Relacionamiento automático de opiniones – UNFPA

1. Contexto del negocio

Los Objetivos de Desarrollo Sostenible (ODS/SDG; Objetivos Globales; [UNDP, 2024]) fueron adoptados por las Naciones Unidas en 2015 como un llamamiento mundial para poner fin a la pobreza, proteger el planeta y garantizar que para el 2030 todas las personas disfruten de tranquilidad y bienestar.

a. Objetivo #3 "Salud y bienestar": Se enfoca en la buena salud como elemento esencial para el desarrollo sostenible, sin tener en cuenta las propiedades de salud global emergentes. Busca mantener enfoques multisectoriales basados

en derecho y con perspectiva de género. Su progreso es determinado por la discrepancia de la expectativa de vida entre países y sus promedios nacionales. Lucha contra las principales causas de enfermedades con riesgo de defunción:

i. Ampliación de las desigualdades económicas y sociales.
ii. Rápida urbanización.
iii. Amenazas para el clima y el medio ambiente.
iv. Lucha continua contra el VIH y otras enfermedades infecciosas.
v. Nuevos problemas de salud (enfermedades no transmisibles, etc.).

b. Objetivo #4 "Educación de calidad": Se enfoca en incrementar el número de menores de edad que acceden a servicios de educación y la tasa de alfabetización, nivelando a su vez brechas de género. Algunos de los principales agentes en contra son los niveles de pobreza, conflictos armados y otras emergencias en países en desarrollo.
c. Objetivo #5 "Igualdad de género": Se enfoca en erradicar la discriminación contra el género femenino en pro del desarrollo sostenible y crecimiento económico, alentando el aumento de mujeres líderes. Específicamente, busca detener:

i. Negación de los derechos laborales que sí tienen los hombres.
ii. Violencia y explotación sexual.
iii. División desigual del trabajo no remunerado.
iv. Exclusión de la toma de decisiones en el ámbito público.
v. Cambio climático.
vi. Conflicto y migración.

2. Oportunidad/problema del negocio

Actualmente, el relacionamiento de los diferentes ODS con la información brindada por los ciudadanos presenta un consumo excesivo de recursos humanos. La interpretación de los datos recopilados requiere de conocimientos especializados, es decir contar con personas especializadas en el contexto para asegurar la calidad del análisis. En otras palabras, el proceso de relacionamiento de los ODS depende estrechamente de la consulta constante a expertos que dominen el entendimiento de todos los contextos sociales alrededor del mundo. Dicha tarea se limita por la disponibilidad de los expertos, ralentizando el análisis de la información y puede resultar costosa a largo plazo. Es así que se busca relacionar de forma automática y eficiente las opiniones de los ciudadanos con los ODS 3, 4 y 5 mediante un modelo de análisis de datos que opere a partir las opiniones en lenguaje natural.

3. Impacto que puede tener en Colombia este proyecto

De los 51.6 millones de habitantes [DANE, 2022] de Colombia, el 51,2% son mujeres, de las cuales el 74,6% han sido víctimas de delitos sexuales en áreas recreativas; además, ha aumentado la tasa de letalidad materna desde 2018 [DANE, 2022]. En este contexto, los ODS 3, 4 y 5 están dirigidos a eliminar necesidades insatisfechas

en materia de anticoncepción, que las mujeres/niñas no sean víctimas de violencia y suprimir las defunciones maternas evitables [UNFPA, 2024]. Por consiguiente, los ODS están dirigidos a una población que representa más de la mitad de la nación.

La analítica del relacionamiento de los diferentes ODS con la información brindada por los ciudadanos entra a jugar un papel crítico frente a esta problemática latente, aportaría a reducirla drásticamente y lucharía contra algo que se ha convertido en la realidad cotidiana del país.

