ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

INTELIGENCIA ARTIFICIAL

II Término, 2021

INTEGRANTES

- Juan José Loor
- Gary Barzola
- Diego Rojas

Análisis sentimental hacia el presidente Guillermo Lasso con respecto a su gestión gubernamental hasta la actualidad

Grupo: 10

Problema

Debido a los recientes escándalos sobre los Panamá papers, problemas en el sistema carcelario, aumento notable de la delincuencia en Guayaquil y alza de combustible, resulta de utilidad saber que tan afectada se ha visto la popularidad del presidente frente al pueblo ecuatoriano y conocer los nuevos índices de aceptación a raíz de estas noticias. Para ello, Twitter posee una enorme cantidad de reacciones y comentarios de ecuatorianos con respecto a la gestión del actual presidente, las cuales son de gran interés para el análisis sentimental de los ciudadanos.

Descripción

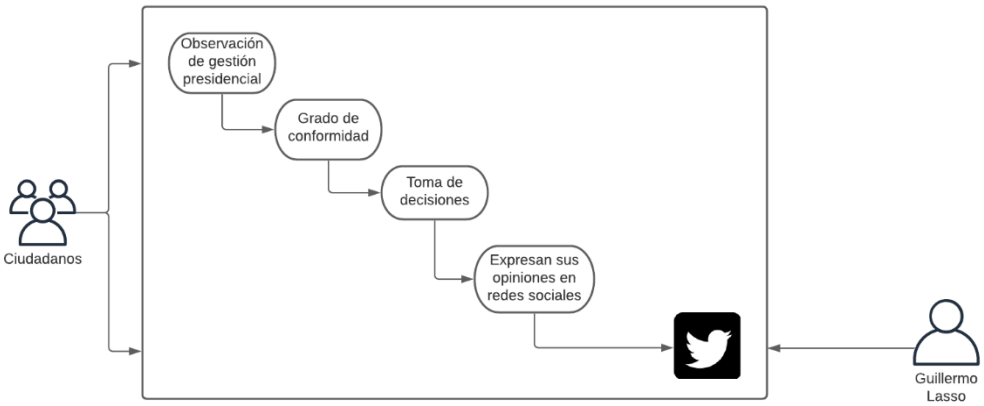
Para este problema haremos uso de técnicas de machine learning, específicamente el uso de regresión logística la cual nos permitirá clasificar el estado de los tweets. Nuestro modelo estará basado en el procesamiento del lenguaje natural para la clasificación del texto extraído de al menos 8000 tweets, por lo que debido a la gran cantidad de comentarios y opiniones como mecanismo de recolección se espera tener acceso al API que provee Twitter para armar nuestra base de datos y así alimentar a nuestro algoritmo. La presente propuesta presenta varios beneficiaros, ecuatorianos y gobernante, los más importante para nuestra propuesta son los ecuatorianos, para que tengan un respaldo y un enfoque más directo ante decisiones a futuro respecto a los gobernantes, los mismos que mediante nuestra solución tiendan a tomar mejores acciones y decisiones que conlleven una mayor popularidad y captación de ciudadanos a su favor.

Objetivo

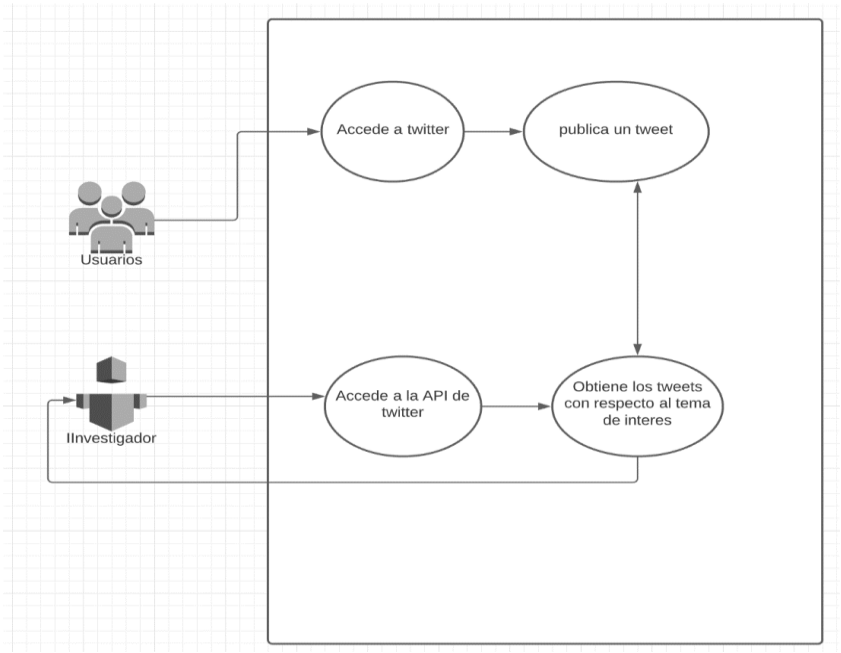
Predecir la polaridad sentimental de comentarios realizados por los ecuatorianos hacia el presidente Guillermo Lasso.

