Universidad de los Andes
Inteligencia de Negocios

Proyecto 1: Analítica de textos para ODS del UNFPA

Etapa 1 – Construcción de modelos de analítica de texto

Grupo 26
Juliana Sofía Ahumada – 201921471
Laura Daniela Arias – 202020621

7 de septiembre de 2024

TABLA DE CONTENIDO

CONTEXTO

En este proyecto se busca apoyar al Fondo de Poblaciones de las Naciones Unidas (UNFPA). El objetivo principal de la organización es relacionar de forma automática opiniones de los ciudadanos con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5. La tarea se realizará por medio del análisis de información textual proporcionada por este actor en la forma de opiniones de distintos ciudadanos sobre los ODS de interés. Esta primera etapa se centrará en la construcción de varios modelos de analítica de textos aplicados al caso.

ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO

Oportunidad/Problema Negocio	Dificultad para analizar automáticamente grandes cantidades de datos textuales provenientes de opiniones ciudadanas.
Objetivos y criterios de éxito desde el punto de vista del negocio.	Crear un sistema automatizado para relacionar las opiniones ciudadanas con los ODS 3 (salud y bienestar), 4 (educación de calidad) y 5 (igualdad de género). El éxito se mide por la precisión, eficiencia y escalabilidad del modelo.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Entidades públicas como el UNFPA, ministerios de educación, salud y género. Los ciudadanos también se benefician al ver sus opiniones reflejadas en políticas públicas más efectivas. Científicos de datos: Encargados de analizar los datos, construir los modelos predictivos y generar los insights.

	Ingenieros de datos: Responsables de procesar y preparar los datos, y de desplegar los modelos en un entorno de producción. Ingenieros de software: Encargados de desarrollar las aplicaciones web o móviles que permitan a los usuarios finales acceder a los resultados del análisis.
Impacto que puede tener en Colombia este proyecto.	Este proyecto puede mejorar la toma de decisiones en políticas públicas, al proporcionar una mejor interpretación de las opiniones ciudadanas en áreas críticas como la salud, educación y género. Facilitará el alineamiento de las políticas con los ODS, promoviendo una mejor calidad de vida, igualdad y bienestar. A largo plazo, ayudará a fomentar una mayor participación ciudadana y transparencia en la gestión pública.
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.	Dado que los datos proporcionados son, más relevantemente, de naturaleza textual y están acompañados de una categorización numérica, concluimos implementar modelos de clasificación supervisada. De esta manera, podemos buscar predecir la calificación que se le va a otorgar a un sitio a partir de los comentarios realizados. Es así como se decidió que cada estudiante implementara un modelo de los 3 siguientes: – Support Vector Machines o SVM – Naive Bayes – Random Forest Acompañados de un *pipeline* para el manejo de los datos.

ENTENDIMIENTO DE LOS DATOS

Estructura
El archivo .xlsx original cuenta con 4049 datos de opiniones de ciudadanos—es decir, 4049 filas—, donde cada entrada cuenta con su opinión textual y su clase de tipo numérico con valores que van del 3 al 5—2 columnas.

Completitud
No se cuenta con datos faltantes dentro de la muestra total. Ningún dato es nulo.

Unicidad
De acuerdo con la información proporcionada por el *Pandas Profiling Report* dentro de los datos no se encontraron datos duplicados.

Validez
Todos los datos proporcionados son válidos para el ejercicio que se ha de realizar.

PREPARACIÓN DE LOS DATOS

Dentro de la preparación de los datos se llevaron a cabo diversas acciones para obtener los textos más óptimos para el desarrollo del trabajo. Para un entendimiento más fácil definimos estas acciones en dos grandes categorías—eliminaciones y conversiones—presentadas a continuación:
Eliminaciones

- Caracteres no-ASCII para evitar problemas con elementos no pertenecientes al alfabeto implementado en el español.
- Puntuación que no aporta a la interpretación de la información durante el análisis de cada palabra.
- Palabras vacías o comunes que no aportan al significado o análisis de los textos (e.g. la, el).
- Prefijos y sufijos para así normalizar el texto.

