Proyecto 1 Etapa 1 – Analítica de textos

Evelin Vanessa Villamil Guerrero – 202113360

Carlos German Monroy Andrade – 201728260

Objetivos

1. Aplicar la metodología de analítica de textos para la construcción de soluciones de analítica alineadas con los objetivos del negocio en un contexto de aplicación.
2. Planear la interacción con un grupo interdisciplinario para identificar usuarios y posibles herramientas a desarrollar para la interacción del resultado del modelo desarrollado.

Introducción

Se quiere desarrollar un modelo de clasificación, con técnicas de aprendizaje automático, que permita relacionar de manera automática un texto según los ODS. Al igual que desarrollar una aplicación que facilite la interacción con el resultado de dicho modelo. El modelo podrá ser utilizado entonces para la interpretación y análisis de la información textual que es recopilada a través de diferentes fuentes por UNFPA en procesos de planeación participativa para el desarrollo a nivel territorial.

Tabla de contenido

1. Entendimiento del negocio y enfoque analítico

Los Objetivos de Desarrollo Sostenible (ODS), también conocidos como Objetivos Mundiales, son un conjunto de 17 objetivos interconectados y 169 metas que fueron adoptados por todos los estados miembros de las Naciones Unidas en septiembre de 2015 como parte de la Agenda 2030 para el Desarrollo Sostenible. Estos objetivos representan un llamado universal a la acción para poner fin a la pobreza, proteger el planeta y asegurar que todas las personas gocen de paz y prosperidad.

El proyecto se enfoca en tres ODS específicos (6, 7 y 16):

Objetivo 6: Garantizar la disponibilidad de agua y su gestión sostenible y el saneamiento para todos. Algunas metas de este objetivo son:

6.1 De aquí a 2030, lograr el acceso universal y equitativo al agua potable a un precio asequible para todos

6.2 De aquí a 2030, lograr el acceso a servicios de saneamiento e higiene adecuados y equitativos para todos y poner fin a la defecación al aire libre, prestando especial atención a las necesidades de las mujeres y las niñas y las personas en situaciones de vulnerabilidad

6.3 De aquí a 2030, mejorar la calidad del agua reduciendo la contaminación, eliminando el vertimiento y minimizando la emisión de productos químicos y materiales peligrosos, reduciendo a la mitad el porcentaje de aguas residuales sin tratar y aumentando considerablemente el reciclado y la reutilización sin riesgos a nivel mundial

6.4 De aquí a 2030, aumentar considerablemente el uso eficiente de los recursos hídricos en todos los sectores y asegurar la sostenibilidad de la extracción y el abastecimiento de agua dulce para hacer frente a la escasez de agua y reducir considerablemente el número de personas que sufren falta de agua

6.5 De aquí a 2030, implementar la gestión integrada de los recursos hídricos a todos los niveles, incluso mediante la cooperación transfronteriza, según proceda

6.6 De aquí a 2020, proteger y restablecer los ecosistemas relacionados con el agua, incluidos los bosques, las montañas, los humedales, los ríos, los acuíferos y los lagos

6.a De aquí a 2030, ampliar la cooperación internacional y el apoyo prestado a los países en desarrollo para la creación de capacidad en actividades y programas relativos al agua y el saneamiento, como los de captación de agua, desalinización,

uso eficiente de los recursos hídricos, tratamiento de aguas residuales, reciclado y tecnologías de reutilización

6.b Apoyar y fortalecer la participación de las comunidades locales en la mejora de la gestión del agua y el saneamiento

Objetivo 7: Garantizar el acceso a una energía asequible, segura, sostenible y moderna. Algunas metas de este objetivo son:

7.1 De aquí a 2030, garantizar el acceso universal a servicios energéticos asequibles, fiables y modernos

7.2 De aquí a 2030, aumentar considerablemente la proporción de energía renovable en el conjunto de fuentes energéticas

7.3 De aquí a 2030, duplicar la tasa mundial de mejora de la eficiencia energética

7.a De aquí a 2030, aumentar la cooperación internacional para facilitar el acceso a la investigación y la tecnología relativas a la energía limpia, incluidas las fuentes renovables, la eficiencia energética y las tecnologías avanzadas y menos contaminantes de combustibles fósiles, y promover la inversión en infraestructura energética y tecnologías limpias

7.b De aquí a 2030, ampliar la infraestructura y mejorar la tecnología para prestar servicios energéticos modernos y sostenibles para todos en los países en desarrollo, en particular los países menos adelantados, los pequeños Estados insulares en desarrollo y los países en desarrollo sin litoral, en consonancia con sus respectivos programas de apoyo

