Análisis de Datos en Twitter

Introducción

Un tema interés en el grupo de sistemas complejos de la carrera de Física de la UMSA es el modelamiento social. Una publicación en particular de la revista Boliviana de Física es la motivación del presente estudio, este artículo se encuentra en el siguiente repositorio online: http://200.7.160.189/docentes/mramirez/SUBIETA17.pdf con el siguiente título:

> CARACTERIZACION DE UN MODELO SOCIAL DISCRETO DE TOMA DE DECISION BASADO EN REDES COMPLEJAS.

Las redes sociales facilitan la interacción entre personas, siendo Twitter una plataforma enfocada en la opinión concisa (menos de 280 caracteres). Convirtiéndose esta red social en una fuente de información directa y abierta de la opinión de cada usuario respecto a cualquier tema.

Siendo que twitter es una especie de pared pública para que cualquier usuario pueda expresar un mensaje, también se publica una cantidad de metadata que permite distinguir quiénes, en qué forma y cuándo ocurren estas interacciones.

Por esta razón considero a Twitter ideal para aplicar el esquema propuesto en el paper mencionado. Siendo este estudio una primera aproximación al análisis de datos sociales.

Antecedentes

En 2019 ocurrió un hecho político de trascendencia internacional en la región latinoamericana: El Estado Plurinacional de Bolivia se quedaba por primera vez desde 2009 con un vacío de poder en el gobierno. Este hecho desato una participación muy activa en redes sociales, particularmente en Twitter.

Algunos estudios muestran que entre el 10–11 de noviembre del 2019 se publicaron alrededor de 2 millones de tuits. Mucha información de esta etapa se ha perdido en la plataforma por lo cual realizar un análisis hoy para contrastar los resultados de estos estudios no es posible.

> Notas públicas sobre el tema en la Web: https://www.eldiario.es/tecnologia/operacion-expulsar-morales-bolivia-twitter_1_1248347.html http://www.cubadebate.cu/noticias/2019/11/17/mas-de-100-000-cuentas-falsas-han-sido-creadas-para-apoyar-el-golpe-de-estado-en-bolivia/ Particularmente este me llama la atención pues es el mas elaborado aunque no comparte ningún método ni base de datos. https://notipress.mx/actualidad/una-campana-de-bots-en-twitter-alimenta-confusion-en-bolivia-2366

Estas publicaciones nos muestran un especial interés sobre la comunicación digital en política en Bolivia.

Objetivos

Este estudio tiene el objetivo principal de explorar el entorno y herramientas que Twitter proporciona. Para lo cual se propone hacer un

estudio de la comunicación de distintos politicos del país dentro de la plataforma de Twitter.

Metodología.

Twitter te permite automatizar el acceso a su plataforma de la misma forma que lo hace un usuario humano en una aplicación: acceder a los tuits públicos, sus respuestas, buscar perfiles, ver ubicaciones y fechas, etc. Limitando a los desarrolladores en únicamente dos sentidos: – El número de peticiones/solicitudes tiene un máximo de 3200 cada 15 minutos además de poder acceder hasta un límite 2 millones de tuits por mes.
– Acceder al historial completo de tuits de un usuario y buscar en el archivo desde el primer tuit hasta el último es una característica de pago. En este estudio no sobrepasamos ni de lejos los límites de tuits procesados ni hacemos uso de funciones de pago. Una vez obtenido un permiso como desarrollador por parte de Twitter (proceso que tiene una serie de pasos y filtros) se puede acceder al API mediante credenciales únicas.

En el siguiente repositorio se muestra el software desarrollado y usado para la extracción, almacenamiento, procesamiento y visualización de un usuario de twitter.

https://github.com/jpcrespo/twanalysis

Procedimiento

El software en resumida explicación hace lo siguiente:

