Proyecto 1

Inteligencia de negocios

Santiago Latorre 202111851 and Andrés Villota 201914885

Departamento de Ingeniería de Sistemas y Computación, Universidad de los Andes

15 de octubre de 2023

Contenido

Entendimiento del Negocio y Enfoque Analítico

OBJETIVO

El propósito del proyecto es proporcionar las herramientas necesarias para llevar a cabo una clasificación **automatizada** de los textos recopilados por el UNFPA durante su labor de seguimiento y evaluación de las políticas implementadas en Colombia. El objetivo principal es desarrollar modelos de aprendizaje automático capaces de categorizar estos textos en función de los Objetivos de Desarrollo Sostenible (ODS) a los que se refieren. Este enfoque ofrecerá ventajas significativas para el UNFPA, ya que el proceso actual de clasificación se realiza manualmente, requiriendo la supervisión de un experto para identificar el ODS al que corresponde cada texto.

La automatización de este proceso permitirá al UNFPA agilizar y optimizar el seguimiento de las políticas implementadas en Colombia. Además, facilitará una comprensión más profunda del impacto de las iniciativas en las poblaciones, lo que posibilitará tomar decisiones informadas sobre la continuidad o la modificación de dichas políticas.

ODS

Para entender el impacto que puede tener este proyecto es necesario entender primero que son los Objetivos de Desarrollo Sostenible (ODS), en especial el 3, 4 y 5 que serán tratados en este proyecto, y el impacto que tendrían en la población colombiana.

Los Objetivos de Desarrollo Sostenible son en total 17 objetivos establecidos por las Naciones Unidas en 2015 como parte de la agenda 2030. Estos objetivos buscan mejorar la calidad de vida de las personas mediante acciones e iniciativas que reduzcan la pobreza, cuiden el planeta y la disminución de las desigualdades.

ODS 3: Salud y Bienestar

El ODS 3 busca promover una vida sana y promover el bienestar para todos en todas las edades. Este objetivo ha cobrado mayor relevancia en los últimos años debido a la pandemia de COVID-19, que afectó a la población mundial. Además, es de vital importancia en Colombia, donde el sistema de salud presenta deficiencias significativas que tienen un impacto negativo en la salud y el bienestar de las poblaciones más vulnerables. Por lo tanto, es crucial implementar iniciativas destinadas a cumplir con este objetivo en Colombia para mejorar la calidad de vida de los colombianos. Por lo tanto, es fundamental realizar un seguimiento continuo de estas iniciativas a través de testimonios y evaluar si están teniendo un impacto positivo en la población.

ODS 4: Educación de Calidad

El ODS 4 busca garantizar una educación inclusiva, equitativa y de calidad y promover oportunidades de aprendizaje durante toda la vida para todos. Este objetivo es bastante importante ya que la educación es un factor clave para acabar con la pobreza, pues es importante para que las personas progresen y tenga mayores y mejores oportunidades laborales y en consecuencia una mejor calidad de vida. Este objetivo es importante para Colombia que siempre ha tenido bajos resultados en los rankings internacionales que califican el nivel de educación de los países. A pesar de los esfuerzos del gobierno nacional por mejorar la situación aún falta mucho para lograr tener una educación de calidad para todos. Por lo tanto, es importante que la ONU por medio de la UNFPA ayude implementando programas de mejora en la educación del país y es importante tener un control y seguimiento de estos programas.

ODS 5: Igualdad de género

El ODS 5 busca lograr la igualdad entre los géneros y empoderar a todas las mujeres y las niñas. Este objetivo es importante ya que durante muchos años los hombres han gozados de beneficios y ventajas frente a las mujeres por el simple hecho de ser hombres. Es importante garantizar la igualdad de género ya que es uno de los pilares claves para garantizar el desarrollo de una población pacífica y próspera. A pesar de que Colombia ha tenido grandes avances respecto a este objetivo, aún falta cerrar muchas brechas para garantizar las mismas condiciones para hombres y mujeres. Es importante generar espacios e iniciativas de apoyo para las mujeres en Colombia y llevar un seguimiento de esta población y los avances con el objetivo de igualdad de género.

