Trabajo Práctico 2 : Críticas cinematográficas

Introducción

En este trabajo práctico vamos a utilizar una colección de críticas cinematográficas en idioma español y vamos a tratar de identificarla como positiva o negativa.

Modalidad de entrega

Notebook

El trabajo debe ser realizado en una notebook *Jupyter* de Python, se espera que la misma contenga **todos** los resultados de la ejecución los cuales siempre deben ser **reproducibles**. La notebook debe respetar la siguiente nomenclatura :

7506R_TP2_GRUPOXX_CHPX_ENTREGA

En el caso que sea estrictamente necesario entregar más de una notebook las mismas deben contar con una numeración correlativa manteniendo un orden lógico entre ellas (7506R_TP2_GRUPOXX_CHPX_ENTREGA_**N1**, 7506R_TP2_GRUPOXX_CHPX_ENTREGA_**N2**, etc)
Las secciones del trabajo deben estar claramente diferenciadas en la notebook utilizando celdas de *markdown*. Se debe incluir una sección principal con el título del trabajo, el número de grupo y el nombre de todos los integrantes.
Todo análisis realizado debe estar acompañado de su respectiva explicación y toda decisión tomada debe estar debidamente justificada. Cualquier hipótesis que sea considerada en el desarrollo del trabajo práctico debe ser detallada y debe estar informada en la entrega. Cualquier criterio que se utilice basado en fuentes externas (papers, bibliografía, etc.) debe estar correctamente referenciado en el trabajo.

Modelos
Todos los modelos entrenados deben ser guardados en un archivo (joblib / pickle) y deben estar disponibles en la entrega para ser utilizado por el equipo docente.

Reportes
Todos los reportes/informes solicitados deben respetar la extensión máxima solicitada, deben estar en formato pdf y deben tener la siguiente nomenclatura

7506R_TP2_GRUPOXX_CHPX_REPORTE.pdf

Repositorio

Cada grupo deberá crear su propio repositorio en github con la siguiente nomenclatura:

7506R-1C2023-GRUPOXX (usen el mismo del TP01)

En dicho repositorio deberá estar disponible la notebook, los modelos entrenados, los conjuntos de datos utilizados para el entrenamiento y cualquier archivo que sea necesario para la correcta ejecución del trabajo.

Competencia Kaggle

El trabajo práctico estará enmarcado en una competencia de **Kaggle**, dónde todos los alumnos deberán participar. Para unirse a la misma deben acceder con el siguiente enlace y conformar los grupos correspondientes. Pueden elegir cualquier nombre que represente al equipo. Se recomienda que mantengan los mismos nombres que en el TP1.

El objetivo de la competencia es hacer la predicción más precisa posible acerca de si una crítica cinematográfica dada en español es positiva o negativa.
Para saber qué tan bien se desempeña un modelo, cada grupo hará su predicción sobre el conjunto de test y la subirá (submit) a **Kaggle**.

Kaggle verificará las predicciones contra el archivo de soluciones utilizando la **métrica F1** y mostrará la posición del equipo en la tabla de puntajes (leaderboard). Pero sólo usará para ello el 60% de sus respuestas. El 40% restante se usará también para calcular su puntaje pero en un tablero privado que sólo pueden ver los docentes y que se revelará al finalizar la competencia (07/12/2023).

Enunciado

Los conjuntos de datos a utilizar **train** y **test** se encuentran disponibles en la competencia de Kaggle y deberán descargarlos desde allí. Allí mismo encontrarán también un archivo de ejemplo de cómo se deben subir las soluciones.
El trabajo consiste en construir diferentes modelos de clasificación, capaces de analizar una porción de texto en lenguaje natural y detectar el sentimiento allí presente, al menos de forma binaria: positivo o negativo.
Para ello habrá que realizar un preprocesamiento del texto para que este pueda ser analizado por los distintos modelos. Se utilizará el modelo de ***bag of words***, o cualquier otro que permita convertir texto en vectores.

Los modelos que se deben construir son los siguientes:

- Bayes Naïve
- Random Forest
- XGBoost
- Un modelo de red neuronal aplicando Keras y Tensor Flow.
- Un ensamble de al menos 3 modelos elegidos por el grupo.

Para cada uno de estos modelos se debe realizar una búsqueda de hiperparametros que optimicen su desempeño en el conjunto de test local (porción del archivo training).
Una vez encontrados dichos hiperparametros, se procederá a hacer un submit a Kaggle. Es decir que habrá al menos 5 submits (uno por cada modelo).

Informe

En el informe se debe explicar de forma clara y sintética los hiperparametros escogidos para cada modelo. En el caso de la red neuronal, se debe indicar la arquitectura y explicar por qué se la eligió. Se deben indicar además: la precisión, el recall y la medida F1 en el conjunto de datos de prueba local (porción del archivo training) y el puntaje obtenido en el tablero público de Kaggle. Pueden encontrar un modelo de informe en el siguiente link.

Fechas de entrega

La fecha final de entrega final del trabajo práctico es el día 07 de diciembre de 2023.

Condiciones de Aprobación

- **Todos** los integrantes de los grupos deben participar de la competencia.
- Se debe tener **al menos 1 submit** por modelo/ensamble pedido.
- Todas las semanas deben participar en la competencia se considerarán las siguientes fechas semanales:

  1°sem: 05/11/2023 al 11/11/2023
  2°sem: 12/11/2023 al 18/11/2023
  3°sem: 19/11/2023 al 25/11/2023
  4°sem: 26/11/2023 al 02/12/2023

La 5° sem, comprendida entre el 03/12/2023 y el 07/12/2023, no será obligatorio subir predicciones a Kaggle ya que pueden utilizar ese tiempo para confeccionar el informe final.

No obstante pueden hacer submissions hasta el 07/12/2023 23:59 hs.