TAREA DE APRENDIZAJE

El tipo de aprendizaje es supervisado ya que es una tarea de clasificación sobre datos etiquetados. Se busca predecir a qué ODS pertenece una opinión dentro de tres posibles resultados: ODS 1, fin de la pobreza; ODS 3, salud y bienestar; ODS 4, educación de calidad. Así, se busca obtener de forma inmediata el tipo de ODS que clasifica el texto.

DECISIONES

Los resultados del modelo se convierten en recomendaciones para el usuario final al poder agrupar las opiniones de los ciudadanos con respecto a los ODS. Se puede ver qué tanto hablan de estos y luego se podría analizar sus opiniones con respecto a su perspectiva.

PROPUESTA DE VALOR

El beneficiario final son los gobiernos locales y las ONG que buscan realizar el seguimiento del progreso de los ODS. También es útil para los ciudadanos ya que sus ideas pueden ser transmitidas de una forma más sencilla. En cuanto a la empresa, esta es el Fondo de Población de las Naciones Unidas (UNFPA), que también colabora con actores territoriales y autoridades públicas.

En cuanto a las dificultades que abordan, estas son principalmente dos. La primera es que hay una gran volumen de opiniones, lo cual dificulta la lectura de las opiniones. Además, la falta de categorización de estas opiniones también puede entorpecer el proceso de toma de decisiones ya que los responsables de cada ODS no tendrían un contenido focalizado.

Por último, los errores que pueden ocurrir es que se realice una clasificación erronea ya que esto llevaría la opinión al departamento incorrecto. También puede ocurrir que los datos tengan ciertos sesgos o incongruencias al momento de ser publicados lo cual lleve a que la opinión sea díficil de interpretar y no sea tan útil para la toma de decisiones.

FUENTES DE DATOS

Para el desarrollo de este proyecto, la fuente de datos principal fue un archivo Excel que contiene textos, los cuales son opiniones, y una columna labels, la cual hace referencia a que ODS se asocia la opinión. Debido a que la información es sacada en su totalidad de ese Excel, no hay uso de APIs, aunque en el despliegue real se podrían implementar para recopilar textos de redes sociales o bases de datos externas.

Esta fuente de datos es suficiente para lograr realizar el objetivo del análisis, pues se tiene todo lo necesario para entrenar los modelos de aprendizaje supervisado (los textos y los labels de cada texto). Sin embargo, es importante resaltar que existe un desbalance de clases, lo cual se debería mirar si es necesario tratar, o si por la naturaleza de los datos siempre se genera un desbalance.

SIMULACIÓN DE IMPACTO

En cuanto a las decisiones correctas, cuando el modelo logra clasificar de manera correcta, se logra ahorrar tiempo en un análisis manual del texto y una mejora en los procesos de categorización de las opiniones. Permitiendo tomar acciones basadas en las categorizaciones del modelo, de esta manera automatizando una parte del proceso. Sin embargo, cuando un texto es clasificado de manera errónea, se pueden generar conclusiones equívocas, sesgar indicadores o afectar las decisiones institucionales basadas en la clasificación.

Los criterios de éxito del modelo son principalmente tener un F1 score alto para todas las clases, lograr un recall alto en clases críticas (como podría serlo la minoritaria) y que el modelo generalice de manera correcta un conjunto de datos de prueba, para verificar que este no tenga un sobre ajuste. Por último, si existen restricciones de equidad debido a que las clases están desbalanceadas. Esto podría llevar al modelo a favorecer sistémicamente la clase con más datos.

APRENDIZAJE (USO DEL MODELO)

En este caso, lo más natural es hacer uso del modelo por lotes, cargando conjuntos de textos y que el modelo realice la clasificación de todos estos a la vez. En cuanto a la frecuencia del uso del modelo, se esperaría usar el modelo de manera periódica, con un uso semanal o mensual dependiendo del volumen de la llegada de datos, es decir que si se espera que se generen muchos datos de manera diaria lo ideal sería usar el modelo diariamente, pero si la llegada de datos es baja se podría usar de manera semanal o incluso mensual. Esta última parte depende principalmente de cuantas opiniones se espera que sean generadas en un periodo de tiempo dado.

CONSTRUCCIÓN DE MODELOS

Se necesita un mínimo de tres modelos para poder comparar resultados y definir cuál de estos es el mejor en categorización de ODS. La actualización de estos debería ser constante con nuevo lenguaje natural, en la segunda etapa del proyecto sé debe reentrenar el modelo. Para el procesamiento de caracteres y el entrenamiento del modelo se tiene aproximadamente dos semanas.

INGENIERÍA DE CARACTERÍSTICAS

Las variables importantes son los ODS 1, 3 y 4, también están los tókenes que se extraeran al procesar los datos, la frecuencia de las diferentes palabras que están en los datos y la relacion de estas con las etiquetas. Por otro lado, la transformación que se van a hacer son, limpieza de datos removiendo los caracteres que no sean ASCII, convirtiendo todo a minúsculas, quitando los signos de puntuación, removiendo las stopwords y intercambiando los numero por palabras y se normalizan las palabras.