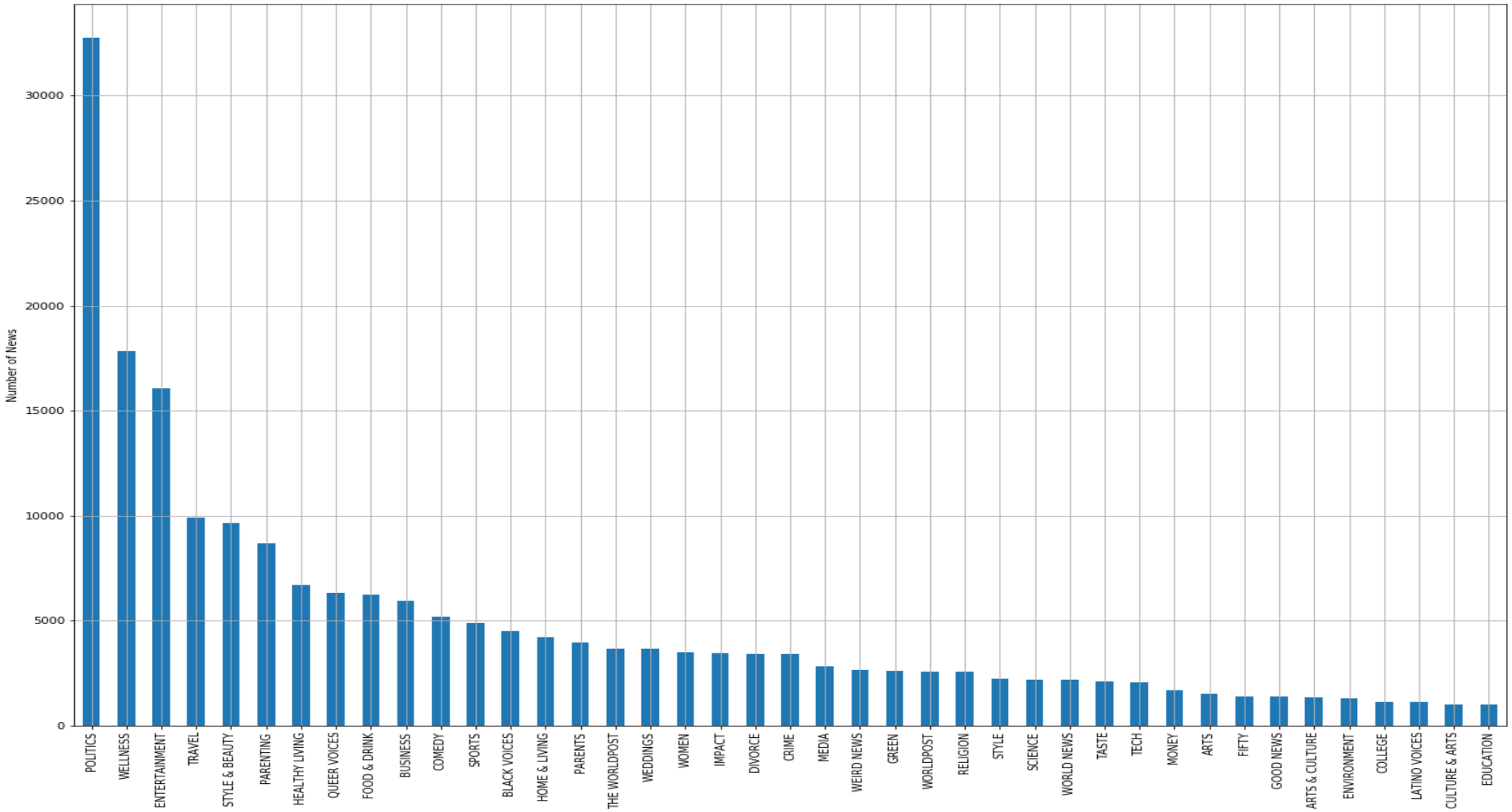
Clasificación de noticias

El objetivo principal es usar el conjunto de datos y aplicar ML para clasificación.

- Para este propósito, primero se debe realizar un análisis del tipo de datos con los cuales se va a trabajar. En este caso en específico trabajaremos con texto.
- Para aplicar los diferentes algoritmos de ML debemos realizar un preprocesamiento de los datos. Es proceso involucra eliminar caracteres especiales, convertir mayúsculas a minúsculas y realizar el proceso de lematización.
- Una vez realizado el preprocesamiento del texto, se deben sacar las características para realizar el entrenamiento del clasificador.
- Finalmente se escoge el clasificador y se evalúa su rendimiento.

