Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Análisis y Visualización de Datos Masivos

Análisis del Sentimiento Político en Twitter durante las Elecciones Congresales 2020 en el Perú.

Trabajo Fin de Máster

Tipo de trabajo:

Presentado por: Alva Segura, Daniel

Director: Baldiris Navarro, Silvia Margarita

Resumen

En este TFM se efectúa la investigación para el desarrollo de un sistema en análisis de sentimiento político referente a las elecciones congresales del 2020 en Perú. El análisis de sentimiento es un área del procesamiento del lenguaje natural teniendo como objetivo el tratamiento de textos de los cuales se extrae un sentimiento u emoción, la cual puede ser positiva, negativa o neutral. El surgimiento de las redes sociales como Twitter ha sido aprovechada por la inteligencia artificial y el Big data para el tratamiento de información con respecto a los tuits.

El objetivo principal de este sistema es realizar un análisis de sentimiento de las elecciones Congresales del 2020, con las herramientas tecnológicas actuales del mercado que, a su vez, solucionen problemas y tareas dentro del análisis de sentimiento.

En el primer objetivo se realizó la investigación de las herramientas actuales en el mercado que solucionan el problema del análisis de sentimiento en este ámbito y un estudio acerca de las tareas del análisis de sentimiento.

El segundo objetivo es extraer los datos de Twitter de octubre 2019 a enero 2020 usando las palabras clave del nombre del partido político y que el tuit sea en español, utilizando técnicas del preprocesamiento de datos para limpiar los tuits.

Se han creado y optimizado modelos como Naive Bayes, Maquina vectores de soporte (SVM) y redes neuronales convolucionales (RNC) utilizando la metodología Cross Industry Standard Process for Data Mining (CRISP-DM) para analizar el sentimiento de dichos datos a nivel de documento, donde se puso en funcionamiento el clasificador que tuvo la mejor predicción de cada una de las clases mencionadas, utilizando un conjunto de corpus que se ajustan a este estudio.

Sobre el marco del desarrollo del sistema, se implementó una arquitectura full stack (MEAN), utilizando Node para el back-end, MongoDB para el almacenamiento de los tuits, y finalmente, se han utilizado un conjunto de graficas por medio de Angular para cada partido político, con la finalidad de mejorar el entendimiento de la información extraída.

Palabras clave: Análisis de Sentimiento. NLP, Aprendizaje supervisado, RNC.

Abstract

In this TFM, research is carried out to the development of a system analysis of political sentiment regarding the congressional elections to 2020 in Peru. Sentiment analysis is part of data processing of natural language, having as objective, the processing texts. Once these texts are extracted, you can get out of them some feeling or emotion; they can be positive, negative or neutral. The appearance of social networks such as Twitter has been used by the artificial intelligence and Big data to processing information regarding tweets.

The main objective in this system is to carry out a sentiment analysis of 2020 Congressional elections, with the technological tools that there are in the market and at the same time to solve problems and tasks within the sentiment analysis.

In the first objective, the research was carried out about the actual tools in the market that solve the sentiment analysis problems, and to study the sentiment analysis task.

The second objective is to extract the Twitter data from October 2019 to January 2020, using the name or keywords of the political party. The tweets needed to be on Spanish. We would use pre-processing techniques to clean the tweets.

The models such as Naive Bayes, Support Vector Machine (SVM) and Convolutional Neural Networks (RNC) have been created and optimized using methods such as the Cross Industry Standard Process for Data Mining (CRISP-DM) to analyze the sentiment from those data at the document level. Here, we put into operation the classifier that has the best prediction to each one of the mentioned classes, using a set of corpus that are adjusted to this study.

About the framework to the system development, it was implemented full stack architecture (MEAN), using Node to the back-end, MongoDB for the storage of tweets, and finally, a set of graphs have been used through Angular for each political party, in order to improve the understanding to the information extracted.

Keywords: Sentiment Analysis, NLP, Supervised Learning, RNC.

Índice de contenidos

Índice de tablas

Índice de figuras

1. Introducción

Los procesos electorales en el Perú han generado espacios de dialogo y debate, ya que en estos se produce una mayor interacción entre candidatos y ciudadanos. En ellos se promueve unas series de estrategias que se enfoquen al desarrollo de un país, las cuales tienen un efecto en el ciudadano teniendo una reacción ante determinadas propuestas u posiciones políticas.

Es por ello, que el aumento de las redes sociales, particularmente Twitter se está estableciendo en una práctica común, especialmente en el tiempo de elecciones, pero hasta el momento no hay un estudio que pueda aprovechar esta forma de opinión en las elecciones congresales en el Perú, convirtiendo estos datos en conocimiento.

En lo que respecta, el análisis de sentimientos es la extracción automatizada de actitudes, opiniones y emociones de fuentes de texto, discurso base de datos. El análisis de sentimientos implica clasificar las opiniones en el texto en categorías como "positivo" o "negativo" o "neutral" (Rogalski, 2019). En este tipo de análisis se puede evidenciar las tendencias políticas positivas o negativas de la población hacia ciertos partidos políticos. Twitter es una red social que cuenta con 290 millones de cuentas a nivel mundial durante el año 2019. (Fernández, 2021) y en Perú solo en el 2015 había un aproximado de 4 millones de cuentas. (Vilca & Vilca, 2020).

Este trabajo de fin de master (TFM) describe el proceso en el desarrollo de una herramienta que analice el sentimiento político en las elecciones congresales de 2020, y cómo se llevó a cabo dicho proceso. Para este desarrollo se han usado modelos de aprendizaje supervisado para predecir los sentimientos.

Se han utilizado y comparado muchas herramientas y Interfaz de programación de aplicaciones (APIS) disponibles en el mercado. Así mismo se han usado un conjunto de tecnologías MEAN, que hace uso de MongoDB, Express y Node para el lado del servidor; y Angular para el lado del cliente para la aplicación del análisis del sentimiento.

1.1. Motivación

El crecimiento del uso de las redes sociales en el Perú en los últimos años ha generado más fuentes de información y opinión, dado la facilidad de acceso e interacción desde nuestras computadoras, tabletas y teléfonos; esto ha causado que se realicen más estudios sobre ellas. También se han convertido en un canal importante para que los políticos se dirijan a la población, haciéndolos más accesibles a sus posibles votantes.

Los métodos de análisis de opinión ayudan a clasificar y comprender los sentimientos de los usuarios sobre un tema de interés. Sin embargo, la gran complejidad de los sistemas sociotécnicos y las características de big data de las redes complejas hacen que el análisis de los eventos de las redes sociales sea una tarea difícil. En este contexto, los estudios de caso de campañas políticas del mundo real son de particular interés porque ayudan a comprender el comportamiento humano, detectar patrones e identificar enfoques genéricos para analizar el comportamiento del usuario en las redes sociales en línea. También los políticos necesitan conocer la opinión de los votantes y el efecto que producen sus campañas en el electorado. Mucha gente emite su voto revisando las opiniones de otros acerca de los candidatos en las redes sociales.

El análisis de sentimiento político está siendo una herramienta de apoyo y en un futuro remplazara a las encuestas tradicionales. Trabajando los datos con más rapidez y menor costo.

Es importante saber que paradigmas, algoritmos y tecnologías utilizar para resolver el problema del análisis de sentimiento, en especial las opiniones más subjetivas y complejas. Se han desarrollado trabajos sobre el análisis de sentimiento en idioma inglés, pero pocos trabajos han sido desarrollados para medir el sentimiento en idioma español. Desde nuestro conocimiento no hay investigaciones para analizar el sentimiento en la política peruana por medio de Twitter y dado que hay un proceso de elecciones congresales 2020, se pretende aprovechar la información generada en esta red social como fuente de información para un trabajo en esta línea.

1.2. Justificación

La red social Twitter ha sido objeto de investigación para el análisis de sentimiento en los últimos años dado que es usada para la comunicación en cuanto a la opinión de valoraciones, actitudes y emociones sobre un tema en concreto, expresando de esta manera los diferentes tipos de opiniones. Cabe destacar, que por medio de la red social se permite publicar

mensajes de texto tomando en cuenta que el Twitter es una red que genera cantidad de datos y mensajes que se vinculan a un evento directo en cualquier parte del mundo. A pesar que las fuentes comunes tales como ; las encuestas electorales o sondeo de opinión siguen vigentes, la llegada del internet y el constante uso de las redes sociales han producido cambios en la forma de comunicarse de las personas (Vásquez Torres & Joyanes Aguilar, 2018). El problema a tratar es la falta de mecanismos de soporte a los candidatos electorales en cuanto a las opiniones de los ciudadanos con respecto a sus partidos políticos, permita mostrar que las redes sociales no solo influyen en la difusión de una determinada campaña electoral, sino también pueden aportar una mejor relación con el proceso político electoral.

Es importante destacar que existen muchas formas de resolver el problema del análisis de sentimiento usando herramientas de procesamiento del lenguaje natural (NLP) como SentiWordNet (http://swn.isti.cnr.it/) y CoreNLP (https://corenlp.run/).De acuerdo con Bin Liu, el análisis de sentimiento ofrece una excelente plataforma para todos los investigadores y de este modo puedan realizar procesos tangibles y centrados de las investigaciones, y que a su vez sea más manejables de lograr grandes avances y poder resolver un problema en el análisis de sentimientos sin tener que cambiar el tema o área de la investigación (Liu, 2015)
En los últimos años han aumentado la valoración de opiniones a través de las redes sociales y esto se debe a la mayor demanda de estas herramientas que simplifican los procesos tradicionales, tales como, las encuestas electorales o sondeo de opinión. Según tendencias digitales “La inclusión de internet en el Perú el 2017 fue de 48%, conllevando significativamente el aumento en el uso de este sistema como medio de información. De todas las personas que usan las redes sociales en el Perú, el 55% publica comentarios” (Vásquez Torres & Joyanes Aguilar, 2018). También en los últimos años la mayoría de los candidatos al congreso en estas elecciones tiene una cuenta en Twitter y la usa de forma regular.
Por último, es conveniente notar, que el interés del análisis de sentimiento se ha intensificado en los últimos años. En la Figura 1.1 muestra la frecuencia de la búsqueda del Sentiment Analisys a nivel mundial desde el año 2010 a la actualidad.

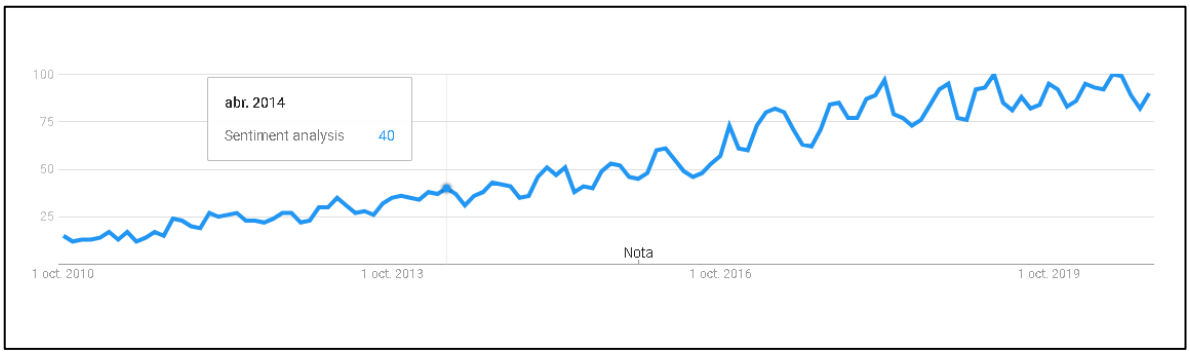

Figura 1.1. Datos relacionados con la búsqueda de keywords - Sentiment Analysis
Fuente: Google Trends (2020).

1.3. Planteamiento del trabajo

Este trabajo de fin de estudios propone el desarrollo en un sistema basado en el análisis de sentimientos de usuarios de Twitter, acerca de las elecciones congresales del año 2020 en Perú. Los datos obtenidos han seguido una serie de tareas indicadas según el desarrollo de este TFM. Se han utilizado varios corpus del taller de análisis semántico de la Sociedad Española para el Procesamiento del Lenguaje Natural SEPLN(TASS). Este es un taller de evaluación experimental dentro de la SEPLN para promover la investigación dentro del área del análisis de sentimientos en las redes sociales, específicamente enfocado en el idioma español. Este corpus está basado en opiniones cortas de texto (Twitter), publicado por personalidades representativas. (Villena Roman, Lana Serrano, Martínez Cámara, & González Cristóbal, 2013). Además, se usó un corpus basado en los partidos políticos peruanos, debido a la poca información en español para la solución de este problema. Este corpus ha sido etiquetado según las reglas de las oraciones y la expresión de los sentimientos del tuit. También ha sido revisado por un especialista en lingüística para aumentar su precisión.

Ante lo expuesto, se utiliza la metodología CRISP-DM para resolver este estudio, utilizando técnicas de preprocesamiento de datos y desarrollo de algoritmos de aprendizaje supervisado, quien, a su vez, es aplicado al corpus para su debida evaluación y de esta manera obtener resultados más concretos y precisos del sentimiento hacia un partido político en específico. Se ha seleccionado el modelo que ha obtenido el resultado más relevante en las métricas de evaluación.

Por otra parte, se busca las palabras relacionadas a un partido político o candidato representativo considerando los retuits ya que el usuario que lee el tuit podrá expresar en este mismo su opinión en cuanto a la publicación. Asimismo, los datos filtrados de la API de Twitter se van almacenando en MongoDB mediante Phyton. Considerando lo antes planteado, la predicción es aplicada por medio del mejor modelo entrenado.

Finalmente, el propósito de este trabajo es construir una herramienta de visualización que permita medir el sentimiento de cada partido político que son expresados por medio de lo tuits, utilizando la arquitectura MongoDB, Express, Angular, y Node (MEAN).

1.4. Estructura del Trabajo

En el siguiente apartado se resume cada uno de los capítulos desarrollados en este TFM.

En el segundo capítulo se describe la problemática de la política actual de la Republica del Perú, su historia en los últimos años y como está estructurado el estado peruano en especial el congreso de la república. Por otra parte, se describe el cierre del congreso en el año 2019, forzando a elecciones congresales extraordinaria y los partidos políticos a participar en las elecciones congresales 2020.

En el tercer capítulo, se describirá lo que es el análisis de sentimientos, las tareas y el nivel del sentimiento, el problema de los lexicones, el uso de herramientas o API, así como también el uso de herramientas comerciales que podrían solucionar el problema del sentimiento político en el Perú. También se describirán los casos en que se usó el análisis de sentimiento político en diversas campañas de elecciones, y el uso masivo de las redes sociales en Latinoamérica y en Perú.

En el cuarto capítulo, se describe el objetivo general y los específicos, la metodología (CRISP-DM), que es usada con frecuencia para la analítica de texto, igualmente se describirá las tareas más importantes a seguir en el análisis de sentimiento.

En el quinto capítulo, se describe el desarrollo de la contribución, siguiendo las tareas descritas en el capítulo anterior, se almacenarán y preprocesarán los datos con las reglas más comunes del NLP, se creará y entrenará un modelo óptimo para la predicción de los resultados por medio de pruebas realizadas a un conjunto de corpus TASS y un corpus en base a los partidos políticos y se mostrará los resultados de los datos trabajados en un dashboard personalizado, por otra parte, se describe el desarrollo de la herramienta.

En el sexto capitulo, se describe las conclusiones basadas en los objetivos planteados y se realiza las recomendaciones para futuros trabajos.

2. Análisis del Contexto

En esta sección se presenta la problemática política del estado peruano en los últimos años, del cual mencionamos los acontecimientos más importantes. También se menciona la estructura del sistema político en el Perú. Además del manejo de las redes sociales en Latinoamérica.

2.1. Situación política en el Perú

En este apartado se hará una reseña breve de la situación política del Perú desde 1990 hasta la actualidad.

2.1.1. Orígenes de la Corrupción 1990-2016

Según estudios sociales la corrupción tiene efecto negativo sobre la estabilidad institucional, la inversión y en consecuencia sobre el crecimiento económico de un determinado país.

En el Perú esta crisis empezó desde la década de 1990 con la elección del presidente Alberto Fujimori y ha ido ahondándose en los gobiernos posteriores, donde se llegó a privatizar empresas claves, se ha malversado donaciones extranjeras, se sobornó a jueces, además unas cuantas compañías privadas (club de la construcción) y extranjeras se beneficiaron con los fondos y gastos públicos del estado peruano (Quiroz, 2019).
Los tentáculos del Fujimorismo (aliados de Fujimori) se han mantenido hasta la actualidad en el control del congreso de la república, el cual representa el poder ejecutivo. El poder Judicial también cayó bajo la corrupción aceptando pagos y coimas (Quiroz, 2019).
En el año 2009, la Corte Suprema de Justicia del Perú condenó a Alberto Fujimori a 25 años de pena privativa de la libertad por delitos de secuestro agravado, lecciones graves y homicidio calificado (Naupari, 2018).

2.1.2. Orígenes de corrupción 2016 a la actualidad

En el año 2016 salió electo el presidente Pedro pablo Kuczynski y con ello las esperanzas de seguir con la reforma del país. El mayor número de congresistas fue ocupado por la bancada Fujimorista llamada ahora Fuerza popular con un total 73 de 130 escaños los cuales fueron elegidos mediante el proceso electoral (Pierina, 2016).

A 20 meses de su mandato, el presidente Kuczynski renuncio debido a que el poder legislativo encabezado por la oposición Fujimorista estuviera cerca de aprobar su dimisión debido al caso Odebrecht y a la publicación de videos y audios por la compra de votos. Esta lucha por

el poder absoluto entre el poder legislativo y el ejecutivo había saldado con la vacancia presidencial (BBC Mundo, 2018).

En marzo del 2018 asumió el mando a la presidencia del Perú, el vicepresidente de la republica Martin Vizcarra, teniendo como objetivo darle a Perú un nuevo comienzo y con una propuesta destinada a superar la crisis de la corrupción, reactivación económica y reformas institucionales. Muchas de sus acciones en contra de la corrupción han tenido la oposición de algunos congresistas (Carvallo, 2019).

2.1.3. Cierre del congreso

En el año 2017 Keiko Fujimori, hija del expresidente Alberto Fujimori decidido a través de su bancada vacar al presidente Kuczynski, fomentando que el denominado Club de la construcción, haciendo uso de toda su maquinaria empresarial y política liquidar todo el poder político del fujimorismo. El escándalo de la corrupción de Odebrecht puso en peligro al empresariado ligado a este club por la red de corrupción, sobornos y sobre costos de las obras (Calderón, 2019).

La guerra entre el presidente Vizcarra y el fujimorismo fue por el control del poder económico, el problema principal del fujimorismo es que estaban desprestigiados ante la mayoría de la población. (Calderón, 2019).

El día 30 de septiembre del 2019, el presidente Martin Vizcarra, determino disolver el parlamento, ante la negativa de la oposición fujimorista de aplicar una reforma de nombramiento de magistrados del tribunal constitucional (Fowks, 2019).

El caso de corrupción de Odebrecht ha salpicado a la mayoría de expresidentes de la república del Perú como consecuencia un expresidente se suicidó, otros dos están con arresto domiciliario. Keiko Fujimori está en prisión, así como sus congresistas destituidos, y la corrupción ha estado a punto de alcanzar al presidente Vizcarra por la cuestionada concesión del aeropuerto de Chincheros (Calderón, 2019).

2.1.4. Elecciones parlamentarias 2020

A pocas horas de disolver el congreso el presidente Martin Vizcarra por medio de una ordenanza de urgencia apela a elecciones para el 26 de enero de 2020. (El Comercio, 2019).

Hay 24 partidos habilitados por el Jurado Nacional de elecciones, de los cuales los más relevantes son (Chillitupa, 2019):

- Fuerza Popular (FP) cuya cabeza de lista es Marta Chávez,
- El Partido Aprista (APRA) cuya cabeza de lista es Mauricio Mulder.
- El partido Morado (PM) cuya cabeza de lista es Francisco Sagasti.

- El partido popular cristiano (PPC) cuya cabeza de lista es Alberto Beingolea.
- Alianza por el Progreso (APP) cuya cabeza de lista es Omar Chehade.
- Podemos Perú (PP) cuya cabeza de lista es Daniel Belizario Urresti Elera.
- Frente Amplio (FA) cuya cabeza de lista es Carlos Enrique Fernández Chacón.

2.2. Sistema político en el Perú

En este apartado se presenta un resumen del sistema político peruano, que servirán de guía para este trabajo.

2.2.1. Estructura del Sistema Político

El Sistema político peruano está dividido en el poder legislativo, el poder ejecutivo, el poder judicial, el gobierno regional, el gobierno local y organizaciones constitucionales autónomos.

El Poder Ejecutivo

El Poder Ejecutivo en el Perú está dirigido por el presidente de la Republica, quien tiene las funciones de Jefe de Estado, representa los intereses de la nación y al mismo tiempo orienta la política estatal de los diversos ministerios que integran el Ejecutivo. Es elegido a través del voto popular.

Encargado de planificar y poner en curso los proyectos programados de desarrollo nacional, a través de proyectos de desarrollo en áreas como en lo económico, la salud, la educación, entre otros. También tiene la facultad de dictar normas, decretos y garantiza el cumplimiento de las leyes y normativas promulgadas por el Congreso (Gobierno del Peru, 2018).

El Poder Legislativo

Este organismo es una asamblea deliberativa y está representado por el Congreso de la República que tiene la exclusiva autoridad de crear leyes, además de administrar el presupuesto del estado y tiene como tres facultades constitucionales: legislar, controlar y representar dentro y fuera del país, así como de ejercer el control político de la nación; todo estos a través de las normativas establecidas en la Constitución vigente. Es la representación de la democracia ya que este está formado por representas elegidos por el pueblo. (Gobierno del Peru, 2018).

El poder judicial

Es un cuerpo del Estado responsable en desempeñar y proveer justicia en la nación de acuerdo con la Constitución y las leyes, avalando la protección de las pertenencias y haciendo justicia en las personas.

En cuanto a la Corte Suprema de Justicia este es un órgano que se encarga de actuar en todo el territorio nacional, juntamente con la Cortes Superiores con un alcance en los distritos judiciales, juzgados de primera instancia y seguidamente, los juzgados de paz letrados y no letrados.
Posteriormente, resuelven litigios, protegen los derechos de los ciudadanos y se hace cumplir las responsabilidades innatas de cada uno de ellos (Gobierno del Peru, 2018).

El Gobierno Regional

Los Gobiernos Regionales administran a cada uno de los departamentos en cuanto a su autonomía política, económica y administrativa en todas las entidades públicas con el fin de enfocar aquellos asuntos que le competan, a ellos, en un marco de estado unitario y descentralizado. Son responsables de la administración de cada departamento según su jurisdicción, con autonomía política, económica y administrativa en las instituciones públicas y de esta manera encaminar las dificultades y logros que le competen, en marco a un estado democrático, unitario y descentralizado. Así mismo, se dividen en dos órganos: Consejo Regional y Gobernador Regional (antes de 2015 se usó el término de Presidente Regional) (Gobierno del Peru, 2018).

El Gobierno Local

Las Municipalidades son los entes reguladores de la administración pública local encargada de las gestiones de desarrollo y regularización de normas y leyes ante una cantidad limitada de ciudadanos en un territorio determinado esto es a nivel provincial y distrital, por lo que se dividen en Municipalidades Provinciales y Municipalidades Distritales. Los alcaldes y regidores son elegidos mediante voto popular por un periodo de 4 años(Gobierno del Peru, 2018).

En la Figura 2.1 se muestra el último organigrama de la estructura del gobierno del Perú.

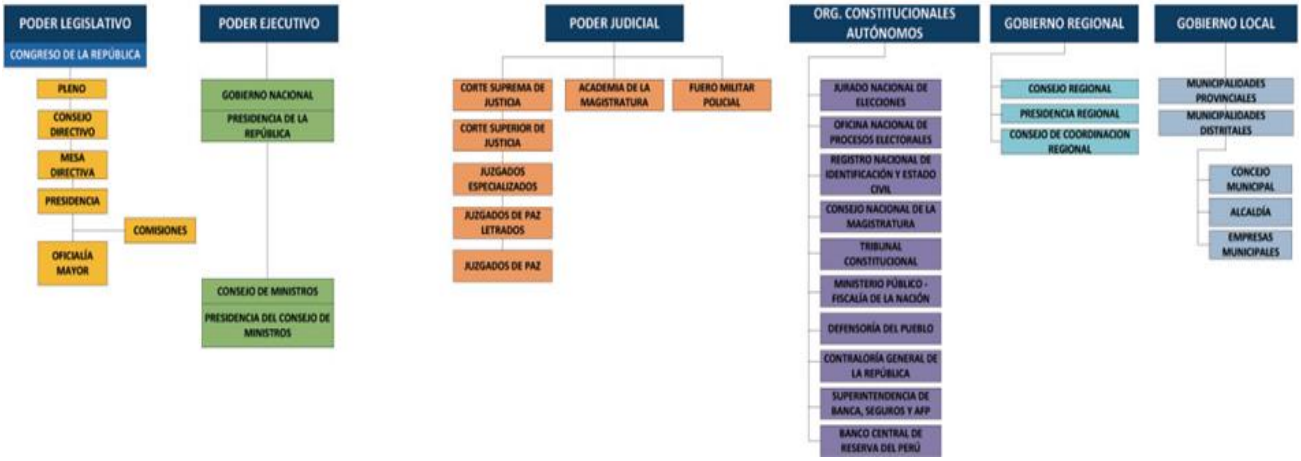

Figura 2.1. Organigrama del estado peruano
Fuente: Secretaría de Gestion Publica. (2014).

2.2.2. Estructura del congreso de la Republica

El Congreso de la República representa al Poder Legislativo; tiene la función de desarrollar la mayor parte del marco legal bajo el cual se rige la administración pública y el sistema de justicia. También tiene otras funciones (Congreso de la Republica, s.f.).

- Función la representación de la nación
- Función la dación de leyes.
- Función de fiscalización y control político
- Función legislativa.
- Función eventual de la reforma de la constitución
- Funciones especiales.

La estructura del congreso de la república está conformada por los siguientes órganos de gobierno legislativo:

- **El Pleno:** Está formado por representantes llamados parlamentarios, siendo estos estos la máxima asamblea deliberativa, encargada de debatir y tomar decisiones que vaya de acuerdo con la constitución y las normas legales.

- **El Consejo Directivo:** Se conforma por representantes de grupos parlamentarios que se denominan en: directivos o portavoces y miembros de la mesa directiva.

- **Mesa Directiva.** Es conformada por el presidente y tres vicepresidentes, que a su vez pueden presidir y debates del pleno, de la comisión permanente y dirigir los consejos directivos.

- **La Presidencia.** Es un cargo electo que se ejerce por un periodo legislativo (1 año), en el cual, a su vez, se realiza el nombramiento del presidente y los integrantes de la mesa directiva, con la finalidad de fijar las funciones y atribuciones del presidente.

- **Las Comisiones.** Sobre este particular, son un grupo de parlamentarios con un determinado propósito o función a realizar, cuyo trabajo principal consiste en el seguimiento de obras y desarrollo de una fiscalización de los órganos estatales de los sectores que componen la administración pública.

2.3. La importancia de la información en las redes sociales

El servicio de Internet ha provocado un avance en la forma de comunicación de las personas, eliminando en gran medida las diferencias geográficas existentes. Esto ha aumentado la velocidad de la información, así como también los modos de comunicación. Lo cual ha allanado el camino para una penetración cultural entre comunidades que están muy separadas geográficamente. Por lo tanto, ha permitido que las personas se sensibilicen sobre temas ajenos a su cultura como las elecciones presidenciales de EE. UU, o u otros temas ajenos a su localidad. Por consiguiente, este fenómeno de sensibilización compartida permite a las personas exhibir emociones de apoyo, empatía, odio y agresiones a través de las redes sociales. Las redes sociales almacenan cantidades de datos textuales, debido a debates y discusiones entre personas pertenecientes a diferentes culturas, religión, estratos sociales y raciales puede utilizarse para encuestas de opinión exhaustivas motivadas para una gran variedad de propósitos (Cambria, Das, Bandyopadhyay, & Feraco, 2017).

En América Latina, una región en proceso de crecimiento tecnológico está descubriendo las funcionalidades de las redes sociales y está convirtiéndose en la región más comprometida en el uso de estas a nivel mundial.
Según el informe presentado el año 2017 por la consultora comScore (https://www.comscore.com/), Latinoamérica se está convirtiendo en la región que usa mayoritariamente las redes sociales a nivel. Las redes de mayor uso son WhatsApp, Facebook, Twitter, LinkedIn, YouTube hasta Pinterest, están influenciando en los usuarios y las empresas de Latinoamérica, que usen estas plataformas como medio de comunicación y también para exponer sus servicios y productos a la población, creando de este modo una

dependencia de las personas a esta forma de comunicación (Vásquez & Joyanes ,2018). En la Figura 2.2 muestra el tiempo de uso de las redes sociales por regiones en un mes.

