Análisis de texto

Leslye Hernández
1819111
Universidad Autónoma de Nuevo León
Facultad de Ciencias Físicas Matemáticas
Maestría en Ciencias de datos

Resumen

Este trabajo presenta un análisis de clasificación de textos basandose en noticias de internet de la plataforma `https://www.bbc.com`, se eligieron noticias de la sección sport e internacionales y se pretende predecir de que tipo de sección pertenece cada nota a través de herramientas de programación como el lenguaje de Python.

Introducción

La clasificación de textos es una tarea fundamental en el campo de procesamiento de lenguaje natural (NLP, por sus siglas en inglés), que tiene como objetivo asignar etiquetas o categorías a documentos o fragmentos de texto. Esta tarea es ampliamente utilizada en aplicaciones como la clasificación de correos electrónicos (spam vs. no spam), análisis de sentimientos, categorización de noticias y clasificación de opiniones de clientes, entre otros.

Enfocándonos en las noticias, estas presentan diferentes secciones, cada una distinguida entre el tema principal y audiencia a la cual va dirigida, en el siguente documento se trabajará con la sección de 'sport' y la sección 'internacional', la cual etiquetaremos como 'news'.

- **Noticias Deportivas:** Se distingue por un lenguaje dinámico, se enfoca en argumentar, criticar o describir eventos, resultados, jugadores, equipos, y todo lo relacionado con el mundo del deporte. Entre sus temas principales se incluyen estadísticas, análisis de rendimiento, entrevistas a deportistas y entrenadores.
- **Noticias Internacionales:** Este tipo de noticias abarca los eventos que ocurren en el mundo. Sus temas van cubren temas relacionados a la política, conflictos bélicos, acuerdos diplomáticos, derechos humanos, eventos sociales o económicos que tengan un impacto global. Su lenguaje es de un tono formal e informativo. La profundidad de la información es importante, ya que se intenta dar contexto completo a los eventos.

Estructura, análisis estadístico

Las noticias fueron extraídas de `https://www.bbc.com/` , desde la sección de 'sport' y 'news' , por medio de la librería `BeautifulSoup`. Se extrajeron un total de 288 notas, de las cuales 153 notas son extraídas desde 'news' y el restante, 195, son de 'sport'. La distribución se aprecia en la figura 1.

Frecuencia

En la figura 2, se muestran las palabras más repetidas dentro de las noticias, se puede resaltar que las que mayormente tienen relevancias adjetivos como verbos, mientras que en la figura 3, están representadas el top 20 de las palabras con mayor frecuencia dentro de las noticias de la sección de deportes, en ella se pueden observar que la palabra que más aparece es 'England', por lo que se puede asumir, que mayormente se habla de los deportes a nivel nacional, ya que, la cadena BBC es de Inglaterra. Por otro lado, en la figura 4, se observa el top 20 de la palabras que más aparece en la sección de noticias internacionales y por el contrario a deportes, la palabra 'Trump', que es el apellido del presidente de U.S.A. , también se aprecian palabras como 'president', 'people', entre otras, lo cual no abre un enfoque que la mayoría de las notas tienen a tocar temas sociales.

Conteo

Palabras

La figura 5, reprsenta el tamaño de las palabras en cada sección de noticias, se observa en como las palabras de la sección de noticias internacionales. 'news', tienen una mayor longitud, esto puede ir ligado con la formalidad que se emplea para describir una situación.

Por otro lado, la figura 6, señala la probabilidad del tamaño del texto por sección y vemos como para la categoría de 'sport', que es de deportes, existe apróximadamente un 0.35 de probabilidad de encontrarse un texto más corto a diferencia de las noticias internacionales que además de tener palabras más extensas en longitud, también tienen textos más largos.

Carácteres

La figura 7, señala el número de carácteres por sección, se observa como la sección de 'sport' suele tener más número de carácteres, lo que se puede interpretar que las oraciónes usalmente son mucho más cortas para explicar el suceso, a diferencia de la sección de 'news' que se emplean más carácteres para expresar la noticia.

Densidad

La densidad de las palabras se refiere a la frecuencia o cantidad de veces que una palabra aparece dentro de un texto, un documento o un conjunto de textos, en relación con el total de palabras o el número total de términos. Es un concepto utilizado principalmente en el análisis de textos, especialmente en procesamiento de lenguaje natural (NLP) y optimización de motores de búsqueda (SEO).

