PRACTICA 2: Reto de Predicción Automática de Correctitud de Pasajes Textuales Respecto a Preguntas Médicas

La clasificación automática de textos es una de las tareas clásicas en el área del Procesamiento del Lenguaje Natural (PLN). Consiste en catalogar documentos o partes de textos de manera automática en función de una característica común. En el día a día, nos encontramos múltiples ejemplos de esta clasificación automática como, por ejemplo, la detección de spam que hacen nuestros clientes de correo.

En este reto se va a abordar la catalogación automática de pasajes textuales (de tamaño párrafo o similar) en función de su correctitud respecto a una pregunta médica (por ejemplo, "¿puede la vitamina C curar el COVID-19?"). Se trata de desarrollar tecnología predictiva que pueda automáticamente estimar si un párrafo está dando la respuesta correcta o no a la consulta médica. Para ello, se proporcionará un conjunto de entrenamiento, con preguntas médicas, pasajes relacionados con ellas y etiquetas de correctitud. El alumnado deberá diseñar soluciones creativas para, a partir del conjunto de entrenamiento, desarrollar tecnología predictiva que a partir de una consulta médica y un texto relacionado sea capaz de estimar si el texto está respondiendo de manera correcta a la pregunta.

La práctica se realizará en grupos de hasta 3 personas. Cada grupo se identificará por un nombre a vuestra elección.

Pasos a seguir:

1) En esta práctica, os proporcionamos un conjunto de pasajes procedentes de páginas web y relacionados con una serie de preguntas médicas (numeradas entre el 1 y el 50). Para cada pregunta médica un pasaje puede ser correcto o incorrecto. La colección se encuentra dividida en conjunto de entrenamiento y test. Para el conjunto de entrenamiento disponéis tanto del texto de la pregunta médica como el pasaje relacionado como la etiqueta de correctitud. Para el conjunto de test sólo disponéis del texto de la pregunta y el pasaje relacionado (y, por tanto, debéis estimar la correctitud con tecnología automática).

Sugerencia: podéis crear un split de validación a partir del conjunto de entrenamiento para comprobar cómo funcionan vuestras diferentes soluciones.

2) El principal objetivo de la práctica se resume en que propongáis diferentes soluciones al problema de estimación de correctitud. Esta parte es libre y cada grupo presentará una o varias soluciones diferentes al problema de clasificación.

A continuación, os presentamos una lista de posibles ideas, pero no son las únicas y se valorará la originalidad de cada grupo y la búsqueda de soluciones alternativas al problema:

- Utilización de algoritmos tradicionales de clasificación (p.e., una SVM, un Naïve Bayes, KNN, etc), de modelos de Deep NLP (e.g. como BERT, RoBERTa, XLNet, etc.) o utilización de estrategias de prompt engineering con Large Language Models como ChatGPT.
Sugerencia: Para los modelos de DeepNLP podeís usar la librería de HuggingFace (https://huggingface.co/docs/transformers/training) o Ernie (https://github.com/labteral/ernie). Tened en cuenta que estos modelos requieren de GPU para hacer inferencia, podéis usar herramientas como Google Colab, para ello.
- Utilización de diferentes estrategias de preprocesado (e.g. eliminar oraciones irrelevantes de los textos) y de extracción de características estilísticas de texto.
- Probar estrategias de "document expansión" para generar palabras clave a partir

de los pasajes y expandir los textos con esas descripciones expandidas (e.g. docTTTTTquery: https://github.com/castorini/docTTTTTquery).
- Ampliar el training data (p.e. crawleando sitios web de preguntas-respuestas médicas).
- Aplicar "transfer learning" usando clasificadores disponibles de alguna otra dimensión y que puedan generalizar bien a esta tarea (p.e. aplicar un clasificador de readability como este https://huggingface.co/valurank/en_readability con la hipótesis de que, por ejemplo, los pasajes correctos tienen un mejor nivel de legibilidad).
- Diseñar estrategias de "factual checking"

Cada equipo deberá especializarse en una determinada técnica o método y hacerse algún tipo de pregunta de investigación que quiera responder mediante sus experimentos y variantes (por ejemplo, para la expansión de textos las preguntas que podría realizarse el equipo podrían ser: ¿puede la expansión de los pasajes ayudar a realizar una mejor clasificación? ¿en qué medida debo expandir los pasajes para mejorar sus representaciones? ¿a qué modelos de clasificación ayuda la expansión?)

3) Cada equipo deberá mandar varias variantes o soluciones ("runs") en un csv con el formato csv que tienen los ficheros de ejemplo (por ejemplo, el fichero random_predictions_for_test_data.csv, que tiene predicciones aleatorias para el conjunto de test).

4) Estas runs serán evaluadas automáticamente por el profesor de la materia y se presentará un "leaderboard" con las soluciones ordenadas por la métrica de accuracy. El objetivo no es ganar la competición y no se penalizará un rendimiento predictivo bajo. El principal objetivo del leaderboard es que todos/as aprendamos y veamos qué estrategias de clasificación funcionan mejor para este problema.

Entregables:
(en un único archivo .ZIP)
1) Guión python (.py)
2) Python Notebook (.pynb) (sed particularmente cautos/as en detallar los resultados de los experimentos, etc) con las diferentes soluciones implementadas
3) Ficheros .csv con las variantes
Es fundamental que el Notebook sea autoexplicativo de todos los pasos (con celdas textuales acompañando a celdas con código y que contenga explícitamente los resultados -sin tener que ejecutar las celdas de nuevo-). Comprobad esto antes de enviar el Notebook. Cualquier proyecto de Analítica de Datos debe ser autodocumentado y sus experimentos fáciles de reproducir. Un aspecto clave en la evaluación de esta práctica reside en la calidad de las explicaciones y documentación que acompañéis al código dentro del Notebook.

Dado que se trata de un trabajo en grupo el notebook debe tener un apartado en el que se explique cómo se ha organizado el trabajo, cómo se han dividido las tareas o retos a afrontar entre las personas del grupo y cómo se ha coordinado la comunicación, escritura de resultados, análisis, programación, etc.

Valoración y Fecha de Entrega:
Esta práctica tiene una valoración de 4 puntos (sobre el total de 7 puntos de la parte práctica de la materia)
Fecha entrega: Habrá dos entregas. Una primera entrega de variantes el **31 de octubre**, a las 23:59h. Esta entrega servirá para generar el primer leaderboard y que podáis ver cómo

funcionan vuestras distintas soluciones y mejorarlas. La entrega definitiva será el **5 de diciembre**, a las 23.59h.

Cada grupo en cada ronda de entrega de variantes sólo podrá enviar 5 ficheros de estimaciones (por ejemplo, los 5 modelos o aproximaciones que entiende son más prometedoras). Cada fichero de variante debe tener un nombre de archivo que identifique claramente qué grupo lo envía y de qué variante se trata (por ejemplo, MOZOS_DE_AROUSA_variante_SVM_kernelRBF.csv).

El fichero ZIP con el código, notebook, etc. sólo se entregará en la ronda final.

Habrá una sesión final de defensa/presentación de resultados a realizar el **12 de diciembre a las 16:00h**. Cada grupo tendrá 10 min para exponer su solución.

Se permiten entregas retrasadas del ZIP final pero se reducirá la puntuación del siguiente modo:

Cada día tarde reduce en un 10% de la máxima nota alcanzable (es decir, cada día tarde resta un 0.4 puntos de la nota que se os asigne al valorar la práctica)