Universidad de los Andes

Inteligencia de Negocios, 2023

Modelo de clasificación de textos según su ODS

Author:
Cuellar Argotty, Juan Esteban
Ortiz Almanza, David Santiago
Vargas Prada, David Santiago

Índice

1. Introducción

En la era actual, donde la sostenibilidad y el desarrollo se han convertido en pilares fundamentales para las organizaciones globales, la Agenda 2030 de las Naciones Unidas destaca como una hoja de ruta esencial para el progreso mundial. Esta agenda, con sus 17 Objetivos de Desarrollo Sostenible (ODS), busca abordar desafíos críticos como la pobreza, la salud, la educación y el impacto ambiental. Sin embargo, para alcanzar estos objetivos, es esencial no solo establecer metas, sino también monitorear y evaluar continuamente el progreso hacia ellas. En este contexto, el Fondo de Población de las Naciones Unidas (UNFPA) ha tomado la iniciativa de colaborar en la clasificación y análisis de información textual relacionada con los ODS. Este documento detalla un proyecto llevado a cabo en colaboración con la Universidad de los Andes, que tiene como objetivo desarrollar un modelo de clasificación basado en técnicas de aprendizaje automático para analizar y categorizar textos según los ODS. A través de este esfuerzo, se busca no solo mejorar la eficiencia en la clasificación de datos, sino también proporcionar insights valiosos que puedan guiar las políticas públicas y las intervenciones relacionadas con el desarrollo sostenible.

2. Entendimiento del negocio y enfoque analítico

2.1. Objetivos y criterios de éxito desde el punto de vista del negocio

Nuestro objetivo principal es implementar un modelo de clasificación, utilizando técnicas de aprendizaje automático, que permita relacionar automáticamente un texto con los Objetivos de Desarrollo Sostenible (ODS). Por lo tanto, el criterio de éxito será que el modelo de clasificación elegido debe ser capaz de procesar y clasificar grandes cantidades de información textual de manera eficiente y precisa (al menos un F1 score de 97 %).

ODS: Los Objetivos de Desarrollo Sostenible (ODS) son 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.

ODS involucrados en el proyecto asignado:

- **ODS 6 (Agua limpia y saneamiento):** Asegurar la disponibilidad y gestión sostenible del agua y el saneamiento para todos.
- **ODS 7 (Energía asequible y no contaminante):** Garantizar el acceso a una energía asequible, segura, sostenible y moderna para todos.
- **ODS 16 (Paz, justicia e instituciones sólidas):** Promover sociedades pacíficas e inclusivas para el desarrollo sostenible, facilitar el acceso a la justicia para todos y crear instituciones eficaces, responsables e inclusivas a todos los niveles.

Impacto en Colombia: En Colombia, los ODS 6, 7 y 16 son clave para nuestro bienestar. El ODS 6 busca garantizar agua limpia para todos, lo que mejora nuestra salud y calidad de vida. El ODS 7 tiene el objetivo de ofrecer energía de forma limpia y accesible, cuidando así nuestro entorno. Mientras que el ODS 16 aspira a crear un ambiente de paz y justicia, donde las instituciones trabajen de manera confiable para todos. Estos tres objetivos en definitiva tienen un impacto muy positivo dado que buscan un futuro más prometedor y equitativo para Colombia.

Sección	Contenido
Oportunidad/problema	Clasificación automática de textos según los ODS relacionados para mejorar la eficiencia en el análisis y la toma de decisiones.
Negocio	UNFPA y sus colaboradores en la planeación participativa para el desarrollo territorial.
Enfoque analítico	Clasificación supervisada de textos usando técnicas de aprendizaje automático. *Técnicas propuestas:* Clasificador Naive Bayes, Regresión Logística, Random Forest y SVM
Organización y rol beneficiado	UNFPA como entidad principal se beneficiará al tener una herramienta que automatice el proceso de clasificación de textos, ahorrando tiempo y recursos.
Contacto con experto externo al proyecto	s.garciap2@uniandes.edu.co - 16 de Octubre

3. Entendimiento y preparación de los datos

3.1. Entendimiento de la calidad de los datos

Para comprender mejor nuestros datos, hemos recurrido a diversas visualizaciones gráficas. Estas representaciones nos permiten evaluar la calidad y características intrínsecas de la información con la que trabajamos.
La primer gráfica nos permite ver que los datos están balanceados debido a que hay la misma cantidad de textos por cada categoría por lo cual no es necesario realizar ninguna imputación de datos.