Objetivo 16: Promover sociedades justas, pacíficas e inclusivas. Algunas metas de este objetivo son:

16.1 Reducir significativamente todas las formas de violencia y las correspondientes tasas de mortalidad en todo el mundo

16.2 Poner fin al maltrato, la explotación, la trata y todas las formas de violencia y tortura contra los niños

16.3 Promover el estado de derecho en los planos nacional e internacional y garantizar la igualdad de acceso a la justicia para todos

16.4 De aquí a 2030, reducir significativamente las corrientes financieras y de armas ilícitas, fortalecer la recuperación y devolución de los activos robados y luchar contra todas las formas de delincuencia organizada

16.5 Reducir considerablemente la corrupción y el soborno en todas sus formas

16.6 Crear a todos los niveles instituciones eficaces y transparentes que rindan cuentas

16.7 Garantizar la adopción en todos los niveles de decisiones inclusivas, participativas y representativas que respondan a las necesidades

16.8 Ampliar y fortalecer la participación de los países en desarrollo en las instituciones de gobernanza mundial

16.9 De aquí a 2030, proporcionar acceso a una identidad jurídica para todos, en particular mediante el registro de nacimientos

16.10 Garantizar el acceso público a la información y proteger las libertades fundamentales, de conformidad con las leyes nacionales y los acuerdos internacionales

16.a Fortalecer las instituciones nacionales pertinentes, incluso mediante la cooperación internacional, para crear a todos los niveles, particularmente en los países en desarrollo, la capacidad de prevenir la violencia y combatir el terrorismo y la delincuencia

16.b Promover y aplicar leyes y políticas no discriminatorias en favor del desarrollo sostenible

Impacto en Colombia

El cumplimiento de estos ODS en Colombia puede tener un impacto significativo en la vida de la población colombiana y en el desarrollo sostenible del país. Aquí hay ejemplos de cómo cada uno de estos ODS puede beneficiar a Colombia:

- **ODS 6 – Agua limpia y saneamiento:** Al garantizar el acceso a agua potable y saneamiento adecuado, se mejora la salud pública, se promueve la seguridad alimentaria, se conservan los ecosistemas acuáticos, se fortalece la resiliencia al cambio climático y se reduce la desigualdad. Además, la gestión sostenible del agua fomenta el turismo sostenible, contribuye al desarrollo económico y la creación de empleo.
- **ODS 7 – Energía asequible y no contaminante:** Puede impulsar el crecimiento económico a través de la promoción de fuentes de energía limpia, reducir la pobreza energética, promover la sostenibilidad ambiental y mitigar el cambio climático. Además, garantizar el acceso universal a la electricidad puede mejorar la calidad de vida en áreas rurales y marginadas,

y la inversión en tecnología y conocimiento puede estimular la innovación en el sector energético colombiano.

- **ODS 16 – Paz, justicia e instituciones sólidas:** Contribuir a la consolidación de la paz, el fortalecimiento del estado de derecho, la promoción de los derechos humanos y la reducción de la violencia. Esto se traduce en un entorno más seguro, una mejor gobernanza, mayor participación ciudadana y el fortalecimiento de las instituciones gubernamentales. Estos aspectos son fundamentales para el proceso de paz en Colombia y para impulsar el desarrollo sostenible en el país.
-

Objetivos de negocio específicos

- **Automatización de la Clasificación de Textos**

Desarrollar un modelo de clasificación de textos basado en técnicas de aprendizaje automático que permita asignar automáticamente un texto a uno de los Objetivos de Desarrollo Sostenible (ODS) específicos. Es importante porque agiliza el proceso de identificar los temas relacionados con los ODS en grandes volúmenes de datos, lo que ahorra tiempo y recursos.

- **Mejora de la Eficiencia en la Evaluación de Políticas Públicas**

Facilitar la identificación y evaluación de políticas públicas relacionadas con los ODS mediante el análisis automatizado de opiniones y comentarios de la población local. Es importante porque la mejora de la eficiencia en la evaluación de políticas públicas permite a UNFPA y otras entidades identificar desafíos y oportunidades más rápidamente, lo que a su vez acelera el proceso de toma de decisiones.