Exploración de datos: Tipo de noticias en conjunto de datos



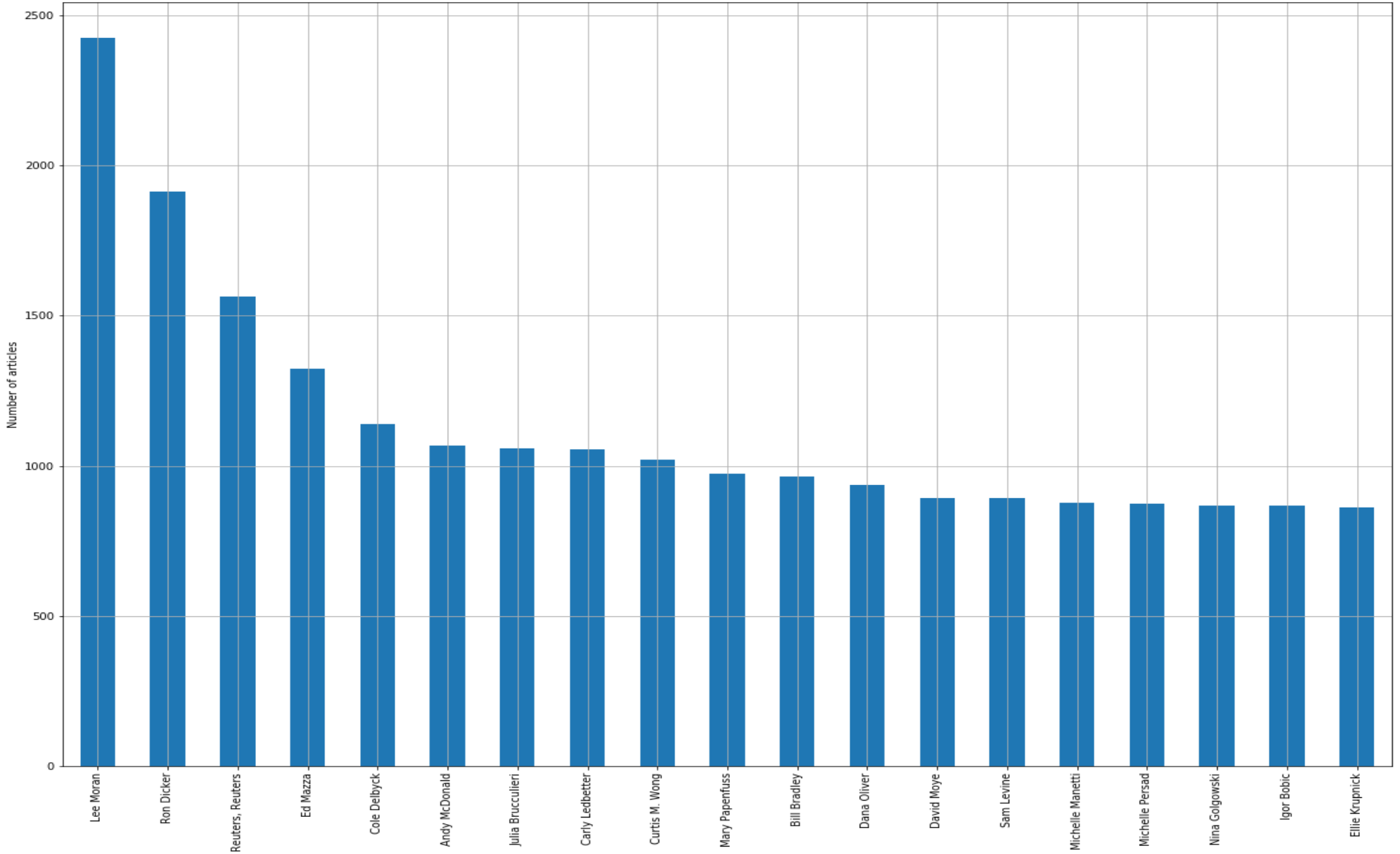
Aspectos importantes de la anterior gráfica:

- Se observa que la mayoría de los artículos corresponden a temas políticos, bienestar y entrenamiento.
- El número de artículos genera un desbalance por parte del clasificador, favoreciendo la clasificación a estas clases principalmente. Sin embargo, para ciertos propósitos como el análisis de noticias enfocadas en un solo tema no alteraría significativamente el resultado.
- Utilizando estas categorías de noticias y usando la librería SKLearn producimos las etiquetas para la clasificación.