- Accede a los 3200 tuits mas recientes de una cuenta a estudiar (target).
- Extrae un archivo en texto plano (.csv) con las fechas y texto del tuit publicado.
- Realiza una visualización de la serie de tiempo.
- Extrae las palabras y realiza conteos para obtener: – Top 10 de palabras más usadas. – Visualización de las 100 palabras más usadas. – Análisis de sentimientos usando un procesador natural de lenguaje NLP (aplicando redes neuronales) analizar el contenido semántico de un tuit y otorgar un valor numérico del sentimiento del mensaje. Esta aproximación usa directamente un framework en español https://github.com/pysentimiento/pysentimiento con muy buen resultado. Un ejemplo de la capacidad de esta herramienta se muestra a continuación.

```
from pysentimiento import create_analyzer
analyzer = create_analyzer(task="sentiment", lang="es")

analyzer.predict("Qué gran jugador es Messi")
# returns AnalyzerOutput(output=POS, probas={POS: 0.998, NEG: 0.002, NEU: 0.000})
analyzer.predict("Esto es pésimo")
# returns AnalyzerOutput(output=NEG, probas={NEG: 0.999, POS: 0.001, NEU: 0.000})
analyzer.predict("Qué es esto?")
# returns AnalyzerOutput(output=NEU, probas={NEU: 0.993, NEG: 0.005, POS: 0.002})
```

  – Serie de tiempo del sentimiento de los tuits publicados.

Resultados

Este estudio demostró el poder del API de Twitter satisfactoriamente.

Se estudiaron los datos extraídos a tres políticos Bolivianos con presencia en Twitter, la elección de estas personas se debe a que se pueden identificar como 'fuentes' directas de información en cada espacio político que representan:

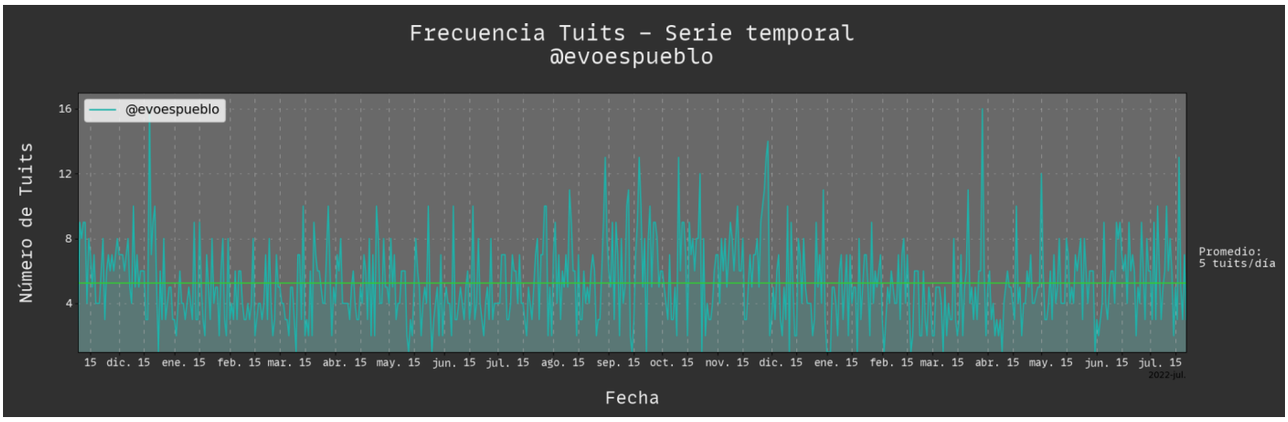
- El expresidente de Bolivia, Evo Morales.
- El gobernador de Santa Cruz y excandidato presidencial, Luis Camacho.

- El expresidente y excandidato presidencial, Carlos Mesa.