- **Identificación de Problemas y Soluciones de Manera Más Rápida y Precisa**

Proporcionar una herramienta que permita identificar problemas y soluciones en relación con los ODS en un contexto territorial de manera más rápida y precisa. Es importante porque la capacidad de identificar problemas y soluciones de manera eficiente ayuda a enfocar los esfuerzos en áreas críticas y a lograr un impacto más significativo en el desarrollo sostenible.

Criterios de Éxito

- **Precisión de la Clasificación de Textos**

Medir la precisión del modelo de clasificación de textos en asignar correctamente los textos a los ODS específicos. Un alto nivel de precisión es fundamental para la utilidad de la herramienta.

- **Eficiencia en el Procesamiento de Datos**

Evaluar la eficiencia del proceso de procesamiento de datos, incluyendo la velocidad de clasificación de textos y la capacidad de manejar grandes volúmenes de datos de manera oportuna.

- **Tiempo de Respuesta en la Identificación de Problemas y Soluciones**

Medir cuánto tiempo se ahorra en la identificación de problemas y soluciones en comparación con enfoques anteriores.

- **Impacto en la Toma de Decisiones**

Evaluar cómo el proyecto contribuye a la toma de decisiones más informadas y efectivas en la implementación de políticas relacionadas con los ODS.

Enfoque Analítico

1. Selección de algoritmos:

Random Forest: Es un algoritmo de aprendizaje supervisado que se utiliza tanto para la clasificación como para la regresión. Este algoritmo se basa en la combinación de múltiples árboles de decisión, donde cada árbol se entrena con una submuestra aleatoria del conjunto de datos de entrenamiento. Durante la predicción, el algoritmo promedia las predicciones de todos los árboles de decisión para obtener la salida final. Puede ser especialmente útil en la clasificación de texto para manejar la alta dimensionalidad de los datos.

GradientBoostingClassifier: El algoritmo Gradient Boosting es una técnica de aprendizaje automático que se utiliza para mejorar la precisión de un modelo de clasificación. Es un enfoque de conjunto (ensemble) que combina varios modelos de aprendizaje débil en un modelo de aprendizaje fuerte. En cada iteración, el algoritmo construye un nuevo modelo débil que se enfoca en los casos que el modelo anterior no ha clasificado correctamente, y agrega ese modelo al conjunto de modelos débiles ya existentes. Es útil para clasificar un texto según los ODS porque permite crear un modelo preciso y robusto al combinar varios modelos débiles que se enfocan en diferentes aspectos de los datos.

Árboles de Decisión: El algoritmo de árboles de decisión es un método de aprendizaje supervisado utilizado para clasificar datos en categorías o predecir valores numéricos. En este algoritmo, se construye un árbol de decisiones a partir de los datos de entrenamiento, donde cada nodo del árbol representa una característica del conjunto de datos y las ramas representan las posibles respuestas

a esa característica. El árbol se construye de forma recursiva, dividiendo el conjunto de datos en subconjuntos más pequeños y homogéneos según ciertas características hasta alcanzar una determinada profundidad o un criterio de parada. Puede resultar útil para entrenar modelos de clasificación de textos según ODS porque permite identificar las características más relevantes para hacer la clasificación, como las palabras más utilizadas

2. Preprocesamiento de Datos:

Tokenización: Se dividirán los textos en palabras o tokens, lo que permitirá analizar las palabras individualmente.

Eliminación de Stop Words: Se eliminarán las palabras comunes (stop words) que no aportan información significativa a la clasificación.

Lematización o Stemming: Se reducirán las palabras a su forma base para eliminar las variaciones de palabras, lo que ayuda a agrupar palabras similares.

Vectorización de Texto: Los textos se convertirán en vectores numéricos utilizando técnicas como TF-IDF (Term Frequency-Inverse Document Frequency) para representar la importancia de las palabras en los textos.

3. Entrenamiento del Modelo:

Recopilación de Datos Etiquetados: Se recopilarán textos relacionados con los ODS que estén etiquetados con la categoría correspondiente (ODS 6, 7 o 16). Estos datos se utilizarán para entrenar y evaluar el modelo.

Selección de Características: Se seleccionarán las características más relevantes, como vectores TF-IDF, para representar los textos en el modelo.

4. Evaluación y Ajuste:

División de Datos: Se dividirán los datos en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo.

Evaluación del Modelo: Se utilizarán métricas de evaluación, como precisión, recall, F1-score y matriz de confusión, para evaluar la capacidad del modelo para clasificar los textos correctamente.