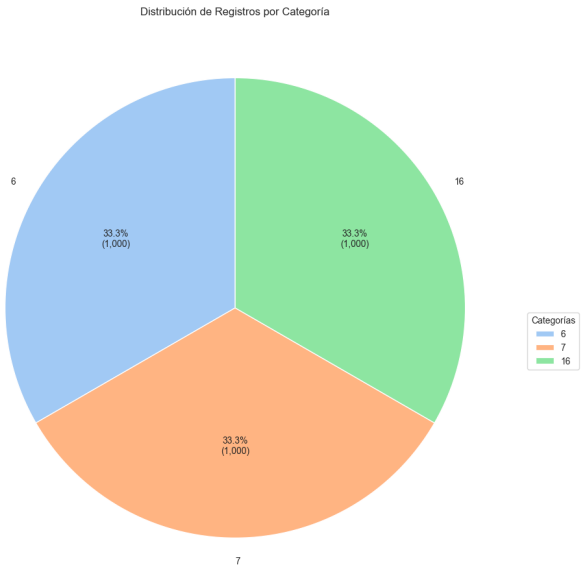

Figura 1: Distribución de textos por categoria

La siguiente gráfica que utilizamos fue con el fin de entender cuales eran las palabras más fuertes presentes en los textos, lo cual nos permitió observar que todas eran conectores que no aportaban y no decían nada del significado de los textos por lo que se decidieron eliminar estas palabras (stop words).

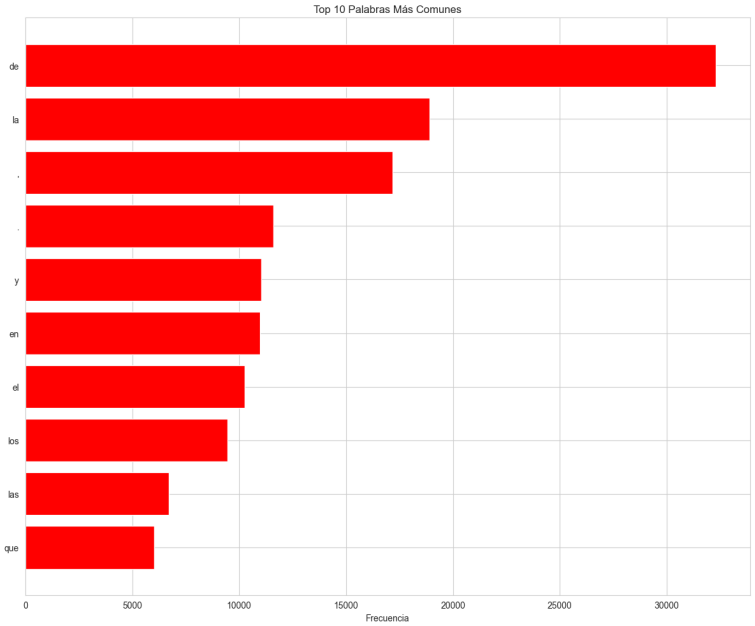

Figura 2: Palabras más fuertes

Después usamos una gráfica para ver cuales eran las palabras mas largas lo que nos permitió observar que eran links que no aportaban nada al momento de realizar el modelamiento por lo que decidimos eliminarlos.