Densidad de la palabra= ($No.VecesQueAparecеLaPalabraEnElTexto/TotalDePalbrasEnElTexto$)*100
En la figura 8, se puede observar la densidad de las palabras por sección, y con el paquete de `spacy` se obtiene:

- Densidad de palabras para 'sport': 0.63
- Densidad de palabras para 'news': 0.59

Esto se interpreta en que para la sección de 'sport', el 63 porciento de las palabras son significativas para el análisis y el 37 porciento restante son stopwords, las cuales no aportan mucho valor en términos de contenido relevante para tareas como análisis de sentimientos, clasificación de texto, etc. Mismo caso para la sección de 'news', esto indica que el 59 porciento son palabras relevantes para su análisis.

Modelos

Para hacer la clasificación correspondiente se utilizó el paquete `sklearn` , con el 40 porciento de la muestra como de entrenamiento y el 60 porciento restante de prueba.

Se utilizaron 4 diferentes modelos en los cuales, se disntiguen por las siguientes características:

- Máquinas de Vectores de Soporte (SVM): Es un modelo utilizado en clasificación binaria
- Random Forest: Este modelo utiliza varios árboles y promedia sus predicciones, es bueno para datos estructurados
- Naïve Bayes: Este modelo está basado en probailidades y es útil cuando los datos pequeños e independientes entre sí.
- Logistic: Este tipo de modelo aprende una relación entre las características del texto y la probabilidad de que pertenezca a una categoría

Los resultados de cada modelo se encuentran en el cuadro 1.

Interpretación:

En el cuadro 1 podemos ver las diferencias entre los modelos, tanto en la precisión, que nos dice si el modelo es confible cuando predice una clase positiva, es decir, que la noticia es de la sección de deportes cuando en realidad es de deportes, entre mayor sea el número , mayor precisión tendrá, para los 4 modelos la precisión es igual y muy alta, por lo que la presición del modelo es buena.

También se tiene F1 - Score, este nos indica que existe un buen equilibrio entre la precisión y recall, es decir, entre las veces que el modelo acierta y el número de veces clasifica de forma correcta la noticia, es muy útil cuando las clases están desequilibradas, en nuestro caso tienen un alto score, pues debido a que ambas categorías tienen muy parecido número de noticias.

Por último tenemos la descripción ROC que son las siglas en inglés de Area Under Curver, que es la área bajo la curva, esta medida distingue las clases y muestra la relación entre la Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR).

Para todos nuestros modelos están cercanos a 1 lo que se interpreta como una gran capacidad para diferenciar entre las secciones de noticias, sin embargo, la que tiene mayor precision es el modelo de Random Forest.

Predicción

Después de entrenar el modelo se realizó un nuevo ejemplo con noticias que no estaban dentro de

1. Noticia: *Netanyahu seeks strong backing from Trump, as first foreign leader to visit.*
   Link: `https://www.bbc.com/news/articles/c62e7d6r08ro`
2. Deportes: *Villa sign Asensio and agree Disasi loan deal.*
   Link: `https://www.bbc.com/sport/football/articles/cj918vvmz0go`

Los resultados se observan en el cuadro 2. Los cuales se analizan que en todos los modelos acierta en la clasificación.

Conclusiones finales

Al elegir modelos para hacer la clasificación de textos hay que tomar en cuenta la longitud del data set y si se tienen con antelación clasificados los textos, pues para un dataset pequeño, los modelos más simples y rápidos como Naive Bayes y Logistic Regression son una buena opción, ya que son menos propensos al sobreajuste y funcionan bien con menos datos. Naive Bayes es especialmente eficiente para la clasificación de texto, mientras que Logistic Regression puede ser útil si las relaciones entre las características son lineales.

Si se tiene más datos y se quiere explorar interacciones más complejas, SVM y Random Forest pueden ser útiles, pero es importante ajustar los hiperparámetros para evitar sobreajuste.

Referencias

- BBC. (2025). Disponible en: `https://www.bbc.com/`.
- Hovy, D. (2020) "Text Analysis in Python for Social Scientists". Cambridge University.
- Sarkar, D. (2016) "Text Analytics with Python". Apress.
- Bengfort, B. (2018) "Applied Text Analysis with Python". OReilly.

Anexos

Cuadro 1: Modelos.