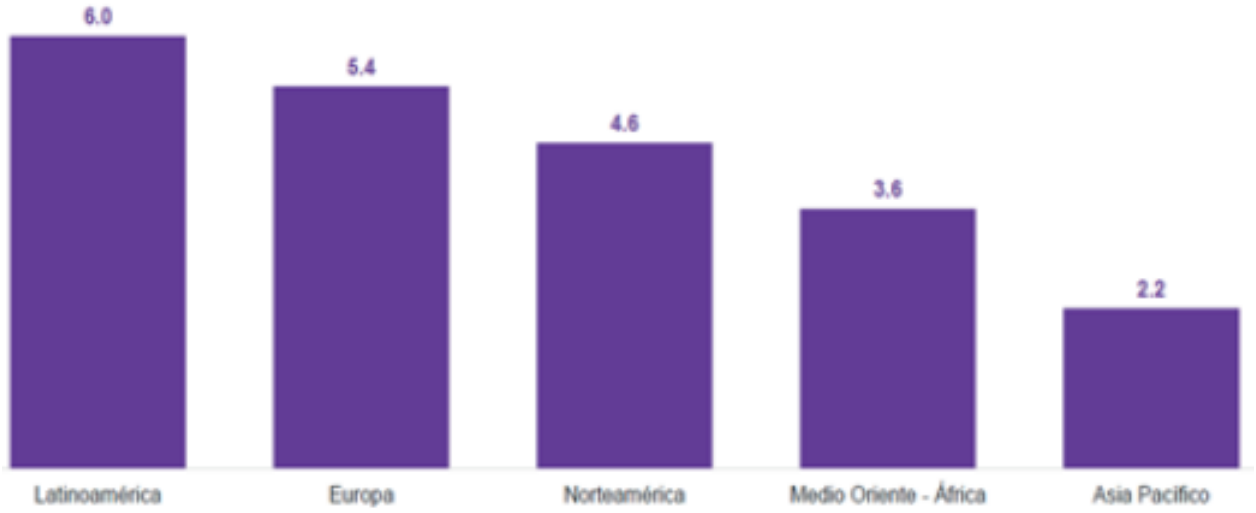

Figura 2.2. Promedio de horas en redes sociales al mes de personas por región
Fuente: Vásquez,J & Joyanes ,L (2018)

Por último, es conveniente notar un estudio completo del panorama de las redes sociales en Latinoamérica, se puede palpar un aumento de estas en todos los niveles: usuario, educativo, político y de empresa. Debe señalarse que la llegada de las redes sociales a Latinoamérica ha sido contundente, pasando de ser el segundo consumidor de Social Media, hasta alcanzar el mayor promedio de uso en horas a nivel global.

3. Estado del Arte

En este capítulo se recoge el estado del arte relativo al análisis de sentimiento, incluyendo las aplicaciones que ya se han desarrollado, así mismo las herramientas, servicios y páginas que abordan este problema.

3.1. Análisis de Sentimientos

En este apartado se presenta un resumen del análisis de sentimientos, sus problemas, las soluciones utilizadas en el mercado que se han utilizado para abordar este problema y las librerías mayormente utilizadas a nivel de programación que trabajan para su solución.

3.2. Definición, antecedentes y retos

El análisis de sentimientos es la disciplina que analiza todas las emociones que un individuo puede expresar, tales como: sentimientos, opiniones, actitudes y valoraciones hacia ciertas entidades y atributos expresados en un texto. Dichas entidades pueden ser: productos o servicios, individuos u organizaciones, eventos y temas o problemas.

Desde el año 2002, la investigación en análisis de sentimientos ha ido en auge, debido de la mayor disponibilidad de un gran volumen de datos de opinión en las redes sociales. Por lo tanto, la industria y las aplicaciones en torno al análisis de sentimientos han crecido desde el año 2006. Por otro lado, el análisis de sentimientos también ofrece numerosos problemas.

Las redes sociales nos han permitido estudiar parte de los temas y opiniones de las personas en particular, mediante un perfil de opinión de cada usuario en las diferentes plataformas de acuerdo a sus necesidades e intereses y opiniones actuales, expresados en sus publicaciones. Dicha información se puede utilizar en muchas aplicaciones, por ejemplo, recomendando productos servicios y determinando a qué candidatos políticos votar. Además, los participantes en las redes sociales no solo pueden publicar mensajes, sino también interactuar entre ellos a través de conversatorios y debates, que involucran sentimientos como el acuerdo y el desacuerdo (o contención). Por ejemplo, los temas sociales y políticos y las opiniones de posiciones opuestas pueden usarse para enmarcar temas políticos y predecir los resultados electorales (Liu, 2015).

“El campo del análisis de sentimientos es una nueva y emocionante dirección de investigación debido a la gran cantidad de aplicaciones del mundo real. Descubrir la opinión de las personas es algo muy relevante a la hora de tomar una decisión. El análisis de sentimientos es el estudio que analiza la opinión y el sentimiento en las personas, tales como; servicios, productos y

entidades, que se encuentran en un texto. Siempre ha sido importante saber lo que piensan los demás. Las personas utilizan sitios de revisión en línea, blogs, foros, sitios de redes sociales, entre otros, para expresar su opinión que aumenta los datos generados por los usuarios en la web. Por lo tanto, ha surgido la necesidad de analizar y comprender estos datos / revisiones generados en línea." (Agarwal & Mittal, 2016)

Es importante señalar que la diferencia entre opinión y sentimientos se trata de que el sentimiento es una actitud, pensamiento o juicio provocado por el mismo, mientras que la opinión es un juicio o una valoración formada en la mente sobre un asunto en particular (Pozzi, Fersini, Messina, & Liu, 2017). En la Tabla 3.1 se especifican los componentes de la opinión.

Tabla 3.1. Componentes de la opinión

Componente	Descripción
e	es la entidad objetivo
a	es el aspecto objetivo de la entidad e sobre el cual se ha emitido la opinión
s	s es el sentimiento de la opinión sobre el aspecto a de la entidad e
h	es el titular de la opinión
t	es el tiempo de publicación de la opinión

Fuente: Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017).

Tomemos un ejemplo que aplica estas características el cual fue posteado por big Jhon el 15 de septiembre del 2011 el cual describe de la siguiente manera "(1) Compré una cámara Samsung y mi amigo trajo una cámara Canon ayer. (2) En la semana pasada, ambos usamos mucho las cámaras. (3) Las fotos de mi Samy no están claras para las tomas nocturnas, y la duración de la batería es corta también. (4) Mi amigo estaba complacido con su cámara y le encanta su calidad de imagen. (5) Quiero una cámara que tenga buena resolución en la toma fotos. (6) Voy a devolverlo mañana." (Cambria, Das, Bandyopadhyay, & Feraco, 2017).

La primera tarea debe extraer las expresiones de entidad (e), Samsung, Samy y Canon. La segunda tarea debe extraer expresiones de aspecto(a) imagen, foto y duración de la batería. La tercera tarea debe encontrar al titular de las opiniones el cual es big John (el autor del blog) y que el titular de las opiniones en la frase (4) es el amigo de big John. La cuarta tarea debe encontrar el momento en que se publicó el blog (15/09/2011). La quinta tarea debe encontrar que la frase (3) da una opinión negativa a la calidad de imagen de la cámara Samsung y una opinión también a su duración de la batería. La frase (4) da una opinión positiva a la cámara Canon en su conjunto y también a su calidad de imagen. La frase (5) aparentemente expresa una opinión positiva, pero no lo hace. (Cambria, Das, Bandyopadhyay, & Feraco, 2017).

3.3. Tareas del Análisis de Sentimiento

El proceso del análisis de sentimiento implica el desarrollo de algunas tareas las cuales son descritas a continuación (Pozzi, Fersini, Messina, & Liu, 2017).

3.3.1. Primera tarea: distinción entre tipo de oraciones.

La primera tarea cuando se trata del análisis de sentimientos generalmente consiste en distinguir entre oraciones subjetivas y objetivas. Un ejemplo de una oración objetiva es "El Samsung Galaxy es un teléfono inteligente", mientras que un ejemplo de una oración subjetiva es " Samsung Galaxy es increíble". La clasificación de polaridad es la tarea que distingue las oraciones que expresan polaridades positivas, negativas o neutrales. Tenga en cuenta que una oración subjetiva puede no expresar ningún sentimiento positivo o negativo (por ejemplo, "supongo que ha llegado"). Por esta razón, debe clasificarse como "neutral".

3.3.2. Segunda tarea: Detección de la polaridad

Está dirigida a detectar texto positivo, negativo o neutral. Los mensajes de redes sociales son uno de los tipos de texto más difíciles de manejar. Esta complejidad se debe principalmente a las siguientes características:

- **Mensajes cortos:** Los mensajes de redes sociales suelen ser muy cortos, pero ricos en semántica incorporada

- **Contenido ruidoso:** un aspecto adicional que debe modelarse explícitamente cuando se trata de análisis de sentimientos en redes sociales se relaciona con textos mal formados, donde el vocabulario, la ortografía y la sintaxis representan un desafío lingüístico. Los mensajes de redes sociales se caracterizan por expresiones coloquiales, abreviaturas, emoticones, alargamiento de palabras, mayúsculas irregulares y expresiones enfáticas, y generalmente no se ajustan a las reglas gramaticales canónicas.

Con respecto al lenguaje utilizado en las redes sociales, el esfuerzo principal de la comunidad de análisis de sentimientos se ha dedicado a capturar y modelar la expresión típica en la red a través de texto, etiquetas de parte del discurso (por ejemplo, adverbios y adjetivos) y contenido paralingüístico (emojis, jerga, hashtags) para obtener modelos de predicción más efectivos. Para la solución de esta tarea se utilizan algoritmos de aprendizaje supervisado y no supervisado (Pozzi, Fersini, Messina, & Liu, 2017).

3.3.3. Tercera tarea: Resumen de opinión

Esta tarea consiste en el resumen de opinión basado en aspectos tiene dos características principales. Primero, captura la esencia de las opiniones: objetivos de opinión (entidades y sus aspectos) y sentimientos sobre ellos. Segundo, es cuantitativo, lo que significa que da el número o porcentaje de personas que tienen opiniones positivas o negativas sobre las entidades y los aspectos. El lado cuantitativo es crucial debido a la naturaleza subjetiva de las opiniones. El resumen de opinión resultante es una forma de resumen estructurado producido a partir del quíntuple de opinión como muestra en la Figura 3.1.

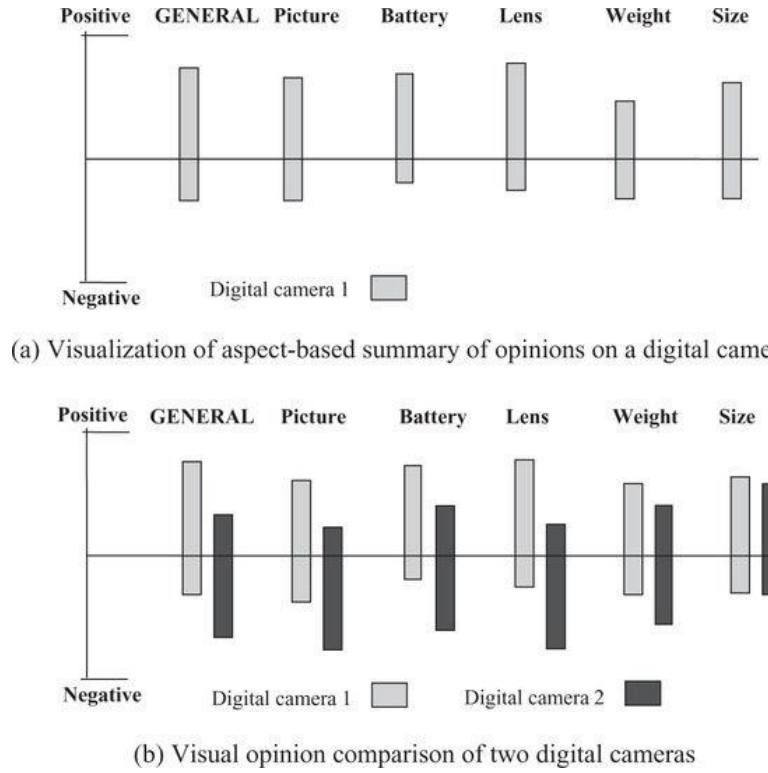

Figura 3.1. Comparación basada en aspectos de una cámara
Fuente: Pozzi, F. A., Fersini, E., Messina, E., & Liu, B (2017)

3.3.4. Cuarta tarea: Técnicas de Visualización

La cuarta tarea consiste en utilizar las técnicas de visualización para ayudar al usuario a explorar una gran cantidad de información. Las técnicas de visualización de datos aprovechan nuestra capacidad de procesamiento de información visual al crear representaciones visuales del gran conjunto de datos. Entre ellos, se propuso un modelo más recientemente, y está diseñado teniendo en cuenta las características únicas de las conversaciones en línea. Por

ejemplo, la Figura 3.2 muestra el análisis de sentimiento acerca del discurso de Obama y la cantidad de tuits analizados.

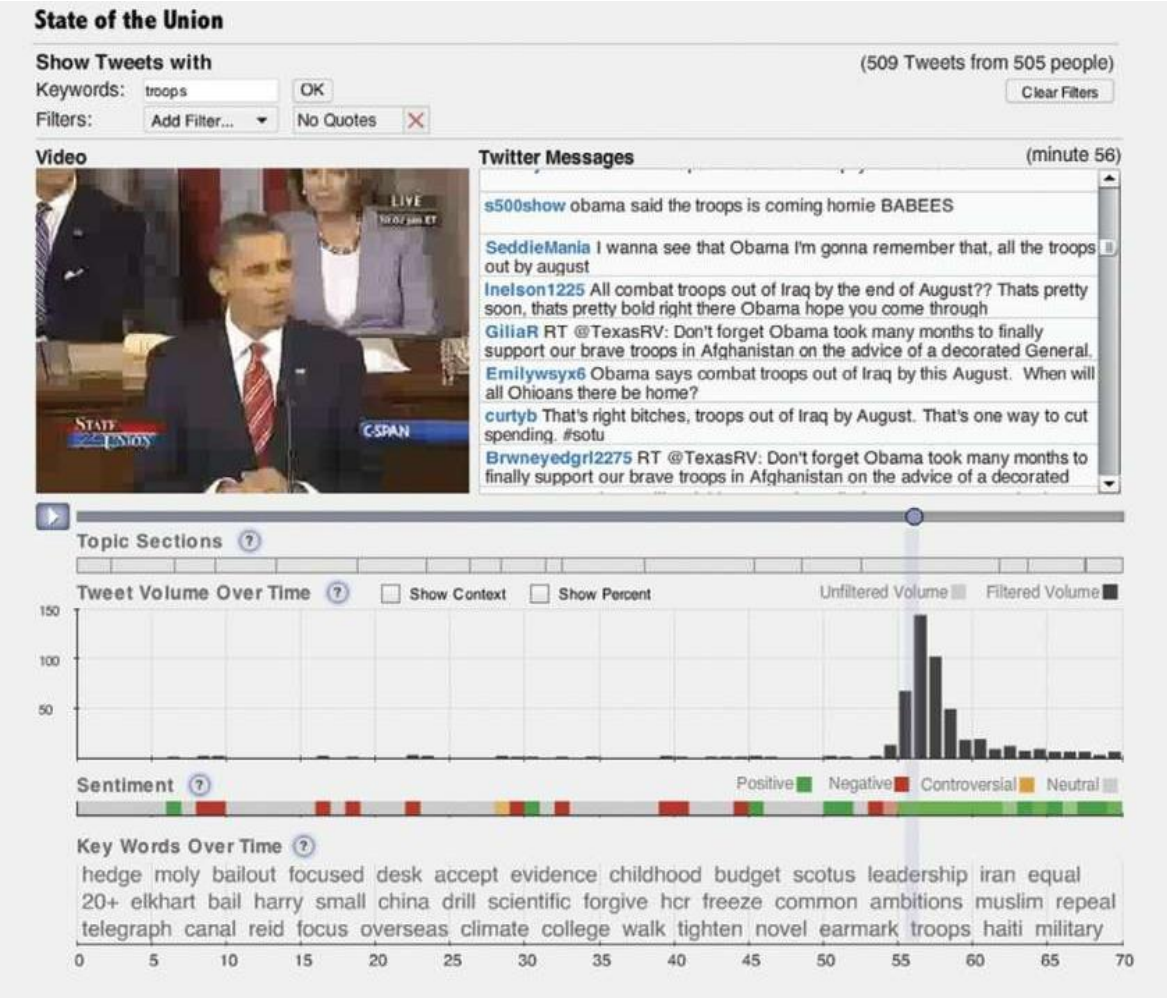

Figura 3.2. Visualización de los sentimientos del discurso de Obama

Fuente: Pozzi, F. A., Fersini, E., Messina, E., & Liu, B (2017)

3.3.5. Quinta tarea: Detección de la ironía

La quinta tarea consiste en la detección de ironía y sarcasmo es descubrir características que nos permitan discriminar textos irónicos (o sarcásticos) de textos no irónicos (o no sarcásticos). El interés en la detección de ironía y sarcasmo en las redes sociales requiere que tengamos datos generados por los usuarios que nos permitan capturar el uso real de dispositivos de lenguaje figurativo de este tipo. Como en la mayoría de las tareas del NLP, la falta de corpus es un problema. Existen dos enfoques principales para la construcción de cuerpos irónicos / sarcásticos: autoetiquetado y crowdsourcing. El primero considera como instancias positivas aquellos textos en los que la autora señala su intención utilizando una etiqueta explícita (por ejemplo, los hashtags #irony y #sarcasm). El crowdsourcing implica la

interacción humana mediante el etiquetado del contenido como irónico (o sarcástico). Principalmente, el proceso de etiquetado se realiza sin una definición o directriz estricta. Por lo tanto, es una tarea subjetiva, donde el acuerdo entre los anotadores es a menudo muy bajo. De esta manera, es posible obtener textos irónicos y sarcásticos potenciales producidos por personas en las redes sociales. Como clasificadores, la regresión logística y las SVM han sido las más utilizadas para la detección del sarcasmo. Los enfoques recientes sobre la detección de sarcasmos consideran la información más allá del texto mismo, explotando la información contextual y la información sobre el usuario (Pozzi, Fersini, Messina, & Liu, 2017).

3.3.6. Sexta tarea: Detección de spam.

Otra Tarea del análisis de sentimiento es la detección de spam. Una característica de las redes sociales es que permite a las personas de cualquier parte del mundo expresar libremente sus puntos de vista y opiniones sin revelar su verdadera identidad y sin temor a consecuencias indeseables. Estas opiniones son por lo tanto muy valiosas. Sin embargo, este anonimato tiene un precio. Facilita que las personas con intenciones maliciosas publiquen opiniones falsas para promover o desacreditar algunos productos, servicios, organizaciones o individuos objetivos sin revelar sus verdaderas intenciones. Estas personas se denominan spammers de opinión y su actividad se denomina spam de opinión. El spam de opinión se ha convertido en un tema resaltante en los medios sociales. Es importante detectar tales actividades para garantizar que las opiniones en la web sean fuentes confiables de información valiosa. A diferencia de la extracción de opiniones positivas y negativas, la detección de spam de opinión no es solo un problema del NLP, sino también un problema de minería de datos, ya que implica analizar los comportamientos de publicación de los revisores (Liu, 2015).

En la Figura 3.3 se presentan graficas estas tareas según han sido identificadas por Pozzi, F. A., Fersini, E., Messina, E., & Liu, B.

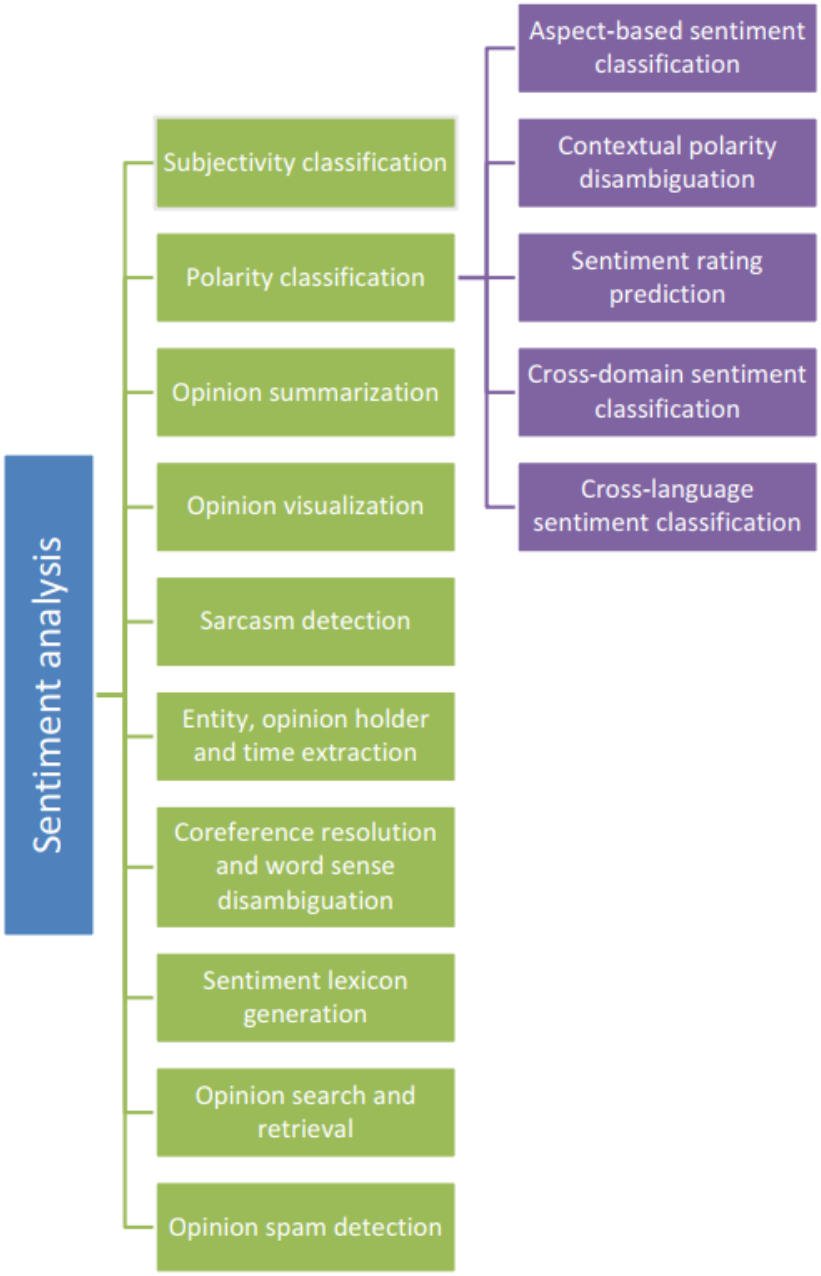

Figura 3.3. Tareas del análisis de sentimiento

Fuente: Pozzi, F. A., Fersini, E., Messina, E., & Liu, B (2017)

3.4. Niveles del Análisis de Sentimiento

La investigación de análisis de sentimientos se ha llevado a cabo principalmente en tres niveles: nivel de documento, nivel de oración y nivel de aspecto.

Nivel de documento La tarea a nivel de documento es clasificar que en un documento exprese la opinión completa si el sentimiento que persiste en el mismo es de manera positiva o negativa, así mismo, se conoce como clasificación de sentimientos a nivel de documento. Por ejemplo, al analizar un producto el algoritmo determinara, si dicha opinión se expresa de manera positiva o negativa hacia el mismo. En todo caso, se presume que, en cada documento sobre una entidad, se pueden expresar opiniones. Cabe destacar, que no es adaptable a documentos que evalúan o comparan entidades múltiples, para las cuales se necesita un análisis más detallado.

Nivel de oración. El siguiente nivel es determinar si cada oración expresa una opinión positiva, negativa o neutral, teniendo en cuenta que "opinión neutral" generalmente significa "sin opinión". En esta perspectiva, el nivel de análisis está estrechamente relacionado con la clasificación de subjetividad identificando oraciones que expresan información objetiva (oraciones objetivas) de oraciones que expresan puntos de vista subjetivos y opiniones (oraciones subjetivas). Sin embargo, la subjetividad no es equivalente al sentimiento o la opinión muchas oraciones objetivas pueden implicar sentimientos u opiniones, por ejemplo, "Compramos el auto y se ha malogrado los frenos". Por el contrario, muchas oraciones subjetivas no pueden expresar ninguna opinión o sentimiento.

Nivel de aspecto. Ni los análisis a nivel de documento ni a nivel de oración descubren exactamente qué le gusta y qué no le gusta a la gente. En otras palabras, no dicen de qué trata cada opinión, es decir, el objetivo de la opinión, una oración puede tener múltiples opiniones, por ejemplo, "a Apple le está yendo muy bien en este país subdesarrollado". No tiene mucho sentido clasificar esta oración como positiva o negativa porque es positiva sobre Apple, pero negativo sobre el país. Para obtener este nivel de resultados detallados, debemos pasar al nivel de aspecto. En lugar de mirar las unidades del lenguaje (documentos, párrafos, oraciones, cláusulas o frases), el análisis de nivel de aspecto mira directamente la opinión y su objetivo (llamado objetivo de opinión). Darnos cuenta de la importancia de los objetivos de opinión nos permite tener una mejor comprensión del problema del análisis de sentimientos. (Liu, 2015).

3.5. Lexicones y sus problemas

Los indicadores más importantes de los sentimientos son las palabras. Por ejemplo, bueno, maravilloso, asombroso y genial son palabras de sentimiento positivo, malo, pobre y espantoso son palabras de sentimiento negativo. Además de las palabras individuales, también hay frases y modismos. Las palabras y frases de opinión son fundamentales para el análisis de opinión. Una lista de estas se llama lexicón de sentimiento (o lexicón de opinión).

Aunque las palabras y frases de opinión son importantes, están lejos de ser suficientes para un análisis preciso de las opiniones, teniendo los siguientes problemas.

Una palabra de sentimiento positivo o negativo puede tener orientaciones o polaridades opuestas en diferentes dominios de aplicación o contextos de oración. Por orientación o polaridad, queremos decir si un sentimiento u opinión es positivo, negativo o neutral. Por lo tanto, decimos que las orientaciones de las palabras de sentimiento pueden ser dominio dependiente o incluso dependiente del contexto de la oración.

Una oración que contenga palabras de sentimiento no puede expresar ningún sentimiento, este fenómeno ocurre en varios tipos de oraciones. Las oraciones de preguntas (interrogativas) y las oraciones condicionales son dos tipos principales, por ejemplo, "¿Puedes decirme qué cámara Samsung es buena?" Y "Si puedo encontrar una buena cámara en la tienda, la compraré". Ambas oraciones contener la palabra de sentimiento buena, pero ninguna expresa una opinión positiva o negativa sobre ninguna cámara específica. Sin embargo, eso no quiere decir que todas las oraciones condicionales y las oraciones interrogativas no expresen ninguna opinión o sentimiento, por ejemplo, "¿Alguien sabe cómo reparar esta terrible impresora?" Y "Si está buscando un buen automóvil.

Las oraciones sarcásticas con o sin palabras de sentimiento u opinión son difíciles de manejar, por ejemplo, “¡Qué gran moto! Dejó de funcionar en tres días”. El sarcasmo no es tan común en las revisiones de los consumidores sobre productos y servicios, pero es común en las discusiones políticas, que dificultan el tratamiento de las opiniones políticas.

Cabe destacar, que muchas oraciones sin palabras de sentimiento pueden implicar sentimientos u opiniones positivas o negativos de sus autores. Por ejemplo, “Esta lavadora usa mucha agua” implica una opinión negativa sobre la lavadora porque usa muchos recursos (agua). Por otra parte, en la mayoría de estas oraciones son en realidad oraciones objetivas que expresan cierta información objetiva, por ejemplo, "Después de dormir en el colchón durante dos días, se ha formado un hueco en el medio" expresa una opinión negativa sobre la calidad del colchón, esta oración puede considerarse objetiva porque establece un hecho, aunque el valle se usa aquí como una metáfora. Como

podemos ver, estas dos oraciones no contienen palabras de sentimiento, pero ambas expresan algo indeseable, lo que indica opiniones negativas (Liu, 2015).

3.6. Aplicaciones

Las aplicaciones de análisis de sentimientos hoy en día se han extendido a varios campos, desde servicios y atención médica hasta servicios financieros, elecciones políticas, y exploración de fauna, para el propósito de este trabajo se detallan las elecciones norteamericanas del presidente Obama el año 2012 y las elecciones Austriacas el año 2016.

3.6.1. Obama y las elecciones norteamericanas 2012

La primera referencia del análisis del sentimiento Político fue realizada por el Candidato Obama a las elecciones presidenciales norteamericanas el año 2012. Obama decidió utilizar los servicios de la empresa Vertica (https://www.vertica.com/) con el fin de obtener mayor margen de competencia frente a sus contrincantes (Méndez, 2015).