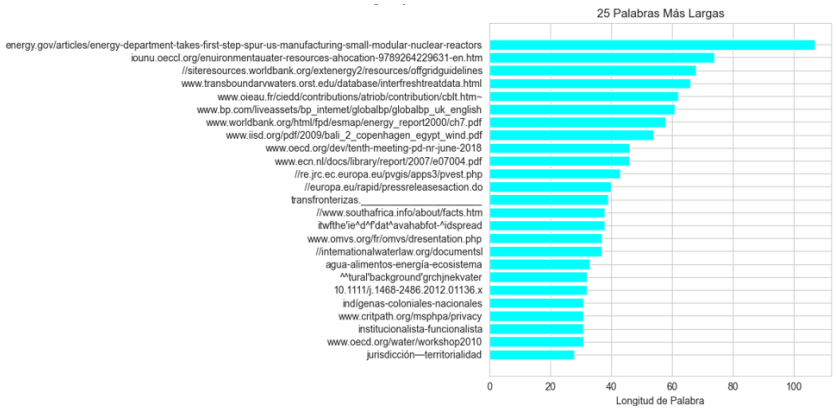

Figura 3: Palabras más largas

La ultima gráfica importante para el entendimiento de los datos fue la de identificación y conteo de caracteres extraños, la cual nos permitió observar que eran bastante frecuentes por lo que era necesario realizar imputación y normalización de los mismos.

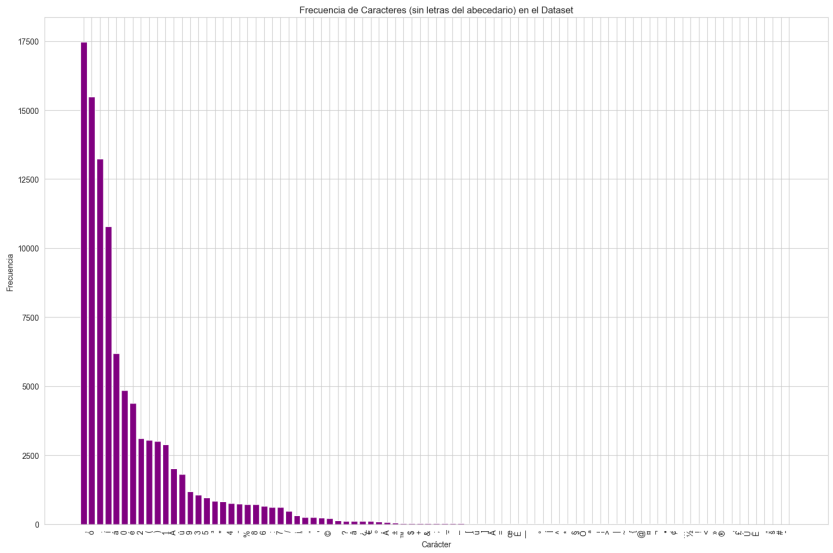

Figura 4: Caracteres extraños

Al culminar la preparación de los datos, abordamos distintas correcciones para optimizar la calidad de la información. En primer lugar, sustituimos aquellos caracteres que emergieron por errores de codificación. A continuación, depuramos el texto al eliminar todos los caracteres que no fueran letras, lo que incluye tanto números como símbolos. Acto seguido, homogeneizamos el conjunto de datos al transformar todo el texto a minúsculas. Luego, llevamos a cabo un proceso de lematización para reducir las palabras a su forma base. Finalmente, para facilitar el análisis y evitar inconsistencias, reemplazamos los caracteres con tildes por sus contrapartes sin ellas.

3.2. Entendimiento de los datos post-preprocesamiento

Para asegurarnos de que el preprocesamiento se llevó a cabo de manera adecuada, generamos diversas gráficas. Estas visualizaciones nos ofrecen una clara perspectiva sobre la estructura y calidad final de los datos por categoría.
En primer lugar, utilizamos una gráfica para saber como quedó la longitud de los textos por cada categoría y podemos observar que hay una diferencia notoria en la longitud los textos que pertenecen al ods 16.

