Informe Proyecto 1 Etapa 2

Automatización y uso de modelos de analítica de textos:

Análisis de sentimientos de películas

Inteligencia de Negocios

2023-10

Integrantes:
Juan Felipe Patiño 201922857
Juan Ricardo Diaz 201922167

1. Proceso de automatización

1.1. Proceso de preparación de datos

Para el proceso de preparación de datos tuvimos en cuenta lo realizado durante la etapa 1. Además, con el objetivo de automatizar esto por medio de un Pipeline tuvimos en cuenta el tipo de preparación de datos necesarios. Esta etapa es de las más importantes y críticas para el proceso de modelos analíticos en textos, y es clave que el entrenamiento seleccionado como mejor opción en la etapa anterior sea el mismo usado en esta etapa. En esta etapa buscamos asegurar que los datos sean relevantes, consistentes y en el formato esperado por el proceso. El primer paso es la recopilación de datos, lo cual manejaremos por la API a desarrollar, pero para el entrenamiento como tal usamos los datos que se entregaron. Para la limpieza tomamos la decisión de eliminar las palabras de menos de 2 usos ya que estas pueden causar un overfitting del modelo. Fuera de esto tomamos la decisión de vectorizar con ayuda de la librería nktl para generar una vectorización sobre todo el abecedario español. Esta vectorización se realizó con el método Bag of Words, el cual fue seleccionado en la etapa anterior como el mejor.

1.2. Construcción del modelo

La construcción del modelo y las decisiones asociadas fueron los pasos realizados en la etapa anterior de la entrega. Es por esto que con los resultados y decisión del modelo anterior construimos el pipeline para el sector de modelos. Recordemos que seleccionamos el un modelo de Random Forest Classifier, el cual tuvo un mejor desempeño respecto al resto, además utilizamos un grid search para los parámetros óptimos. Estos parámetros son de 200 estimadores, con un mínimo de 5 para un split, una profundidad de árbol máxima de 5 y bajo el criterio de entropía. De esa manera el pipeline construido contiene el modelo de mejor desempeño de la etapa anterior. Usualmente se realizará pasos como selección de características donde se filtran ciertas columnas, pero en este caso la única característica es el texto en español, y está es la que es va a recibir en el modelo, por lo cual no es necesario este paso para nuestra pipeline, pues está únicamente recibe el texto. Otros pasos relevantes, como lo son la preparación de datos de entrenamiento y prueba, el entrenamiento del modelo, la evaluación y ajuste del mismo son los pasos que ya realizamos en la etapa anterior del proyecto, por lo cual ya contamos con estos pasos realizados y se pueden referenciar en la entrega de la etapa 1.

1.3. Persistencia del modelo

1.4. Acceso por medio de API

2. Desarrollo de la aplicación y justificación

(40%) Desarrollo de la aplicación y justificación. Descripción del usuario/rol de la organización que va a utilizar la aplicación, la conexión entre esa aplicación y el proceso de negocio que va a apoyar (si aplica), y la importancia que tiene para ese rol la existencia de esta aplicación. Adicionalmente, el desarrollo de una aplicación web o móvil para interactuar con el resultado del modelo (y su calidad) a partir de un texto o textos datos por el usuario. En este punto debe contar con el apoyo del grupo de estadística que le fue asignado y mostrar evidencia de las mejoras que logró gracias a esa interacción

3. Resultados

Video de máximo siete (7) minutos con la descripción y visualización en la aplicación de los resultados del modelo, que permita a un rol dentro de la organización comprenderlos y usarlos. El video publicado en el padlet respectivo (ver condiciones de entrega), debe simular la interacción del usuario final con la aplicación y describir dos acciones que puede realizar como resultado de dicha interacción, haciendo énfasis en la forma como el resultado del modelo aporta en esas acciones. En este punto debe contar con el apoyo del grupo de estadística que le fue asignado y mostrar evidencia de las mejoras que logró gracias a esa interacción.

4. Trabajo en Equipo

4.1. Roles

Integrante	**Rol**
Juan Felipe Patiño	Líder de Proyecto, Ingeniero de Datos
Juan Ricardo Diaz	Ingeniero de software responsable del

	diseño de la aplicación y resultados, Ingeniero de software responsable de desarrollar la aplicación final

Juan Felipe Patiño, como Líder de Proyecto, estuvo a cargo de coordinar las fechas de reuniones y de trabajo, llenar los formatos de pre-entrega y dividir el trabajo a realizarse de manera equitativa y justa. Además, estuvo pendiente de los avances constantes y notificó al grupo sobre los detalles de la entrega. Finalmente en este rol fue el que subió y entregó todos los entregables pertinentes. Por otro lado como Ingeniero de Datos, tuvo la labor de el proceso de automatización del modelo, en especial el proceso de construcción del modelo y preparación de datos, traduciendo la etapa anterior a un pipeline autocontenido y desplegado como un modelo.joblib con su modelo.py para acceder a este, la cual se interconecta con la API posteriormente.

Por otro lado Juan Ricardo, como Ingeniero de software tuvo la responsabilidad del diseño de aplicación y resultados. Para esto se enfocó en el desarrollo de un frontend funcional y desarrollar la integración de este con el backend desarrollado por el ingeniero de datos. Para esto se realizó un trabajo iterativo, donde se consultaba con terceros o con el compañero del equipo para revisar la accesibilidad y usabilidad del producto, y ajustarlo adecuadamente ante cualquier problema.

Con los roles de cada uno asignados y desarrollados, cada integrante se planteó utilizar una misma cantidad de tiempo individual y un tiempo grupal de reuniones y de apoyo mutuo. Este tiempo individual de desarrollo fue menor a 5 horas cada uno, dependiendo ya de la velocidad de desarrollo y condiciones del software y su entorno, mientras que de manera grupal es tuvieron múltiples reuniones, alrededor de 3 formales y múltiples informales, de cada una alrededor de 1 hora

Entre los retos principales estaba el tiempo disponible, ya que si bien se tuvo el enunciado disponible con antelación no fue hasta después de terminar la etapa anterior que se pudo iniciar a avanzar en esta. Además, para esta ocasión fue necesario un mayor grado de investigación sobre el desarrollo de front end local, para poder desplegar la aplicación. Entonces el reto de obtener el tiempo para desarrollar la actividad de manera completa y de buen nivel fue un gran reto. Durante el transcurso de la segunda etapa se llevaron a cabo múltiples reuniones, donde vale la pena recalcar una primera de decisión de tema, una segunda de división de roles, dos reuniones posteriores de avances y retroalimentación mutua y una reunión final para finalización de entregables y posterior entrega.

4.2. División de puntuación

La división de puntos, acorde a la planeación y realización del proyecto se realizará de manera equitativa. Si bien en algunos tramos un integrante había avanzado más que el otro, analizando la situación como una suma del total el trabajo fue equitativo entre ambos. Es por esto que dividiremos los puntos 50-50.