

DOCUMENTO



Tecnológico
de Monterrey



Reporte técnico - Dashboard de Sentimientos

Jessica Viridiana Gómez Otero A01378414

Itzel Themis Vargas Tejada A01378973

Romina Rivas Escobar A01378229

Emilio Villacis Mora A01750148



Índice

Introducción

Elementos técnicos

- 1) Elementos del curso integrados en el proyecto y 2) Ejecución de concepto complejos y herramientas del curso
- 3) Innovación de la solución
- 4) Entendimiento y análisis de los datos
- 5) Adquisición y cantidad de los datos
- 6) Limpieza de datos
- 7) Asertividad y validación de datos
- 8) Asertividad y validación de modelos de datos
- 9) Fundamentación de selección de métodos usados en el proyecto
Descripción general de las fases:
- 10) Calidad de la visualización de la información

Reporte técnico - Dashboard de Sentimientos

Introducción

Para la construcción de esta herramienta se empieza partiendo de la información en tiempo real de tipo texto presente en tweets que estén vinculados con Cuautitlán Izcalli, de ello se realiza un proceso de análisis de sentimiento y texto que es finalmente desplegado como las visualizaciones del tablero que presentan datos valiosos para entender qué están sintiendo y expresando los ciudadanos de cuautitlán Izcalli en ese momento.

El análisis y procesamiento que corre en un punto intermedio entre que se recuperan los tweets y hasta convertirse en datos mostrados en el tablero ocurre en Python mediante un proceso de clasificación de nuevos casos con una librería llamada pysentimiento.

La ficha técnica de la librería puede ser consultada en: [Pysentimiento - Ficha Técnica](#)

Ahora bien, la calidad y validez de la información presentada en el Dashboard de Sentimientos depende directamente del proceso de construcción y tratamiento de los datos con el que fue alimentado. Para ello se presenta el reporte técnico que permita comprobar paso a paso el correcto desempeño de éste.

Elementos técnicos

1) Elementos del curso integrados en el proyecto y 2) Ejecución de concepto complejos y herramientas del curso

La tabla a continuación muestra cada uno de los elementos que fueron aplicados en el proyecto y como estos dependen de un concepto clave y herramienta para su desarrollo.

Aplicación al proyecto	1) Concepto	2) Herramienta
Obtención de datos	Web scraping	<ul style="list-style-type: none">- Twitter API Standard v1.1- Python (tweepy, requests)
Preprocesamiento de datos	Regex Transformación de texto	<ul style="list-style-type: none">- Python (regex, sklearn)
Procesamiento de datos	Generación de modelos Análisis de Sentimiento	<ul style="list-style-type: none">- Python librerías como: datetime,

		pandas, numpy, transformers
Base de datos	Almacenamiento datos	- Azure, SQL
Modelo para análisis de sentimiento	Métodos de evaluación de modelos	- Out-sample error (manual test data) - Accuracy
Asistente virtual	Chatbot	- IBM watson assistant
Dashboard descriptivo	Análisis exploratorio de la data (EDA)	- Python (pandas, numpy, sklearn, matplotlib) - PowerBI
Presentación	Storytelling y presentación de impacto	- Canva

3) Innovación de la solución

Más allá de un proyecto de análisis de texto buscamos que el valor agregado se encuentre en temas más allá de lo común para ofrecer una solución lo más útil y de valor posible:

Para lo cual identificamos dos puntos clave a buscar: primero una automatización del proceso para que el usuario final sólo tenga que visualizar nuestro dashboard sin la necesidad de correr código o manualmente subir datos nuevos y dos, una automatización de datos en tiempo real sin la necesidad de una persona que deba hacer la consulta. Con ello los temas que nos permitieron atender la situación son:

- Actualización de datos nuevos en tiempo real: Se construyó un trigger que permite hacer el proceso de recuperar los datos (web scrap) automáticamente y sin la necesidad de intervención.
- Seguridad y recuperación de los datos: Clave para futuros análisis se instaló una base de datos en Azure que nos garantiza una protección y almacenamiento de todos los tweets que se trabajen. La base puede ser consultada cuando se desee y tiene el valor de que los datos están en la nube y por lo cual se puede migrar y conectar a múltiples herramientas para otros propósitos.
- Automatización del dashboard: Mediante una conexión de nuestros datos en la base de datos en Azure con la herramienta del dashboard que es PowerBI, tenemos una conexión en real que no requiere de almacenamiento en local (es decir un archivo que se descarga en la computadora y luego se sube a la