El equipo de Obama en 18 meses unifico toda la información relevante para la campaña con los potenciales votantes, recopilando listas de los encuestadores y de los voluntarios y escaneando perfiles en las redes sociales extrayendo datos básicos como edad, sexo, raza, zona de residencia, nivel de ingreso, hasta la inclusión de datos de preferencias, consumo, y amistades. El equipo de Obama analizo cuatro flujos distintos de datos de los votantes indecisos de los estados donde Obama y Ronney (Candidato Opositor) estaban igualados. A partir de este análisis se descubrió que para ganar en el estado de Florida fue indispensable cautivar a la población femenina menor a los 35 años, se analizó las preferencias de este grupo y se descubrió que tenían ciertas características en común como la predilección por ciertas series de televisión o pertenecían a una red social determinada (Bayo, 2015).

3.6.2. Análisis político en Twitter durante las elecciones presidenciales norteamericanas 2016

El conflicto político entre Hillary Clinton y Donald Trump se reflejó en las discusiones entre los usuarios en las redes sociales como Twitter. Aunque los tuits de los candidatos pueden llegar a un gran número de usuarios, los debates disputados en Twitter y en otras redes sociales muestran que no todos los usuarios tienen el mismo sentimiento con respecto a los mensajes de los candidatos.

Se definieron las siguientes condiciones para la identificación de tuits políticos:

- El tuit es el retuit de un candidato
- El tuit apunta al menos a un candidato
- El tuit menciona al menos un candidato
- El tuit tiene el nombre propio de un candidato

Si una de estas condiciones se cumplió para un candidato, entonces ese tuit se consideró como un tuit político para ella o para él. Si alguna de estas condiciones se cumplía para ambos candidatos, entonces ese tuit se consideraba político para ambos.
Se definió tres clases de polaridad de sentimientos positivos, neutrales y negativos hacia ambos candidatos (Malik, Ayeena & Kapoor, Divya & Singh, Amit. ,2016).
Se recolectaron 4.9 millones de tuits publicados por 18,450 usuarios, sus perfiles y sus relaciones con otros usuarios. Un perfil de usuario de Twitter se compone de los siguientes atributos: nombre, descripción del perfil, foto y ubicación. Una línea de tiempo es el conjunto de tuits que publicó un usuario de Twitter. El candidato Donald Trump tiene una cuenta de Twitter identificada por el usuario @realDonaldTrump y en noviembre de 2016 tenía alrededor de 17,1 millones de seguidores. Por otro lado, la candidata Hillary Clinton, identificada por el usuario @HillaryClinton, fue seguida por aproximadamente 11,6 millones de usuarios en noviembre de 2016. Para abordar este problema, se dividió el análisis de sentimientos en dos enfoques. En el primer enfoque, realizamos un análisis de sentimiento considerando todo el texto del tuit. Por lo tanto, el sentimiento de texto se asignó a un candidato si y solo si el tuit era sobre ese candidato. el mismo enfoque para los tuits no políticos. En el segundo enfoque, identificaron las palabras relacionadas con cada candidato si y solo si el tuit era sobre ambos candidatos. Por lo tanto, realizaron un análisis de sentimientos considerando solo las palabras relacionadas con los candidatos, usando la herramienta Stanford Parser. (Caetano, J., Lima, H., Santos, M. *et al.*, 2018).

3.6.3. Análisis político en Twitter durante las elecciones Austriacas 2016

También se han hecho estudios sobre las elecciones presidenciales austriacas de 2016, utilizando ciencia de redes y análisis de sentimientos utilizando datos de Twitter. En las elecciones presidenciales del 2016, Austria ha sido testigo de las opiniones polarizadoras entre sus ciudadanos. El Partido de la Libertad de Austria con Norbert Hofer, y su candidato opositor, un ex miembro del Partido Verde, Alexander Van der Bellen, se enfrascaron en una lucha por la presidencia.

En este análisis, se encontró un patrón claro que muestra que los tuits emocionales (tanto negativos como positivos) se retuitean. En este proyecto se utilizó SentiStrength(https://nlp.stanford.edu/software/lex-parser.shtml), que se basa en un léxico de palabras sentimentales, una lista de expresiones idiomáticas y una lista de emoticones. En particular, SentiStrength se ocupa de la corrección ortográfica, las palabras de refuerzo, la negación y la puntuación repetida para asignar finalmente dos puntuaciones para cada tuit individual.

Al aplicar técnicas de NLP, se encontró evidencia de que los candidatos usaron campañas negativas para obtener más seguidores. En particular, este estudio distingue entre información errónea e información negativa, y encontramos que la información negativa en estas elecciones en particular obtuvo un alto número de retuits y likes. Sin embargo, también se verifico que la propagación de información errónea puede tener efectos negativos en el candidato.

Con respecto al comportamiento de tuiteo de los seguidores de los candidatos, se utilizó técnicas de análisis de sentimientos y análisis de redes para construir comunidades de los candidatos, y examinamos qué temas los usuarios de las redes sociales asocian con cada candidato. Estos resultados muestran una clara distinción en cómo los usuarios perciben a ambos candidatos. En particular, se descubrió que los hashtags con una connotación negativa se han asociado predominantemente con el candidato Norbert Hofer. Este fenómeno se puede observar en las redes de hashtag de idioma inglés como en alemán (Kušen & Strembeck, 2017).

3.6.4. Análisis político en Twitter durante las elecciones de Uruguay 2019

En América latina, el caso más conocido es de las elecciones nacionales de Uruguay realizada por la editorial El País, juntamente con IGV (http://www.igv.com.uy/). La cual es una empresa tecnológica dedicada a soluciones de hardware y software. El País e IGV publicaron un algoritmo de inteligencia artificial (IA) para analizar datos del discurso de más de 30.000 usuarios que opinaron en Twitter sobre las elecciones. Debido a la gran cantidad de tuits a procesar se automatizo el proceso. El proceso constó de tres pasos:

- Recolección
- Clasificación
- Visualización

Los resultados de este análisis se almacenaron en tiempo real. Para la clasificación de los tuits utilizaron la tecnología Watson de IBM como herramienta de TI. Para el almacenamiento y rapidez grafico de los datos al instante utilizaron la aplicación de la nube de IBM (IBM Cloud). Se analizaron segundo a segundo un total de 61.632 tuits desde las 00:00 h hasta las 23:59 h del domingo 27 de octubre del 2019. Usaron las frases de búsqueda los términos y frases que hacen referencia a cada candidato (Verdier & Rocha, 2019).

3.7. Aplicaciones y herramientas

En este apartado se mencionarán las herramientas existentes actuales en el mercado, así como las librerías que usan para la programación.

3.7.1. Aplicaciones comerciales

El blog de Brand24 hizo una evaluación de las 17 mejores herramientas existentes en el mercado (Rogalski, 2019), de las cuales para este estudio solo mencionaremos las que pueden usarse en el sentimiento político.

- **Brand24:** Es una aplicación de pago que hace un monitoreo de las redes sociales más utilizadas como Twitter, YouTube, foros, noticias. Tiene la opción de polaridad de los sentimientos., además de mostrar las menciones en escala de tiempo e incluso una demo. Por otra parte, está disponible en más de 24 idiomas incluido el inglés y el español, y su vez permite guardar los resultados de sus consultas (https://brand24.com/).

  Esta página también tiene la opción para visualizar los comentarios más populares y una búsqueda de palabras dentro del resultado de los posts.

  Para probar esta página se utilizó el nombre de Martin Vizcarra como se muestra en la Figura 3.4 para la investigación, el cual arroja un resultado neutral, también indica un cuadro con la cantidad de post en las redes sociales analizadas, además de tener un tablero de conteo de las opiniones de sentimiento positivo y negativo y su porcentaje en relación con los posts analizados, así también el detalle de cada opinión con su respectivo sentimiento. En la página se muestra una representación temporal en el intervalo de 1 mes de los tuits acerca del presidente Martin Vizcarra.

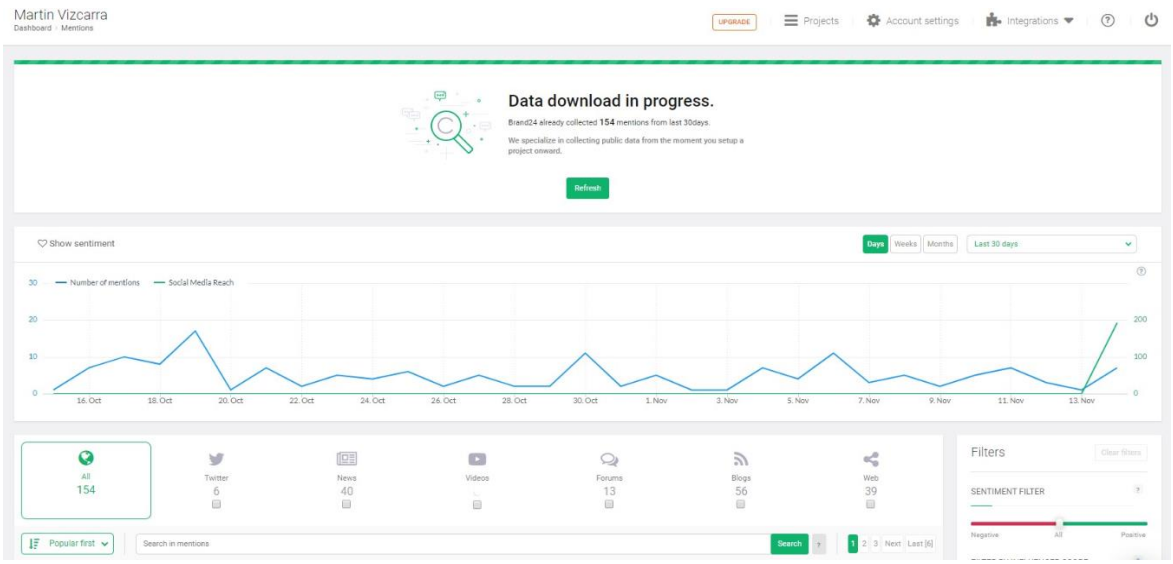

Figura 3.4. Análisis de Sentimientos en Brand24
Fuente: Página web de Brand24 https://brand24.com/

- **Social Mención:** Es una página gratuita que también hace un monitoreo de las redes sociales de forma gratuita, agregando el contenido del posteo de los usuarios en toda la web. Tiene la opción de filtro por Blog, Microblog y preguntas, además soporta 15 idiomas como el inglés y el español. Así mismo, todas las menciones están marcadas con puntos verdes, rojos o grises, y no guarda el resultado de las consultas.
  Para probar esta página se utilizó el nombre de Martin Vizcarra para el análisis, el cual arrojó un resultado neutral, también hay un tablero que indica la intensidad del sentimiento, además de tener un gráfico de palabras más utilizadas en las opiniones de sentimiento, así como un gráfico de los usuarios con mayor número de posteo en los que referencian al presidente, la cual se muestra en la Figura 3.5 (http://www.socialmention.com/).

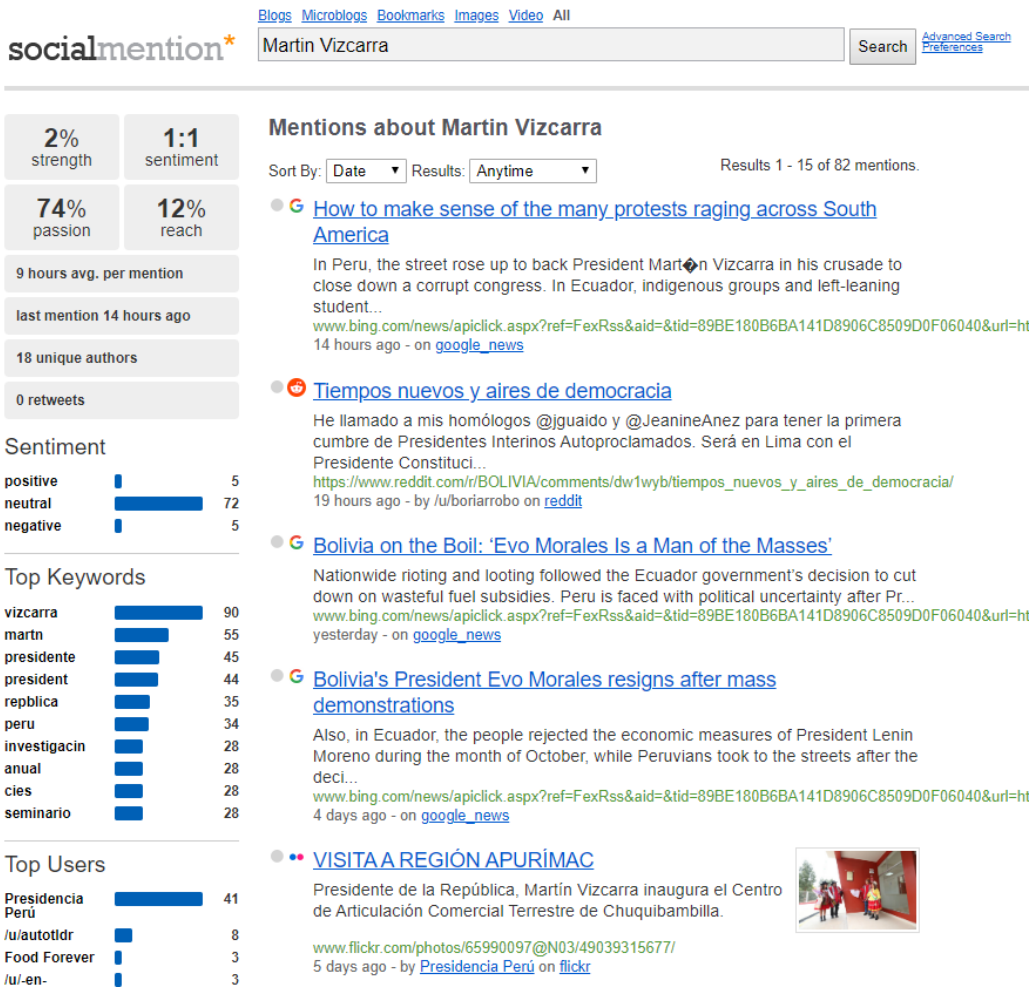

Figura 3.5. Análisis de Sentimientos en Social Mención

Fuente: Página web de Social Mención: http://www.socialmention.com/

- **Social Searcher:** Es una página que tiene varias herramientas como Social Buzz, Google Social Search, Media Monitoring, entre otras. En cuanto al análisis de sentimientos muestra en colores los sentimientos respectivos y puede aplicar filtros de opinión para que solo pueda ver un conjunto particular. Además, puedes ver la línea de tiempo de los posts generados, además de determinar la polaridad de los sentimientos por cada red social y una ratio de todas las redes analizadas, así como una tabla con la cantidad de los usuarios. Otra característica que tiene es que puedes ver la cantidad de post por red social y las palabras clave más utilizadas. Está disponible en dos maneras; gratuito y de pago(https://www.social-searcher.com/).

  Para probar esta página se utilizó el nombre de Martin Vizcarra para el análisis, el cual arrojo un resultado de 16 tuits positivos, 27 negativos,214 neutrales, sumando un total 257 posts y 124 usuarios involucrados, además de visualizar el porcentaje de la polaridad por cada red social, como se visualiza en la Figura 3.6.

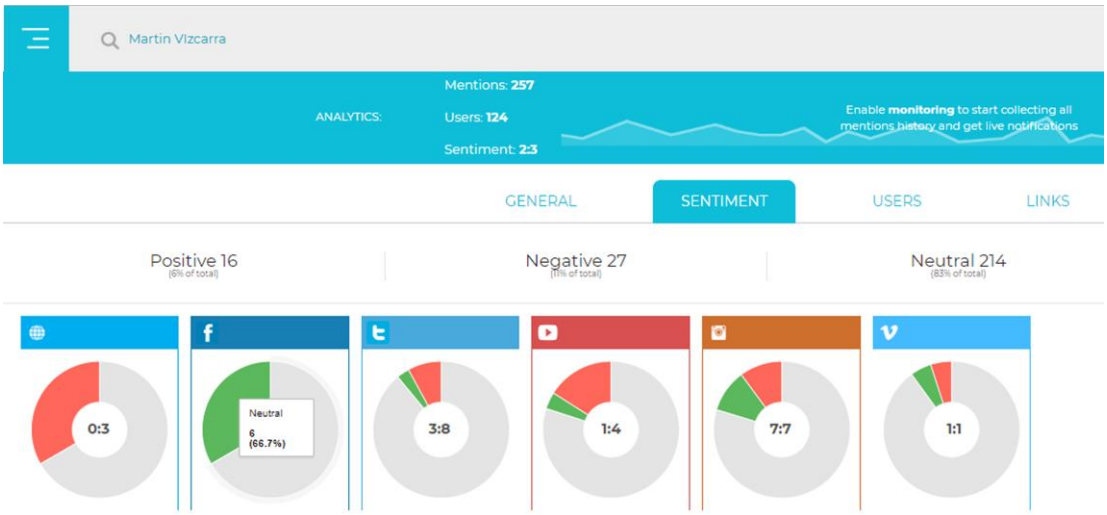

Figura 3.6. Análisis de Sentimientos en Social Search
Fuente: Página web de Social Search: https://www.social-searcher.com/

- **Tweet Sentiment Visualization**: Es una página de gratuita que analiza los tuits de la palabra seleccionada, además analiza mapas de calor, línea de tiempo, nube de palabras. Esta herramienta está desarrollada en Java (https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/).

  Para probar esta página, se utilizó el nombre de Martin Vizcarra para el análisis, el cual arrojó un resultado, los tuits desagradables se dibujan como círculos azules a la izquierda y los tuits agradables como círculos verdes a la derecha. Así como el resultado de la nube de palabras como se visualiza en la Figura 3.7.

Figura 3.7. Análisis de Sentimientos en Tweet Sentiment Visualization
Fuente: Página web de sentiment viz:
https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

- **Talkwalker's Quick Search:** Es un motor de búsqueda que hace un monitoreo de las redes sociales, que te ayuda a encontrar ideas de contenido y descubrir tendencias. Muestra las tendencias por genero por edad y por los idiomas utilizados, además el número de los usuarios con mayor participación (https://www.talkwalker.com/quick-search-form/) .

En la Tabla 3.2. se muestra una comparación de las aplicaciones comerciales analizadas.

Tabla 3.2. Comparativa de Soluciones

Herramienta	Soporta Idiomas (Español)	Twitter	Otras Redes	Polaridad Sentimiento en escala	Gratuito	Algoritmo utilizado	Guarda el Resultado de las consultas
Brand 24	Si	Si	Si	Si	No	Caja Negra	Si
Social Mencion	Si	Si	Si	Si	Si	Caja Negra	No
Social Searcher	Si	Si	Si	Si	No	Caja Negra	Si
Tweet Sentiment Visualization	Si	Si	No	Si	Si	Caja Negra	No
Talkwalker's Quick Search	Si	Si	Si	Si	No	Caja Negra	Si

Fuente: Elaboración Propia

3.3.2 APIs y librerías

En cuanto a las herramientas o APIs o librerías utilizadas para el análisis de sentimiento hay muchas que son ampliamente utilizadas, de las cuales las más relevantes para este estudio se describen a continuación:

Scikit-Learn:

Es una librería de software libre que usa Python para el uso de operaciones de Big data como el NLP, aprendizaje supervisado, aprendizaje no supervisado, preprocesamiento, selección y evaluación del modelo, inspección, visualizaciones y utilidades de carga de conjuntos de datos(https://scikit-learn.org/).

Textblob:

Es una librería de Python para el NLP y para la extracción de frases nominales, tiene las siguientes características(https://textblob.readthedocs.io/en/dev/).

- Análisis de los sentimientos.
- Clasificación (Naive Bayes, árbol de decisión).
- Traducción y detección de idiomas con tecnología de Google Translate.
- Tokenización.
- Frecuencias de palabras y frases.
- Inflexión de palabras (pluralización y singularización) y lematización.
- Corrección ortográfica.
- Agregar nuevos modelos o idiomas a través de extensiones.
- Integración con WordNet.

Intellexer:

Es una API que permite a los desarrolladores integrar productos semánticos en aplicaciones de consumo o empresariales o servicios web utilizando XML o JSON. También incluye soluciones del NLP para análisis de sentimientos, reconocimiento de entidades con nombre, resumen de múltiples documentos, extracción de texto, comparación de documentos, detección de idioma, corrección ortográfica y más. Funciona con cualquier componente de software que emita solicitudes HTTP y utiliza de un conjunto de reglas lingüísticas, estadísticas y semánticas complejas para la predicción del sentimiento. (https://www.intellexer.com/sentiment_analyzer.html) (Berlind, 2016).

SentimentAnalyzer:

Semantic Analyzer es una biblioteca de código abierto para averiguar lo que utiliza el análisis de texto sin procesar en busca de pistas sobre sentimientos positivos o negativos. Esta biblioteca devuelve una probabilidad de sentimiento y una puntuación junto con los indicadores de predicción (verdadero para positivo, falso para negativo). Desarrollado por ML.NET(https://www.nuget.org/packages/SentimentAnalyzer/) (con .NET Standard 2.0).

Repustate:

Es una API que realiza análisis de sentimientos en tiempo real. Las respuestas (response) de la API están en formato JSON, que analizan la petición (request) de un fragmento de un texto. También tiene lo opción de analizar una gran cantidad de documentos de texto a través de su API masiva. Las puntaciones del sentimiento varían de -1 (negativo) a 1(positivo) con un puntaje de 0 (neutral). Es aplicable a nivel semántico y se puede configurar añadiendo reglas

y clasificaciones. También soporta múltiples lenguajes como el inglés y español. Utiliza algoritmos de aprendizaje automático, los cuales son reentrenados constantemente. (https://www.repustate.com/sentiment-analysis/).

Microsoft Text Analytics:

Está basado en un servicio en la nube, el cual puede proporcionar funciones de procesamiento de lenguaje natural (NLP), que a su vez incluyen; análisis de sentimientos, minería de opiniones, la extracción de frases claves, detección de idiomas y reconocimiento en las entidades con su respectivo nombre. La API emplea un algoritmo de clasificación en aprendizaje automático para generar una valoración de sentimiento entre 0 y 1. El modelo utilizado está pre-entrenado con un corpus de texto y asociaciones de sentimientos. Utiliza una combinación de técnicas para el análisis, que incluye procesamiento de texto, análisis de parte del discurso, colocación de palabras y asociaciones de palabras (https://azure.microsoft.com/es-es/services/cognitive-services/text-analytics/). Esta API forma parte de Azure Cognitive Services, que es un conjunto de algoritmos de aprendizaje automático y de inteligencia artificial en la nube.

Pyspark:

Apache Spark está escrito en lenguaje de programación Scala, para admitir Python con Spark, la comunidad de Apache Spark lanzó una herramienta, Pyspark, también puede trabajar con RDD en el lenguaje de programación Python. Es gracias a una biblioteca llamada Py4j que pueden lograr esto. A través de este conjunto de librerías se pueden realizar uso de operaciones de Big data como el NLP. Por otra parte, es compatible con un conjunto amplio de herramientas de un alto nivel que incluyen Spark SQL para SQL y DataFrames, MLlib para el aprendizaje automático (https://pypi.org/project/pyspark/).

En la Tabla 3.3 se resume las características más relevantes de las herramientas y APIs analizadas.

Tabla 3.3. Características del producto

Herramienta	Licencia	Escalable con algoritmos	Análisis Semántico	Compatible con Wordnet	Soporta el Idioma Español
Scikit-Learn	Libre	SI	SI	SI	SI
Textblob	Libre	SI	SI	SI	NO
Intellexer	Comercial	NO	SI	NO	SI
Sentiment Analyzer	Libre	NO	NO	SI	NO
Repustate	Comercial	NO	SI	NO	SI
Microsoft Text Analytics	Comercial	NO	NO	SI	SI
Pyspark	Libre	SI	SI	NO	SI

Fuente: Elaboración Propia

4. Objetivos concretos y metodología de trabajo

Considerando el estado del arte y los trabajos preliminares, el presente trabajo se ha planteado los siguientes objetivos, la metodología a usar y la arquitectura propuesta del software del análisis de sentimiento para este TFM.

4.1. Objetivo general

Desarrollar un sistema para el análisis de los sentimientos políticos de los ciudadanos del Perú, que utilice como corpus un conjunto de tuits extraídos la red social Twitter durante las elecciones congresales 2020 de Perú.

4.2. Objetivos específicos

- Investigar las herramientas del estado de arte que resuelven el problema del análisis de Sentimiento.
- Extraer los datos de Twitter de octubre de 2019 a enero 2020 en torno a los principales partidos políticos del Perú.
- Limpiar los datos de Twitter utilizando técnicas de preprocesamiento de datos.
- Analizar el sentimiento sobre los datos obtenidos usando técnicas de aprendizaje supervisado.
- Desarrollar una herramienta de visualización que permita la extracción de los tuits analizados de los ciudadanos.

4.3. Metodología de trabajo

En este capítulo se indica el proceso realizado para desarrollar la herramienta que determine la polaridad e intensidad de los sentimientos de las elecciones congresales en el Perú el año 2020.

El propósito de la aplicación es polarizar o clasificar textos cortos o tuits de carácter general, en (positivo, neutro, negativo). Para la clasificación se adopta una metodología hibrida combinada a la metodología CRISP-DM la cual tiene las siguientes fases (Figura 4.1).

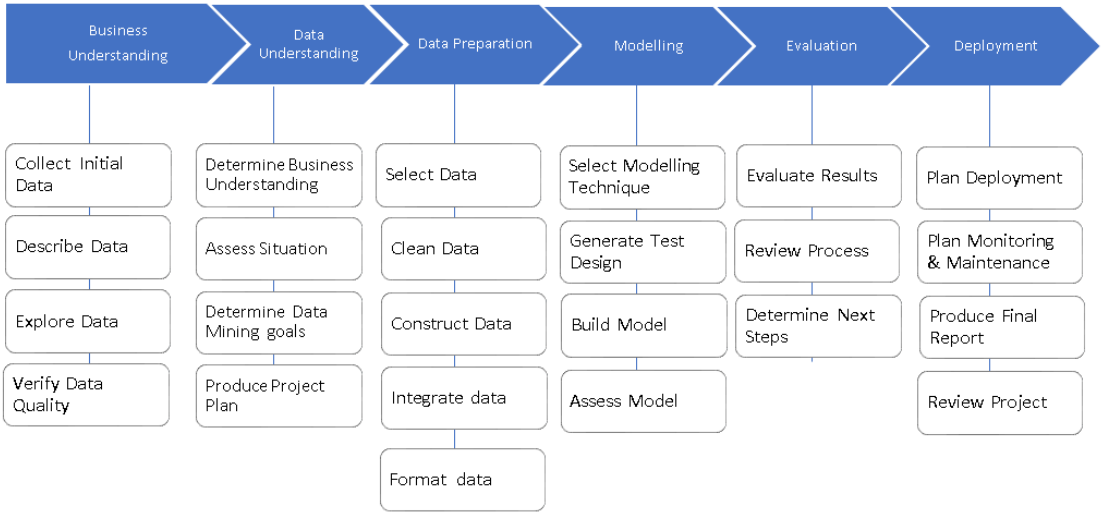

Figura 4.1. Metodología CRISP-DM

Fuente: Página web de IBM : https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/modeler_crispdm_ddita-gentopic1.html

De la metodología se identificaron las siguientes tareas a realizar (Tabla 4.1):

Tabla 4.1. Tareas a realizar

Tareas a realizar	Requisitos
Business Understanding	
Tarea 1. Conseguir los API keys y Token en la cuenta de Twitter	Tener una cuenta de Twitter
Tarea 2. Elección y creación de los Corpus de entrenamiento	Que la data sea en español, de preferencia de Perú y sea de los partidos políticos relevantes
Tarea 3. Elegir los datos a evaluar	No se evaluarán todos los campos de los datos extraídos
Data Preparation	
Tarea 4 Preprocesamiento: Hacer el proceso de, Normalización, Stemming y Tokenización	Analizar las tareas de preprocesamiento que más se ajusten a la data.
Modeling	
Tarea 5. Desarrollo de modelos de aprendizaje supervisado	Se evaluarán 4 modelos de aprendizaje supervisado
Evaluation	
Tarea 6. Analizar con las métricas de evaluación el mejor algoritmo construido para esta problemática.	Se utilizarán las métricas de evaluación Precision, Accuracy, Recall.
Deployment	
Tarea 7. Almacenar los parámetros del mejor modelo para ser usado por una API y función de Phyton	
Tarea 8. Desarrollar una herramienta MEAN la cual pueda realizar las siguientes visualizaciones. • Polaridad por partido. • Nube de palabras por partido. • Línea de tiempo de los tuits por partido. • Línea de menciones de los tuits por partido. • Usuarios con más publicaciones por partido.	Esta tarea será complementada con otra metodología llamada Ágil, para la construcción de la herramienta

Fuente: Elaboración Propia

A continuación, se describen cada una de las tareas realizadas.

4.3.1. Business Understanding

En esta fase se elegiría los corpus de testeo y entrenamiento de la aplicación y se consumirá la API de Twitter. Preferentemente los corpus a utilizar para este TFM deben ser en español y relacionados con la política peruana o del Perú.