4. Objetivos

a. Automatizar un proceso replicable para aplicar la metodología de analítica de textos en la construcción de modelos analíticos.
b. Desarrollar una aplicación que utilice un modelo analítico basado en aprendizaje automático y sea de interés para una organización, empresa o institución y en particular para un rol existente en alguna de ellas

5. Proceso de automatización

a. Proceso de preparación de datos

El preprocesamiento de texto fue una parte fundamental en la construcción de un modelo de aprendizaje automático para el proyecto. En nuestro caso, usamos la librería NLTK para tokenización y eliminación de stopwords, y Spacy para la lematización de las palabras en español.

i. Clase de preprocesamiento personalizada (TextPreprocessor)
Generemos una clase llamada TextPreprocessor, que extiende las clases base de Scikit-learn (BaseEstimator y TransformerMixin). Esto nos permite integrar el preprocesamiento como parte del pipeline de Scikit-learn.
El flujo de la clase fue el siguiente:
   1. Eliminación de valores nulos: primero, eliminamos cualquier texto que sea None o nulo.
   2. Tokenización: usamos nltk.word_tokenize() para dividir el texto en palabras individuales.
   3. Conversión a minúsculas: convertimos todas las palabras a minúsculas para asegurar que el modelo no distinga entre mayúsculas y minúsculas.
   4. Eliminación de puntuación: usamos expresiones regulares para eliminar cualquier carácter de puntuación.
   5. Normalización: normalizamos los caracteres utilizando unicodedata.normalize(), lo que elimina acentos y caracteres especiales.

6. Eliminación de stopwords: usamos las listas de stopwords de NLTK para español, y añadimos algunas palabras personalizadas como 'mas'.
7. Lematización: utilizamos Spacy para reducir las palabras a su forma base o raíz, por ejemplo, "corriendo" se convierte en "correr".
8. Conversión a cadena: finalmente, unimos las palabras en una cadena de texto de nuevo para que pueda ser vectorizada por el modelo.

Este proceso asegura que los datos de entrada se normalicen y se reduzcan al mínimo para que el modelo pueda aprender mejor.

b. Construcción del modelo

El siguiente paso fue generar un pipeline que incluya el preprocesamiento de texto y la clasificación. Utilizamos un modelo SVM con el kernel "RBF" (rbf) para clasificar textos en categorías de los ODS.

i. Pipeline de Scikit-learn
   1. El pipeline incluye primero la clase TextPreprocessor que acabamos de describir.
   2. Luego, aplicamos la vectorización de texto con TfidfVectorizer para convertir los textos en una representación numérica basada en TF-IDF.
   3. Finalmente, usamos svm.SVC() como clasificador, configurado para devolver probabilidades (probability=True).

c. Persistencia del modelo

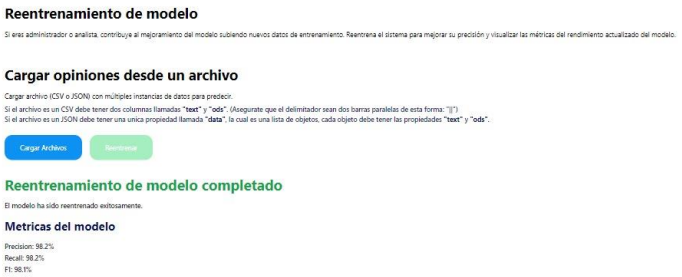
Una vez que el modelo está entrenado, lo guardamos en un archivo .joblib para su uso posterior. Esto nos permite cargar el modelo en cualquier momento sin tener que entrenarlo nuevamente.

d. Acceso por medio de API

Luego de entrenar y guardar el modelo, generamos una API utilizando FastAPI para realizar predicciones en nuevas opiniones. El modelo previamente guardado se carga usando joblib.load().

i. Endpoint de predicciones: recibe una lista de opiniones y las procesa utilizando el pipeline cargado.
   1. Las predicciones y las probabilidades asociadas a cada opinión son devueltas como respuesta JSON.
   2. Como resultado, podemos ver esta imagen de cómo funciona el resultado de nuestra aplicación después de haberle pasado palabras significativas al modelo.

ii. Endpoint de reentrenamiento: se adjuntan nuevos datos para ser procesados por el modelo.
    1. Los datos se reciben en un csv o JSON como una lista de opiniones y su ODS asociado.
    2. El resultado de la aplicación son las métricas resultantes después de que el modelo procesara los datos.