Modelo	Precisión	F1 - Score	ROC
Random Forest	0.9870	0.9934	0.9913
Multinomial	0.9870	0.9934	0.9897
Logistic	0.9870	0.9934	0.9888
SVM	0.9870	0.9934	0.9882

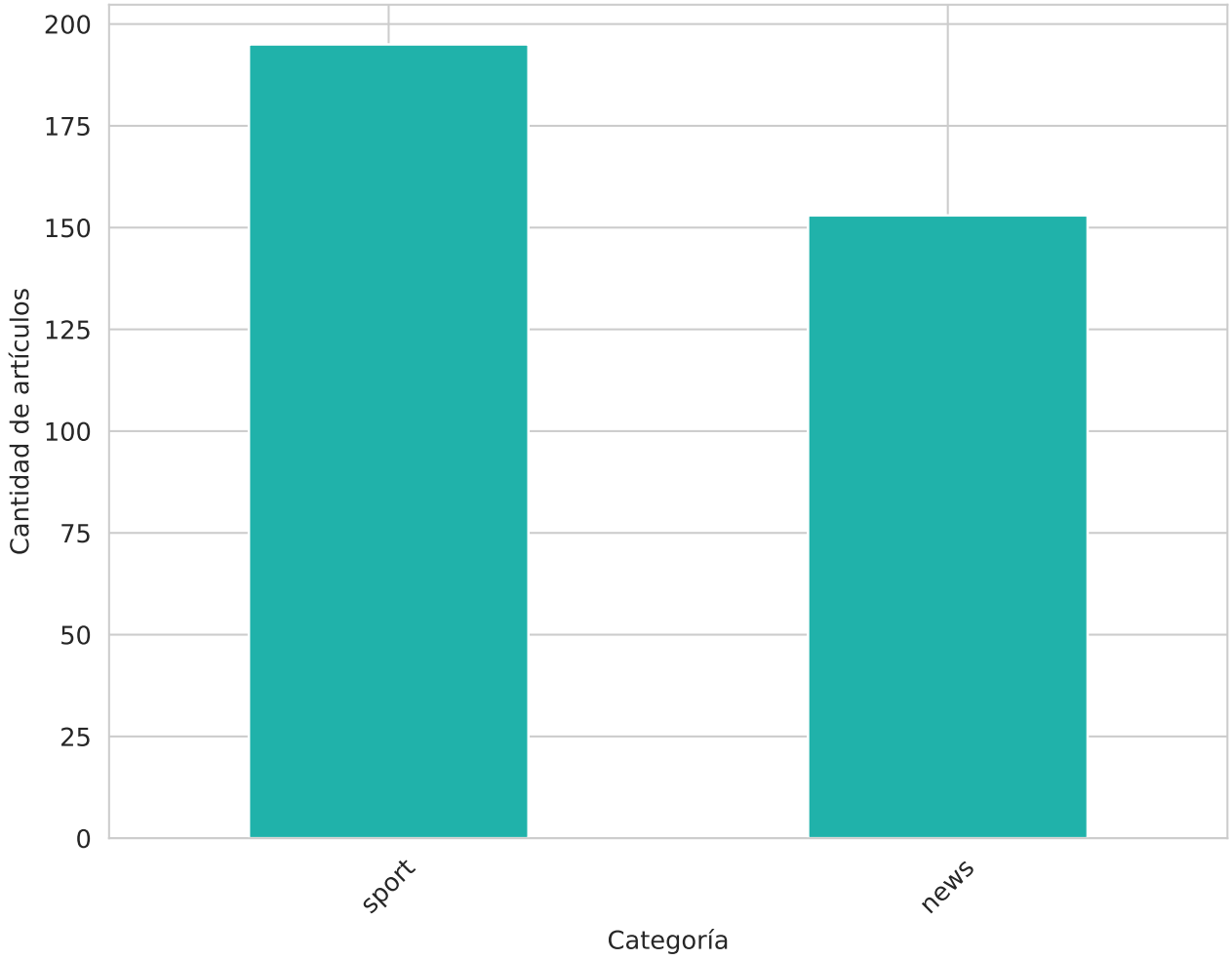

Figura 1: Histograma de número de artículos por sección - Elaboración propia.

