Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

ANÁLISIS DE INFORMACIÓN PÚBLICA DE CARÁCTER POLÍTICO PROVENIENTE DE REDES SOCIALES UTILIZANDO CIENCIA DE DATOS Y TÉCNICAS ESTADÍSTICAS DE REGRESIÓN

Javier Estuardo Navarro Delgado

Asesorado por el Ing. Erick Carlos Roberto Navarro Delgado

Guatemala, enero de 2020

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

ANÁLISIS DE INFORMACIÓN PÚBLICA DE CARÁCTER POLÍTICO PROVENIENTE DE REDES SOCIALES UTILIZANDO CIENCIA DE DATOS Y TÉCNICAS ESTADÍSTICAS DE REGRESIÓN

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

JAVIER ESTUARDO NAVARRO DELGADO
ASESORADO POR EL ING. ERICK CARLOS ROBERTO NAVARRO DELGADO

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, ENERO DE 2020

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Christian Moisés de la Cruz Leal
VOCAL V	Br. Kevin Armando Cruz Lorente
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANA	Inga. Aurelia Anabela Cordova Estrada
EXAMINADOR	Ing. Marlon Francisco Orellana López
EXAMINADOR	Ing. Sergio Arnaldo Méndez Aguilar
EXAMINADOR	Ing. Luis Fernando Espino Barrios
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

ANÁLISIS DE INFORMACIÓN PÚBLICA DE CARÁCTER POLÍTICO PROVENIENTE DE REDES SOCIALES UTILIZANDO CIENCIA DE DATOS Y TÉCNICAS ESTADÍSTICAS DE REGRESIÓN

Tema que me fuera asignado por la Dirección de la Escuela de Ingeniería en Ciencias y Sistemas, con fecha abril 2019

Javier Estuardo Navarro Delgado

Guatemala, 30 de diciembre de 2019

Ingeniero
Carlos Azurdia
Escuela de Ciencias y Sistemas
Facultad de Ingeniería
Universidad de San Carlos de Guatemala

Respetable Ingeniero Azurdia:

Por este medio hago de su conocimiento, que como asesor del trabajo de graduación del estudiante de la carrera de Ingeniería en Ciencias y Sistemas, **Javier Estuardo Navarro Delgado**, quien se identifica con el código único de identificación **2986 47664 0101** y con el registro académico **201513630**, hago constar que ha finalizado todos los capítulos del trabajo de investigación titulado: **"ANÁLISIS DE INFORMACIÓN PÚBLICA DE CARÁCTER POLÍTICO PROVENIENTE DE REDES SOCIALES UTILIZANDO CIENCIA DE DATOS Y TÉCNICAS ESTADÍSTICAS DE REGRESIÓN"**, el cual he tenido la oportunidad de revisar y doy mi aprobación al mismo.

Agradeciendo su atención a la presente, me es grato suscribirme.

F: _____

Ing. Erick Carlos Roberto Navarro Delgado
Catedrático, Organización de lenguajes y compiladores 2
Universidad de San Carlos de Guatemala, Facultad de Ingeniería

Erick Carlos Roberto Navarro Delgado
Ingeniero en Ciencias y Sistemas
No. Colegiado: 16465

Universidad San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 24 de enero de 2020

Ingeniero
Carlos Gustavo Alonzo
Director de la Escuela de Ingeniería
En Ciencias y Sistemas

Respetable Ingeniero Alonzo:

Por este medio hago de su conocimiento que he revisado el trabajo de graduación del estudiante **JAVIER ESTUARDO NAVARRO DELGADO** con carné **201513630 y CUI 2986 47664 0101** titulado **"ANÁLISIS DE INFORMACIÓN PÚBLICA DE CARÁCTER POLÍTICO PROVENIENTE DE REDES SOCIALES UTILIZANDO CIENCIA DE DATOS Y TÉCNICAS ESTADÍSTICAS DE REGRESIÓN"** y a mi criterio el mismo cumple con los objetivos propuestos para su desarrollo, según el protocolo aprobado.

Al agradecer su atención a la presente, aprovecho la oportunidad para suscribirme,

Atentamente,

Ing. Carlos Alfredo Azurdia
Coordinador de Privados
y Revisión de Trabajos de Graduación

ESCUELA DE INGENIERÍA EN CIENCIAS Y SISTEMAS

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÌA
ESCUELA DE INGENIERÍA EN
CIENCIAS Y SISTEMAS
TEL: 24188000 Ext. 1534

El Director de la Escuela de Ingeniería en Ciencias y Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor con el visto bueno del revisor y del Licenciado en Letras, del trabajo de graduación, ***"ANÁLISIS DE INFORMACIÓN PÚBLICA DE CARÁCTER POLÍTICO PROVENIENTE DE REDES SOCIALES UTILIZANDO CIENCIA DE DATOS Y TÉCNICAS ESTADÍSTICAS DE REGRESIÓN"*** *realizado por el estudiante, JAVIER ESTUARDO NAVARRO DELGADO, aprueba el presente trabajo y solicita la autorización del mismo.*

"ID Y ENSEÑAD A TODOS"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DIRECCION DE INGENIERIA EN CIENCIAS Y SISTEMAS

MSc. Ing. Carlos Gustavo Alonzo
Director
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 30 de enero de 2019

Universidad de San Carlos
de Guatemala

Facultad de Ingeniería
Decanato

DTG. 032.2020

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería en Ciencias y Sistemas, al Trabajo de Graduación titulado: **ANÁLISIS DE INFORMACIÓN PÚBLICA DE CARÁCTER POLÍTICO PROVENIENTE DE REDES SOCIALES UTILIZANDO CIENCIA DE DATOS Y TÉCNICAS ESTADÍSTICAS DE REGRESIÓN,** presentado por el estudiante universitario: **Javier Estuardo Navarro Delgado,** y después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

Inga. Anabela Cordova Estrada
Decana

Guatemala, enero de 2020

/gdech

ACTO QUE DEDICO A:

Dios	Por tantas bendiciones recibidas, por escuchar mis plegarias aun sabiendo que lo buscaba más en momentos adversos, por su infinito amor y misericordia.
Mi madre	Sandra Delgado. Por su amor, esfuerzos y sacrificios. Por esas llamadas cuando ya era tarde, por arrullarme en sus brazos, por esos masajes de espalda y por sus cuidados.
Mis hermanos	Erick y Joaquín Navarro. Por ser además de mis hermanos, mis mejores amigos, por las discusiones sin sentido, las charlas interesantes, por esas muestras de cariño esporádicas pero importantes y por su amor.
Mis abuelos	Carmen García y Carlos Delgado. Por su compañía, comprensión, cariño y apoyo. Por estar siempre conmigo y siempre estar dispuestos a ofrecerme esa apreciada taza de café.
Mi tío	Guillermino Mejía. Por enseñarme tantas cosas importantes y apoyarme en todo momento.

Mi tío Antonio Delgado. Por su apoyo a lo largo de mi carrera, sus valiosos consejos y por ayudarme a salir adelante.

Mi novia Heather Salamanca, por su apoyo incondicional, sus palabras de ánimo, por despertarse temprano preocupada por mis asuntos a realizar, por todas esas risas, por cada viaje, cada aventura, por cada abrazo, por brindarme paz y felicidad, pero sobre todo por su amor.

Mi ahijado Miguel Hernández. Por ser ese ejemplo de vida y enseñarme tanto, por aconsejarme y motivarme a ser un profesional.

Mi familia Por compartir conmigo los buenos y los malos momentos, por alegrarse de mis triunfos y apoyarme incondicionalmente.

Mis amigos Por hacer de mi vida, una experiencia amena, llena de risas, momentos memorables, por brindarme su apoyo desinteresado, su interés y cariño. En especial agradezco a: Adriana Oliva, Axel González, Brayan Flores, Carlos Torres, Celeste Duarte, Cesar Morales, Erick Mendoza, Fredy Ramírez, Jairo García, José Bautista, Kevin Cruz, Lucia Rodríguez, Luis Azurdia, Pablo Hernández, Rebeca Valladares.

AGRADECIMIENTOS A:

Guatemala	La tierra que me vio crecer, espero que mi profesión y mis conocimientos sean útiles para tu crecimiento.
Universidad de San Carlos de Guatemala	Por ser mi casa de estudios, la universidad del estado que me abrió sus puertas para ser uno más de sus alumnos.
Facultad de Ingeniería	Por haberme brindado tantos conocimientos y experiencias de vida.
Departamento de Matemática	Por haberme brindado la oportunidad de crecer profesional, laboral y personalmente, agradezco en especial a: Ing. Arturo Samayoa, Lic. Carlos Morales, Ing. Francisco García, Inga. Helen Ramírez, Ing. José Saquimux e Ing. Renaldo Girón por los conocimientos compartidos y sus valiosas lecciones.
Mi asesor	Ing. Erick Carlos Roberto Navarro Delgado, gracias por compartir conmigo sus conocimientos y por su apoyo en la elaboración de este trabajo.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
%	Porcentaje

GLOSARIO

Almacén de datos Un almacén de datos es una colección de datos orientada a un determinado ámbito, integrado, no volátil y variable en el tiempo, que sirve como soporte en la toma de decisiones.

API API es una sigla que procede de la lengua inglesa y que alude a la expresión Application Programming Interface (cuya traducción es Interfaz de Programación de Aplicaciones). El concepto hace referencia a los procesos, las funciones y los métodos que brinda una determinada biblioteca de programación a modo de capa de abstracción para que sea empleada por otro programa informático.

Internet Es una red de acceso público que interconecta muchas redes de computadoras a nivel mundial y transmite información utilizando IP, el protocolo de internet.

Python Lenguaje de programación multiparadigma, que soporta programación orientada a objetos, imperativa y funcional, es un lenguaje interpretado, dinámico y multiplataforma.

Red social	Según la RAE es un servicio de la sociedad de la información que ofrece a los usuarios una plataforma de comunicación a través de internet para que estos generen un perfil con sus datos personales, facilitando la creación de comunidades con base a criterios comunes y permitiendo la comunicación de sus usuarios.
Software	Según la RAE, el software es un conjunto de programas, instrucciones y reglas informáticas que permiten ejecutar distintas tareas en una computadora.
URL	Uniform Resource Locator. Es un localizador de recursos uniforme, el recurso al que dicho localizador apunta puede variar a lo largo del tiempo.

RESUMEN

Con el auge de las redes sociales, cada vez más personas utilizan medios digitales para emitir su opinión sobre temas políticos y sociales, por ese motivo las mismas producen cantidades voluminosas de datos de carácter público que pueden ser procesados y analizados, para estudiar el comportamiento de los usuarios en las redes sociales.

Aplicando ciencia de datos y análisis sentimental a las opiniones públicas y comentarios realizados por guatemaltecos acerca de los candidatos que participan en las elecciones presidenciales, es posible construir un modelo predictivo de regresión que permita predecir qué partido tendrá más posibilidades de ser electo en base a la reputación del mismo en redes sociales.

OBJETIVOS

General

Analizar información pública de carácter político proveniente de redes sociales, utilizando ciencia de datos y técnicas estadísticas de regresión.

Específicos

1. Determinar el criterio más adecuado para la selección de una muestra representativa de la opinión publica referente a la política en Guatemala, proveniente de la red social Twitter.
2. Seleccionar un método de análisis sentimental apropiado para realizar un análisis de la muestra recopilada, conformada por comentarios y opiniones de los usuarios de la red social Twitter.
3. Fijar un perfil de popularidad en función de los resultados obtenidos en el análisis realizado para cada candidato presidencial, en un intervalo de tiempo previo a las elecciones generales en la red social Twitter.
4. Construir un modelo predictivo de regresión, utilizando los perfiles de cada partido político en los últimos tres periodos electorales y posterior a esto, determinar la calidad de ajuste del modelo construido.
5. Establecer la predicción de popularidad para cada candidato en el siguiente periodo electoral.

INTRODUCCIÓN

Desde los inicios de los sistemas de información, una de las consecuencias más notables de su implementación es la generación de datos y su almacenamiento, su procesamiento es de vital importancia, ya que permite obtener como resultado información que puede ser utilizada para realizar mejoras a procesos o productos, evaluar un mercado en específico e inducir los cambios estratégicos necesarios para lograr aumentar la competitividad de una empresa o incluso, podría permitir la construcción de un modelo predictivo relacionado a un fenómeno específico, que podría ser climático, social, económico, entre otros.

Los modelos predictivos de regresión representan la relación entre dos o más variables y permiten realizar predicciones de los valores que tomará la variable dependiente, en función de una o más variables independientes. El objeto del modelo es evaluar la probabilidad de que un elemento similar al estudiado de una muestra diferente exhiba, un comportamiento cercano al esperado.

Las redes sociales pueden brindar información relacionada a la opinión pública y permite definir un perfil de reputación de grupos, paginas o personas individuales, un claro ejemplo de la influencia que poseen las redes sociales se observó en Guatemala, el 25 de abril de 2015, cuando el resultado de la creación de un evento en Facebook, fue una masiva protesta en la que participaron alrededor de 15 mil personas que estaban reunidas en la Plaza de la Constitución, solicitando la renuncia del presidente Otto Pérez, publicaciones

negativas respecto a la labor realizada por el mandatario reflejaban una mala reputación lo que posteriormente se resumió en la renuncia a su cargo.

Eventos de esa índole pueden ser predichos mediante modelos de predicción estadísticos. En el presente trabajo se mostrarán los elementos necesarios para la construcción de un modelo predictivo de regresión utilizando los datos obtenidos de la red social Twitter y que sean públicos, para realizar una predicción del comportamiento de la población guatemalteca que posea un usuario en dicha red social, y realice comentarios u opiniones que permitan la creación de un perfil para cada partido político en base a la reputación que sus candidatos posean.

Esto se realizará llevando a cabo un análisis sentimental de todas las publicaciones relacionadas a dichos candidatos, y se podrá predecir qué partido tendrá mayores posibilidades de ser electo tomando como criterio principal, el perfil de reputación que posea cada candidato.

1. INFORMACIÓN PÚBLICA DE CARÁCTER POLITICO PROVENIENTE DE LA RED SOCIAL TWITTER

En este capítulo se introducirá al concepto de red social y cómo es posible acceder a la información que los usuarios comparten en estas plataformas. Primero ser procederá a presentar el concepto de una red social, seguido de una breve explicación al porqué es posible acceder a cierta información de los usuarios y por último, una breve introducción una red social en específico Twitter, esto debido a que en el caso de estudio que se presentará más adelante, se utilizará dicha red social, en tal introducción, se dará énfasis a la política de privacidad que maneja esta.

1.1. Red social

A continuación, se procederá a definir este concepto, sus características y por último, su funcionamiento básico.

1.1.1. Definición

Una red social se define de una forma genérica como un conjunto de actores y de enlaces que los relacionan. “Los actores, tales como personas, organizaciones o cualquier otra entidad social, se conectan por relaciones de amistad, parentesco, intercambio financiero o de información; es decir, por motivos sociales, cotidianos o profesionales.”[1]

[1] CARBALLAR, José Antonio. *Social Media. Marketing personal y profesional*. p.71.

Es importante resaltar de la definición, que debido a los enlaces entre los actores se genera información, esta información puede o no ser compartida con los demás actores, estos criterios se definirán más adelante y están en función de la plataforma y sus políticas internas.

1.1.2. Características

En general las redes sociales se caracterizan por ser grupos de personas que:

- Comparten un interés personal o profesional, una afición o bien cualquier otro tema concreto. Este elemento incentiva la relación.
- Se comunican a través de Internet, generalmente utilizando una plataforma online e independiente del dispositivo, este servicio es brindado por la agrupación a la que pertenecen, por ejemplo, Twitter.
- Se mantienen relaciones estables y normadas, estas por supuesto son diferentes a las normas de comportamiento habituales existentes para relaciones presenciales.

1.1.3. Funcionamiento

"Las redes sociales online permiten a sus usuarios construir un perfil público o semipúblico y articular una lista de contactos con los que relacionarse. Más aún, los usuarios pueden intercambiar mensajes de texto o contenido multimedia que ponen a disposición de los demás usuarios, de una forma más o menos restringida (en función de los filtros o privacidad definida)."[2]

[2] CARBALLAR, José Antonio. *Social Media. Marketing personal y profesional*. p.71.

1.2. Información pública en redes sociales

Así como el estado guatemalteco posee una Ley de Acceso a la Información Pública, que tiene por objeto garantizar a cualquier persona solicitar el acceso a información pública, una red social posee normas y acuerdos de privacidad que permiten el acceso a información pública de sus usuarios, estos deciden qué información será de carácter público y cual no, esta situación genera una fuente de datos lista para ser analizada, pero debe ser regulado por una ley vigente como la anteriormente mencionada dependiendo la región en la que se desee acceder a la información.

A pesar de que no existe una normativa que haya sido adoptada de forma internacional, existen modelos de leyes provistos por organizaciones gubernamentales como la Organización de Estados Americanos (OEA), que regulan el acceso a información pública en función de la jurisdicción y organismos correspondientes del Gobierno, estos modelos afectan a toda organización que represente legalmente una plataforma de red social, esta debe construir una política de privacidad que cumpla y sea regida por estos modelos.