Modelos completos de análisis y definición del problema:

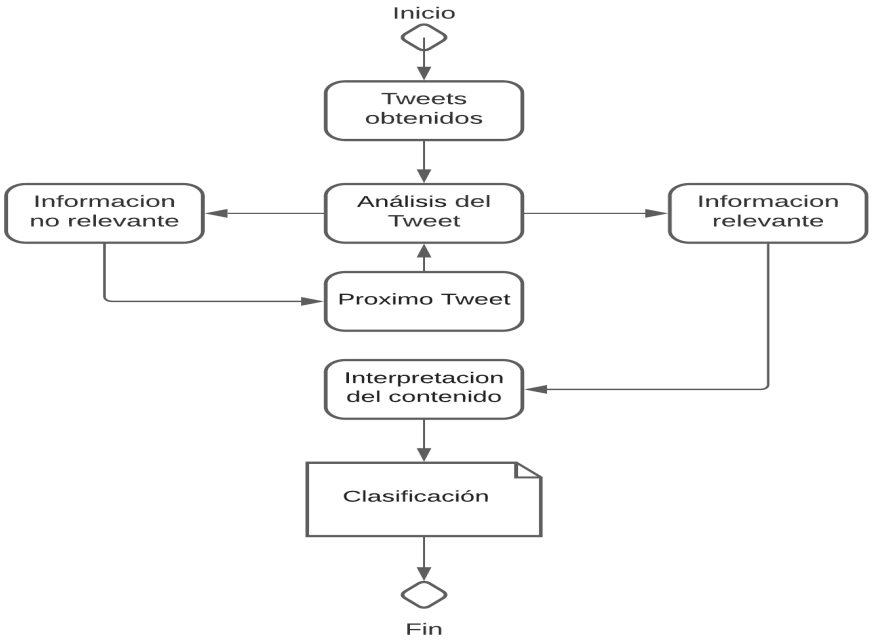
Los ciudadanos del Ecuador leen, escuchan y observan las gestiones realizadas por el presidente Guillermo Lasso y la de sus funcionarios, gestiones que generan un grado de conformidad hacia las personas según las acciones, sean estas buenas o malas, mismas que les permiten tomar decisiones a los ciudadanos que en muchos casos suelen ser protestas al gobierno o en el peor de los casos golpes de estados. Muchas de las opiniones de conformidad y desconformidad son expresadas a través de redes sociales sin filtro, donde no existen restricciones de expresar lo que se piense sobre una u otra persona. La red social más conocida para estos fines es Twitter, la cuál será usada como fuente de datos.



Los datos serán obtenidos a través de un acceso a la API de Twitter, la cual nos permitirá extraer los tweets en base al tema deseado y almacenarlos en una base de datos para su posterior filtrado e interpretación. A continuación, se muestra un diagrama del flujo de extracción de datos antes nombrado:

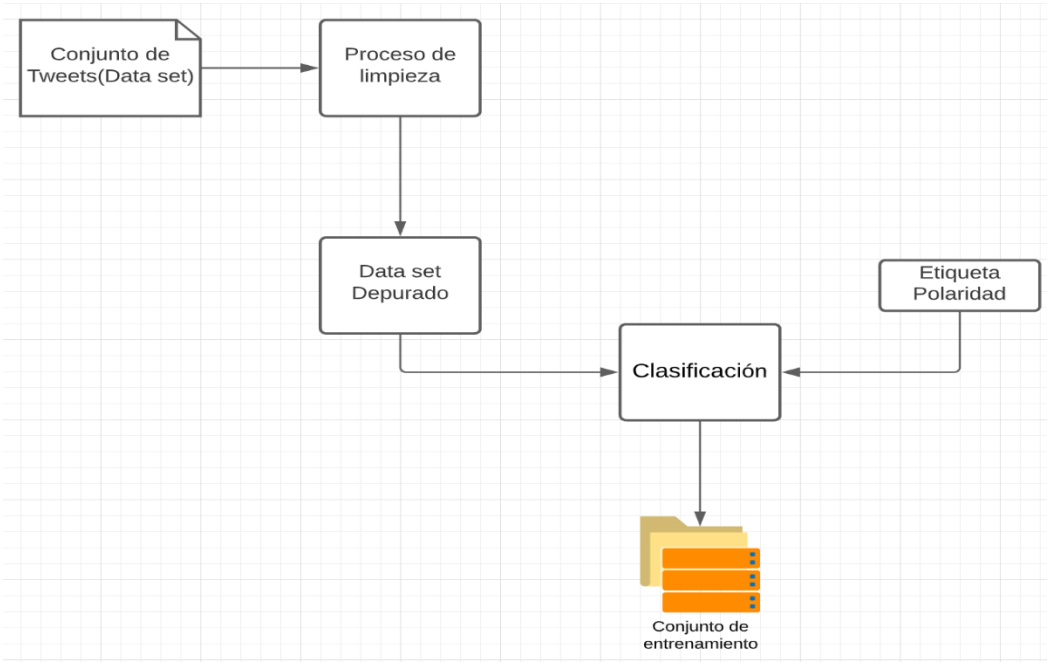


El siguiente diagrama muestra los pasos para el filtrado e interpretación:

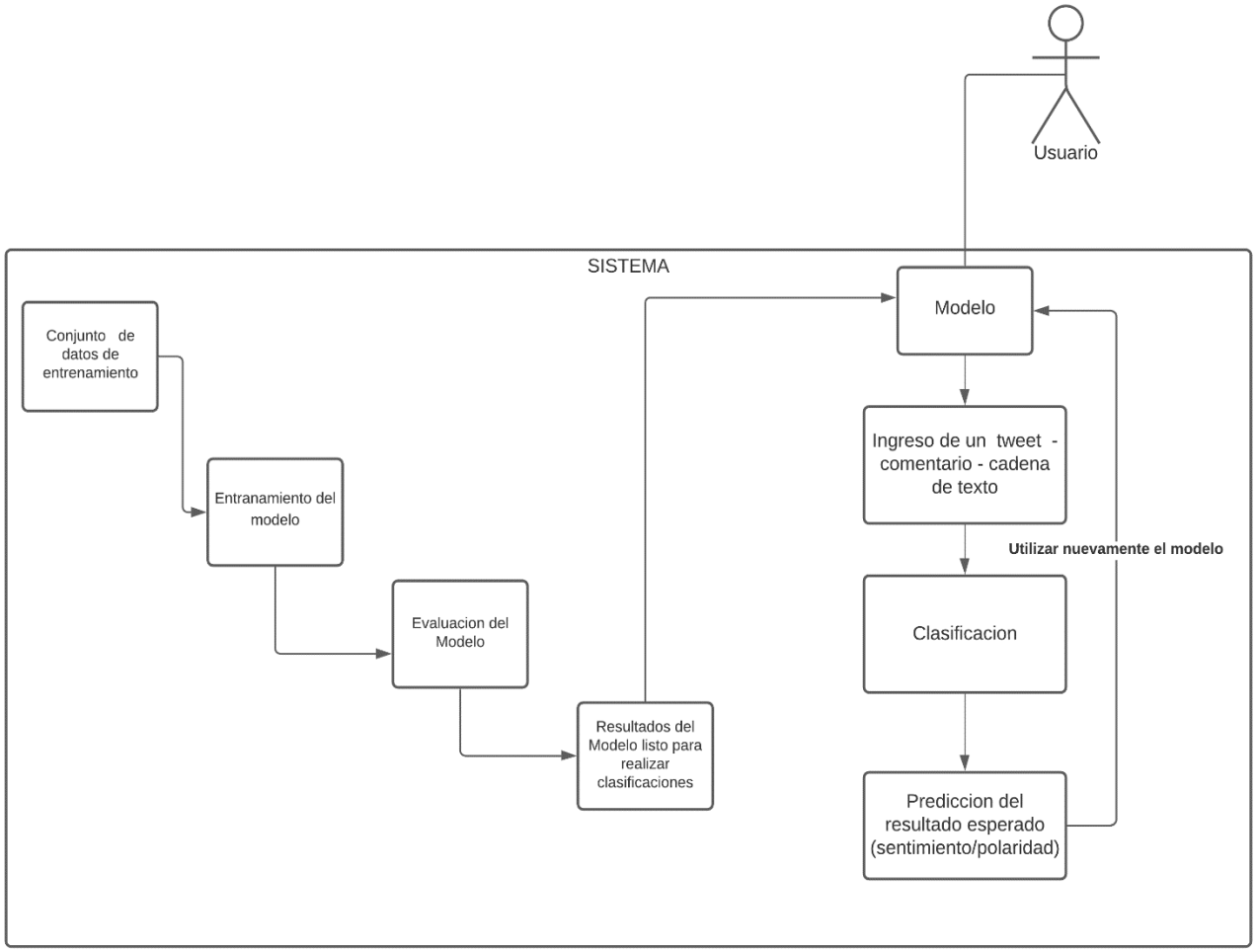


Modelos del diseño de la solución:

Nuestro modelo de análisis está basado en el procesamiento del lenguaje natural, específicamente en el área de la Clasificación del texto, dado un conjunto de tweets (dataset) se procedió a realizar una limpieza de este, luego de tener el dataset depurado le asignamos una etiqueta clasificadora a cada tweet con 3 posibles valores: positivo, neutro, negativo, según su polaridad. Al final obtuvimos el conjunto de datos de entrenamiento con todos los tweets y su clasificación respectiva, mediante el siguiente diagrama podemos visualizar este proceso:



Para la solución de este problema hicimos uso de técnicas de Machine Learning, donde el modelo de diseño de la solución es el Análisis de Sentimientos, el cual se basa en un modelo de Aprendizaje Supervisado que aprendió del conjunto de entrenamientos previamente obtenido, en conjunto a un modelo de Regresión Logística, nuestra solución nos permite dado un tweet, comentario, o cadena de texto que le pasemos a nuestro programa (entrada), obtener un resultado, que básicamente es la clasificación correcta de dicha entrada, en otras palabras el sentimiento/polaridad esperado de esa cadena de texto ingresada. . El siguiente diagrama representa el proceso antes descrito:



Descripción de la Implementación

Para este proyecto hicimos uso del lenguaje de programación Python, ya que es elegante, fácil de leer, escribir y además por la facilidad de usar librerías matemáticas de alto computo como lo es Sklearn y Pandas.

Pandas, la cual es una poderosa biblioteca de análisis y manipulación de datos, misma que usamos para la lectura y manipulación de nuestro dataset, además de un archivo de stopwords el cuál sirvió para quitar relevancia a palabras usadas como conectores dentro de un tweet.

Sklearn, es un módulo de Python que integra algoritmos clásicos de aprendizaje automático con el objetivo de brindar soluciones simples y eficientes a problemas de aprendizaje que sean reutilizables en varios contextos.

En este proyecto precisamente hicimos uso de:

- sklearn.feature_extraction para obtener una lista de stopwords, pero en inglés.
- sklearn.model_selection.train_test_split para dividir los datos que usaremos para el entrenamiento y para hacer pruebas. Los hiper-parametros pasados a esta función son test_size=0.1, el cual nos permite escoger el 10% de nuestro dataset para pruebas como buena práctica y random_state=4, el cual permite mezclar los datos de entrenamiento de forma aleatoria.
- sklearn.feature_extraction.CountVectorizer, permite convertir la matriz de tweet en tokens con los hiper-parametros: max_df que sirve para ignorar los términos que tengan una frecuencia mayor o igual al umbral 0.9, min_df que ignora términos con frecuencia inferior al umbral 0.01 y ngram_range el cual nos sirve para escoger un rango de palabras, en nuestro caso (1,2).
- sklearn.linear_model.LogisticRegression, del cual hacemos uso de los hiper-parametros por defecto, exceptuando multi_class el valor que define la multi clase, el cual se escogió ‘multinomial’ dado que tratamos con tokens que no son binarios. Max_iter.

Luego se procedió a ajustar el modelo con los datos de entrenamientos obtenidos como resultado de la función train_test_split y el vector objetivo que en nuestro caso es la polaridad que se desea obtener como respuesta de usar el método fit y predict de nuestro clasificador LogisticRegression.

Por último, se creó una interfaz haciendo uso de la librería tkinter, donde el usuario puede ingresar un texto simulando un tweet para posteriormente presionar un botón que da paso a la clasificación del comentario haciendo uso de la lógica antes mencionada.