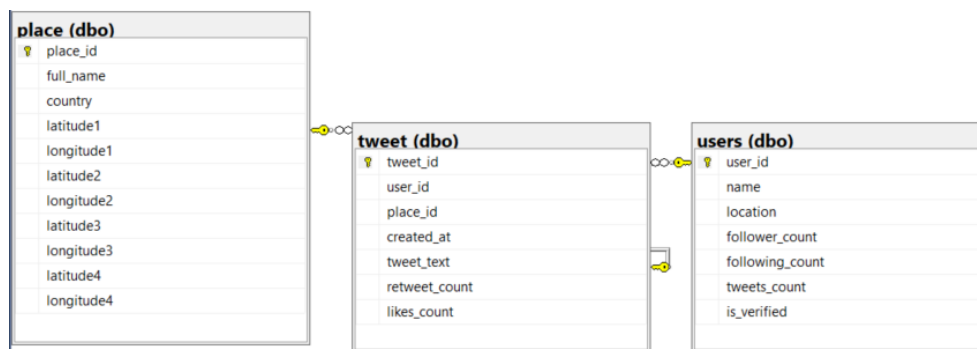
herramienta de visualización) de los datos y con ello se permite siempre poder consultar el tablero con cualquier dato/tweet nuevo simplemente entrando a ver el tablero.

- El proceso de análisis y clasificación realizado con modelos probados y entrenados que mostraron altos parámetros de desempeño: Mediante la función de Pysentimiento, librería que utiliza modelos pre-entrenados de transformers para distintas tareas de SocialNLP.
- Visualización: PowerBi

4) Entendimiento y análisis de los datos

De la extracción definimos datos claves para poder realizar un análisis sólido, estas fueron de tres áreas grandes, aplicando tres documentación de la API de Twitter que mediante sus llaves nos permitió recuperar estos datos.

Se generó igual el siguiente diccionario: [Diccionario de datos Tweepy](#) el cual contiene de las áreas, tres tablas y sus variables definidas así como la documentación utilizada para la extracción.

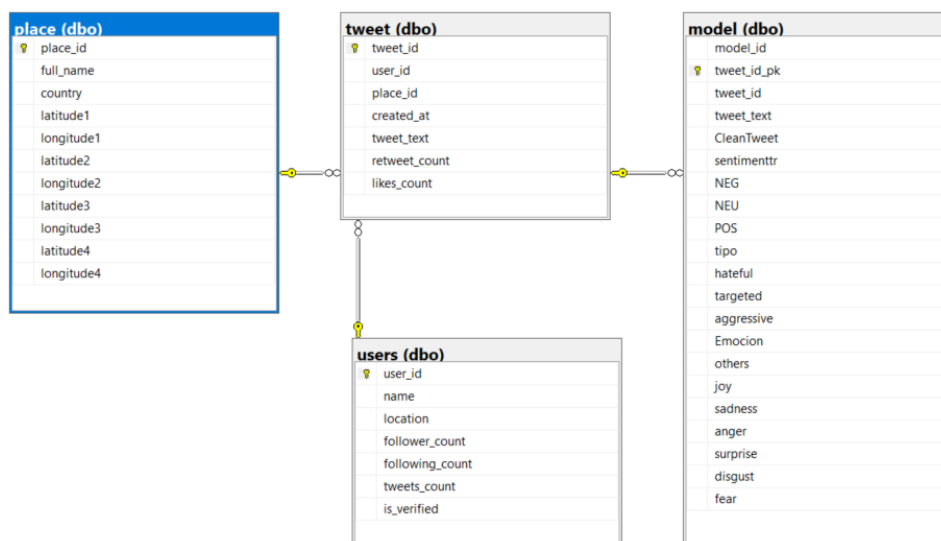


5) Adquisición y cantidad de los datos

- tweets: 36,330
- place: 56
- user: 20,898
- model: 11,618
- Total con model: 68,902
- Total sin model: 57,284
- fecha de inicio: 2022-04-14 04:35:08.000
- fecha de fin: 2022-06-02 14:33:51.000

Los datos adquiridos del web scarp del modelo anterior dan un total de 36,330 tweets del periodo 2022-04-14 al 2022-06-02. Adicionalmente y para fines del análisis se creó una nueva tabla con nuevas columnas que nos permitirán entender mejor el caso de estudio y fungieron como nuevas variables para encontrar correlaciones e información importante.

Con ello se conservan la misma cantidad de casos/filas pero se agrega la nueva tabla “model”. Teniendo como estructura de datos final el siguiente modelo:



6) Limpieza de datos

La limpieza de datos fue aplicada una vez que se recuperaron del web scrap en su búsqueda por palabras clave.

Pasos limpieza para análisis de texto:

1) Eliminar tweets retwitteados (rt)

En este paso, se eliminaron los re-tweets debido a que no se asegura que lo que el usuario exprese textualmente. Esto puede verse afectado por un gran número de conteos solo por el hecho de ser viral por no mencionar que en realidad puede ser otro tema que se hable en el contexto.