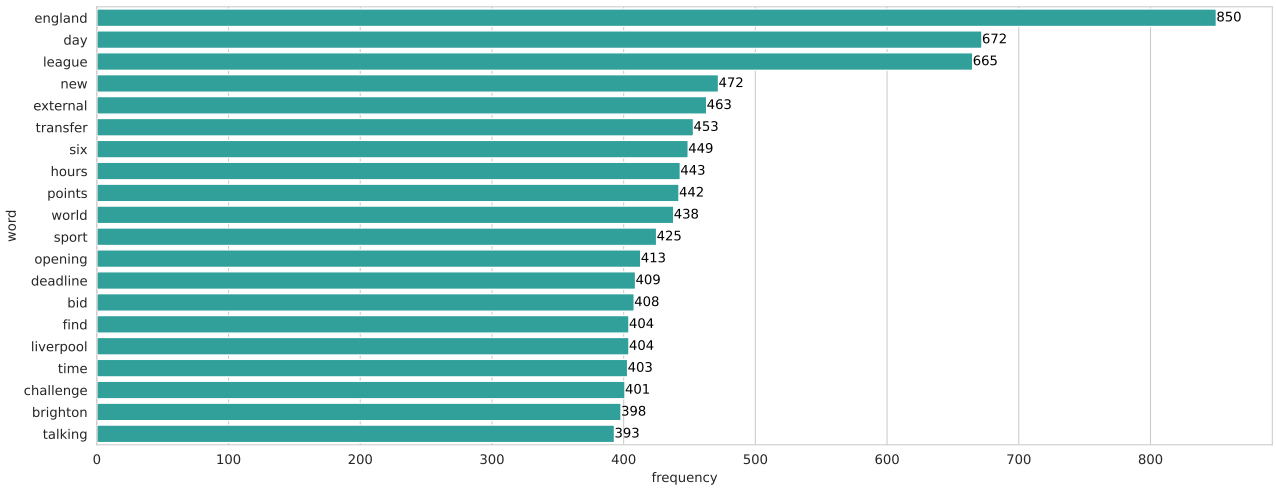

Figura 2: Top 20 de palabras con mayor frecuencia - Elaboración propia.

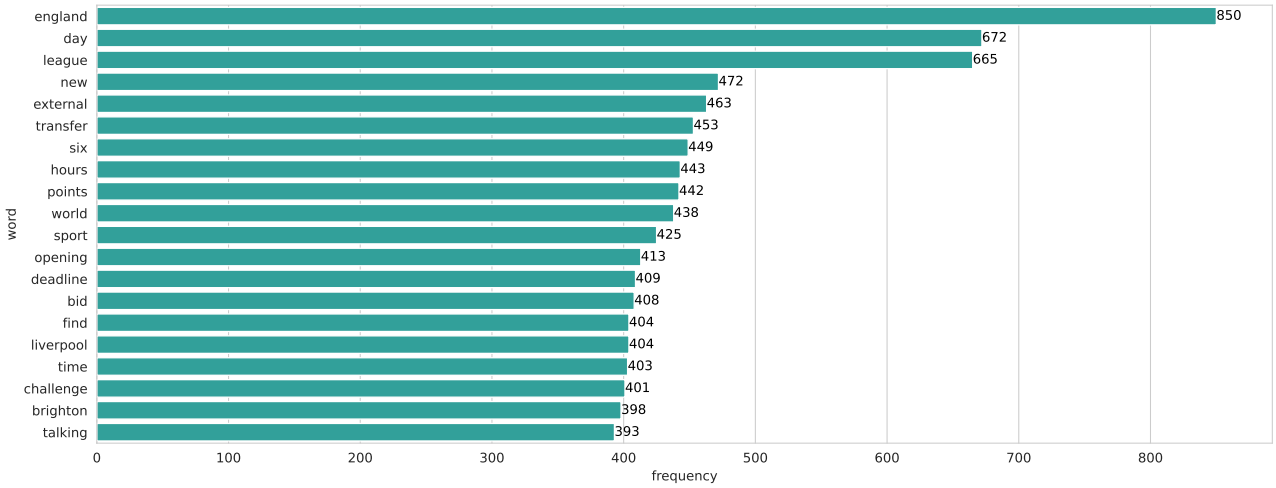

Figura 3: Top 20 de palabras con mayor frecuencia de la sección de deportes - Elaboración propia.

