PROYECTO BI – ETAPA 1

Boris N. Reyes R.

202014743

ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO

OPORTUNIDAD/PROBLEMA NEGOCIO

El UNFPA enfrenta el reto de procesar grandes volúmenes de opiniones de comunidades locales para identificar y priorizar problemas relacionados con la salud, la educación y la igualdad de género (ODS 3, 4 y 5). La dificultad radica en analizar esta información de manera efectiva para tomar decisiones informadas. Implementar un modelo analítico basado en la analítica de textos puede automatizar la clasificación de estas opiniones, permitiendo identificar rápidamente las preocupaciones prioritarias, priorizar recursos y acciones de manera precisa, y mejorar la toma de decisiones. Esto no solo acelera el análisis, sino que también proporciona insights más profundos para una respuesta más efectiva a las necesidades comunitarias.

OBJETIVOS Y CRITERIOS DE ÉXITO DESDE EL PUNTO DE VISTA DEL NEGOCIO

Objetivos

1. **Optimizar la Identificación de Problemas Críticos:**
   Automatizar la clasificación de opiniones locales para identificar problemas vinculados a los ODS 3, 4 y 5.
2. **Mejorar la Toma de Decisiones Informadas:**
   Priorizar recursos y acciones basados en datos textuales, alineando intervenciones con las necesidades reales de la comunidad.
3. **Incrementar la Eficiencia Operativa:**
   Reducir el tiempo y esfuerzo manual en el análisis de datos textuales, permitiendo a los equipos centrarse en la implementación de soluciones.

Criterios de Éxito

1. **Precisión en la Clasificación:**
   El modelo debe alcanzar una precisión alta ($\geq 85\%$) en la clasificación de los ODS.
2. **Alineación con Necesidades Locales:**
   Las acciones deben reflejar claramente las preocupaciones identificadas, validadas por la predicción sobre los datos de pruebas y a través de una retroalimentación.
3. **Impacto Medible:**

Mejoras tangibles en indicadores clave de salud, educación e igualdad de género, atribuibles a las decisiones basadas en el modelo.

ORGANIZACIÓN Y ROL DENTRO DE ELLA QUE SE BENEFICIA CON LA OPORTUNIDAD DEFINIDA

Organización:

UNFPA (Fondo de Población de las Naciones Unidas)

Rol que se Beneficia:

Analistas de Programas y Coordinadores de Proyectos del UNFPA:

Los Analistas de Programas y Coordinadores de Proyectos son responsables de diseñar, implementar y supervisar las iniciativas asociadas a los ODS. Estos roles requieren una comprensión profunda de las necesidades de la población y la capacidad de priorizar intervenciones basadas en datos relevantes.

Beneficios para el Rol:

El modelo analítico mejorará la toma de decisiones al ofrecer insights precisos y dirigidos. Reducirá el tiempo de análisis manual, permitiendo un enfoque mayor en la implementación de proyectos. Optimizará la asignación de recursos hacia problemas críticos, aumentando la efectividad de las intervenciones.

IMPACTO QUE PUEDE TENER EN COLOMBIA ESTE PROYECTO

El proyecto tiene el potencial de generar un impacto significativo en varias áreas clave de desarrollo en Colombia:

1. Mejoras en la Salud y Bienestar (ODS 3):

- **Impacto:** Al identificar y abordar problemas críticos relacionados con la salud mediante un análisis preciso de las opiniones locales, el proyecto puede contribuir a mejorar el acceso a servicios de salud, reducir la mortalidad materna e infantil, y combatir enfermedades. Esto puede tener un efecto directo en el bienestar general de las comunidades, especialmente en regiones vulnerables.

2. Aumento en la Calidad de la Educación (ODS 4):

- **Impacto:** Al clasificar y priorizar las preocupaciones educativas expresadas por la población, el proyecto puede facilitar la implementación de políticas y programas que mejoren la calidad educativa. Esto es crucial para reducir la desigualdad en el acceso a la educación y asegurar que más niños y jóvenes en Colombia reciban una educación de calidad.

3. Promoción de la Igualdad de Género (ODS 5):

- **Impacto:** El proyecto puede ayudar a identificar barreras específicas para la igualdad de género, como la violencia de género o la discriminación en el lugar de trabajo, y guiar las acciones para eliminarlas. Esto fortalecerá los esfuerzos hacia la igualdad de género y empoderará a las mujeres y niñas en Colombia.

Este impacto se verá reflejado en el largo plazo, con mejoras en la calidad de vida, la equidad y la sostenibilidad en las comunidades colombianas, haciendo de este proyecto una iniciativa crucial para el progreso social y económico del país.