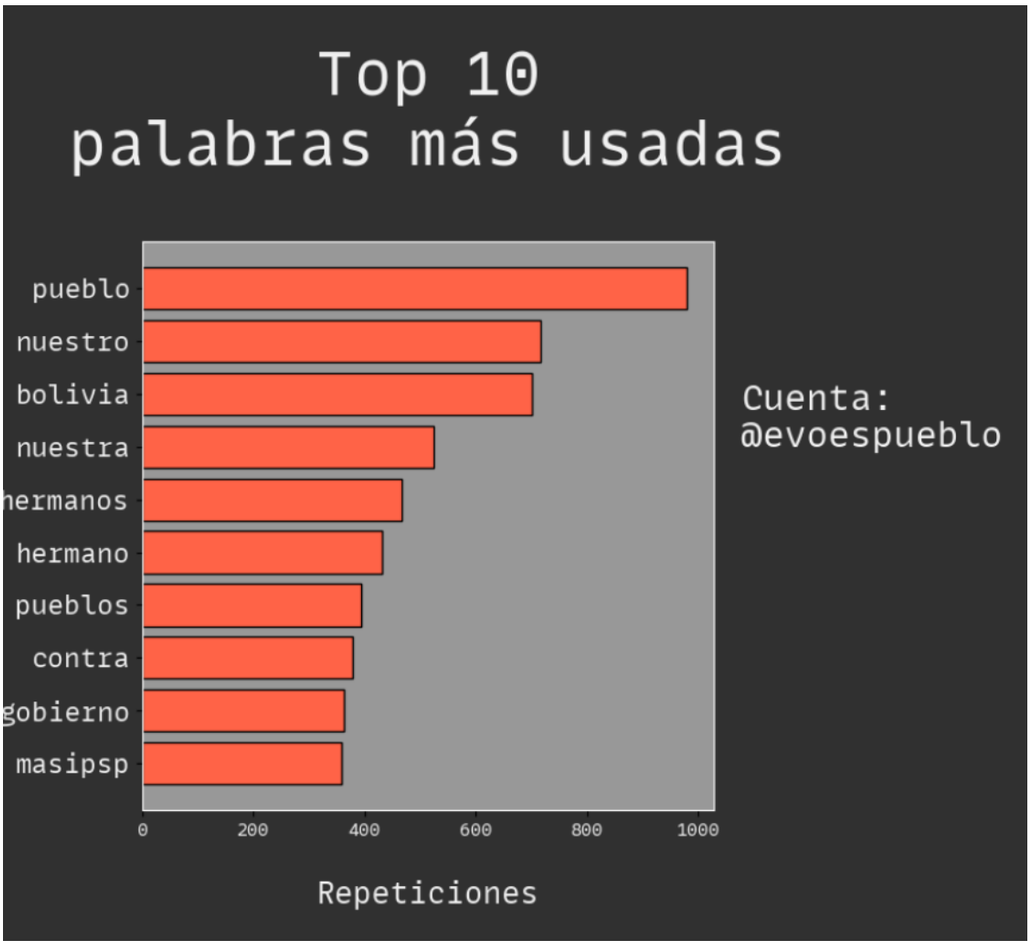
Evo Morales

La cuenta de Morales es quién mas tuitea de entre los tres. Teniendo un promedio de 5 tuits/día, Camacho de 3 tuits/día y Mesa de 2 tuits/día.