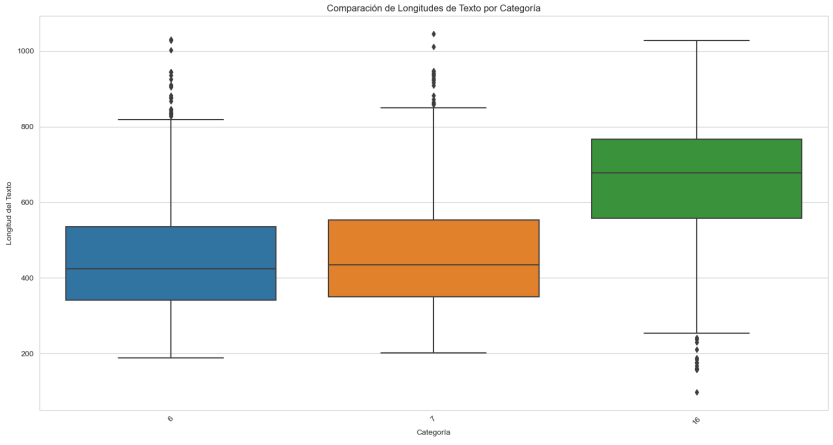

Figura 5: Longitud de textos por categoría

Después realizamos una gráfica para observar las palabras mas fuertes por cada categoría y vemos que las palabras de ahora si tienen relación con el tema abordado por cada ods, lo que indica que el preprocesamiento fue adecuado.

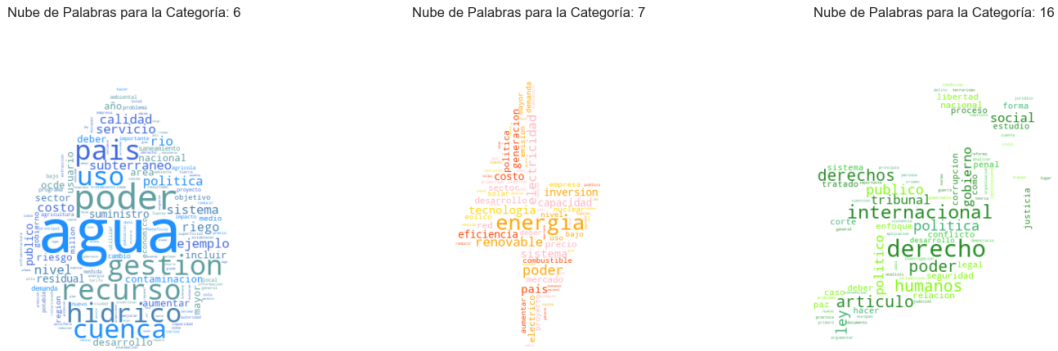

Figura 6: Palabras mas fuertes por categoría

Finalmente, realizamos una gráfica para observar los bigramas y trigramas mas frecuentes en los textos después del preprocesamiento de los datos y podemos observar que todos son conjuntos de palabras con sentido y relevantes para determinar el tema tratado en cada texto, lo que termina de confirmar que el preprocesamiento realizado fue correcto.

