

PREDICTION TASK



Type of task:
Es un problema de **clasificación supervisada de texto**. El objetivo es asignar automáticamente una etiqueta (categoría) a un documento textual en función de su contenido.

Entity on which predictions are made:
Cada **documento de texto preprocesado** (ejemplo: reseñas, comentarios, publicaciones, noticias) es la unidad de análisis sobre la cual el modelo hace predicciones.

Possible outcomes:
El modelo predice la **clase o categoría** a la que pertenece un documento. En este caso, se tienen **tres categorías** (etiquetas 1, 3 y 4, según tus datos).

When are outcomes observed?:
Los outcomes se observan **inmediatamente después de ejecutar el modelo** sobre un documento nuevo. Una vez procesado el texto (tokenización, lematización, vectorización TF-IDF), el modelo predice la clase en tiempo de inferencia.

DECISIONS



How are predictions turned into actionable recommendations?
Las predicciones del modelo permiten clasificar automáticamente los documentos de texto en sus respectivas categorías. Esto facilita que el usuario final no tenga que revisar manualmente cada documento, sino que pueda enfocarse directamente en el grupo de interés.

Parameters of the process / application:
Predicción automática de categoría: El modelo asigna la etiqueta más probable según el contenido del texto.

- 1. **Umbral de confianza:** Si la predicción tiene baja confianza, se puede enviar al usuario para **validación manual**.
- 2. **Interfaz de uso:** El sistema devuelve la etiqueta asignada y puede integrarse en un **dashboard** o sistema de gestión de documentos para la toma de decisiones.

- Decisions enabled for the end-user:**
- Clasificación rápida de grandes volúmenes de texto.
 - Priorización de documentos según su categoría.
 - Reducción del tiempo y costo de procesamiento manual.
 - Mejora de la precisión en análisis y reportes derivados de los datos.

VALUE PROPOSITION



End beneficiary:
El modelo está diseñado para **analistas, investigadores o empresas** que trabajan con grandes volúmenes de texto y necesitan **clasificación rápida y precisa** sin depender de procesos manuales.

- Pain points addressed:**
- 1. **Tiempo** → actualmente clasificar manualmente grandes cantidades de texto es lento y costoso.
 - 2. **Errores humanos** → el modelo reduce la subjetividad y mejora la consistencia en la clasificación.
 - 3. **Escalabilidad** → permite procesar miles de documentos de manera automática, algo inviable manualmente.

- Integration with workflow:**
- Se puede integrar en un **dashboard interactivo** donde el usuario carga los textos y recibe la categoría asignada.
 - Puede conectarse con un **sistema de gestión documental** para organizar automáticamente los archivos entrantes
 - Posibilidad de generar **reportes analíticos en tiempo real** para apoyar decisiones estratégicas.

- User interfaces:**
- Dashboard web con visualizaciones (ej: métricas de clasificación, matrices de confusión).
 - API que devuelve las predicciones para ser consumidas por otros sistemas.

DATA COLLECTION



NO APLICA

DATA SOURCES







- Internal sources:**
- **Corpus de documentos proporcionado en el proyecto:** textos ya etiquetados en distintas categorías (labels) que sirven como conjunto de entrenamiento y validación.
 - **Preprocesamiento interno:** limpieza de texto, tokenización, eliminación de stopwords, lematización.

- External sources (opcional/futuro):**
- **APIs públicas** (ejemplo: noticias, artículos, redes sociales) para incrementar el corpus y robustecer el modelo.
 - **Repositorios académicos** (papers, informes técnicos) que permitan ampliar las clases existentes.

- Observed outcomes:**
- La etiqueta de cada documento (label) ya incluida en el dataset, que representa la **variable objetivo** usada en el entrenamiento.

- Transformations:**
- Conversión del texto a representaciones numéricas mediante **TF-IDF**.
 - Limitación del vocabulario a un número máximo de features relevantes (ej: 5000 palabras).
 - Generación de matrices dispersas (sparse) para reducir consumo de memoria.