Serie de Tiempo

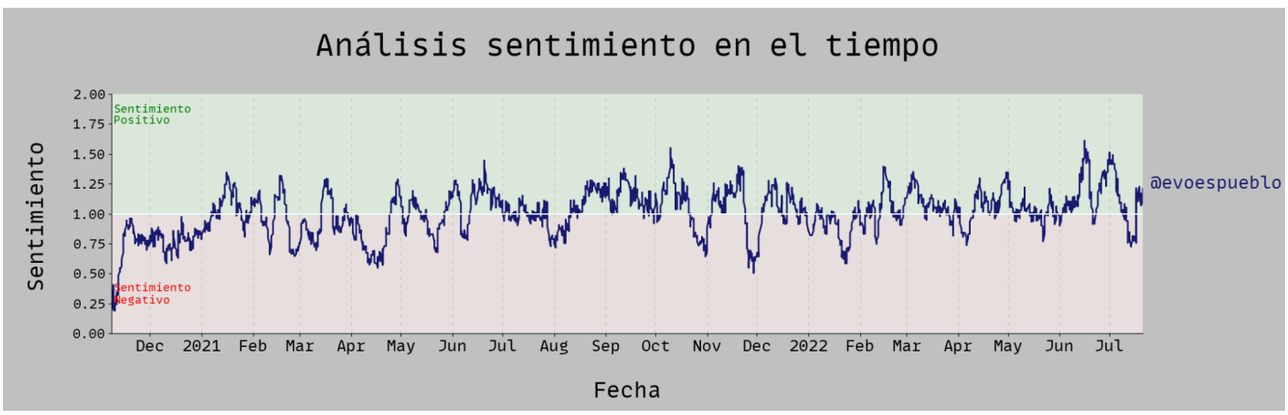


Top 10



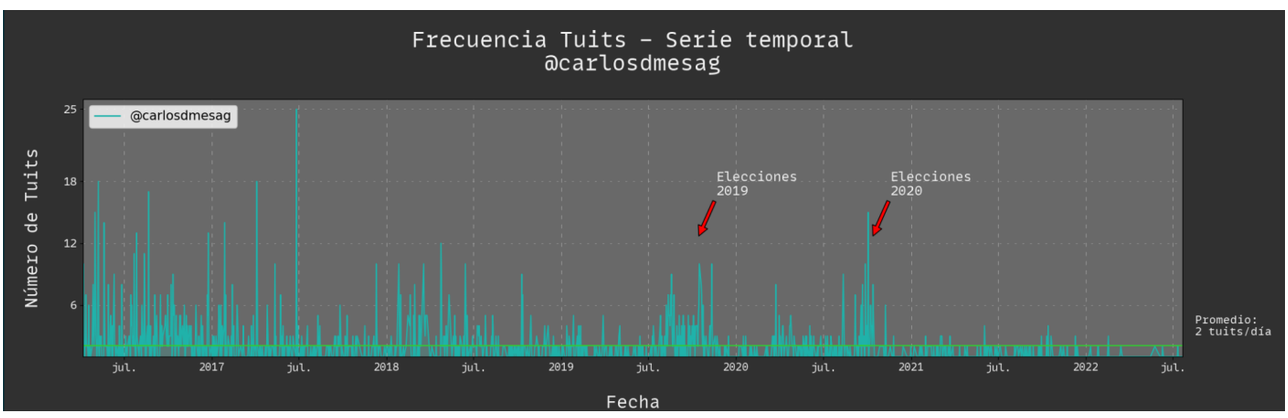
WordCloud

Sentimiento

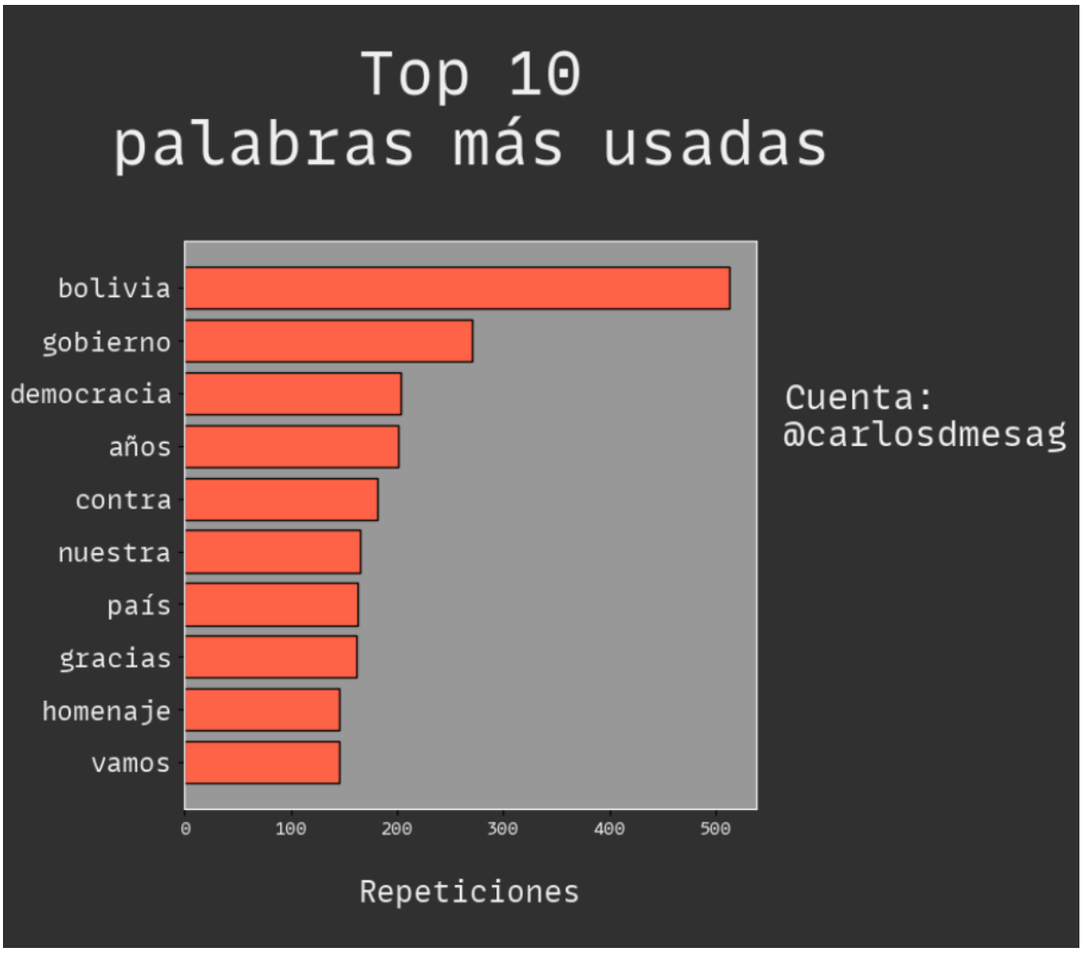


Carlos Mesa

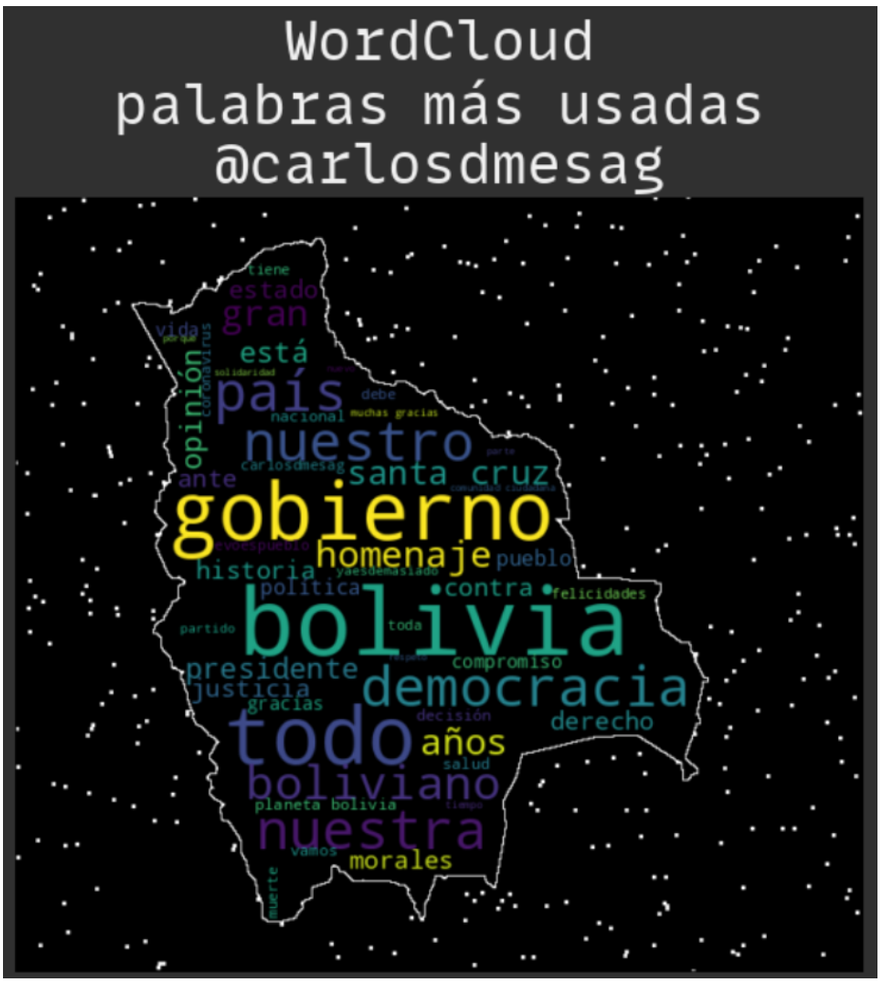
Serie de Tiempo



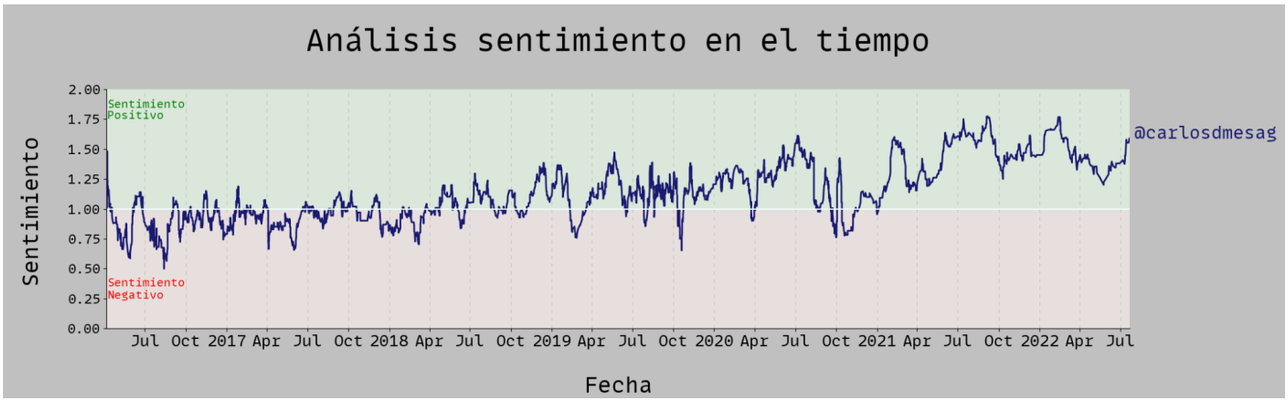
Top 10



WordCloud

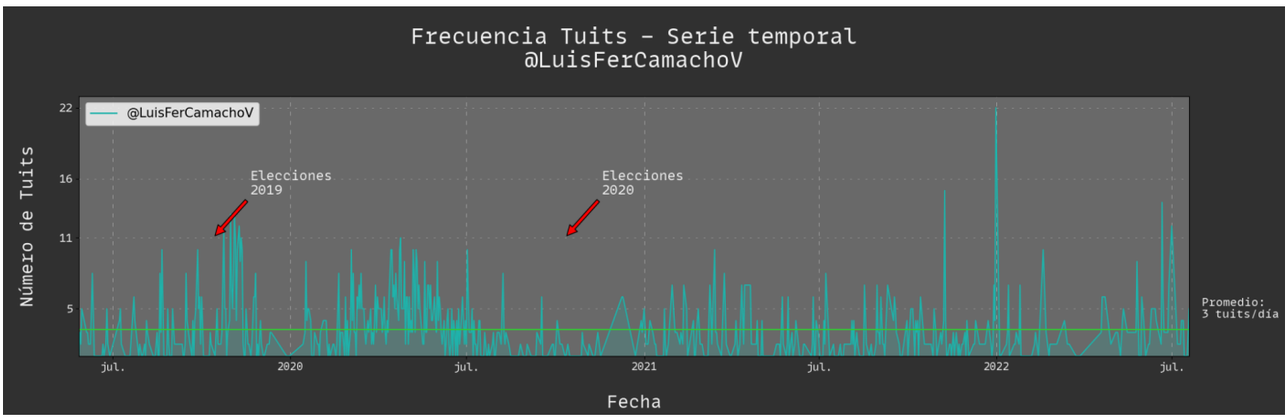


Sentimiento

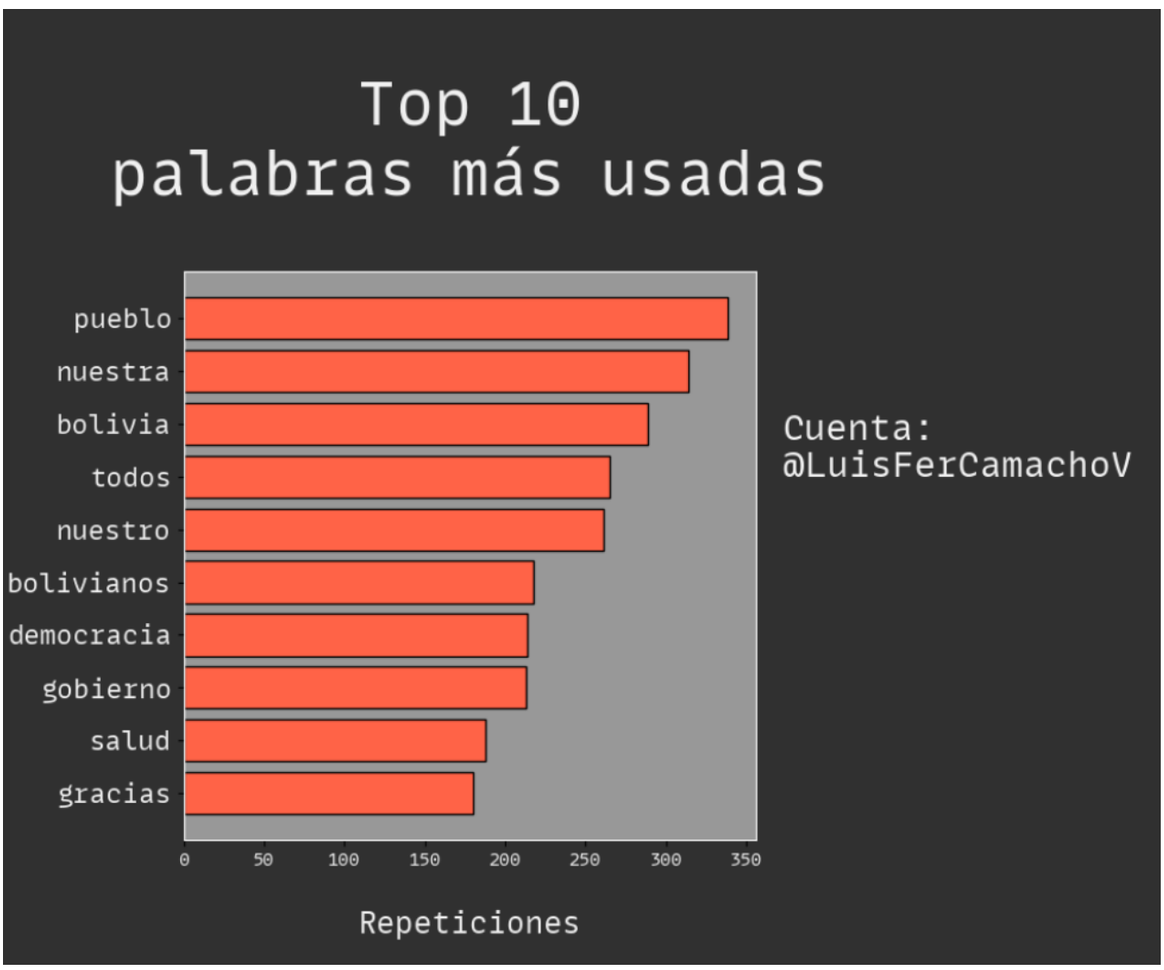