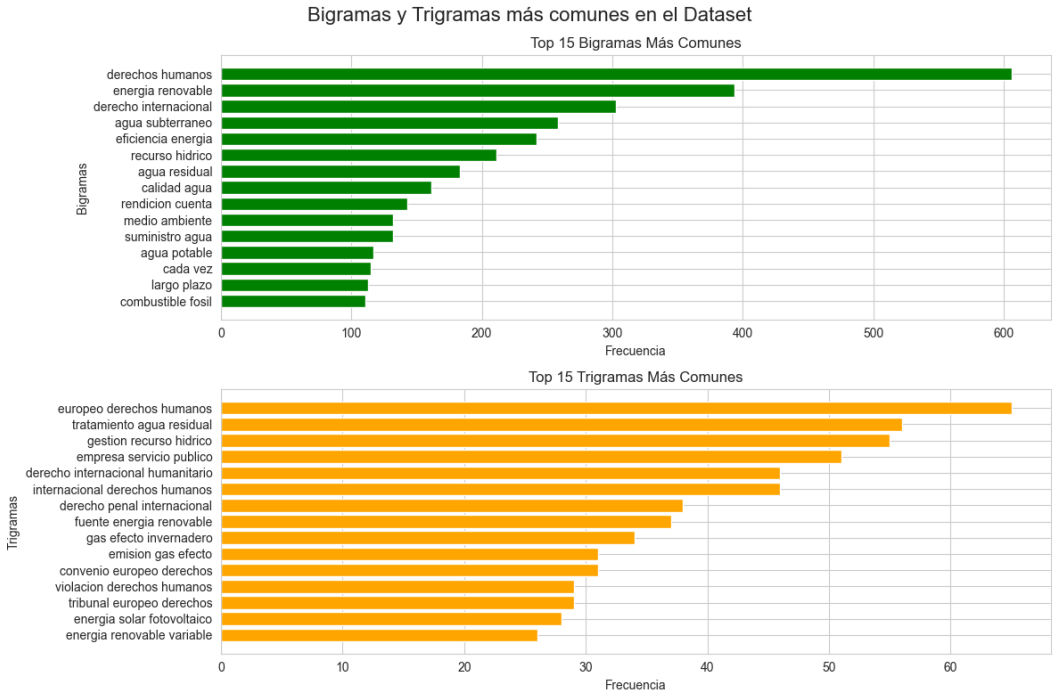

Figura 7: Bigramas y Trigramas más frecuentes

3.3. Procesamiento para el modelado

Representaciones Textuales:
Nos centramos en dos representaciones textuales primordiales:

- Bag of Words (BoW): Esta representación convierte el texto en una matriz donde cada fila representa un documento y cada columna indica la frecuencia de una palabra en dicho documento.

- TF-IDF (Frecuencia de Término - Frecuencia Inversa de Documento): Esta representación pondera los términos según su importancia en un documento en relación con una colección de documentos.

La representación de un texto es crucial para transformar información no estructurada en un formato comprensible para algoritmos de machine learning. Los modelos de clasificación, en su mayoría, operan sobre vectores numéricos, por lo que es esencial convertir el texto en vectores que representen adecuadamente su contenido y significado. Esta conversión permite que los modelos puedan discernir y aprender patrones dentro de estos vectores numéricos. Por tanto, una correcta representación textual no solo facilita la comprensión del contenido por parte de los algoritmos, sino que también es determinante para el desempeño y precisión de los modelos de clasificación.

4. Modelado y evaluación

En el proceso de modelado y evaluación, empleamos diversas técnicas y algoritmos de aprendizaje automático con el objetivo de obtener los mejores resultados posibles. A continuación, se presenta un desglose detallado de nuestro enfoque:

4.1. Algoritmos de aprendizaje

Para cada una las representaciones planteadas, para los siguientes modelos llevamos a cabo una búsqueda exhaustiva de hiperparámetros mediante la técnica de grid search y validación cruzada de 5 pliegues.

1. Clasificador Naive Bayes:
    - Es un clasificador basado en el teorema de Bayes con suposición de independencia entre características.
    - Elegimos este algoritmo por su rapidez y eficiencia en problemas de alta dimensión.
    - Utilizamos el modelo MultinomialNB de la biblioteca sklearn.
    - Se realizó una búsqueda en malla para optimizar hiperparámetros como alpha (parámetro de suavizado), fit_prior, class_prior, entre otros.
    - Se seleccionó el F1 score, con un promedio ponderado, como métrica de evaluación.
    - Tras ajustar el modelo con los datos, se presentaron los resultados de validación cruzada para cada combinación de hiperparámetros, destacando los hiperparámetros óptimos y su F1 score correspondiente.
    - Finalmente, después de elegir los hiperparametros se procedió a probar el mejor modelo encontrando en el conjunto test para evaluar el modelo con datos no etiquetados.
2. Regresión Logística:
    - Se utiliza para la clasificación binaria o multiclase, proporcionando probabilidades asociadas a cada clase.
    - Su elección es ideal para problemas que requieren interpretabilidad y una frontera de decisión lineal.
    - Empleamos el modelo LogisticRegression de la biblioteca sklearn.
    - Se llevó a cabo una búsqueda en malla para encontrar los hiperparámetros ideales como C, class_weight, multi_class, penalty, y solver.
    - Se realizó una validación cruzada de 5 pliegues para cada representación y conjunto de hiperparámetros, presentando el F1 score promedio y los resultados de cada pliegue.
    - Al final, tras seleccionar los hiperparámetros, se evaluó el modelo óptimo usando el conjunto de datos de prueba con información no etiquetada.
3. Random Forest:
    - Es un algoritmo de ensemble que combina múltiples árboles de decisión para obtener predicciones precisas.