"El concepto de privacidad se encuentra vinculado al de la intimidad que es un derecho garantizado en los principales organismos interamericanos y universales de derechos humanos que adquiere en Sillicon Valley una nueva perspectiva, con retos que exigen un balance entre el derecho a la intimidad del individuo y los avances en el mundo de las Tecnologías de Información y Comunicación"[3].

[3] Organización de los Estados Americanos. *Protección de datos personales*. http://www.oas.org/es/sla/ddi/proteccion_datos_personales.asp. Consulta: julio de 2019.

1.3. Red social Twitter

"Twitter es una red social creada por Jack Dorsey, Evan Williams y Biz Stone en 2006. Una de las características más relevantes de Twitter es que se trata de una red abierta, esto quiere decir que no es necesario estar registrado para acceder a la información de sus usuarios."[4] Esta situación implica que Twitter sea difusor que va más allá del grupo cercano de relaciones personales.

1.3.1. Características

De acuerdo con Carballar, en su obra *Social Media*, al considerar que las actividades fundamentales relacionadas con la información son dos: la investigación, para recabar información y la comunicación, para ofrecer información. Twitter es una buena herramienta para ambas actividades. Además de esta característica es importante mencionar las siguientes:

- Permite comunicar sentimientos, ideas, información de una forma muy rápida y casi desde cualquier sitio.
- Permite tomar una fotografía, añadir un comentario y crear una tendencia pública.
- Es ágil, está al alcance de todos, por ejemplo, una nota de prensa, se publica con suerte, el día siguiente, pero con Twitter, se realiza en tiempo real.
- Efectivo para el marketing relacional, este se refiere a la gestión adecuada de las relaciones con los clientes, Twitter es útil para crear el sentimiento de comunidad de una marca.
- Permite obtener información sobre un mercado en específico, sobre los competidores.

[4] Twitter. *Política de privacidad de Twitter*. https://twitter.com/es/privacy/previous/version_12. Consulta: julio de 2019.

- Permite mejorar la relación con los clientes, ofreciendo un medio alternativo flexible, rápido y económico.

1.3.2. Audiencia

"En Twitter, toda persona que tenga acceso a internet puede ser audiencia de esta, pues por la naturaleza de la plataforma no se requiere registro para tener acceso a la información pública que cumpla con la política de privacidad de dicha red social."[5]

1.3.3. Política de privacidad

Tal como se menciona en la sección anterior, cualquier persona tiene acceso a la información de Twitter, pero únicamente un usuario registrado podrá enviar un Tweet, es así como le llama dicha plataforma a un mensaje de 140 caracteres o menos, que es público de forma predeterminada y que puede incluir fotos, videos y enlaces a otros sitios web. "Lo que se publica en Twitter puede verse en todo el mundo de manera instantánea."[6]

1.3.3.1. Recopilación y uso de información

Básicamente la plataforma recopila la información enlistada a continuación para proporcionar, comprender y mejorar los servicios que provee:

- Información básica de la cuenta y de contacto, como nombre de usuario, dirección de correo electrónico o número de teléfono.
- Información adicional, como agenda de contactos.

[5] Twitter. *Política de privacidad de Twitter*. https://twitter.com/es/privacy/previous/version_12. Consulta: julio de 2019.
[6] Ibid.

- Comentarios, usuarios seguidos, listas, perfil y otra información pública, Twitter está diseñado para compartir la información con el mundo, en esta información se basará el caso de estudio.
- Mensajes directos y comunicaciones no públicas.
- Ubicación, enlaces, cookies, datos de widgets, etc.

1.3.3.2. Cesión y revelación de información

Twitter no revela datos personales, excepto en las siguientes circunstancias:

- Consentimiento o indicación del usuario: siguiendo las instrucciones del usuario, como cuando el usuario autoriza a un cliente o una aplicación web de un tercero acceder a su cuenta, o cuando indica que desea compartir sus comentarios con un negocio.
- Proveedores de servicios como, por ejemplo: Google Analytics, pero este cumplirá con la misma política de privacidad.
- Transacciones comerciales, al efectuar algún pago, Twitter revelará información de pago.
- Información pública o no personal, como el usuario.
- Otras abarcan, daños, normas de prioridad, como órdenes judiciales.

1.3.3.3. Operaciones en todo el mundo

Las operaciones de Twitter entre la Unión Europea y Estados Unidos se ven reguladas por los principios de blindaje de la privacidad entre los previamente mencionados, con esto los participantes están sujetos a las fuerzas de investigación y cumplimiento de la Comisión de Comercio Federal de los EE. UU., y a las de los otros cuerpos autorizados.

2. CIENCIA DE DATOS

"La ciencia de datos es el estudio de dónde proviene la información, qué representa y cómo se puede convertir en un recurso valioso para la creación de estrategias empresariales y de Tecnologías de la Información"[7]. La extracción de grandes cantidades de datos estructurados y no estructurados para identificar patrones puede ayudar a una organización a controlar los costos, aumentar la eficiencia, reconocer nuevas oportunidades de mercado y aumentar la ventaja competitiva de la organización.

El campo de la ciencia de datos emplea matemáticas, estadística y disciplinas informáticas, e incorpora técnicas como el aprendizaje automático, el análisis sentimental de conglomerados, la extracción de datos y la visualización de estos.

2.1. Importancia de la ciencia de datos

En la actualidad las organizaciones tratan con enormes volúmenes de datos, estos llegan a ser *zetta bytes* o incluso *yotta bytes* que pueden ser estructurados y no estructurados, esto se da todos los días. Cuando estos datos son analizados mediante herramientas de la tecnología de la información, se convierten en una ventaja competitiva, porque si estos datos, tratan sobre la competencia, brinda la oportunidad de implementar planes estratégicos.

[7] ROUSE, Margaret. *Ciencia de datos*. https://searchdatacenter.techtarget.com/es/definicion/Ciencia-de-datos. Consulta: julio de 2019.

2.2. Big Data, fuente de datos para la ciencia de datos

Big Data, es un término que hace referencia a grandes volúmenes de datos. Muchas son las definiciones que entidades y organizaciones han dado para el término Big Data, pero todas ellas se pueden resumir a un conjunto de datos cuyo tamaño supera considerablemente la capacidad de captura, almacenamiento, gestión y análisis del software convencional de bases de datos.

Sin embargo, el concepto no hace referencia simplemente al tamaño de la información, sino también a la variedad del contenido y a la velocidad con la que los datos se generan, almacenan y analizan. De acuerdo con la empresa Gartner, en el glosario de términos publicado en su sitio web oficial, estas dimensiones son las “3V” que describe el concepto de Big Data, es decir volumen, velocidad y variedad de los datos:

- Volumen. Un gran volumen de datos que se generan diariamente en las empresas y organizaciones de todo el mundo.
- Velocidad. Se trata de los flujos de datos, la creación de registros estructurados y la disponibilidad para el acceso y la entrega de estos. Es decir, que tan rápido se están produciendo los datos, así como la rapidez en la que se trata de satisfacer la demanda de éstos.
- Variedad. Big data ha de tener la capacidad de combinar una gran variedad de información digital en los diferentes formatos en los que se puedan presentar.

Hasta 2003 la humanidad había generado cinco exabytes de información a lo largo de toda su historia. Lo dijo Eric Schmidt, CEO de Google, durante una conferencia en 2010. En 2007 se generaron 281 exabytes, según las investigadoras Hardy y Williams, y apenas cuatro años más tarde se alcanzaron

los 1800 exabytes. Estos datos provienen de distintas fuentes, a continuación, daremos algunos ejemplos de estos:

- Los sensores que se utilizan en el complejo de compras para recopilar información del comprador.
- Los posts que las personas hacen en las plataformas de redes sociales.
- Las fotos y videos digitales que son capturadas por los usuarios de celulares o dispositivos electrónicos con cámara.
- La transacción de compra que se realiza a través del comercio electrónico.

El término Big Data es utilizado para referirse a estos enormes volúmenes de datos.

2.3. Big Data para el análisis predictivo

Tanto Big Data como el análisis predictivo se emplean fundamentalmente en marketing y ventas. Los usos más extendidos son:

- Afianzamiento de las campañas de marketing
- Lanzamiento de nuevos productos
- Aumento de la cartera de clientes
- Expansión de mercados

3. ANÁLISIS SENTIMENTAL

El análisis sentimental es una de las herramientas con las que cuenta la ciencia de datos para generar información. A continuación, se definirán los conceptos básicos que este análisis requiere.

3.1. Procesamiento de Lenguajes Naturales

"El Procesamiento de Lenguajes Naturales(PLN) es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano."[8]

A continuación, se definen algunos de los componentes del PLN, no todos los componentes que se describen se aplican siempre, porque dependerá del objetivo de la aplicación y son los siguientes:

- Análisis léxico: se construye en base a una entrada determinada una lista de componentes léxicos y se reporta todo aquel componente, que no cumpla con las reglas definidas en el lenguaje.
- Análisis sintáctico: se verifica que la lista de componentes léxicos cumplan con la estructura del modelo gramatical empleado.
- Análisis semántico: proporciona una interpretación a la entrada, tras haber eliminado los errores causados por ambigüedades.
- Análisis pragmático: incorpora el análisis del contexto de uso a la interpretación final, como, por ejemplo, se incluye el lenguaje figurado.

[8] MORENO, Antonio. *Procesamiento del lenguaje natural*. https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/. Consulta: junio de 2019.

Según el sitio web Planeta Chatbot, las siguientes técnicas son algunas de las más comunes del PLN:

- Normalización: coloca todas las palabras en igualdad, por ejemplo; convirtiendo todas las palabras de mayúsculas o minúsculas.
- La eliminación de palabras: la eliminación de palabras ocurre con frecuencia con artículos como el, la, y, a, etc., que no contribuyen en gran medida a entender el texto y pueden eliminarse.
- Stemming: elimina los afijos de las palabras para obtener la raíz de la palabra, por ejemplo; si las entradas son bibliotecario" y "bibliotecas" genera como salida "biblioteca".
- Lematización: similar a la anterior, pero puede obtener la forma canónica de una palabra en función de su lema, por ejemplo, nadar es el lema de nadé o nadaste.
- Etiquetado de partes del discurso: asigna etiquetas para sustantivos, pronombres, verbos, adjetivos, adverbios, etc.
- N-gramas: son una secuencia continua de palabras adyacentes en una oración, necesarias para obtener el significado correctamente, por ejemplo: 'Machine learning' es un bi-grama.
- TF: llamada frecuencia de término y se calcula por el número de veces que aparece una palabra en un mensaje o una oración, para indicar la importancia de esa palabra.
- Reconocimiento de entidades nombradas: como, por ejemplo, lugares, fechas, personas, etc.

Es importante tener en cuenta que la precisión de lo anterior depende de la cadena de entrada utilizada para efectuar el entrenamiento del modelo. Por lo tanto, un modelo entrenado con comentarios y uno entrenado con mensajes de Twitter dará resultados diferentes.

> Dependiendo del tipo de problema que necesita resolverse usando NLP, se puede identificar un conjunto de datos de entrenamiento apropiado para capacitar a un modelo o se debe identificar un modelo pre-entrenado que utiliza un conjunto de datos similar. Además, un modelo entrenado con texto en inglés no puede usarse para procesar otro texto complejo de lenguaje de caracteres, el modelo es específico del idioma en el que se ha formado.[9]

3.2. Definición de análisis sentimental

El análisis de sentimientos es un campo dentro del Procesamiento de Lenguajes Naturales que construye sistemas que intentan identificar y extraer opiniones dentro del texto, para que esto sea posible es necesario contar con métricas o atributos objetivos, por ejemplo:

- Polaridad: si el hablante expresa una opinión positiva o negativa
- Asunto: de lo que se habla
- Titular de opinión: la persona o entidad que expresa la opinión

"Las empresas utilizan el análisis de sentimientos para analizar automáticamente las respuestas de las encuestas, las revisiones de productos, los comentarios en las redes sociales y similares para obtener información valiosa sobre sus marcas, productos y servicios."[10]

Con la ayuda de los sistemas de análisis de sentimientos, esta información no estructurada podría transformarse automáticamente en datos estructurados de opiniones públicas sobre productos, servicios, marcas, políticas o cualquier tema sobre el que las personas puedan expresar opiniones. Estos datos pueden ser muy útiles para aplicaciones comerciales como análisis de

[9] SOTO, Marvin. Episodio 1: *Procesamiento de lenguaje natural*. https://planetachatbot.com/1-procesamiento-de-lenguaje-natural-1443ff471ed0. Consulta: junio de 2019.

[10] MonkeyLearn. *Sentiment analysis: complete guide*. https://monkeylearn.com/sentiment-analysis/. Consulta: octubre de 2019.

mercadotecnia, relaciones públicas, revisiones de productos, puntaje de promotores netos, retroalimentación de productos y servicio al cliente.

3.3. Opiniones

Antes de entrar en más detalles, primero se va a dar una definición de opinión. El texto se puede clasificar en dos categorías principales: hechos y opiniones. Los hechos son expresiones objetivas de algo. Las opiniones suelen ser expresiones subjetivas que describen los sentimientos, las valoraciones y los sentimientos de las personas hacia un tema.

El análisis de sentimientos, al igual que muchos otros problemas de PNL, se puede modelar como un problema de clasificación en el que se deben resolver dos subproblemas:

- Clasificar una oración como subjetiva u objetiva, conocida como clasificación de subjetividad.
- Clasificar una oración como una opinión positiva, negativa o neutral, conocida como clasificación de polaridad.

En una opinión, la entidad a la que se refiere el texto puede ser un objeto, sus componentes, sus aspectos, sus atributos o sus características. También podría ser un producto, un servicio, un individuo, una organización, un evento o un tema.

Observar el siguiente ejemplo: "La duración de la batería de esta cámara es demasiado corta", se puede deducir que se está expresando una opinión negativa sobre una característica (duración de la batería), de una entidad (cámara).

3.3.1. Opiniones comparativas versus directas

Hay dos tipos de opiniones: directas y comparativas. Las opiniones directas dan una opinión sobre una entidad directamente, por ejemplo: "La calidad de imagen de la cámara A es pobre". Esta opinión directa expresa una opinión negativa sobre la cámara A.

En opiniones comparativas, la opinión se expresa comparando una entidad con otra, por ejemplo: "La calidad de imagen de la cámara A es mejor que la de la cámara B." Generalmente, las opiniones comparativas expresan similitudes o diferencias entre dos o más entidades, utilizando una forma comparativa o superlativa de un adjetivo o adverbio. En el ejemplo anterior, hay una opinión positiva sobre la cámara A y a la inversa, una opinión negativa sobre la cámara B.

3.3.2. Opiniones explícitas versus implícitas

Una opinión explícita sobre un tema es una opinión expresada explícitamente en una oración subjetiva. La siguiente oración expresa una opinión explícita positiva: "La calidad de voz de este teléfono es increíble". Una opinión implícita sobre un tema es una opinión implícita en una oración objetiva. La siguiente oración expresa una opinión negativa implícita: "El auricular se rompió en dos días". Dentro de las opiniones implícitas, se podrían incluir metáforas que pueden ser el tipo de opinión más difícil de analizar, porque incluyen una gran cantidad de información semántica.

3.4. Alcance

Según Monkey Learn, el análisis de sentimientos se puede aplicar en diferentes niveles de alcance:

- El análisis de sentimiento a nivel de documento obtiene el sentimiento de un documento o párrafo completo.
- El análisis de sentimiento a nivel de oración obtiene el sentimiento de una sola oración.
- El análisis de sentimiento a nivel de sub-oración obtiene el sentimiento de sub-expresiones dentro de una oración.

3.5. Tipos de análisis sentimental

Hay muchos tipos de análisis de sentimientos y las herramientas disponibles van desde sistemas que se centran en la polaridad (positivo, negativo, neutral), a sistemas que detectan sentimientos y emociones (enojados, felices, tristes, entre otros.) o identifican intenciones (le interesa o no le interesa). En la siguiente sección, se cubrirán las más importantes.

3.5.1. Análisis de sentimiento de grano fino

Si se quiere medir el nivel de polaridad de una opinión, es posible agregar además de las categorías positiva, neutral o negativa, las categorías de muy positiva y muy negativa. Puede realizarse una analogía con la calificación que se le da a los hoteles que va de una máxima de 5 estrellas (el más lujoso y con los mejores estándares de atención al cliente), a una mínima de 1 estrella (con los estándares más básicos).

Algunos sistemas también proporcionan diferentes grados de polaridad al identificar si el sentimiento es positivo o negativo y si este está relacionado con algún sentimiento en particular, como ira, tristeza o preocupaciones (es decir, sentimientos negativos) o felicidad, amor o entusiasmo (es decir, sentimientos positivos).

3.5.2. Detección de emociones

La detección de emociones tiene como objetivo detectar emociones como, felicidad, frustración, enojo, tristeza y demás. Muchos sistemas de detección de emociones recurren a los léxicos (es decir, listas de palabras y las emociones que transmiten) o algoritmos complejos de aprendizaje automático.

