Universidad de los Andes
Ingeniería de Sistemas y Computación

Proyecto Etapa 1 - Construcción de modelos de analítica de textos

Caso Fondo de Poblaciones de las Naciones Unidas (UNFPA)

Integrantes

- Mateo Calderon
- Juan Ramirez
- Daniela Castrillón

Sección 1 - Entendimiento del negocio y enfoque analítico

Elemento	Descripción
Oportunidad/problema negocio	Vincular automáticamente las opiniones ciudadanas con los Objetivos de Desarrollo Sostenible (ODS) para mejorar la eficiencia en el tratamiento de las mismas y la precisión en la identificación de problemáticas a nivel local.
Objetivos y criterios de éxito desde el punto de vista del negocio	Crear un modelo analítico eficiente que clasifique las opiniones de los ciudadanos según los ODS relevantes. La evaluación o criterios de éxito se basan en las métricas de rendimiento de precisión, recall y F1-score junto con una matriz de confusión. Se espera contar con un modelo con precisión mínima del 85%.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	La UNFPA es el principal beneficiario. Dentro de la organización, los analistas de datos y los responsables de política que utilizarán el sistema para tomar decisiones informadas basadas en datos.
Impacto que puede tener en Colombia este proyecto	El proyecto podría aumentar la transparencia y la eficacia de las políticas públicas en Colombia, asegurando que las decisiones se basen en una comprensión precisa de las opiniones y necesidades ciudadanas. Esto puede llevar a una mayor satisfacción ciudadana y mejores resultados en términos de desarrollo sostenible.
Enfoque analítico Descripción de la categoría de análisis, tipo y tarea de aprendizaje e incluya las técnicas y algoritmos que propone utilizar	La categoría de análisis a utilizar es predictiva para anticipar cómo se pueden clasificar automáticamente las opiniones de los ciudadanos en relación con los ODS específicos. El tipo de aprendizaje es supervisado ya que el modelo se entrena utilizando un conjunto de datos etiquetados. La tarea de aprendizaje es de clasificación,

	donde cada opinión será clasificada en una de las categorías predefinidas que corresponden a los ODS específicos. Las técnicas y algoritmos propuestos se dividen en dos partes: Procesamiento de texto y Modelado. Para el caso del procesamiento se utilizará TF-IDF para transformar el texto en una matriz de características numéricas, destacando la importancia de términos específicos. Para el modelado se hará uso de los árboles de decisión, Support Vector Machines (SVM) y Naive Bayes.

Sección 2 – Entendimiento y preparación de los datos

La solución de esta sección se encuentra en el notebook.

Sección 3 – Modelado y evaluación

La solución de esta sección se encuentra en el notebook.

Sección 4 – Resultados

Durante el desarrollo de la solución se implementaron un total de 3 algoritmos de aprendizaje automáticos con el fin de identificar el mejor modelo que permite relacionar de forma automática opiniones de los ciudadanos con los ODS 3, 4 y 5 del Fondo de Poblaciones de las Naciones Unidas (UNFPA).

Los algoritmos usados para la solución son:

- Decision Three
- Support Vector Machines (SVM)
- Naive Bayes

A continuación, se presentan las métricas de desempeño obtenidas para cada modelo y su respectivo análisis.

Decision Three

- Accuracy

  Esta métrica nos indica que el 91.36% de las veces, el modelo hizo la predicción correctamente, siendo una buena tasa de acierto.

- Matriz de Confusión

    - ODS 3: 228 correctamente clasificadas; 22 incorrectamente clasificadas en otras clases.
    - ODS 4: 246 correctamente clasificadas; 22 incorrectamente clasificadas en otras clases.
    - ODS 5: 266 correctamente clasificadas; 26 incorrectamente clasificadas en otras clases.