- **Tarea 1.- Conseguir los API keys y Token en la cuenta de Twitter**

Para esto se debe tener una cuenta de Twitter. La API free de Twitter presenta ciertas limitaciones en cuanto al tiempo de captura de datos, es decir con relación al análisis de Hashtag solo pueden recopilarse los tuits de 7 días antes. Otras opciones que presenta son versiones de paga como la Premium y Enterprise, con capacidad de captura más amplia en relación con el tiempo. Para estas versiones ingresar al siguiente link: https://developer.twitter.com/en/pricing.html.

La Figura 4.2 muestra la culminación del proceso de creación de la API de Twitter, la cual nos da los keys y tokens para poder trabajar en la extracción de datos. La cual nos dará las siguientes credenciales: Consumer Key, Consumer Secret. Access Token. Access Secret.

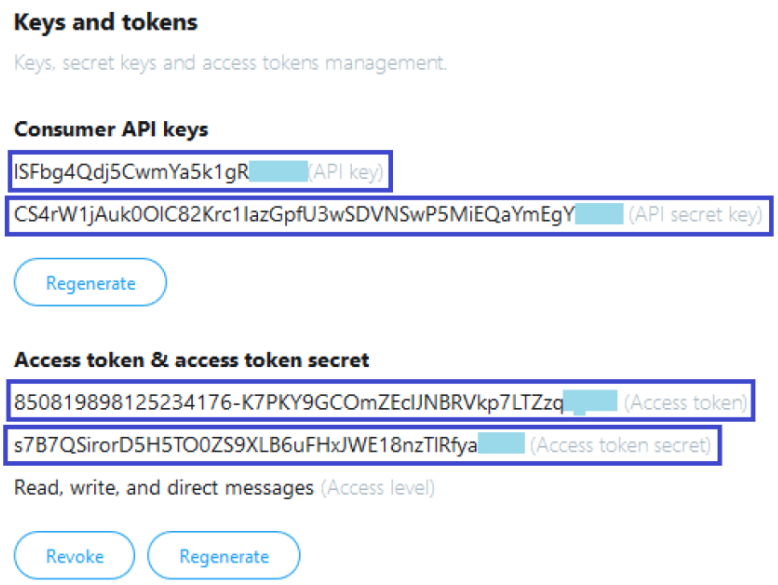

Figura 4.2. Creación de una API en Twitter
Fuente: Elaboración Propia

- **Tarea 2.- Corpus a utilizar**

Para esta sección se utilizará un corpus del Taller de Análisis Semántico en SEPLN(TASS) el cual es muy usado para las pruebas de del análisis de sentimiento en español.

Debido a que la API de Twitter solo considera 140 caracteres en la búsqueda por defecto se considera la propiedad tweet_mode='extended' y el lenguaje español (lang='es'). Se puede mostrar. Esta configuración también se aplicará para los datos de los partidos políticos. El resultado de esta búsqueda se grabará en un archivo JSON, para su evaluación, dicho archivo muestra una estructura como en la Figura 4.3.

```
"created_at": "Wed Jan 22 19:25:25 +0000 2020",
"id": 1220064946081406979,
"id_str": "1220064946081406979",
"full_text": "RT @Socialistas90: Junto a @CosseCarolina y compa\u00f1eras y compa\u00f1eros frenteamplistas
"truncated": false,
"display_text_range": [
    0,
    140
],
```


Figura 4.3. Estructura del resultado de un tuit

Fuente: Elaboración Propia

Se utilizará esta información para ayudar a la construcción del modelo de clasificación A partir de tener estos datos se clasificará cada tuit según la polaridad las cuales son P, NEU, N (positivo, neutro, negativo).

Después de obtener los datos se buscarán si algún dato de los campos de texto o usuario tienen valor nulo y de no pasar el 1% de la muestra se eliminarán, para este caso no se encontraron datos con valor nulo, de ser el caso habría un sesgo en los resultados.

Realizado esto se hace una clasificación manual de los tuits según las palabras o sus orientaciones, utilizando las emociones primarias, secundarias y terciarias de Parrott (2001), el cual se muestra en la Tabla 4.2. (Estos términos han sido traducidos al español).

Tabla 4.2 Emociones primarias, secundarias y terciarias de Parrott (2001)

Emoción primaria	Emoción Secundaria	Emoción Terciaria
Ira	disgusto	Desprecio, odio, repulsión
	envidia	Celos
	exasperación	frustración
	Irritabilidad	Agravación, agitación, molestia, parche, malhumorado, gruñón
	Rabia	Ira, amargura, disgusto, ferocidad, furia, odio, hostilidad, indignación, resentimiento, desprecio, rencor, venganza
	Tormento	
Miedo	Horror	Alarma, susto, histeria, mortificación, pánico, conmoción, terror.
	Nerviosismo	Ansiedad, aprensión. (miedo), angustia, temor, suspenso, inquietud, preocupación

Gozo	Alegría	Diversión, dicha, jovialidad, deleite, disfrute, felicidad, júbilo, euforia, satisfacción, éxtasis, euforia
	Satisfacción	Placer
	Contentamiento	
	Cautivo	éxtasis
	Optimismo	esperanza
	Orgullo	Triunfo
	Alivio	
	Ánimo	Entusiasmo, euforia, emoción, alivio.
Amor	Afecto	Adoración, atracción, cariño, compasión, simpatía, sentimentalismo, ternura.
	Deseo	
	Lujuria / deseo sexual	Deseo, enamoramiento, pasión.
Tristeza	Decepción	Consternación, disgusto
	Negligencia	Alienación, derrotismo, abatimiento, vergüenza, nostalgia, humillación, inseguridad, insulto, aislamiento, soledad, rechazo.
		Depresión, desesperación, melancolía, miseria, infelicidad, aflicción.
	Vergüenza	Culpa, arrepentimiento, remordimiento
	Sufrimiento	Agonía, angustia, dolor
	Simpatía	
Sorpresa		Asombro

Fuente: Liu, B (2015)

El lenguaje de anotación y representación de emociones (EARL) propuesto por la Red de Interacción Humano-Máquina sobre Emociones (HUMAINE) (HUMAINE, 2006) ha clasificado cuarenta y ocho emociones en diferentes tipos de orientaciones o valencias positivas y negativas las cuales están descritas en la Tabla 4.3 Sin embargo, según el autor se debe tener en cuenta que algunas emociones no tienen orientaciones positivas o negativas, por ejemplo, sorpresa e interés. Algunos psicólogos consideraron que no deberían considerarse como emociones (Ortony y Turner, 1990) simplemente porque no tienen orientaciones o valencias positivas o negativas. Por la misma razón, no se usan comúnmente en el análisis de sentimientos. (Liu, 2015).

Tabla 4.3. Datos de Twitter por palabra clave de los partidos políticos

Negative and forceful	Negative and passive	Quiet positive
Anger	Boredom	Calm
Annoyance	Despair	Content
Contempt	Disappointment	Relaxed
Disgust	Hurt	Relieved
Irritation	Sadness	Serene
Negative and not in control	Positive and lively	Caring
Anxiety	Amusement	Affection
Embarrassment	Delight	Empathy
Fear	Elation	Friendliness
Helplessness	Excitement	Love

Fuente: Liu, B (2015)

- **Tarea 3.- Eleccion de datos a evaluar**

Los datos para entrenar son una combinación de los datos del TASS y un corpus personalizado. El corpus personalizado es basado en los datos extraídos de la API Twitter, como por ejemplo los datos del partido morado(#PartidoMorado), cuya estructura final es la siguiente.

- El campo full_text: el cual contiene texto del tuit en modo extendido.
- El campo POS: El cual contiene la polaridad del tuit clasificado en tres categorías P (positivo), negativo (N), y neutro (NEU).

Una vez reconocida la polaridad se procederá al etiquetado del tuit para la construcción del corpus, como se visualiza en la Tabla 4.4. Para este proceso se pidió la ayuda de un doctor en lingüística para la verificación correcta del etiquetado del corpus.

Tabla 4.4. Ejemplo de etiquetado

tweetId	Texto del tuit. (full_text)	Polaridad (POS)	Emoción
1219484966935060483	- ¿qué haces aquí? - me dijiste que había reunión del #partidomorado. https://t.co/bgalmkkldk	NEU	
1218686355431153665	julioguzmanperu con convicción mi voto sigue siendo por el #partidomorado https://t.co/u6z7iwmdmb	P	Simpatía
1220072402832375809	esta es la gente del #partido morado. encubridores de acoso sexual y estafadores de estudiantes universitarios	N	disgusto

Fuente: Elaboración propia

Para el entrenamiento y evaluación de nuestro algoritmo se ha usado los siguientes corpus del TASS (http://tass.sepln.org/tass_data/download.php).

- **General corpus train:** Ofrece 7219 tuits escritos en español por diversos personajes y celebridades conocidas en el ámbito de la política, economía, comunicación y cultura y que fueron obtenidos entre noviembre de 2011 y marzo de 2012.

- **Politics corpus**: Presenta 2500 tuits separados durante la campaña electoral de las Cortes Generales de España de 2011. Estos mensajes dan a conocer los cuatro partidos políticos que más sobresalen en aquel momento: PP, PSOE, IU y UPyD.

- **InterTASS Perú corpus train:** Contiene 1000 tuits escritos en español. Estos mensajes mencionan temas tales como política, economía, comunicación y cultura del Perú (2019).

- **InterTASS Perú corpus dev:** Contiene 500 tuits escritos en español. Estos mensajes mencionan temas tales como política, economía, comunicación y cultura del Perú (2019).

- **Corpus Partido Morado:** Es un corpus personalizado con 30468 tuits, está basado en la información de Twitter y escrito en español.

El formato de los corpus TASS es en XML, pero por motivos de flexibilidad se ha usado el formato CSV, creando un nuevo corpus a partir de la unión de todos los demás. En resumen, la Tabla 4.5. muestra el número de tuits por cada nivel de clasificación y a la colección a la que pertenece. No se ha considerado la clase NONE de dichos corpus.

Tabla 4.5. Distribución de polaridad de los partidos políticos

Corpus	Positivo(P)	Negativo(N)	Neutros (NEU)	Total
General Corpus Train	2884	2182	670	5736
Politics Corpus	639	698	941	2278
Corpus Tass 2019 Perú train	231	242	166	639
Corpus Tass 2019 Perú dev	95	106	61	262
Corpus Partido Morado	7523	14441	8504	30468
TOTALES	11372(29%)	17669(45%)	10342(26%)	39383(100%)

Fuente: Elaboración propia

4.3.2. Data Preparation

En cada proceso de entrenamiento de algoritmos, es conveniente que la información sea limpia y normalizada, con el propósito que se eliminen los datos que influyan de una manera negativa, ya que es habitual conseguir mensajes con carencias ortográficas, caracteres duplicados, el uso incorrecto de las mayúsculas y minúsculas., así como el uso de jerga y abreviaturas en la redacción. A continuación, se describe la tarea más importante de esta fase.

- **Tarea 4.-Preprocesamiento**

La tarea de preprocesamiento es de las más largas que se encontraron en este proceso, debido a que la red de Twitter tiene muchas formas de expresar, y el texto muchas veces presenta problemas gramaticales que es necesario corregir para mejorar los datos a analizar. Para el proceso de preprocesamiento de datos se utilizó el siguiente flujo (Figura 4.4). Un mal formato de los datos puede causar pronósticos erróneos en la solución del problema.

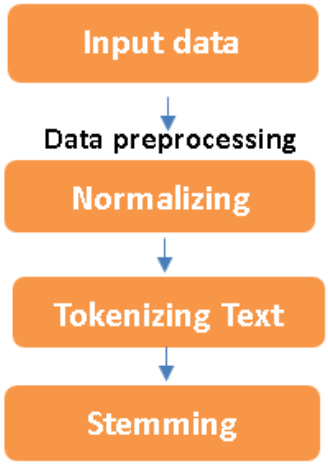

Figura 4.4. Flujo de preprocesamiento de datos
Fuente: Elghazaly, Tarek & Mahmud, Amal & Hefny, Hesham. (2016). Political Sentiment Analysis Using Twitter Data. 1-5. doi:10.1145/2896387.2896396.

Se necesita contar con un proceso para normalizar el texto tanto para el corpus de entrenamiento y pruebas, por eso hay que crear una función con determinadas reglas para efectuar dicha limpieza, evitando que este proceso al ser aplicado no pierda la polaridad inicial del mensaje.

- **Eliminación de tildes:**
  La mayoría de los usuarios peruanos no utilizan adecuadamente el uso de las tildes, por eso al usarlas incorrectamente el algoritmo puede considerarla de una manera distinta. En este caso se remplazan las vocales con tilde por vocales sin estas.
- **Eliminación de usuarios, enlaces, hashtags y caracteres ASCI extraños:**
  Hay muchos símbolos que son escritos por equivocación y que no aportan sentido a la oración por lo que se sustraerán de los mensajes de Twitter. Los usuarios generalmente son precedidos por el @ por ejemplo @MauricioMulder. También se eliminan, las palabras reservadas dentro del retuit cuyas palabras contengan la letra "RT".
- **Normalización de mayúsculas y minúsculas:**
  Se normalizan las palabras que estén en mayúsculas o que contengan una mayúscula, o en caso contario la convierte en letras minúsculas.
- **Tratamiento de la duplicidad de caracteres:**
  Muchas veces los usuarios de Twitter y en general de las redes sociales suelen repetir los caracteres o acentuar dicha palabra como por ejemplo "Gaaanamos" o "Traidoreees", dichas palabras no pueden ser reconocidas por el algoritmo de aprendizaje automático. Por ende, no aportar al sentido semántico de la oración.

Con esta función se normalizarán dichas palabras.

- **Eliminación de números:**
  Muchas veces los números no aportan información relevante para la identificación de la polaridad del sentimiento, por lo que se procederá a eliminarlos de los mensajes de Twitter a analizar.
- **Normalización de risas y Jergas:**
  El aumento del uso de las plataformas de internet y la rapidez por comunicarse o expresar una opinión ha generado un cambio en la forma de expresarse por ejemplo las jergas predominan en las conservaciones son “q”, “TQM”,” jaja”. Muchas de estas expresiones al ser normalizadas pueden aportar sentido de polaridad al mensaje transmitido por el usuario.

En la Tabla 4.6 se muestra un ejemplo de preprocesamiento.

Tabla 4.6. Ejemplo de preprocesamiento

tweetId	Texto Normal.	Texto Preprocesado
12189805933436 39552	RT @CesarBejarano21: @CayetanaAljovin Por eso no podemos volver a votar por #FuerzaPopular #Apra #PodemosPerú #VamosPerú #PerúPatriaSegura	Por eso no podemos volver a votar por fuerza popular apra podemosperu Vamos Perú Perupatriasegura
12178661923896 23808	#AHORA @DanielUrresti1 candidato al Congreso por #PodemosPerú ofrece conferencia de prensa luego de que el @JNE_Peru lo reincorporara a las elecciones #Elecciones2020 #EleccionesCongresales https://t.co/JGopHhmZTi	ahora candidato al congreso por podemosperu ofrece conferencia de prensa luego de que el lo reincorporara a las elecciones Elecciones2020 elecciones Congresales

Fuente: Elaboración propia

Después de normalizar los textos se procede con la etapa de tokenización, en donde las frases se fragmentan en pequeñas unidades llamadas tokens. Una parte importante de esta tokenización es el tratamiento de emoticones.

- **Tratamiento de emoticones:**

  Los usuarios de Twitter muchas veces expresan sentimientos en manera de emoticones. Por eso es necesario que se remplace el carácter del emoticón por su significado textual en español, por ejemplo ☺ significa felicidad ☹ significa tristeza. Con esta función se remplazarán dichos emoticones por su significado.

- **Lematización:**

  Es un proceso lingüístico que modifica el lema de una palabra con el uso del diccionario, esto sirve para obtener una oración gramatical no sujeta a variabilidad, por ejemplo, la palabra ganadores podría ser remplazada por la palabra ganador, por otra parte, sabemos que juego, juegas, canta, jugamos, jugáis, juegan son formas diferentes(relacionar) de un mismo verbo(jugar); de igual manera, niña, niño, niñita, niños, y otras más, son distintas formas del vocablo niño. Lo que realiza la lematización es poder obviar las diferencias y juntar todas estas variantes en un solo término. Po otra parte el lema son todas aquellas frases que expresan una idea o algún pensamiento referente a algo o alguien.

- **Stemming:**

  Es un método que permite reducir una palabra a su raíz por medio de supresión de sufijos e inflexiones. Es parecido a la lematización, pero los resultados no son muchas veces las palabras de un idioma. Un ejemplo es picante y picar tienen como raíz pic. Una desventaja del stemming es que pueden “recortar” demasiado la raíz y encontrar relaciones entre palabras que realmente no existen (overstemming). También puede suceder que deje raíces demasiado extensas o específicas, y que tengamos más bien un déficit de raíces (understemming).

- **Stopwords:**

  Son aquellas palabras que no poseen un significado alguno o que son vacíos, entre las que se encuentran están; los artículos, los pronombres, las preposiciones, los adverbios, e inclusive algunos verbos. Cabe destacar, dentro del procesamiento de datos no se consideran por su nulo significado y a veces pueden dañar el procesamiento de su lenguaje natural. Para ello, podemos eliminarlos almacenando una lista de palabras las cuales son consideradas como Stopword.

4.3.3. Modeling

En esta fase se buscará modelar los algoritmos de aprendizaje supervisado para la solución de este TFM.

- **Tarea 5.- Desarrollo de modelos de Aprendizaje Supervisado**

Muchas veces es complicado elegir que algoritmo usar para el análisis de sentimientos, sin embargo, basándonos en la publicación en la publicación de Towards Data Science, (Rao, 2019) en donde hicieron las pruebas con diferentes algoritmos y librerías del cual tenemos lo siguiente: que las herramientas Textblob por sí sola no es tan eficiente en cuanto a la predicción del sentimiento, se consideró el método de Naive Bayes(NB), las maquinas Vectores de Soporte (SVM), Random Forest y las redes neuronales convolucionales(RNC). Pero no solo importa tener el mejor resultado, sino que sean escalables y fáciles de interpretar y que tengan rapidez en cuanto al desarrollo de estos.

- **Random Forest:**

  Es una mejora del algoritmo árbol de Decisión, Random Forest es un tipo de Ensamble en Machine Learning que obtiene buenos resultados en los problemas de análisis de sentimiento.

  Los modelos Random Forest se definen como conjuntos de combinación de árboles, donde cada uno muestra los datos de una manera distinta mediante bootstrapping. Posteriormente para obtener la nueva observación de una predicción, se debe agregar todos los datos individuales para que formen un modelo.
  Por otra parte, todos los métodos que se basan en árboles, ha de convertirse en uno de los ámbitos predictivos relevantes, debido a los buenos resultados que estos generan en problemas de clasificación y regresión.(Amat Rodrigo, 2020)
  Un grupo de modelos “débiles”, se combinan en un modelo robusto. Cada árbol da una clasificación (vota por una clase). Y el resultado es la clase con mayor número de votos en todo el bosque (forest). En la Figura 4.5. se visualiza un ejemplo de modelo de random forest.

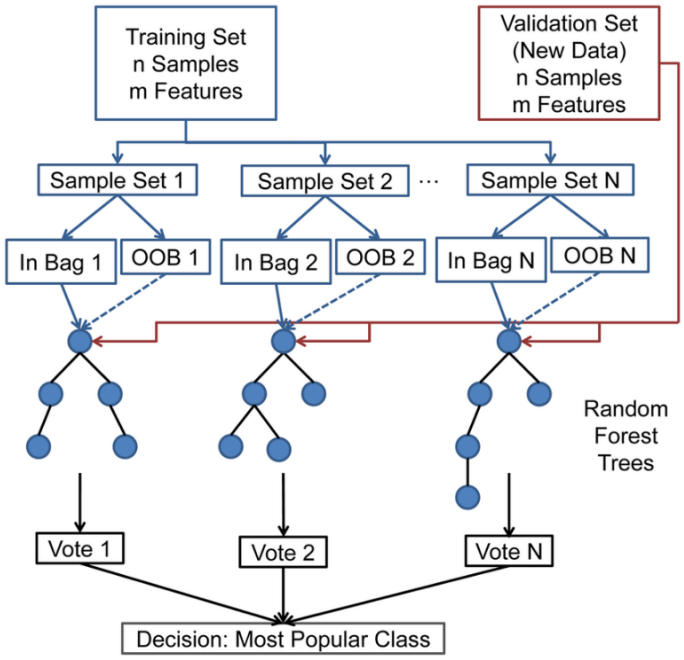

Figura 4.5. Modelo de Random Forest

Fuente: Pagina web de bookdown.org

https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html#random-forest

- **Naive Bayes:**

  Realiza la clasificación asumiendo la independencia entre las características (ingenuo) y calcula las clases según la probabilidad bayesiana

  La familia de algoritmos Naive Bayes están fundados en el célebre Teorema de Bayes, el cual dice lo siguiente:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

  Donde:

  P(H) es la probabilidad a priori esencialmente basado en lo que conocemos del problema y cuyos valores puede tomarse como hipótesis.

  P(D|H) es el grado de presunción condicional de una hipótesis H, dado los datos D, específicamente, la probabilidad de obtener D dado que H es cierta.

  P(D) es el grado de presunción marginal o evidencia, de los valores predictores, en otras palabras, cual es la probabilidad de observar los datos D promediando en comparación con el resto de los datos H.

  P(H|D) es la probabilidad a posteriori representa lo que se conoce de H despues de

recolectar los datos de D. Es el resultado lógico de haber usado un conjunto de datos, un grado de presunción y un a priori.

En términos simples, se usa para indicar que la variable dependiente H, está condicionada por varias variables independientes D = {d1, d2, …, dn}, las cuales se asumen que las probabilidades de los datos D son independientes de cualquier otro, expresado en la siguiente formula:

$$Posterior = \frac{Probabilidad\ x\ Anterior}{Evidencia}$$

La ecuación nos indica la probabilidad de que una hipótesis H sea cierta si algún evento D ha sucedido. (Camacho, 2020)

El modelo Naive Bayes es uno de los algoritmos más comunes para la clasificación de texto. El algoritmo Naive Bayes asume que todas las variables predictoras son independientes entre sí. En pocas palabras, asume que la presencia de una característica en particular no está relacionada con ninguna otra característica en los datos. Esta suposición no siempre es correcta en la vida real

En la Figura 4.6 se muestra un modelo de Naive Bayes

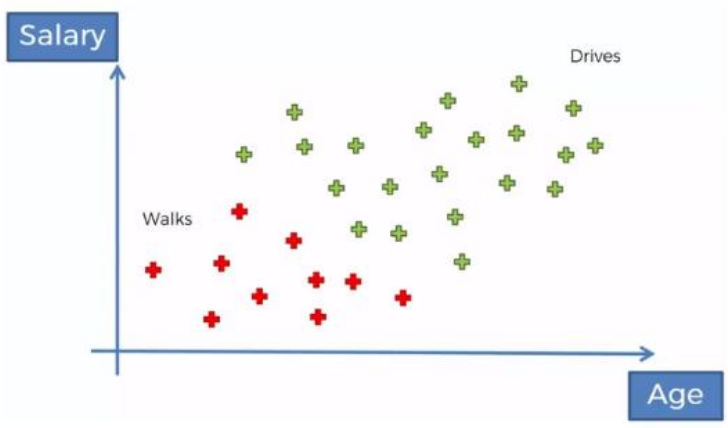

Figura 4.6. Modelo de Naive Bayes

Página de Towards Data Science
https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed

- **SVM:**

  Las máquinas de vectores de soporte (SVM) son muy similares a la regresión logística en términos de cómo optimizan una función de pérdida para generar un límite de decisión entre puntos de datos. Sin embargo, la principal diferencia es el uso de "funciones del núcleo", es decir, funciones que transforman un espacio de decisión complejo y no lineal en uno que tiene una mayor dimensionalidad, de modo que se pueda encontrar un hiperplano apropiado que separe los puntos de datos. El clasificador SVM busca maximizar la distancia de cada punto de datos desde este hiperplano usando "vectores de soporte" que caracterizan cada distancia como un vector.

  Una característica clave de las SVM es el hecho de que utilizan una pérdida de bisagra en lugar de una pérdida logística. Esto lo hace más robusto para los valores atípicos en los datos, ya que la pérdida de bisagra no diverge tan rápidamente como una pérdida logística. (Rao, 2019). En la Figura 4.7 se muestra el ejemplo de un SVM.

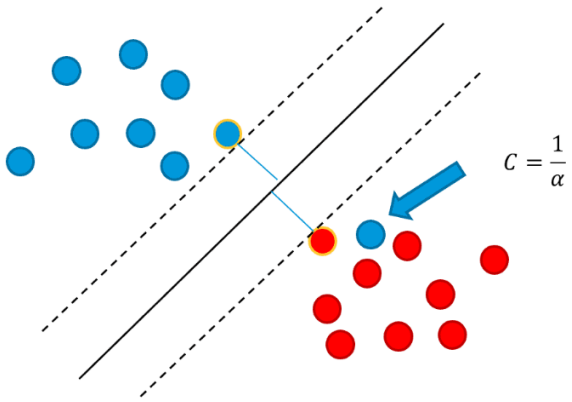

Figura 4.7. Modelo de SVM

Página web de IAArtificial.Net
https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/

- **Redes neuronales convolucionales:**

  Son una clase de redes neuronales artificiales las cuales se pueden aplicar para la solución de problemas NLP, su principal ventaja es que cada parte de la red necesita entrenarse para ejecutar una labor, reduciendo a la vez un numero de capas ocultas, de manera que aumenten la rapidez del entrenamiento.

  Arquitectura:

  Una red neuronal convolucional que consta de capas convolucionales y de reducción alternadas, y al finalmente tiene capas de conexión total como una red perceptrón

multicapa. Son utilizadas para el tratamiento de imágenes, pero también se pueden aplicar a tareas del NLP. En la Figura 4.8. se muestra un ejemplo de una red neuronal convolucional

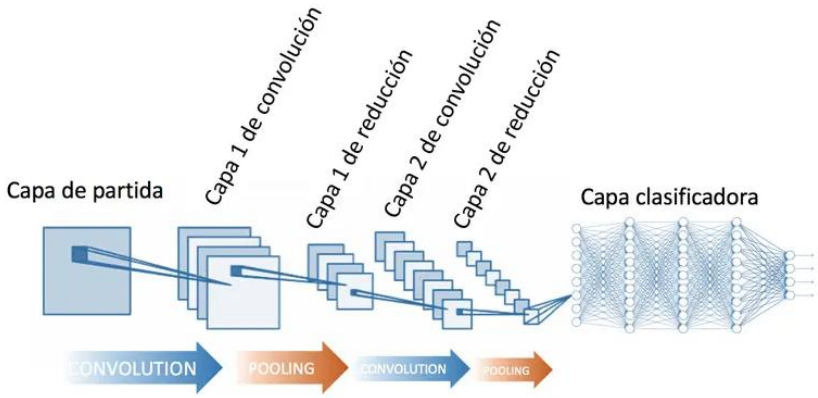

Figura 4.8. Modelo de una red convolucional
Página web de Diego Calvo
https://www.diegocalvo.es/red-neuronal-convolucional/

Para el caso del análisis de sentimientos es necesario transformar las palabras en vectores utilizando one hot encoding, buscando una representación numérica en las palabras. Esto significará que el algoritmo creará una serie de relaciones durante el proceso. Las redes neuronales crean una relación llamada Word embedding, las cuales son usadas en este tipo de relaciones como en la Figura 4.9 .

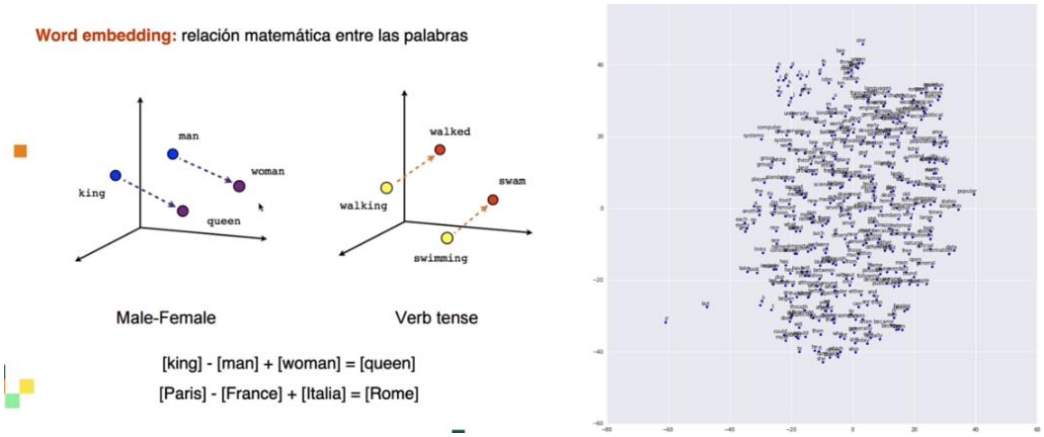

Figura 4.9. Word embedding
Fuente: Elaboración propia

Una de las preguntas que se hace en este tipo de soluciones.