Enfoque analítico

Dadas las necesidades de UNFPA de saber en qué categoría están clasificados cada uno de los textos suministrados y teniendo en cuenta la base de datos que se nos da con las etiquetas del ODS al cual corresponde cada una se decidió darle el siguiente enfoque analítico para el problema actual:

Tipo de aprendizaje: Supervisado
Tarea de aprendizaje: Clasificación
Técnica de aprendizaje: Árbol de Decisión

Este enfoque va a permitir cumplir con la meta del negocio de determinar el ODS del texto relacionado, es decir con base en los testimonios de la población saber que ODS se está buscando resolver por medio de las políticas que se están implementando.

Preparación de los Datos

Para garantizar resultados precisos y consistentes, es necesario homogenizar los textos que se van a utilizar tanto para la preparación como para la aplicación del modelo de clasificación. Esto implica cambiar todo al mismo formato y mantener solo aquella información que pueda ser de utilidad para la construcción del modelo (por ejemplo, elementos como puntuación y palabras vacías no aportan nada para el proceso de aprendizaje, pero si pueden llegar a afectar negativamente la precisión)

Un paso adicional que se puede realizar para garantizar resultados óptimos es reducir la dimensionalidad de la información. Para esto se utiliza funciones de normalización de texto y procesamiento de lenguaje natural que reduzcan las palabras a su forma raíz y así variaciones de la misma estén representadas por el mismo vector. (Por ejemplo, que "consistente" sea representado de la misma forma que "consistentemente")

Habiendo realizado estos pasos, ya se cuenta con un conjunto de datos óptimo para construir modelos de clasificación a continuación se evidencia una pequeña visualización del cambio realizado al comparar la columna "Textos_espanol" y "palabras"

	Textos_espanol	sdg	palabras
0	Por ejemplo, el nÃºmero de consultas externas ...	3	ejemplo naomero consulta externa especialista ...
1	En 2007, el gobierno central financió directam...	3	dos mil siete gobierno cent financio directa s...
2	Claramente, hay muchos otros factores en juego...	3	clara fact juego particul variedad barrera det...
3	Por ejemplo, el estado australiano de Victoria...	3	ejemplo australiano victor creo agenc promoc s...
4	El consumo anual de alcohol se estima en 15,7 ...	3	consumo an alcohol estim 157 litro ano hombr m...
...	...	...	...
2995	Un caucus efectivo se basa en fuertes vínculos...	5	cauc efectivo bas fuert vinculo grupo nac muj ...
2996	Por el contrario, el porcentaje de hogares en ...	5	contrario porcentas hog edad trabaj hombr trab...
2997	El análisis utilizará una gama de medidas que ...	5	analis utilizar gam medida refles barrera basi...
2998	La capacitación económica, el apoyo y, a veces...	5	capacitac economic apoyo vec incentivo abord j...
2999	Esto ha sido reconocido por los comitÃ©s en la...	5	sido reconocido comita atencia3n prestad salud...

Figura 1: Preparación de los datos

Se procede a separar el conjunto de datos en train y test como es usual en tareas de aprendizaje supervisado, con el fin de poder realizar una evaluación optima del desempeño del modelo construido

Modelado y Evaluación

Vectorización

Se probarán 3 diferentes algoritmos para realizar la vectorización, esto con el fin de garantizar la construcción del modelo más preciso posible:

Bag of Words: Una técnica bastante simple que representa la información como una colección de palabras sin considerar el orden o la estructura de esas palabras dentro del documento. Sus principales ventajas son la simplicidad, la versatilidad y la eficiencia, mientras que su principal desventaja es la pérdida de significado que, dependiendo del contexto, puede ser importante.