Una de las desventajas es que la forma en que las personas expresan sus emociones varía mucho. Algunas palabras que normalmente expresan enojo como carajo o muerte también podrían expresar felicidad.

3.5.3. Análisis de sentimiento basado en aspectos

Por lo general, al analizar la opinión de los usuarios en una entidad en específico, como, por ejemplo, un producto, puede ser de interés no solo conocer la polaridad positiva, neutral o negativa sobre dicho producto, sino también sobre qué aspectos o características particulares del producto hablan. De esto se trata el análisis del sentimiento basado en aspectos.

Un claro ejemplo de este tipo de análisis sentimental es la siguiente opinión: "La duración de la batería de esta cámara es demasiado corta", esta oración expresa una opinión negativa sobre la cámara, pero más precisamente, sobre la duración de la batería, que es un aspecto particular de la cámara.

3.5.4. Análisis de intenciones

Este tipo de análisis es diferente a los anteriores porque el objetivo es reconocer lo que las personas quieren lograr al redactar un texto en lugar de lo

que las personas expresan directamente con ese texto. A continuación, algunos ejemplos:

- “Su centro de atención telefónica es un desastre. He estado en espera durante 20 minutos".
- "Me gustaría saber cómo cambiar mi chip".
- "¿Puedes ayudarme a llenar este formulario?"

3.6. Retos y dificultades técnicas

El desarrollo en la teoría de la computación, en especial de los compiladores han permitido a una maquina ser capaz de analizar determinados lenguajes formales, una situación totalmente diferente es que una maquina analice una entrada en lenguaje natural, por ejemplo, un ser humano no tiene problemas para detectar las quejas en determinada entrada. Sin embargo, las máquinas pueden tener algunos problemas para identificarlas. A veces la acción deseada puede inferirse del texto, pero a veces, inferirla requiere algún conocimiento contextual.

3.7. Importancia del análisis sentimental

De acuerdo con Monkey Learn, alrededor del 80% de los datos mundiales no están estructurados y no están organizados de manera predefinida. La mayor parte de esto proviene de datos de texto, como correos electrónicos, boletos de soporte, chats, redes sociales, encuestas, artículos y documentos. Estos textos suelen acarrear una dificultad alta al momento de tener la intención de analizarlos, ya que requieren tiempo y recursos para poder comprender y clasificar.

3.7.1. Ventajas

Algunas de las ventajas del análisis del sentimiento incluyen lo siguiente:

- Escalabilidad: el análisis de sentimientos permite procesar datos a escala de manera eficiente y rentable.
- Análisis en tiempo real: se puede usar el análisis de sentimientos para identificar información crítica que permita el conocimiento de la situación en escenarios específicos, en tiempo real.
- Criterios consistentes: al utilizar un sistema de análisis de sentimientos centralizado, las empresas pueden aplicar los mismos criterios a todos sus datos. Esto ayuda a reducir errores y mejorar la consistencia de los datos.

3.7.2. Desventajas

Es importante resaltar algunas desventajas o riesgos: si la muestra seleccionada no representa de forma satisfactoria a la población los datos no revelarán información significativa, o si el contexto que se le da no es el adecuado, la consecuencia será la misma.

3.8. Librería TextBlob

Una herramienta disponible para el análisis de lenguaje natural y en especial para el análisis sentimental es TextBlob. “TextBlob es una biblioteca de Python para procesar texto. Proporciona una API simple para sumergirse en tareas comunes de procesamiento del lenguaje natural como el etiquetado de

parte del discurso, extracción de frases nominales, análisis de sentimientos, clasificación, etc.”[11]

[11] LORIA, Steven. *TextBlob: Simplified text processing*. https://textblob.readthedocs.io/en/dev. Consulta: octubre de 2019.

4. TECNICAS ESTATISTICAS DE REGRESIÓN

A continuación, se mostrarán las definiciones más relevantes relacionadas a técnicas estadísticas que se utilizarán. En un análisis de regresión simple existen dos tipos de variables:

- Variable explicativa: es la variable independiente del modelo, puede estar formada por un vector de una sola característica o puede ser un conjunto de n características, atributos o dimensiones.
- Variable de respuesta: es la variable dependiente del modelo, es decir, esta depende de la variable explicativa.

Al realizar un análisis de regresión se espera obtener una función de la variable explicativa en términos de la variable de respuesta, mientras sea lo más sencilla posible, esta debe ser capaz de describir adecuadamente la variación de dicha variable explicativa.

4.1. Regresión lineal

Un modelo de regresión lineal permite analizar la relación entre la variable de respuesta y un conjunto de variables explicativas. “Esta relación se expresa como una ecuación que predice la variable de respuesta como una función lineal de los parámetros. Estos parámetros se ajustan para que la medida de ajuste sea óptima. Gran parte del esfuerzo en la adaptación del modelo se centra en minimizar el error.”[12]

[12] ESPINO, Carlos. *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo*. p.18.

Ya que el modelo de regresión lineal depende directamente de los parámetros, es necesario definir una forma de encontrar dichos parámetros, para fines prácticos se utilizará un método específico, y se denomina mínimos cuadrados ordinarios, este método permitirá encontrar los parámetros del modelo minimizando la suma de distancias verticales entre las respuestas observadas en la muestra y las respuestas del modelo, valuando la función explicativa con los valores muestrales.

4.1.1. Tipos de regresión lineal

Existen múltiples clasificaciones para un modelo de regresión lineal, pero para fines de este trabajo de investigación se definirán los siguientes tipos elementales de una regresión lineal, tomando como base la necesidad de relacionar una o más variables independientes con una variable dependiente de las mismas.

4.1.1.1. Regresión lineal simple

Una regresión lineal simple se define como un modelo con una variable dependiente e independiente cuantitativas cuya relación es caracterizada por una línea recta, la función es del tipo $\hat{y}_i = w_0 + w_1 x_i + \epsilon$, en donde w_0 es el intercepto o corte con el eje de la variable dependiente y w_1 es la pendiente, que en este caso es conocida como coeficiente de regresión. Y por último está el error ϵ que representa los factores no controlados que se llamaran como una perturbación o error aleatorio.

4.1.1.2. Regresión lineal múltiple

Ya se describió la forma más básica que puede tomar un modelo de regresión, pero si se llegará tener más variables independientes n, es en tal caso que surge la necesidad de un modelo de regresión múltiple, por ser necesaria una función que dependa de una o más variables explicativas, por tanto se deberá definir la función predictiva de forma generalizada para un caso de regresión múltiple utilizando la siguiente expresión: $\hat{y}_i = w_0 + w_1 x_{1i} + \cdots + w_n x_{ni} + \epsilon$. Para que sea posible encontrar la función que permite optimizar la distancia entre los valores observados y los esperados para cuando se tienen n dimesiones. Esta tarea se resume a encontrar los coeficientes de regresión w.

4.1.2. Calidad de ajuste

Según Pereira González, Tras haber determinado el modelo idóneo para predecir el comportamiento de un fenómeno específico mediante una regresión lineal o múltiple se requiere definir la calidad de ajuste del modelo y la medición que se realizará mediante el mismo. Algunas de las métricas que permiten caracterizar la calidad de ajuste de un modelo de regresión, son las siguientes:

4.1.2.1. Suma total de cuadrados

Sum Squared Total (SST), por sus siglas en inglés, es una variación de los valores de una variable dependiente, en esencia, cuantifica la variación total en una muestra y está dada por: $SST = \sum(y_i - \bar{y})^2$. Donde y_i es el valor en una muestra y $\bar{y}$ es la media de una muestra.

4.1.2.2. Suma de cuadrados de regresión

Sum Squared Regression (SSR), por sus siglas en inglés, describe que tan bien un modelo de regresión representa los datos modelados, si esta métrica toma valores muy elevados, indicaría que el modelo no se ajusta bien a los datos, esta se puede determinar de la siguiente forma: $SSR = \sum(\hat{y}_i - \bar{y})^2$. Donde $\hat{y}_i$ es el valor estimado por el modelo de regresión y $\bar{y}$ es la media de una muestra.

4.1.2.3. Coeficiente de determinación

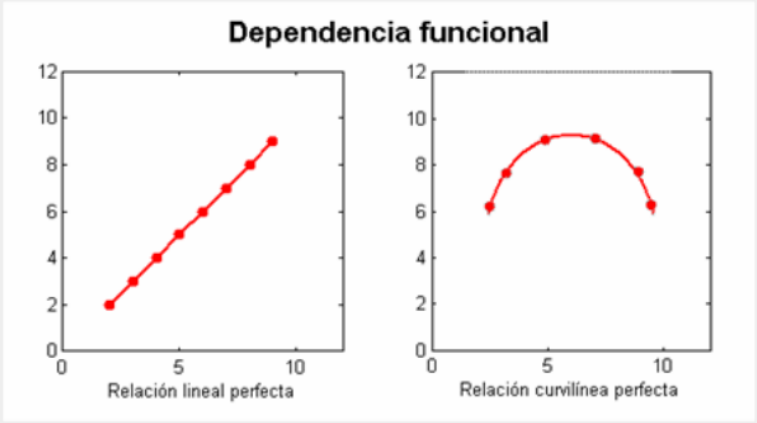
R^2, representa el coeficiente de determinación, que es la proporción de la variabilidad de la variable de respuesta que queda explicada por el modelo. Este coeficiente esta dado por: SSR/SST. Para el modelo de regresión lineal coincide con el cuadrado del coeficiente de correlación lineal $R^2 = r^2$. Cuanto más se acerque a uno, mayor fiabilidad tendrán las predicciones.

4.2. Dependencia lineal entre variables independientes

Un modelo de regresión lineal, sin importar su tipo, debe cumplir con la premisa básica de que todas sus variables explicativas sean linealmente independientes entre ellas. Una variable i es linealmente dependiente de una variable j si es posible definir i como una combinación lineal de j, cuando un conjunto de datos presenta este fenómeno se dice que existe colinealidad. La forma de detectar fácilmente si existe dependencia entre las variables es con un diagrama de dispersión, como se observa en la figura 1.

Se basa en la idea de representar cada par de variables conformado por $(x_{1i}, x_{1i}), \ldots, (x_{1i}, x_{ni})$, en diferentes planos cartesianos para así obtener tantos diagramas de dispersión como variables independientes existan para una única variable x_{1i} con respecto al resto de variables independientes. Con esta representación es posible inferir si existe una relación morfológica entre dichas representaciones.

Figura 1. **Diagramas de dispersión y la dependencia de variables**



Fuente: PEREIRA GONZÁLEZ, Augusto. *Análisis predictivo de datos mediante técnicas de regresión estadística*. p. 21.

5. CASO DE ESTUDIO

5.1. Descripción

Para la aplicación de los conceptos teóricos expuestos en los capítulos anteriores, será necesario entrar a un contexto político, y se ve influenciado por las elecciones presidenciales que se llevan a cabo en Guatemala cada cuatro años, porque las redes sociales desenvuelven un papel muy importante en la opinión pública, es posible que las mismas provean información importante sobre la popularidad o aceptación de cada uno de sus principales integrantes, y en consecuencia de cada agrupación política.

Se procederá a extraer información que pertenezca a la categoría pública de la red social Twitter, específicamente de la información que pertenezca a páginas de divulgación, de medios de comunicación, de debate político, periódicos o incluso de noticias. Para esto se utilizará la API que ofrece Twitter *for developers* y posteriormente para realizar el análisis sentimental se utilizará la API TextBlob, con los datos obtenidos se realizará un perfil para cada partido político, con el cual se obtendrá un *ranking* de estos.

Con los *rankings* de popularidad construidos para los últimos 3 periodos electorales se procederá a realizar un análisis de regresión lineal simple, para determinar un modelo predictivo del comportamiento de los usuarios de esta famosa red social, se utilizarán únicamente tres periodos electorales, debido a la antigüedad de la red social utilizada y el reciente auge que esta ha tenido.

5.2. Extracción de datos

En esta sección se describirá todo el proceso necesario para la extracción de datos, desde la selección de la muestra hasta el procedimiento de descarga mediante la utilización de la API que ofrece la red social Twitter.

5.2.1. Criterio para la selección de la muestra

Para seleccionar la muestra que representará a todos los usuarios de Twitter que escriban comentarios relacionados a política que puedan aportar a la creación del perfil de popularidad, para cada partido se utilizará cada uno de los nombres de los candidatos que demostraron tener la mayor popularidad en periodos electorales previos, para realizar una búsqueda de todos los comentarios que los mencionen.

5.2.1.1. Descripción del criterio

Para llevar a cabo el proceso se procederá a realizar una búsqueda utilizando cada uno de los nombres de los principales miembros de cada asociación política, este método es directo porque se puede verificar directamente los comentarios que mencionen a los integrantes, aunque exista cierta incertidumbre generada por los homónimos, para fines de este trabajo de investigación se hará la suposición que esta incertidumbre tiende a ser cero.

5.2.1.2. Selección de los integrantes más representativos de cada partido para los últimos tres periodos electorales

Para la selección de los integrantes más representativos se eligió a los candidatos a presidente y para la selección de las agrupaciones políticas por analizar, se seleccionaron los primeros cinco partidos con más votos exceptuando el primero que fue electo, esto considerando que no es válida la reelección en Guatemala.

A continuación, se presentan de forma tabular los resultados de los últimos tres periodos electorales:

Para el 2011, los cuatro partidos con más votos se observan en la tabla I.

Tabla I. **Resultados de elecciones de 2011**

Partido político	Candidato
Libertad Democrática Renovada-LIDER	Candidato 1
Compromiso, Renovación y Orden-CREO	Candidato 2
Unión del Cambio Nacional- UCN	Candidato 3
Visión con Valores-Encuentro por Guatemala-VIVA-EG	Candidato 4

Fuente: Tribunal Supremo Electoral de Guatemala. https://www.tse.org.gt/memoria-electoral-2011.pdf. Consulta: octubre de 2019.

Para el 2015, los cuatro partidos con más votos se observan en la tabla II.

Tabla II. **Resultados de elecciones de 2015**

Partido Político	Candidato
Unidad Nacional de la Esperanza-UNE	Candidato 5
Libertad Democrática Renovada-LIDER	Candidato 1
FUERZA	Candidato 6
Visión con Valores-VIVA	Candidato 7

Fuente: Tribunal Supremo Electoral de Guatemala.
https://www.tse.org.gt/images/memoriaselec/me2015.pdf. Consulta: octubre de 2019.

Para el 2019, los cinco partidos con más votos se observan en la tabla III.

Tabla III. **Resultados de elecciones de 2019**

Partido Político	Candidato
Unidad Nacional de la Esperanza-UNE	Candidato 5
Partido Humanista de Guatemala-PHG	Candidato 8
Movimiento para la Liberación de los Pueblos-MLP	Candidato 9
Partido de Avanzada Nacional-Podemos-PAN	Candidato 10
Movimiento político - WINAQ	Candidato 11

Fuente: Tribunal Supremo Electoral de Guatemala. https:// elecciones2019.tse.org.gt/resultados.
Consulta: octubre de 2019.

Tras haber recopilado esta información, ya se tiene lo que se requería, se tomarán los nombres de cada candidato a presidente de los partidos descritos anteriormente, para realizar la extracción de datos utilizando la API de Twitter.

5.2.2. Obtención de datos desde red social Twitter

Para la utilización de la API que ofrece la red social Twitter es necesario contar con una cuenta de desarrollador otorgada por la plataforma, tras obtener

la aprobación, es posible crear una App y con la cual, será posible generar credenciales para poder acceder a todas las herramientas disponibles incluyendo la API, todo este procedimiento se detalla en el apéndice 1.

Lo siguiente es ejecutar un script que permita la descarga de los comentarios utilizando el conjunto de credenciales generadas y dada una cadena de entrada que representa el criterio de búsqueda, en este caso cada uno de los nombres de los candidatos seleccionados, este proceso se detalla en el apéndice 2.

5.3. Manejo de los datos

Contando con la estructura provista por la solicitud a la API, se cuenta con una estructura de tipo Cursor, esta estructura cuenta con más cosas de las que se requieren, por eso es necesario la utilización de una librería que facilite el manejo de estos datos y permita el almacenamiento de estos.

Para realizar la solución a un caso de estudio particular fue necesario instalar herramientas que permitieran el análisis de los datos.

5.3.1. Interprete de Python 3

Lo primero que se procedió a realizar es la instalación del intérprete de Python3, se descarga el instalador desde la página oficial de Python https://www.python.org/downloads/, el método de instalación varía entre los distintos sistemas operativos soportados.

5.3.2. Librería xlswriter

Xlswriter es una librería auxiliar para el manejo de archivos de extensión xls o xlsx, que contiene una serie de procedimientos hechos para la fácil escritura y manejo de estos.