- Es una excelente opción debido a su robustez, capacidad para manejar overfitting y para determinar la importancia de las características, y por eso fue escogido.
- Se utilizó el modelo RandomForestClassifier de sklearn.
- Se efectuó una búsqueda en malla para optimizar hiperparámetros tales como el número de árboles, profundidad máxima, mínimas muestras para dividir, entre otros.
- El rendimiento se evaluó utilizando el F1 score.
- Después de determinar los hiperparámetros adecuados, pusimos a prueba el modelo más efectivo con el conjunto de datos no etiquetados.

4. Máquina de Soporte Vectorial (SVM):

- SVM busca el hiperplano óptimo para separar las clases en un espacio de alta dimensión.
- Se escoge por su capacidad de manejar espacios complejos y por su alta precisión en escenarios donde las clases están bien diferenciadas.
- Empleamos el modelo SVC de sklearn.
- Se realizó una búsqueda en malla para determinar los mejores hiperparámetros como C, kernel, degree y gamma.
- La eficacia del modelo se evaluó mediante el F1 score.
- Una vez escogidos los hiperparámetros, se llevó a cabo una prueba con el modelo más prometedor en el conjunto de datos de prueba que contenía datos sin etiquetar.

A lo largo del proceso de modelado y evaluación, garantizamos que nuestras decisiones fueran impulsadas por los datos y orientadas a lograr el máximo rendimiento posible para nuestros modelos. Se eligió el F1 score, una media armónica de precisión y sensibilidad, como nuestra métrica de evaluación principal debido a su eficacia en la medición de la precisión de los modelos de clasificación, especialmente en situaciones donde pueden existir desequilibrios de clases.

A continuación, mostraremos los resultados de los cuatro modelos seleccionados y su desempeño utilizando las dos representaciones de texto (TF-IDF y BoW) que demostraron ser más efectivas, basándonos en los f1-scores registrados.

Algorithm	**Representation**	**F1 Score**		
		Validation Set (K-Fold C=5)	**Test Set**	**Difference Validation Test**
Logistic Regression	**BoW**	98,45 %	98,33 %	-0,12 %
	TF-IDF	98,74 %	97,99 %	-0,75 %
Random Forest Tree	**BoW**	98,70 %	97,84 %	-0,86 %
	TF-IDF	98,41 %	**99,17 %**	0,76 %
C-Support Vector Classification	**BoW**	98,41 %	98.50 %	0,09 %
	TF-IDF	98,74 %	98,33 %	-0,41 %
Multinomial Naive Bayes	**BoW**	98,54 %	98,00 %	-0,54 %
	TF-IDF	98,54 %	97,67 %	-0,87 %

En la tabla se presentan los F1-scores obtenidos para diversas combinaciones de algoritmos y representaciones vectoriales, evaluados en dos conjuntos distintos: entrenamiento y prueba. La finalidad de este proceso dual de evaluación fue determinar la capacidad de generalización de cada modelo frente a instancias no observadas previamente. A través de este análisis, se determinó que el algoritmo Random Forest Tree, utilizando la representación TF-IDF, logró el F1-score más elevado en el conjunto de prueba, evidenciando su superioridad en términos de adaptabilidad a datos desconocidos. En función de estos resultados, seleccionamos el Random Forest Tree con representación TF-IDF para la clasificación subsiguiente de datos no etiquetados.

4.2. Experimentaciones que no se incluyeron en el notebook final

Además de las técnicas y modelos previamente mencionados, se exploraron otras representaciones y modelos en las etapas iniciales del proyecto. Aunque estos enfoques no se incluyeron en el notebook final debido a diversas razones, como la eficiencia, la precisión o la complejidad, es esencial mencionarlos para proporcionar una visión completa de todo el proceso de experimentación:

Representaciones Textuales Adicionales:

- GPT-2 word embeddings: Se experimentó con la representación generada por el modelo GPT-2, conocido por su capacidad para generar texto coherente y contextual. Aunque es poderoso, su implementación puede ser más compleja y no siempre garantiza mejoras significativas para todas las tareas.
- Hashing: La técnica de hashing se probó como una alternativa para convertir el texto en una representación numérica. Aunque es eficiente en términos de memoria, puede no ser tan preciso como otras representaciones debido a posibles colisiones de hashing.
- PCA (Análisis de Componentes Principales): Se utilizó PCA para reducir la dimensionalidad de las representaciones textuales y conservar la mayor cantidad de información posible en un espacio de características reducido. Aunque puede ser útil para visualizar datos y reducir el tiempo de cálculo, en algunos casos, la reducción de dimensionalidad puede llevar a una pérdida de información crítica.

