ser recolectados son:

Datos recolectados del objeto Tweet a través de las API de Twitter.

Fuente: Elaboración propia.

DATO	TIPO	DESCRIPCIÓN
id_str	String	La representación de cadena del identificador único para el Tweet.
created_at	String	Hora UTC cuando se creó el Tweet.
source	String	Aplicación utilizada por el usuario para publicar el Tweet, como una cadena con formato HTML.

DATO	TIPO	DESCRIPCIÓN
coordinates	Coordinates	Representa la ubicación geográfica del Tweet según lo informado por el usuario o la aplicación empleada.
text	String	El texto real del Tweet en formato UTF-8.

3.5.3 Almacenamiento de Datos

El “Anexo II. Diseño de la Base de Datos” contiene la definición del modelo entidad-relación de la base de datos que servirá para el almacenamiento de los datos recolectados. El Anexo contiene también la descripción de los campos de cada una las tablas que representan a las entidades definidas.

Las entidades de la base de datos son:

1. Entidad Semilla, contiene los datos de las cuentas semilla identificadas en la etapa 1 del flujo de datos.
2. Entidad Usuario, contiene los metadatos de cuentas de usuario recuperadas en la etapa 3 del flujo de datos.
3. Entidad Relación, contiene la relación entre usuarios de Twitter que se recuperan en la etapa 3 del flujo de datos. La relación está definida como: el usuario X sigue al usuario Y.
4. Entidad Tweet, contiene los datos de Tweets recuperados en la etapa 5 del flujo de datos.
5. Entidad Sentimiento, contiene los resultados del análisis de sentimiento de Tweets relativos al tema de análisis obtenidos en la etapa 7 del flujo de datos.
6. Entidad Tema, contiene las palabras claves relativas al tema de análisis del modelo que serán empleadas en la etapa 6 del flujo de datos.

7. Entidad Hito, contiene datos de eventos de referencia relativos al tema de análisis del modelo. Estos eventos sirven para realizar comparaciones entre resultados obtenidos a través de la aplicación del modelo y otras fuentes.

3.5.4 Flujo de Datos

El flujo de datos es la columna vertebral del modelo y describe las etapas que deben seguirse para su aplicación.

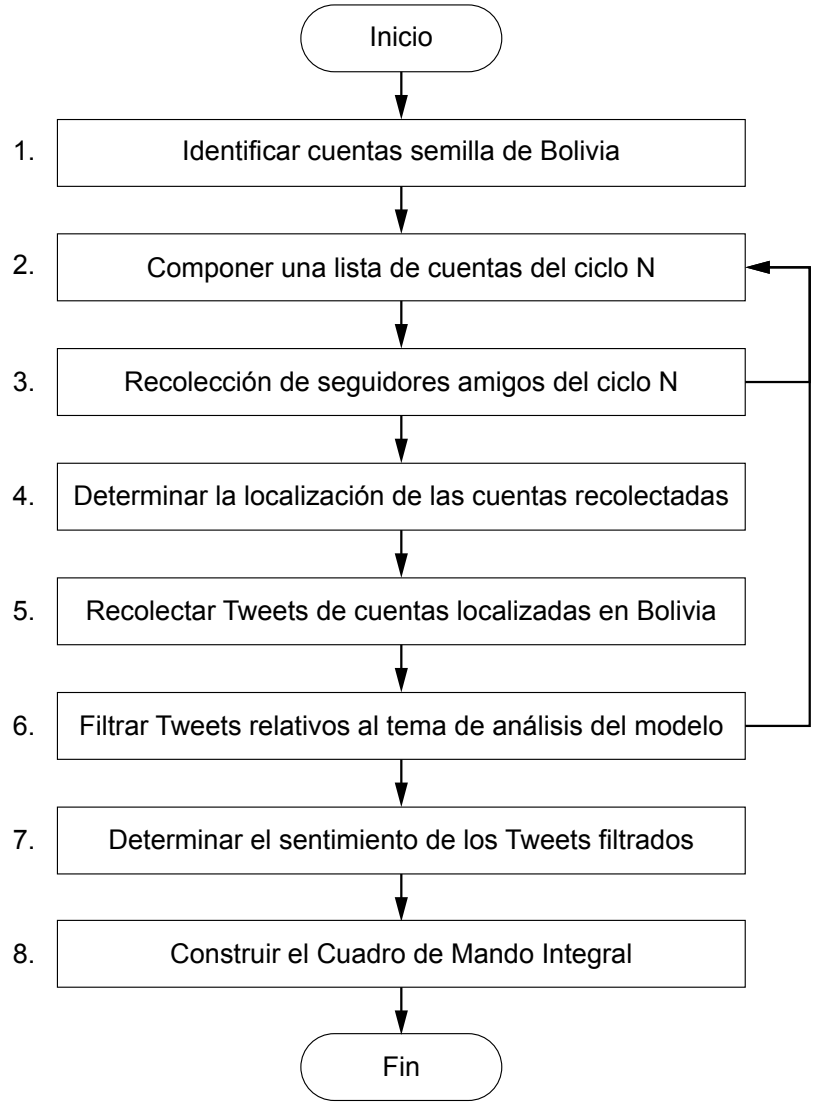

Figura 18: Flujo de datos del modelo propuesto.

Fuente: Elaboración propia.

A continuación, se describen las etapas del flujo de datos:

1. El punto de partida de la recolección de datos es la identificación de cuentas semilla (*seed-accounts*). Una cuenta semilla es una cuenta popular en Bolivia con una gran cantidad de seguidores.

   Existen distintas fuentes disponibles en Internet que proveen información de uso en redes sociales con propósitos de mercadeo para grandes y pequeñas empresas.

2. Para la recolección de cuentas de usuario se debe compilar una lista de cuentas de las que se obtendrán las cuentas que les siguen y las cuentas a quienes siguen.

   Se debe obtener los metadatos de las cuentas semilla en Twitter. La lista de cuentas semilla comprenderá las cuentas del ciclo 1 del proceso de recolección de cuentas. Los metadatos son recuperados a través del método "*GET users/show*" de las API de Twitter.

3. Una vez compilada la lista de cuentas del ciclo 1, se deben recuperar las cuentas de los seguidores y amigos de cada una de las cuentas de este ciclo.

   Este proceso puede repetirse N veces, donde las cuentas de un ciclo N son de la combinación de los seguidores y amigos de un ciclo N-1. El criterio para determinar N está sujeto a la cantidad de cuentas esperadas, la cantidad de cuentas nuevas recuperadas en cada ciclo y al tiempo destinado para la recolección de cuentas.

   Las cuentas de seguidores y amigos son recuperadas a través de los métodos "*GET followers/list*" y "*GET friends/list*" respectivamente de las API de Twitter.

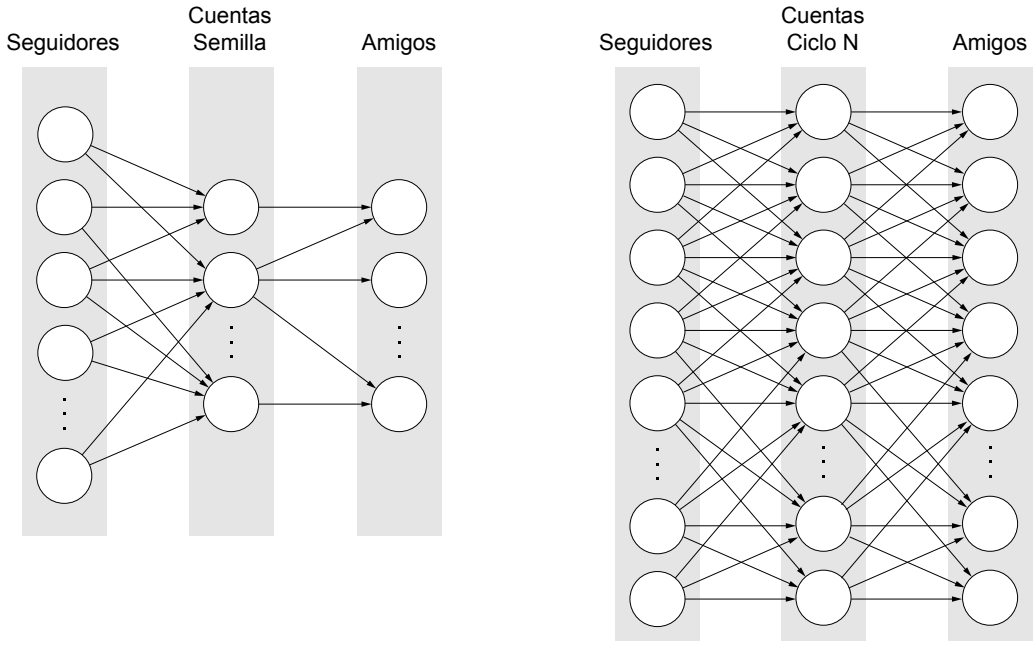

Figura 19: Ilustración de recolección de seguidores y amigos.

Fuente: Elaboración propia.

4. El paso siguiente es determinar la localización de los usuarios de todas las cuentas recuperadas en todos los ciclos. Esto se debe hacer en dos partes:

    a. Determinar la localización de los usuarios a partir de los metadatos de sus cuentas. Por ejemplo: el usuario con identificador “19514282” publicó que su localización es “La Paz - Bolivia”. En esta etapa, se deben excluir las cuentas cuya localización señale puntos fuera del territorio nacional.

    Se debe tomar en cuenta que la localización se obtiene del atributo *location* del objeto Usuario, el cual contiene la ubicación definida por el usuario para el perfil de su cuenta. Este valor no necesariamente es una ubicación, ni es analizable por máquina.

    b. Deducir la localización de los usuarios que no hubieran publicado una localización, cuando el atributo *location* no tiene valor, o que

la localización sea genérica. Por ejemplo: el usuario con identificador “242925003” publicó que su localización es “Bolivia”.

La deducción de la localización se hará a través de la detección de comunidades en grafos. Para el efecto, se deben cargar todas las cuentas localizadas a través de sus metadatos, las cuentas sin localización y las cuentas con localización genérica, en una base de datos de grafos y aplicar sobre ellas algoritmos de detección de comunidades tomando como referencia las localidades ya detectadas.

Esta deducción es una aproximación que parte del supuesto de si un usuario, que no hubiera declarado su localización, sigue a una gran cantidad de usuarios de una determinada localidad, entonces existe una alta probabilidad de que el usuario pertenezca a la misma localidad.

La localización es equivalente a cualquiera de los nueve departamentos de Bolivia: Beni, Chuquisaca, Cochabamba, La Paz, Oruro, Pando, Potosí, Santa Cruz y Tarija.

5. Habiéndose identificado las cuentas de usuario localizadas en Bolivia, lo que sigue es la recuperación de los últimos Tweets que cada usuario hubiera publicado.

   Los Tweets son recuperados a través del método “*GET statuses/user_timeline*” de las API de Twitter.

6. El paso siguiente, de filtrado de Tweets, consiste en identificar a los Tweets que guarden relación con el tema de análisis.

El tema de análisis es definido a partir de un conjunto de palabras claves, o *keywords*, que deben ser buscados en el valor recuperado del atributo *text* del objeto Tweet.

Si se determina que la cantidad de Tweets relativos al tema de análisis no es suficiente para obtener resultados satisfactorios, es posible volver al paso 3 para ampliar la cantidad de cuentas de usuario y continuar con las nuevas cuentas recolectadas.

7. La determinación de la polaridad de los Tweets filtrados se hace a través de algoritmos de clasificación de Aprendizaje Automático, propios del Procesamiento del Lenguaje Natural y del Análisis de Sentimiento. Este paso es la punta del iceberg del presente trabajo y le da su nombre.

8. El proceso concluye con la construcción del Cuadro de Mando Integral. Los indicadores clave del cuadro de mando deben responder las preguntas que propiciaron el tema de análisis.

3.5.5 Aplicación de Teoría de Grafos

El procesamiento de datos descrito en la etapa 4 del flujo de datos comprende la distribución de los usuarios de Twitter según sus localizaciones declaradas y deducidas.

Para ello se aplican técnicas de Teoría de Grafos como se describe a continuación:

1. Se deben exportar los datos de localización de los usuarios y sus relaciones de la base de datos PostgreSQL a archivos intermedios en formato CSV (*comma-separated values*). La documentación de

PostgreSQL contiene el procedimiento para hacer esta exportación[101].

2. Se deben importar los datos de los archivos intermedios en formato CSV a la base de datos de grafos. La documentación de Neo4j contiene el procedimiento para hacer esta importación[102].

3. Se deben aplicar algoritmos de detección de comunidades en la base de datos grafos creada a partir de los archivos intermedios en formato CSV. La documentación de Neo4j contiene las rutinas para recomendadas para la aplicación de los algoritmos[103] y la estrategia para exportar los resultados en nuevos archivos intermedios en formato CSV[104].

4. Se deben importar los resultados de la aplicación de los algoritmos a la base de datos PostgreSQL. Para determinar la localización de los usuarios no localizados se utilizan las comunidades asociadas a las localidades ya conocidas por clasificados por cantidad de usuarios por comunidad. El procedimiento es el mismo de que la exportación de datos de PostgreSQL.

El “Capítulo IV. Aplicación del Modelo” describe de forma detallada el uso de los algoritmos de detección de comunidades en una base de datos de grafos en Neo4j.

3.5.6 Aplicación de Minería de Opinión

El procesamiento de datos descrito en la etapa 7 del flujo de datos comprende el análisis de datos para identificar los Tweets que expresan

[101] https://www.postgresql.org/docs/current/sql-copy.html
[102] https://neo4j.com/docs/cypher-manual/3.5/clauses/load-csv/
[103] https://neo4j.com/docs/graph-algorithms/current/algorithms/community/
[104] https://neo4j.com/docs/operations-manual/current/tools/cypher-shell/

opiniones relativas al tema de análisis y para determinar la polaridad de dichas opiniones.

Para ello se aplican técnicas de Minería de Opinión propios del Procesamiento del Lenguaje Natural como se describe a continuación:

1. Se deben segmentar los Tweets en palabras, *tokenization* en inglés, para su comparación con las palabras claves que definen el tema de análisis.

2. Se deben construir dos listas de Tweets para entrenamiento que contengan opiniones relativas al tema de análisis. La primera lista debe ser de Tweets con opiniones positivas y la segunda lista debe ser de Tweets con opiniones negativas.

3. Se deben construir dos listas adicionales de Tweets para evaluación que contengan opiniones relativas al tema de análisis. Las listas deben ser similares a las listas de entrenamiento.

4. Se debe construir un sistema de clasificación de texto basado en clasificadores Naive Bayes. Estos clasificadores deben ser entrenados con las listas de Tweets de entrenamiento y evaluados con las listas de Tweets de evaluación.

   La polaridad de los Tweets se determina con la ejecución de clasificadores entrenados. La documentación de la librería TextBlob contiene las rutinas recomendadas para la construcción del sistema de clasificación de texto[105].

El “Capítulo IV. Aplicación del Modelo” describe de forma detallada el uso de las librerías NLTK y TextBlob en *scripts* Python.

[105] https://textblob.readthedocs.io/en/dev/classifiers.html

3.5.7 Aplicación de Análisis de Negocios

Los modelos analíticos predictivos forman parte del Análisis de Negocios, *Business Analytics* en inglés, y permiten modelar relaciones entre variables y comprender cómo se comporta la variable dependiente a medida que cambia la variable independiente.

Las líneas de tendencia son aplicaciones de las funciones matemáticas de dichos modelos. El coeficiente de determinación R^2 establece la medida en que la tendencia determinada explica la variable real.

Para construir líneas de tendencia a partir de los resultados del Procesamiento de Datos, se deben determinar los valores de los indicadores identificados en la operacionalización de variables a partir de consultas SQL en la base de datos PostgreSQL: “Promedio de las polaridades de las opiniones” por “Cuenta de usuario de Twitter”, “Localización” y “Periodo de tiempo”.

Se debe emplear la clase *LinearRegression* de la librería de Aprendizaje Automático SciKit-Learn[106] para aplicar regresión lineal de mínimos cuadrados ordinarios. Esta clase cuenta con el método *predict* para obtener pronósticos a partir de un modelo lineal construido.

Los resultados deben ser presentados de forma gráfica en un Cuadro de Mando Integral. Para ello la librería gráfica Plotly[107] cuenta con varios componentes gráficos, entre ellos: *Pie*, *Bar* y *Scatter* para tortas, barras y líneas respectivamente.

Otra alternativa para el análisis de datos y la construcción del Cuadro de Mando Integral es exportar los datos de la base de datos en archivos

[106] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
[107] https://plot.ly/python/

intermedios en formato CSV[108] que luego pueden ser importados a hojas de cálculo de Microsoft Excel. La herramienta "Línea de Tendencia" o *Trendline*[109] de Excel proporciona un conjunto de opciones para determinar líneas de tendencia para conjuntos de datos. La función TENDENCIA o TREND[110] de Excel permite hacer pronósticos con base en tendencias lineales respecto de una o más variables independientes.

El "Capítulo IV. Aplicación del Modelo" describe de forma detallada el uso de las librerías SciKit-Learn y Plotly en *scripts* Python.

[108] https://www.postgresql.org/docs/current/sql-copy.html
[109] https://support.office.com/es-es/article/opciones-de-l%c3%adnea-de-tendencia-en-office-92157920-fee4-4905-bc89-6a0f48152c52?ui=es-ES&rs=es-ES&ad=ES
[110] https://support.office.com/es-es/article/funci%c3%b3n-tendencia-e2f135f0-8827-4096-9873-9a7cf7b51ef1?ui=es-ES&rs=es-ES&ad=ES

CAPITULO IV. APLICACIÓN DEL MODELO

La versión demostrativa operativa del modelo constituye una prueba de concepto del mismo. Esta versión es una implementación resumida construida con el propósito de verificar que el modelo es capaz de ser explotado en entornos de producción.

Conforme lo expuesto, la arquitectura de la versión demostrativa es:

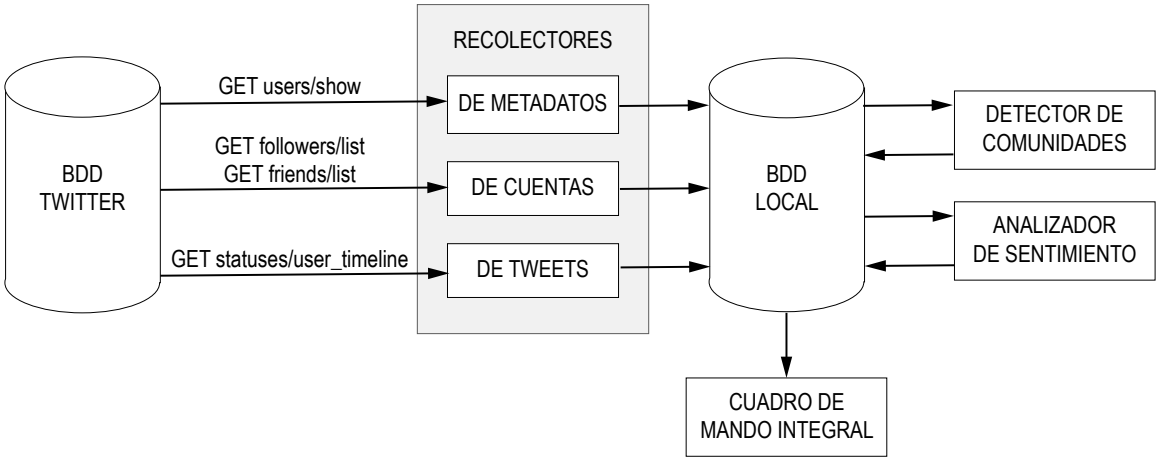

Figura 20: Arquitectura de la versión demostrativa.