2) Eliminar formatos y links (html y cualquier link a páginas web)

Se eliminaron direcciones web (urls) con sus respectivos formatos de texto (http, https, www, etc) puesto que no aportaban valor al objetivo principal de análisis de texto. En este sentido y, para lograr un mejor entendimiento del contexto, fueron removidas del análisis.

3) Eliminar usuarios que no aportan valor (ej. cuentas oficiales de gobierno, políticos, etc)

Se identificaron los usuarios de cuentas políticas y de gobierno, con ello se procedió a remover todos los tweets que provenían de ellas debido a que no son parte del foco de estudio que busca entender el sentimiento de la población, estos únicamente nos permitiran tener información predeterminada que no expresa ningún sentimiento de la población de Cuautitlán Izcalli.

4) Eliminar puntuación y caracteres especiales, solo se conserva alfanumérico

Con el objetivo de tener un texto más limpio y entendible, se removieron todo tipo de puntuaciones y caracteres especiales (" # \$ % & ' () *). Esto permitió disponer de un texto mayormente legible y para uso analítico.

5) Eliminar stopwords

En este paso se eliminaron palabras vacías. Término que hace referencia a las palabras de uso común que brindan poca o ninguna información única que se pueda usar para la clasificación, es decir, son palabras que generalmente son ignoradas por el motor de búsqueda como: “en”, “la”, “es”, etc. Estas palabras se eliminan para ahorrar espacio en la base de datos y el tiempo de procesamiento.

6) Eliminar palabras de contexto

Finalmente, se eliminaron palabras de contexto, debido a que no son necesarias realmente dentro del análisis puesto que son palabras que fueron parte del filtro de obtención de la información.

7) Asertividad y validación de datos

Confiables:

Uso de red social que expresa sentimientos y opiniones (Twitter)

Extracción directa sin manipulación de terceros (Raw data de API Twitter)

Validez: Proceso de preparación y limpieza

-Búsqueda por tópicos de interés/ palabras clave

- Excluir casos que no aportan valor

8) Asertividad y validación de modelos de datos

Podemos validar la confiabilidad y precisión del modelo de datos seleccionado que se aplica con Pysentimiento primeramente por el proceso de extracción y limpieza de datos que fue cuidadosamente realizado y explicado anteriormente.

En segundo punto, se selecciona gracias a una comparativa de diversos modelos públicos disponibles. La intención de no construir nuestro propio modelo se debe a que la investigación realizada nos permitió entender que NLP es un tema complejo que requeriría de tiempo para entrenar una amplia cantidad de datos que volvieran sólido el análisis. Por ello la mejor opción fue tomar los modelos más usados y entrenados disponibles y a partir de ellos realizar una comparativa de precisión con la técnica out-sample que nos permite ver que sin duda es esta selección de pysentimiento la mejor opción, aunado a que mostró

bondades para el análisis tales como análisis de emociones y detección de odio en el texto con sus respectivas probabilidades, por no mencionar que tiene como modelos base a algoritmos desarrollados por google mismos que son entrenados con millones de datos y que constantemente sigue aprendiendo. Estos modelos utilizan técnicas de enmascaramiento de palabras en modo bidireccional, lo cual permite entender el contexto ya sea de forma gramatical como de coherencia desde el lado izquierdo hacia el derecho y viceversa.

Este modelo ha sido utilizado en diversos proyectos con el objetivo de evaluar el sentimiento en redes sociales con respecto a diferentes temas, entre los que se encuentran un análisis de la violencia contra las mujeres políticas en redes sociales por ONU Mujeres Uruguay y sobre opiniones y creencias sobre las vacunas en redes sociales bolivianas.

Modelos	IDIOMA	Que predice	Matriz	Precisión
sentiment-analysis	ES	NEGATIVE / POSITIVE	Predicted NEGATIVE POSITIVE Actual NEG 1698 309 NEU 2862 897 POS 171 62	78.57%
finiteautomata/bertweet-base-sentiment-analysis	ES	NEG / NEU / POS	Predicted NEG NEU POS Actual NEG 68 1914 25 NEU 50 3649 60 POS 2 194 37	62.5%
sagorsarker/codeswitch-spa-eng-sentiment-analysis-linc e	ES	Label 0 = Negativo Label 1 = Positivo	Predicted LABEL_0 LABEL_1 Actual NEG 756 1251 NEU 901 2858 POS 7 226	43.8%
j-hartmann/sentiment-roberta-large-english-3-classes	INGLÉS (Tweets traducidos)	negative / neutral / positive	Predicted negative neutral positive Actual NEG 1128 823 56 NEU 1030 2550 179 POS 6 142 85	62.7%
pysentimiento	ES	NEU / NEG/ POS	Predicted NEG NEU POS Actual NEG 1757 246 4 NEU 1432 2247 80 POS 14 90 129	68.8%

