//introducción culia pajera y motivación

//??

Buenos días… (introducción, presentar a todos y weas)

La base de datos que estudiará nuestro grupo consiste en un archivo que contiene 1.3 millones de comentarios sarcásticos de Reddit. Para contextualizar, Reddit es una red social en la cual los usuarios pueden participar subiendo todo tipo de contenido, e interactúan a través de comentarios presentes en cada publicación. Por convención, si un usuario redacta un comentario que debe ser interpretado de forma sarcástica, se escribe el tag “/s” al final de este para evitar cualquier amiguedad. Nuestra motivación para estudiar esta base de datos es poder encontrar patrones que permitan caracterizar comentarios sarcásticos sin que sea necesaria la presencia de este tag, y de esta forma obtener información importante que nos permita deducir cómo en un futuro una máquina podría detectar el sarcasmo al procesar el lenguaje natural.

//Caracterización de la base de datos, ver en kaggle o en Excel, lo que sea más cómodo

//jose

Nuestro dataset consta de una tabla con 1.3 millones de filas y 10 columnas. Entre las más relevantes se encuentran el comentario en sí, junto con el nombre del usuario que redactó el comentario, y el subreddit en el que fue publicado. Un subreddit es uno de los tantos foros que existen en Reddit, y en el cual la discusión gira en torno a un tema en particular, ya sea política, noticias, series o videojuegos específicos, etc. Además, se tiene el puntaje del comentario, que es calculado a través de los likes y dislikes recibidos por este. También se tiene la fecha y la hora de la publicación, y finalmente se tiene el comentario padre, es decir, aquel al que el comentario sarcástico hace referencia. A través del estudio de este dataset se obtuviero la siguiente información, la cual será explicada por…

//exploración parte 1

//javo

Lo primero que se hizo fue analizar las palabras presentes en cada tipo de comentario. Para esto, primero se modificó el texto de cada comentario para eliminar mayúsculas (y caracteres especiales?) y de esta forma obtener una cuenta más precisa. Este proceso se llevó a cabo tanto para los comentarios padre como para los comentarios sarcásticos. Como resultado de este análisis se obtuvieron los siguientes gráficos. El primero ilustra las 20 palabras más frecuentes en los comentarios no-sarcásticos, mientras que el segundo hace lo mismo para los comentarios sarcásticos. Como se evidencia en los gráficos, hay bastante similitud entre ambas figuras. Sin embargo, podemos notar que en los comentarios sarcásticos existe la presencia de las palabras “yeah”, “well”, “right”,”you’re”, “sure” y “can’t” dentro del top 20, por lo que podemos inferir que la presencia de una o más de estas palabras es un fuerte indicador de la presencia de un comentario que debe ser interpretado de forma sarcásrica.

//exploración parte 2 y 3

//Sali

Posterior a esto, se buscó la cantidad de subreddits con mayor participación, ya que se consideró que este sería un dato necesario si en algún momento se decidiera usar el subreddit de procedencia como objeto de estudio. Al llevar a cabo el análisis se obtuvieron los siguientes resultados:

Finalmente, se procedió a medir el largo de los comentarios sarcásticos en comparación a los comentarios padre, para determinar si existe alguna correlación entre la longitud de un comentario y la probabilidad de que su contenido sea sarcástico. A través de esto se obtuvieron los siguientes gráficos. Como se puede notar…

//preguntas

// ¿?

A partir de los datos estudiados, surgen interrogantes más complejas que podrían ser respondidas, tales como:

¿Qué tan necesario es el contexto para determinar si un comentario es sarcástico? ¿La presencia de un comentario padre podría facilitar la detección de este, o se puede analizar el contenido de un comentario sarcástico sin necesariamente conocer el contexto?

¿En qué subreddits el uso de sarcasmo es mayor proporcionalmente al número total de comentarios? Para esto se necesitaría una base de datos con el número de comentarios totales en cada subreddit, pero teniendo estos datos se puede llegar a conclusiones interesantes, como por ejemplo qué tipos de tema dan lugar a discusiones más tóxicas, y bajo qué temas se llevan a cabo discusiones más serias.

¿Es posible entrenar un clasificador para que pueda discernir entre un comentario sarcástico y uno no sarcástico?¿Bajo qué modelo podría llevarse a cabo este proceso?