- Reporte de Clasificación

ODS	Precisión	recall	f1-score
3	0.90	0.91	0.91
4	0.91	0.92	0.91
5	0.92	0.91	0.92

El modelo ofrece interpretabilidad y es capaz de identificar patrones significativos en los datos, aunque su precisión es inferior en comparación con otros modelos, lo que puede limitar su aplicabilidad en escenarios donde la precisión es crítica.

Naive Bayes

- Accuracy

    El resultado que nos da el modelo es de un accuracy del 96,79%. Lo cual nos muestra que el algoritmo hizo un muy buen trabajo y fue bastante efectivo para poder clasificar los textos.

- Matriz de Confusión

    - ODS 3: 241 correctamente clasificadas; 9 incorrectamente clasificadas en otras clases.
    - ODS 4: 261 correctamente clasificadas; 7 incorrectamente clasificadas en otras clases.
    - ODS 5: 282 correctamente clasificadas; 10 incorrectamente clasificadas en otras clases.

- Reporte de Clasificación

ODS	Precisión	recall	f1-score
3	0.99	0.96	0.98
4	0.96	0.97	0.96

5	0.96	0.97	0.96

Aunque ligeramente menos preciso que el SVM, Naive Bayes es efectivo para clasificar los comentarios y ofrece un buen balance entre precisión y recall, lo que lo convierte en una opción viable para tareas donde se valora la rapidez y simplicidad del modelo.

Support Vector Machines (SVM)

- Accuracy

  Se evidencia un valor de 97,65% de accuracy, este alto valor de precisión sugiere que el modelo tiene un alto rendimiento general y es muy efectivo para esta tarea de clasificación multiclase.

- Matriz de Confusión

  - ODS 3: 247 correctamente clasificadas; 3 incorrectamente clasificadas en otras clases.
  - ODS 4: 259 correctamente clasificadas; 9 incorrectamente clasificadas en otras clases.
  - ODS 5: 285 correctamente clasificadas; 7 incorrectamente clasificadas en otras clases.

- Reporte de Clasificación

ODS	Precisión	recall	f1-score
3	0.99	0.99	0.99
4	0.97	0.97	0.97
5	0.98	0.98	0.98

La alta precisión y capacidad de discriminación del SVM lo convierte en el modelo más adecuado para esta tarea, ofreciendo predicciones confiables y consistentes que pueden guiar las decisiones estratégicas de la organización.

Conclusiones

Opiniones sobre Salud y Bienestar (ODS 3)

Las palabras clave más importantes identificadas para este ODS incluyen términos como salud, sanitario, atención, mental, paciente, medico, y hospital. Estas palabras reflejan temas centrales de salud y bienestar, incluyendo la atención médica, enfermedades, y tratamiento.

Las opiniones que mencionan palabras relacionadas con la salud, atención médica, y tratamiento indican áreas de preocupación o satisfacción en la prestación de servicios de salud.

Basado en esta información la organización puede usar esta información para identificar áreas que requieren mejora, como el acceso a servicios médicos, la calidad del tratamiento, y el apoyo a la salud mental. Invertir en programas de educación sanitaria y mejorar la cobertura médica podría ser beneficioso.

Opiniones sobre Educación de Calidad (ODS 4):

Para el ODS 4, las palabras clave como educación, estudio, escuela, aprendizaje y docente destacan temas relacionados con la educación y el aprendizaje. Estas palabras sugieren un enfoque en la calidad educativa, formación de docentes, y el acceso a la educación.

Las palabras clave relacionadas con educación y aprendizaje en los comentarios suelen indicar el nivel de satisfacción o insatisfacción con los sistemas educativos.

La organización podría enfocarse en mejorar la calidad de la educación y la formación de docentes. También puede ser útil identificar brechas en el acceso a la educación y desarrollar programas para abordar esas brechas.

Opiniones sobre Igualdad de Género (ODS 5):

En el caso del ODS 5, las palabras clave relevantes como educación, estudio, escuela, aprendizaje, y docente también aparecen, indicando que la educación y el acceso a oportunidades educativas son cruciales para la igualdad de género.