TF-IDF: Una técnica que cuantifica la importancia de los términos (tanto palabras como frases) dentro de un documento, y la compara con otros documentos. Esto con el fin de determinar términos discriminativos que diferencien un texto de otro (por ejemplo, términos que se repitan en una gran cantidad de documentos proveen menos información con respecto a la categoría del documento que otros que solo se vean en un tipo de categoría). Esta técnica es particularmente útil cuando se manejan documentos donde el lenguaje utilizado difiera mucho de una categoría a otra, pues le resultará sencillo encontrar los términos discriminativos, y con base a ello realizar predicciones acertadas.

Doc2Vec: Una extensión del modelo Word2Vec utilizado para el procesamiento del lenguaje natural. Mientras que Word2Vec está diseñado para aprender representaciones vectoriales de palabras, Doc2Vec aprende representaciones vectoriales de documentos enteros o párrafos. Logrando capturar el significado semántico de los documentos en un espacio vectorial continuo. Esta herramienta es muy útil cuando se trabaja en aplicaciones donde el contenido y el contexto del documento es importante para realizar la tarea de clasificación adecuadamente.

Construcción de Modelos

Al igual que con los modelos de vectorización, se utilizarán 3 algoritmos de clasificación de textos diferentes, con el fin de ampliar el espacio de prueba y así poder determinar con mayor precisión el modelo óptimo para la tarea a realizar.

Random Forest Classification

Random Forest Classifier es un algoritmo que combina las predicciones de múltiples modelos base (en este caso árboles de decisión) para realizar predicciones más precisas y robustas. En tareas de clasificación, este algoritmo hace una votación por mayoría para determinar la etiqueta final. Es decir, cada árbol realiza independientemente el proceso de clasificación, y con los resultados de cada uno se hace una votación donde el ganador es la clase con más votos. Random Forest Classifier también puede medir la importancia de cada característica (en este caso palabras o frases) para realizar las predicciones, y así determinar cuáles son las que más influencian en el resultado. Finalmente, es un algoritmo capaz de manejar valores atípicos, datos ruidosos y campos vacíos gracias a que agrega predicciones de múltiples árboles.

Logistic Regression

La regresión logística es un algoritmo de clasificación que modela la relación entre una variable categórica dependiente y una o más variables independientes mediante la estimación de la probabilidad de un resultado dado. En la regresión logística multinomial se hace uso de la función Logit Multinomial, la cual permite calcular la probabilidad relativa de pertenecer a una categoría particular en comparación con una categoría de referencia, facilitando la comprensión y cuantificación de la relación entre las variables independientes y las categorías de la variable dependiente en el modelo. Se trata de un algoritmo con una alta interpretabilidad que permite encontrar las variables con mayor impacto con facilidad. Sin embargo, éste siempre asume una relación lineal entre las variables independientes y dependientes, de manera que es sensible a valores atípicos, y puede no funcionar óptimamente en relaciones más complejas.

Multinomial NB

Multinomial Naive Bayes es una variación del algoritmo Naive Bayes, el cual parte de la suposición que las variables (en este caso palabras o frases) son condicionalmente independientes. Multinomial hace referencia al modelo de distribución de probabilidad utilizado para representar la frecuencia de las características en los documentos.

El clasificador Multinomial Naive Bayes construye un modelo de frecuencia para cada clase o categoría. Esto implica contar cuántas veces aparece cada palabra o término clave en los documentos de cada categoría. Específicamente, se mide la frecuencia de términos en cada categoría. Una vez cuenta con esta información, se puede realizar predicciones mediante el cálculo de la probabilidad condicional de que un documento pertenezca a cada categoría dada su representación de características, donde elige aquella con la probabilidad más alta como la predicción. La principal desventaja está en que está fundamentada casi exclusivamente en la frecuencia de las palabras, de manera que elementos como el orden y el contexto no son considerados incluso cuando puedan ser importantes.