Fuente: Elaboración propia.

4.1 Determinación del Tema de Análisis

En octubre del presente año se llevarán a cabo las elecciones generales para elegir al presidente y al vicepresidente del estado, uno de los temas más polémicos del año es la reelección o no de los actuales mandatarios.

El tema de análisis para la versión demostrativa del modelo es:

> "Opinión pública boliviana, a favor o en contra, de la potencial reelección de los actuales mandatarios."

Este tema asegura la disponibilidad de información adicional hechas por empresas especializadas para realizar comparaciones de resultados.

Las palabras clave que describen el tema de análisis son:

Palabras clave del Tema de Análisis del modelo.

Fuente: Elaboración propia.

PALABRAS CLAVE		
evoespueblo	21f	somosmas
evo	boliviadijono	fuerzaevo
ema	noesno	boliviadigna
linera	boliviadiceno	boliviadicesi
agl	evonuncamas	procesodecambio
	elecciones2019	2020-2015

La definición de las palabras clave se basó en los reportes de Temas de Tendencia en Twitter publicadas en el portal **TrendsMap.com**[111] a nivel de Hashtags al 10 de junio de 2019. Además de contemplar palabras que identifican a los actuales mandatarios, la elección balanceó la cantidad de Hashtags de temáticas a favor y en contra de los actuales mandatarios.

Se prevé el análisis de Tweets aplicado exclusivamente sobre palabras compuestas por letras y números, descartando signos de puntuación y de control, razón por la cual las palabras clave no llevan el signo numeral que precede a los Hashtags.

Los resultados de los sondeos de opinión detallados en el apartado “2.2.3.1 Sondeos de Intención de Voto” fueron cargados en la tabla `twitter_milestone` y servirán para comparar resultados con uno de los indicadores clave del Cuadro de Mando Integral.

[111] https://www.trendsmap.com/local/bolivia

4.2 Preparación Entorno de Trabajo

Para la preparación del entorno de trabajo se siguieron los pasos descritos en:

a) “Anexo III. Creación del Entorno de Desarrollo” el cual contiene los detalles de la preparación de los recursos tecnológicos empleados en la aplicación del modelo.

b) “Anexo IV. Creación de Cuenta de Desarrollo de Twitter” el cual contiene los detalles de la creación de una cuenta de desarrollador de Twitter. El procedimiento concluye con la obtención de las credenciales de la aplicación creada en Twitter.

4.3 Identificación de Cuentas Semilla

Para definición de las cuentas semilla se utilizaron los reportes de estadísticas de uso de Twitter hecha por SocialBakers.com[112] al 10 de junio de 2019, tal como se detalla en el “Anexo V. Definición de Cuentas Semilla”.

Se encontraron 63 cuentas de Twitter con más seguidores en Bolivia entre cuentas de marcas, de celebridades, de comunidad, de entretenimiento, de medios de comunicación, de lugares, de sociedad y de deportes.

Tabla 10: Resumen de recolección de cuentas semillas.
Fuente: Elaboración propia.

CUENTAS SEMILLA	TOTAL AMIGOS	TOTAL SEGUIDORES
63	80,516	6,714,107

Las cantidades totales de amigos y seguidores no son de usuarios únicos ya que un usuario puede seguir a más de una cuenta semilla.

[112] https://www.socialbakers.com/statistics/twitter/profiles/bolivia/

Siguiendo el diseño de la base de datos descrito en el “Anexo II. Diseño de la Base de Datos”, las cuentas semillas fueron insertadas en la tabla `twitter_seed`.

A continuación, los segmentos principales del *script* Python para la obtención de los metadatos de las cuentas semilla. Los primeros tres segmentos del *script* corresponden a: importación de librerías, definición de credenciales de Twitter y conexión a la base de datos local. Estos segmentos se repetirán en los siguientes *scripts*, razón por la cual serán omitidos.

Script para obtención de metadatos de cuentas semillas.

Fuente: Elaboración propia.

Script Python	
1	`import tweepy as tw`
2	`import psycopg2 as pg`
3	`consumer_key = 'NA28xye1...'`
4	`consumer_secret = 'qgJ6DXQH...'`
5	`access_token = '11309732...'`
6	`access_token_secret = 'YqK6Ds8G...'`
7	`connection = pg.connect(user="postgres", password="postgres",`
8	`    host="127.0.0.1", port="5432", database="sentodb")`
9	`cursor = connection.cursor()`
9	`cursor.execute("SELECT screen_name FROM twitter_seed")`
10	`rows = cursor.fetchall()`
11	
12	`for row in rows:`
13	`    user = api.get_user(screen_name=row[1])`
14	`    cursor.execute("INSERT INTO twitter_user VALUES ('" +`
15	`        user.id_str+"','"+user.screen_name+"', '"+`
16	`        str(user.created_at)+"','"+user.location+"', '"+`
17	`        str(user.followers_count)+"',"+str(user.friends_count)+`
18	`        str(user.protected) + ", null, null, 'UNCHECKED_')")`
19	
20	`cursor.close()`

En principio se establece conexión a la base de datos local con la cuenta por defecto y se establece conexión a las API Twitter empleando las credenciales obtenidas con la cuenta de desarrollo de Twitter.

4.4 Recolección de Seguidores y Amigos

La recolección de seguidores y amigos comprende 2 ciclos: el primer ciclo es la recolección de seguidores de cuentas semilla y el segundo ciclo comprende la recolección de seguidores y amigos de las cuentas recolectadas en el primer ciclo.

Este proceso es el más largo de la aplicación del modelo debido a las bajas tasas de transferencia de los métodos de las API de Twitter. Se ejecutó en un periodo de 60 días, del 24 de junio al 22 de agosto de 2019, y obtuvo un total de 4,031,077 de cuentas únicas de usuarios de Twitter.

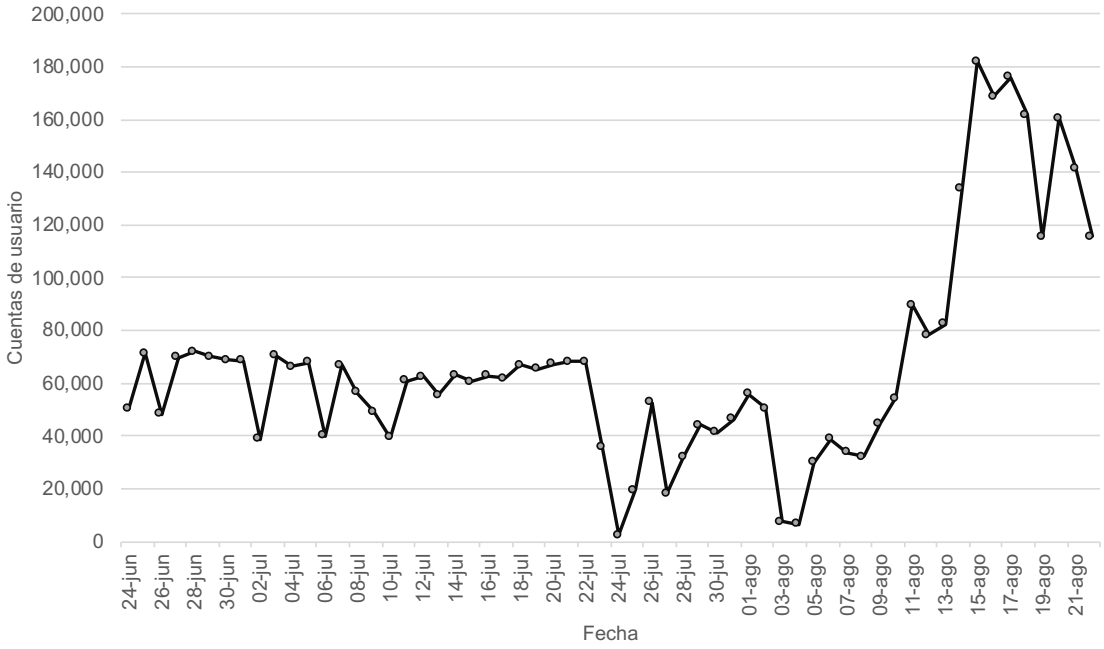

Figura 21: Evolución de la recolección de cuentas de Twitter.

Fuente: Elaboración propia.

El primer ciclo se ejecutó del 24 de junio al 8 de agosto de 2019 y el segundo ciclo se ejecutó del 9 de agosto al 22 de agosto de 2019.

Para la recolección de seguidores de las cuentas semilla, el primer ciclo, se emplearon las credenciales de sola cuenta de desarrollo de Twitter. Siendo

que este proceso tomó mucho tiempo, se crearon 2 cuentas de desarrollo de Twitter adicionales para ampliar indirectamente la tasa de transferencia, aunque esta no es una práctica recomendada por Twitter. Los resultados son apreciables en el segundo ciclo de recolección.

Tabla 11: Cuentas de usuario de Twitter recolectadas por ciclo.

Fuente: Elaboración propia.

INDICADOR	CICLO 1	CICLO 2	TOTAL
Cuentas recolectadas	2,325,631	1,705,446	4,031,077
Días de recolección	46	14	60

Es importante destacar que las cantidades de cuentas indicadas anteriormente no reflejan el esfuerzo del proceso. El proceso recolectó en total datos de 33,598,048 cuentas de Twitter, sin embargo, sólo 4,031,077 son cuentas únicas.

A continuación, los segmentos principales del *script* Python empleado para la recolección de seguidores y amigos, empleado en ambos ciclos.

Script para recolección de seguidores y amigos.

Fuente: Elaboración propia.

Script Python

```
1  cursor.execute("SELECT user_id FROM twitter_user WHERE status =
2      'UNCHECKED_'")
3  rows = cursor.fetchall()
4
5  for row in rows:
6      for users_in_page in tw.Cursor(api.followers, id=row[0],
7          count=200).pages():
8
9          for user in users_in_page:
10             cursor.execute("INSERT INTO twitter_user VALUES ('"+
11                 user.id_str+"','"+user.screen_name + "','"+
12                 str(user.created_at)+"','"+user.location+"','"+
13                 str(user.followers_count)+"',"+
14                 str(user.friends_count)+","+
15                 str(user.protected)+",null,null,'UNCHECKED_')")
16             cursor.execute("INSERT INTO twitter_following "+
17                 "VALUES ('"+user.id_str+"','"+row[0]+"')")
```

Script Python

```
18      for users_in_page in tw.Cursor(api.friends, id=row[0],
19          count=200).pages():
20
21          for user in users_in_page:
22              cursor.execute("INSERT INTO twitter_user VALUES ('"+
23                  user.id_str+"','"+user.screen_name + "','"+
24                  str(user.created_at)+"','"+user.location+"','"+
25                  str(user.followers_count)+"',"+
26                  str(user.friends_count)+","+
27                  str(user.protected)+",null,null,'UNCHECKED_')")
28              cursor.execute("INSERT INTO twitter_following "+
29                  "VALUES ('"+user.id_str+"','"+row[0]+"')")
30      cursor.execute("UPDATE twitter_user SET status = 'CHECKED_FF'
31          WHERE user_id = "+row[0]);
32  cursor.close()
```

El primer segmento del *script* obtiene la lista de las cuentas de las cuales se obtendrán los seguidores y amigos.

El segundo segmento del *script* recolecta cuentas de seguidores y el tercer segmento recolecta cuentas de amigos. El *script* no incluye las capturas de excepciones por llave primaria duplicada, esta excepción es arrojada cuando se intenta insertar una cuenta ya existente. El *script* tampoco incluye las capturas de excepciones de las invocaciones de las API de Twitter que comprenden condiciones de error atribuibles al estado de las cuentas consultadas e indisponibilidad de los servicios de Twitter.

El último segmento del *script* actualiza el estado de las cuentas de las que ya se obtuvieron seguidores y amigos.

4.5 Determinación de Localizaciones

La localización es equivalente al departamento en que se encuentra un usuario de Twitter: Beni, Chuquisaca, Cochabamba, La Paz, Oruro, Pando, Potosí, Santa Cruz o Tarija.

Se inició el proceso clasificando la información contenida en el campo `location` de la tabla `twitter_user`, con el siguiente resultado:

a) Con localización, que son cuentas cuya localización es una población dentro del territorio boliviano.

b) Localizados en Bolivia, que son cuentas cuya localización indica de forma genérica “Bolivia” sin incluir una localidad específica.

c) Sin localización, que son cuentas que no tienen un valor en el campo de localización.

d) Localizados en el exterior, que son cuentas cuya localización es una población fuera del territorio boliviano o un país extranjero.

Tabla 12: Resumen de clasificación de localización.

Fuente: Elaboración propia.

TIPO	CUENTAS	CUENTAS [%]
CON LOCALIZACIÓN	83,918	2.08%
LOCALIZADOS EN BOLIVIA	134,909	3.35%
SIN LOCALIZACIÓN	2,099,487	52.08%
LOCALIZADOS EN EL EXTERIOR	1,712,763	42.49%
TOTAL	**4,031,077**	**100.00%**

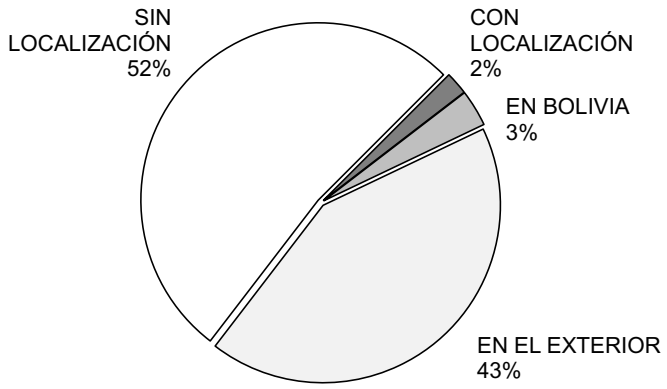

Figura 22: Resumen de clasificación de localización.

Fuente: Elaboración propia.

La primera etapa de la identificación consistió en determinar el departamento de las cuentas CON LOCALIZACIÓN, es decir, de los usuarios que publicaron su localización. Para el efecto, se usó la lista de municipios declarados en el Decreto Supremo Nº 1672[113] de Aprobación de Resultados del Censo de Población y Vivienda 2012, con el siguiente resultado:

Tabla 13: Cuentas con localización publicada por el usuario en Bolivia.

Fuente: Elaboración propia.

DEPARTAMENTO	CUENTAS	CUENTAS [%]
BENI	1,498	1.79%
CHUQUISACA	4,479	5.34%
COCHABAMBA	12,177	14.51%
LA PAZ	29,168	34.76%
ORURO	2,687	3.20%
PANDO	531	0.63%
POTOSÍ	1,587	1.89%
SANTA CRUZ	27,729	33.04%
TARIJA	4,062	4.84%
TOTAL	**83,918**	**100.00%**

La segunda etapa de la identificación empezó con la exportación de las cuentas de usuario Con Localización, Localización en Bolivia y Sin Localización, que sigan a usuarios con la misma clasificación de localización.

Comandos SQL para exportación de datos de cuentas.

Fuente: Elaboración propia.

Comando SQL	
1	`\COPY (`
2	`    SELECT a.user_id, a.screen_name, a.detected_location`
3	`    FROM twitter_user a, twitter_user b, twitter_following c`
4	`    WHERE a.user_id = c.user_id AND b.user_id = c.follows_to`
5	`        AND a.detected_location!='EXTERIOR'`

[113] https://medios.economiayfinanzas.gob.bo/MH/documentos/NORMAS-Y-DECRETOS/DS-2012/DS1672CENSO.pdf

Comando SQL	
6	AND b.detected_location!='EXTERIOR')
7	TO 'twitter_user.csv' DELIMITER ',' CSV HEADER;
8	\COPY (
9	SELECT a.user_id, b.user_id
10	FROM twitter_user a, twitter_user b, twitter_following c
11	WHERE a.user_id = c.user_id AND b.user_id = c.follows_to
12	AND a.detected_location!='EXTERIOR'
13	AND b.detected_location!='EXTERIOR')
14	TO 'twitter_following.csv' DELIMITER ',' CSV HEADER;

Se exportaron 2,020,593 cuentas de la base de datos y 14,111,514 relaciones archivo a archivos con formato CSV desde la terminal interactiva `psql` de la base de datos PostgreSQL. Los archivos fueron depositados en el directorio de importación de Neo4j: `neo4j-community-3.5.8/import/`.

Los datos de los archivos fueron importados en una base de datos de grafos desde el navegador de Neo4j[114].

Comandos Cypher para importación de datos de cuentas.

Fuente: Elaboración propia.

Comando Cypher	
1	USING PERIODIC COMMIT
2	LOAD CSV WITH HEADERS FROM "file:/twitter_user.csv" AS row
3	CREATE (n:User)
4	SET n = row
5	USING PERIODIC COMMIT
6	LOAD CSV WITH HEADERS FROM "file:/twitter_following.csv" AS row
7	MATCH (n:User {user_id: row.user_id})
8	MATCH (m:User {user_id: row.follows_to})
9	MERGE (n)-[:FOLLOWS]->(m)

Las cuentas fueron importadas como nodos etiquetados como `User` y las relaciones fueron importadas con la etiqueta `FOLLOWS`.

Dada la cantidad de cuentas y sus relaciones, no es posible desplegar de forma gráfica el grafo completo. Para ilustrar el grafo, la figura siguiente

[114] http://localhost:7474/browser/

muestra el grafo de los usuarios seguidos por el usuario con la cuenta `richardrojasi`. El grafo está compuesto por 106 nodos y 2,768 relaciones.