IMPACT SIMULATION 	MAKING PREDICTIONS 	BUILDING MODELS 	FEATURES 
<p>Costos de decisiones incorrectas:</p> <ul style="list-style-type: none">Una clasificación errónea puede llevar a malinterpretar la categoría del documento, lo cual afecta el análisis posterior (ej. reportes, decisiones basadas en datos).Esto puede generar pérdida de confianza en el sistema y tiempo adicional para corrección manual. <p>Ganancias de decisiones correctas:</p> <ul style="list-style-type: none">Clasificación automática rápida y precisa de documentos.Reducción del tiempo de análisis manual.Mayor consistencia y escalabilidad en la organización de textos. <p>Datos para simulación de impacto antes de despliegue:</p> <ul style="list-style-type: none">Se utilizan las matrices de confusión y las métricas (accuracy, precision, recall, F1) obtenidas en la fase de validación.Esto permite estimar cuántos documentos se clasificarán correctamente y en qué categorías tienden a producirse errores. <p>Criterios de despliegue:</p> <ul style="list-style-type: none">El modelo debe superar un umbral mínimo de desempeño (ej: F1 ≥ 0.95).Debe demostrar consistencia entre clases (no solo buen desempeño en la clase mayoritaria).Balance entre precisión y recall, dependiendo del caso de uso. <p>Fairness constraints:</p> <ul style="list-style-type: none">Se evalúa que el modelo no favorezca sistemáticamente una clase sobre otra.En caso de detectar sesgo, se podrían aplicar técnicas de balanceo de clases (ej. oversampling, undersampling).	<ul style="list-style-type: none">Modo de uso del modelo: El modelo entrenado (SVM en este caso) se despliega junto con el vectorizador TF-IDF. Cada nuevo documento pasa por el mismo pipeline de preprocesamiento y se transforma en un vector para que el modelo lo clasifique.Flujo de predicción:<ul style="list-style-type: none">Entrada del usuario: un documento de texto sin etiqueta.Preprocesamiento: limpieza, tokenización, eliminación de stopwords, lematización.Vectorización: transformación del texto en representación TF-IDF utilizando el vectorizer.pkl entrenado.Predicción: el modelo SVM asigna la categoría más probable (ODS correspondiente).Salida: etiqueta final junto con la probabilidad/confianza de la predicción.Interfaces posibles:<ul style="list-style-type: none">Batch prediction: se cargan múltiples documentos a la vez (ej. un archivo CSV) y el sistema devuelve las categorías asignadas a cada documento.Online prediction: el usuario ingresa un documento en tiempo real (ej. desde un formulario web o API) y recibe la predicción inmediatamente.Outputs esperados:<ul style="list-style-type: none">Categoría ODS asignada al documento (ej. clase 1, 3 o 4).Nivel de confianza (score de probabilidad).Reporte agregado cuando se procesan muchos documentos (ej. distribución de documentos por clase).	<p>How many models are needed in production? Se necesita un único modelo en producción (el SVM) junto con el vectorizador TF-IDF ya entrenado.</p> <p>When should they be updated? El modelo debería actualizarse cada cierto período (ej. cada semestre o cada año) o cuando se disponga de nuevos datos representativos que puedan cambiar el lenguaje o las tendencias de los textos.</p> <p>How much time is available for this (including featurization and analysis)?</p> <ul style="list-style-type: none">El proceso completo de reentrenamiento (limpieza → vectorización → ajuste del modelo → evaluación) se estima en minutos a pocas horas, dependiendo del volumen de datos.En producción, las predicciones son inmediatas (milisegundos por texto). <p>Which computation resources are used?</p> <ul style="list-style-type: none">Entrenamiento: puede ejecutarse en CPU estándar (no requiere GPU, ya que SVM y TF-IDF escalan bien para datasets medianos).Producción: suficiente con un servidor ligero o contenedor Docker para exponer un endpoint de clasificación.	<p>What representations are used for entities at prediction time? Los textos se representan como vectores numéricos usando la técnica TF-IDF (Term Frequency – Inverse Document Frequency). Cada palabra del vocabulario se convierte en una característica con un peso que refleja su importancia en el documento frente al corpus completo.</p> <p>What aggregations or transformations are applied to raw data sources?</p> <ol style="list-style-type: none">Limpieza y normalización: conversión a minúsculas, eliminación de signos de puntuación, números y caracteres especiales.Tokenización: separación del texto en palabras individuales.Eliminación de stopwords: palabras muy frecuentes sin valor semántico relevante.Lematización: reducción de cada palabra a su forma base o raíz.Vectorización TF-IDF: transformación del texto procesado en un vector de hasta 5000 características para capturar la importancia relativa de los términos.

MONITORING

Which metrics and KPIs are used to track the ML solution's impact once deployed, both for end-users and for the business?
How often should they be reviewed?



Version 1.2. Created by Louis Dorard, Ph.D. Licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).
Please keep this mention and the link to ownml.co when sharing.

[OWNML.CO](https://ownml.co)