Aunque la información sobre igualdad de género no aparece tan prominente en las palabras clave, la presencia de términos relacionados con la educación sugiere que la igualdad de oportunidades educativas es importante.

Basado en este análisis se podrían implementar políticas y programas que promuevan la igualdad de género en el acceso a la educación y oportunidades de formación puede ser una manera efectiva de abordar este ODS.

Esta información es útil para Fondo de Poblaciones de las Naciones Unidas (UNFPA) ya que les permite:

Optimización de Recursos: Entender qué temas son más recurrentes en los comentarios permite a la organización asignar recursos de manera más efectiva, centrándose en las áreas con mayor impacto.

Mejora de Servicios: Las áreas identificadas como problemáticas pueden ser mejoradas mediante la implementación de estrategias basadas en los comentarios de los usuarios.

Toma de Decisiones Informadas: Analizar las palabras clave y sus frecuencias ayuda a tomar decisiones basadas en datos reales y opiniones de los usuarios, lo cual es fundamental para el desarrollo de estrategias efectivas.

Sección 5 – Mapa de actores relacionado con el producto de datos creado

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
UNFPA	Cliente/Financiador	Identificar las principales problemáticas y evaluar posibles soluciones	Si el modelo no funciona, identificará erróneamente los ODS, dificultando las problemáticas y soluciones que se necesitan
Entidades públicas	Beneficiario/ Cliente	Evaluar los ODS relacionados con la información de los ciudadanos para poder hacer acciones que contribuyan a solucionar las inquietudes de la ciudadanía con diversos temas	Problemas con la identificación de los ODS que lleve a que las entidades gubernamentales centren sus esfuerzos en temas que para la ciudadanía no son tan relevantes como otros, perdiendo tiempo y recursos
Expertos	Proveedores	Aportan el conocimiento para basado en la información y aportan en la interpretación de los resultados	Si el modelo no es confiable, el experto deberá entrar a validar que todos los resultados que se tengan sean los correctos, dificultando su trabajo
Científicos de datos	Proveedores	Garantiza las métricas de calidad del modelo, lo que asegura su correcto funcionamiento para	Si alguna etapa del proceso no está bien realizada se corre el riesgo de que el modelo

		la tarea. Además, de la correcta preparación y análisis de los datos	(entendiéndolo como producto final) no funcione de la mejor manera, lo que haga que la organización no lo pueda usar para su objetivo
Ciudadanos	Beneficiarios	Recibe ayudas y medidas tomadas por los gobiernos para lograr solucionar los problemas que la población identifico	No lograrán satisfacer sus necesidades haciendo que los recursos destinados en pro de su beneficio terminen en cosas que no son primordiales para la ciudadanía

Sección 6 – Trabajo en equipo

- **Líder del proyecto:** Daniela Castrillón es la estudiante encargada de la gestión del proyecto, define las fechas de las reuniones, pre-entregables del grupo y asigna las tareas para que la carga sea equitativa. Finalmente, subirá la entrega del grupo.
  Durante el desarrollo del proyecto de desarrollaron las siguientes reuniones:
  - Lanzamiento y planeación
  - Ideación
  - Seguimiento
  - Finalización
- **Líder de negocio:** Daniela Castrillón es la estudiante encargada de velar por la resolución del problema identificado y que esté alineado con la estrategia del negocio para el cual se plantea el proyecto. Se debe garantizar que le producto sea comunicado de forma correcta.
- **Líder de datos:** Mateo Calderon es el estudiante encargado de gestionar los datos que se van a usar en el proyecto y de asignar las tareas sobre los datos. Además, es el encargado de dejarlo disponible para el grupo y se debe garantizar la entrega en el repositorio de git.
- **Líder de analítica:** Juan Ramirez es el encargado de gestionar las tareas de analítica del grupo. El se encargará de verificar que los entregables cumplan con los estándares de análisis y que se obtenga el mejor modelo según las restricciones existentes.