Modelos Adicionales:

- Red Neuronal preentrenada BERT: Se experimentó con una red neuronal preentrenada específicamente para la clasificación de textos, denominada BERT. Aunque estos modelos preentrenados pueden ser muy potentes, su implementación y ajuste requieren una gran cantidad de recursos y tiempo. Además, es crucial tener un conjunto de datos adecuado para afinarlos correctamente.

Es importante destacar que la elección de las representaciones y modelos finales se basó en una combinación de precisión, eficiencia y simplicidad. Aunque algunas de las técnicas y modelos mencionados anteriormente son avanzados y sofisticados, la decisión de no incluirlos en el notebook final se tomó después de una evaluación cuidadosa, priorizando aquellos que ofrecían el mejor equilibrio entre rendimiento y practicidad para el proyecto en cuestión.

5. Resultados

Los resultados de nuestro proyecto han evidenciado el potencial del modelo de C-Support Vector Classification con la representación TF-IDF para clasificar de manera eficiente textos según los Objetivos de Desarrollo Sostenible (ODS) relevantes. Esta eficiencia se refleja en el f1-score, que combina precisión y exhaustividad, y es esencial para garantizar una clasificación precisa de los textos.

Para UNFPA, esta herramienta es una oportunidad significativa para mejorar la eficiencia en la interpretación y análisis de la información textual recolectada a través de diferentes fuentes en procesos de planeación participativa. Recomendamos que UNFPA integre este modelo en sus flujos de trabajo y considere ampliar su uso en otras áreas que se beneficien de la clasificación automática de textos. Además, sería provechoso desarrollar una aplicación intuitiva que facilite aún más la interacción con los resultados del modelo y proporcione a los usuarios un acceso más sencillo a la información.

La verdadera ventaja para UNFPA radica en la capacidad del modelo para destilar grandes volúmenes de texto en categorías ODS claras, permitiendo a la organización concentrarse en estrategias y acciones específicas relacionadas con estos objetivos. Este enfoque basado en datos puede ser una piedra angular en los esfuerzos de UNFPA para promover un desarrollo sostenible más efectivo a nivel territorial.

6. Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido

Organización Beneficiada: Ministerio de Medio Ambiente y Desarrollo Sostenible

Descripción: Esta entidad gubernamental se encarga de la formulación, adopción, dirección y coordinación de las políticas, planes generales, programas y proyectos del sector administrativo del Medio Ambiente y Desarrollo Sostenible. La clasificación automática de textos relacionados con los ODS 6 (Agua limpia y saneamiento), 7 (Energía asequible y no contaminante) y 16 (Paz, justicia e instituciones sólidas) puede ayudarles a monitorear y evaluar las políticas públicas y su impacto social en relación con estos objetivos.

Mapa de Actores:

1. **Rol dentro de la organización:** Directores de Política Ambiental
    - Tipo de actor: Decisor
    - Beneficio: Toma de decisiones informadas basadas en la clasificación y análisis de textos relacionados con los ODS.
    - Riesgo: Tomar decisiones basadas en clasificaciones incorrectas o sesgadas.
2. **Rol dentro de la organización:** Equipo de Investigación y Análisis
    - Tipo de actor: Usuario-cliente
    - Beneficio: Acceso a herramientas automatizadas que facilitan el análisis y clasificación de grandes volúmenes de texto.
    - Riesgo: Dependencia excesiva en la herramienta sin validación manual.
3. **Rol dentro de la organización:** Organizaciones No Gubernamentales (ONGs) asociadas
    - Tipo de actor: Proveedores/Beneficiados
    - Beneficio: Colaboración en proyectos basados en datos y acceso a insights derivados del modelo.
    - Riesgo: Interpretación errónea de los datos clasificados que podría afectar sus iniciativas.
4. **Rol dentro de la organización:** Ciudadanía en general
    - Tipo de actor: Beneficiarios finales
    - Beneficio: Políticas y proyectos mejor informados que aborden directamente las preocupaciones y necesidades relacionadas con los ODS 6, 7 y 16.
    - Riesgo: Desinformación o malentendidos si el modelo clasifica incorrectamente los textos.
5. **Rol dentro de la organización:** Entidades Financieras y Donantes
    - Tipo de actor: Financiadores
    - Beneficio: Inversión en proyectos basados en datos con potencial de alto impacto en los ODS.
    - Riesgo: Inversión en proyectos que no aborden adecuadamente los ODS debido a clasificaciones erróneas.