- ✓ ¿Cómo entrenamos esta matriz?
- ✓ ¿Cuántas capas se utilizarán para nuestra red neuronal?

En el lenguaje humano, las palabras que rodean a una palabra dada es lo que le

llamamos el contexto, y como humanos damos un contexto u otro dependiendo de lo que estamos hablando. Los bigramas y trigramas son usado en este tipo de relaciones las cuales son uniones de dos y tres palabras respectivamente

Para el entrenar la red neuronal se necesita utilizar un corpus de texto que tenga muchos datos.

Para el uso de las capas de las capas de convolución en Tensorflow no utilizaríamos la convolución al 2D que se suele utilizar en imágenes, sino que pasaríamos a utilizar la convolución uno de que sería la capa donde se aplica el filtro de convolución en una sola dimensión.

En resumen, los elementos para poder crear la RNC son los siguientes.

Entrada: Es el texto a analizar.

Capa De Convolución: procesa la salida de neuronas que están conectadas en "regiones locales" de entrada, calculando el producto escalar entre sus pesos y una pequeña región a la que están conectados en el volumen de entrada. Aquí usaremos, por ejemplo, tenemos filtro de altura 4, filtro de altura 3, filtro de altura 2, esto significa que queremos asociar grupo de 2 ,3 y 4 palabras, cuya salida es un vector.

Capa relu : aplicará la función de activación en los elementos de la matriz.

Pool ó Subsampling: Hará una reducción en las dimensiones alto y ancho, pero se mantiene la profundidad.

Capa Tradicional: red de neuronas feedforward que conectará con la última capa de subsampling y finalizará con la cantidad de neuronas que queremos clasificar.

Cada uno de estos modelos tienen una forma particular de solución, para esto se debe balancear cada una de las clases para la fiabilidad de los resultados. También se ha evitado el sobreentrenamiento (overfitting) debido que los modelos no solo se ajustan a aprender casos particulares, sino para el reconocimiento de nuevos datos. Cada uno de los modelos ha utilizado el mismo corpus de entrenamiento, Si el modelo generado no es lo suficientemente bueno, se utiliza otros parámetros hasta lograr una mejor precisión.

Almacenamiento de modelos: con el fin de no estar entrenando el modelo cada vez que llega una nueva información, dicho modelo se debe serializar, para su uso en cualquier aplicación.

4.3.4. Evaluación

En esta fase se eligió cuatro métricas para la evaluación del algoritmo de análisis de sentimiento.

- **Tarea 6.- Métricas de evaluación**

Son usadas para medir el rendimiento del modelo entrenado, antes de usarlo en producción. Si no se realiza una evaluación apropiada del modelo puede darse el caso de que este genere malas predicciones. Esto sucede porque, en casos como éste, los modelos no memorizan, sino que guardan información.

- **Exactitud (Accuracy)**

Es la principal medida la cual se basa en el porcentaje de los textos clasificados correctamente. Esta medida la obtenemos mediante la siguiente formula:

$$\textbf{Accuracy} = \frac{true\ postitives + true\ negative}{\text{Tamaño del Corpus}}$$

- **Precisión**

Se usa para medir la proporción de los textos clasificados correctamente por el algoritmo, mide el porcentaje de textos clasificados correctamente del conjunto total de documentos que el algoritmo evalúa. La fórmula para obtener la precisión de los textos evaluados es la siguiente Donde true positives es el número de textos clasificados correctamente y false positives son los textos clasificados por el algoritmo de forma incorrecta.

$$Precision = \frac{\text{True Positives}}{\text{True Positives+ False positives}}$$

- **Exhaustividad (Recall)**

Se usa para determinar qué porcentaje de los textos seleccionados forman parte del conjunto objetivo. Es decir, el porcentaje de textos que el sistema clasifica correctamente sobre el conjunto total de textos que se deben clasificar. En el caso de la exhaustividad por polaridad, se comprueban los textos clasificados correctamente sobre el conjunto de textos total que pertenecen a polaridad valorada. La fórmula para obtenerla es la siguiente:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+ False positives}}$$

- **Medida-F (F1-measure)**

Como último, la medida-F combina la precisión y la exhaustividad obteniendo una medida de desempeño general, en la que no se sacrifica la precisión o la exhaustividad. Cuya formula es la siguiente. Donde la Precisión y Recall fueron obtenidos previamente.

$$F1 = 2 * \frac{\text{precisión} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.3.5.Deployment

Es la ultima fase donde se despliega el modelo, el cual es usado para la predicción del sentimiento.

Tarea 7.- Despliegue del modelo

Se almacena el modelo ya optimizado y de esta manera puede ser utilizado en una API, para que a su vez sea consumida en la aplicación.

4.4. Desarrollo de la herramienta

Para el desarrollo de la aplicación, el autor eligió una arquitectura en donde el back-end (lado cliente) está separado del front-end (parte del servidor). Se hizo esta decisión ya que es una solución flexible e independiente una de la otra (Tarea 8).

4.4.1. Back-end

La mayoría de las aplicaciones de hoy en día necesitan un back-end que puedan comunicar las bases de datos con servicios externos e internos, aplicaciones móviles y de stemming entre otras.

Para el back-end se está utilizando la API que se ha desarrollado en el entorno de ejecución multiplataforma de Node. El cual está basado en el lenguaje de programación JavaScript, asíncrono. Node es sobresaliente para el manejo de muchas operaciones de I/O (Input/Output) se puede utilizar para crear aplicaciones que implementen real time en sus servicios. También permite transmitir una gran cantidad de datos en fragmentos en forma secuencial. Otra ventaja del Node es su escalabilidad.

Otra herramienta que ha sido utilizada es la API de Twitter la cual es consumida para los nuevos tuits. Flask utiliza el los pesos del modelo ajustado una vez entrenado juntamente con el tokenizer.

4.4.2. Front-end

Para la selección del front-end hay muchas variantes, pero se eligió Angular por ser un framework muy usado en la actualidad y debido a la experiencia del autor en este software. La curva de aprendizaje no es tan alta como en otras herramientas del front-end. Tiene también librerías para la visualización de gráficos y demás componentes como los controladores web. Se utiliza la inyección de dependencias para comunicarse con el back-end.

4.4.3. Base de Datos

Para este tipo de aplicaciones la mejor opción es utilizar una base de datos no relacional. Se guardarán los datos de los tuits con su respectiva polaridad. Se utiliza para cuando se tiene necesidad de volumen, velocidad y variabilidad.

5. Desarrollo específico de la contribución

El desarrollo especifico de esta contribución explica el desarrollo del software para el análisis de sentimiento de los partidos políticos.

5.1. Requisitos de software

Se identificaron los siguientes requisitos:

- Diseñar una herramienta web para que pueda ser usada mediante cualquier dispositivo.
- Desarrollar un tutorial para el uso de la herramienta.
- Verificar la usabilidad de la herramienta.

5.2. Análisis de requisitos

El análisis de requisitos establece prioridades para las funcionalidades y alcance de la aplicación, con la finalidad de agregarlas según el valor para el cliente En el desarrollo del aplicativo se están considerando las tres siguientes estructuras:

1. Análisis de riesgos. Se han analizado los riesgos más importantes del proyecto y se ha tomado las medidas tomadas para minimizarlos.
2. Historias de usuario. Las historias de usuario simbolizan los requerimientos de la aplicación expresados en un lenguaje sencillo.
3. Diagrama de arquitectura. Explica en detalle la comunicación entre los componentes más relevantes de la aplicación y su relación con otros elementos

5.2.1. Análisis de riesgos

Se identificaron los riesgos más importantes para este desarrollo software:

Viabilidad tecnológica de los objetivos planteados.

El diseño de una aplicación implica el desarrollo de una aplicación interactiva que pueda ser usada de varias maneras

Se analizaron varias propuestas tecnológicas tanto en el desarrollo del modelo como en el desarrollo de la aplicación. Pero por motivos de tiempo se eligieron tecnologías que se

abarcaron en este master y las que domina el autor como MongoDB, Node, Express, Angular (MEAN) y Phyton.

Incertidumbre debida a la elección de una tecnología de desarrollo determinada.

La elección de una tecnología determinada indica la viabilidad de funciones y actualizaciones que soporta esa plataforma, además el desarrollador se hizo una evaluación sobre sus capacidades. Se publicará en un entorno local y se versionará las versiones del proyecto con Git. Para la gestión de tiempos se mantendrá una comunicación con la tutora.

5.2.2. Historias de usuario

Las historias de usuario identifican los requisitos de la aplicación desde el punto de vista y lenguaje del propio usuario. Además, describen el resultado deseado y tienen una prioridad asociada a la funcionalidad del aplicativo.

Los usuarios de esta aplicación son principalmente personas interesadas en la política peruana.

En la Tabla 5.1 se describen las historias de usuario junto con un indicador de prioridad siendo A mayor prioridad y C menor prioridad.

Tabla 5.1 Historia de Usuario

ID	Historia de usuario	Prioridad
	Requisitos hardware/software	
1	El usuario necesita que la aplicación se ejecute vía web	A
2	El usuario necesita que la aplicación pueda ser manejado en las versiones de Chrome, Firefox, Explorer, etc..	C
3	El usuario necesita que se utilice las tecnologías de JavaScript, CSS y otras tecnologías como Node.	A
4	El usuario necesita que se pueda utilizar haciendo un registro previo para el uso del sistema.	A
5	El usuario quiere que los datos se almacenen para mejorar su facilidad de uso	C
	Requisitos de usabilidad/accesibilidad	
6	El usuario necesita que cada partido político tenga una interfaz diferente	B
7	El usuario necesita una aplicación fácil de usar	A
8	El usuario quiere que se controle mediante ratón o de forma táctil.	B
9	El usuario amerita que la interfaz sea muy sencilla, con el mínimo de controles.	B

	Forma de programar (No Funcional)	
10	Se necesita tener una tecnología de front-end independiente del back-end para una mejor distribución.	A

Fuente: Elaboración propia

5.2.3. Diagrama de arquitectura

El diagrama de la arquitectura se divide en 2 momentos, el ingreso de datos y la arquitectura de la aplicación,

Ingreso de Datos: Se ha buscado mediante la API de Twitter, las palabras relacionadas a los partidos políticos, ya mencionadas en el capítulo anterior. El script de Phyton consume la API de Twitter y graba en un archivo JSON por cada partido político. Este archivo es transformado con los datos relevantes de Twitter para la base de datos de MongoDB. Los datos relevantes para este TFM son presentados en la Tabla 5.2.

Tabla 5.2. Diccionario de datos del tuit

Atributo	Tipo	Descripción
created_at	String	Hora UTC cuando se creó el tuit.
id_str	String	El identificador único del tuit.
Full_text	Text	El texto UTF-8 del cual se hace la publicación
source	Text	Utilizada para postear el tuit. Tienen un valor fuente de web.
user	Objeto	El usuario que publicó este tuit. Para este caso utilizaremos screen_name y location
retweet_count	Entero	Número de veces que se ha retuiteado este tuit.
favorite_count	Entero	Indica aproximadamente cuántas veces le han gustado a este tuit los usuarios de Twitter.

Fuente: Página web de Twitter:
https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

Arquitectura de la aplicación:

En la arquitectura de la aplicación utiliza las siguientes tecnologías.

MongoDB

Es una base de datos NoSQL de carácter general, basada en documentos para el desarrollo de aplicaciones modernas. Utiliza estructuras de datos BSON (una especificación similar a JSON) con un esquema dinámico, lo que facilita su acceso desde las aplicaciones.

El modelo de documento se asigna a los objetos en el código de su aplicación, lo que facilita el trabajo con los datos. Las consultas ad hoc, la indexación y la agregación en tiempo real brindan formas poderosas de acceder y analizar sus datos.

MongoDB junto con un paquete de NPM llamado Mongoose proporcionan un acceso fácil y sencillo para guardar, obtener, actualizar y eliminar datos.

Express

Express es el estándar en todas las aplicaciones servidor creadas en Node. Así mismo, es un framework gestiona las peticiones HTTP de entrada y las encamina. Cabe destacar, que también forma parte del conjunto de aplicaciones MEAN y es el más utilizado en este tipo de soluciones. Es gratuito, flexible y de código abierto, bajo la licencia MIT. Por otra parte, está diseñado para crear aplicaciones web, API y móviles. Cuenta con miles de métodos y middleware a su disposición para que la creación de una API sea sólida es rápida y sencilla.

Node

Node es un entorno JavaScript de desarrollo de libre distribución, multiplataforma. Usa el motor V8 de Google, es asíncrono y gran uso de I/O de datos. Está basado en eventos; creado para ser útil en programas altamente escalables.

Angular

Angular es un framework modular y escalable de libre distribución desarrollado por Google para facilitar la creación y programación de aplicaciones web de una sola página, las webs SPA (Single Page Application). Es importante señalar que está basado en el patrón MVC(Modelo-Vista-Controlador). Además, el lenguaje principal de programación Angular es Typescript, de esta manera toda la sintaxis y el modo de hacer las cosas en el código es el mismo, lo que añade coherencia y consistencia a la información.

En la Figura 5.1 explica las tecnologías usadas para la arquitectura de la solución y su relación entre cada componente, La petición (Request) es lanzada por Angular y escuchada y por Node y Express los cuales escuchan todas las peticiones que lleguen por orden. A través de Mongoose lee los documentos de la Base de Datos de MongoDB y devuelve una colección. Se han añadido middlewares para devolver el resultado de la petición (Response).

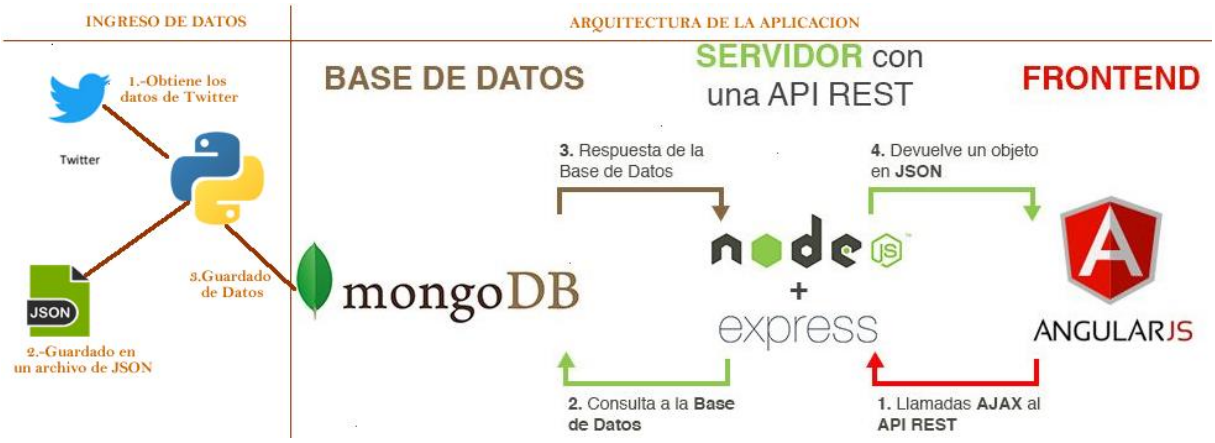

Figura 5.1. Arquitectura de la solución

Fuente: Elaboración propia

5.2.4. Modelo de base de Datos

El modelo de base de datos nos indica como se almacenan los datos, para este TFM se eligió la base de datos MongoDB, debido a que es una base moderna y distribuida que puede escalar en forma vertical (RAM y CPU) y forma horizontal(nodos). Está basada en colecciones y documentos. Las principales colecciones de datos son.

- **Alianza por el progreso (APP):**

t_app: Contiene los datos relevantes del Twitter del partido político APP juntamente con la polaridad del mensaje descrito.
analisis_sentimiento_app: Contiene el resumen de la polaridad del partido político APP.
analisis_tendencia_app : Contiene los términos más relevantes la frecuencia que se utilizan para el partido APP.
t_usuarios_app: Contiene los usuarios con más likes tanto en opiniones negativas como positivas del partido APP.

- **Acción Popular (AP):**

t_ap: Contiene los datos relevantes del Twitter del partido político AP juntamente con la polaridad del mensaje descrito.
analisis_sentimiento_ap: Contiene el resumen de la polaridad del partido político AP.
analisis_tendencia_ap : Contiene los términos más relevantes la frecuencia que se utilizan para el partido AP.
t_usuarios_ap: Contiene los usuarios con más likes tanto en opiniones negativas como positivas del partido AP.

- **Apra:**

t_apra: Contiene los datos relevantes del Twitter del partido político Aprista juntamente con la polaridad del mensaje descrito.
analisis_sentimiento_apra: Contiene el resumen de la polaridad del partido político Aprista.
analisis_tendencia_apra: Contiene los términos más relevantes la frecuencia que se utilizan para el partido Aprista.
t_usuarios_apra: Contiene los usuarios con más likes tanto en opiniones negativas como positivas del partido Aprista.

- **Frente Amplio:**

t_partido_frente_amplio: Contiene los datos relevantes del Twitter del partido político Frente Amplio juntamente con la polaridad del mensaje descrito.
analisis_sentimiento_frente_amplio: Contiene el resumen de la polaridad del partido político Frente Amplio.
analisis_tendencia_frente_amplio: Contiene los términos más relevantes la frecuencia que se utilizan para partido Frente Amplio.
t_usuarios_frente_amplio: Contiene los usuarios con más likes tanto en opiniones negativas como positivas del partido Frente Amplio.

- **Fuerza Popular:**

t_partido_fuerza_popular: Contiene los datos relevantes del Twitter del partido político Fuerza Popular juntamente con la polaridad del mensaje descrito.
analisis_sentimiento_fuerza_popular: Contiene el resumen de la polaridad del partido político Fuerza Popular.
analisis_tendencia_fuerza_popular: Contiene los términos más relevantes la frecuencia que se utilizan para partido Fuerza Popular.
t_usuarios_fuerza_popular: Contiene los usuarios con más likes tanto en opiniones negativas como positivas del partido Fuerza Popular.

- **Partido Morado:**

t_partido_morado: Contiene los datos relevantes del Twitter del partido Morado juntamente con la polaridad del mensaje descrito.
analisis_sentimiento_partido_morado: Contiene el resumen de la polaridad del partido Morado.
analisis_tendencia_partido_morado: Contiene los términos más relevantes la frecuencia que se utilizan para partido Morado.

t_usuarios_partido_morado: Contiene los usuarios con más likes tanto en opiniones negativas como positivas del partido Morado

- **Podemos Perú:**

t_podemos_peru: Contiene los datos relevantes del Twitter del partido político Podemos Perú juntamente con la polaridad del mensaje descrito.
analisis_sentimiento_podemos_peru: Contiene el resumen de la polaridad del partido político Podemos Perú.
analisis_tendencia_podemos_peru: Contiene los términos más relevantes la frecuencia que se utilizan para partido Podemos Perú.
t_usuarios_podemos_peru: Contiene los usuarios con más likes tanto en opiniones negativas como positivas del partido Podemos Perú.

- **Partido Popular Cristiano (PPC):**

t_ppc: Contiene los datos relevantes del Twitter del partido político PPC juntamente con la polaridad del mensaje descrito.
analisis_sentimiento_ppc: Contiene el resumen de la polaridad del partido político PPC.
analisis_tendencia_ppc: Contiene los términos más relevantes la frecuencia que se utilizan para partido PPC.
t_usuarios_ppc: Contiene los usuarios con más likes tanto en opiniones negativas como positivas del partido PPC.
usuario: Contiene los datos de los usuarios, que son para el logueo de la aplicación.

La Figura 5.2. muestra el modelo de la base de datos con las tablas descritas anteriormente

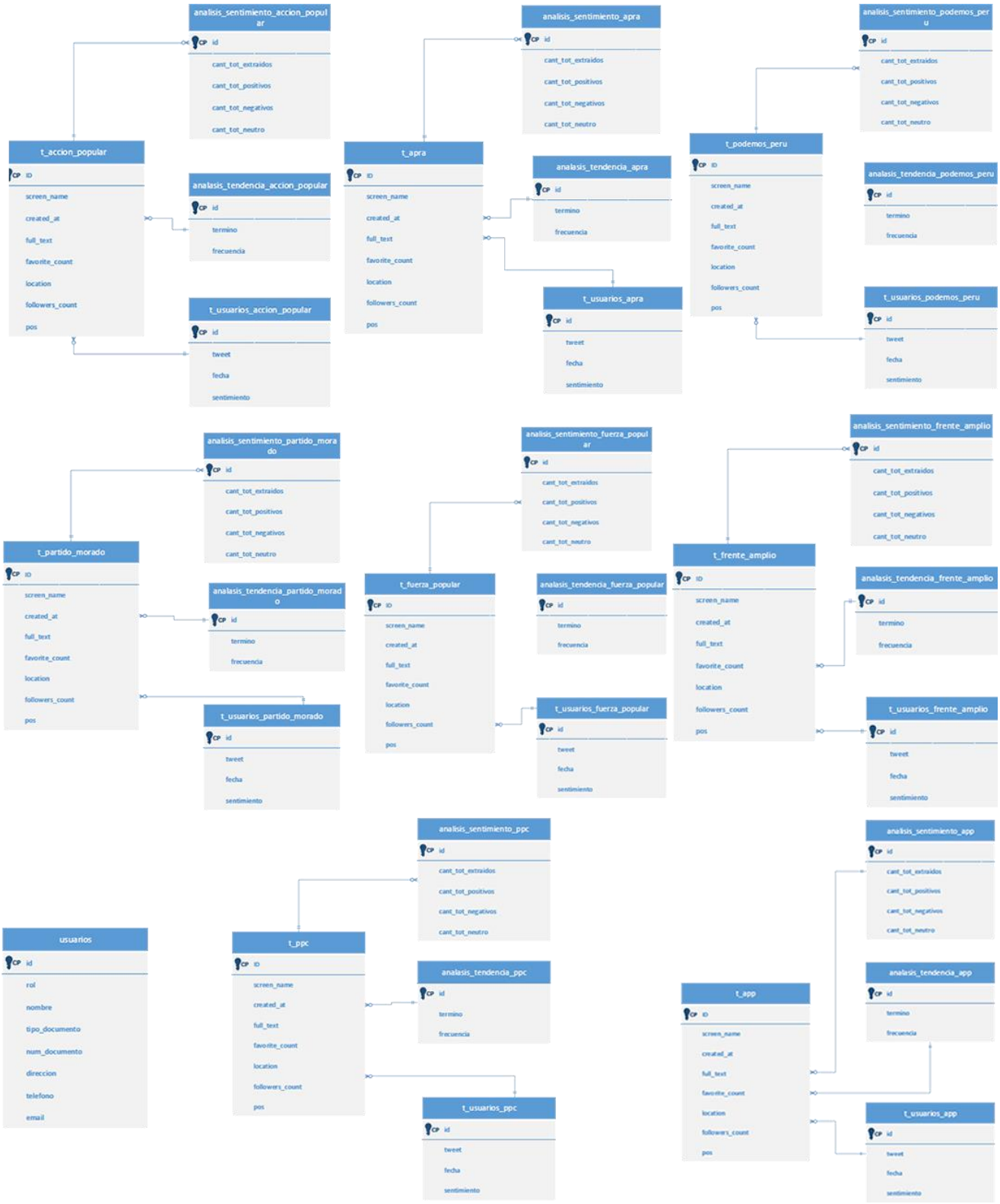

Figura 5.2. *Modelo de base de datos*

Fuente: Elaboración propia

Se ha almacenado 8 colecciones de tuits pertenecientes a los partidos políticos más influyentes en la actualidad del Perú, del cual ya se describió en el estado del arte. Utilizando la API de Twitter con la herramienta Phyton se realizó la búsqueda desde el 01-10-2019, fecha en que se oficializo las elecciones congresales extraordinarias, hasta el 22-02-2020. Cabe mencionar que solo se recolectaron los tuits en español.

Por otra parte, se ha creado scripts en Phyton para grabar los datos resultantes de la API de Twitter a la base de datos MongoDB, estos datos han sido transferidos a MongoDB mediante la librería pymongo. Se ha utilizado funciones propias de Phyton para filtrar y transformar los datos relevantes, y luego ser grabados de forma persistente

En la Tabla 5.3. se visualiza las palabras claves que se usaron para la extracción de datos de los partidos políticos.

Tabla 5.3. Datos de Twitter por palabra clave de los partidos políticos

Partido	Palabra usada para la búsqueda
Acción Popular	#AccionPopular
Alianza para el Progreso	Alianza por el progreso
Apra	Mauricio Mulder
Fuerza Popular	#FuerzaPopular
Frente Amplio	#FrenteAmplio
Partido Morado	#PartidoMorado
Podemos Perú	#PodemosPeru
PPC	#PPC

Fuente: Elaboración propia

5.3. Descripción de la herramienta software desarrollada

La herramienta desarrollada fue en dos partes tanto la del back-end como la del front-end, para mantener independencia y facilitar el trabajo.

5.3.1. Back-end

La aplicación está desarrollada en el back-end con Node y Express, la cual una vez levantada los endpoints, lanzaran peticiones para la diversa integración con el front-end. Express funciona mediante el manejo de middlewares, nuestra carpeta de configuración del back-end quedara de la siguiente manera (Figura 5.3.):

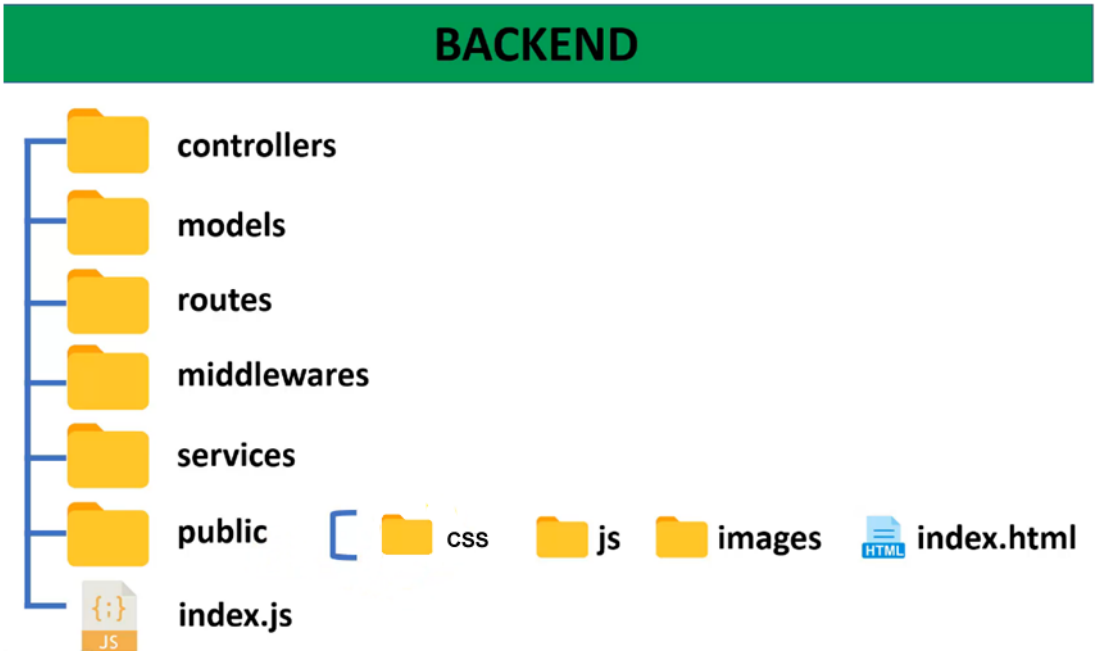

Figura 5.3. Carpetas del back-end
Fuente: Elaboración propia

El contenido de las carpetas es la siguiente:

- La carpeta models se crearán los modelos de la aplicación.
- La carpeta controllers cargaremos todos nuestros controladores las cuales tendrán las funciones a utilizar.
- La carpeta routes contienen las direcciones y el tipo de petición (REQUEST) a utilizar.
- La carpeta middlewares contendrá los middlewares que verificaran los accesos por rol.