Gracias a las estructuras provistas por pandas y la facilidad en la escritura de hojas de cálculo, utilizando xlswriter es posible almacenar la fuente de datos de forma tabular en archivos de extensión .xlsx para su posterior análisis, este procedimiento se detalla en el apéndice 3

5.4. Análisis sentimental

La siguiente etapa en el proceso es realizar el análisis del archivo de extensión .xlsx que contiene los datos relacionados a los comentarios obtenidos, este procedimiento puede verse claramente en el apéndice 4, mediante un script de código Python que recoge la información de la base de datos, en este caso el archivo .xlsx y haciendo uso de la librería TextBlob, de modo que retorna el grado de polaridad, donde un numero negativo, representa polaridad negativa, cero, representa polaridad neutra y un número positivo, representa polaridad positiva.

5.4.1. Tipo de análisis sentimental a utilizar

Se utilizará el tipo de análisis sentimental de grano fino, por su simplicidad y por la limitada cantidad de posibles resultados de polaridad: positivo, negativo o neutro.

5.4.2. Descripción de criterio para cálculo del perfil popularidad

El perfil de popularidad del partido se efectuará en base al candidato presidencial que lo represente. La ponderación para cada candidato será calculada de la siguiente forma: $popularidad = positivas - negativas + 0 * neutrales$ esto se realizará para los últimos tres periodos electorales.

5.5. Representando datos

Tras haber realizado el análisis sentimental para todos los candidatos, se tabularon los datos, luego se representaron gráficamente utilizando gráficos de barras, a continuación, se presentan las gráficas que representan los resultados para cada uno de los candidatos analizados. En la tabla IV se muestra una representación gráfica de los resultados obtenidos.

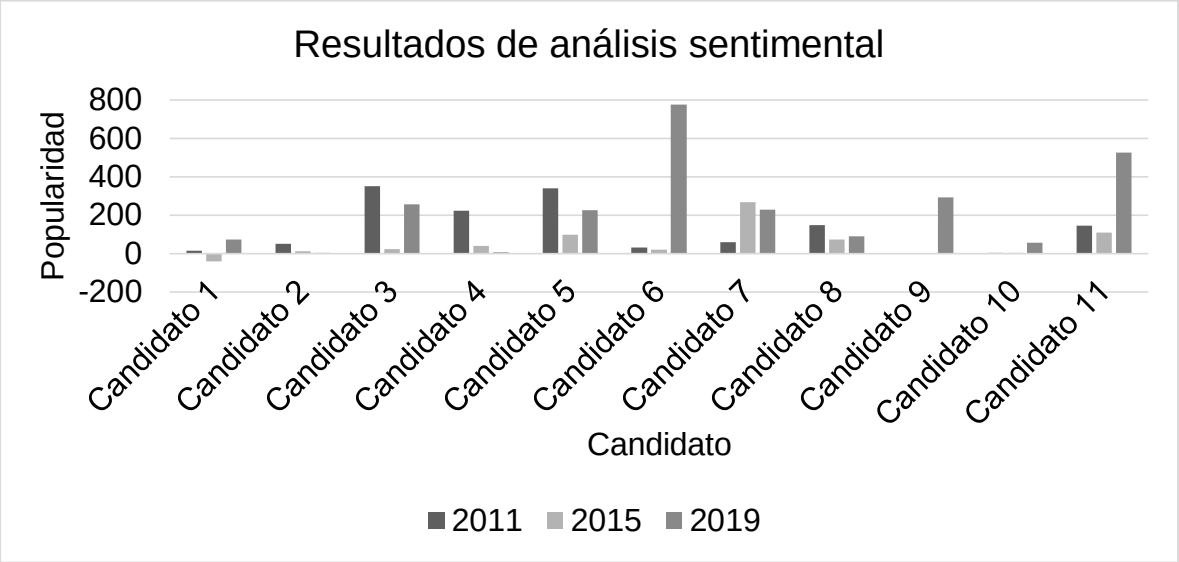
Tabla IV. **Resultados del análisis sentimental**

Resultados de popularidad			
Candidato	**2011**	**2015**	**2019**
Candidato 1	16	-40	74
Candidato 2	50	11	3
Candidato 3	350	23	257
Candidato 4	224	39	7
Candidato 5	340	98	227
Candidato 6	32	21	775
Candidato 7	60	268	229
Candidato 8	147	72	90
Candidato 9	0	0	292
Candidato 10	1	2	57
Candidato 11	144	108	525

Fuente: elaboración propia.

A continuación, en la figura 2, se presenta la representación gráfica para los resultados de la tabla IV.

Figura 2. **Representación gráfica de los resultados del análisis sentimental**



Fuente: elaboración propia.

5.6. Análisis de regresión lineal

Tras representar individualmente cada periodo se procede a realizar el análisis de regresión lineal para cada candidato, esto para estimar su punteo para el siguiente periodo electoral, en el año 2023.

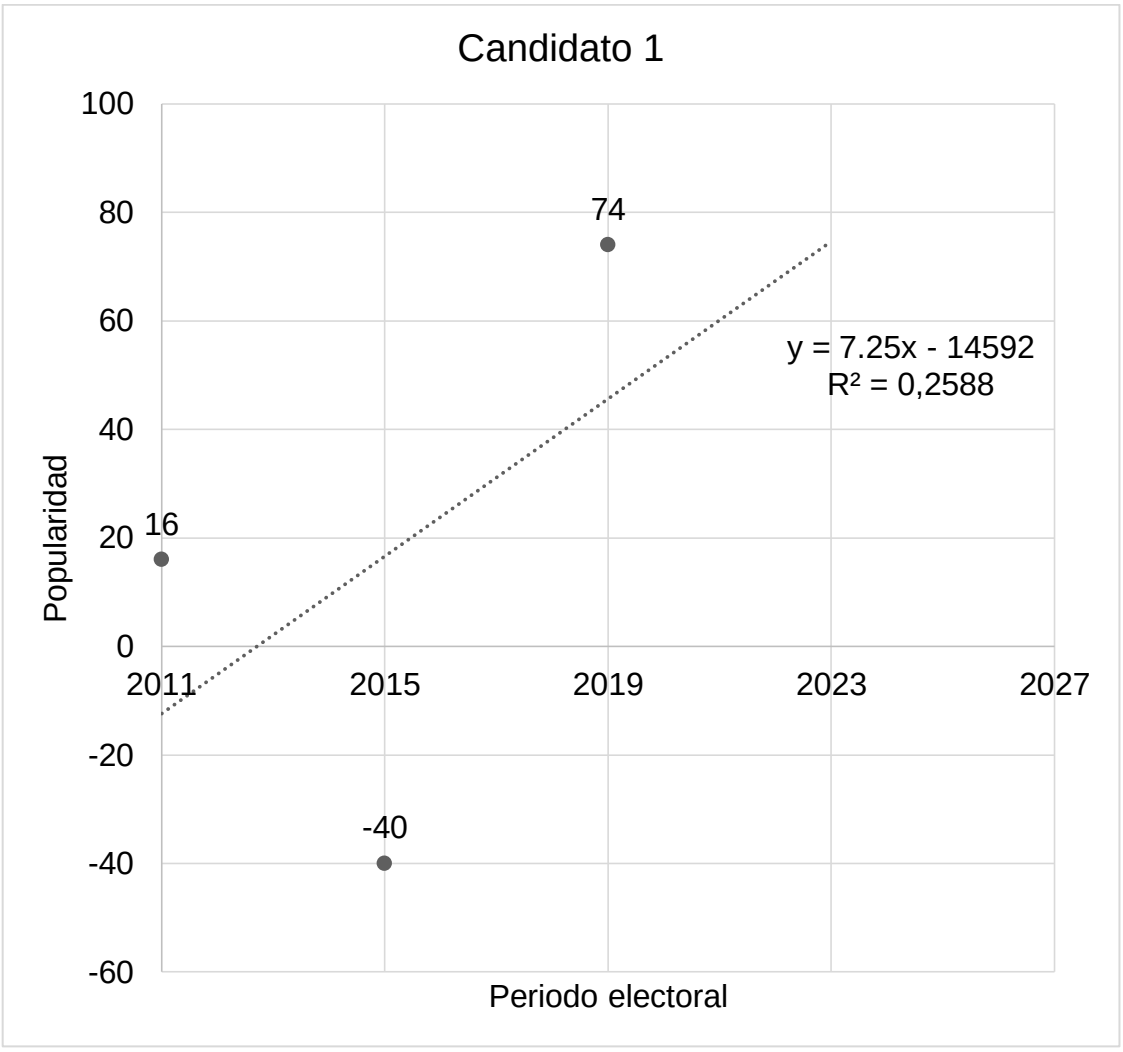
5.6.1. Modelo predictivo para cada candidato

A continuación, se presenta cada uno de los análisis de regresión lineal simple para los candidatos seleccionados.

5.6.1.1. Candidato 1

Se observa en la figura 3 el modelo de predicción para el candidato por el partido Líder.

Figura 3. **Análisis de regresión lineal candidato 1**

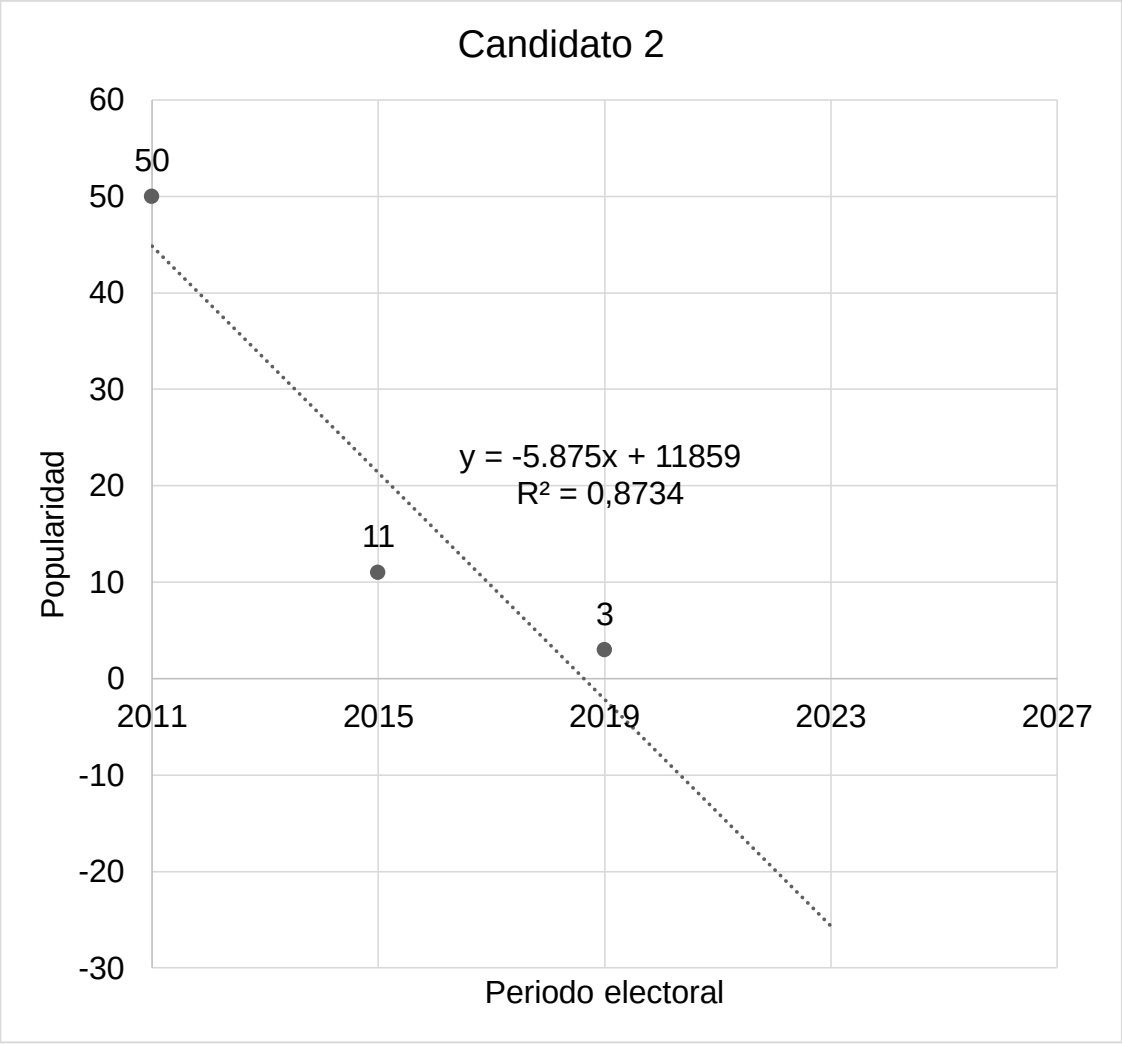


Fuente: elaboración propia.

5.6.1.2. Candidato 2

Se observa en la figura 4 el modelo de predicción para el candidato por el partido Creo.

Figura 4. **Análisis de regresión lineal candidato 2**

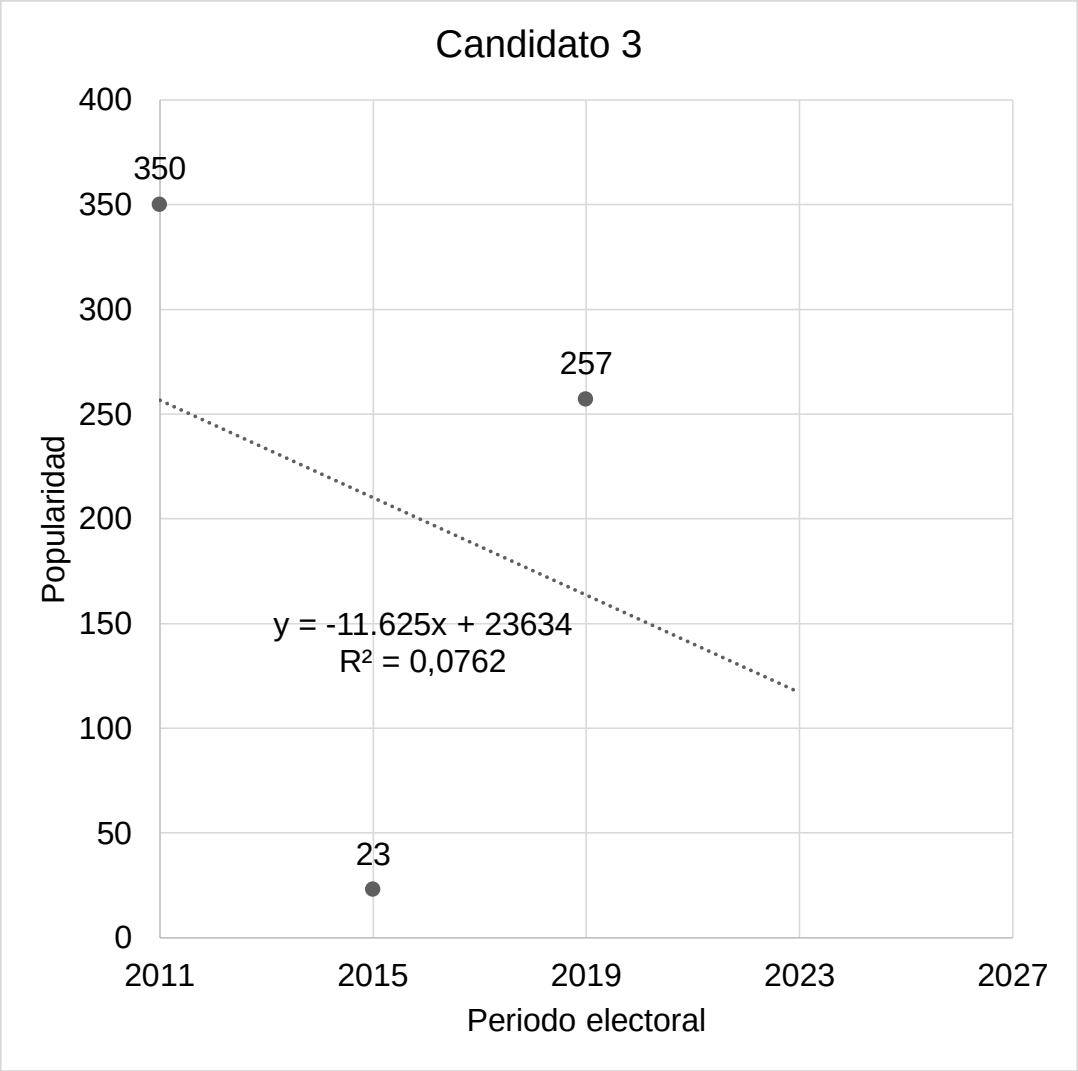


Fuente: elaboración propia.

5.6.1.3. Candidato 3

Se observa en la figura 5 el modelo de predicción para el candidato por el partido UCN.