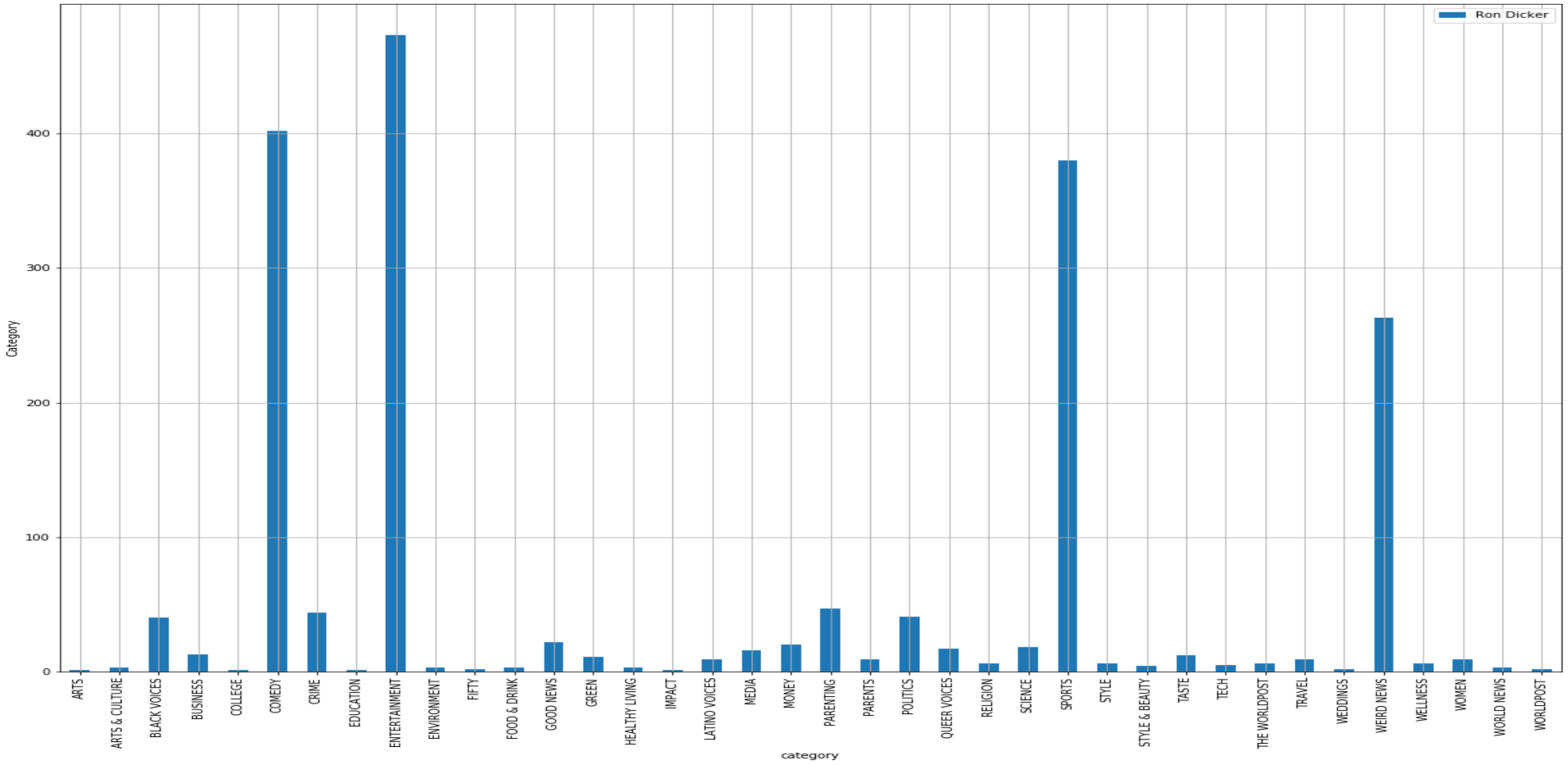
Escogencia de las características a evaluar:

- Uno de los aspectos importantes para lograr una buena clasificación en Machine Learning es escoger adecuadamente las características que el clasificador usará.
- Para esto, se utilizaron varios aspectos de procesamiento del lenguaje natural, usando la librería NLTK para realizar el procesamiento de los títulos de las noticias y las descripciones breves.
- Además, observando los autores de los artículos y las categorías a las que escriben frecuentemente se podría obtener información relevante.

Autores con más artículos:



Categoría por escritor:



Clasificador:

- Usando, el autor, el título de las noticias y la breve descripción de las noticias se vectorizó el texto.
- Para la clasificación de texto, se ha observado que algoritmos como KNN, naive Bayes y SVM logran una buena clasificación.
- En este caso se utilizó el clasificador de naive Bayes. Utilizado un conjunto de entrenamiento del 70% del total de las 200853.
- Evaluando en el conjunto de prueba se obtuvo que la exactitud de entrenamiento es del 62%.