Sol V. Anselmo Agustin H. Mattiussi Agustina S. Ortu M. Mercedes Baron

Clasificación por Temáticas de las Noticias de la BBC

Resumen

Este trabajo tiene como objetivo el análisis de artículos de noticias de la BBC y a partir de esto, proponer una solución para facilitar el etiquetado de los mismos y que este se realice con mayor eficiencia mediante la identificación por categorías y tendencias teniendo en cuenta el contenido del artículo. De esta manera, el etiquetado adecuado de los artículos atraerá a quienes están interesados en cada tema específico.

Se investigará si es posible mediante herramientas de procesamiento de lenguaje natural identificar la temática de los artículos según el lenguaje utilizado. También se quiere realizar extracción de información sobre cada temática y lograr identificar tendencias, personas relevantes a la categoría, eventos de importancia, entre otras cuestiones de interés.

Para lograr esto se trabajará con un corpus de artículos de la BBC sobre diversas temáticas.

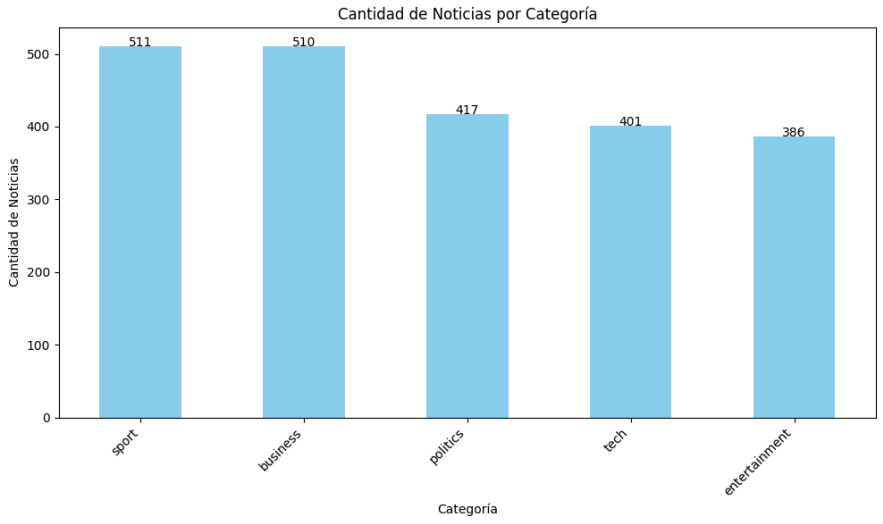
Corpus

Este conjunto de datos comprende artículos de noticias de la BBC a texto completo, cada uno categorizado en una de las cinco categorías distintas: negocios, entretenimiento, política, deporte y tecnología. Consta de 2225 documentos del sitio web de noticias de la BBC correspondientes a historias de 2004-2005. Los datos están etiquetados, lo que proporciona una clasificación clara para cada artículo, lo que los hace adecuados para diversas tareas de procesamiento del lenguaje natural, como la clasificación de textos, el modelado de temas y el análisis de sentimientos.

El corpus cuenta con los siguientes campos:

- **data**: texto de la noticia.
- **labels**: una etiqueta que clasifica a la noticia según su temática (las categorías son excluyentes, es decir, un artículo presenta solo una etiqueta).

A continuación podemos ver una figura que nos muestra la cantidad de noticias que hay para cada categoría:



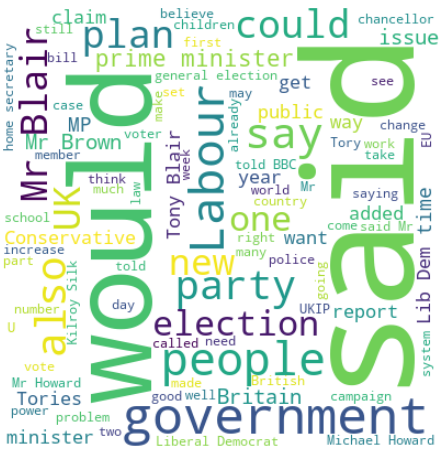
Sol V. Anselmo Agustin H. Mattiussi Agustina S. Ortu M. Mercedes Baron

Análisis exploratorio

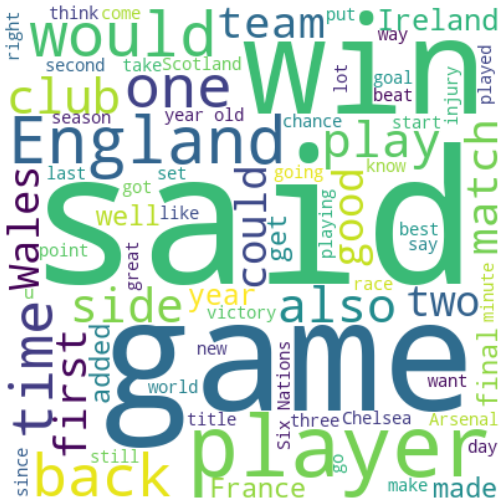
A continuación se muestra la representación de cloud of words de todos los artículos para cada temática:

Negocios

Entretenimiento



Política



Deporte

Tecnología

De esta manera, podemos comprender rápidamente el contenido del dataset, visualizando de forma clara e intuitiva las palabras claves utilizadas en las noticias de las diferentes categorías.

Luego se realizó un análisis similar pero para las palabras que comienzan con mayúscula para ver si se pueden detectar títulos, nombres, lugares, etc.

Sol V. Anselmo Agustin H. Mattiussi Agustina S. Ortu M. Mercedes Baron

Negocios

Entretenimiento

Política

Deporte

Tecnología

Propuesta de análisis

Para detectar la temática de un artículo mediante técnicas de Procesamiento de Lenguaje Natural, se evaluarán las siguientes técnicas de análisis.

- **BERT (Bidirectional Encoder Representations from Transformers)**: Se realizará un proceso de ajuste fino (fine-tuning) de un modelo BERT previamente entrenado en artículos con etiquetas conocidas. Posteriormente, este modelo podrá ser utilizado para clasificar noticias desconocidas y determinar su temática. Cabe aclarar que este modelo, que tiene contextos limitados, puede usarse ya que la longitud de los artículos no supera las 10 líneas.
- **TF-IDF (Term Frequency-Inverse Document Frequency)**: Cálculo la frecuencia de palabras y frases clave en los reportes de la BBC de cada temática. El objetivo es identificar qué tan importante es un término en el documento. La idea sería obtener las keywords del artículo, luego agruparlas por categorías y ver dentro de cada categoría las keywords que más frecuentes.