Figura 5. **Análisis de regresión lineal candidato 3**

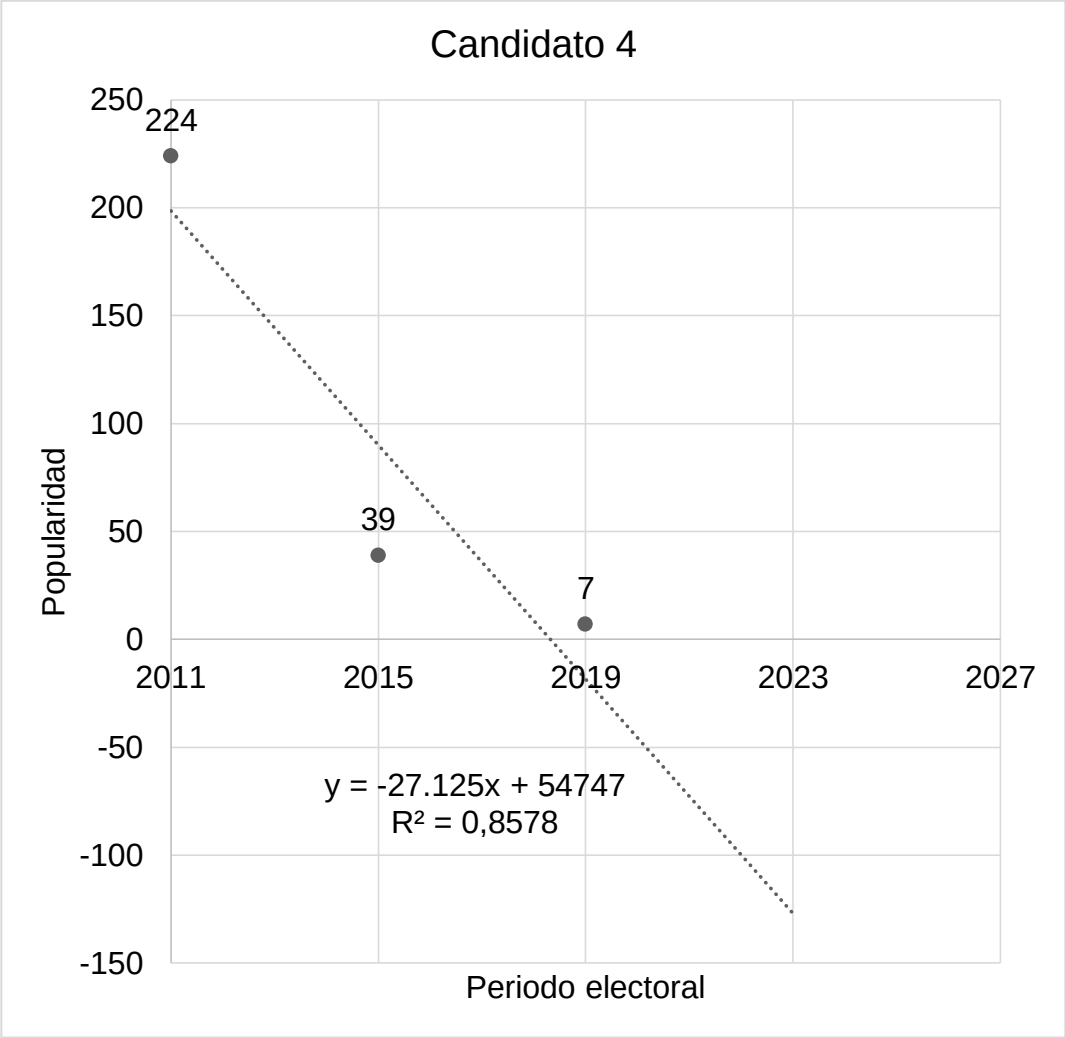


Fuente: elaboración propia.

5.6.1.4. Candidato 4

Se observa en la figura 6 el modelo de predicción para el candidato por el partido VIVA.

Figura 6. **Análisis de regresión lineal candidato 4**

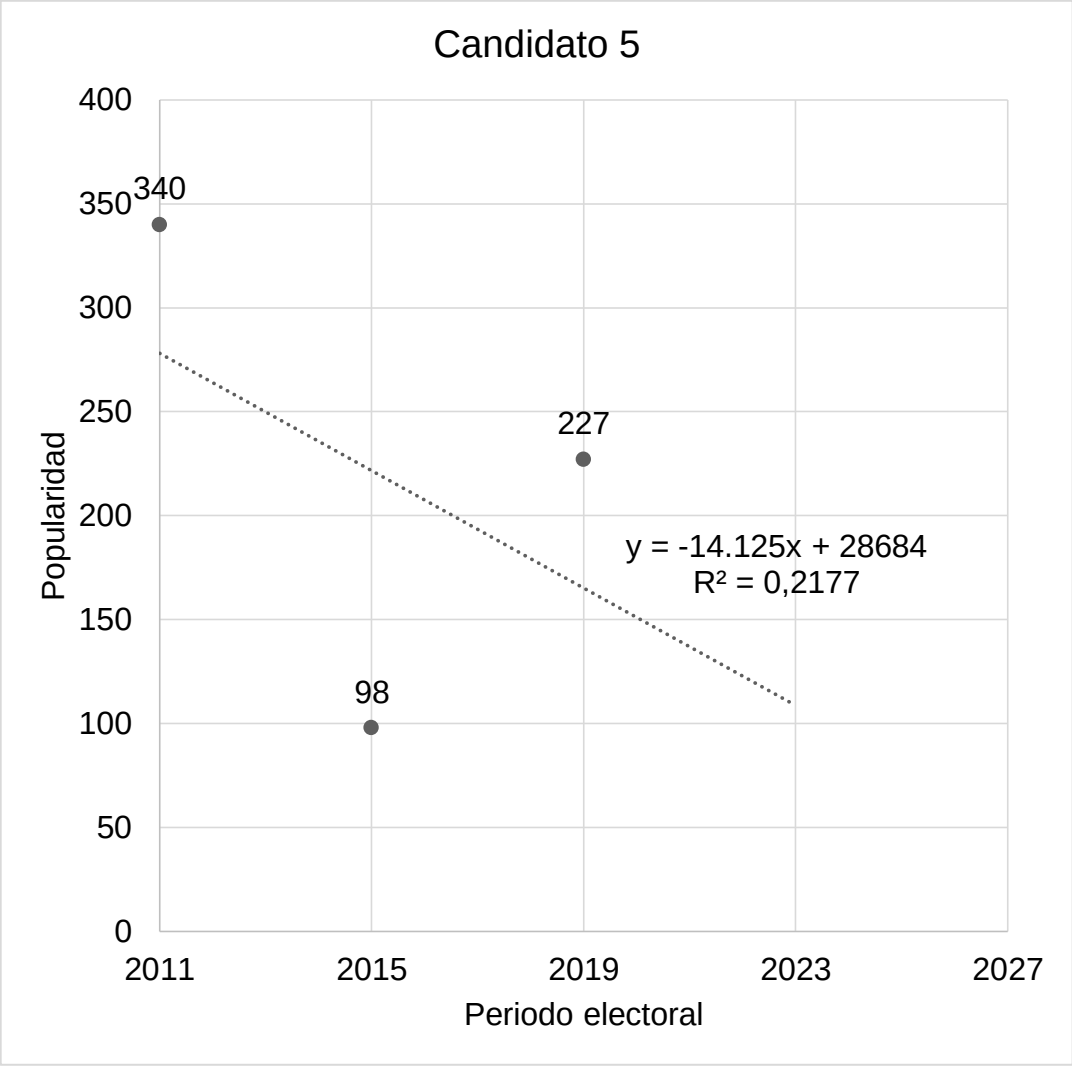


Fuente: elaboración propia.

5.6.1.5. Candidato 5

Se observa en la figura 7 el modelo de predicción para el candidato por el partido UNE.

Figura 7. **Análisis de regresión lineal candidato 5**

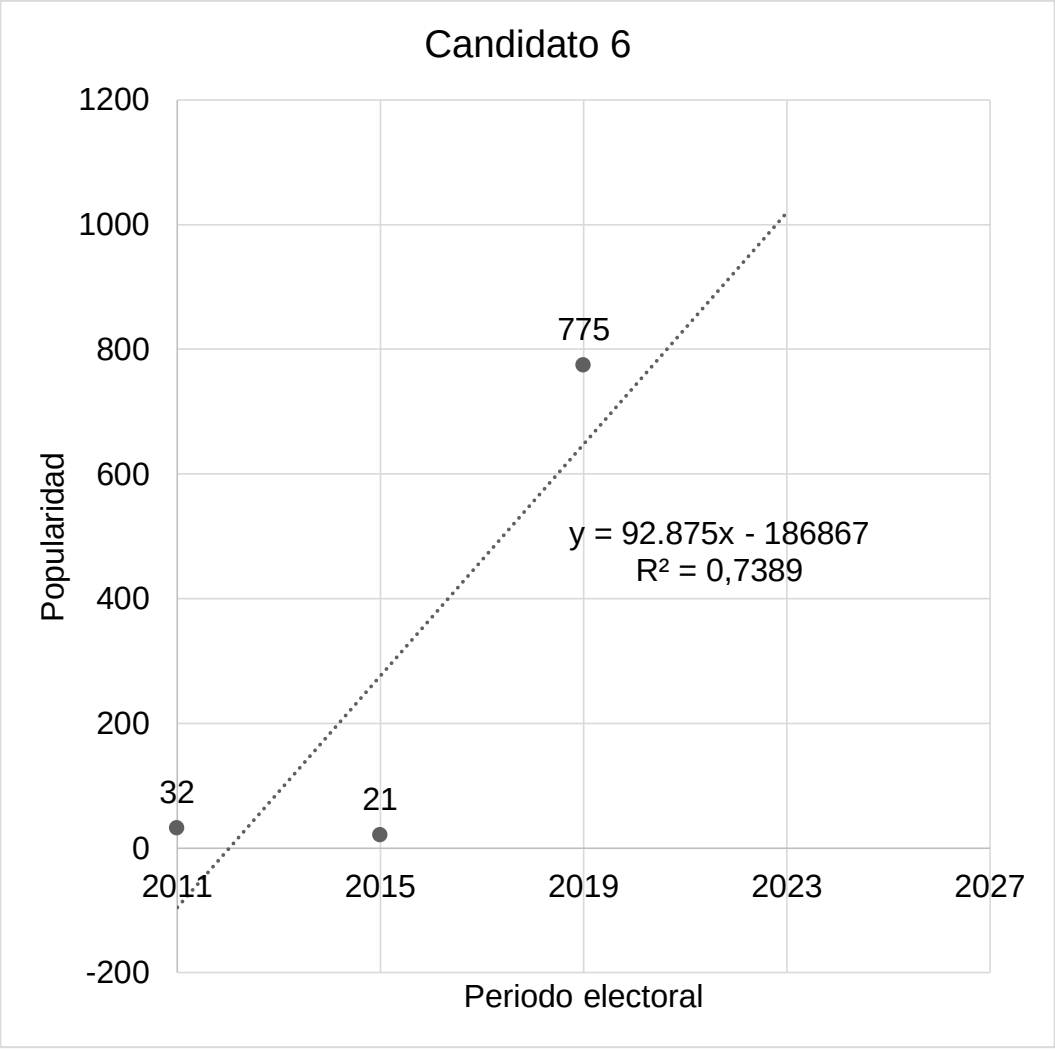


Fuente: elaboración propia.

5.6.1.6. Candidato 6

Se observa en la figura 8 el modelo de predicción para el candidato por el partido Fuerza.

Figura 8. **Análisis de regresión lineal candidato 6**

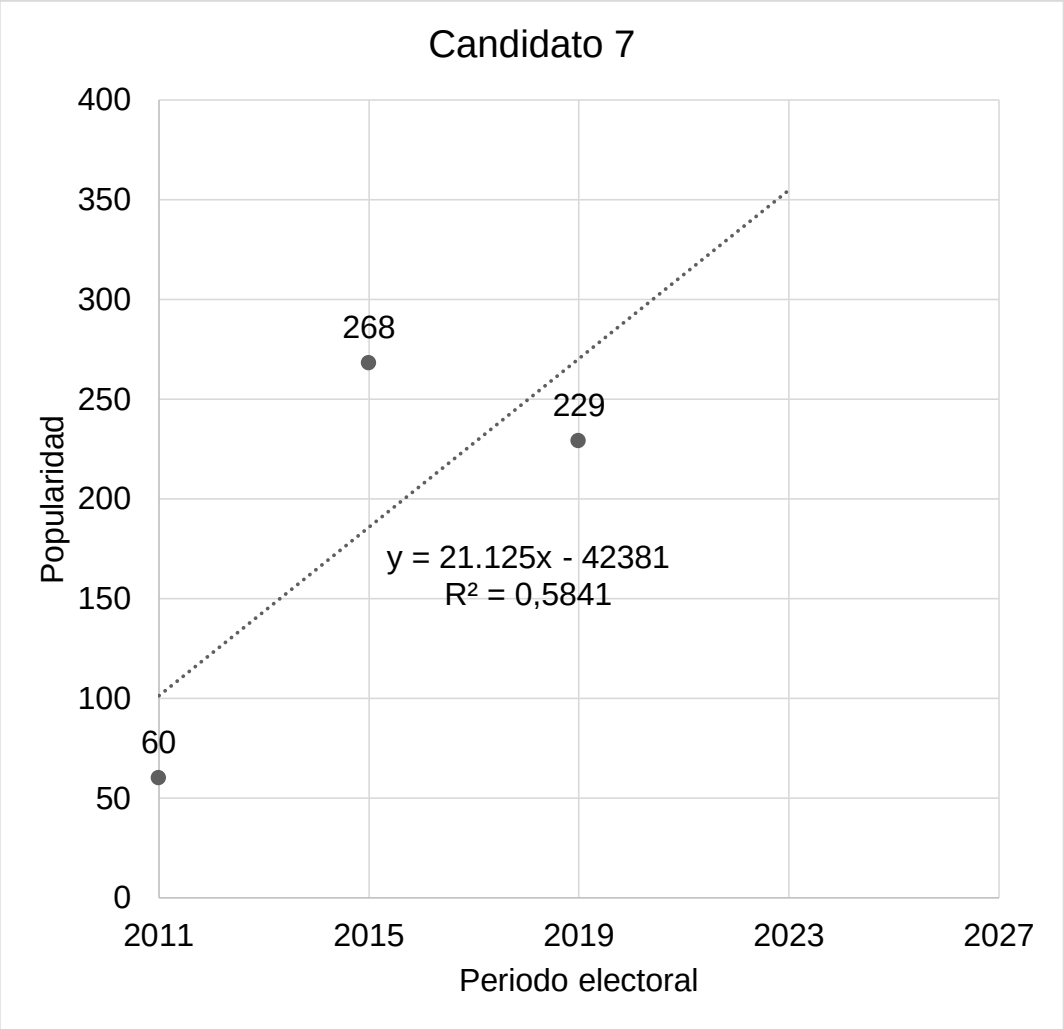


Fuente: elaboración propia.

5.6.1.7. Candidato 7

Se observa en la figura 9 el modelo de predicción para el candidato por el partido VIVA.

Figura 9. **Análisis de regresión lineal candidato 7**

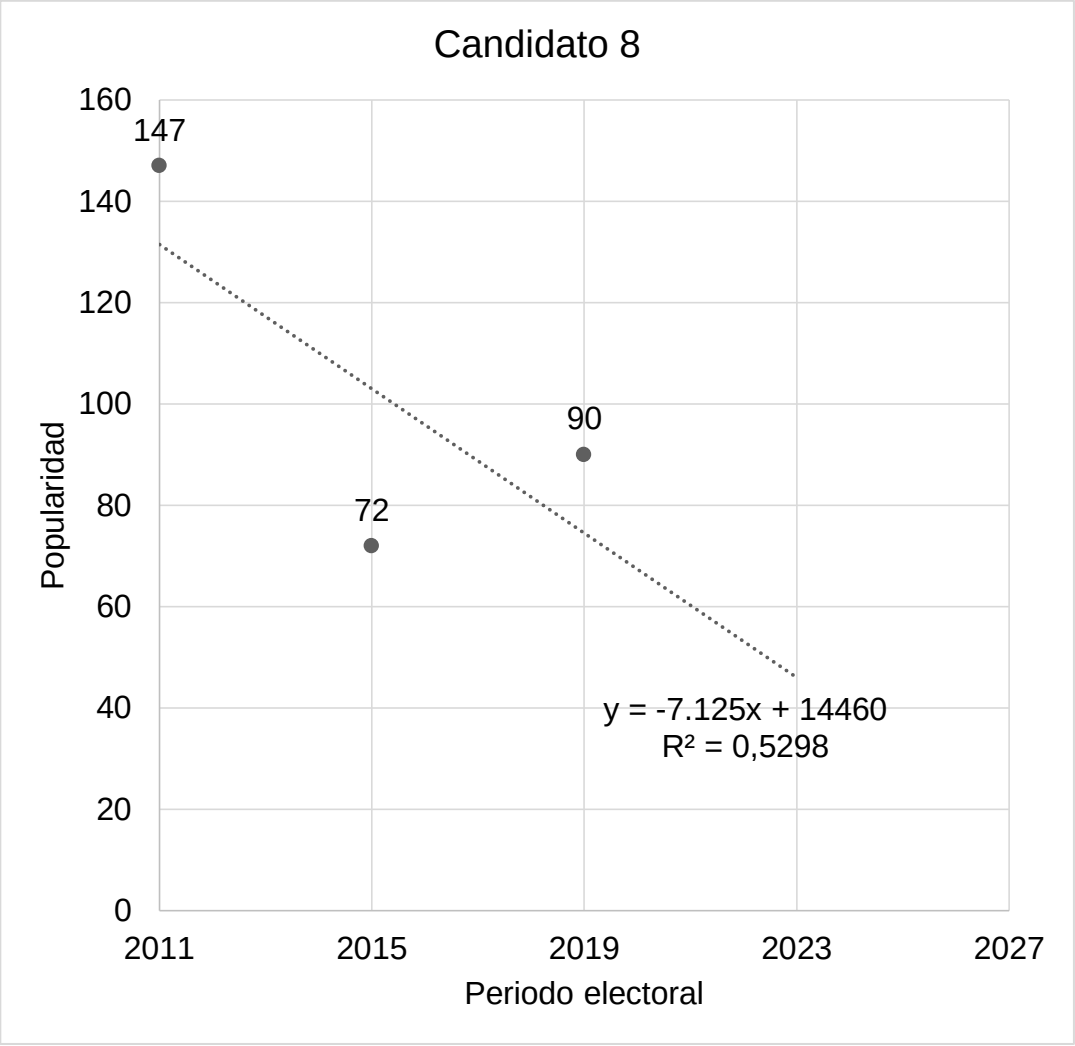


Fuente: elaboración propia.