6. Desarrollo de la aplicación

a. Usuarios

i. El endpoint de predicciones está dirigido a miembros pertenecientes a entidades públicas colombianas, con las que trabaja la UNFPA, que recopilan opiniones de los ciudadanos para caracterizar el desarrollo de los ODS en zonas geográficas específicas del país. Este usuario trabajaría en zonas donde hay poblaciones con altas tasas de letalidad debido a enfermedades, menores de edad que no tengan acceso a la educación y desigualdad de género. Mediante la recopilación de datos en forma de opiniones, el usuario busca predecir a qué ODS hacen referencia los testimonios de los ciudadanos. De esta manera, podría identificar cuál es el ODS presente más crítico de la región y por el cual se debería empezar a trabajar en la mitigación de las problemáticas con el fin de cumplir las metas propuestas para el 2030.

ii. El endpoint de reentrenamiento está dirigido a un usuario administrador de la aplicación o analista que busca añadir opiniones etiquetadas con el ODS asociado al que hacen referencia para que el modelo vaya aprendiendo sobre los nuevos datos que recibe.

b. Conexión con el proceso de negocio

Como se había planteado anteriormente, un criterio de éxito del proyecto es identificar factores sociales presentes dentro de las opiniones de los ciudadanos que diluciden acerca de cuáles son los problemas con mayor influencia dentro del desarrollo de una zona del país. Si la recopilación de datos se hiciera por todas las zonas objetivo alrededor del territorio nacional, las entidades públicas podrían distinguir las problemáticas que impactan el desarrollo de cada región con el fin de tomar de decisiones informadas para la implementación de medidas priorizadas que apoyen a la eliminación de las necesidades insatisfechas expresadas por los ODS.

c. Importancia y justificación que tiene la existencia de la aplicación para el usuario

A pesar de que cada ODS cubre una necesidad específica, el escenario de acción es amplio. Adicionalmente, quedan menos de 7 años para lograr la calidad de vida objetivo planteada por la UNFPA para 2030. En consecuencia, la identificación oportuna de las problemáticas definidas por los ODS en las zonas nacionales de estudio a través de las predicciones realizadas por la aplicación permite dirigir los esfuerzos de desarrollo sostenible de manera eficiente. De igual manera, el reentrenamiento permite que el modelo no quede obsoleto y siga generando valor para el negocio sin importar la constante información que recibiría.

7. Análisis de opciones de reentrenamiento

a. Fine-tuning [IBM, 2023]: pondera sobre el modelo ya entrenado para tomarlo como referencia para conjuntos de datos de menor de tamaño, donde cada uno representa un escenario específico de los datos dentro del contexto donde se aplicaría el modelo.
    i. Ventaja: su adaptabilidad logra que sean mayormente adecuados para los contextos en los que se implementa. En sí, es personalizado, entendiendo patrones de comportamiento complejos.
    ii. Desventaja: su diversidad implica que necesita de diferentes configuraciones específicas para lograr el rendimiento objetivo de un modelo.
b. Domain adaptation [IEEE, 2023]: tiene como objetivo encontrar un modelo matemático enfocado en llegar al dominio de un conjunto de datos aplicado a partir de un modelo entrenado sobre un conjunto de datos origen. Requiere que ambos dominios, con diferentes distribuciones de los datos, tengan algunas características similares.
    i. Ventaja: Además de reducir el posible impacto (sesgo) de los cambios de dominio, es capaz de mantener un buen desempeño en el nuevo dominio.

ii. Desventaja: trabaja adecuadamente de un dominio origen a dominio destino, aunque no es degrada su desempeño al intentar devolverse desde el dominio destino al dominio origen.

c. Continual learning [NEPT, 2022]: el modelo es entrenado por conjuntos de datos individuales aisladamente; los datos históricos no están dentro. Es decir que cada conjunto de datos es usado como única muestra y no se suele repetir para el entrenamiento. Este enfoque tiene como propósito el entrenamiento efectivo de modelos usando aproximaciones prácticas.
    i. Ventaja: es adecuado para escenarios donde varios usuarios desean procesar información ligeramente diferente, como los estilos de redacción, por lo que se ajusta gradualmente a la información.
    ii. Desventaja: encontrar la forma adecuada en que debe configurarse adecuadamente el modelo para evitar que haga sobreajuste sobre los datos actuales u olvide los anteriores no es fácil.
    iii. Razón de elección: dado el contexto en el que se usa el modelo, se decidió implementar esta opción de reentrenamiento por su capacidad de adaptarse al flujo de datos dinámico, como lo son las opiniones, de manera rápida.