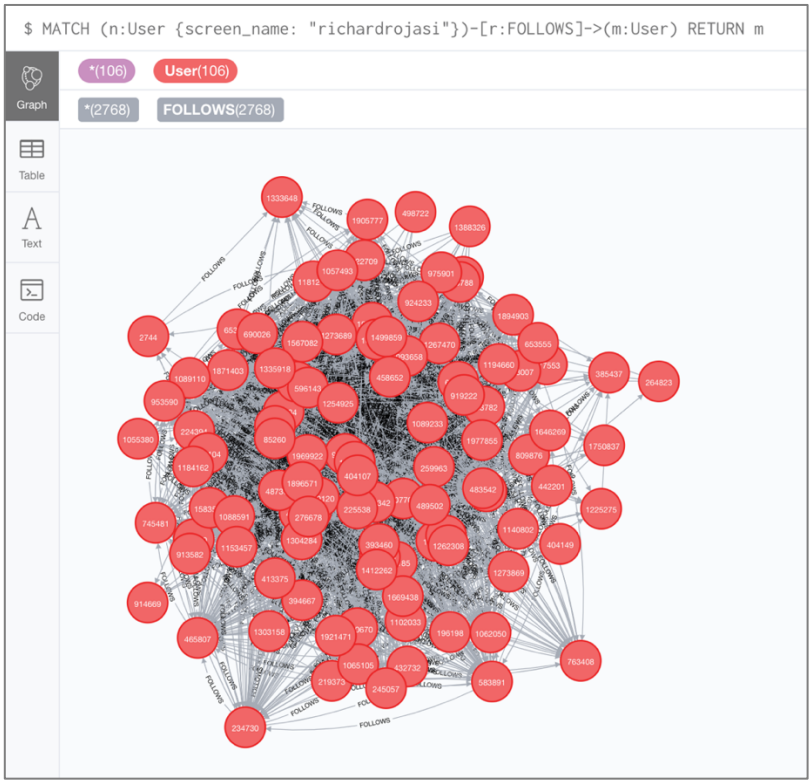

Figura 23: Ilustración del grafo de un usuario.

Fuente: Captura del navegador de Neo4j.

Se ejecutaron los algoritmos de detección de comunidades aplicables al grafo: algoritmo de Louvain sin comunidades predefinidas, algoritmo de Louvain con comunidades predefinidas, y algoritmo de Propagación de Etiquetas.

A continuación, los comandos Cypher para la obtención de estadísticas de ejecución de cada uno de los algoritmos indicados para determinar cuál es el más efectivo.

Comandos Cypher para estadísticas de algoritmos de detección de comunidades.

Fuente: Elaboración propia.

	Comando Cypher
1	`CALL algo.louvain('User', 'FOLLOWS',`
2	`    {write:true, writeProperty:'community'})`
3	`YIELD nodes, communityCount, iterations, loadMillis,`

Comando Cypher	
4	computeMillis, writeMillis;
5 6 7 8	CALL algo.louvain('User', 'FOLLOWS', {communityProperty: 'detected_location', write:true, writeProperty:'community'}) YIELD nodes, communityCount, iterations, loadMillis, computeMillis, writeMillis;
9 10 11 12 13	CALL algo.labelPropagation('User', 'FOLLOWS', {direction: "OUTGOING", iterations: 10, writeProperty: 'community', write: true}) YIELD nodes, communityCount, iterations, loadMillis, computeMillis, writeMillis;

Las estadísticas indican que el algoritmo de Louvain sin comunidades predefinidas detecta la menor cantidad de comunidades en el grafo. Sus estadísticas indican que el algoritmo ejecutado sobre un grafo de 2,020,593 nodos, detecta 85 comunidades en 4 iteraciones, la carga de datos es hecha en 11,655 milisegundos, la ejecución del algoritmo es hecha en 57,005 milisegundos y la escritura de resultados es hecha en 25,051 milisegundos.

Tabla 14: Estadísticas de algoritmos de detección de comunidades.

Fuente: Elaboración propia.

Algoritmo	Comuni-dades	Iteracio-nes	Carga [mseg]	Ejecución [mseg]	Escritura [mseg]
Louvain sin comunidades predefinidas	85	4	11,655	57,005	25,051
Louvain con comunidades predefinidas	109	4	6,778	13,674	24,957
Propagación de Etiquetas	322,329	9	2,557	5,362	88

Se guardó el comando de ejecución del algoritmo de Louvain sin comunidades predefinidas en un archivo llamado `algo_louvain.cql` cuyo contenido es:

Comando Cypher de ejecución del algoritmo de Louvain.

Fuente: Elaboración propia.

Comando Cypher

```
1  CALL algo.louvain.stream('User', 'FOLLOWS', {})
2  YIELD nodeId, community
3  RETURN algo.asNode(nodeId).user_id AS user_id, community
4  ORDER BY community
```

Se ejecutó el algoritmo de Louvain desde línea de comando empleando el utilitario `cypher-shell` de Neo4j. El resultado se exportó al archivo `algo_result.csv`. El comando ejecutado es:

Comando Shell de exportación de resultados del algoritmo de Louvain.

Fuente: Elaboración propia.

Comando Shell

```
1  cat algo_louvain.cql | cypher-shell -u neo4j -p password \
2      > algo_result.csv
```

Se importaron los datos del archivo `algo_result.csv` a una tabla auxiliar que contiene los campos `user_id` y `community` que es el identificador de la comunidad detectada por el algoritmo de Louvain, este campo tiene un valor numérico entre 0 y 84. El comando para la importación desde la terminal interactiva `psql` es:

Comando SQL para importación de datos de comunidades detectadas.

Fuente: Elaboración propia.

Comando SQL

```
1  \COPY aux_user_louvain1 FROM 'algo_result.csv' DELIMITER ','
2      CSV HEADER;
```

Las comunidades fueron posicionadas por cantidad de cuentas respecto de las localizaciones conocidas. Se asume como localización de la comunidad a la que ocupa la primera posición. Por ejemplo, las 4 primeras posiciones de la comunidad 2 ordenadas por cantidad de cuentas son: Santa Cruz, La Paz, Cochabamba y Chuquisaca, con lo que se asume que la localidad de la comunidad 2 es Santa Cruz.

Tabla 15: Ejemplo de posicionamiento de localizaciones de la comunidad 2.

Fuente: Elaboración propia.

POSICIÓN	LOCALIZACIÓN	CUENTAS
1	SANTA CRUZ	15,751
2	LA PAZ	12,920
3	COCHABAMBA	4,441
4	CHUQUISACA	1,920

A continuación, el resumen de las comunidades por localización. Los departamentos de Oruro, Pando y Potosí no lograron imponerse en ninguna comunidad.

Tabla 16: Resumen de comunidades por departamento de Bolivia.

Fuente: Elaboración propia.

DEPARTAMENTO	TOTAL COMUNIDADES	COMUNIDADES
BENI	2	17, 45
CHUQUISACA	6	3, 36, 66, 71, 77, 78
COCHABAMBA	14	9, 11, 12, 37, 41, 42, 49, 60, 69, 74, 75, 76, 81, 84
LA PAZ	40	0, 1, 4, 5, 6, 7, 14, 16, 19, 23, 25, 26, 27, 28, 29, 30, 31, 32, 34, 35, 39, 40, 43, 44, 46, 51, 53, 56, 57, 58, 62, 63, 65, 68, 72, 73, 79, 80, 82, 83
SANTA CRUZ	21	2, 8, 10, 13, 15, 20, 21, 22, 24, 33, 47, 48, 50, 52, 54, 55, 59, 61, 64, 67, 70
TARIJA	2	18, 38

Siendo el problema de esta etapa la identificación de la localización de las cuentas localizadas en Bolivia de forma genérica y de las cuentas que no tienen localización, la distribución resultante de cuentas por departamento es:

Tabla 17: Distribución de cuentas localizadas en Bolivia y sin localización.

Fuente: Elaboración propia.

DEPARTAMENTO	EN BOLIVIA	SIN LOCALIZACIÓN
BENI	0	94
CHUQUISACA	1,501	14,531
COCHABAMBA	7,974	69,028
LA PAZ	105,339	1,286,697
ORURO	0	0
PANDO	0	0
POTOSI	0	0
SANTA CRUZ	19,082	423,027
TARIJA	937	8,493
NINGUNO	**76**	**297,617**

Existe un remanente de 76 cuentas localizadas en Bolivia de forma genérica que no forman parte de ninguna comunidad y 297,617 cuentas sin localización que tampoco forman parte de ninguna comunidad. Estas cuentas, al igual que las cuentas localizadas en el exterior, dejan de formar parte de la población de estudio del modelo.

El estado final de las cuentas con localizaciones declaradas combinadas con las cuentas con localizaciones deducidas es:

Tabla 18: Distribución final de cuentas por departamento.

Fuente: Elaboración propia.

DEPARTAMENTO	CUENTAS	CUENTAS [%]
BENI	1,592	0.08%
CHUQUISACA	20,511	1.02%
COCHABAMBA	89,179	4.41%
LA PAZ	1,421,204	70.34%
ORURO	2,687	0.13%

DEPARTAMENTO	CUENTAS	CUENTAS [%]
PANDO	531	0.03%
POTOSÍ	1,587	0.08%
SANTA CRUZ	469,838	23.25%
TARIJA	13,492	0.67%
TOTAL	**2,020,621**	**100.00%**

4.6 Recolección de Tweets

El proceso de recolección de Tweets se ejecutó en un periodo de 12 días, del 6 al 17 de septiembre, periodo en el cual se recolectaron 105,125,489 Tweets de 2,020,621 cuentas de usuario de Twitter.

Para la recolección de Tweets se emplearon las credenciales de las 3 cuentas de desarrollo de Twitter que se obtuvieron en la etapa de Recolección de Seguidores y Amigos.

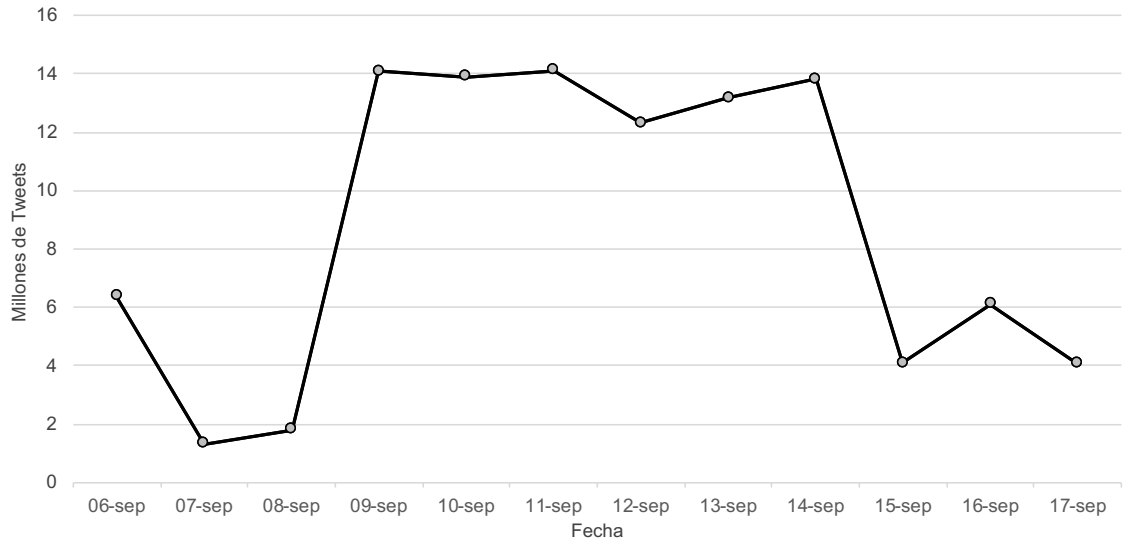

Tabla 19: Evolución de la recolección de Tweets.

Fuente: Elaboración propia.

A continuación, los segmentos principales del *script* Python empleado para la recolección de los 200 últimos Tweets de cada cuenta de usuario. La ejecución del *script* se hizo por localización detectada.

Script para recolección de Tweets.

Fuente: Elaboración propia.

Script Python

```
cursor.execute("SELECT user_id FROM twitter_user WHERE status in
    ('UNCHECKED_','CHECKED_FF') AND detected_location in ("+l+")")
rows = cursor.fetchall()

for row in rows:
    for messages_in_page in tw.Cursor(api.user_timeline,
        user_id=row[0], tweet_mode="extended", include_rts="true",
        count=200, page=1).pages():

        for msg in messages_in_page:
            cursor.execute("INSERT INTO twitter_tweet VALUES ('"+
                row[0]+"', '"+str(msg.id)+"', '"+
                str(msg.created_at)+"', '"+msg.source+"', '"+
                str(msg.coordinates)+"', '"+msg.full_text+
                "', NULL, NULL)")
        break
    cursor.execute("UPDATE twitter_user SET status = status||'MM'
        WHERE user_id = "+row[0]);

cursor.close()
```

El primer segmento del *script* obtiene la lista de las cuentas de las cuales se obtendrán los Tweets y explora la lista.

El segundo segmento del *script* recolecta los últimos 200 Tweets de cada usuario sin cortes en el contenido del texto del Tweet. El *script* no incluye las capturas de excepciones de las invocaciones de las API de Twitter que comprenden condiciones de error atribuibles al estado de las cuentas consultadas e indisponibilidad de los servicios de Twitter.

El último segmento del *script* actualiza el estado de las cuentas de las que ya se obtuvieron Tweets.

4.7 Determinación de Polaridad de Tweets Recolectados

Para la construcción de las listas de Tweets para entrenamiento y para evaluación del sistema de clasificación de texto, se obtuvo una muestra aleatoria de 10,000 Tweets que contengan cualquiera de las palabras clave.

A continuación, el comando SQL ejecutado en la terminal interactiva `psql` de la base de datos PostgreSQL para exportar los Tweets de la tabla `twitter_tweet` a un archivo con formato CSV, seguido del comando de sistema operativo empleado para obtener la muestra aleatoria de 10,000 Tweets.

Comando SQL para exportar Tweets de la base de datos.

Fuente: Elaboración propia.

Comando SQL

```
1  \COPY (SELECT user_id, tweet_id, text FROM twitter_tweet)
2      TO 'twitter_tweet.csv' DELIMITER ',' CSV;
```

Comando Shell para extraer filas de un archivo de texto.

Fuente: Elaboración propia.

Comando Shell

```
1  egrep '<keywords>' twitter_tweet.csv | head -10000 > \
2      tweets_related.csv
```

A través de una revisión manual, se eligieron los primeros 500 Tweets con opiniones positivas y los primeros 500 Tweets con opiniones negativas. La lista de Tweets para entrenamiento fue compuesta con 400 Tweets con opiniones positivas y 400 Tweets con opiniones negativas. La lista de Tweets para evaluación fue compuesta con los restantes 100 Tweets con opiniones positivas y los restantes 100 Tweets con opiniones negativas.

Las listas de Tweets para entrenamiento y para evaluación tienen 800 y 200 Tweets respectivamente. Cada texto de los Tweets está acompañado de la

clasificación recibida: "pos" para opiniones positivas y "neg" para opiniones negativas. Se convirtieron los textos a minúsculas y se quitaron las tildes y diéresis en vocales.

Ejemplos de Tweets con opiniones positiva y negativa.

Fuente: Elaboración propia.

Ejemplo de Lista de Entrenamiento	
1	#estadolaico sigamos avanzando sigamos con el proceso de cambio velando por los derechos de todos los bolivianos y esperando lo objetivo de lo subjetivo,pos
2	acaso usted protege los derechos humanos cuando reprime jovenes estudiantes que solo piden que se respete el voto popular del #21f donde se nego la posibilidad de re postularse,neg

El archivo con la lista de entrenamiento es llamado `training_data.csv` y el archivo con la lista de evaluación es llamado `test_data.csv`.

A continuación, la primera parte del *script* Python para la construcción de un sistema de clasificación de texto basado en clasificadores Naive Bayes.

Script para construcción de clasificador Naive Bayes.

Fuente: Elaboración propia.

Script Python	
1	`import re`
2	`import csv`
3	`from textblob import TextBlob`
4	`from textblob.classifiers import NaiveBayesClassifier`
5	`with open("training_data.csv", "r") as fp:`
6	`    cl = NaiveBayesClassifier(fp, format="csv")`
7	`with open("test_data.csv", "r") as fp:`
8	`    accuracy = cl.accuracy(fp, format="csv")`
9	`    print("Accuracy: {0}".format(accuracy))`

El primer segmento del *script* contiene la importación de librerías a ser usadas.

El segundo segmento contiene la creación del clasificador entrenado a partir de los mensajes contenidos en la lista de entrenamiento.

El tercer segmento contiene la evaluación del clasificador entrenado hecho con la lista de evaluación. **La evaluación tiene 95% de precisión**, esto quiere decir que la clasificación que hizo el clasificador coincide el 95% de las veces con la clasificación que ya tienen los Tweets de entrenamiento.

A continuación, la segunda parte del *script* Python para la determinación de la polaridad de los Tweets.

Script para determinar la polaridad de Tweets.

Fuente: Elaboración propia.

Script Python

```
1   replacements = [(u'\xc1', 'A'), (u'\xc9', 'E'), (u'\xcd', 'I'),
2       (u'\xbf', 'O'), (u'\xda', 'U'), (u'\xdc', 'U'),
3       (u'\xe1', 'a'), (u'\xe9', 'e'), (u'\xed', 'i'),
4       (u'\xf3', 'o'), (u'\xfa', 'u'), (u'\xfc', 'u')]
5
6   def remove_accents(str):
7       for a, b in replacements:
8           str = str.replace(a, b)
9       return str
10  cursor.execute("SELECT keyword FROM twitter_issue")
11  keywords = [row[0] for row in cursor.fetchall()]
12  result_file = open('twitter_sentiment.csv', 'w')
13
14  with open('twitter_tweet.csv') as csv_file:
15      csv_reader = csv.reader(csv_file, delimiter=",")
16
17      for row in csv_reader:
18          msg = re.sub('http[a-zA-Z0-9:/.]+', '', row[5])
19          blob = TextBlob(remove_accents(msg.lower()), classifier=cl)
20
21          if len(blob.words) > 1:
22              is_related = False
23              for word in blob.words:
24                  if word in keywords:
25                      is_related = True
26                      break
27
28              if is_related:
29                  acc = cl.classify(msg)
30                  result_file.write(f'{row[0]},{row[1]},
31                                        {row[2]},{acc}\n')
32  result_file.close()
```

El primer segmento del *script* contiene las definiciones necesarias para una función de eliminación de tildes y diéresis en vocales a través de codificación Unicode.

El segundo segmento contiene la definición de las palabras clave del tema de análisis obtenida de la tabla `twitter_issue`.

El último segmento contiene la determinación de la polaridad de los Tweets contenidos en el archivo exportado de la base de datos PostgreSQL. Incluye la eliminación de cadenas de URL y conversión del texto de Tweets a minúsculas. El paso anterior a la determinación de la polaridad es la segmentación del texto de los Tweets en palabras, *tokenization*, para su comparación con las palabras claves. El resultado son los identificadores del Tweet y la clasificación dada por el clasificador entrenado, estos datos son escritos en un archivo de texto con formato CSV.

El proceso concluye con la importación de los resultados de la determinación de polaridad a la tabla `twitter_sentiment` de la base de datos PostgreSQL y la creación de una vista con las estadísticas de polaridad calculada por cuenta de usuario.

Comando SQL para importar resultados de determinación de polaridad.

Fuente: Elaboración propia.