Luis Camacho

Serie de Tiempo

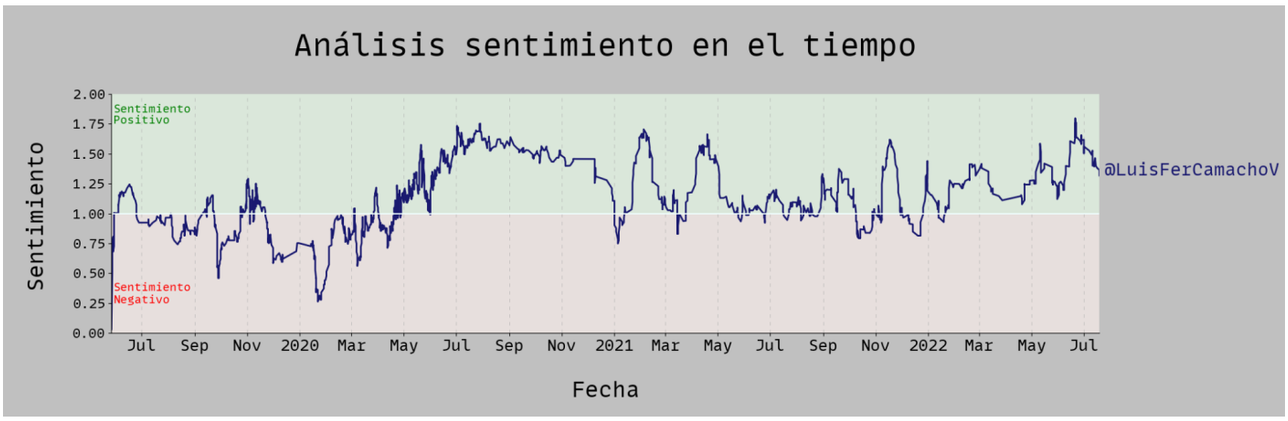


Top 10



WordCloud

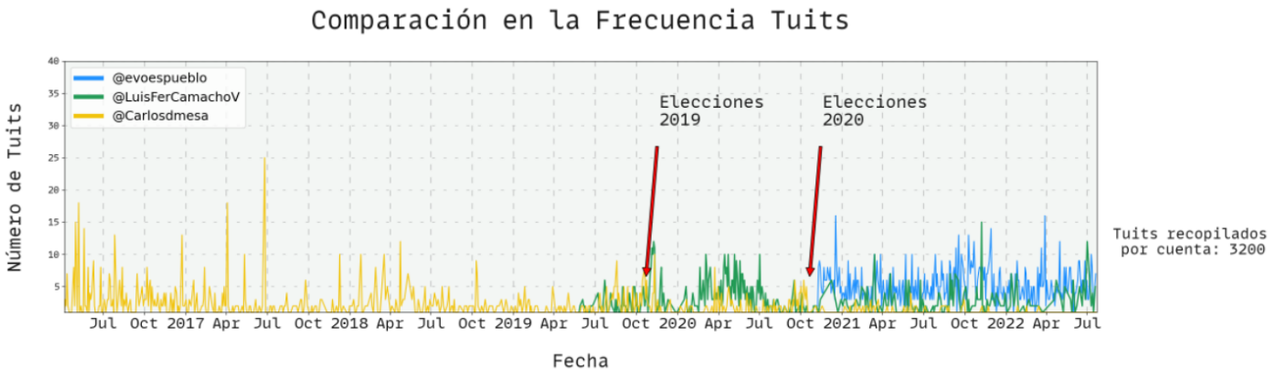
Sentimiento



Comparativas

Esta gráfica muestra como se pueden comparar distintos usuarios. Cabe notar tres casos:

1. Carlos Mesa es de quien más tuits antiguos se pudo rescatar.
2. Luis Camacho es la cuenta más nueva, por lo que para él si se pudo analizar todo el historial.
3. Evo Morales tuitea con tal frecuencia que el límite de 3200 tuits se acaba hace un año.



Conclusiones y perspectivas

- La extracción de datos es rápida y se tienen muchas opciones para filtrar búsquedas.
- Existen otros frameworks que quizá sean más adecuados para realizar los análisis. El NLP es un área de las ciencias de la computación en pleno desarrollo y muchas soluciones se pueden probar, por citar un par:
  - El paquete stanza de la Universidad de Standford.
  - OpenIA, el proyecto de IA de Elon Musk.
- Al no tener una perspectiva social, pues este estudio solo se limita al procesamiento de datos, podría ser de utilidad para algún analista que tenga en cuenta mas variables como fechas, eventos, declaraciones, etc. En el gráfico del punto de Comparaciones por ejemplo se pretende mostrar como cambió la frecuencia de tuiteo respecto a las convocatorias a elecciones.
- El posterior análisis a realizarse pasa a estudiar ya no un usuario específico sino a todo usuario que opine respecto a un tema popular. Por ejemplo, me gustaría elaborar un medidor del sentimiento en Twitter respecto a un activo como Bitcoin y tener un parámetro más para valorar la volatilidad.