7. Roles y Tareas

1. **David Santiago Ortiz Almanza - Líder de Proyecto:**
    - *Tareas:* Coordinación general del proyecto, establecimiento de cronogramas, organización de reuniones y supervisión de las entregas. Además, fue el encargado de mediar en las decisiones y resolver conflictos.
    - *Tiempo dedicado:* 30 horas.

- *Algoritmo trabajado:* Regresión logística y redes neuronales para análisis en textos.
- *Retos:* Mantener al equipo alineado y motivado, así como garantizar que los plazos se cumplieran.
- *Soluciones:* Implementación de herramientas de gestión de proyectos como Trello y sesiones de brainstorming para mantener la creatividad y el compromiso del equipo.

2. **David Santiago Vargas Prada - Líder de Negocio:**
   - *Tareas:* Definición de objetivos estratégicos, identificación de stakeholders clave y asegurar que el proyecto cumpla con las expectativas del negocio. También fue el enlace con expertos externos (s.garciap2).
   - *Tiempo dedicado:* 30 horas.
   - *Algoritmo trabajado:* Naive Bayes Model y SVM para clasificación de textos según su relevancia para los ODS.
   - *Retos:* Asegurar que el modelo se alinee con las necesidades reales del negocio y que los resultados sean comunicables y comprensibles para los stakeholders.
   - *Soluciones:* Organización de talleres con stakeholders y sesiones de feedback con el equipo técnico.
3. **Juan Esteban Cuellar Argotty - Líder de Datos y Analítica:**
   - *Tareas:* Curación y preparación de los datos, implementación de algoritmos y validación de los modelos. Supervisó la calidad del análisis y la interpretación de los resultados.
   - *Tiempo dedicado:* 30 horas.
   - *Algoritmo trabajado:* Random Forest para clasificación de textos e identificación de características clave en los textos relacionados con los ODS.
   - *Retos:* Lidiar con datos no estructurados y garantizar la integridad y calidad de los mismos.
   - *Soluciones:* Uso de herramientas avanzadas de procesamiento de lenguaje natural y colaboración estrecha con el líder de negocio para entender las necesidades de datos.

7.1. Reuniones Realizadas

- **Reunión de lanzamiento:** Establecimos la visión del proyecto, definimos roles y trazamos un plan inicial.
- **Reunión de ideación:** Discutimos las primeras exploraciones de los datos y definimos la dirección estratégica del proyecto.
- **Reuniones de seguimiento:** Monitoreamos el progreso, discutimos desafíos y ajustamos el enfoque según sea necesario.
- **Reunión de finalización:** Evaluamos el trabajo realizado, consolidamos la documentación y definimos los pasos a seguir.

7.2. Distribución de Puntos

Dada la contribución y el compromiso de cada integrante, la distribución de los 100 puntos es:

- David Santiago Ortiz Almanza: 33.3 puntos.
- David Santiago Vargas Prada: 33.3 puntos.
- Juan Esteban Cuellar Argotty: 33.3 puntos.

7.3. Puntos a Mejorar

Para futuras entregas, consideramos esencial fortalecer la comunicación interdisciplinaria, establecer revisiones técnicas más frecuentes y explorar nuevas fuentes de datos para enriquecer el modelo.

Referencias

[1] ¿Qué son los objetivos de desarrollo sostenible -ODS? - rendición de cuentas - función pública. (s.f.).https://www.funcionpublica.gov.co/web/murc/que-son-los-objetivos-de-desarrollo-sostenible-ods-1

[2] Sklearn.decomposition.PCA. (s. f.). scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[3] Nkitgupta. (2021). Text-Representations. Kaggle. https://www.kaggle.com/code/nkitgupta/text-representations