Para la conexión con MongoDB se hará empleando la librería Mongoose. La versión del Node será la 10.15, el primer middleware será Morgan, el cual será una dependencia que permite ver las aplicaciones del navegador o del cliente. Para la autenticación de los usuarios se ha creado un endpoint y otro endpoint para el registro de usuarios.
La lista de la Tabla 5.4. contiene todos los endpoints utilizados en el sistema, los cuales son:

Tabla 5.4. Lista de endpoints

Ruta	función	Parámetros	Método HTTP	Autorización
usuario/add	Permite añadir un usuario al sistema	Body: rol (String-30) nombre (String-50) tipo_documento (String-20) num_documento (String-20) direccion (String-70) telefono (String-20) email (String-50) password (String-64)	POST	Administrador
usuario/list	Lista los usuarios del sistema	Query ?valor=texto	GET	Administrador
usuario/query	Realiza la búsqueda de un usuario por las palabras relacionadas a su nombre.	Query ? _id=texto	GET	Administrador
usuario/update	Permite actualizar los datos del usuario del sistema	Body _id (String) rol (String-30) nombre (String-50) tipo_documento (String-20) num_documento (String-20) direccion (String-70) telefono (String-20) email (String-50) password (String-64)	PUT	Administrador
usuario/remove	Permite eliminar lógicamente el usuario	Body _id (String)	DELETE	Administrador
usuario/activate	Permite dar de alta a un usuario	Body _id (String)	PUT	Administrador

usuario/deactivate	Permite dar de baja a un usuario	Body _id (String)	PUT	Administrador
usuario/login	Permite la autenticación del usuario al sistema	Body email (String-50) password (String-64)	POST	Administrador
analisis_sentimiento/query_cantidad_post_app	Lista la cantidad por polaridad del partido APP	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_positivos_app	Lista el top de las 5 publicaciones más positivas del partido APP	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_usuario_app	Lista el top de los 5 usuarios con más publicaciones del partido APP	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_negativos_app	Lista el top de los 5 post más negativos del partido APP	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_serie_app	Lista los términos con más frecuencia del partido APP	Query ?_id=texto	GET	Administrador Usuario
analisis_sentimiento/query_menciones_app	Contiene la lista de polaridad por partido político (APP) por fecha	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_cantidad_post_apra	Lista la cantidad por polaridad del partido Aprista	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_positivos_apra	Lista el top de los 5 post más positivos del partido Aprista	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_usuario_apra	Lista el top de los 5 usuarios con más publicaciones del partido Aprista	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_negativos_apra	Lista el top de los 5 post más negativos del partido Aprista	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_serie_apra	Lista los términos con más frecuencia del partido Aprista	Query ? _id=texto	GET	Administrador Usuario

analisis_sentimiento/query_menciones_apra	Contiene la lista de polaridad por partido político (Apra) por fecha	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_cantidad_post_ap	Lista la cantidad por polaridad del partido AP	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_positivos_ap	Lista el top de los 5 post más positivos del partido AP	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_usuario_ap	Lista el top de los 5 usuarios con más publicaciones del partido AP	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_negativos_ap	Lista el top de los 5 post más negativos del partido AP	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_serie_ap	Lista los términos con más frecuencia del partido AP	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_menciones_ap	Contiene la lista de polaridad por partido político (AP) por fecha	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_cantidad_post_fzap	Lista la cantidad por polaridad del partido Fuerza Popular	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_positivos_fzap	Lista el top de los 5 post más positivos del partido Fuerza Popular	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_usuario_ fzap	Lista el top de los 5 usuarios con más publicaciones del partido Fuerza Popular	Query ?_id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_negativos_fzap	Lista el top de los 5 post más negativos del partido Fuerza Popular	Query ?_id=texto	GET	Administrador Usuario
analisis_sentimiento/query_serie_fzap	Lista los términos con más frecuencia del partido Fuerza Popular	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_menciones_fzap	Contiene la lista de polaridad por partido político (Fuerza Popular) por fecha	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_cantidad_post_fampli o	Lista la cantidad por polaridad del partido Frente Amplio	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_positivos_famplio	Lista el top de los 5 post más positivos del partido Frente Amplio	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post	Lista el top de los 5 usuarios con más	Query ? _id=texto	GET	Administrador Usuario

usuario famplio	publicaciones del partido Frente Amplio			
analisis_sentimiento/query_post_negativos_famplio	Lista el top de los 5 post más negativos del partido Frente Amplio	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_serie_famplio	Lista los términos con más frecuencia del partido Frente Amplio	Query ?_id=texto	GET	Administrador Usuario
analisis_sentimiento/query_menciones_famplio	Contiene la lista de polaridad por partido político (Frente Amplio) por fecha	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_cantidad_post_ppc	Lista la cantidad por polaridad del partido PPC	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_positivo_ppc	Lista el top de los 5 post más positivos del partido PPC	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_usuario_ppc	Lista el top de los 5 usuarios con más publicaciones del partido PPC	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_negativos_ppc	Lista el top de los 5 post más negativos del partido PPC	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_serie_ppc	Lista los términos con más frecuencia del partido PPC	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_menciones_ppc	Contiene la lista de polaridad por partido político (PPC) por fecha	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_cantidad_post_pmorado	Lista la cantidad por polaridad del partido Morado	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_positivos_pmorado	Lista el top de los 5 post más positivos del partido Morado	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_usuario_pmorado	Lista el top de los 5 usuarios con más publicaciones del partido Morado	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_negativos_pmorado	Lista el top de los 5 post más negativos del partido Morado	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_serie_pmorado	Lista los términos con más frecuencia del partido Morado	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_menciones_pmorado	Contiene la lista de polaridad por partido político (partido morado) por fecha	Query ? _id=texto	GET	Administrador Usuario

analisis_sentimiento/query_cantidad_post_pperu	Lista la cantidad por polaridad del partido Podemos Perú	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_positivos_pperu	Lista el top de los 5 post más positivos del partido Podemos Perú	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_usuario_ pperu	Lista el top de los 5 usuarios con más publicaciones del partido Podemos Perú	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_post_negativos_pperu	Lista el top de los 5 post más negativos del partido Podemos Perú	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_serie_pperu	Lista los términos con más frecuencia del partido Podemos Perú	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/query_menciones_pperu	Contiene la lista de polaridad por partido político (Podemos Perú) por fecha	Query ? _id=texto	GET	Administrador Usuario
analisis_sentimiento/analisis_texto	Analiza el sentimiento de un texto	Query ? _id=texto	GET	Administrador Usuario

Fuente: Elaboración propia

Para el testeo de las APIS se han hecho pruebas con la herramienta Postman, a manera de probar que las respuestas de los endpoints sea la correcta. En cada petición los datos son parseados en el formato JSON, a través de Express. Se ha utilizado el IDE de Visual Studio Code por ser uno de los más flexibles y fáciles de usar tanto en el front-end. como en el back-end.

5.3.2. Frond-end

En el desarrollo de la aplicación web se usó Angular en la versión 11. La configuración de las carpetas es la siguiente (Figura 5.4).

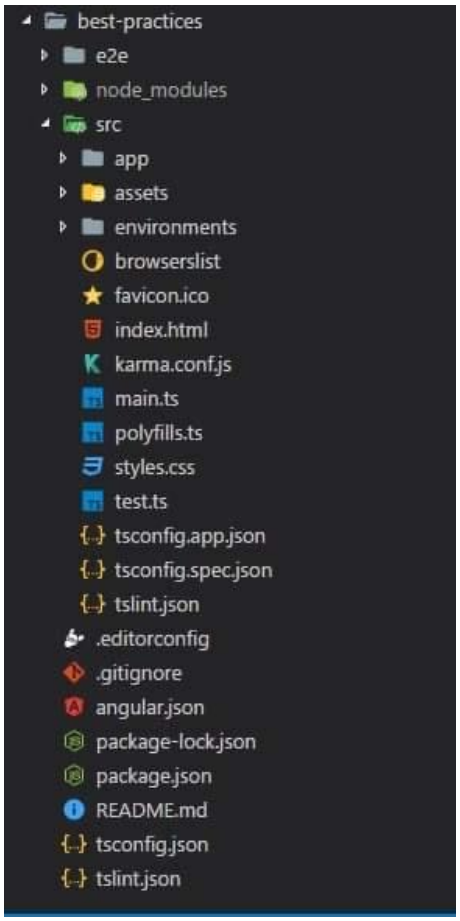

Figura 5.4. Carpetas del front-end
Fuente: Elaboración propia

En la carpeta src se ubicaron todas los archivos y configuraciones, en la subcarpeta componentes, se crearon las vistas de dashboard para cada partido político, 8 en total. Asi mismo, se utilizarán para los grafico la librería Chart.JS. Por tema de tiempo no desplegara en un sitio web. Solo se realizaron las pruebas en un servidor local. También se ha insertado librerías para que sea responsivo (adaptable a cualquier dispositivo).

La ventana inicial es la de logeo, la cual se muestra antes de acceder al sistema (Figura 5.5) Para ingresar al sistema el usuario necesita contar con sus credenciales: usuario y contraseña

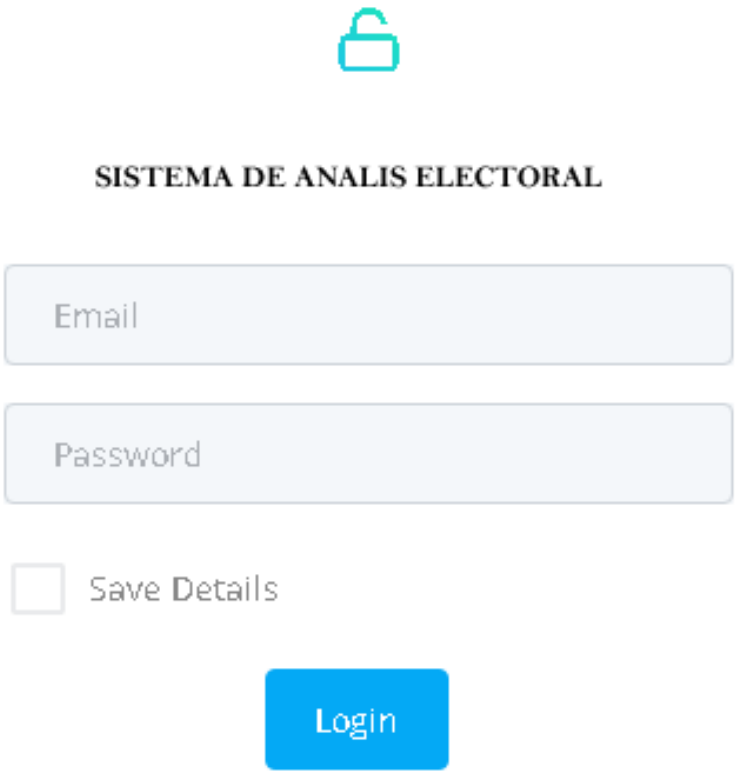

Figura 5.5. Pantalla del logeo del sistema
Fuente: Elaboración propia.

El sistema también tiene un listado de usuarios, registro actualización y eliminación de los mismos, los cuales sirven para el acceso y el control de la aplicación. En la figura 5.6 se muestra una pantalla para el registro de los usuarios.

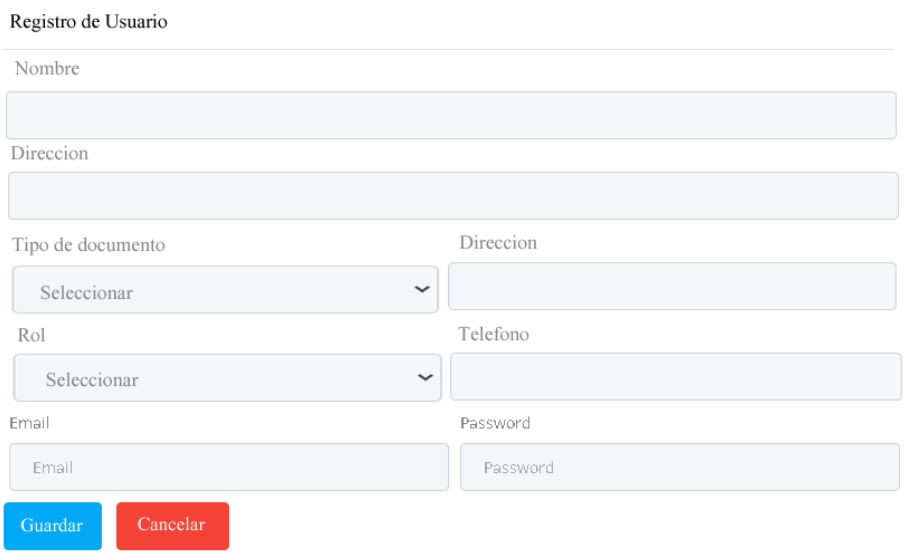

Figura 5.6. Registro de Usuario

Fuente: Elaboración propia

Se desarrollo un Dashboard con los resultados por cada partido político. En el Dashboard del partido de Acción Popular de un total de 474 tuits analizados se puede ver que la clasificación del sistema es la siguiente; 83 tuits positivos,298 tuits negativos y 93 tuits neutros. El día con más publicaciones fue el 19/01/2020. El usuario con más publicaciones fue AccionPopular con 31 tuits. También se ha obtenido las menciones de los usuarios por polaridad juntamente con la fecha de su publicación. En la Figura 5.7 se muestra dichos resultados.

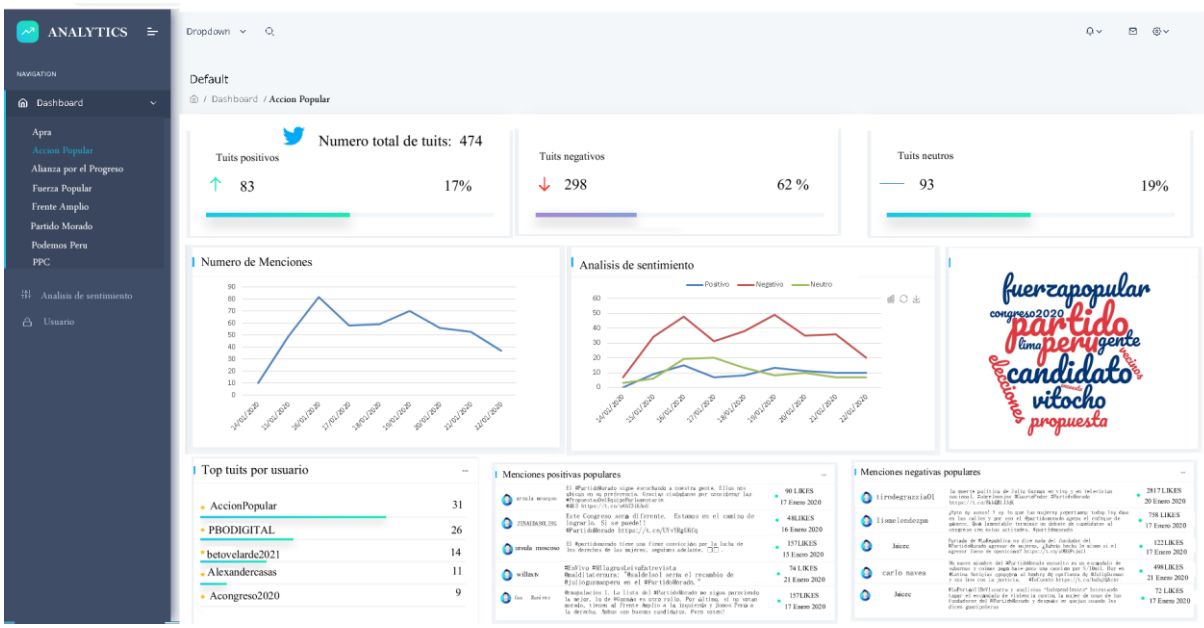

Figura 5.7. Dashboard de Acción Popular

Fuente: Elaboración propia

En el Dashboard del partido de Alianza por el progreso de un total de 132 tuits analizados se puede ver que la clasificación del sistema es la siguiente; 24 tuits positivos,54 tuits negativos y 54 tuits neutros. El día con más publicaciones fue el 19/01/2020. El usuario con más publicaciones fue PeruPressTv con 4 tuits. También se ha obtenido las menciones de los usuarios por polaridad juntamente con la fecha de su publicación. En la Figura 5.8 se muestra dichos resultados.

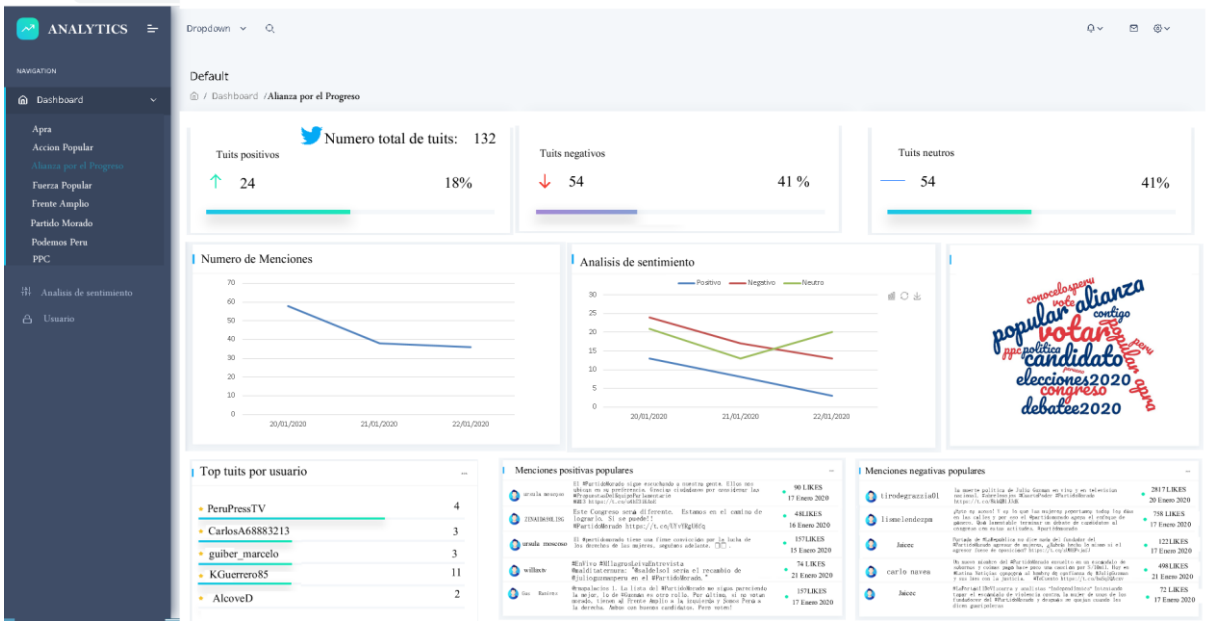

Figura 5.8. Dashboard Alianza por el Progreso

Fuente: Elaboración propia

En el Dashboard del partido Aprista por el progreso de un total de 628 tuits analizados se puede ver que la clasificación del sistema es la siguiente; 148 tuits positivos,397 tuits negativos y 83 tuits neutros. El día con más publicaciones fue el 20/01/2020. El usuario con más publicaciones fue Winston_Aguilar con 10 tuits También se ha obtenido las menciones de los usuarios por polaridad juntamente con la fecha de su publicación. En la Figura 5.9 se muestra dichos resultados.

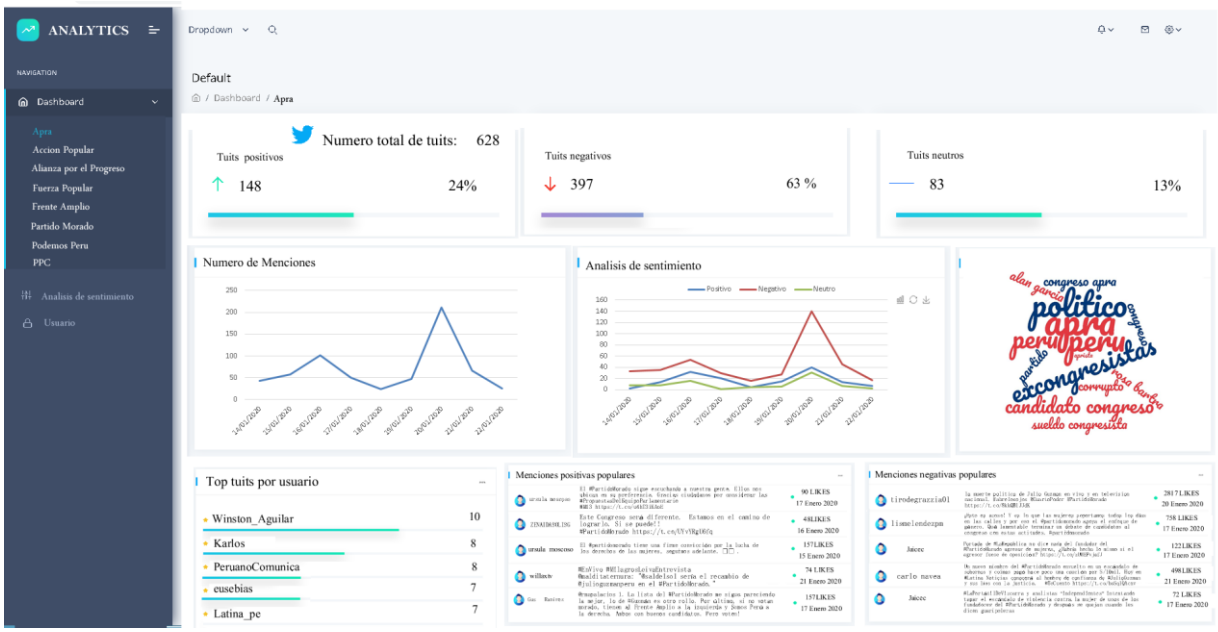

Figura 5.9. Dashboard Apra
Fuente: Elaboración propia

En el Dashboard del partido Frente Amplio por el progreso de un total de 487 tuits analizados se puede ver que la clasificación del sistema es la siguiente; 188 tuits positivos,205 tuits negativos y 94 tuits neutros. El día con más publicaciones fue el 17/01/2020. El usuario con más publicaciones fue maritabeti con 17 tuits. También se ha obtenido las menciones de los usuarios por polaridad juntamente con la fecha de su publicación. En la Figura 5.10 se muestra dichos resultados.

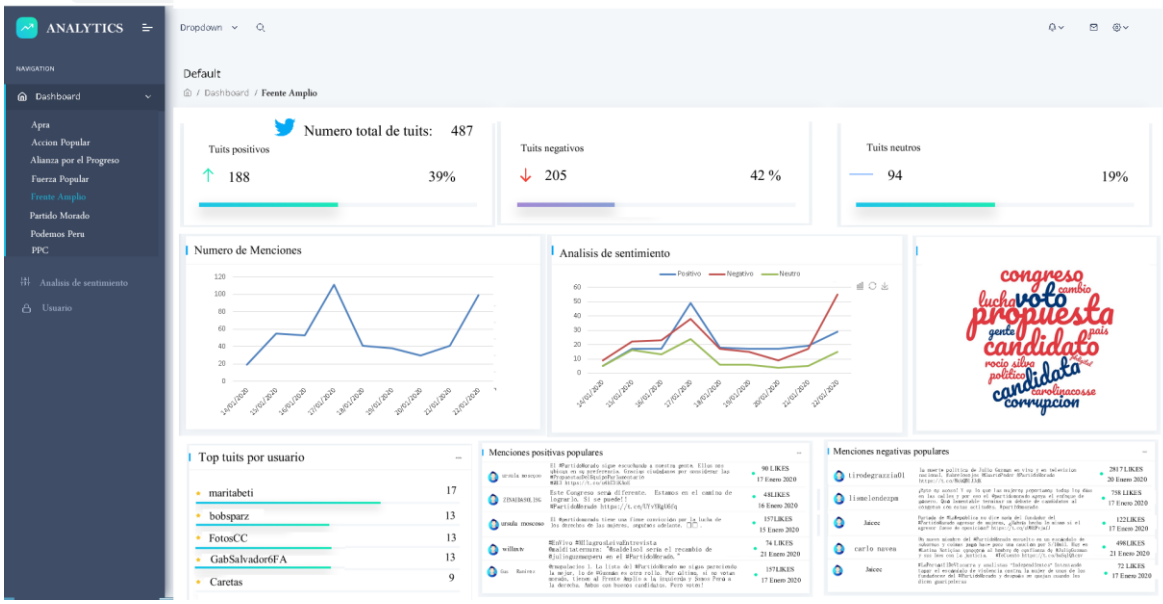

Figura 5.10. Dashboard Frente Amplio
Fuente: Elaboración propia

En el Dashboard del partido Fuerza Popular por el progreso de un total de 609 tuits analizados se puede ver que la clasificación del sistema es la siguiente; 243 tuits positivos,277 tuits negativos y 89 tuits neutros. El día con más publicaciones fue el 17/01/2020. El usuario con más publicaciones fue fplimanorte con 26 tuits. También se ha obtenido las menciones de los usuarios por polaridad juntamente con la fecha de su publicación. En la Figura 5.11 se muestra dichos resultados.

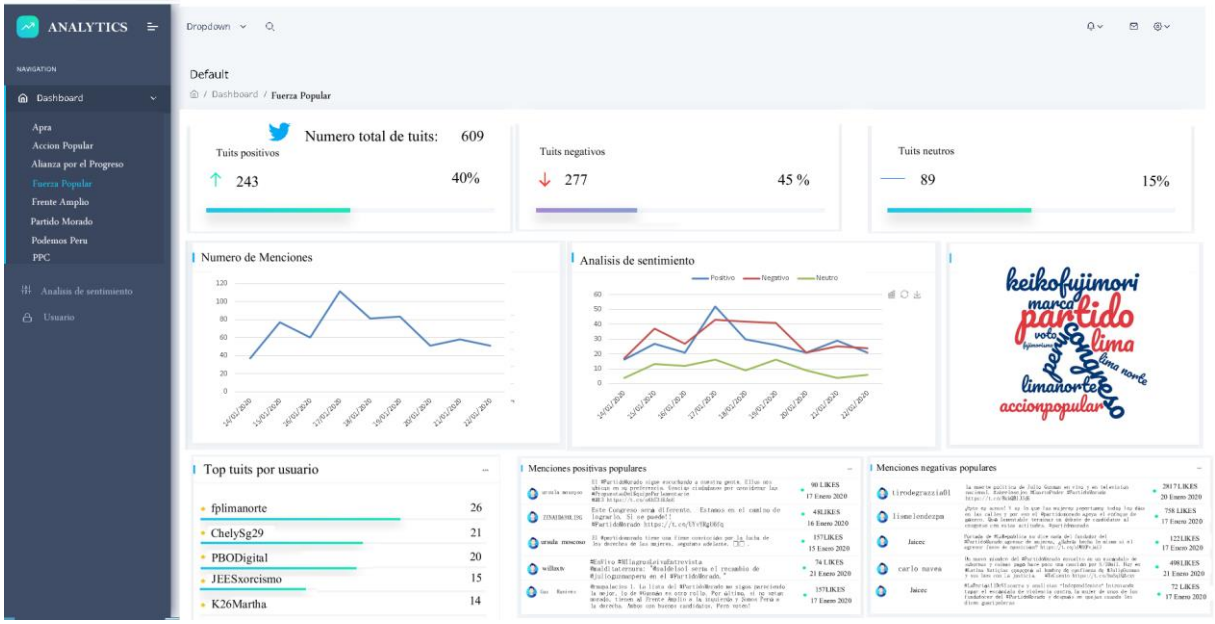

Figura 5.11. Dashboard Fuerza Popular
Fuente: Elaboración propia.

En el Dashboard del partido Morado por el progreso de un total de 1602 tuits analizados se puede ver que la clasificación del sistema es la siguiente; 421 tuits positivos,999 tuits negativos y 182 tuits neutros. El día con más publicaciones fue el 20/01/2020. El usuario con más publicaciones fue hersegami con 91 tuits. También se ha obtenido las menciones de los usuarios por polaridad juntamente con la fecha de su publicación. En la Figura 5.12 se muestra dichos resultados.

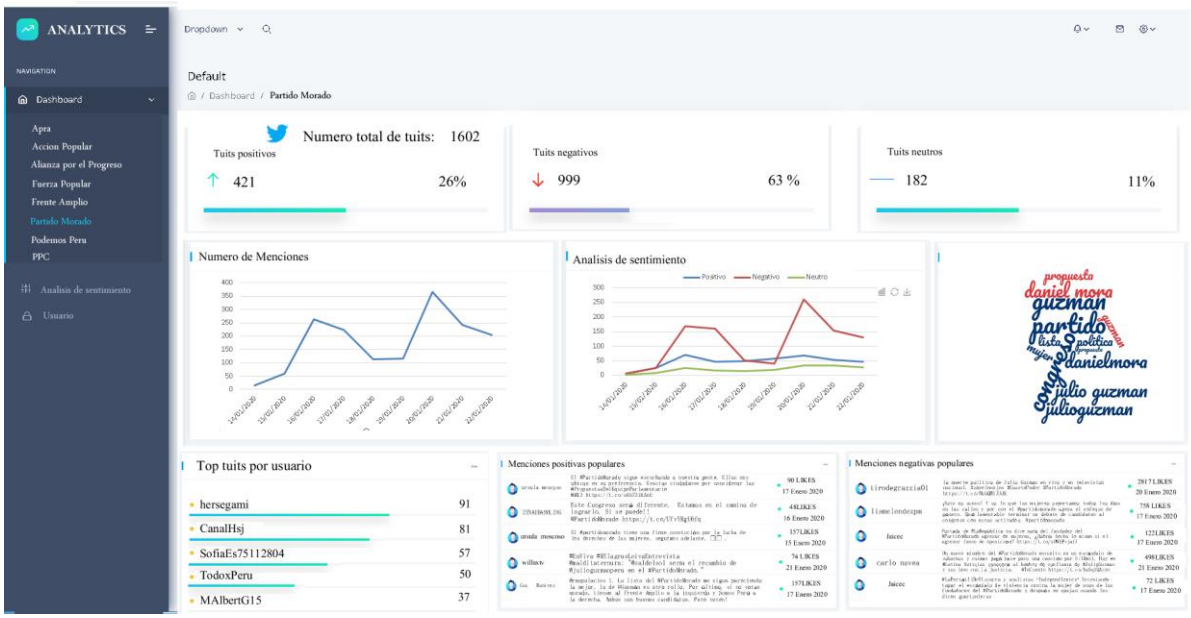

Figura 5.12. Dashboard Partido Morado

Fuente: Elaboración propia.