5.6.1.8. Candidato 8

Se observa en la figura 10 el modelo de predicción para el candidato por el partido PHG.

Figura 10. **Análisis de regresión lineal candidato 8**

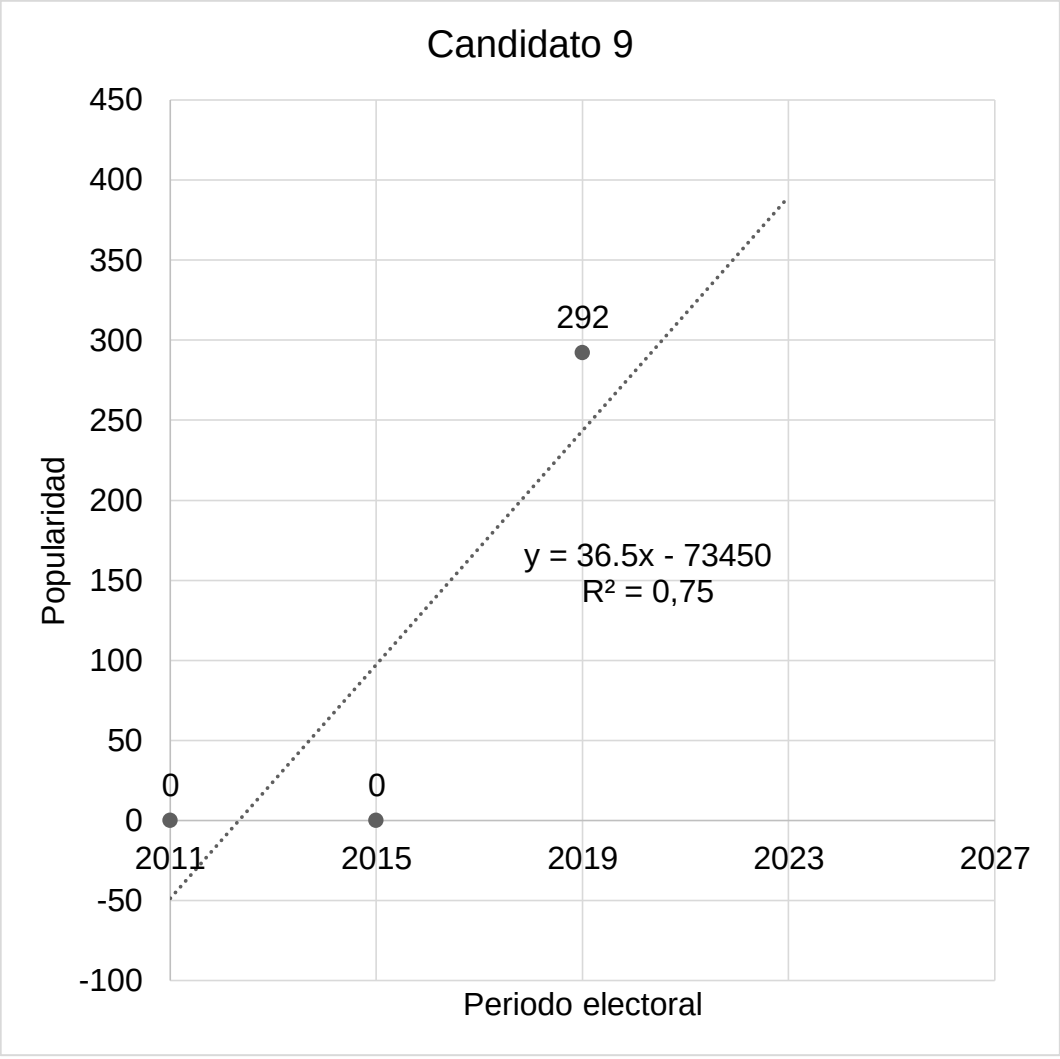


Fuente: elaboración propia.

5.6.1.9. Candidato 9

Se observa en la figura 11 el modelo de predicción para el candidato por el partido MLP.

Figura 11. **Análisis de regresión lineal candidato 9**

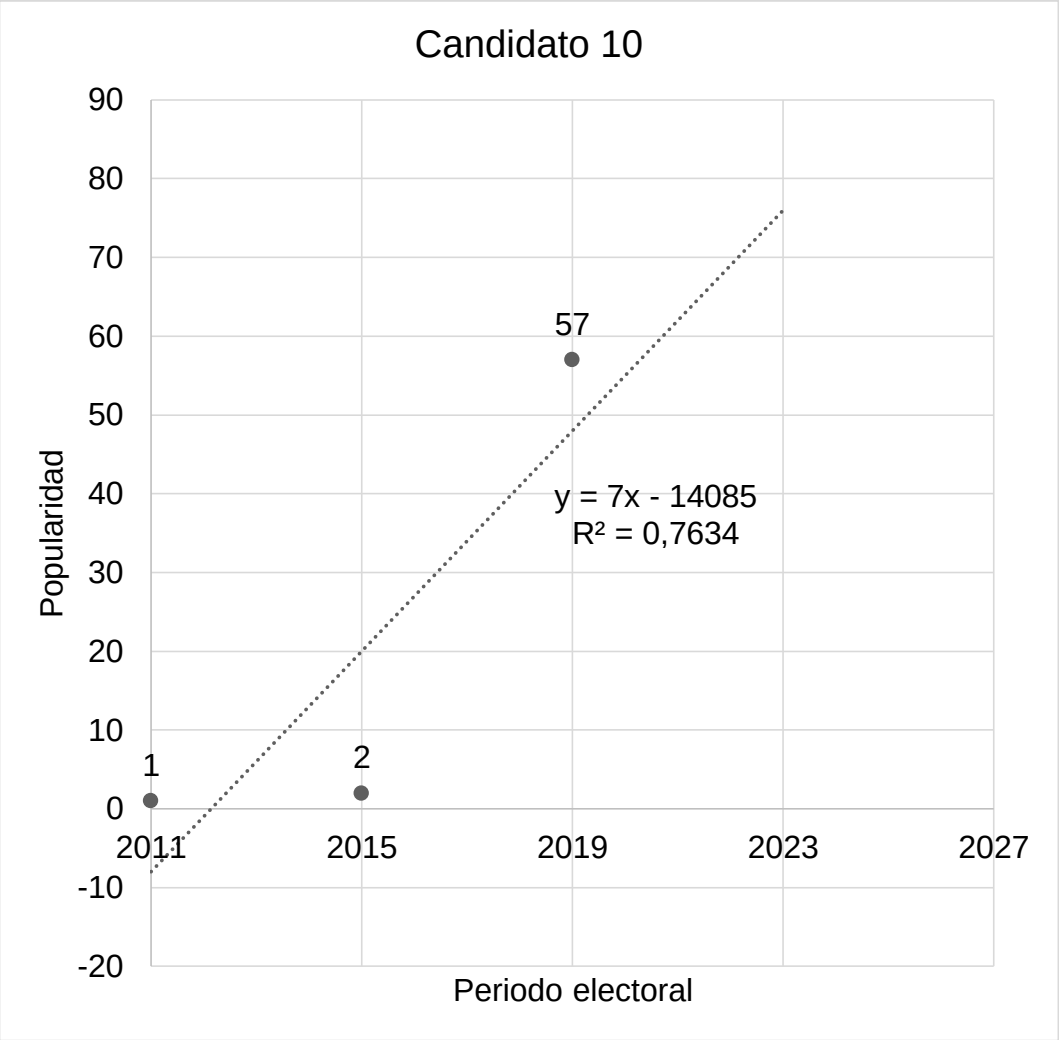


Fuente: elaboración propia.

5.6.1.10. Candidato 10

Se observa en la figura 12 el modelo de predicción para el candidato por el partido PAN.

Figura 12. **Análisis de regresión lineal candidato 10**

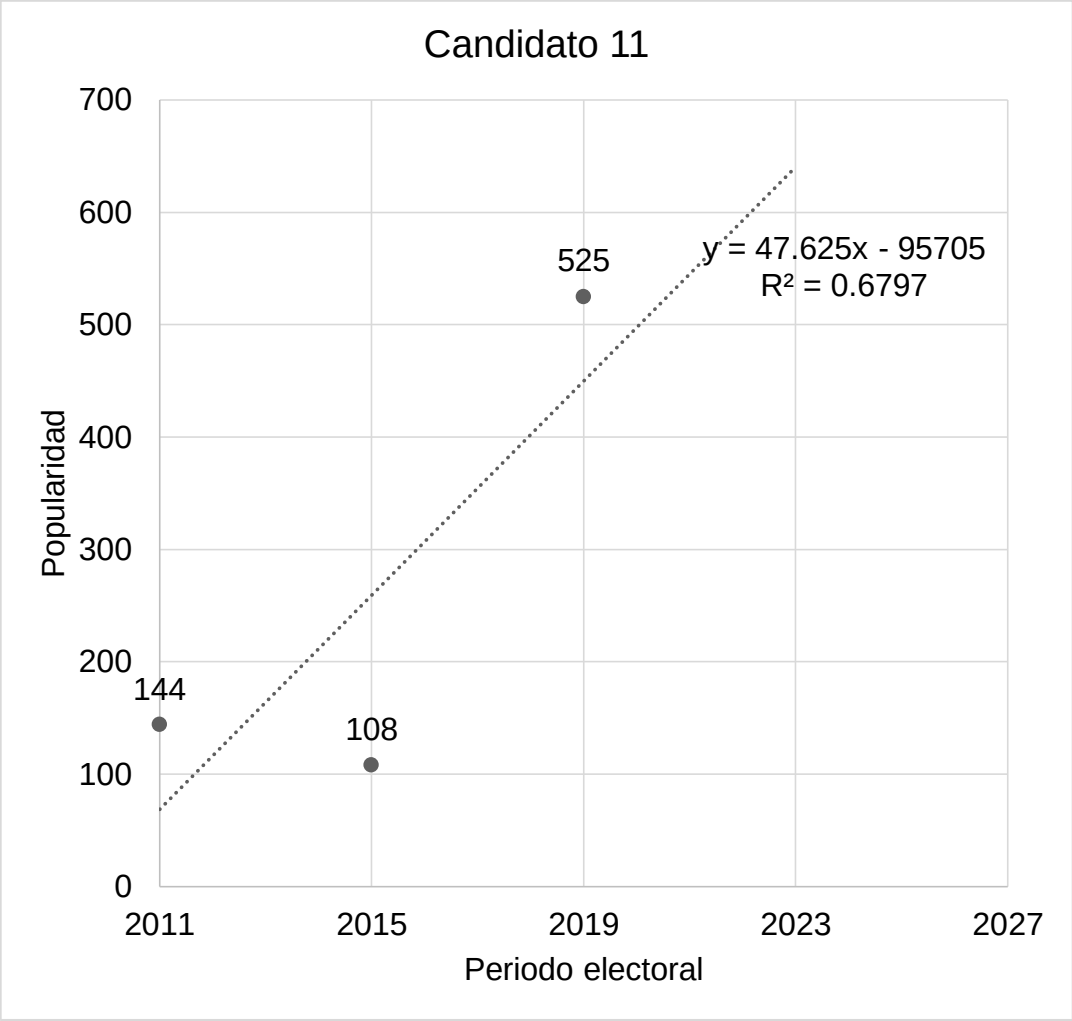


Fuente: elaboración propia.

5.6.1.11. Candidato 11

Se observa en la figura 13 el modelo de predicción para el candidato por el partido WINAQ.

Figura 13. **Análisis de regresión lineal candidato 11**



Fuente: elaboración propia.

5.6.2. Predicción periodo electoral 2023

Tras haber obtenido el modelo lineal predictivo de regresión es posible valuar la variable independiente (el año), en esta y obtener una predicción para tal año, al realizarlo con cada una de las rectas de regresión se obtuvieron los resultados que se muestran en la tabla VI.

5.6.2.1. Calidad de ajuste del modelo

La calidad de ajuste del modelo se representa mediante el coeficiente de determinación, en la tabla V es posible observar cada uno de los coeficientes para los análisis de los candidatos:

Tabla V. **Coeficientes de determinación**

Predicción periodo electoral 2023	
Candidato	**Coeficiente de determinación**
Candidato 1	0,2588
Candidato 2	0,8734
Candidato 3	0,0762
Candidato 4	0,8578
Candidato 5	0,2177
Candidato 6	0,7389
Candidato 7	0,5841
Candidato 8	0,5298
Candidato 9	0,75
Candidato 10	0,7634
Candidato 11	0,6797

Fuente: elaboración propia.

En la tabla V se observa que los coeficientes de determinación en su mayoría no se acercan a uno, y el modelo no está lo suficientemente entrenado, esto se debe a que únicamente se tienen tres observaciones, el modelo podría tener una mejor calidad de ajuste si el análisis se realizara en un futuro con mayor cantidad de observaciones. Otra opción sería evaluar la implementación de otro tipo de regresión dependiendo de la tendencia de los datos.

5.6.2.2. Resultados de regresión

A continuación, se presentan los resultados obtenidos del análisis de regresión para cada candidato.

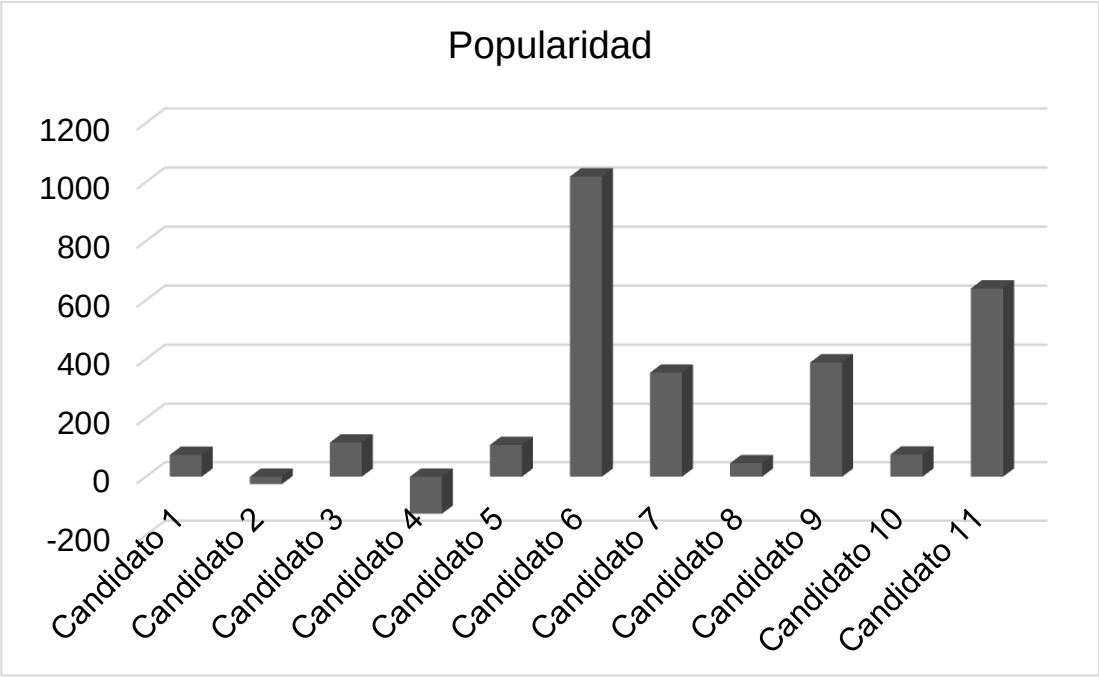
Tabla VI. **Resultados del análisis de regresión para cada candidato**

Predicción periodo electoral 2023	
Candidato	**Popularidad**
Candidato 1	75
Candidato 2	-26
Candidato 3	117
Candidato 4	-127
Candidato 5	109
Candidato 6	1 019
Candidato 7	355
Candidato 8	46
Candidato 9	389
Candidato 10	76
Candidato 11	640

Fuente: elaboración propia.