	Comando SQL
1 2	`\COPY twitter_sentiment FROM 'twitter_sentiment.csv'` `    DELIMITER ',' CSV;`

3
4
5
6
7
8
9
	`create view aux_user_sentiment as`
	`select user_id,`
	`    avg(case when sentiment='pos' then 1 else -1 end) sent_avg,`
	`    stddev(case when sentiment='pos' then 1 else -1 end) sent_dev,`
	`    count(1) opinions`
	`from twitter_sentiment`
	`group by user_id;`

Comando SQL	
10	`create view aux_user_sentiment_daily as`
11	`select user_id, created_at::date dd,`
12	`avg(case when sentiment='pos' then 1 else -1 end) sent`
13	`from twitter_sentiment`
14	`group by user_id, created_at::date;`

De acuerdo a la operacionalización de variables, la opinión de una cuenta de usuario es el promedio de las opiniones emitidas a través de la cuenta, donde 1 es una opinión positiva y -1 es una opinión negativa.

La desviación estándar de las opiniones permite determinar el grado de seguridad de una cuenta de usuario respecto de sus opiniones.

4.8 Construcción del Cuadro de Mando Integral

El Cuadro de Mando Integral está compuesto de los siguientes indicadores clave:

1. Tasa de cuentas de usuario que emiten opinión respecto del tema de análisis.

2. Tasa de opiniones a favor, en contra y neutras de cuentas de usuario que emiten opinión respecto del tema de análisis total.

3. Tasas de opiniones a favor, en contra y neutras de cuentas de usuario que emiten opinión respecto del tema de análisis total por departamento.

4. Tasas de opiniones a favor y en contra de cuentas de usuario que emiten opinión respecto del tema de análisis de forma diaria comparada con resultados de sondeos de opinión.

La definición de los indicadores está en función del tema de análisis y la experiencia en el tema de quienes aplican el modelo.

Para obtener el primer indicador clave, el comando SQL es el siguiente:

Comando SQL para el primer indicador del Cuadro de Mando.

Fuente: Elaboración propia.

Comando SQL

```
select case when s.user_id is null then 'NO OPINAN' else 'OPINAN'
       end acc_type, count(1) accounts
from twitter_user u full outer join aux_user_sentiment s
     on u.user_id = s.user_id
where u.detected_location not in
      ('EXTERIOR','PEDNING-BO','PENDING-EMPTY')
group by case when s.user_id is null then 'NO OPINAN' else
        'OPINAN' end;
```

Para obtener el segundo indicador clave, el comando SQL es el siguiente:

Comando SQL para el segundo indicador del Cuadro de Mando.

Fuente: Elaboración propia.

Comando SQL

```
select case when sent_avg > 0 then 'A FAVOR' when sent_avg = 0
       then 'NEUTRA' else 'EN CONTRA' end sentiment,
       count(1) accounts
from aux_user_sentiment
group by case when sent_avg > 0 then 'A FAVOR' when sent_avg = 0
         then 'NEUTRA' else 'EN CONTRA' end;
```

Para obtener el tercer indicador clave, el comando SQL es el siguiente:

Comando SQL para el tercer indicador del Cuadro de Mando.

Fuente: Elaboración propia.

Comando SQL

```
select u.detected_location,
       sum(case when sent_avg < 0 then 1 else 0 end) against,
       sum(case when sent_avg > 0 then 1 else 0 end) on_favor,
       sum(case when sent_avg = 0 then 1 else 0 end) neutral
from twitter_user u, aux_user_sentiment s
where u.user_id = s.user_id and u.detected_location not in
      ('EXTERIOR','PENDING-BO','PENDING-EMPTY')
group by u.detected_location;
```

Para obtener el cuarto indicador clave, el comando SQL es el siguiente:

Comando SQL para el cuarto indicador del Cuadro de Mando.

Fuente Elaboración propia.

Comando SQL

```
1  select to_char(s.dd,'dd/mm/yyyy') dd,
2         sum(case when sent > 0 then 1 else 0 end)*1.0/count(1),
3         m.value_1 on_favor,
4         sum(case when sent < 0 then 1 else 0 end)*1.0/count(1),
5         m.value_2 against,
6         sum(case when sent = 0 then 1 else 0 end)*1.0/count(1),
7         m.value_3 neutral
8  from twitter_user u, aux_user_sentiment_daily s left outer join
9       twitter_milestone m on s.dd = m.posted_at
10 where u.user_id = s.user_id and u.detected_location not in
11       ('EXTERIOR','PENDING-BO','PENDING-EMPTY') and
12       s.dd between '04/25/2019'::date and '09/11/2019'::date
13 group by s.dd, m.value_1, m.value_2, m.value_3
14 order by s.dd;
```

El cuarto indicador clave está delimitado para hacer comparaciones y obtener tendencias en el periodo de tiempo en el cual se cuentan con datos de sondeos de opinión: del 25 de abril al 11 de septiembre de 2019.

Para construir el marco del Cuadro de Mando Integral se empló el marco de trabajo Python para construir aplicaciones web Dash[115]. Los componentes gráficos se construyeron con la librería gráfica Plotly[116].

Dado que las tendencias de opinión tienen un comportamiento lineal, la aplicación de la regresión lineal para obtener pronósticos se hizo empleando la clase *LinearRegression* de la librería de Aprendizaje Automático SciKit-Learn[117].

[115] https://dash.plot.ly/

[116] https://plot.ly/python/

[117] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

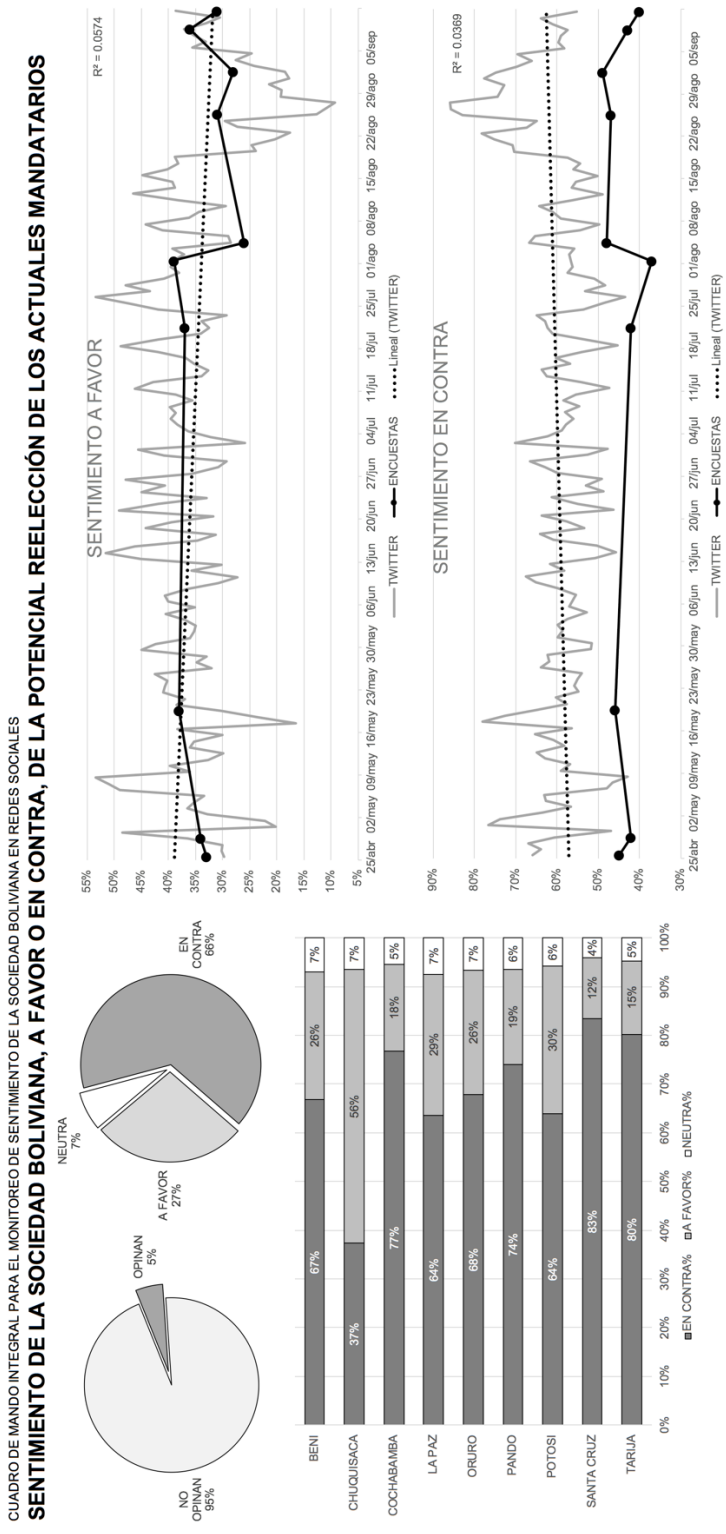

Tabla 20: Ilustración del Cuadro de Mando Integral propuesto.

Fuente: Elaboración propia.

CAPITULO V. MARCO DE RESULTADOS

5.1 Resultados de la Aplicación del Modelo

5.1.1 Indicadores del Cuadro de Mando Integral

A continuación, los resultados de los indicadores clave definidos para la versión demostrativa del modelo:

1. La tasa de cuentas de usuario que emiten opinión respecto del tema de análisis es 5.04%.

Tabla 21: Tasa de cuentas de usuario que emiten opinión.

Fuente: Elaboración propia.

TIPO USUARIOS	CUENTAS	CUENTAS [%]
OPINAN	101,933	5.04%
NO OPINAN	1,918,688	94.96%
TOTAL	**2,020,621**	**100.00%**

2. De la tasa de opiniones a favor, en contra y neutras de cuentas de usuario que emiten opinión respecto del tema de análisis, la tasa de opiniones en contra es 65.52%.

Tabla 22: Tasa de opiniones emitidas respecto del tema de análisis.

Fuente: Elaboración propia.

TIPO OPINIÓN	CUENTAS	CUENTAS [%]
EN CONTRA	66,791	65.52%
A FAVOR	28,053	27.52%
NEUTRA	7,089	6.95%
TOTAL	**101,933**	**100.00%**

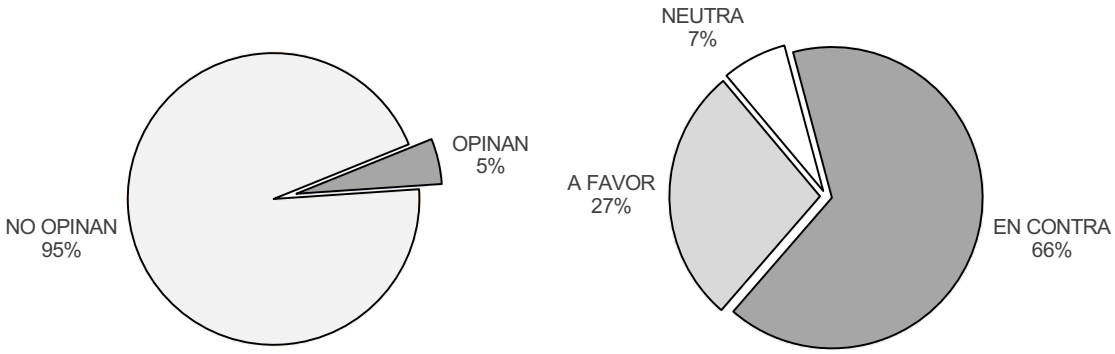

Figura 24: Tasas de quienes emiten opinión y el tipo de opiniones que emiten.

Fuente: Elaboración propia.

3. De las tasas de opiniones a favor, en contra y neutras de cuentas de usuario que emiten opinión respecto del tema de análisis total por departamento, las opiniones en contra son la mayoría en 8 de 9 departamentos.

Tabla 23: Tasa de opiniones emitidas por departamento.

Fuente: Elaboración propia.

DEPARTAMENTO	EN CONTRA	A FAVOR	NEUTRA
BENI	66.82%	26.27%	6.91%
CHUQUISACA	37.29%	56.19%	6.52%
COCHABAMBA	76.80%	17.74%	5.47%
LA PAZ	63.64%	28.94%	7.41%
ORURO	67.81%	25.63%	6.56%
PANDO	74.03%	19.48%	6.49%
POTOSI	63.95%	30.23%	5.81%
SANTA CRUZ	83.48%	12.48%	4.04%
TARIJA	80.26%	15.07%	4.67%
TOTAL	**65.52%**	**27.52%**	**6.95%**

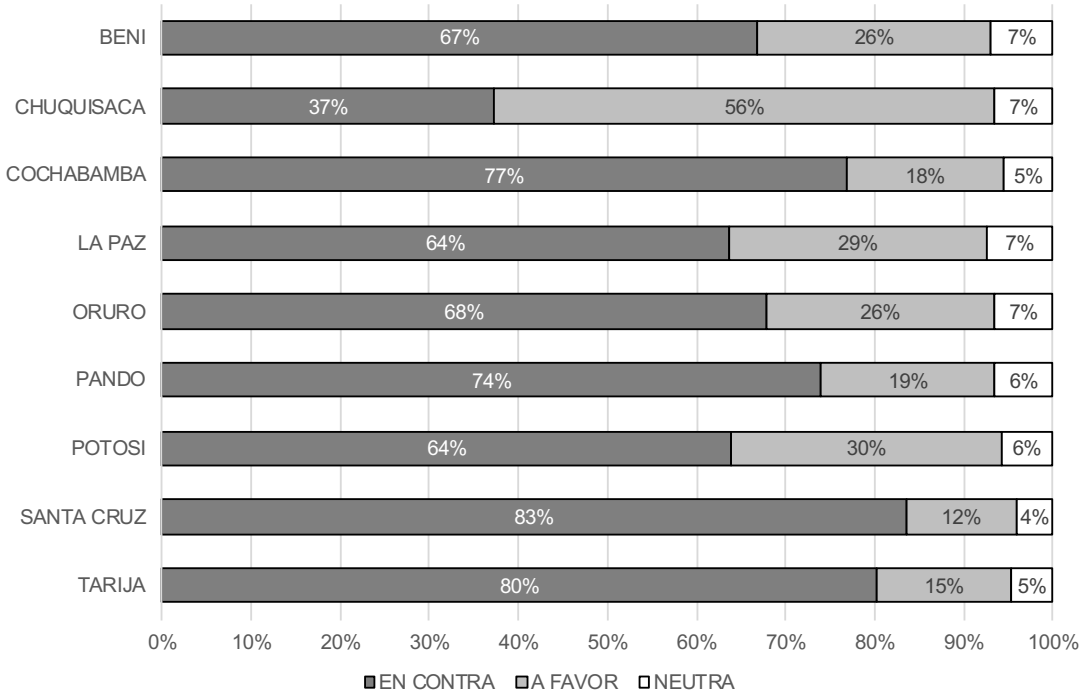

Figura 25: Opiniones a favor, en contra y neutras por departamento.

Fuente: Elaboración propia.

4. De las tasas de opiniones a favor y en contra de cuentas de usuario que emiten opinión respecto del tema de análisis de forma diaria comparada con resultados de sondeos de opinión, es fácil determinar que las opiniones a favor tienen una tendencia descendente.

   El último punto de la tendencia coincide con el resultado del sondeo de opinión hecho el 11 de septiembre de 2019 por la iniciativa “Tu Voto Cuenta”.

   A su vez, la tendencia de las opiniones en contra es ascendente y está por encima de todos los resultados de los sondeos de opinión.

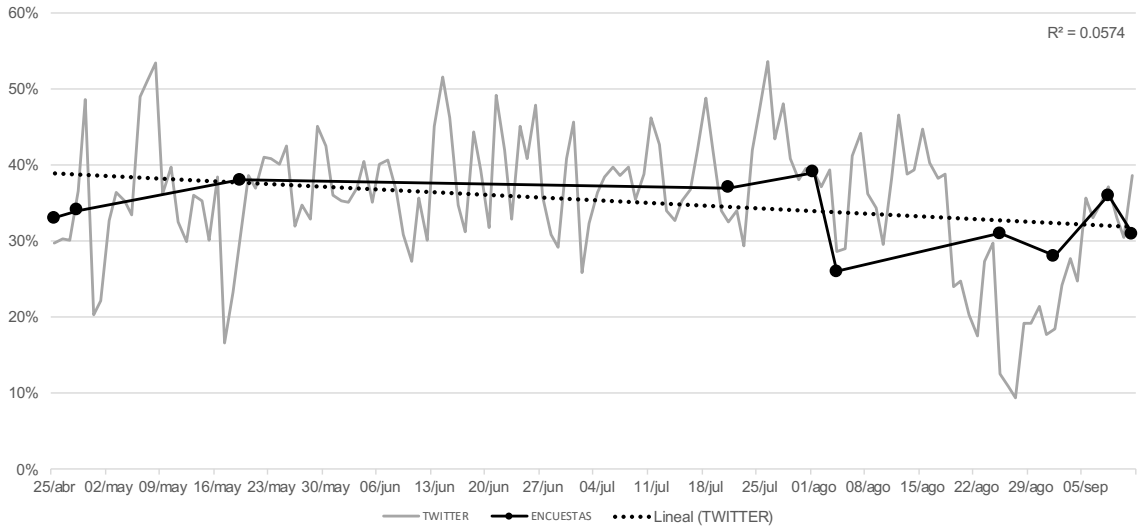

Figura 26: Tendencia de tasa diaria de opiniones a favor comparada.

Fuente: Elaboración propia.

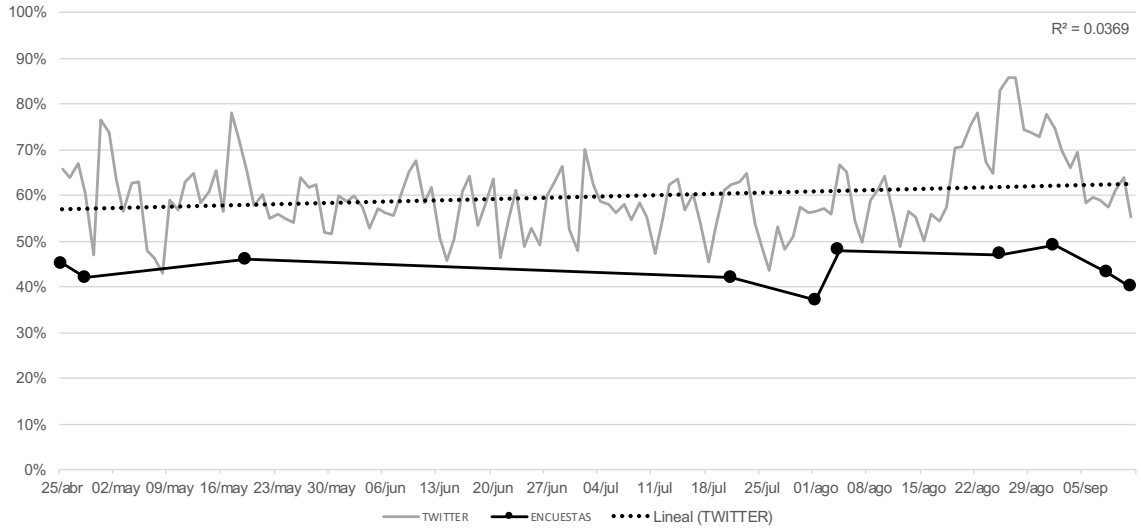

Figura 27: Tendencia de tasa diaria de opiniones en contra comparada.