8. Trabajo en equipo

a. Roles
    i. Líder de proyecto: Santiago Tapias.
        1. Tareas realizadas
            a. Resolución de problemas en APIs.
            b. Mejoras en el pipeline del endpoint 1.
            c. Aporte en el backend.
            d. Aporte en el front-end del proyecto del endpoint 1.
        2. Horas dedicadas: 24.
        3. Retos enfrentados: la comunicación entre el front y al API era fluida, pero cuando se involucraba el joblib, los errores aparecieron y no eran fáciles de corregir.
            a. Solución planteada: como el pipeline no serializaba las funciones, se decidió realizar la definición de una clase que implementara parte de las funciones deseadas.
    ii. Ingeniero de datos: Felipe Lancheros.
        1. Tareas realizadas
            a. Construcción de pipeline.
            b. Calidad de la automatización del modelo analítico.
        2. Horas dedicadas: 24.
        3. Retos enfrentados: rutas de decisión dentro del pipeline.
            a. Solución planteada: dado que un pipeline es como una caja negra, no se puede acceder a una parte intermedia del proceso que realiza. Por eso, dentro de

lo que probé, realicé segmentación del pipeline en puntos donde se requiriera la intervención manual.

iii. Ingeniero de software: Obed Cabanzo.
    1. Tareas realizadas
        a. Diseño de la aplicación.
        b. Proceso de construcción de la aplicación.
        c. Desarrollo de la aplicación final.
    2. Horas dedicadas: 24.
    3. Retos enfrentados: uso de pipelines.
        a. Solución planteada: revisión de la documentación con el fin de tener mayor claridad frente a los problemas en la implementación.

b. Distribución de puntos
    i. Santiago Tapias: 33.
    ii. Felipe Lancheros: 33.
    iii. Obed Cabanzo: 33.

c. Puntos a mejorar
    i. Debemos realizar un proceso de diseño con antelación para lograr la solución de posibles dudas lo más pronto posible para que la implementación sea más directa.
    ii. Es importante planear el tiempo que se va a dedicar en el desarrollo del proyecto para evitar posibles cruces con otros asuntos.

9. Referencias

a. [UNDP, 2024]: Objetivos de Desarrollo Sostenible | Programa De Las Naciones Unidas Para El Desarrollo. (n.d.). UNDP. https://www.undp.org/es/sustainable-development-goals
b. [UNFPA, 2024]: UNFPA en Colombia. (n.d.). UNFPA-Colombia. https://colombia.unfpa.org/es/unfpa-en-colombia
c. [DANE, 2022]: DANE. (2022). Mujeres y hombres: brechas de género en Colombia. Recuperado 3 de septiembre de 2024, de https://www.dane.gov.co/files/investigaciones/genero/publicaciones/mujeres-y-hombre-brechas-de-genero-colombia-resumen-ejecutivo-2daEdicion.pdf
d. [DANE, 2022]: DANE. (2022). Mujeres y hombres: brechas de género en Colombia. Recuperado 3 de septiembre de 2024, de https://www.dane.gov.co/files/investigaciones/genero/publicaciones/
e. [NEPT, 2022]: Wojcik, M. (2024, August 22). Continual Learning: methods and application. neptune.ai. https://neptune.ai/blog/continual-learning-methods-and-application
f. [IEEE, 2023]: P. Singhal, R. Walambe, S. Ramanna and K. Kotecha, "Domain Adaptation: Challenges, Methods, Datasets, and Applications," in IEEE Access, vol. 11, pp. 6973-7020, 2023, doi: 10.1109/ACCESS.2023.3237025.

g. [IBM, 2023]: Bergmann, D. (2023, 12 junio). Fine Tuning. IBM. https://www.ibm.com/es-es/topics/fine-tuning