Por último, en la figura 14 se procede a presentar los resultados del análisis de regresión en un gráfico de barras que representa la predicción para el 2023.

Figura 14. **Predicción para periodo electoral 2023**



Fuente: elaboración propia.

5.6.2.3. Interpretación de los resultados de regresión

Se puede observar en la figura 14 que el candidato con más popularidad será el candidato 6 del partido LIDER, pero el 11 de agosto de 2019, en segunda vuelta, el candidato a presidente de dicho partido quedó electo, por lo que queda descartado de la predicción debido a que es prohibida la reelección en Guatemala, según el artículo 187 de la Constitución Política de Guatemala, resultado que concuerda con los resultados de la tabla III, porque el candidato con mayor popularidad en 2019 previo a las elecciones fue el candidato 6, por lo tanto el candidato con mayor popularidad será el candidato 11, del partido WINAQ seguido por el candidato 9, del partido MLP.

5.7. Fiabilidad del modelo considerando factores tecnológicos, sociales y económicos de Guatemala

Es importante resaltar que factores tienen impacto en la fiabilidad del modelo propuesto, a continuación, se enlistan los principales:

- Factores tecnológicos: existe gran cantidad de población que vota en las elecciones que posee un celular inteligente, pero no tienen un usuario en la red social Twitter.
- Factores sociales: no en todos los casos los comentarios dan a conocer la polaridad correcta debido al vocabulario de los guatemaltecos, en la muestra tomada se observaron comentarios de polaridad negativa que en un contexto más específico serian de polaridad positiva.
- Factores económicos: existe gran cantidad de población que vota en las elecciones que no posee recursos económicos para comprar un celular inteligente que posea la red social Twitter instalada, y esto implica que tampoco tienen un usuario en dicha plataforma, y no pueden ser parte de la muestra.

Estos factores tienen un gran impacto en la fiabilidad del modelo, porque esta reduce.

CONCLUSIONES

1. El criterio más adecuado de selección para una muestra representativa de la opinión publica guatemalteca referente a la política proveniente de la red social Twitter, es el de extraer todos los comentarios que contienen el nombre de cada candidato.

2. El método de análisis sentimental seleccionado para realizar un análisis de la muestra recopilada es el de polaridad.

3. Se construyó el perfil de popularidad de cada candidato a la presidencia por partido, utilizando como puntos a favor cada *tweet* con polaridad positiva, en contra cada *tweet* con polaridad negativa y se ignoró la polaridad neutra.

4. Se construyó un modelo predictivo de regresión utilizando los perfiles de cada partido político en los últimos tres periodos electorales, con una calidad de ajuste medida con el coeficiente de determinación de cada recta de regresión, el menor ajuste lo tuvo el modelo para el candidato 3, del partido UCN y el que tuvo el mejor ajuste fue el del candidato 2 del partido CREO.

5. Tras analizar la predicción de popularidad para cada partido se concluye que para el siguiente periodo electoral el candidato 11, del partido WINAQ será el que tendrá mayor aceptación por parte de los usuarios de Twitter.

RECOMENDACIONES

1. Para la selección de la muestra se debe tomar en cuenta que no todos los comentarios relacionados a la política brindan una opinión acerca de un integrante específico de algún partido.

2. Al realizar el análisis sentimental con el método de polaridad, no debe ser de interés sobre qué trate el *tweet* realizado, porque con el simple hecho de que un usuario opine negativamente de un candidato la popularidad de este se reduce o de lo contrario si es positiva la opinión, esta aumenta.

3. La popularidad debe ser ponderada basándose en los comentarios positivos y negativos.

4. Para mejorar la calidad de ajuste se puede considerar aplicar algún otro modelo de regresión diferente al lineal, esto puede decidirse al conocer la tendencia de los datos.

5. La popularidad definida por una sola red social no determina la popularidad en otra red social, esto debido a que el tipo de usuarios de cada red social varia y con esto sus opiniones.

BIBLIOGRAFÍA

1. CARBALLAR, José Antonio. Social Media. *Marketing personal y profesional.* México : Alfaomega, 2013. 245 p.

2. ESPINO, Carlos. *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo.* Tesis de grado de Ing. Universidad de Cataluña , 2017. 65 p.

3. GARTNER. *Gartner Glossary.* [en línea]. <https://www.gartner.com/en/information-technology/glossary/big-data>. [Consulta: 01 de julio de 2019].

4. LORIA, Steven. TextBlob: *Simplified text processing*. [en línea]. <https://textblob.readthedocs.io/en/dev>. [Consulta: 01 de octubre de 2019].

5. MONKEYLEARN. *Sentiment analysis*: Complete guide. [en línea]. <https://monkeylearn.com/sentiment-analysis/>. [Consulta: 01 de octubre de 2019].

6. MORENO, Antonio. *Procesamiento del lenguaje natural.* [en línea]. <https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/>. [Consulta: 15 de junio de 2019].

7. Organización de Estados Americanos. *Proteccion de datos personales.* [en línea].

<http://www.oas.org/es/sla/ddi/proteccion_datos_personales.asp>. [Consulta: 01 de julio de 2019].

8. PEREIRA, Augusto. *Análisis predictivo de datos mediante técnicas de regresión estadística.* Tesis de maestría. Universidad Complutense de Madrid, 2010. 61 p.

9. ROUSE, Margaret. *Ciencia de datos.* [en línea]. <https://searchdatacenter.techtarget.com/es/definicion/Ciencia-de-datos>. [Consulta: 01 de julio de 2019].

10. SOTO, Marvin. Episodio 1: *Procesamiento de lenguaje natural*. [en línea]. <https://planetachatbot.com/1-procesamiento-de-lenguaje-natural-1443ff471ed0>. [Consulta: 01 de junio de 2019].

11. Twitter. *Política de privacidad de Twitter*. [en línea]. <https://twitter.com/es/privacy/previous/version_12>. [Consulta: 01 de julio de 2019].

APÉNDICES

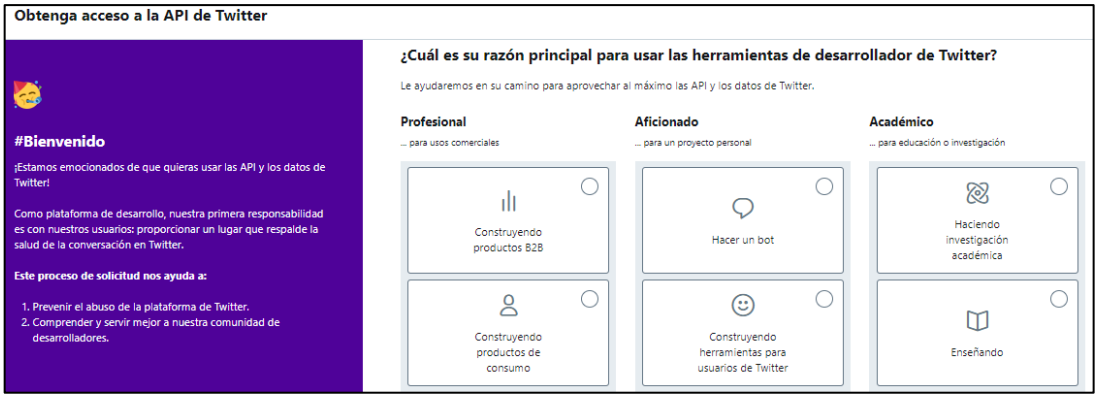
Apéndice 1. **Utilización de la API de Twitter**

La API de Twitter para extraer datos es brindada por la plataforma bajo estrictas normas, para aplicar a crear una cuenta de desarrollador es necesario realizar un proceso, básicamente se resume a los siguientes pasos:

Aplicar para la creación de una cuenta de desarrollador

Para acceder a las herramientas que ofrece Twitter a los desarrolladores es necesario llenar un formulario, que inicia solicitando al aplicante que defina la razón por la que desea aplicar, tal como se observa en la figura 1.

Figura 1. **Formulario de aplicación Twitter *Developer***



Fuente: *Twitter Developer.* https://developer.twitter.com/en/application/use-case. Consulta: julio de 2019.

Continuación apéndice 1.

El proceso requiere que el aplicante posea una cuenta en Twitter y brinde la información personal, luego se tiene que detallar el uso que se le dará a la información que se consuma mediante las herramientas, posterior a eso se requerirá que se revisen los datos y que se acepten las condiciones. Básicamente el formulario sigue el proceso observado en la figura 2.

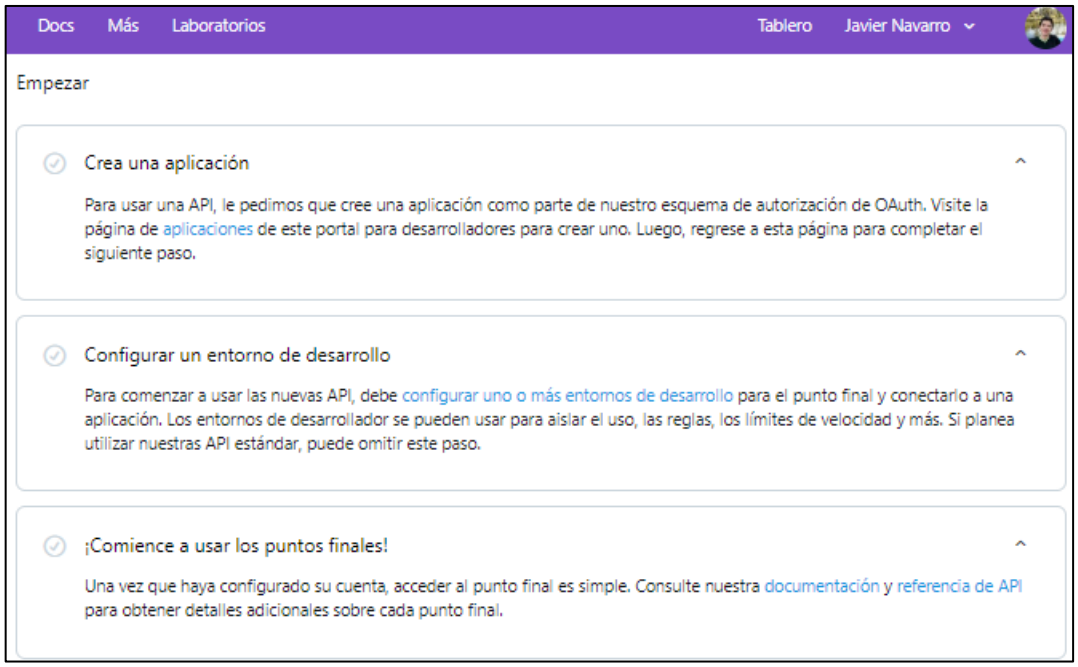
Figura 2. **Flujo de formulario de aplicación, Twitter *Developer***

Fuente: *Twitter Developer.* https://developer.twitter.com/en/application/login?useCase=12. Consulta: julio de 2019.

Continuación apéndice 1.

Si la aplicación es aprobada por Twitter, se recibirá un enlace vía correo electrónico que mostrará un tablero de trabajo como el que se observa en la figura 3.

Figura 3. **Tablero de la cuenta de desarrollador de Twitter**



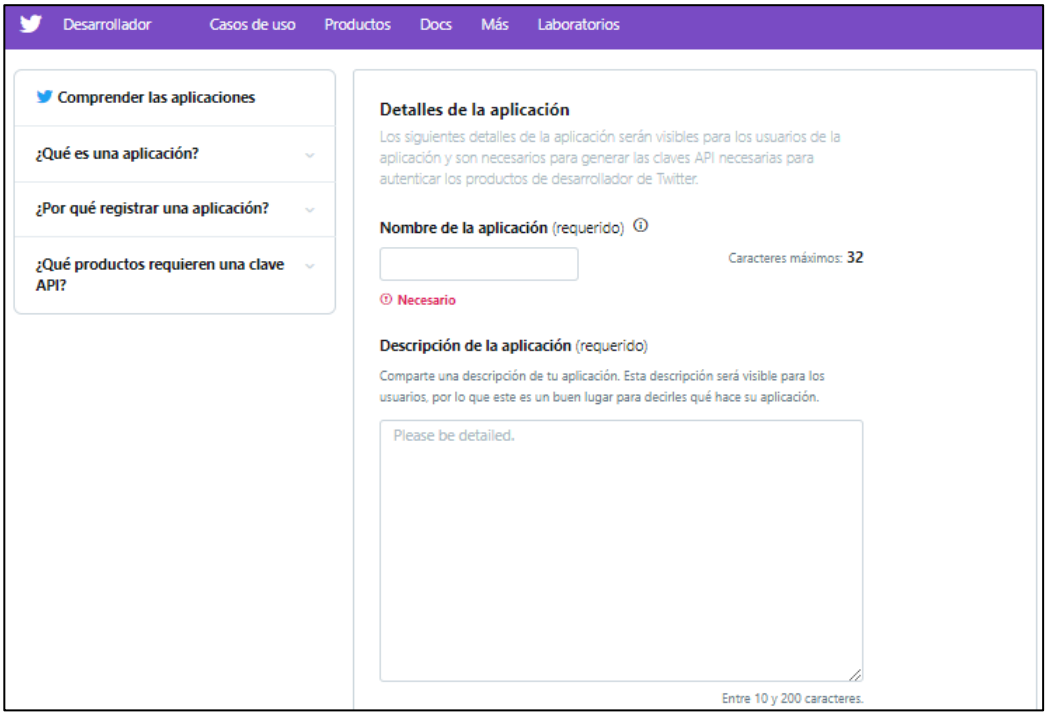
Fuente: *Twitter Developer.* https:// developer.twitter.com/en/account/get-started. Consulta: julio de 2019.

Crear una aplicación y obtener credenciales

Para crear una aplicación es necesario dirigirse al tablero, y luego llenar un pequeño formulario como el que se observa en la figura 4.

Continuación apéndice 1.

Figura 4. **Formulario para crear una App en Twitter**



Fuente: Twitter Developer. https:// developer.twitter.com/en/account/get-started. Consulta: julio de 2019.

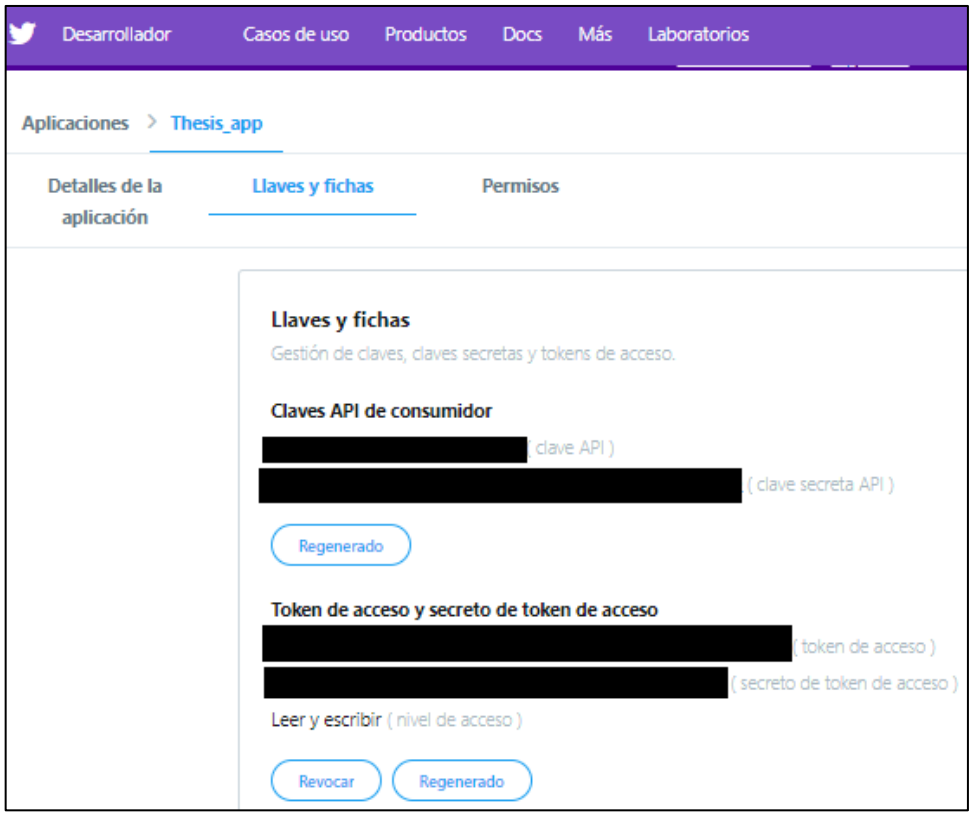
Posterior a la creación de la App, será posible generar las credenciales que permitan el acceso a la API y así que sea posible descargar los comentarios que conformarán la fuente de datos, estas credenciales son conformadas por los siguientes elementos:

- Clave de la API
- Clave secreta de la API
- *Token* de acceso
- *Token* de acceso secreto

Continuación apéndice 1.