ENFOQUE ANALÍTICO. DESCRIPCIÓN DE LA CATEGORÍA DE ANÁLISIS (DESCRIPTIVO, PREDICTIVO, ETC.), TIPO Y TAREA DE APRENDIZAJE E INCLUYA LAS TÉCNICAS Y ALGORITMOS QUE PROPONE UTILIZAR

Categoría de Análisis:

1. Predictivo:

- El proyecto busca predecir la clasificación de nuevas opiniones en relación con los ODS 3, 4 y 5, basándose en un conjunto de datos etiquetados previamente.

Tipo de Análisis:

1. Supervisado:

- Dado que el conjunto de datos incluye etiquetas ("sdg") para cada opinión, el enfoque supervisado es el más adecuado para entrenar el modelo.

Tarea de Aprendizaje:

1. Clasificación:

- La tarea principal es clasificar las opiniones en una de las tres categorías correspondientes a los ODS (3, 4, 5).

Técnicas y Algoritmos Propuestos:

1. Procesamiento de Lenguaje Natural (PLN):

- **Técnica:** Tokenización, lematización y eliminación de stop words para preparar los textos para análisis.
- **Herramientas:** NLTK, SpaCy.

2. TF-IDF (Term Frequency-Inverse Document Frequency):

- **Técnica:** Asignar pesos a palabras clave para destacar las más relevantes en las opiniones, ayudando en la clasificación.

3. Modelos de Clasificación:

- **Algoritmos:**
    - **Naive Bayes:** Adecuado para la clasificación de texto por su simplicidad y efectividad.
    - **Support Vector Machines (SVM):** Eficaz para la clasificación de opiniones en múltiples categorías.
    - **Random Forest:** Algoritmo de ensamble que combina múltiples árboles de decisión para mejorar la precisión de la clasificación.

El éxito del modelo se medirá por su capacidad para clasificar correctamente un conjunto de datos de prueba, asegurando que las nuevas opiniones se asignen correctamente a su respectivo ODS.

ENTENDIMIENTO Y PREPARACIÓN DE LOS DATOS

Se realizó un análisis sobre la calidad de los datos para asegurar que cumplieran con los criterios de completitud, unicidad, validez y consistencia. No se identificaron problemas por lo que no se tomaron medidas correctivas en esta parte.

El preprocesamiento de los textos incluyó normalización y limpieza, donde los textos fueron convertidos a minúsculas, se eliminaron signos de puntuación, y se realizaron correcciones de caracteres extraños resultantes de problemas de codificación. Sin embargo, surgió un problema con caracteres acentuados como "número," que fue procesado incorrectamente como "naomero" en ciertos pasos. Para resolverlo, se buscó emplear técnicas adicionales para asegurar que los caracteres especiales fueran correctamente interpretados, **pero no fue posible una corrección apropiada**. Por otra parte, se realizó la tokenización y eliminación de stopwords para reducir el ruido en los datos.

MODELADO Y EVALUACIÓN

En la Fase 3 del proyecto, se aplicaron diversos algoritmos de aprendizaje automático para la tarea de clasificación de textos en relación con los Objetivos de Desarrollo Sostenible (ODS) del UNFPA. A continuación, se describe cada uno de los algoritmos empleados y se explica por qué son apropiados para el procesamiento de textos y su clasificación en los ODS.

Naive Bayes

Este algoritmo de aprendizaje automático supervisado se utiliza para tareas de clasificación, como la clasificación de texto. Utiliza principios de probabilidad para realizar tareas de clasificación. Es útil cuando se dispone de un gran número de características (palabras) y se requiere un modelo eficiente y fácil de interpretar por lo que es una opción viable para el proyecto.

Support Vector Machine (SVM)

"SVM funciona correlacionando datos a un espacio de características de grandes dimensiones de forma que los puntos de datos se puedan categorizar, incluso si los datos no se puedan separar linealmente de otro modo. Se detecta un separador entre las categorías y los datos se transforman de forma que el separador se puede extraer como un hiperplano. Tras ello, las características de los nuevos datos se pueden utilizar para predecir el grupo al que pertenece el nuevo registro." (IBM, Funcionamiento de SVM, 2024).

La capacidad del modelo para encontrar el margen óptimo entre clases ayuda a mejorar la precisión de la clasificación, por lo que se considera para el análisis de opiniones sobre los ODS.

Random Forest

El tercer algoritmo es Random Forest. Este algoritmo "Combina los resultados de múltiples árboles de decisión para llegar a un resultado único. Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja tanto problemas de clasificación como de regresión."(IBM, What is random forest?). La técnica de agregación de múltiples árboles permite capturar patrones complejos siendo una buena opción para el proyecto.