Ajuste del Modelo: Se realizarán ajustes en los hiperparámetros de los algoritmos, la profundidad de los árboles en Random Forest, Gradient Boosting Classifier y Árboles de decisión, para mejorar el rendimiento del modelo.

Validación Cruzada: Se puede aplicar validación cruzada para evaluar la capacidad de generalización del modelo en diferentes conjuntos de datos.

La evaluación y ajuste se repetirá hasta que se obtenga un modelo preciso que pueda automatizar la clasificación de textos relacionados con los ODS de manera efectiva y se cumplan los criterios de éxito del proyecto.

Tabla con la información solicitada:

Oportunidad/problema Negocio	Automatización de la Clasificación de Textos relacionados con los ODS
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	Se propone utilizar técnicas de aprendizaje automático, incluyendo Random Forest, Gradient Boosting Classifier Árboles de Decisión. Se aplicará preprocesamiento de datos, que incluye tokenización, eliminación de stop words, lematización, y vectorización de texto. El modelo se entrenará con datos etiquetados y se evaluará utilizando métricas como precisión, recall y F1-score.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Fondo de Poblaciones de las Naciones Unidas (UNFPA)
Contacto con experto externo al proyecto	Emilio Sánchez - e.sanchez1123 Alejandra villa - ma.villa **Fecha reunión:** 18/10/2023 **Canal:** Zoom

2. Entendimiento y preparación de los datos

Inicialmente, se leyó el archivo con la biblioteca pandas y se asignó a una variable llamada "data_t". Luego, se identificó el idioma de los textos y se eliminaron los textos en inglés, seguido de la eliminación de la columna "idioma". Se verificó la ausencia de valores nulos y duplicados en el conjunto de datos. Posteriormente, se realizó una limpieza de los textos que incluyó la corrección de palabras mal codificadas, la eliminación de caracteres especiales y puntuación, la tokenización de las palabras y la normalización del texto, incluyendo la obtención de raíces y lemas de las palabras. Finalmente, se convirtió la columna "words" en una sola cadena de texto y se guardó el nuevo DataFrame en un archivo CSV llamado "clean_cat_6716.csv". Estos pasos nos permitieron preparar los datos para su posterior análisis y modelado.

3. Modelado y evaluación

Random Forest (Evelin Villamil): Se eligió el algoritmo Random Forest debido a su capacidad para manejar múltiples características de texto, lo que es esencial en la clasificación de textos según los ODS. El modelo se entrenó utilizando la representación TF-IDF de los textos. Se realizaron las siguientes etapas:

- División de datos en conjuntos de entrenamiento y prueba.
- Vectorización del texto utilizando TF-IDF.
- Entrenamiento del modelo Random Forest.
- Visualización de la importancia de las características.
- Predicciones en los conjuntos de entrenamiento y prueba.

Los resultados de la evaluación se presentaron en términos de métricas de precisión, recall y F1-score para los conjuntos de entrenamiento y prueba. Se observó que el modelo alcanzó una precisión perfecta en el conjunto de entrenamiento, lo que indica una capacidad excepcional para clasificar correctamente las instancias de entrenamiento. En el conjunto de prueba, el modelo alcanzó una alta precisión, recall y F1-score, con valores en torno a 0.97, lo que demuestra su capacidad para generalizar en datos no vistos. Además, se realizó una validación cruzada con 5 iteraciones, y el modelo obtuvo un puntaje promedio de aproximadamente 0.98, lo que confirma su rendimiento consistente y sólido.

Gradient Boosting Classifier (Carlos Monroy): El algoritmo Gradient Boosting es una técnica de aprendizaje automático que combina varios modelos de aprendizaje débil en un modelo de aprendizaje fuerte. Es una técnica de conjunto que se utiliza para mejorar la precisión de un modelo de clasificación. El modelo se entrenó utilizando la representación TF-IDF de los textos, que convierte el texto en vectores numéricos ponderados, lo que es adecuado para el procesamiento de texto. A continuación, se presentan las etapas seguidas:

- División de datos en conjuntos de entrenamiento y prueba.
- Vectorización del texto utilizando TF-IDF.
- Entrenamiento del modelo Gradient Boosting.
- Visualización de la importancia de las características.
- Predicciones en los conjuntos de entrenamiento y prueba.