Fuente: Elaboración propia.

En resumen, el 65.52% de la sociedad boliviana está en contra de la potencial reelección de los actuales mandatarios. Esta opinión se refleja en 8 de los 9 departamentos de Bolivia.

De acuerdo a las tendencias de las opiniones a favor y en contra, al 20 de octubre de 2019:

- 29.87% estará a favor de la reelección de los actuales mandatarios.
- 64.17% estará en contra de la reelección de los actuales mandatarios.

5.1.2 Validación de la Muestra

Aplicando el marco teórico del apartado “2.1.4.1 Definición de la Muestra” y los valores mínimo y máximo de los errores muestrales detallados en el apartado “2.2.3.1 Sondeos de Intención de Voto”, se comprueban los fundamentos por los cuales las empresas especializadas en sondeos de opinión en Bolivia tomaron muestras entre 800 y 2,250 individuos.

Tabla 24: Muestras de Sondeos de Intención de Voto según error muestral.
Fuente: Elaboración propia.

PARÁMETRO	VALOR MÍNIMO	VALOR MÁXIMO
Grado de confianza (Z)	95%	95%
Probabilidad de éxito (p)	50%	50%
Error muestral permitido (e)	2.07%	3.47%
Tamaño de la muestra (n)	**2,241**	**798**

Siendo que el presente trabajo obtuvo una muestra de 101,933 cuentas de usuario de Twitter, el error muestral calculado con base en el mismo marco teórico es 0.31%.

Tabla 25: Error muestral del presente trabajo.
Fuente: Elaboración propia.

PARÁMETRO	VALOR
Grado de confianza (Z)	95%
Probabilidad de éxito (p)	50%
Tamaño de la muestra (n)	101,933
Error muestral (e)	**0.31%**

5.1.3 Resultados Adicionales

Se determinó que, de los más 2 millones de cuentas de usuario recolectadas, el 14.40% de las cuentas recolectadas fueron creadas el año 2016. El año 2016 muchas autoridades de estado crearon sus cuentas en Twitter, entre ellas la del primer mandatario, la cual fue creada el 15/04/2016.

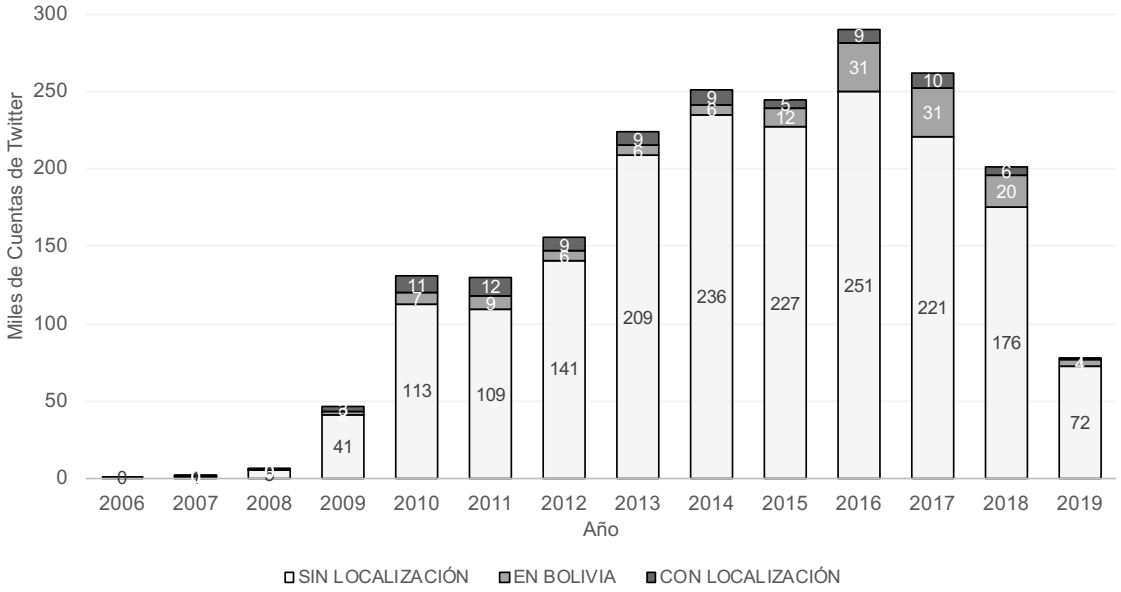

Figura 28: Cantidad de Cuentas de Twitter por año de creación.

Fuente: Elaboración propia.

Se confirma el carácter público de la información publicada en Twitter, el 97.53% de las cuentas recolectadas son públicas, las opiniones de estas cuentas son de acceso público. Sólo el 2.41% de las cuentas recolectadas son privadas.

De los más de 105 millones de Tweets recolectados, el 35.64% de los Tweets fueron escritos el año 2019. Esto sugiere que, si bien existe un importante número de usuarios, aproximadamente 2 millones, la mayoría no publica Tweets.

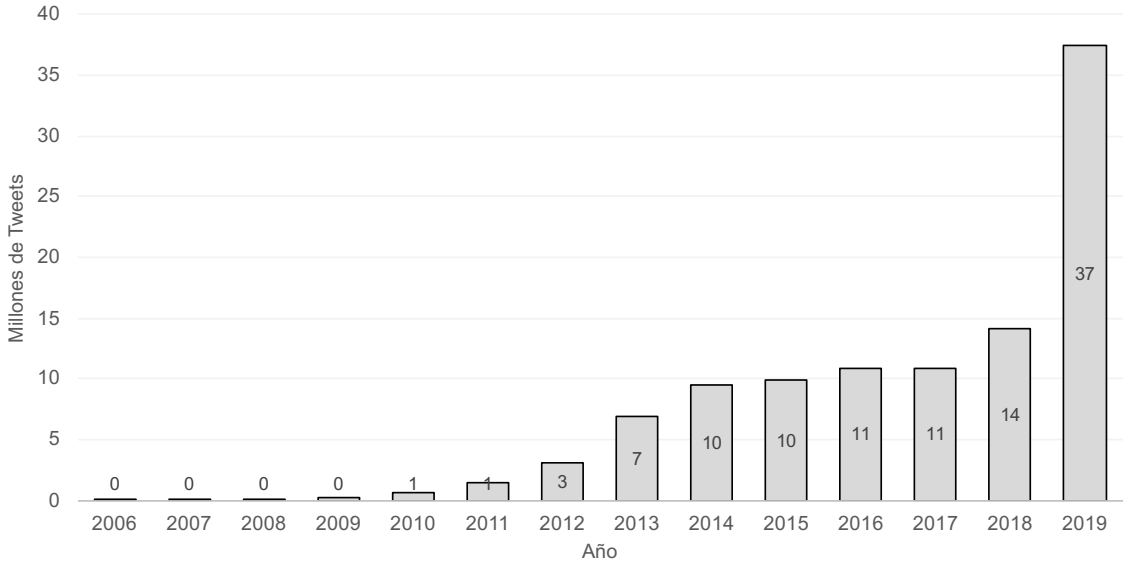

Figura 29: Cantidad de Tweets por año de creación.

Fuente: Elaboración propia.

Por otro lado, la aplicación más utilizada para publicar Tweets es la aplicación Web de Twitter seguida por la aplicación móvil para Android.

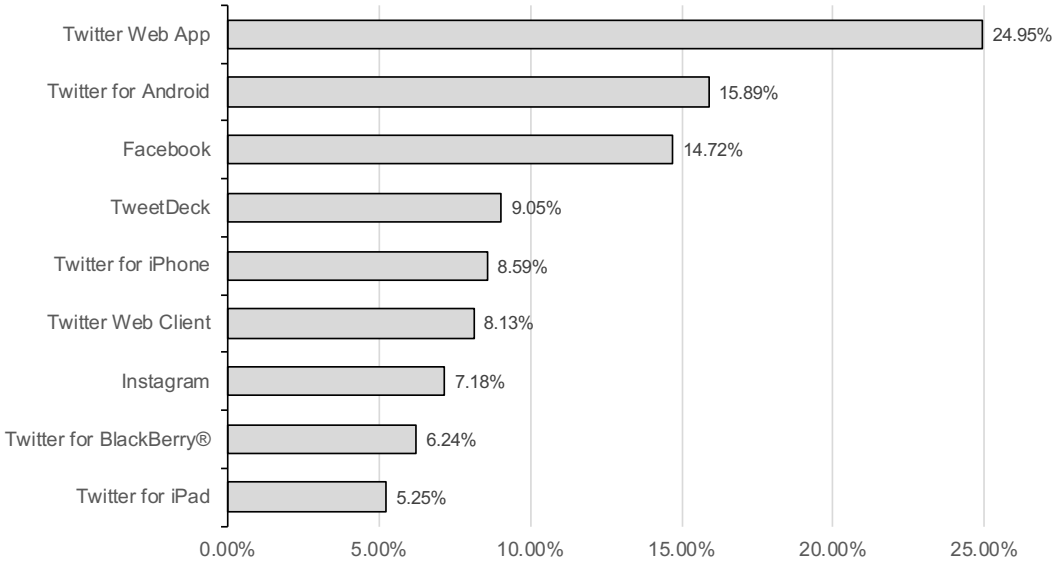

Figura 30: Primeros 9 dispositivos usados para publicar Tweets.

Fuente: Elaboración propia.

Con base en la disponibilidad de la ubicación geográfica de los Tweets emitidos en el territorio nacional, se determinó la distribución geográfica de los usuarios de Twitter a lo largo de los principales municipios del país.

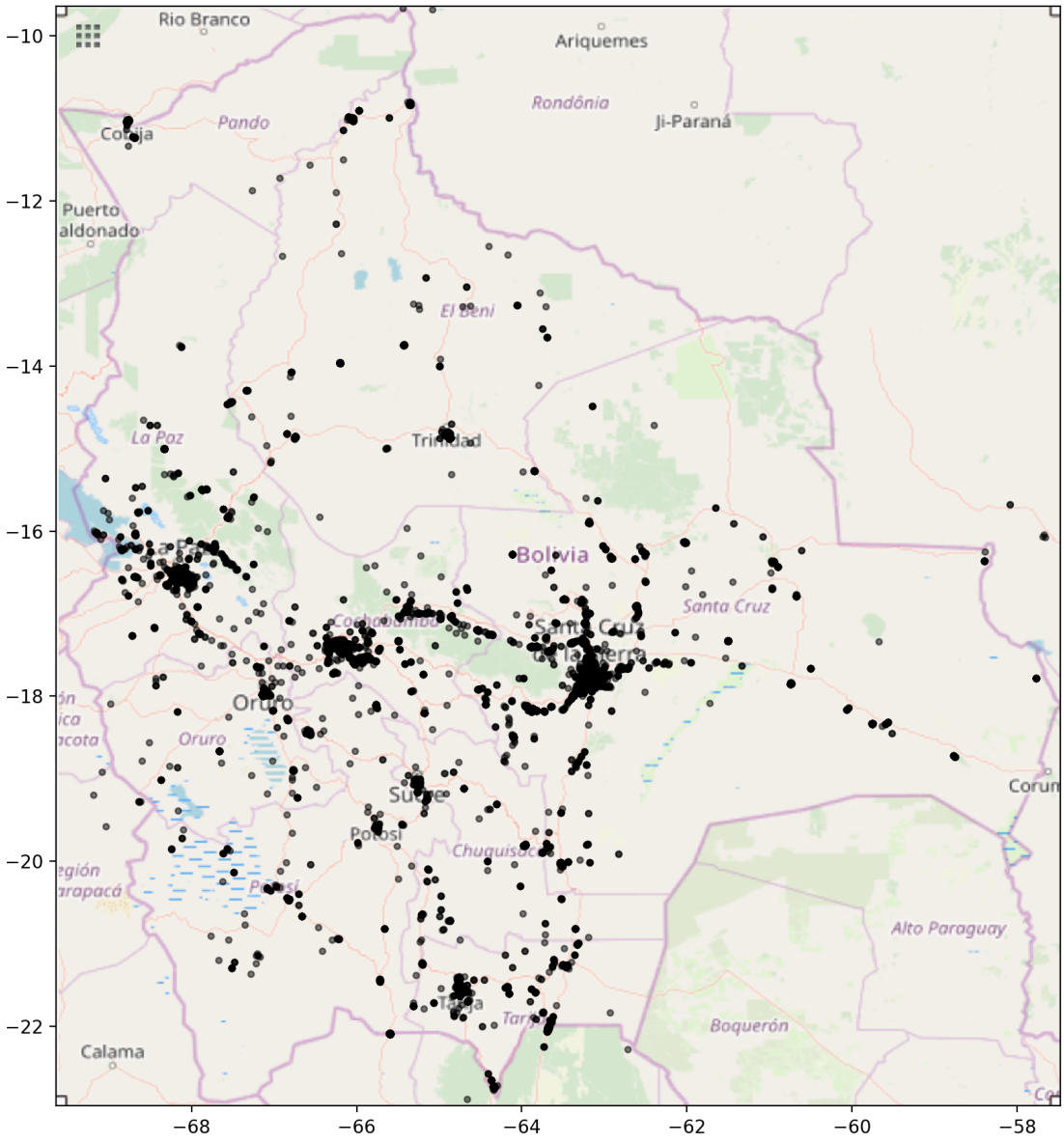

Figura 31: Localizaciones desde las cuales se emiten Tweets en Bolivia.

Fuente: Elaboración propia.

Esta representación gráfica se logró empleando la librería Python Matplotlib[118] y las coordenadas de ubicación geográfica de 98,766 Tweets emitidos por cuentas de usuario bolivianas.

[118] https://matplotlib.org

Producto del análisis de la desviación estándar de las opiniones a favor y en contra por cuenta de usuario, se determinó que 49.12% de las cuentas de usuario de Twitter que emitieron opinión respecto del tema de análisis, no son constantes en cuanto a sus opiniones. Es decir, cerca de la mitad de los usuarios no está seguros de su opinión promedio, son usuarios que emiten opiniones positivas y negativas.

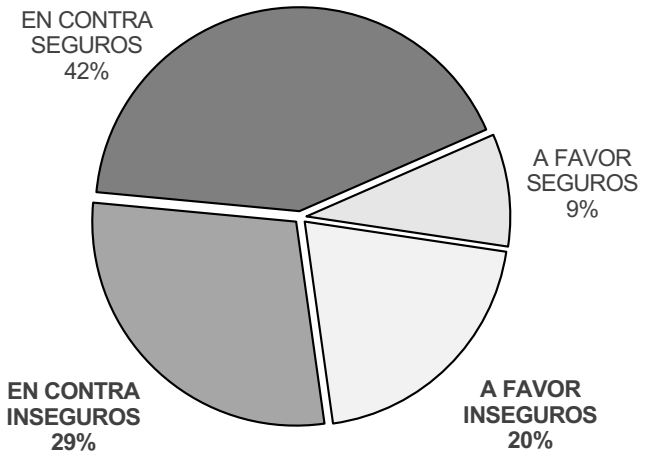

Figura 32: Clasificación de opiniones según seguridad de opinión.

Fuente: Elaboración propia.

5.1.4 Tiempos y Costos

El presente trabajo se ejecutó conforme el cronograma propuesto en el Perfil de Tesis, en un plazo de 148 días calendario.

Tuvo un costo de 324 horas hombre y 1,608 horas máquina, lo cual representa un costo total subvencionado de Bs. 32,400.00.

El proceso más costoso, en cuanto a horas máquina, fue la Recolección de Seguidores y Amigos, el cual ocupó el 41% del tiempo del proyecto. La base de datos de cuentas de usuario obtenida en este proceso representa el resultado con mayor valor del presente trabajo.

5.2 Conclusiones de la Investigación

5.2.1 Estado de los Objetivos

El objetivo general fue cumplido: se desarrolló un Modelo de Cuadro de Mando Integral para el monitoreo de sentimiento de la sociedad boliviana, respecto de una temática específica, con base en las opiniones expresadas por las personas en la red social Twitter.

El modelo se encuentra desarrollado en el “Capítulo III. Modelo Propuesto”.

En cuanto al estado de los objetivos específicos, todos se cumplieron tal como se describe a continuación:

a) Se investigaron las capacidades de interacción e integración a través de las Interfaces de Programación de Aplicaciones de Twitter.

   La investigación de estas capacidades está descrita en el apartado “3.3 Interacción e Integración con Twitter”.

b) Se seleccionaron los recursos tecnológicos de Ciencia de Datos adecuados para aplicar tareas de Análisis de Sentimiento.

   La selección está descrita en el apartado “3.4 Selección de Recursos Tecnológicos”.

c) Se diseñó el Modelo de Cuadro de Mando Integral para la explotación de la base de datos de Twitter aplicando Teoría de Grafos y Minería de Opinión con un enfoque de Análisis de Negocios.

   El diseño del modelo se encuentra en el apartado “3.5 Diseño del Modelo”. Las aplicaciones de Teoría de Grafos y Minería de Opinión están descritas en los apartados “3.5.5 Aplicación de Teoría de Grafos” y “3.5.6 Aplicación de Minería de Opinión” respectivamente. La

aplicación del enfoque de Análisis de Negocios está descrito en el apartado “3.5.7 Aplicación de Análisis de Negocios”.

d) Se implementó una versión demostrativa operativa del Modelo propuesto.

   La implementación de la versión demostrativa está descrita en el “Capítulo IV. Aplicación del Modelo”.

e) Se evaluaron los resultados de la versión demostrativa y se establecieron conclusiones.

   La evaluación de resultados y el establecimiento de conclusiones están desarrollados en el “Capítulo V. Marco de Resultados”.

5.2.2 Estado de la Hipótesis

Los resultados de la aplicación del modelo demuestran que **la hipótesis del presente trabajo es cierta**: la información que proporciona Twitter, a través de sus API, es suficiente para construir una muestra válida para una medición estadística a nivel nacional.

Los sondeos de opinión descritos en el apartado “2.2.3.1 Sondeos de Intención de Voto” se hicieron con muestras entre 800 y 2,250 individuos, con excepción del sondeo de la iniciativa “Tu Voto Cuenta” que se hizo con una muestra de 14,238 individuos.

La muestra obtenida en el presente trabajo es de 101,933 cuentas de usuario de Twitter, localizados en Bolivia, que emitieron opinión respecto del tema de análisis. Una muestra así tiene 0.31% de error muestral.