RESULTADOS

Resultados de los Modelos

- **Naive Bayes**
    - **Precisión General (Accuracy):** 0.97
    - **Precisión (Macro Avg):** 0.98
    - **Recall (Macro Avg):** 0.97
    - **F1-Score (Macro Avg):** 0.97
- **Support Vector Machine (SVM)**
    - **Precisión General (Accuracy):** 0.9827
    - **Precisión (Macro Avg):** 0.9828

- **Recall (Macro Avg):** 0.9827
- **F1-Score (Macro Avg):** 0.9827

- **Random Forest**
  - **Precisión General (Accuracy):** 0.97
  - **Precisión (Macro Avg):** 0.97
  - **Recall (Macro Avg):** 0.97
  - **F1-Score (Macro Avg):** 0.97

Análisis de Métricas

El modelo SVM logró una precisión general del 98.27%, superando a Naive Bayes y Random Forest, lo que demuestra una alta tasa de clasificación correcta. Con una precisión del 98.28%, SVM asegura una alta tasa de predicciones positivas correctas, crucial para orientar estrategias de manera precisa. El recall también es alto, con un 98.27%, indicando que el modelo identifica casi todos los casos positivos. El F1-score del 98.27% muestra un equilibrio excelente entre precisión y recall, evitando errores de clasificación significativos.

El elevado rendimiento en precisión y recall del modelo SVM facilita una clasificación precisa de las opiniones en relación con los ODS, mejorando la identificación de áreas que necesitan intervención. El alto F1-score asegura una detección efectiva de problemas y minimiza los errores de clasificación, lo que es esencial para diseñar estrategias que aborden de manera eficiente las preocupaciones comunitarias.

El modelo SVM ha identificado palabras clave relacionadas con los ODS mediante técnicas de Procesamiento de Lenguaje Natural (PLN) y TF-IDF. Las palabras identificadas reflejan preocupaciones en salud, educación y género. Basado en estos resultados, se recomienda desarrollar estrategias específicas para cada ODS: mejorar los servicios de salud y prevención para el SDG 3, optimizar la infraestructura educativa y capacitación para el SDG 4, y promover la igualdad de género a través de iniciativas y políticas para el SDG 5.

La identificación precisa de opiniones permite enfocar los recursos y esfuerzos de manera más efectiva, aumentando el impacto positivo en la comunidad. Utilizar el modelo SVM para clasificar opiniones asegura una asignación óptima de recursos hacia áreas con mayores necesidades. La retroalimentación del análisis permite ajustar y mejorar continuamente los programas y políticas, alineándose mejor con las necesidades de la comunidad y los objetivos de desarrollo sostenible.

MAPA DE ACTORES

Rol en la Organización	Tipo de Actor	Beneficio	Riesgo
Coordinador de programas	Usuario, cliente	Mejora del manejo de recursos en proyectos asociados a los ODS 3, 4 y 5	Si el modelo se equivoca se pueden tomar decisiones poco acertadas generando pérdidas de recursos
Equipos de procesamiento manual de opiniones	Usuario, cliente	Reduce tiempo y recursos destinados a personal lo que permite tomar decisiones en corto tiempo	Si el modelo presenta errores puede conducir a la toma de decisiones equivocadas
Donantes/Inversionistas	Inversor	Que el UNFPA tome mejores decisiones aumenta la probabilidad de éxito en sus proyectos dando resultados positivos a los inversionistas	Pérdida de recursos a raíz de enfoques y decisiones erróneas ocasionando frustración e inconformidad

TRABAJO EN EQUIPO

El proyecto fue desarrollado por una sola persona. Se destinaron días para avanzar en las diferentes etapas y lograr avances constantes. El espaciar el trabajo permitió generar un análisis consciente del avance del proyecto tomando medidas correctivas. Desde la fase de investigación y exploración para el correcto entendimiento del negocio hasta la generación de los datos finales para entregar al cliente, el proceso fue detallado y documentado buscando compartir de la manera más completa las decisiones que motivaron cada fase.

FUENTES DE INFORMACIÓN

https://www.unfpa.org/

https://www.ibm.com/mx-es/topics/naive-bayes#:~:text=El%20clasificador%20Na%C3%AFve%20Bayes%20es,para%20realizar%20tareas%20de%20clasificaci%C3%B3n.

https://developer.ibm.com/tutorials/awb-classifying-data-multinomial-naive-bayes-algorithm/

https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works