Conversiones

- Conversión a minúsculas de todo el texto para evitar problemas de interpretación por parte de los modelos de palabras iguales con capitalización diferente.
- Conversión de números a su redacción textual para que se acomoden mejor al contexto que se está manejando.
- Lematización de los verbos para reducir la variación de palabra y facilitar el procesamiento e interpretación de los textos.
- Tokenización con el fin de diferenciar mejor las palabras y ofrecer un mejor análisis sobre estas.

MODELADO Y EVALUACIÓN

Como se mencionó anteriormente los 3 modelos que implementamos son SVM, Naive Bayes y Random Forest.

Modelo SVM
El objetivo de un modelo SVM es encontrar un hiperplano en un espacio N-dimensional (donde N es el número de características) que mejor separe las clases de datos. Este hiperplano se elige de manera que maximice la distancia entre los puntos más cercanos de las clases, que se llaman vectores de soporte. De ahí su nombre, *Support Vector Machine*.

Reporte de Entrenamiento para Support Vector Machines

	precision	recall	f1-score	support
3	1.00	1.00	1.00	997
4	1.00	1.00	1.00	1086
5	1.00	1.00	1.00	1156
accuracy			1.00	3239
macro avg	1.00	1.00	1.00	3239
weighted avg	1.00	1.00	1.00	3239

Ilustración 1. Reporte de Entrenamiento para SVM

La anterior imagen ilustra el reporte del entrenamiento del modelo, que presenta métricas asombrosas.

Modelo Naive Bayes

Un modelo Naive Bayes es un clasificador probabilístico basado en el teorema de Bayes y en la suposición "ingenua" de independencia condicional entre las características. En el aprendizaje, se calculan las probabilidades a priori de cada clase y las probabilidades condicionales de cada característica dada cada clase en el conjunto de entrenamiento. Para predecir la clase de un nuevo ejemplo, se calcula la probabilidad posterior de cada clase utilizando el teorema de Bayes y la suposición de independencia condicional, y se selecciona la clase con la mayor probabilidad posterior.

Reporte de Entrenamiento para Naive Bayes

	precision	recall	f1-score	support
3	0.99	0.98	0.99	997
4	0.98	0.99	0.99	1086
5	0.98	0.98	0.98	1156
accuracy			0.99	3239
macro avg	0.99	0.99	0.99	3239
weighted avg	0.99	0.99	0.99	3239

Ilustración 2. Reporte de Entrenamiento para NB

La anterior imagen ilustra el reporte del entrenamiento del modelo, el cual cuenta con valores menores a los del modelo anterior pero que no dejan de ser extremadamente positivos. La precisión para este modelo es de 0.9861068230935474.

Modelo Random Forest

Por último, tenemos el modelo Random Forest. Este modelo funciona combinando un conjunto de árboles de decisión para realizar una predicción. Cada árbol en el bosque se entrena de forma independiente con una muestra aleatoria del conjunto de entrenamiento y una selección aleatoria de características. Para realizar una predicción de regresión, cada árbol en el bosque emite una predicción individual, y la predicción final del bosque se obtiene promediando las predicciones de los árboles individuales.

```
Métricas de rendimiento en el conjunto de entrenamiento:
Error cuadrático medio - Mean Squared Error (MSE): 0.012210713183081198
Error absoluto medio - Mean Absolute Error (MAE): 0.040688484100030835
R cuadrado - R-squared (R^2): 0.9815632129276198
```

Ilustración 3. Métricas de Rendimiento para RF

La anterior imagen ilustra las métricas de rendimiento del modelo. Se observa como este cuenta con un buen ajuste (MSE y MAE bajos) y que puede explicar la variabilidad de los datos de manera efectiva (R^2 cercano a 1).