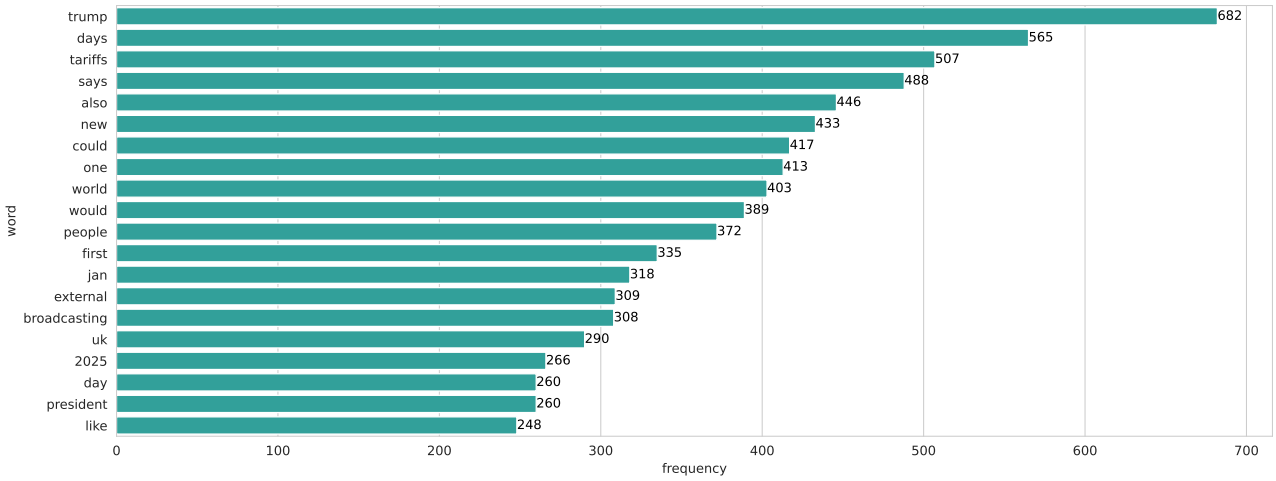

Figura 4: Top 20 de palabras con mayor frecuencia de la sección de noticias internacionales. - Elaboración propia.