Si bien la distribución de las cuentas de usuario por departamentos es una aproximación basada en algoritmos de Teoría de Grafos, la muestra

obtenida por el presente trabajo es una muestra válida que representa a la población boliviana a nivel nacional tal como se puede evidenciar en la distribución geográfica de los Tweets presentada en el apartado “5.1.3 Resultados Adicionales”.

La construcción de la muestra no entra en conflicto con la legislación vigente, en cuanto a la privacidad de los datos de los usuarios de Twitter, y puede ser hecha a un costo menor que un sondeo de opinión encargado a una empresa especializada.

Por otro lado, la estimación hecha por la AGETIC establece que aproximadamente 1.1 millones de personas usan Twitter con base en una encuesta hecha el año 2016. El presente trabajo encontró un poco más de 2 millones de cuentas de usuario de Twitter lo cual respalda la estimación hecha por la AGETIC.

5.2.3 Limitaciones del Trabajo

Si bien los objetivos se cumplieron, la aplicación del modelo tiene limitaciones:

1. La muestra obtenida no asegura la representatividad de áreas rurales. De acuerdo al INE, el área rural considera a los centros poblados con menos de 2,000 habitantes y a las zonas propiamente dispersas.

2. El modelo no analiza respuestas a Tweets, Tweets favoritos, imágenes o videos.

5.3 Recomendaciones de la Investigación

En la búsqueda de información complementaria a los sondeos de opinión, el presente trabajo puede ser escalado a distintos niveles.

A un nivel productivo, los resultados del modelo permitirían a distintos actores de la sociedad conocer la opinión pública boliviana respecto de una temática de interés regional o nacional y tomar decisiones en consecuencia.

Entre los aspectos que pueden ser abordados en la continuidad del modelo están:

a) **Explotación**. El presente trabajo demuestra que el Análisis de Sentimiento puede ser empleado para identificar a las personas que no tienen una opinión constante respecto de una temática específica. Un análisis más minucioso de las opiniones de estos grupos de personas, permitiría identificar los aspectos aceptados o cuestionados de la temática. Esta información serviría de retroalimentación para estrategias de promoción, capacitación o información.

b) **Implementación en producción**. La implementación de un sistema informático dedicado a la explotación de los datos de Twitter es factible. El paso más costoso es la construcción de la base de datos de cuentas de usuario, sin embargo, una vez alcanzado este punto, la recolección de Tweets y la actualización de cuentas de usuario puede hacerse de forma diaria.

   Un sistema de este tipo puede soportar N temas de análisis los cuales pueden ser evaluados de forma paralela con resultados disponibles en función del grado de participación de los individuos de la muestra.

   Se puede ampliar la cantidad datos recolectados desde Twitter con el uso de las API Premium[119] o las API Empresariales[120].

[119] https://developer.twitter.com/en/premium-apis.html
[120] https://developer.twitter.com/en/enterprise.html

c) **Clasificación de cuentas de usuario**. Existen propuestas para determinar características de los usuarios de Twitter a partir de la descripción dada por los usuarios en sus perfiles[121]. Esta información podría permitir hacer clasificaciones de cuentas por género, edad o perfil profesional entre otras categorías. Estas clasificaciones permitirían obtener muestras más precisas para estudios más especializados.

d) **Generalización**. Facebook ofrece *API Graph*[122] la cual es una herramienta similar a la API de Twitter. Esta API es la principal forma de ingresar y extraer datos en la plataforma de Facebook. Está basada en HTTP y las aplicaciones de terceros pueden usarla para consultar datos y otras tareas siempre y cuando las condiciones de privacidad de sus usuarios lo permitan.

   WhatsApp ofrece *API WhatsApp Business*[123] la cual está dirigida a medianas y grandes empresas. El cliente de esta API admite las funciones ya conocidas de las aplicaciones de WhatsApp para Android, iOS y Web. Esta API permite la implementación de un servidor que puede enviar y recibir mensajes e integrar este proceso con sistemas de gestión de clientes.

   Youtube ofrece *YouTube Data API*[124] la cual está destinada a programadores que deseen crear aplicaciones que puedan interactuar con los contenidos de YouTube. Esta API proporciona recursos para obtener información de canales, videos, comentarios y respuestas a comentarios.

---

[121] https://pdfs.semanticscholar.org/0148/bbc80ea2f2526ab019a317639b4fb357f399.pdf
[122] https://developers.facebook.com/docs/graph-api
[123] https://www.whatsapp.com/business/api?lang=es
[124] https://developers.google.com/youtube/v3

Instagram ofrece *Instagram API Graph*[125] la cual permite a las aplicaciones acceder a datos en cuentas de Instagram para empresas y cuentas de creadores de Instagram. Con esta API se pueden crear aplicaciones para administrar comentarios, hashtags y medir interacciones sociales con otros usuarios de Instagram.

[125] https://developers.facebook.com/docs/instagram-api?locale=es_ES

ANEXO I. RELEVAMIENTO DE LAS API DE TWITTER

A continuación se describen los métodos relevantes al modelo, las descripciones fueron traducidas al español.

MÉTODO GET users/show

Devuelve la información sobre el usuario especificado por el parámetro *user_id* o *screen_name* requerido. En cuanto a las cuentas privadas, la cuenta con la cual se hace la consulta debe seguir al usuario privado para poder ver su estado.

Método para obtención de Metadatos de Cuentas.

Fuente: Documentación de API de Twitter.

PARÁMETRO	REQUERIDO	DESCRIPCIÓN
user_id	Requerido	El identificador del usuario del cual se solicita información.
screen_name	Requerido	El nombre de pantalla del usuario del cual se solicita información.
include_entities	Opcional	El nodo *entities* no será incluido en la información cuando se establece el valor del parámetro en *false*.

Ejemplo de invocación del método:

TIPO	EJEMPLO
Solicitud	`GET https://api.twitter.com/1.1/users/show.json ?screen_name=twitterdev`
Respuesta	`{user-object}`

MÉTODO GET followers/list

Devuelve una colección paginada de objetos de usuario para los usuarios que siguen al usuario especificado. Por defecto, los resultados son devueltos en páginas o grupos de 20 usuarios y se pueden navegar múltiples páginas de resultados mediante el uso del valor *next_cursor* en solicitudes posteriores.

En paralelo, se cuenta con el método ***GET followers/ids*** el cual devuelve una colección paginada de identificadores de usuario únicamente para cada usuario que sigue al usuario especificado. Ambos métodos tienen el mismo límite de velocidad de uso.

Método para obtención de Seguidores.

Fuente: Documentación de API de Twitter.

PARÁMETRO	REQUERIDO	DESCRIPCIÓN
user_id	Opcional	El identificador del usuario del cual se solicita información.
screen_name	Opcional	El nombre de pantalla del usuario del cual se solicita información.
cursor	Semi-opcional	Hace que los resultados se dividan en páginas. Si no se proporciona ningún *cursor*, se asumirá un valor de -1, que es la primera página. La respuesta de la API incluirá un *previous_cursor* y *next_cursor* para permitir la paginación de un lado a otro.
count	Opcional	El número de usuarios a devolver por página, hasta un máximo de 200. El valor predeterminado es 20.
skip_status	Opcional	Cuando se establece en *true, t* o *1*, los estados no se incluirán en los objetos de usuario devueltos. Si se establece en cualquier otro valor, se incluirán los estados.

PARÁMETRO	REQUERIDO	DESCRIPCIÓN
include_user_entities	Opcional	El nodo *entities* no será incluido en la información cuando se establece el valor del parámetro en *false*.

Ejemplo de invocación del método:

TIPO	EJEMPLO

Solicitud	`GET https://api.twitter.com/1.1/followers/list.json`
	`?cursor=-1`
	`&screen_name=twitterdev`
	`&skip_status=true`
	`&include_user_entities=false`

Respuesta	`{`
	`"users": [`
	`{user-object},`
	`{user-object},`
	`{user-object}`
	`],`
	`"next_cursor": 1489467234237774933,`
	`"next_cursor_str": "1489467234237774933",`
	`"previous_cursor": 0,`
	`"previous_cursor_str": "0"`
	`}`

MÉTODO GET friends/list

Devuelve una colección paginada de objetos de usuario para cada usuario que el usuario especificado está siguiendo. En esta relación, el usuario especificado es "amigo" del usuario al que sigue.

En paralelo, se cuenta con el método ***GET friends/ids*** el cual devuelve una colección paginada de identificadores de usuario únicamente por cada usuario que el usuario especificado está siguiendo. Ambos métodos tienen el mismo límite de velocidad de uso.

Método para obtención de Amigos.

Fuente: Documentación de API de Twitter.

PARÁMETRO	REQUERIDO	DESCRIPCIÓN
user_id	Opcional	El identificador del usuario del cual se solicita información.
screen_name	Opcional	El nombre de pantalla del usuario del cual se solicita información.
cursor	Semi-opcional	Hace que los resultados se dividan en páginas. Si no se proporciona ningún *cursor*, se asumirá un valor de -1, que es la primera página. La respuesta de la API incluirá un *previous_cursor* y *next_cursor* para permitir la paginación de un lado a otro.
count	Opcional	El número de usuarios a devolver por página, hasta un máximo de 200. El valor predeterminado es 20.
skip_status	Opcional	Cuando se establece en *true, t* o *1*, los estados no se incluirán en los objetos de usuario devueltos. Si se establece en cualquier otro valor, se incluirán los estados.
include_user_entities	Opcional	El nodo *entities* no será incluido en la información cuando se establece el valor del parámetro en *false*.

Ejemplo de invocación del método:

TIPO	EJEMPLO

Solicitud	`GET https://api.twitter.com/1.1/friends/list.json`
	`?cursor=-1`
	`&screen_name=twitterdev`
	`&skip_status=true`
	`&include_user_entities=false`

TIPO	EJEMPLO

Respuesta	<pre>{
	  "users": [
	      {user-object},
	      {user-object},
	      {user-object}
	  ],
	  "next_cursor": 0,
	  "next_cursor_str": "0",
	  "previous_cursor": 1333504313713126852,
	  "previous_cursor_str": "1333504313713126852"
	}</pre>

MÉTODO GET statuses/user_timeline

Devuelve una colección de los Tweets más recientes publicados por el usuario indicado por los parámetros *screen_name* o *user_id*.

Las líneas de tiempo o *timelines* de los usuarios pertenecientes a usuarios protegidos solo se pueden ser solicitados cuando el usuario autenticado es dueño de la línea de tiempo o es un seguidor aprobado por el usuario propietario de la misma.

La *timeline* devuelta es el equivalente al visto como perfil de usuario en Twitter. Este método solo puede devolver hasta 3,200 de los Tweets más recientes de un usuario. Los *retweets* de otros usuarios por parte del usuario se incluyen en este total, independientemente de si el parámetro *include_rts* esté establecido en *false* al invocar este método.

Método para obtención de Tweets.

Fuente: Documentación de API de Twitter.

PARÁMETRO	REQUERIDO	DESCRIPCIÓN
user_id	Opcional	El identificador del usuario del cual se solicita información.
screen_name	Opcional	El nombre de pantalla del usuario del cual se solicita información.

PARÁMETRO	REQUERIDO	DESCRIPCIÓN
since_id	Opcional	Devuelve resultados con un identificador mayor que el identificador especificado, es decir, devuelve los Tweets más recientes que el especificado con *since_id*. Existen límites para la cantidad de Tweets a los que se puede acceder a través de la API. Si se alcanza el límite de Tweets desde *since_id*, entonces se forzará a que *since_id* sea el identificador más antigüo disponible.
count	Opcional	Especifica el número de Tweets para recuperar, hasta un máximo de 200 por solicitud distinta. Se incluyen los *retweets* en el resultado, incluso si no se incluye el parámetro *include_rts* en la solicitud. Se recomienda que siempre envíe el parámetro *include_rts* con el valor 1 cuando use este método.
max_id	Opcional	Devuelve resultados con un identificador menor o igual al identificador especificado, es decir, devuelve los Tweets anteriores a *max_id*.
trim_user	Opcional	Cuando se establece en *true*, *t* o *1*, cada Tweet devuelto en una *timeline* incluirá un objeto de usuario que incluye solo el identificador de autor del Tweet. Se debe omitir este parámetro para recibir el objeto de usuario completo.
exclude_replies	Opcional	Este parámetro evitará que aparezcan respuestas en la *timeline* devuelta. El uso de *exclude_replies* con el parámetro de *count* significará que recibirá hasta *count* Tweets.
include_rts	Opcional	Cuando se establece en *false*, la *timeline* eliminará todos los *retweets*.

PARÁMETRO	REQUERIDO	DESCRIPCIÓN
		Si se usa el parámetro *trim_user* junto con *include_rts*, los *retweets* seguirán conteniendo un objeto de usuario completo.

Ejemplo de invocación del método:

TIPO	EJEMPLO

Solicitud	`GET https://api.twitter.com/1.1/statuses/`
	`  user_timeline.json`
	`  ?screen_name=twitterapi`
	`  &count=2`

Respuesta	`[`
	`  {tweet-object},`
	`  {tweet-object},`
	`  {tweet-object}`
	`]`

OBJETO USUARIO (user-object)

El objeto Usuario contiene metadatos de la cuenta de usuario de Twitter que describen al usuario de Twitter al que se hace referencia. Los usuarios pueden crear Tweets, Retweets, citar otros Tweets de Usuarios, responder a Tweets, seguir a Usuarios, ser mencionados en Tweets y pueden agruparse en listas.

En general, los valores de metadatos de las cuentas de usuario son relativamente constantes. Algunos campos nunca cambian, como la identificación del usuario *id_str* y cuándo se creó la cuenta. Otros metadatos pueden cambiar ocasionalmente, como el nombre de pantalla *screen_name*, el nombre para mostrar, la descripción, la ubicación y otros detalles del perfil. Algunos metadatos cambian con frecuencia, como el número de Tweets que la cuenta ha publicado en *statuses_count* y su número de seguidores en *followers_count*.

Objeto Usuario de Twitter.

Fuente: Documentación de API de Twitter.

ATRIBUTO	TIPO	DESCRIPCIÓN
id	Int64	La representación entera del identificador único para este usuario.
id_str	String	La representación de cadena del identificador único para este usuario. Las implementaciones deberían usar este atributo en lugar del número entero grande dado en el atributo *id*.
name	String	El nombre del usuario, tal como lo ha definido el usuario. No necesariamente es el nombre real de una persona. Normalmente tiene un límite de 50 caracteres, pero está sujeto a cambios.
screen_name	String	El nombre de pantalla, el identificador o el alias con el que este usuario se identifica. Los nombres de pantalla son únicos, pero están sujetos a cambios. Se debe utilizar *id_str* como identificador de usuario siempre que sea posible. Por lo general, tiene un máximo de 15 caracteres, pero algunas cuentas históricas pueden existir con nombres más largos.
location	String	La ubicación definida por el usuario para el perfil de esta cuenta. No necesariamente es una ubicación, ni es analizable por máquina. Este campo ocasionalmente es interpretado de forma difusa por el servicio de búsqueda de Twitter.
derived	Array of Objects	*API empresarial de colección de metadatos de enriquecidos para el usuario.* Proporciona los metadatos del perfil geográfico del usuario enriquecidos.

ATRIBUTO	TIPO	DESCRIPCIÓN
url	String	Una URL proporcionada por el usuario en asociación con su perfil.
description	String	La cadena UTF-8 definida por el usuario que describe su cuenta.
protected	Boolean	Cuando es *true*, indica que este usuario ha elegido proteger sus Tweets.
verified	Boolean	Cuando es *true*, indica que el usuario tiene una cuenta verificada.
followers_count	Int	El número de seguidores que esta cuenta tiene actualmente.
friends_count	Int	El número de usuarios que esta cuenta está siguiendo, también conocidos como sus "amigos".
listed_count	Int	El número de listas públicas de las que este usuario es miembro.
favourites_count	Int	La cantidad de Tweets que le han gustado a este usuario durante la vida útil de la cuenta.
statuses_count	Int	El número de Tweets, incluidos Retweets, emitidos por el usuario.
created_at	String	La fecha y hora UTC en que se creó la cuenta de usuario en Twitter.
profile_banner_url	String	La URL basada en HTTPS que apunta a la representación web estándar del banner del perfil cargado del usuario.
profile_image_url_https	String	Una URL basada en HTTPS que apunta a la imagen de perfil del usuario.
default_profile	Boolean	Cuando es *true*, indica que el usuario no ha alterado el tema o el fondo de su perfil de usuario.
default_profile_image	Boolean	Cuando es *true*, indica que el usuario no ha cargado su propia imagen de perfil y en su lugar se usa una imagen predeterminada.

ATRIBUTO	TIPO	DESCRIPCIÓN
withheld_in_countries	Array of String	Cuando está presente, indica una lista de códigos de país en mayúsculas de dos letras de los que se retiene este contenido. Twitter admite los valores de la lista de códigos de país definidos en el estándar ISO 3166-1 alpha-2. Más detalles: *https://www.iso.org/obp/ui/*
withheld_scope	String	Cuando está presente, indica que el contenido retenido es un "usuario".

OBJETO TWEET (tweet-object)

Los Tweets son el bloque de construcción atómico básico de todas las cosas de Twitter. También son conocidos como "actualizaciones de estado". El objeto Tweet tiene una larga lista de atributos, incluidos los atributos fundamentales como *id*, *created_at* y *text*.

Los objetos Tweet también son el objeto principal de varios objetos secundarios. Los objetos secundarios incluyen: *user*, *entities* y *extended_entities*. Los Tweets que están geo-etiquetados tendrán un objeto secundario *place*.

Objeto Tweet de Twitter.

Fuente: Documentación de API de Twitter.

ATRIBUTO	TIPO	DESCRIPCIÓN
id	Int64	La representación entera del identificador único para este Tweet.
id_str	String	La representación de cadena del identificador único para este Tweet. Las implementaciones deberían usar este atributo en lugar del entero grande del atributo *id*.

ATRIBUTO	TIPO	DESCRIPCIÓN
created_at	String	Hora UTC cuando se creó este Tweet.
text	String	El texto real del Tweet en formato UTF-8.
source	String	Aplicación utilizada por el usuario para publicar el Tweet, como una cadena con formato HTML.
truncated	Boolean	Indica si el valor del texto se truncó, por ejemplo, como resultado de un retweet que excede el límite de longitud del texto del Tweet original de 140 caracteres.
in_reply_to_status_id	Int64	Si el Tweet representado es una respuesta, este campo contendrá la representación entera del identificador del Tweet original.
in_reply_to_status_id_str	String	Si el Tweet representado es una respuesta, este campo contendrá la representación de cadena del identificador del Tweet original.
in_reply_to_user_id	Int64	Si el Tweet representado es una respuesta, este campo contendrá la representación entera del identificador del usuario autor del Tweet original.
in_reply_to_user_id_str	String	Si el Tweet representado es una respuesta, este campo contendrá la representación de cadena del identificador del usuario autor del Tweet original.
in_reply_to_screen_name	String	Si el Tweet representado es una respuesta, este campo contendrá el nombre de pantalla del usuario autor original del Tweet.
user	User Object	El usuario que publicó este Tweet.
coordinates	Coordinates	Representa la ubicación geográfica de este Tweet según lo informado por el usuario o la aplicación empleada.