Las características de hardware usado para la ejecución de este proyecto son:

- Memoria RAM de16GB
- Procesador AMD Ryzen 7 5800H
- Sistema Operativo Windows 10 Pro 64bits
- Disco SSD 1T
- Tarjeta Madre ROG Strix G513QC_G513QC

Resultados

- El desarrollo del modelo de análisis sentimental respecto a la gestión del presidente Guillermo Lasso, se implementó haciendo uso de técnicas de machine learning como el aprendizaje supervisado junto a la regresión logística, el cual fue capaz de clasificar si el tweet es positivo, negativo o neutro con una precisión del 77%. De esta forma el modelo quedó listo para ser usado por grandes conjuntos de tweets respecto al presidente y medir el nivel de aceptación actual que posee.

- Debido a la distribución de los datos, nuestro modelo cuenta con una frecuencia pequeña de tweets clasificados como positivos, lo cual genera que tienda a clasificar en su mayoría a los tweets como negativos y neutros. Esto debido a que la red social Twitter en general, tiene más noticias y quejas antes que mensajes de apoyo.

- Este modelo agrega una nueva forma de clasificar sentimientos en Twitter en el idioma español e inglés, descartando stopwords personalizados más los de la librería sklearn, con el fin de centrarse más en lo que quiere transmitir el usuario en cada tweet.

- La creación de una interfaz amigable con el usuario para el uso sencillo del modelo de análisis sentimental implementado.

- El modelo de análisis sentimental desarrollado resulta de utilidad y puede servir de guía en nuevos modelos de análisis sentimental de la gestión de distintos políticos que gobiernan el estado ecuatoriano, para poder conocer sus índices de aceptación y de esta forma poder contribuir de forma positiva en la toma de decisiones de los ecuatorianos respecto a sus futuros gobernantes.

Conclusiones

- Realizar un modelo de análisis sentimental en Twitter requiere una gran cantidad de datos, y su obtención resulta larga y complicada debido a la inmensa cantidad de información que genera Twitter. Al momento de clasificar tal volumen nos encontramos con el problema de no obtener datos de manera homogénea y la mayoría requirió de un proceso de estandarización y limpieza.

- Al definir los hiper-parametros del modelo, tuvimos que cambiar los datos de min_df y max_df para que les dé importancia a las palabras que aparecen en el dataset desde el 1% al 90 % para obtener así una mayor eficiencia.

- Se evidenció que al aumentar a más de 10 el número de iteraciones hasta converger a una respuesta, nuestro modelo ya no aprende más, producto del cual no existe un incremento ni decremento de la eficiencia del modelo.

Definiciones

- Polaridad: La clasificación de la polaridad tiene como objetivo obtener una puntuación que indique si el texto expresa una opinión positiva o negativa, dentro de un rango donde 0 indicaría una carga subjetiva neutra, 1 una carga subjetiva positiva y -1 una carga subjetiva negativa.

- Depurar Texto: Se refiere al proceso de corregir errores ortográficos y separar los diferentes ámbitos del tweet extraído, como fecha, texto, usuario.

Referencias

[1] Singhal, G. (2020, 1 julio). Building a Twitter Sentiment Analysis in Python. Pluralsight.[En línea] Disponible en: https://www.pluralsight.com/guides/building-a-twitter-sentiment-analysis-in-python. [Accedido: 10-Dic-2021].

[2] Editor. (2021, 1 septiembre). Gestión, imagen y credibilidad de Guillermo Lasso han caído más de 10 puntos en un mes, según Perfiles de Opinión. Pichincha Comunicaciones EP. [En línea]. Disponible en: https://www.pichinchacomunicaciones.com.ec/gestion-imagen-y-credibilidad-de-guillermo-lasso-han-caido-mas-de-10-puntos-en-un-mes-segun-perfiles-de-opinion/. [Accedido: 10-Dic-2021].

[3] Scikit-learn. 2022. API Reference. [En línea] Disponible en: https://scikit-learn.org/stable/modules/classes.html#module-sklearn [Accedido: 16-Ene-2022].

[4] Levatić J, Ceci M, Kocev D, Džeroski S. Semi-supervised learning for multi-target regression. Sep 2014 (pp. 3-18).

[5] Liu, B., 2011. Supervised learning. In Web data mining (pp. 63-132).