En el Dashboard del partido Podemos Perú por el progreso de un total de 44 tuits analizados se puede ver que la clasificación del sistema es la siguiente; 17 tuits positivos,13 tuits negativos y 14 tuits neutros. Los días con más publicaciones fueron el 16,17 de enero de 2020. El usuario con más publicaciones fue CeciliaTV con 8 tuits. También se ha obtenido las menciones de los usuarios por polaridad juntamente con la fecha de su publicación. En la Figura 5.13 se muestra dichos resultados.

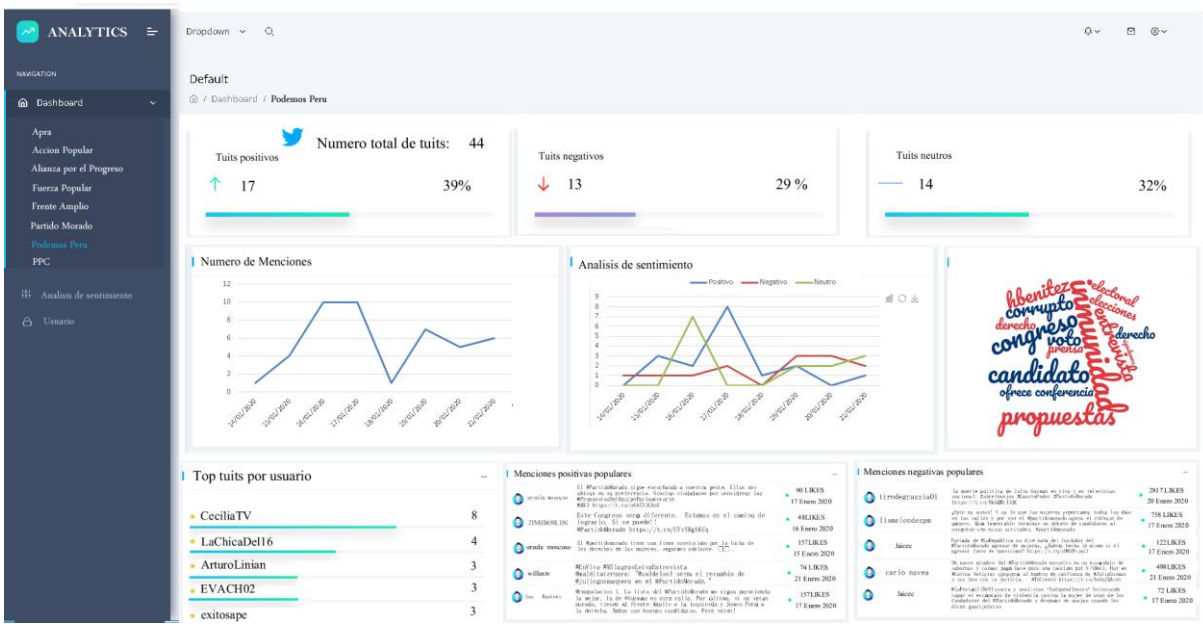

Figura 5.13. Dashboard Podemos Perú

Fuente: Elaboración propia.

En el Dashboard del partido Popular Cristiano de un total de 474 tuits analizados se puede ver que la clasificación del sistema es la siguiente; 254 tuits positivos,144 tuits negativos y 76 tuits neutros. El día con más publicaciones fue el 17/01/2020. El usuario con más publicaciones fue ppc_peru con 30 tuits. También se ha obtenido las menciones de los usuarios por polaridad juntamente con la fecha de su publicación. En la Figura 6.14 se muestra dichos resultados.

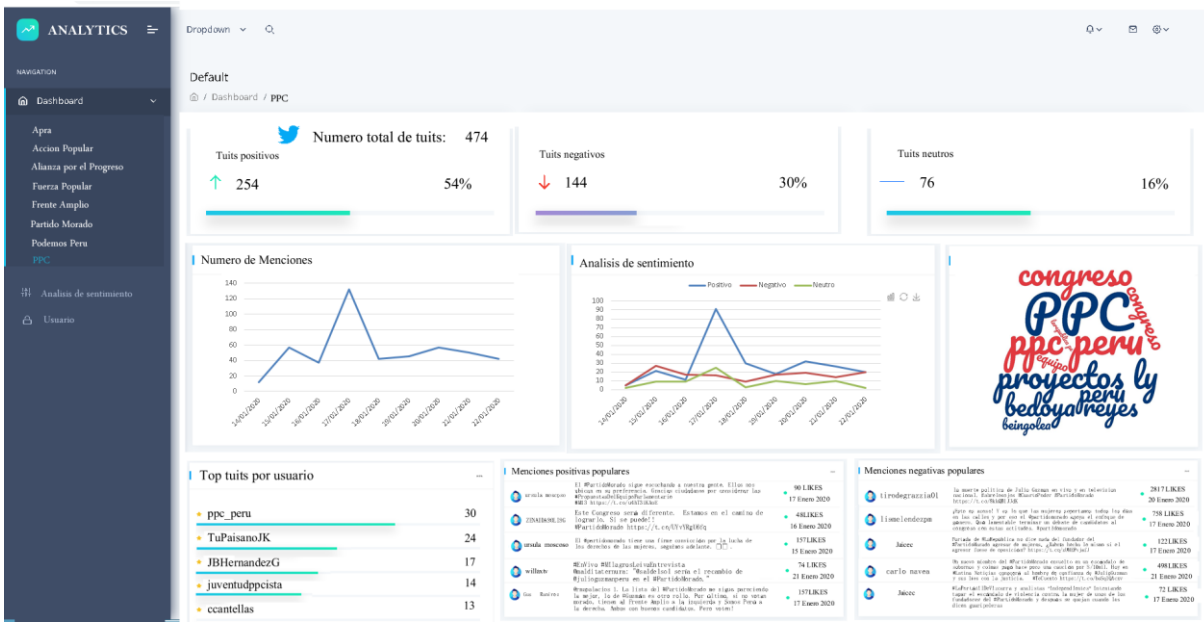

Figura 5.14. Dashboard PPC
Fuente: Elaboración propia.

Esta aplicación también cuenta con un analizador de sentimientos que es la versión almacenada del modelo RNC, cuya implementación está en una API. En este podemos ver el análisis de una frase política en particular. Al presionar el botón *Analizar* el sistema te dará un resultado que puede ser positivo, negativo y neutral (Figura 6.15).

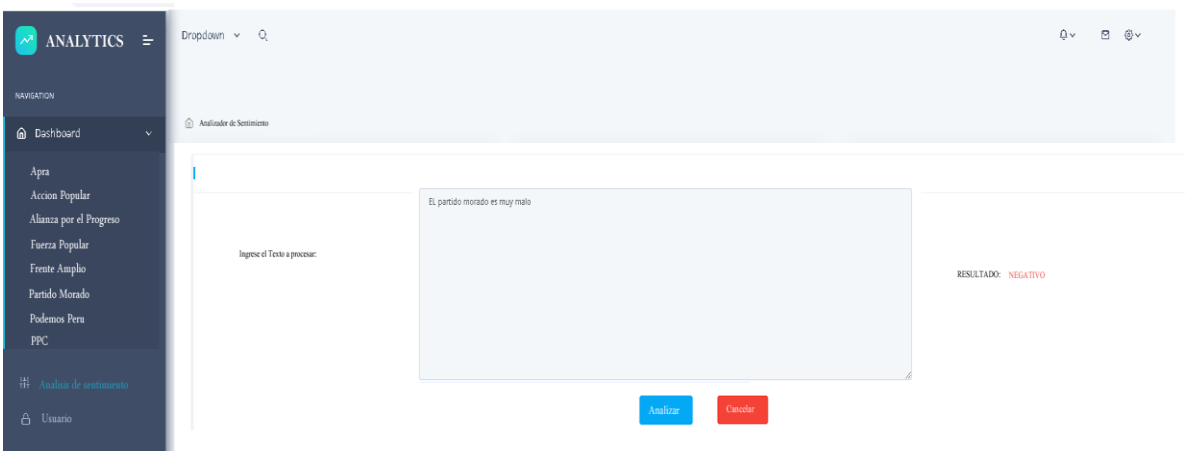

Figura 5.15. Analizador de Sentimientos
Fuente: Elaboración propia.

5.4. Evaluación

En este apartado se evaluará tanto el desarrollo del algoritmo como lo de la herramienta web.

5.4.1. Evaluación del Algoritmo

Para evaluar el algoritmo se utilizaron todos los corpus mencionados en capítulos anteriores. Para la evaluación se utilizó los siguientes algoritmos y librerías: Red neuronal convolucional, Naive Bayes, Regresión lineal, Maquinas vectores de Soporte, Textblob, se evaluó cada uno de los modelos por separado.

5.4.2. Medidas de Evaluación

En esta sección se analiza los resultados de la evaluación de los algoritmos. Para la elección de un modelo se debe elegir el que tenga el mejor F1-score, tomándose los valores máximos que obtuvo cada modelo.

Se puede observar que el modelo de la Red neuronal convolucional tiene el mayor F1- score con 0.94. Por otro lado, el modelo Naive Bayes se encuentra en el segundo lugar. En relación al tercer lugar, se puede ver al modelo SVM. Y en último lugar, el modelo Random Forest no es muy utilizado para problemas de NLP. Los algoritmos han sido probados por el entorno de Google Colab ya que te permite usar muchas librerías de manera online. En este proceso se ha aplicado toda la fase de preprocesado de datos y otras características. En la Tabla 6.5 se puede se aprecian los valores máximos de cada modelo.

Tabla 5.5. Comparativa de métricas

Modelo o Librería.	Acuracy	F1	Precision	Recall
SVM	0.87	0.86	0.87	0.86
RNC	0.94	0.94	**0.94**	0.94
Naive Bayes	0.87	0.86	0.88	0.87
Random Forest	0.62	0.57	0.63	0.62
Textblob	No se pudo evaluar la herramienta debido a problemas técnicos de la misma			

Fuente: Elaboración Propia

Matriz de Confusión y curva Roc de los modelos propuestos.

La matriz de confusión es una métrica que se utiliza en algoritmos de aprendizaje supervisado con la finalidad de permitir una mejor visualización del desempeño del modelo. Para este caso se ha aplicado la matriz de confusión a la red neuronal convolucional del cual se puede apreciar la predicción para cada clase, siendo la clase N (negativa), NEU (neutral) y P (positiva).

Los datos de entrenamiento con el cual se obtuvo el mejor escore fueron:

Para la clase positiva los datos clasificados correctamente (True Positivos) son 2061, siendo clasificados incorrectamente 42 datos para la clase neutra y 38 clasificados incorrectamente para la clase negativa, los cuales sumados son 50 (False Positives).

Para la clase neutra los datos clasificados correctamente (True Positivos) son 1677, siendo clasificados incorrectamente 48 datos para la clase positiva y 69 clasificados incorrectamente para la clase negativa, los cuales sumados son 117 (False Positives).

Para la clase negativa los datos clasificados correctamente (True Positivos) son 3178, siendo clasificados incorrectamente 85 datos para la clase neutra y 45 clasificados incorrectamente para la clase positiva, los cuales sumados son 130 (False Positives).

En la Figura 5.16 se muestra el detalle de la matriz de confusión de este modelo (RNC).

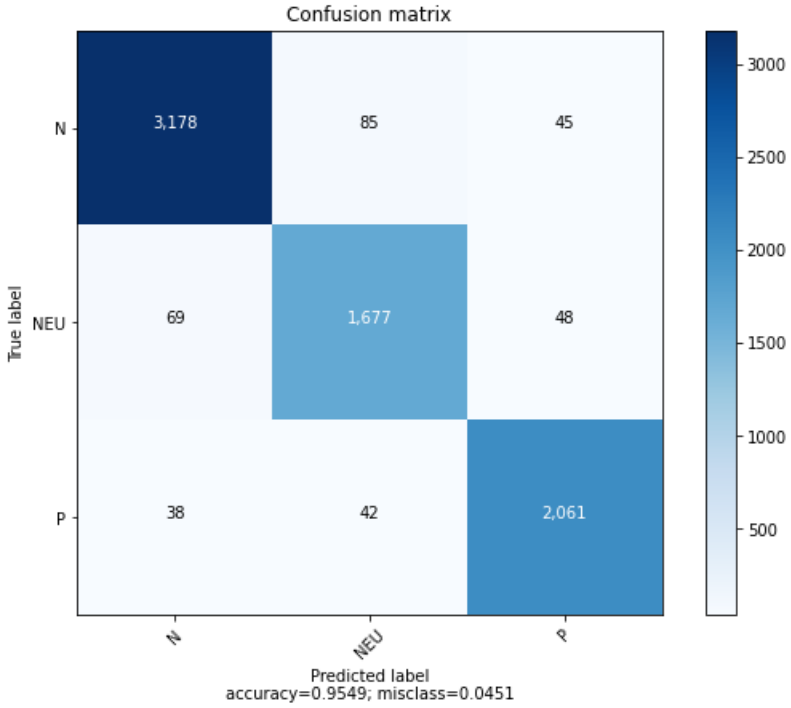

Figura 5.16 Matriz de Confusión para RNC

Fuente: Elaboración Propia

La curva de característica operativa del receptor (ROC) es un gráfico que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. Para tener una mejor idea la curva cuando tiene valores de

- De 0.5 a 0.6: Es un valor malo
- De 0.6 a 0.75 es un valor regular
- De 0.7 a 0.97 es un valor bueno.
- De 0.97 a 1 es un valor excelente.

En la figura se puede ver que la curva ROC está por encima de un valor bueno (0.7). Cuando el valor es más cercano a uno mejor es la predicción de la clase. En este caso el valor obtenido es 0.99 Un valor óptimo para la predicción. El algoritmo con mayor exactitud ha sido utilizado para predecir el contenido de los tuits. En la Figura 5.17. se muestra dicho resultado.

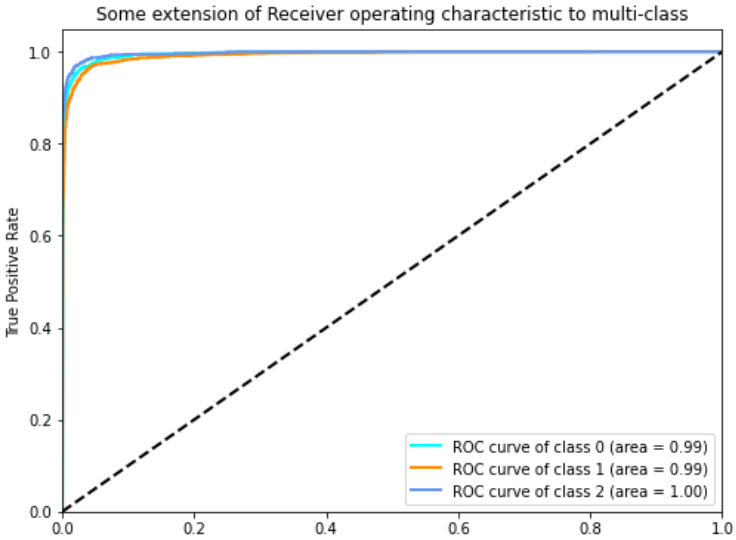

Figura 5.17. Curva ROC Modelo RNC

Fuente: Elaboración Propia

5.4.3. Evaluación de la herramienta

En la siguiente tabla se realiza la comparación de esta solución con las del mercado, los dashboard de las aplicaciones analizadas se encuentran en el Anexo 01. Las funcionalidades a comparar son las siguientes (Tabla 5.6).

Tabla 5.6. Comparativa de funcionalidades

Funcionalidad	Social Search	Brand24	Analytics
Logeo	Si	Si	Si
Grafica por Menciones	No	Si	Si
Nube de Palabras	Si	Si	Si
Top de Usuarios	Si	Si	Si
Múltiples Vistas	No	Si	Si
Analizador de sentimiento	No	No	Si

Línea de tiempo de tuits neutros	No	No	Si
Línea de tiempo de tuits positivos y negativos	Si	Si	Si
Filtro por fecha	SI	Si	No
Tuits por región	No	No	No
Tuits por dispositivo	Si	No	No
Es responsivo	Si	Si	Si

Fuente: Elaboración Propia

La aplicación cuenta con un analizador de sentimiento, y careciendo de un filtro por fecha. y de tuits por dispositivo.

6. Conclusiones y trabajo futuro

La conclusión de este TFM describe como se cumplieron los distintos objetivos descritos en los capítulos anteriores.

6.1. Conclusiones

En la investigación realizada del análisis de sentimiento político en Twitter durante las elecciones congresales 2020 en Perú, se llegó a las siguientes conclusiones:

En el primer objetivo se logró investigar las soluciones para resolver el análisis de sentimiento en español de las cuales Social Searcher y Brand 24 son los mejores resultados de pago en el mercado, también se ha logrado identificar las librerías de código abierto que más se adaptan a la solución de la misma, de la cual se utilizaron Scikit-Leam, Pyspark.

En el segundo objetivo se logró extraer los datos de Twitter desde 01 octubre 2019 a 22 enero 2020 usando como parámetros de búsqueda las palabras clave relacionado con del partido político o candidato, fecha de publicación de tuit y que sea escrito en español, por otra parte, mediante la librería de pymongo se almaceno en la base de datos MongoDB, los campos más relevantes que trae la API de Twitter.

En el tercer objetivo se logró limpiar los datos de Twitter utilizando varias técnicas de preprocesamiento como identificación de emoji, caracteres no deseados, acentos, algunas jergas o contracciones de texto, para posteriormente aplicar la reducción de características como eliminación de stopwords, steaming y la tokenización del texto.

En el cuarto objetivo se identificaron 4 modelos de aprendizaje supervisado, Random Forest, Naive Bayes, SVM y redes neuronales convolucionales, los cuales fueron entrenados mediante un conjunto de corpus en español, por otra parte, se han utilizado tres clases de polaridad; positivo, negativo y neutro, mediante la metodología CRISP-DM se logró determinar que el algoritmo de redes neuronales de convolución tiene los mejores resultados para las métricas de evaluación que los demás modelos de SVM, Naive Bayes, Random Forest. Se logro almacenar dicho modelo para ser usado mediante una API y/o función de Phyton para analizar el sentimiento de los datos de los tuits. Asimismo, el reconocimiento del lenguaje ha sido a través de documento.

En el quinto objetivo cabe destacar también, que se desarrolló una aplicación basada en la tecnología MEAN, del cual se crearon un conjunto de endpoints que fueron consumidos por Angular para el dashboard de cada partido político, que permite visualizar la polaridad de los sentimientos por fecha, el número de menciones y los usuarios top con más publicaciones,

juntamente con las publicaciones negativas y positivas top (las que tienen más likes).y esta tiene una página de logeo del cual se necesita contar con un usuario para su acceso.

En último término, el reto de aprender las tecnologías de Pyspark, Phyton, Node, Express, Angular, MongoDB y otras librerías ha sido un esfuerzo grande en este TFM

En resumen, se desarrolló un sistema para el análisis de sentimiento políticos en Twitter referente a las elecciones congresales de 2020.

6.2. Líneas de trabajo futuro

Para las mejoras de este TFM se podría considerar los siguientes puntos:

- Se puede elaborar un corpus de emociones, calificando a cada frase con una emoción determinada. Por ejemplo, se pude reconocer una emoción primaria (miedo, alegría) como las descritas en este TFM para una oración determinada. Las emociones expresan sentimientos.

- Para mejorar este sistema, se puede pueden guardar las palabras clave de los partidos políticos, generar la consulta a través del streaming API de Twitter y guardarlas en la Base de datos mientras la data es actualizada y transformada por el algoritmo de aprendizaje automático. Así se tendría la información predictiva en tiempo real, junto con la información histórica.

- Es importante tener un mejor impacto en las predicciones de las redes sociales, se deben consideran las demás redes predominantes Facebook y YouTube. Estas redes tienen sus propias APIS de las cuales se pueden extraer datos para el análisis de sentimiento.

- En lo que respecta mejorar el análisis de sentimiento también se pueden utilizar los corpus de imágenes y así abarcar mucho más las diferentes entradas de las publicaciones de las redes sociales, tanto para reconocer emociones, así como las imágenes de texto.

- Cabe destacar también, que se puede entrenar el algoritmo para que detecte el sarcasmo e ironía. Esta es una tarea desafiante del NLP. Debido a que el sarcasmo puede cambiar la polaridad de una oración y, por lo tanto, afectar negativamente el rendimiento de detección de la polaridad. Para esto se debe fusionar el conocimiento lingüístico con el aprendizaje automático.

- Es necesario revisar de alguna manera el spam de opinión para garantizar que las opiniones en la web o redes sociales sean fuentes confiables de información valiosa.

7. Bibliografía

Agarwal, B., & Mittal, N. (2016). Prominent Feature Extraction for Sentiment Analysis. London: Springer. doi:10.1007/978-3-319-25343-5

Amat Rodrigo, J. (octubre de 2020). cienciadedatos.net. Recuperado de https://www.cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_c50

Bayo, C. E. (11 de diciembre de 2015). Público. Recuperado de https://www.publico.es/actualidad/arma-total-obama-vencer-romney.html#ixzz2BWjzUjgG

BBC Mundo. (21 de marzo de 2018). BBC Mundo. Recuperado de https://www.bbc.com/mundo/noticias-america-latina-43481060

Berlind, D. (01 de diciembre de 2016). ProgrammableWeb. Recuperado de https://www.programmableweb.com/api/intellexer

Caetano, J., Lima, H., Santos, M. et al. Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election. J Internet Serv Appl 9, 18 (2018). doi:10.1186/s13174-018-0089-0

Calderón, L. (13 de octubre de 2019). Quiénes ganaron con la crisis política. El Diario, pág. 2.

Camacho, A. (04 de mayo de 2020). jacobsoft.com.mx. Recuperado de https://www.jacobsoft.com.mx/es_mx/clasificador-naive-bayes/

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). A Practical Guide to Sentiment Analysis (Vol. 5). Cham: Springer International Publishing AG.

Carvallo, F. (29 de julio 2019). Estamos solo para decidir si lo que plantea Vizcarra es lo que Perú merece [Audio podcast]. Recuperado de https://open.spotify.com/episode/65IAJA8R7mFUWVHediQsN7?si=tWzMOMQ2SB-GVOo1GeuusQ

Chillitupa, R. (2019). Fuji Sustos. Caretas, 17-18.

Congreso de la Republica. (s.f.). Parlamento Abierto. Recuperado de http://parlamentoabierto.pe/congreso.html

El Comercio. (30 de septiembre de 2019). El Comercio. Recuperado de https://elcomercio.pe/politica/vizcarra-oficializa-disolucion-del-congreso-y-convoca-a-elecciones-para-enero-de-2020-noticia/

Elghazaly, Tarek & Mahmud, Amal & Hefny, Hesham. (2016). Political Sentiment Analysis Using Twitter Data. 1-5. doi:10.1145/2896387.2896396

Fernández, R. (09 de febrero de 2021). Statista. Recuperado de https://es.statista.com/estadisticas/636174/numero-de-usuarios-mensuales-activos-de-twitter-en-el-mundo/

Fowks, J. (01 de octubre de 2019). EL PAIS. Recuperado de https://elpais.com/internacional/2019/10/01/america/1569885710_959879.html

Gerrod Parrott, W. (2001). Emotions in Social Psychology. Philadelphia: Psychology Press.

Gobierno del Peru. (20 de septiembre de 2018). Gobierno del Peru. Recuperado de https://www.gob.pe/estado/

Kušen, E., & Strembeck, M. (2017). Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. Online Social Networks and Media, 49. doi:10.1016/j.osnem.2017.12.002

Liu, B. (2015). Sentiment Analysis. New York: Cambridge University Press.

Malik, Ayeena & Kapoor, Divya & Singh, Amit. (2016). Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election. doi:10.1186/s13174-018-0089-0

Maldonado, A. (22 de mayo de 2017). Parlamento Abierto. Recuperado de http://parlamentoabierto.pe/partidos.html

Méndez, F. (03 de septiembre de 2015).FORBES. Recuperado de https://forbes.es/emprendedores/7560/como-el-big-data-ayudo-a-obama-a-ganar/

Naupari, M. (03 de octubre de 2018). RPP NOTICIAS. Recuperado de https://rpp.pe/politica/judiciales/video-por-que-alberto-fujimori-fue-condenado-a-25-anos-de-prision-noticia-955392

Pierina, P. B. (28 de julio de 2016). BBC Mundo. Recuperado de https://www.bbc.com/mundo/noticias/2016/06/160531_america_latina_peru_elecciones_fujimorismo_decisivo_ppb

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017). Sentiment Analysis in Social Networks. Cambridge: Elsevier Inc.

Quiroz, A. W. (2019). Historia de la Corrupción en el Peru. Lima: IEP INSTITUTO DE ESTUDIOS PERUANOS.

Rao, P. (04 de septiembre de 2019). Towards Data Science. Recuperado de https://towardsdatascience.com/fine-grained-sentiment-analysis-in-python-part-1-2697bb111ed4

Rogalski, K. (03 de marzo de 2019). Brand24. Recuperado de https://brand24.com/blog/best-sentiment-analysis-tools/

Romero, J. (11 de junio de 2019). Jorge Romero. Recuperado de https://jorgeromero.net/acerca-jorge-romero/

Tendencias Digitales. (01 de septiembre de 2017). Tendencias Digitales. Recuperado de https://tendenciasdigitales.com/internet-y-los-medios-sociales-en-peru/

Vásquez Torres, J., & Joyanes Aguilar, L. (2018). Tendencias, Oportunidades y Retos del uso de las Redes Sociales en Latinoamerica: Caso Centroamerica y Panamá. 6th Engineering, Science and Technology Conference (2017) (pág. 11). Panama: Knowledge E.

Verdier, D., & Rocha, M. (29 de octubre de 2019). El País. Recuperado de: https://www.elpais.com.uy/informacion/politica/elecciones-uruguay-analisis-sentimiento-redes-sociales-resultado-inteligencia-artificial.html

Vilca, & Vilca, K. (21 de marzo de 2020). Quantico. Recuperado de https://www.quanticotrends.com/blog/quanticotrends/como-esta-compuesto-hoy-el-universo-de-twitter-en-peru/

Villena Roman, J., Lana Serrano, S., Martínez Cámara, E., & González Cristóbal, J. (21 de marzo de 2013). SEPLN. Recuperado de http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4657

8. Anexos

Anexo I. Resultado de las búsquedas por las herramientas comerciales y/o gratuitas

Se utilizaron las herramientas comerciales Brand24, Social Search, sentiment viz y como punto de referencia, para obtener nuestros resultados.

1.Partido Acción Popular.

Se analizó a la fecha de 22/01/2020 usando la herramienta Brand24, realizando la búsqueda con la palabra #AccionPopular obteniendo el siguiente resultado (Figura 8.1);27 tuits, de los cuales 10 fueron positivos y 7 negativos.

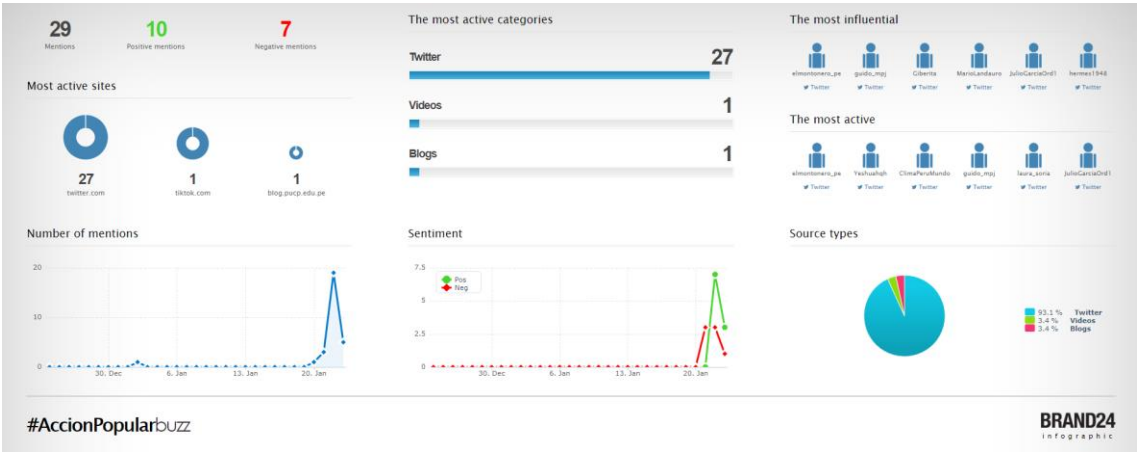

Figura 8.1. Resultado del Análisis de Sentimiento con la Herramienta Brand24 del Partido Político Acción Popular.