RESULTADOS

Luego de realizar todo el proceso anterior y construir los 3 modelos propuestos para atender la situación expuesta, llegamos a los siguientes resultados:

Contamos con 3 modelos de clasificación de texto bastante favorables. Todos estos, al ser construidos a partir del mismo conjunto de datos y el mismo procesamiento de los mismos, cuentan con la misma base. Ergo, podemos realizar una recomendación de modelo informada a las entidades relacionadas. Contando con una precisión del 100%, sería instintivo seleccionar el modelo SVM como el indicado. Sin embargo, señalamos que esta "perfectitud" en la predicción de los datos podría indicar un posible sobre ajuste en el modelo, por lo que recomendamos probar el modelo con un conjunto de datos distintos y, por el momento, hacer uso del modelo Naive Bayes construido para alcanzar los objetivos propuestos.

MAPA DE ACTORES RELACIONADO CON EL PRODUCTO DE DATOS CREADO

Rol dentro de la empresa	**Tipo de actor**	**Beneficio**	**Riesgo**
Fondo de Poblaciones de las Naciones Unidas (UNFPA)	Cliente/Financiador	Uso más eficiente de los recursos asignados a proyectos de salud, educación y género, al basar las decisiones en datos analíticos precisos y automatizados.	Si el modelo no entrega resultados de calidad o no es rentable, los financiadores pueden perder confianza en la inversión realizada y en el retorno esperado.
Ciudadanos	Cliente/Usuario	Mejor atención y políticas de salud más efectivas y adaptadas a las necesidades reales de la población, con	Si el modelo analítico tiene un sesgo o mal funcionamiento, las políticas podrían no reflejar las

		base en la interpretación automática de sus opiniones.	necesidades reales de la población, causando una respuesta inadecuada a las problemáticas de salud.
Entidades Públicas como el Ministerio de Educación y Ministerios de Igualdad	Beneficiado	Mejor comprensión y alineación de las políticas públicas con los ODS 4 (Educación de calidad) y 5 (Igualdad de género), ayudando a focalizar recursos donde más se necesitan.	El riesgo es que el análisis automatizado pueda malinterpretar ciertas opiniones y priorizar mal las necesidades, lo que podría derivar en políticas públicas ineficaces o mal enfocadas.
Empresas de Tecnología	Proveedor	Participación en la creación e implementación del modelo, lo que les permite generar ingresos y obtener reputación en el desarrollo de soluciones analíticas para el sector público.	Si el proyecto fracasa o los resultados no son los esperados, puede dañar la reputación de los proveedores o impactar sus contratos futuros con otras entidades gubernamentales.

TRABAJO EN EQUIPO

Dentro de lo que fue la elaboración del proyecto, nuestro compañero Carlos Andrés Medina Cardozo decidió abandonar el grupo sobre el medio día del 7 de septiembre de 2024 sin aviso o explicación. Debido a esto, para la elaboración de los 3 modelos entonces la carga terminó repartida de la siguiente manera:

- Laura Arias: SVM y Random Forest
- Juliana Ahumada: Naive Bayes

Líder de proyecto: Laura Arias
Líder de negocio: Juliana Ahumada
Líder de datos: Laura Arias
Líder de analítica: Juliana Ahumada

Reuniones

- Reunión de lanzamiento y planeación
    - Fecha: Agosto 27.

    - Se revisó el enunciado y se definió la dinámica de trabajo.
- Reunión de ideación
    - Fecha: Septiembre 1.
    - Definición de los modelos a realizar y quién realiza qué tarea.
- Reuniones de seguimiento
    - De por si no se realizaron reuniones de seguimiento formales, sino que se mantuvo una comunicación constante por medio de un grupo de whatsapp y el compartir de los archivos e información requerida.
- Reunión de finalización:
    - Fecha: Septiembre 7.
    - Reunión generada a raíz de buscar terminar la entrega en conjunto de la mejor manera posible.