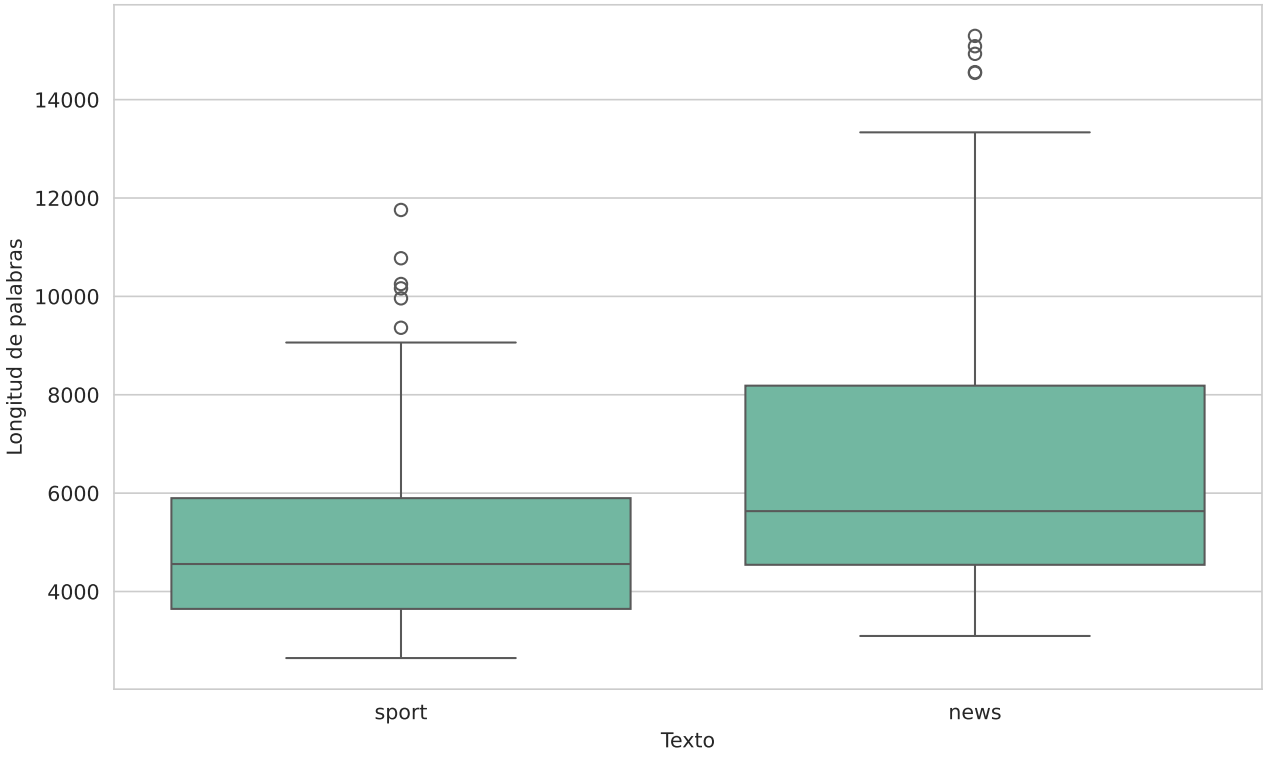

Figura 5: Longitud de palabras por sección. - Elaboración propia.

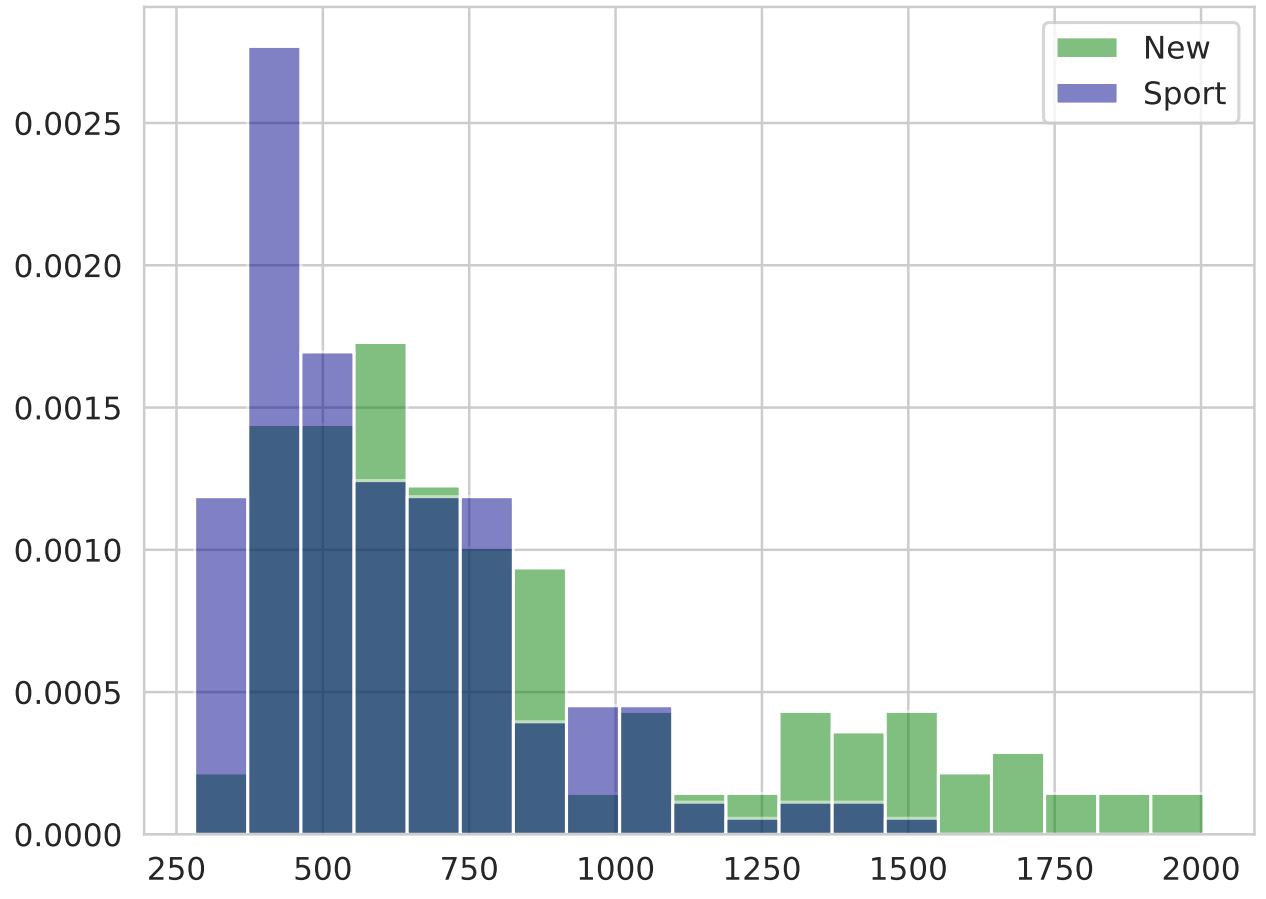

Figura 6: Conteo de palabras por sección. - Elaboración propia.