Se analizó, un segundo resultado a la fecha de 22/01/2020 usando la herramienta Social Search, realizando la búsqueda con la palabra #AccionPopular obteniendo el siguiente resultado (Figura 8.2); 174 tuits, de los cuales 21 fueron positivos y 34 negativos.

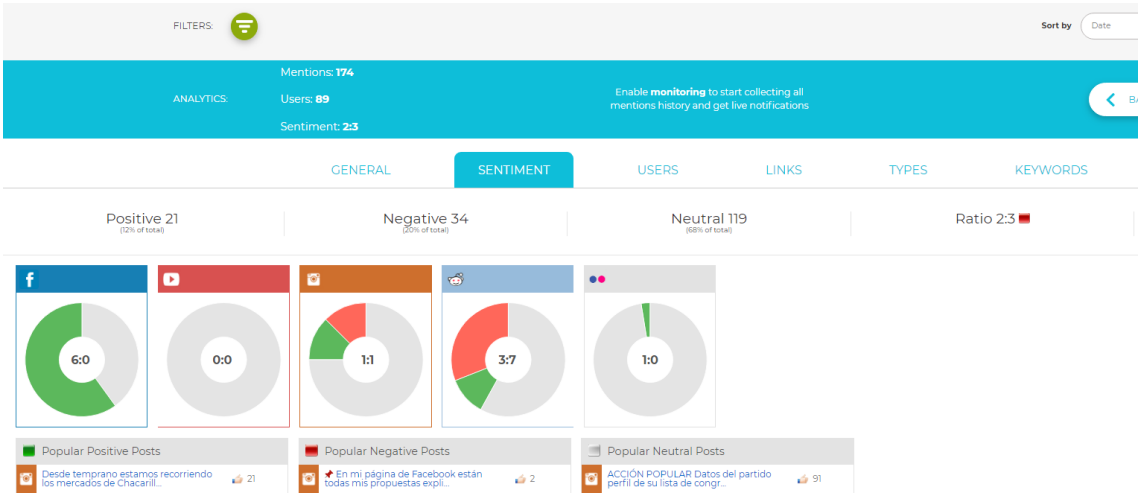

Figura 8.2. Resultado del Análisis de Sentimiento con la Herramienta Social Search del Político Acción Popular.

2. Partido Alianza por el progreso.

Se analizó a la fecha de 22/01/2020 usando la herramienta Brand24, realizando la búsqueda con la palabra Alianza por el Progreso obteniendo el siguiente resultado (Figura 8.3); 4 tuits, de los cuales 3 fueron positivos y 1 negativo.

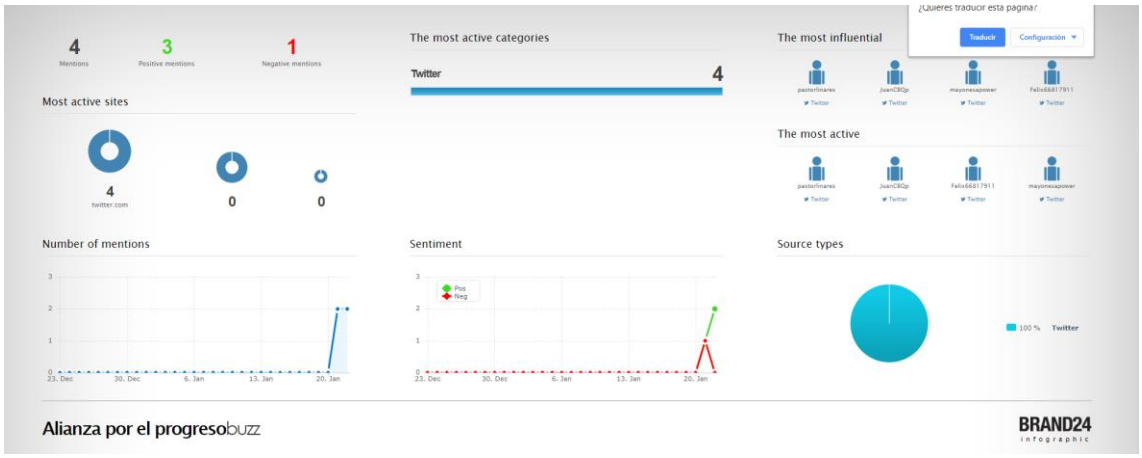

Figura 8.3. Resultado del Análisis de Sentimiento con la Herramienta Brand24 del Partido Político Alianza por el Progreso.

Así mismo, se aplicó a un segundo resultado con fecha de 22/01/2020 usando las la herramienta Social Search, realizando la búsqueda con la palabra Alianza por el Progreso obteniendo el siguiente resultado (Figura 8.4); 584 tuits, de los cuales 96 fueron positivos y 92 negativos.

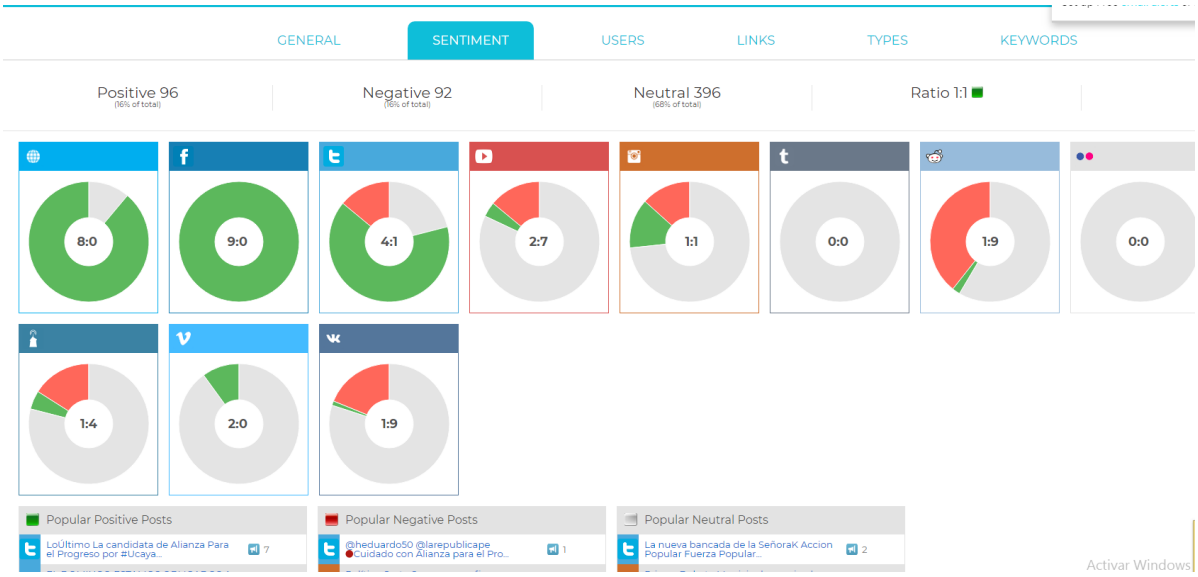

Figura 8.4. Resultado del Análisis de Sentimiento con la Herramienta Social Search del Partido Político Partido Político Alianza por el Progreso.

3.Partido Aprista Peruano.

Se analizó a la fecha de 22/01/2020 usando la herramienta Brand24, analizando la búsqueda con la palabra Mauricio Mulder que es el principal candidato de este partido, obteniendo el siguiente resultado (Figura 8.5); 84 tuits, de los cuales 17 fueron positivos y 25 negativos. No se buscó por la palabra #Apra debido a que no se dieron resultados óptimos.

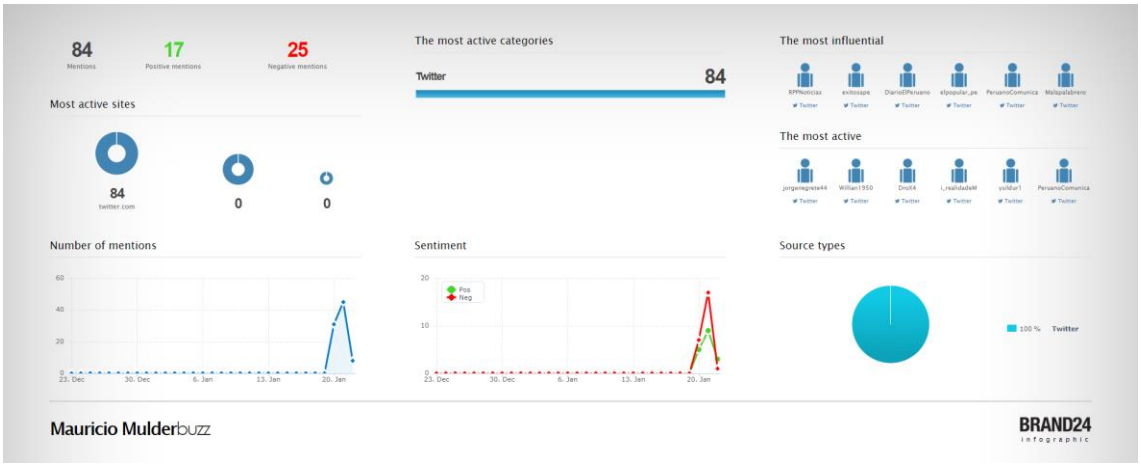

Figura 8.5. Resultado del Análisis de Sentimiento con la Herramienta Brand24 del Partido Político Aprista.

Así mismo, se aplicó a un segundo resultado con fecha de 22/01/2020 usando las la herramienta Social Search, realizando la búsqueda con la palabra Mauricio Mulder obteniendo el siguiente resultado (Figura 8.6); 385 tuits, de los cuales 26 fueron positivos y 40 negativos.

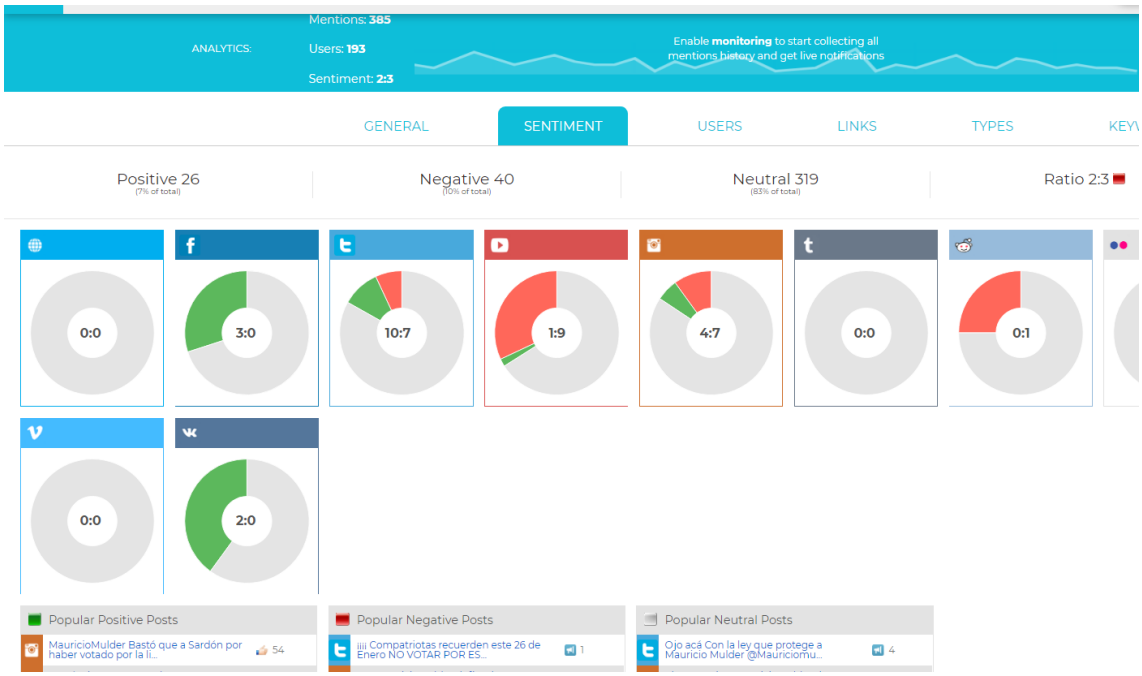

Figura 8.6. Resultado del Análisis de Sentimiento con la Herramienta Social Search del Partido Aprista.

4.Partido Frente Amplio.

Se analizó a la fecha de 22/01/2020 usando la herramienta Brand24.Se hizo la búsqueda por la palabra #FrenteAmplio obteniendo el siguiente resultado. (Figura 8.7); 98 tuits, de los cuales 47 fueron positivos y 15 negativos.

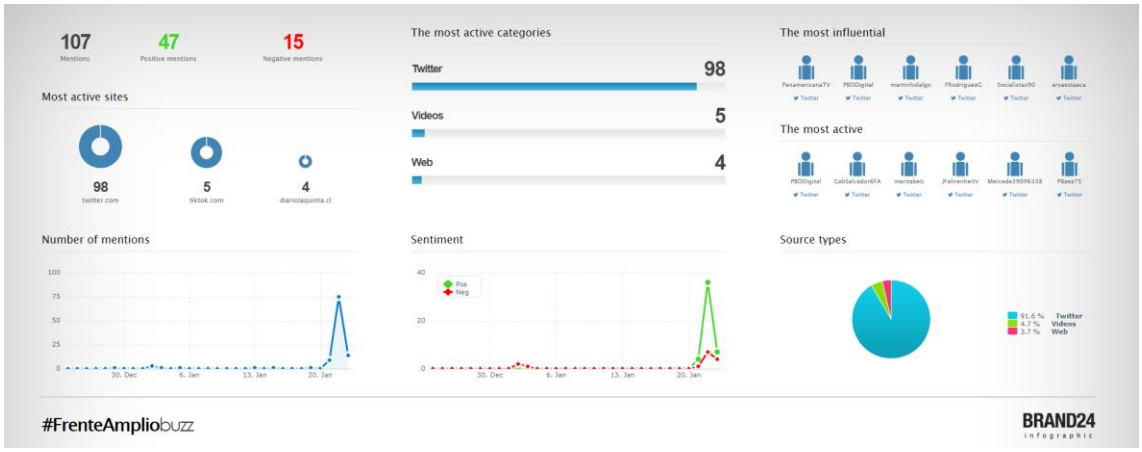

Figura 8.7. Figura: Resultado del Análisis de Sentimiento con la Herramienta Brand24 del Partido Político Frente Amplio.

Se analizó a la fecha de 22/01/2020 usando la herramienta Social Search. Se hizo la búsqueda por la palabra #FrenteAmplio obteniendo el siguiente resultado (Figura 8.8); 428 tuits, de los cuales 41 fueron positivos y 55 negativos.

Figura 8.8. Resultado del Análisis de Sentimiento con la Herramienta Social Search del Partido Frente Amplio.

5.Partido Fuerza Popular.

Se analizó a la fecha de 22/01/2020 usando la herramienta Brand24, realizando la búsqueda con la palabra #FuerzaPopular obteniendo el siguiente resultado (Figura 8.9); 100 tuits, de los cuales 37 fueron positivos y 16 negativos.

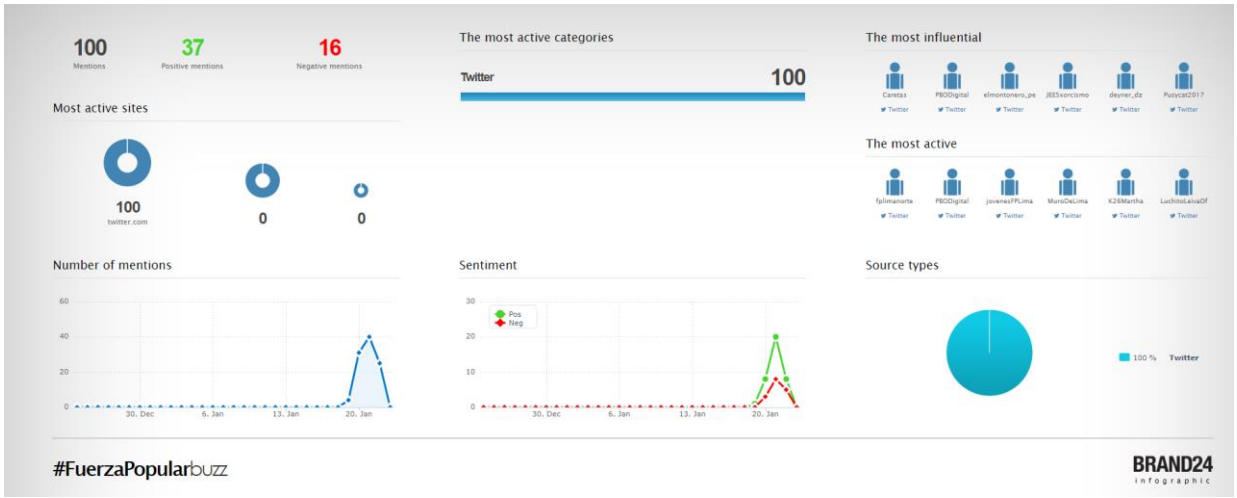

Figura 8.9. Resultado del Análisis de Sentimiento con la Herramienta Brand24 del Partido Político Fuerza Popular.

De igual modo, se analizó un segundo resultado a la fecha de 22/01/2020 usando la herramienta Social Search, realizando la búsqueda con la palabra #FuerzaPopular obteniendo el siguiente resultado (Figura 8.10); 318 tuits, de los cuales 94 fueron positivos y 27 negativos.

Figura 8.10. Resultado del Análisis de Sentimiento con la Herramienta Social Search del Partido Fuerza Popular.

6. Partido Podemos Perú.

Se analizó a la fecha de 22/01/2020 usando la herramienta Brand24.Se hizo la búsqueda por la palabra #PodemosPeru obteniendo el siguiente resultado (Figura 8.11); 34 tuits, de los cuales 16 fueron positivos y 4 negativos.

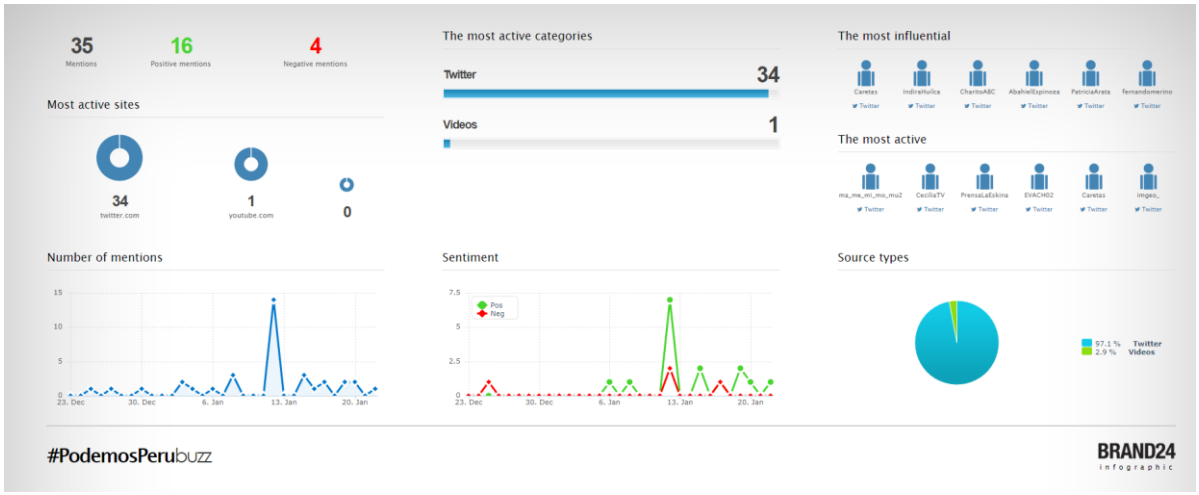

Figura 8.11. Resultado del Análisis de Sentimiento con la Herramienta Brand24 del Partido Político Podemos Perú.

Se analizó, un segundo resultado a la fecha de 22/01/2020 usando la herramienta Social Search, realizando la búsqueda con la palabra #PodemosPeru obteniendo el siguiente resultado (Figura 8.12); 157 tuits, de los cuales 7 fueron positivos y 11 negativos.

.

Figura 8.12. Resultado del Análisis de Sentimiento con la Herramienta Social Search del Partido Político Podemos Perú.

7.Partido Morado.

Se analizó a la fecha de 22/01/2020 usando la herramienta Brand24, realizando la búsqueda con la palabra #PartidoMorado obteniendo el siguiente resultado (Figura 8.13); 50 tuits, de los cuales 14 fueron positivos y 11 negativos.

Figura 8.13. Resultado del Análisis de Sentimiento con la Herramienta Brand24 del Partido Político Morado.

Se analizó, un segundo resultado a la fecha de 22/01/2020 usando la herramienta Social Search, realizando la búsqueda con la palabra #PartidoMorado obteniendo el siguiente resultado (Figura 8.14); 253 tuits, de los cuales 14 fueron positivos y 45 negativos.

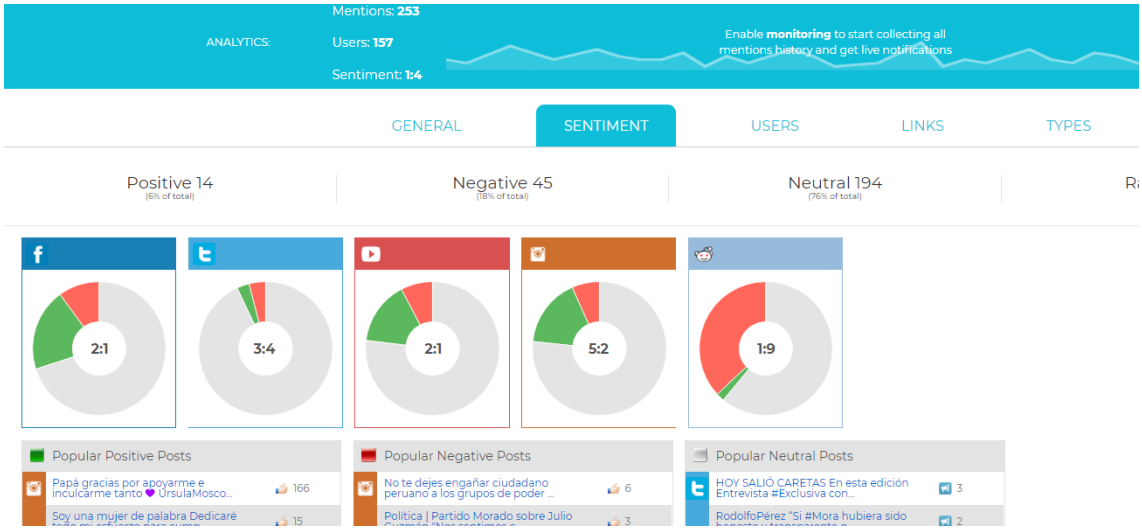

Figura 8.14. Resultado del Análisis de Sentimiento con la Herramienta Social Search del Partido Político Morado.

8.Partido Popular Cristiano.

Se analizó a la fecha de 22/01/2020 usando la herramienta Brand24, realizando la búsqueda con la palabra #PPC obteniendo el siguiente resultado (Figura 8.15); 99 tuits, de los cuales 52 fueron positivos y 10 negativos.

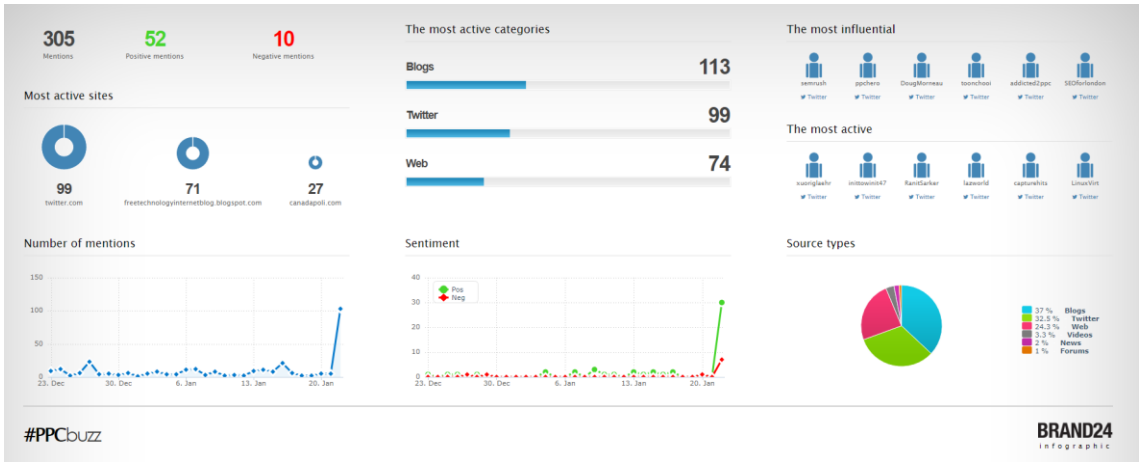

Figura 8.15. Resultado del Análisis de Sentimiento con la Herramienta Brand24 del Partido Político Popular Cristiano.

Se analizó, un segundo resultado a la fecha de 22/01/2020 usando la herramienta Social Search, realizando la búsqueda con la palabra #PPC obteniendo el siguiente resultado (Figura 8.16); 372 tuits, de los cuales 198 fueron positivos y 37 negativos.

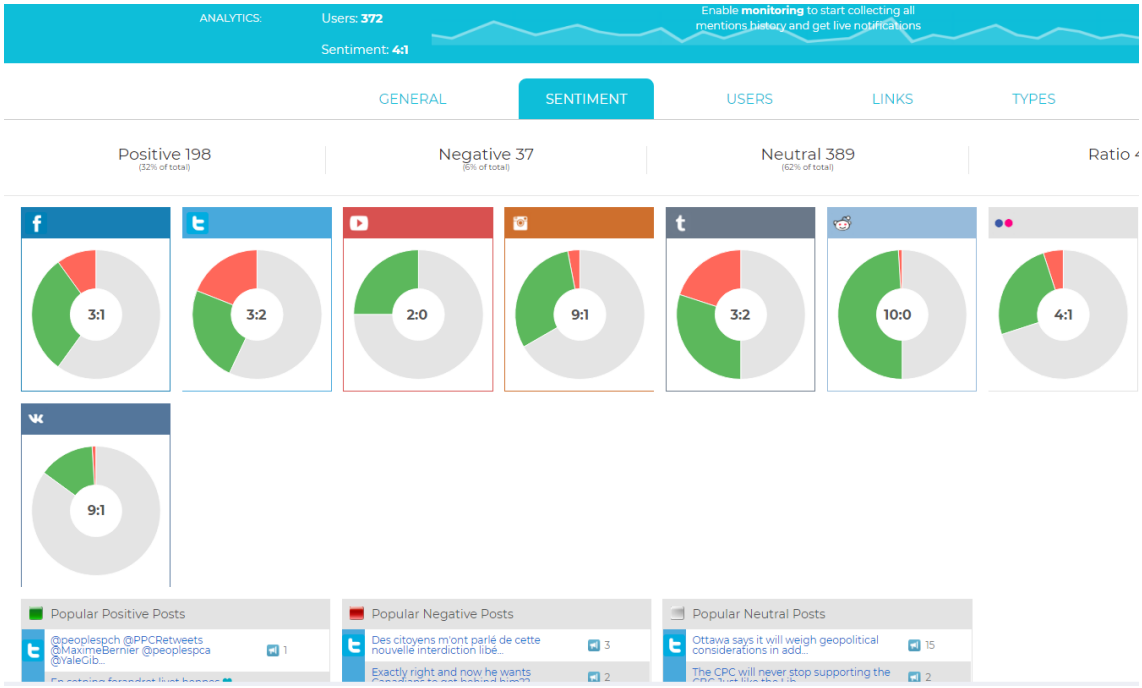

Figura 8.16. Resultado del Análisis de Sentimiento con la Herramienta Social Search del Partido Político Popular Cristiano

Anexo II. Código fuente

Los archivos de las pruebas realizadas en este TFM se encuentran desarrollados en Phyton versión 3.6. con ayuda de las siguientes librerías.

Tensorflow/Keras: Es una librería de código abierto soportado por Phyton para redes neuronales artificiales.

Plotly: Es una biblioteca gratuita de Phyton la cual crea gráficos interactivos.

Numpy: Es una biblioteca que sirve para dar soporte a arreglos y matrices. También incluye funciones matemáticas

OS: Es una biblioteca permite hacer operaciones con directorios y archivos temporales.

El código del proyecto y los dataset pueden ser consultados en Github.

https://github.com/danielalva2008/TFM

La estructura del proyecto es la siguiente.

- **Carpeta "DATA DE TWITTER":** Contiene la información extraída de la API Twitter en formato JSON para los partidos políticos.
- **Carpeta "DATASETS":** Contiene los corpus que se han usado en este proyecto para la evaluación de los algoritmos
- **Carpeta "SCRIPTS":** Contiene los scripts de Phyton descritos en este TFM para la evaluación de los algoritmos descritos en este proyecto.