ATRIBUTO	TIPO	DESCRIPCIÓN
		La matriz de coordenadas internas sigue el formato geoJSON (longitud primero, luego latitud). Más detalles: *https://geojson.org/*
place	Geo Object	Cuando está presente, indica que el Tweet está asociado (pero no necesariamente desde) un lugar.
quoted_status_id_str	String	Este campo solo aparece cuando el Tweet es una cita. Este campo contiene el valor entero del identificador del Tweet citado.
is_quote_status	Boolean	Indica si se trata de un Tweet citado.
quoted_status	Tweet	Este atributo solo aparece cuando el Tweet es una cita. Este atributo contiene el objeto Tweet del Tweet original que fue citado.
retweeted_status	Tweet	Este atributo contiene una representación del Tweet original que fue retuiteado. Los retweet de retweet no muestran representaciones del retweet intermediario, sino solo el Tweet original.
quote_count	Integer	Indica aproximadamente cuántas veces los usuarios de Twitter han citado este Tweet.
reply_count	Int	Número de veces que este Tweet ha sido respondido.
retweet_count	Int	Número de veces que este Tweet ha sido retuiteado.
favorite_count	Integer	Indica aproximadamente cuántas veces le han gustado los usuarios de Twitter a este Tweet.
entities	Entities Objects	Entidades que han sido analizadas del texto del Tweet.

ATRIBUTO	TIPO	DESCRIPCIÓN
extended_entities	Extended Entities	Contiene una matriz de metadatos multimedia de fotos nativas, un video o un GIF animado en el Tweet.
favorited	Boolean	Indica si este Tweet ha sido marcado como favorito por el usuario autenticado.
retweeted	Boolean	Indica si este Tweet ha sido retuiteado por el usuario autenticado.
possibly_sensitive	Boolean	Este campo solo aparece cuando un Tweet contiene un enlace. El significado del campo no se refiere al contenido del Tweet en sí, sino que es un indicador de que la URL contenida en el Tweet puede contener contenido o medios identificados como contenido confidencial.
filter_level	String	Indica el valor máximo del parámetro *filter_level* que se puede usar y aún transmitir este Tweet.
lang	String	Cuando está presente, indica un identificador de idioma correspondiente al idioma detectado por la máquina del texto del Tweet, o si no se pudo detectar ningún idioma. El identificador sigue la normal BCP 47. Más detalles: *https://tools.ietf.org/html/bcp47*
matching_rules	Array of Rule Objects	Presente en productos filtrados como Twitter Search y PowerTrack. Proporciona la identificación y la etiqueta asociadas con la regla que coincide con el Tweet.

CÓDIGOS DE RESPUESTA

La API estándar de Twitter devuelve códigos de estado HTTP además de códigos y mensajes de error basados en JSON.

A continuación, la lista de los códigos de estado HTTP relevantes al modelo.

Códigos de respuesta de métodos de API de Twitter.

Fuente: Documentación de API de Twitter.

CÓDIGO	TEXTO	DESCRIPCIÓN
200	OK	Atención exitosa de la solicitud.
400	Bad Request	La solicitud no es válida o no es posible atenderla. La respuesta es acompañada de un mensaje de error con mayores detalles. Las solicitudes sin autenticación se consideran inválidas y generarán esta respuesta.
401	Unauthorized	Credenciales de autenticación faltantes o incorrectas.
403	Forbidden	La solicitud es correcta, pero se ha rechazado o no se permite el acceso. Un mensaje de error adjunto explicará las razones del rechazo. Este código se utiliza cuando se rechazan solicitudes debido a límites de actualización.
404	Not Found	El URI solicitado no es válido o el recurso solicitado, como un usuario, no existe.
429	Too Many Requests	Se devuelve cuando no se puede atender una solicitud debido a que se ha superado el límite de velocidad de la aplicación para el método.
500	Internal Server Error	Error temporal del método invocado. Por ejemplo, en una situación de alta carga o si un método tiene problemas temporalmente.

CÓDIGO	TEXTO	DESCRIPCIÓN
502	Bad Gateway	Twitter no está disponible o se está actualizando.
503	Service Unavailable	Los servidores de Twitter están activos, pero sobrecargados de solicitudes. Se debe intentar nuevamente más tarde.

El detalle completo de códigos de respuesta API estándar de Twitter se encuentra en:

https://developer.twitter.com/en/docs/basics/response-codes.html

ANEXO II. DISEÑO DE LA BASE DE DATOS

Para el almacenamiento de los datos requeridos por el modelo se definieron seis entidades, ésta entidades guardan datos de usuarios, Tweets y eventos relativos al tema de análisis del modelo.

El diagrama entidad-relación de la base de datos es:

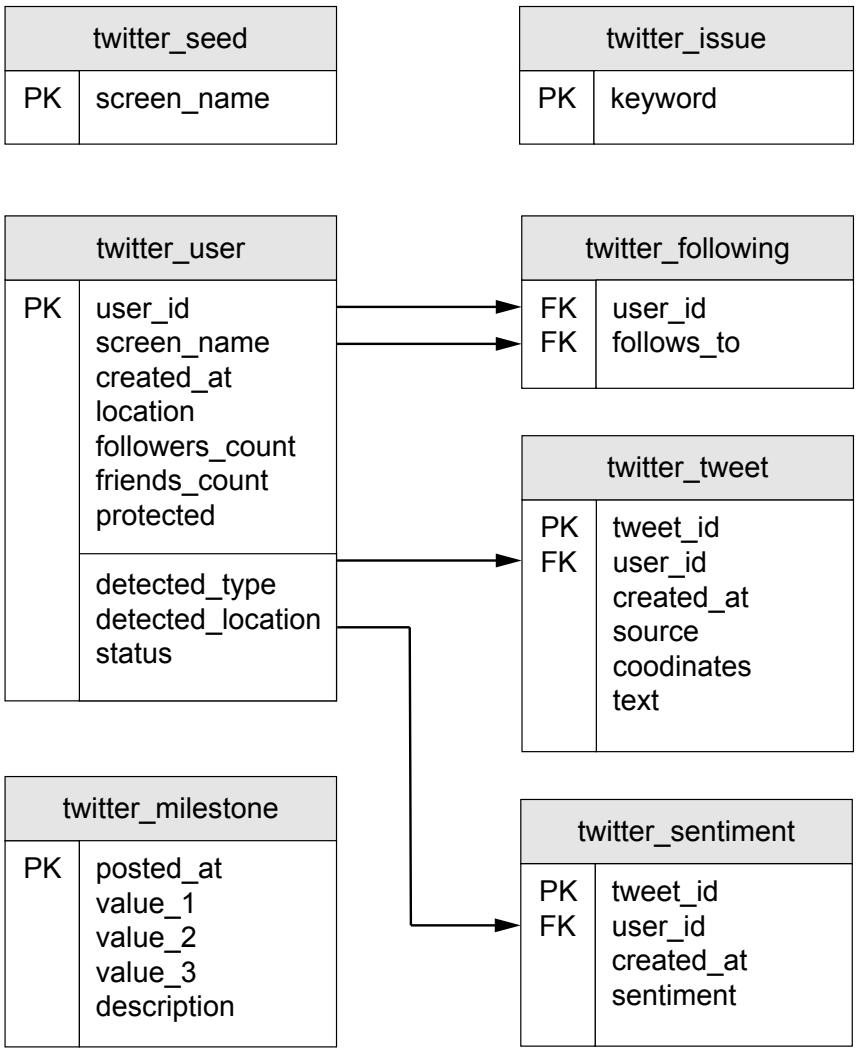

Figura 33: Modelo entidad-relación de la base de datos del modelo.

Fuente: Elaboración propia.

ENTIDAD SEMILLA (twitter_seed)

Contiene los datos de las cuentas semilla para el modelo.

Entidad Semilla de la base de datos del modelo.

Fuente: Elaboración propia.

CAMPO	TIPO	DESCRIPCIÓN
screen_name	varchar(50)	Nombre de pantalla, el identificador o el alias con el que el usuario semilla se identifica.

ENTIDAD USUARIO (twitter_user)

Contiene los datos básicos de los usuarios de Twitter, los cuales son relevantes para el modelo.

Entidad Usuario de la base de datos del modelo.

Fuente: Elaboración propia.

CAMPO	TIPO	DESCRIPCIÓN
user_id	varchar(30)	Identificador único del usuario de Twitter.
screen_name	varchar(50)	Nombre de pantalla, el identificador o el alias con el que este usuario se identifica.
created_at	timestamp	Fecha y hora en la que el usuario fue creado.
location	varchar(150)	La ubicación definida por el usuario para su perfil.
followers_count	integer	Número de seguidores que tiene el usuario en el momento de la consulta a las API de Twitter.
friends_count	integer	Número de cuentas que siguen al usuario en el momento de la consulta a la API de Twitter.
found_at	timestamp	Fecha y hora en la que el usuario fue obtenido a través de las API de Twitter.

CAMPO	TIPO	DESCRIPCIÓN
protected	varchar(20)	Indicador de cuenta protegida: • True=El usuario es privado, no se puede obtener Tweets, seguidores y amigos. • False=El usuario es público, se puede obtener Tweets, seguidores y amigos. • Undefined=Sin determinar por error en la consulta a la API de Twitter o por inhabilitación de Twitter.
detected_type	varchar(10)	Tipo de usuario detectado en el momento en que se hizo la recuperación de datos a través de las API de Twitter. • User suspended=Usuario suspendido. • User not found=Usuario no encontrado. • User private=Usuario privado. • User public=Usuario público.
detected_location	varchar(30)	Departamento donde se encuentra el usuario, detectado producto del análisis del campo *location*. • BENI • CHUQUISACA • COCHABAMBA • LA PAZ • ORURO • PANDO • POTOSI • SANTA CRUZ • TARIJA

CAMPO	TIPO	DESCRIPCIÓN
		• EXTERIOR=Usuario del exterior del país. • PENDING-BO=Usuario del país, pendiente de determinación. • PENDING-EMPTY=Usuario sin localización, pendiente de determinación.
status	varchar(50)	Estado del usuario luego del procesamiento para obtener seguidores, amigos y Tweets. • CHECKED_FF=Cuenta procesada para obtención de seguidores y amigos. • CHECKED_MM=Cuenta procesada para obtención de Tweets. • CHECKED_FFMM=Cuenta procesada para obtención de seguidores, amigos y Tweets. • UNCHECKED=Cuenta no procesada para obtención de seguidores, amigos o Tweets.

ENTIDAD RELACION (twitter_following)

Contiene la relación entre Usuarios de Twitter. La relación es usuario con el identificador *user*_id sigue al usuario con el identificador *follows_to*.

Entidad Relación de la base de datos del modelo.

Fuente: Elaboración propia.

COLUMNA	TIPO	DESCRIPCIÓN
user_id	varchar(30)	Identificador único del usuario que sigue al usuario con identificador *following_to*.

COLUMNA	TIPO	DESCRIPCIÓN
follows_to	varchar(30)	Identificador único del usuario a quien el usuario con *user_id* sigue.

ENTIDAD TWEET (twitter_tweet)

Contiene los datos básicos de Tweets escritos o compartidos por los usuarios.

Entidad Tweet de la base de datos del modelo.

Fuente: Elaboración propia.

COLUMNA	TIPO	DESCRIPCIÓN
user_id	varchar(30)	Identificador único del usuario.
tweet_id	varchar(30)	Identificador único del Tweet.
created_at	timestamp	Fecha y hora en la que el Tweet fue creado.
source	varchar(30)	Aplicación utilizada para publicar el Tweet.
coordinates	varchar(200)	Ubicación geográfica del Tweet según lo informado por el usuario o la aplicación empleada.
text	varchar(350)	Texto del Tweet sin caracteres de control o símbolos.

ENTIDAD SENTIMIENTO (twitter_sentiment)

Contiene los resultados del análisis de sentimiento de Tweets relativos al tema de análisis del modelo.

Entidad Sentimiento de la base de datos del modelo.

Fuente: Elaboración propia.

COLUMNA	TIPO	DESCRIPCIÓN
user_id	varchar(30)	Identificador único del usuario.
tweet_id	varchar(30)	Identificador único del Tweet.

COLUMNA	TIPO	DESCRIPCIÓN
created_at	timestamp	Fecha y hora en la que el Tweet fue creado.
sentiment	varchar(10)	Clasificación determinada del sentimiento, a favor o en contra, del Tweet respecto del tema de análisis del modelo. • pos=El Tweet está a favor del tema. • neg=El Tweet está en contra del tema.

ENTIDAD TEMA (twitter_issue)

Contiene las palabras claves relativas al tema de análisis del modelo. Estas palabras son empleadas para determinar si un Tweet es relativo o no al tema de análisis del modelo.

Entidad Tema de la base de datos del modelo.

Fuente: Elaboración propia.

COLUMNA	TIPO	DESCRIPCIÓN
keyword	varchar(50)	Palabra clave relativa al tema de análisis del modelo.

ENTIDAD HITO (twitter_milestone)

Contiene los eventos de referencia relativos al tema de análisis del modelo de carácter estadístico. Estos eventos sirven para comparar o comprender el comportamiento de los indicadores evaluados.

Entidad Hito de la base de datos del modelo.

Fuente: Elaboración propia.

COLUMNA	TIPO	DESCRIPCIÓN
posted_at	timestamp	Fecha y hora en la cual el evento fue publicado.

COLUMNA	TIPO	DESCRIPCIÓN
value_1	numeric	Primer valor estadístico publicado por el evento.
value_2	numeric	Segundo valor estadístico publicado por el evento.
value_3	numeric	Tercer valor estadístico publicado por el evento.
description	varchar(200)	Descripción adicional del evento.

ANEXO III. CREACIÓN DEL ENTORNO DE DESARROLLO

El entorno de desarrollo del proyecto se construyó en una estación de trabajo con el sistema operativo MacOS Sierra (versión 10.12.6), la cual tiene instalado el intérprete Python versión 2.7.10 y el ejecutor de aplicaciones Java – JRE versión 1.8.0.

Para la creación del entorno de desarrollo se siguieron los siguientes pasos:

1) Se descargó e instaló la última versión disponible de Python (versión 3.7.3) del sitio: *https://www.python.org/downloads/*.

   Se siguieron los pasos indicados por el instalador (archivo `python-3.7.3-macosx10.9.pkg`).

2) Se descargó e instaló la última versión disponible de la plataforma Anaconda para la gestión de contenidos (versión 2019.03) del sitio:

   https://www.anaconda.com/distribution/#download-section

   Se siguieron los pasos indicados por el instalador (archivo `Anaconda3-2019.03-MacOSX-x86_64.pkg`).

3) Se descargó e instaló la última versión disponible del IDE PyCharm (versión *Community* 2019.1.3) del sitio:

   https://www.jetbrains.com/pycharm/download/#section=mac

   Se siguieron los pasos indicados por el instalador (archivo `pycharm-community-2019.1.3.dmg`).

4) Se creó el proyecto Sento con el IDE PyCharm con el intérprete Python 3.7 y el entorno virtual Sento.

La estructura del proyecto es:

a. Directorio `core` para *scripts* Python de recolección de datos.

b. Directorio `cfg` para archivos de configuración.

c. Directorio `log` para archivos de bitácoras de ejecución.

d. Directorio `sql` para archivos SQL.

e. Directorio `cql` para archivos CQL (*Cypher Query Language*).

f. Directorio `tmp` para archivos temporales.

5) Se instaló el módulo Python para conexión con bases de datos PostgreSQL (`psycopg2`) y módulo Python para acceder a la API de Twitter (`tweepy`), ambos en el entorno virtual Sento.

   La instalación se hizo desde el mismo IDE PyCharm cuando se definió el *script* principal del proyecto y el IDE habilitó la opción de instalación automática de los módulos declarados.

6) Se instaló el entorno interactivo Jupyter Notebook desde el gestor Anaconda Navigator, en el entorno virtual Sento, para el desarrollo de código Python dedicado a Ciencia de Datos.

7) Se descargó e instaló la última versión disponible del motor de bases de datos orientada a grafos Neo4j (versión *Community Edition* 3.5.8) del sitio:

   https://neo4j.com/download-center/#community

   Se descomprimió y desempaquetó el archivo `neo4j-community-3.5.8-unix.tar.gz`.

   Se descargó también del mismo sitio las librerías de procedimientos para el uso de algoritmos (archivo `neo4j-graph-algorithms-`

`3.5.8.1-standalone.zip`). El archivo descomprimido se movió al directorio `neo4j-community-3.5.8/plugins/`.

8) Se agregó la configuración de procedimientos para el uso de algoritmos de detección de comunidades en el archivo `neo4j.conf`:

```
dbms.security.procedures.unrestricted=apoc.*, algo.*
dbms.security.procedures.whitelist=apoc.*, algo.*
```

9) Se ejecutó el siguiente *script* para iniciar el motor:

```
$ neo4j-community-3.5.8/bin/neo4j start
```

Se abrió la siguiente dirección en un navegador web para desplegar el navegador de grafos:

http://localhost:7474/

La contraseña por defecto de la cuenta neo4j es `neo4j`.

El repositorio para la importación de datos se encuentra en el directorio `neo4j-community-3.5.8/import`.

10) Se instaló el módulo del Kit de Herramientas para Procesamiento del Lenguaje Natural NTLK (`nltk`) en el entorno virtual Sento.

La instalación se hizo desde el mismo IDE PyCharm, el IDE habilitó la opción de instalación automática. La instalación de los Datos del Kit se hizo siguiendo las instrucciones descritas en el sitio:

https://www.nltk.org/data.html

11) Se instaló el módulo de la librería para Procesamiento del Lenguaje Natural TextBlob (`textblob`) en el entorno virtual Sento.

La instalación de la librería se hizo siguiendo las instrucciones dadas el siguiente sitio desde la Terminal del IDE:

https://textblob.readthedocs.io/en/dev/install.html

12) Se instalaron las librerías del framework Python para crear aplicaciones web Dash (`dash`, `dash_core_components`, `dash_html_components`) en el entorno virtual Sento.

Además, se instalaron las librerías para Minería de Datos y Análisis de Datos scikit-learn (`sklearn`, `numpy`) y las librerías gráficas Plotly (`plotly`) en el entorno virtual Sento.

Las instalaciones de las librerías se hicieron siguiendo las instrucciones dadas en el siguiente sitio desde la Terminal del IDE:

https://dash.plot.ly/installation

https://scikit-learn.org/stable/install.html

13) Se descargó e instaló la última versión disponible de la herramienta de desarrollo y administración de bases de datos multiplataforma DBeaver (versión *Community Edition* 6.1.0) del sitio:

https://dbeaver.io/download/

Se siguieron los pasos indicados por el instalador (archivo `dbeaver-ce-6.1.0-macos.dmg`).