Accuracy por categoría:

Neg = 87.54%

Pos = 55.36%

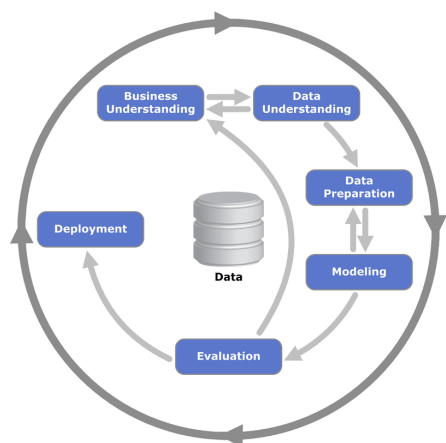
Neu = 59.77%

Es importante mencionar que a pesar de no tener precisiones altas por categoría de positivo y neutral , la precisión de la categoría de Neg aporta un valor del 87.54%, mismo que apoya al interés y al objeto de estudio del reto proporcionado

9) Fundamentación de selección de métodos usados en el proyecto

Metodología CRISP-DM (Cross Industry Standard Process for Data Mining)

Metodología utilizada para la gestión de nuestro proyecto ya que incluye el ciclo de vida de un proyecto estándar de análisis de datos



Este modelo cubre las fases del proyecto, las tareas respectivas y cómo se relacionan estas tareas. A partir de esta metodología se logra el tratamiento de los datos adecuado para la elaboración de modelos el cual es el objetivo final de nuestro proyecto.

A continuación se puede observar el ciclo vital del modelo dividido en 6 partes:

Cabe recalcar que la secuencia de las fases mostradas son completamente flexibles y las flechas solo representan las posibles y más comunes relaciones y dependencias entre las tareas de cada fase. La metodología se presenta como cíclica ya que durante este tipo de proyectos se puede iterar y mejorar el proceso conforme se realizan avances en las tareas y fases.

Descripción general de las fases:

1. Business Understanding. Definición de necesidades del cliente: durante esta fase logramos obtener un entendimiento de la problemática que nuestro cliente intenta resolver, esto a través de investigaciones profundas y consultas a expertos/profesores. Se obtuvo la definición del objetivo de nuestro proyecto.
2. Data Understanding. Estudio y comprensión de los datos: Durante esta fase se desarrollo un método de recolección de datos, esté siendo web scraping. A partir de la recolección de los datos se hizo un análisis de estos utilizando diferentes métodos como creación de queries(SQL) y un estudio exploratorio de los datos. A partir de esto se logró generar una planeación para la futura creación de la base de datos y la forma en la que se desarrollaría el modelo en pasos futuros.

3. *Data Preparation. Análisis de los datos y selección de características:* A lo largo de esta fase se obtuvo la definición de los procesos de limpieza y procesamiento de los datos. Paso durante el cual se obtuvo la materia prima de nuestros modelos.


a. **Metodología base del procedimiento: Data cleansing.**

A lo largo de nuestro proyecto se llevaron a cabo diferentes pasos específicos para asegurar la exactitud, coherencia, validez y uniformidad de nuestros datos. Estos pasos específicos se obtuvieron al realizar varias iteraciones a la metodología CRISP. De manera general se siguieron los siguientes pasos:

- Tratamiento de duplicados
 - Verificación de datos nuevos
 - Actualización constante
 - Implementación de una entrada de datos coherente
4. *Modeling. Modelado:* Se aplican todas las diferentes técnicas de modelado. Durante este punto es común regresar al punto anterior y generar cambios para poder crear modelos de una forma más eficiente o precisa.
5. *Evaluación. Evaluación (obtención de resultados):* Se generan distintas evaluaciones a los modelos creados con el fin de la selección del mejor modelo al caso del proyecto. Aquí también se evalúa la forma en la el modelo logra cumplir con el objetivo del proyecto.
6. *Deployment. Despliegue (puesta en producción):* A partir de esta parte inicia el monitoreo de toda la solución. Durante esta fase se identifican puntos de mejora y el proyecto continúa mientras salta de fase en fase. De esta forma se logra evaluar constantemente el proyecto y sus resultados.

10) Calidad de la visualización de la información

Los tableros desarrollados en PowerBi muestran de manera gráfica e interactiva la información más importante obtenida. Se cuenta con una tabla que explica cada visualización y su respectivo uso. Puede ser consultado en:

 Dashboard sentimientos - Calidad de la visualización de la información

La fecha de elaboración del presente documento tiene corte al 6 de junio del 2022.