Juntos conforman el método de autenticación utilizado por la API para regular el acceso a esta. Se genera al ingresar a las opciones de la App, tal como se aprecia en la figura 5.

Figura 5. **Formulario para generar credenciales para la App**



Fuente: *Twitter Developer.* https://developer.twitter.com/en/apps/17071191. Consulta: julio de 2019.

En la figura 5 se muestra que con esto se logró tener acceso a la API que ofrece Twitter, y ya es posible la extracción de los datos.

Apéndice 2. **Descarga de comentarios utilizando la API de Twitter**

Se utilizará el lenguaje de programación Python para desarrollar un script con el que se logre el objetivo de construir la fuente de datos compuesta por los comentarios, esta tarea se describe a continuación:

Instanciar la API

Para instanciar un objeto que represente la API es necesario autenticarse previamente con las credenciales obtenidas, ver apéndice 1. En la figura 6 se observa el código en Python para instanciar un objeto de la API.

Figura 6. **Instancia de la API utilizando Python**

```python
#!/usr/bin/python
# -*- coding: utf-8 -*-

# Librería que representa la API de Twitter para Python
import tweepy
# Librería que gestiona las fechas
import datetime
# Libreria que permite gestionar la escritura de archivos xlsx
import xlsxwriter
# Libreria con herramientas del sistema
import sys

# Credenciales para el uso de la API de Twitter

consumerKey = "LuDB26QlP7U9FyAXDVU471yIz"
consumerSecret = "nrvODcrQFrNqjhk2rvgfPzlEkb9ogBfOa1Y3tuzNRcpTheDiJQ"
accessToken = "723932016065040386-E8uFaNaYBrPm0OB0WLN93P1HRwfXnQ9"
accessTokenSecret = "1XSh7u0tdWehRNkwCC509vSzaarbOOi3dP337G1nLbnSs"

# Se utilizar un gestor de autenticación con la clave y el secret del usuario
auth = tweepy.OAuthHandler(consumerKey, consumerSecret)
# Se setea el token de acceso junto con su secret
auth.set_access_token(accessToken, accessTokenSecret)

# Se define el objeto que representa una instancia de la API
api = tweepy.API(auth)
```

Fuente: elaboración propia.

Continuación apéndice 2.

Obtención de datos y filtrado

Se procede a utilizar la funcionalidad *search* que ofrece la API para realizar una búsqueda avanzada, básicamente se requiere poder buscar los comentarios que involucren opiniones relacionadas a los candidatos seleccionados como muestra en periodos de tiempo definidos (2011, 2015 y 2019). Ver figura 7.

Figura 7. **Utilización de la funcionalidad search de la Api**

```
# Se define el criterio de busqueda
query = 'Sandra Torres'
# La fecha de inicio y la fecha final, para obtener un periodo electoral determinado
startDate = datetime.datetime(2019, 1, 1, 0, 0, 0)
endDate =   datetime.datetime(2019, 9, 1, 0, 0, 0)

# Una lista que contendrá los Tweets que cumplan con el intervalo de tiempo
tweets = []
# Una lista temporal de tweets sin filtrar
tmpTweets = api.search(q=query)
# se filtra la lista
for tweet in tmpTweets:
    if tweet.created_at < endDate and tweet.created_at > startDate:
        tweets.append(tweet)

while (tmpTweets[-1].created_at > startDate):
    tmpTweets = api.search(q=query, max_id = tmpTweets[-1].id)
    for tweet in tmpTweets:
        if tweet.created_at < endDate and tweet.created_at > startDate:
            tweets.append(tweet)
```

Fuente: elaboración propia.

Al finalizar este segmento de código se tendrá en la lista comentarios de los datos ya filtrados. El siguiente paso en el proceso es construir una base de datos sencilla, que almacene toda esta información de forma tabular.

Apéndice 3. **Manejo de los datos y su almacenamiento**

Este apéndice muestra cómo se llevó a cabo el manejo de los datos extraídos de la red social.

Limpieza del contenido de los comentarios

Hay cosas que pueden descartarse de los comentarios porque no representan cambios en la polaridad de este tras realizar el análisis sentimental. En la figura 8 se muestra la función en código Python encargada de esta tarea.

Figura 8. **Limpieza de los comentarios**

```
37      def clean_tweet(self, tweet):
38          '''
39          Funcion que limpia los tweets eliminando enlaces y caracteres especiales.
40          '''
41          return  ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)",
42                                  " ", tweet).split())
```

Fuente: elaboración propia.

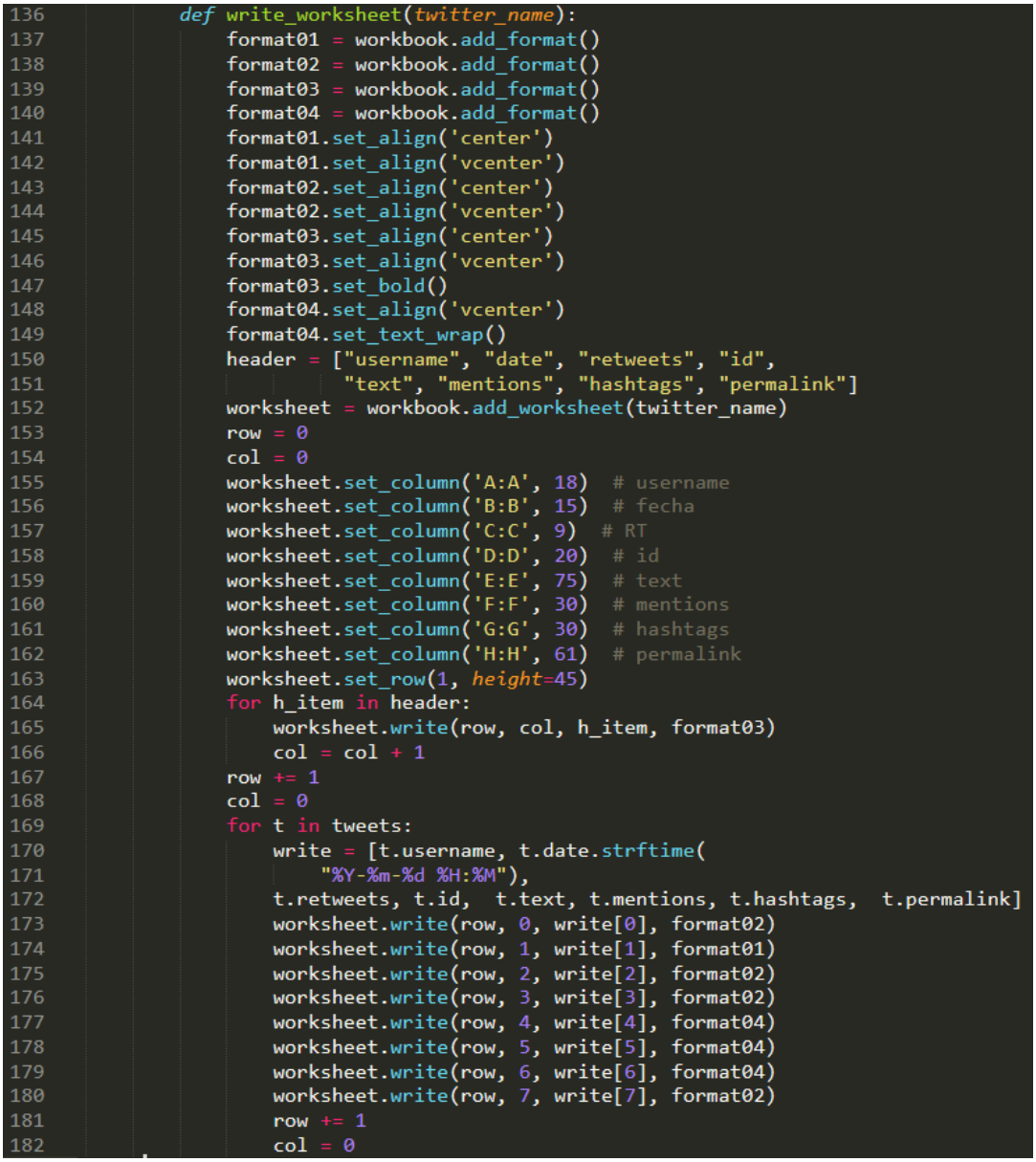
Esta limpieza se realiza utilizando expresiones regulares de los enlaces y funciones propias de las cadenas de Python.

Escritura del archivo xlsx

La escritura del archivo xlsx se realizó mediante la utilización de la librería xlswriter, en este se encuentra organizada la información para su posterior análisis, la función en lenguaje Python que permite lo anteriormente descrito se muestra en la figura 9.

Continuación apéndice 3.

Figura 9. **Manejo de los datos en un archivo xlsx**



```
136         def write_worksheet(twitter_name):
137             format01 = workbook.add_format()
138             format02 = workbook.add_format()
139             format03 = workbook.add_format()
140             format04 = workbook.add_format()
141             format01.set_align('center')
142             format01.set_align('vcenter')
143             format02.set_align('center')
144             format02.set_align('vcenter')
145             format03.set_align('center')
146             format03.set_align('vcenter')
147             format03.set_bold()
148             format04.set_align('vcenter')
149             format04.set_text_wrap()
150             header = ["username", "date", "retweets", "id",
151                       "text", "mentions", "hashtags", "permalink"]
152             worksheet = workbook.add_worksheet(twitter_name)
153             row = 0
154             col = 0
155             worksheet.set_column('A:A', 18)  # username
156             worksheet.set_column('B:B', 15)  # fecha
157             worksheet.set_column('C:C', 9)  # RT
158             worksheet.set_column('D:D', 20)  # id
159             worksheet.set_column('E:E', 75)  # text
160             worksheet.set_column('F:F', 30)  # mentions
161             worksheet.set_column('G:G', 30)  # hashtags
162             worksheet.set_column('H:H', 61)  # permalink
163             worksheet.set_row(1, height=45)
164             for h_item in header:
165                 worksheet.write(row, col, h_item, format03)
166                 col = col + 1
167             row += 1
168             col = 0
169             for t in tweets:
170                 write = [t.username, t.date.strftime(
171                     "%Y-%m-%d %H:%M"),
172                 t.retweets, t.id,  t.text, t.mentions, t.hashtags,  t.permalink]
173                 worksheet.write(row, 0, write[0], format02)
174                 worksheet.write(row, 1, write[1], format01)
175                 worksheet.write(row, 2, write[2], format02)
176                 worksheet.write(row, 3, write[3], format02)
177                 worksheet.write(row, 4, write[4], format04)
178                 worksheet.write(row, 5, write[5], format04)
179                 worksheet.write(row, 6, write[6], format04)
180                 worksheet.write(row, 7, write[7], format02)
181                 row += 1
182                 col = 0
```

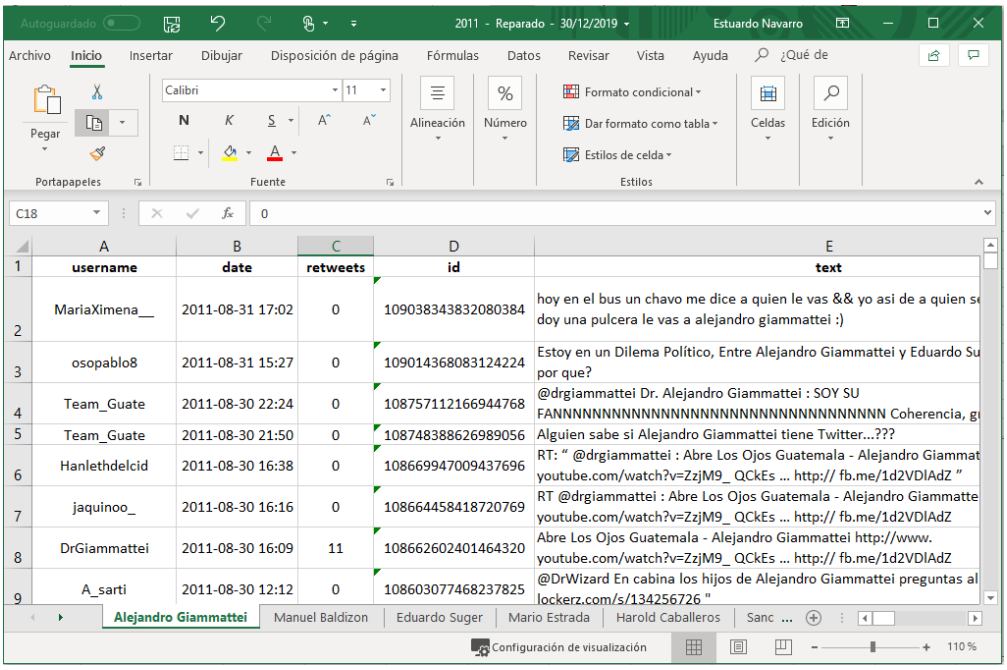
Fuente: elaboración propia.

El código en la figura 9, en esencia crea una nueva hoja en el archivo xlsx porque cada hoja representará a cada candidato, se da formato a cada columna, se definen los encabezados y posteriormente se escriben los datos, el análisis se realiza únicamente sobre el contenido de los comentarios pero para posibles referencias futuras, se almacena el usuario que lo escribió, así como las menciones y etiquetas del Tweet, esto permite que el análisis sentimental sea escalable y más detallado.

Visualización de los datos

En la figura 10 se muestra como luce la fuente de datos ya almacenada en un archivo xlsx.

Figura 10. **Datos almacenados en un archivo xlsx**



	A	B	C	D	E
1	username	date	retweets	id	text
2	MariaXimena__	2011-08-31 17:02	0	109038343832080384	hoy en el bus un chavo me dice a quien le vas && yo asi de a quien se doy una pulcera le vas a alejandro giammattei :)
3	osopablo8	2011-08-31 15:27	0	109014368083124224	Estoy en un Dilema Político, Entre Alejandro Giammattei y Eduardo Su por que?
4	Team_Guate	2011-08-30 22:24	0	108757112166944768	@drgiammattei Dr. Alejandro Giammattei : SOY SU FANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN Coherencia, g
5	Team_Guate	2011-08-30 21:50	0	108748388626989056	Alguien sabe si Alejandro Giammattei tiene Twitter...???
6	Hanlethdelcid	2011-08-30 16:38	0	108669947009437696	RT: " @drgiammattei : Abre Los Ojos Guatemala - Alejandro Giammat youtube.com/watch?v=ZzjM9_ QCkEs ... http:// fb.me/1d2VDlAdZ "
7	jaquinoo_	2011-08-30 16:16	0	108664458418720769	RT @drgiammattei : Abre Los Ojos Guatemala - Alejandro Giammatte youtube.com/watch?v=ZzjM9_ QCkEs ... http:// fb.me/1d2VDlAdZ
8	DrGiammattei	2011-08-30 16:09	11	108662602401464320	Abre Los Ojos Guatemala - Alejandro Giammattei http://www. youtube.com/watch?v=ZzjM9_ QCkEs ... http:// fb.me/1d2VDlAdZ
9	A_sarti	2011-08-30 12:12	0	108603077468237825	@DrWizard En cabina los hijos de Alejandro Giammattei preguntas al lockerz.com/s/134256726 "

Fuente: elaboración propia.

Apéndice 4. **Utilización de librería TextBlob para análisis sentimental**

Para el análisis sentimental se requiere leer el archivo que contiene las entradas, para eso se utilizará la librería xlrd, el análisis se basará en la polaridad, por cada polaridad positiva se suma un punto, de lo contrario si es negativa se resta un punto, en caso sea neutra simplemente no se suma ni se resta, esta funcionalidad se logra con el código mostrado en la figura 11.

Figura 11. **Análisis sentimental usando TextBlob**

```
# Se importan las librerias necesarias para leer archivos xlsx
import xlrd, re
# Se importa libreria para analisis sentimental
from textblob import TextBlob

def get_tweet_sentiment(tweet):
    analysis = TextBlob(tweet)
    # se evalua la polaridad del sentimiento
    if analysis.sentiment.polarity > 0:
        return 1
    elif analysis.sentiment.polarity == 0:
        return 0
    else:
        return -1
# Almacena la direccion del archivo por analizar
loc = ("2011.xlsx")
# Esta variable llevará control del punteo acumulado
punteo=0
# Se abre el archivo
wb = xlrd.open_workbook(loc)
# Se itera sobre todas las hojas del archivo
for x in xrange(0,len(wb.sheets())):
    # Cada hoja representa un candidato nuevo, el punteo se reinicia
    punteo=0
    sheet = wb.sheet_by_index(x)
    print('Punteo para: '+sheet.name+' es: ')
    # Se itera sobre todas las filas para sumar punteo
    for i in xrange(0,sheet.nrows):
        punteo=punteo+get_tweet_sentiment(sheet.cell_value(i, 4))
    print(punteo)
```

Fuente: elaboración propia.

Posterior a ejecutar este *script* es posible tabular los datos obtenidos para representarlos gráficamente y luego realizar el análisis de regresión.