ANEXO IV. CREACIÓN DE CUENTA DE DESARROLLO DE TWITTER

Para la creación de una cuenta de desarrollo en Twitter se siguieron los siguientes pasos:

1) Se creó una cuenta de correo electrónico Gmail desde el sitio *https://accounts.google.com/SignUp* con los siguientes datos:

    a. Nombre y Apellidos: Sentiment BSC

    b. Nombre de usuario: *sentimentbsc@gmail.com*

    El procedimiento seguido es el indicado en:

    https://support.google.com/mail/answer/56256?hl=es-419 (español)

    https://support.google.com/mail/answer/56256?hl=en (inglés)

2) Se creó una cuenta en Twitter desde el sitio *https://twitter.com/i/flow/signup* con los siguientes datos:

    a. Nombre: sentimentbsc

    b. Correo electrónico: *sentimentbsc@gmail.com*

    El procedimiento seguido es el indicado en:

    https://help.twitter.com/es/using-twitter/create-twitter-account (español)

    https://help.twitter.com/en/using-twitter/create-twitter-account (inglés)

3) Se solicitó una cuenta de desarrollo en Twitter desde el sitio *https://developer.twitter.com/en/apply/user* con los mismos datos de la cuenta de Twitter. Se indicó que la solicitud corresponde a una investigación de carácter académico y que se hará solamente análisis de datos de Twitter. Se indicó que no se usará la cuenta para la publicación de contenidos (Tweet, Retweet o Me gusta), no se mostrará

información de Twitter fuera de Twitter y que no se proporcionará información de Twitter a entidades gubernamentales.

4) Se aceptaron los términos del Acuerdo de Desarrollador y la Política de Desarrollador de Twitter.

   Los términos del Acuerdo de Desarrollador se encuentran en línea en:

   https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

5) Se confirmó la solicitud de cuenta de desarrollo a través del correo electrónico que envió Twitter para validar la misma. Más adelante se recibió la confirmación de que la solicitud fue aprobada vía correo electrónico.

6) Se creó una aplicación desde el sitio *https://developer.twitter.com/en/apps* con los siguientes datos:

   a. Nombre de la aplicación: Monitor de Sentimiento
   b. Descripción: Monitor de Sentimiento
   c. URL del sitio: *https://www.twitter.com*
   d. Uso de la aplicación: Determinar si Twitter es una mejor fuente para encuestas de opinión.
   e. Los demás datos requeridos en el formulario se dejaron en blanco.

7) Se obtuvieron las credenciales de la aplicación en la sección "*Keys and tokens*" en la misma interfaz de Twitter. Las credenciales son:

   a. Dos claves de consumidor de la API (*Consumer API keys*) denominadas "*API key*" y "*API secret key*".

b. Dos claves de acceso (*Access token & access token secret*) denominadas "*Access token*" y "*Access token secret*".

ANEXO V. DEFINICIÓN DE CUENTAS SEMILLA

Para definición de las cuentas semilla (*seed-accounts*) se utilizaron los reportes de estadísticas de uso de Twitter por tipo de cuenta del portal de **SocialBakers.com** tal como se detalla a continuación:

1) Se recuperó la lista de las 10 cuentas de Marcas con más seguidores en Bolivia desde la página:

   https://www.socialbakers.com/statistics/twitter/profiles/bolivia/brands/

Tabla 26: Lista de Marcas con más seguidores al 10/06/2019.
Fuente: SocialBakers.com

Nº	CUENTA	AMIGOS	SEGUIDORES
1	BoABolivia (@BoABolivia)	16,867	39,314
2	HuaweiMobileBolivia (@HuaweiMobileBo)	36	33,288
3	ENTEL S.A. (@entelbo)	163	24,935
4	Tigo Bolivia (@Tigo_Bolivia)	173	17,295
5	Coca-Cola Bolivia (@CocaColaBo)	18	15,826
6	VIVA (@vivabo)	1,886	13,529
7	Cedib (@cedib_com)	6,993	12,734
8	Paceña (@CervezaPacena)	3,561	10,259
9	ATT Bolivia (@ATTBolivia)	392	7,907
10	Cambio Sí Bolivia (@CambioSiBolivia)	1,932	5,510

2) Se recuperó la lista de las 10 cuentas de Celebridades con más seguidores en Bolivia desde la página:

   https://www.socialbakers.com/statistics/twitter/profiles/bolivia/celebrities/

Tabla 27: Lista de Celebridades con más seguidores al 10/06/2019.

Fuente: SocialBakers.com

Nº	CUENTA	AMIGOS	SEGUIDORES
1	Carlos D. Mesa Gisbert (@carlosdmesag)	380	445,664
2	Carlos Valverde B. (@CFValverde)	813	196,995
3	Ximena Herrera (@ximherrera)	895	177,960
4	Mike Amigorena (@Mik3amigorena)	66	116,513
5	Chavo Salvatierra (@chavo_mx1)	406	90,512
6	Carlos Emilio Lampe (@CarlosLampe1)	136	32,938
7	JUAN CARLOS ARANA (@JUANCARLOSARANA)	1,925	26,168
8	Adriana Caicedo (@adrianitaca)	2,800	14,937
9	Marcelo Moreno (@MM9oficial)	14	7,149
10	Gisely Ayub (@GiselyAyub)	214	730

3) Se recuperó la lista de las 10 cuentas de Comunidad con más seguidores en Bolivia desde la página:

https://www.socialbakers.com/statistics/twitter/profiles/bolivia/community/

Tabla 28: Lista de Comunidades con más seguidores al 10/06/2019.

Fuente: SocialBakers.com

Nº	CUENTA	AMIGOS	SEGUIDORES
1	Unitel Bolivia (@unitelbolivia)	16	**386,840**
2	Futbol de Bolivia (@futbolbolivia)	312	**200,523**
3	ComandoBolívarChávez (@CBolivarChavez)	396	**176,119**
4	Eulogio Del Pino (@delpinoeulogio)	98	**94,553**
5	Forever Stiven (@foreverstiven)	88	**72,105**
6	EntrenadorLatino (@EntrenadorLatin)	1,199	**56,085**
7	Congreso dl Pueblos (@C_Pueblos)	4,585	**38,114**

Nº	CUENTA	AMIGOS	SEGUIDORES
8	L (@LaMalaPalabra)	488	**27,490**
9	mclanfranconi (@mclanfranconi)	12,492	**27,018**
10	Pat Noticias (@patnoticias)	12	**17,639**

4) Se recuperó la lista de las 5 cuentas disponibles de Entretenimiento con más seguidores en Bolivia desde la página:

https://www.socialbakers.com/statistics/twitter/profiles/bolivia/entertainment/

Tabla 29: Lista de Cuentas de Entretenimiento con más seguidores al 10/06/2019.
Fuente: SocialBakers.com

Nº	CUENTA	AMIGOS	SEGUIDORES
1	CALLE 7 FANS (@FansC7tc)	333	74,057
2	Festival Cine Bogotá (@Bogocine)	1,314	21,442
3	CineCenterSC (@cinecentersc)	649	2,595
4	Don Multicine (@DonMulticine)	364	1,788
5	NotivisionBO (@NotivisionBO)	1,047	782

5) Se recuperó la lista de las 10 cuentas de Medios de Comunicación con más seguidores en Bolivia desde la página:

https://www.socialbakers.com/statistics/twitter/profiles/bolivia/media/

Tabla 30: Lista de Medios de Comunicación con más seguidores al 10/06/2019.
Fuente: SocialBakers.com

Nº	CUENTA	AMIGOS	SEGUIDORES
1	EL DEBER (@diarioeldeber)	42	432,284
2	La Razón Digital (@LaRazon_Bolivia)	255	420,390
3	ATB (@ATBDigital)	65	355,226
4	Página Siete (@pagina_siete)	459	332,789

Nº	CUENTA	AMIGOS	SEGUIDORES
5	ERBOL (@ErbolDigital)	210	285,057
6	Pat Bolivia (@patboliviatv)	21	276,930
7	Agencia Fides (ANF) (@noticiasfides)	1,012	215,064
8	Los Tiempos (@LosTiemposBol)	290	207,757
9	El Día (@Diario_ElDia)	488	138,325
10	Bolivia TV Oficial (@Canal_BoliviaTV)	995	41,658

6) Se recuperó la lista de las 5 cuentas disponibles de Lugares con más seguidores en Bolivia desde la página:

https://www.socialbakers.com/statistics/twitter/profiles/bolivia/place/

Tabla 31: Lista de Cuentas de Lugares con más seguidores al 10/06/2019.
Fuente: SocialBakers.com

Nº	CUENTA	AMIGOS	SEGUIDORES
1	SoySantaCruzBolivia (@SoySCBolivia)	3,941	**27,421**
2	Turismo Bolivia (@ViceminTurismo)	1,803	**10,757**
3	ZonaFrancadeBogotá (@ZFbta)	1,424	**5,198**
4	Boliviaentusmanos (@boliviaetm)	146	**970**
5	hoybolivia (@hoybolivia)	6	**900**

7) Se recuperó la lista de las 10 cuentas de Sociedad con más seguidores en Bolivia desde la página:

https://www.socialbakers.com/statistics/twitter/profiles/bolivia/society/

Tabla 32: Lista de Cuentas de Sociedad con más seguidores al 10/06/2019.
Fuente: SocialBakers.com

Nº	CUENTA	AMIGOS	SEGUIDORES
1	Evo Morales Ayma (@evoespueblo)	28	549,587
2	Samuel Doria Medina (@SDoriaMedina)	728	166,703

Nº	CUENTA	AMIGOS	SEGUIDORES
3	Min. de Comunicación (@mincombolivia)	435	126,209
4	Tuto Quiroga (@tutoquiroga)	78	55,487
5	Min. Economía (@EconomiaBo)	59	43,196
6	Rubén Costas (@RubenCostasA)	497	43,178
7	Salud Bolivia (@MinSaludBolivia)	326	31,824
8	Ministerio Justicia (@MinJusticiaBol)	357	30,890
9	Luis Revilla Herrero (@LuisRevillaH)	1,740	30,638
10	Oscar Ortiz Antelo (@OscarOrtizA)	1,784	30,587

8) Se recuperó la lista de las 3 cuentas disponibles de Deportes con más seguidores en Bolivia desde la página:

https://www.socialbakers.com/statistics/twitter/profiles/bolivia/sport/

Tabla 33: Lista de Cuentas de Deportes con más seguidores al 10/06/2019.
Fuente: SocialBakers.com

Nº	CUENTA	AMIGOS	SEGUIDORES
1	Club Bolívar (@Bolivar_Oficial)	28	219,479
2	Club The Strongest (@ClubStrongest)	65	128,144
3	FBF (@fbf_oficial)	171	20,146

9) Se combinaron las cuentas de las 8 listas y se compuso la lista de 63 cuentas semilla para la aplicación del modelo.

GLOSARIO

Modelo. Arquetipo, prototipo o punto de referencia para imitarlo o reproducirlo.

Ciencia de Datos. Disciplina de hacer útiles a los datos a partir de uso combinado de técnicas de estadísticas, análisis de datos y aprendizaje automático con el fin de extraer valor de los datos. Es también conocido como Descubrimiento de Conocimiento o *Knowledge Discovery* en inglés. En inglés es denominado *Data Science*.

Cuadro de Mando Integral. Herramienta que permite conocer a golpe de vista la situación global de una realidad en un momento dado a partir de indicadores clave de rendimiento o KPI. En inglés es denominado *Balance Scorecard*.

Indicador clave o KPI. Valor medible de desempeño, rendimiento o comportamiento de un proceso. Son usados generalmente para medir la eficacia con que una empresa logra sus objetivos.

Monitoreo. Observación del curso de uno o varios parámetros para detectar tendencias o posibles anomalías.

Sentimiento. Estado afectivo o disposición emocional hacia una cosa, un hecho o una persona.

Opinión. Juicio o valoración que se forma una persona respecto de algo o de alguien.

Análisis de Sentimiento o Minería de Opinión. Métodos de lingüística computacional orientadas a identificar y extraer información subjetiva de contenidos registrados en escenarios digitales.

Polaridad. En Análisis de Sentimiento, es la propiedad positiva, negativa o neutra de una oración.

Grafo. Diagrama que representa mediante puntos y líneas las relaciones entre pares de elementos y que se usa para resolver problemas lógicos, topológicos y de cálculo combinatorio.

Redes Sociales. Estructuras sociales integradas por personas, organizaciones o entidades que se encuentran conectadas entre sí por relaciones de distintos tipos.

Twitter. Servicio de microblogging con sede en California - EE.UU. Se llama así también a la red social de microblogging.

Tweet. Publicación o estado registrado en la red social Twitter hecho por un usuario. El estado puede contener una opinión del usuario.

Hashtag. Una palabra o frase precedida por un signo numeral, utilizado en redes sociales para identificar mensajes sobre un tema específico.

Cuenta de Twitter. Conjunto de datos empleados para registrar y autenticar a un usuario de Twitter.

Usuario de Twitter. Persona que hace uso de la plataforma de microblogging Twitter. Un usuario de Twitter puede tener una o más cuentas de Twitter.

Aprendizaje Automático. Aplicación de Inteligencia Artificial que permite a los sistemas computacionales la capacidad de aprender y mejorar automáticamente a partir de la experiencia sin ser programados explícitamente para ello. En inglés es denominado *Machine Learning*.

Inteligencia Artificial. Teoría para el desarrollo de sistemas computacionales capaces de realizar tareas que simulen la inteligencia humana.

Algoritmo. Conjunto ordenado y finito de operaciones que permite hallar la solución de un problema.

Script. En informática, es un conjunto de instrucciones o comandos interpretados guardados en un archivo.

Comando. Es una instrucción dada por un usuario que le dice a una computadora que haga algo.

Metadato. Dato que describe otro dato.

BIBLIOGRAFÍA

Kotu, V., y Deshpande, B. 2019. *Data Science, Concepts and Practice*. (Ciencia de Datos, Conceptos y Práctica). Ed. Elsevier. EE.UU.

Peng, R., y Matsui, E. 2016. *The Art of Data Science, A Guide for Anyone Who Works with Data*. (El Arte de la Ciencia de Datos, Una Guía para Quien Trabaje con Datos). Ed. Leanpub. Canada.

Robinson, I., Webber, J. y Eifrem, E. 2015. *Graph Databases, New Opportunities for Connected Data*. (Bases de Datos de Grafos, Nuevas Oportunidades para Datos Conectados). Ed. O'Reilly. EE.UU.

Missaoui, R. Y Idrissa, S. 2014. *Social Network Analysis – Community Detection and Evolution*. (Análisis de Redes Sociales – Detección y Evolución de Comunidades). Ed. Springer. Canada.

Bird, S., Klein, E. y Loper, E. 2009. *Natural Language Processing with Python*. (Procesamiento del Lenguaje Natural con Python). Ed. O'Reilly. EE.UU.

Pozzi, F., Fersini, E., Messina, E. y Liu, B. 2017. *Sentiment Analysis in Social Networks*. (Análisis de Sentimiento en Redes Sociales). Ed. Elsevier. EE.UU.

Sarkar, D. 2016. *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insight from Your Data*. (Análisis de Texto con Python: Un Enfoque Práctico del Mundo Real para Obtener Información Procesable de sus Datos). Ed. Apress. India.

Evans, J. 2017. *Business Analytics*. (Análisis de Negocios). Ed. Pearson. EE.UU.

Blanco, C. 2011. Encuesta y Estadística, Métodos de Investigación Cuantitativa en Ciencias Sociales y Comunicación. Ed. Brujas. Argentina.

Marradi, A., Archenti, N., Piovani, J.I. 2007. Metodología de las Ciencias Sociales. Ed. Emecé. Argentina.

Miranda, C., Guzmán, J. y Salcedo, D. 2016. Minería de Opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles. En revista Nº 56 Procesamiento de Lenguaje Natural, marzo de 2016, pp 25-32.

Sampieri, R., Fernández, C., Baptista, P. 2014. Metodología de la Investigación. Ed. McGraw Hill. México.

AGETIC, Agencia de Gobierno Electrónico y Tecnologías de Información y Comunicación. 2018. Estado de las Tecnologías de Información y Comunicación en el Estado Plurinacional de Bolivia. Obtenido de: *https://agetic.gob.bo/pdf/estadotic/AGETIC-Estado-TIC.pdf* Visitado en: 15/04/2019.

AGETIC, Agencia de Gobierno Electrónico y Tecnologías de Información y Comunicación. 2017. Encuesta Nacional de Opinión sobre Tecnologías de Información y Comunicación (TIC). Obtenido de: *https://agetic.gob.bo/pdf/dia_internet_encuesta.pdf* Visitado en: 15/04/2019.

Peñaranda, U. y Condori, M. 2018. Estudio: El oficialismo está perdiendo la batalla de las redes sociales. Obtenido de: *http://www.brujuladigital.net/politica/estudio-el-oficialismo-esta-perdiendo-la-batalla-de-las-redes-sociales* Visitado en: 15/04/2019.

Socialbakers.com. 2019. *Twitter statistics - Society in Bolivia*. (Estadísticas de Twitter – Sociedad en Bolivia). Obtenido de: *https://www.socialbakers.com/statistics/twitter/profiles/bolivia/society/* Visitado en: 15/04/2019.

Statista.com. 2019. *Number of Twitter users in selected countries in Latin America from 2014 to 2020*. (Número de usuarios de Twitter en países seleccionados en América Latina desde 2014 hasta 2020). Obtenido de: *https://www.statista.com/statistics/303931/twitter-users-latin-american-countries/* Visitado en: 15/04/2019.

Willige, A. 2017. *Is Twitter better at predicting elections than opinion polls?* (¿Es Twitter mejor que las encuestas de opinión para predecir las elecciones?). Obtenido de: *https://www.weforum.org/agenda/2017/02/twitter-opinion-polls-election-prediction/* Visitado en: 15/04/2019.

Porcaro, G. Y Müller, H., 2016. *Tweeting Brexit: Narrative building and sentiment analysis*. (Tuiteando acerca del Brexit: Construcción narrative y análisis de sentimiento). Obtenido de: *http://bruegel.org/2016/11/tweeting-brexit-narrative-building-and-sentiment-analysis/* Visitado en: 15/04/2019.

Herranz, A. 2015. ¿Predice el big data sobre redes sociales mejor que las encuestas quiénes ganan las elecciones? Obtenido de: *https://www.xataka.com/aplicaciones/predice-el-big-data-sobre-redes-sociales-mejor-que-las-encuestas-quienes-ganan-las-elecciones* Visitado en: 15/04/2019.

Kremers, B. 2012. *Predicting elections with Twitter and social media. Is it possible?* (Predicción de elecciones con Twitter y redes sociales. ¿Es posible?). Obtenido de: *https://www.buzztalkmonitor.com/blog/predicting-elections-with-twitter-and-social-media-is-it-possible/* Visitado en: 15/04/2019.