En el conjunto de entrenamiento, el modelo logró una ejecución sobresaliente al clasificar todos los casos correctamente, lo que se refleja en una precisión, recall y F1-score de 1. Esto significa que el modelo no cometió errores en la predicción de la clase correcta en el conjunto de entrenamiento, lo que es un indicador de su alta precisión y ausencia de falsos positivos o falsos negativos en este conjunto. En el conjunto de prueba, el modelo mantuvo un sólido rendimiento con una precisión,

recall y F1-score de aproximadamente 0.96. Esto sugiere que el modelo puede identificar correctamente la mayoría de los textos relacionados con los ODS en el conjunto de prueba. Aunque estas métricas son ligeramente más bajas en comparación con el modelo Random Forest, siguen siendo muy buenas. Además, durante la validación cruzada con 5 iteraciones, el modelo obtuvo un puntaje promedio de alrededor de 0.966, lo que indica un buen desempeño general en la precisión de la predicción de las clases.

Árboles de Decisión (Evelin Villamil, Carlos Monroy)

El algoritmo de Árboles de Decisión es un método de aprendizaje supervisado que se adapta bien a esta tarea, ya que es capaz de identificar las características más relevantes en los textos para la clasificación. El modelo se entrenó utilizando la representación TF-IDF de los textos, que convierte el texto en vectores numéricos ponderados, lo que es adecuado para el procesamiento de texto. A continuación, se presentan las etapas seguidas:

- División de datos en conjuntos de entrenamiento y prueba.
- Vectorización del texto utilizando TF-IDF.
- Entrenamiento del modelo Gradient Boosting.
- Visualización de la importancia de las características.
- Predicciones en los conjuntos de entrenamiento y prueba.

En el conjunto de entrenamiento, el modelo logró una precisión del 100%, lo que indica que clasificó correctamente todos los casos en este conjunto. Sin embargo, en el conjunto de prueba, la precisión, recall y F1-score fueron de alrededor del 0.93, lo que aún es un rendimiento muy sólido. Estas métricas indican que el modelo es capaz de identificar la mayoría de los textos relacionados con los ODS en el conjunto de prueba, aunque con una ligera disminución en comparación con los modelos Random Forest y Gradient Boosting Classifier. Además, durante la validación cruzada con 5 iteraciones, el modelo obtuvo un puntaje promedio de aproximadamente 0.936, lo que confirma su buen desempeño en términos de precisión general en la predicción de las clases.

4. Resultados

El algoritmo Random Forest, al ser un método de ensamblado de árboles de decisión, se destacó en comparación con otros enfoques. Su capacidad para combinar múltiples árboles de decisión individuales, basados en muestras aleatorias de datos de entrenamiento y características aleatorias, permitió mejorar la precisión y generalización del modelo. En el contexto de clasificación de textos según los Objetivos de Desarrollo Sostenible (ODS), donde las características pueden ser numerosas y altamente dimensionales debido al procesamiento de texto, este algoritmo demostró ser eficiente al manejar grandes cantidades de datos sin un alto riesgo de sobreajuste.

Los resultados obtenidos con el modelo basado en TF-IDF, particularmente en términos de recall y F1-score, indican que este enfoque es capaz de identificar de manera precisa un mayor número de casos positivos, lo que es esencial en la clasificación de textos relacionados con ODS. La capacidad de identificar relaciones no lineales entre características y etiquetas también se tradujo en un rendimiento superior en comparación con los modelos basados en árboles de decisión o Gradient Boosting.

En cuanto a la validación cruzada, el valor promedio de alrededor de 0.9804 es un indicativo sólido de que el modelo basado en TF-IDF es capaz de generalizar eficazmente a nuevos datos, lo que es esencial en aplicaciones del mundo real.

Los resultados y métricas de calidad obtenidos sugieren que este modelo puede ser una herramienta valiosa para una organización. Puede automatizar la clasificación de textos según los ODS y realizar análisis automatizados de opiniones, lo que podría mejorar la eficiencia operativa y respaldar la toma de decisiones. Esta automatización no solo ahorra tiempo y recursos, sino que también tiene el potencial de mejorar la comprensión de las necesidades y opiniones de la comunidad local, lo que es fundamental en el contexto de los ODS.

Nota:

Los literales ‘Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido’ y ‘Trabajo en equipo’ se incluyen en documentos aparte y en la wiki del repositorio.

Referencias

- Miluska.Jara. (2020, 10 diciembre). *Objetivos y metas de Desarrollo sostenible - Desarrollo sostenible*. Desarrollo Sostenible. https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/
- *DANE - Objetivos de Desarrollo Sostenible - ODS*. (s. f.). https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-informacion/objetivos-de-desarrollo-sostenible-ods