Resultados

A continuación, se determinará cual es el modelo construido con mejor precisión, Recall y F1, pues este será el que se utilizará para realizar las nuevas predicciones.

```
La precisión máxima es 0.9792, y corresponde a la variable(s): BoW_NB
El recall máximo es 0.9789, y corresponde a la variable(s): tfidf_RandomForest, BoW_NB
El F1-score máximo es 0.9789, y corresponde a la variable(s): BoW_NB
```

Figura 2: Mejor modelo

Como se puede evidenciar, la combinación que arroja las mejores métricas es vectorización por Bag of Words y un algoritmo de Multinomial Naive Bayes. No resulta impropio que esta haya sido la mejor combinación, pues estas herramientas trabajan bastante bien en conjunto.

Representación de características: Para comenzar, Bag of Words representa el texto de los documentos como vectores de características donde cada dimensión corresponde a un término del documento, y Multinominal NB está diseñado para manejar este tipo de datos, haciéndolo una elección usual.

Modelamiento de probabilidades condicionales: Por otro lado, Multinominal NB logra modelar las probabilidades condicionales mediante la probabilidad de encontrar frecuencias de términos específicas en cada clase, cosa que se alinea con la representación de BoW.

Manejo de información discreta: Los datos de texto son inherentemente discretos, y tanto BoW como Multinominal NB están diseñados para trabajar con estos datos de conteo discreto.

Otras conclusiones adicionales que se puede obtener de los resultados son:

Los términos discriminativos, que permiten diferenciar entre las categorías según su frecuencia en un texto con respecto a otros, no están jugando un papel significativo en este conjunto de datos, pues las herramientas anteriormente mencionadas que se enfocan en estos elementos presentaron métricas menos favorables que la mejor.

Para este conjunto de datos, el contexto de los documentos no es relevante para construir un modelo de clasificación altamente preciso, pues tanto BoW como Multinomial Naive Bayes dejan de lado esta información para enfocarse en la frecuencia de los términos, y aun así fueron el modelo ganador.

Trabajo en Equipo

Después de una reunión de trabajo inicial donde se asignaron los roles se decidió que Santiago Latorre sería el líder del proyecto y de analítica, mientras que Andrés Villota sería el líder del negocio y de los datos. Estos roles se asignaron teniendo en cuenta las fortalezas de cada uno de los integrantes y sus habilidades. A pesar de que cada uno de los miembros tenía unas funciones asignadas, el trabajo se fue realizando de manera conjunta por parte de los dos, donde se hacían revisiones periódicas del progreso de cada uno para buscar problemáticas en el desarrollo y encontrar soluciones de manera conjunta. Debido a la magnitud del proyecto se decidió empezarlo con tiempo y el tiempo que se le dedicó a este fue de alrededor de 10 horas de trabajo cada uno.
Debido a que se hicieron reuniones y revisiones conjuntas del avance del proyecto y que ambos miembros participaron activamente de estas, no solo para resolver sus problemas y asignaciones si no para buscar solución a todos los problemas que el proyecto presentó, se decidió en la reunión final que el trabajo se manejó de forma equitativa por lo que la repartición de los puntos de haría 50-50.

Mapa de Actores

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
UNFPA	Usuario	Apoya la labor de evaluar las políticas implementadas e identificar los impactos que tienen en las poblaciones	Si se tiene un modelo erróneo, puede identificarse un testimonio o declaración de un ODS diferente al que se trata en realidad y esto conducirá a conclusiones equivocada
Gobierno nacional	Proveedor	Revisa y garantiza que los programas implementados estén funcionando correctamente y generen beneficios en la población colombiana	Si se da un mal manejo puede traer consecuencias negativas para la población y generar problemas en el país
Población	Beneficiado	Recibe ayuda de los programas financiados por la ONU donde se busca que tengan mejores condiciones de salud, educación e igualdad	Si se implementan los programas de manera incorrecta esto puede traer consecuencias económicas y sociales graves en la población.
ONU	Financiador	Ayuda a identificar los proyectos que tienen mayor impacto positivo e invertir mas dinero en estos proyectos y similares	Si el modelo está equivocado puede invertir en proyectos que no están dando resultados positivos y estar gastando dinero en vano