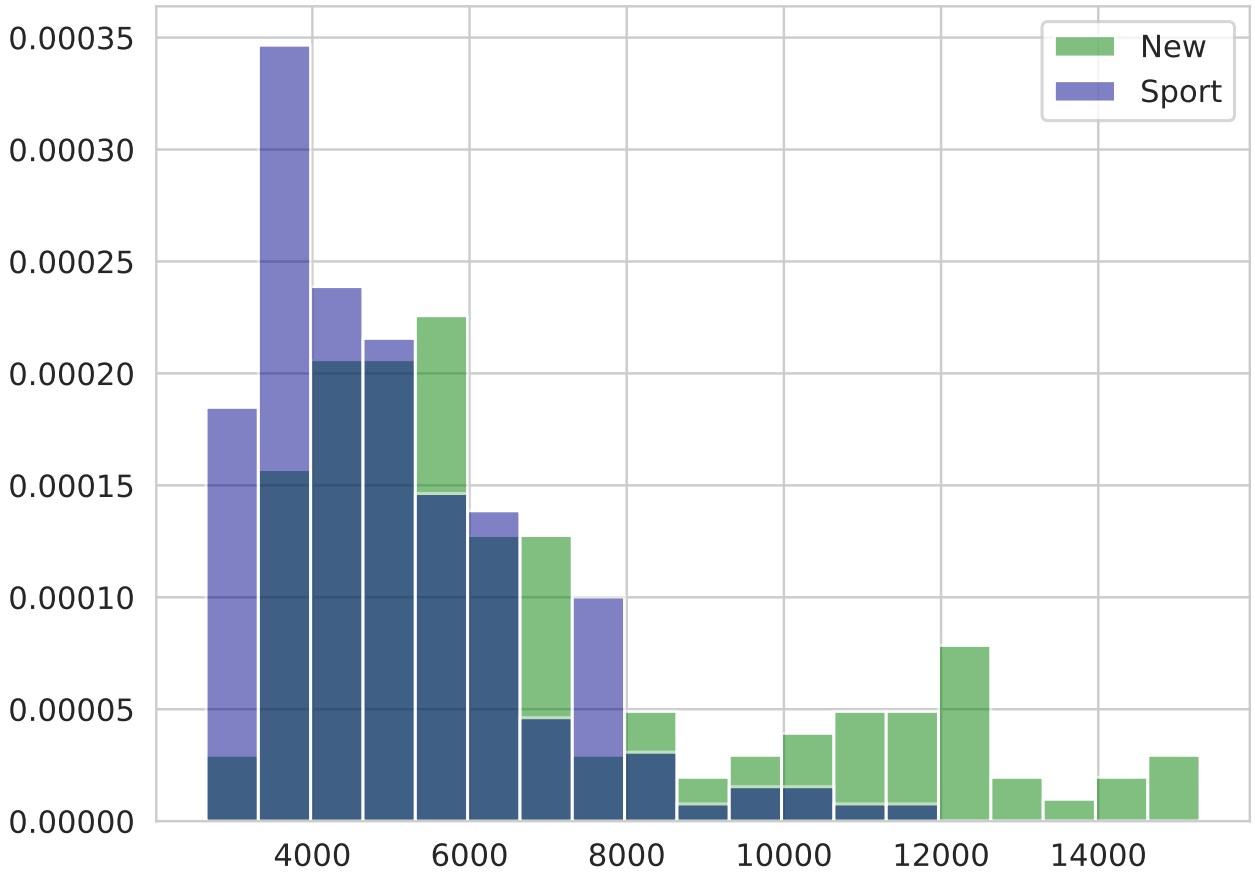

Figura 7: Conteo de cáracteres por sección. - Elaboración propia.

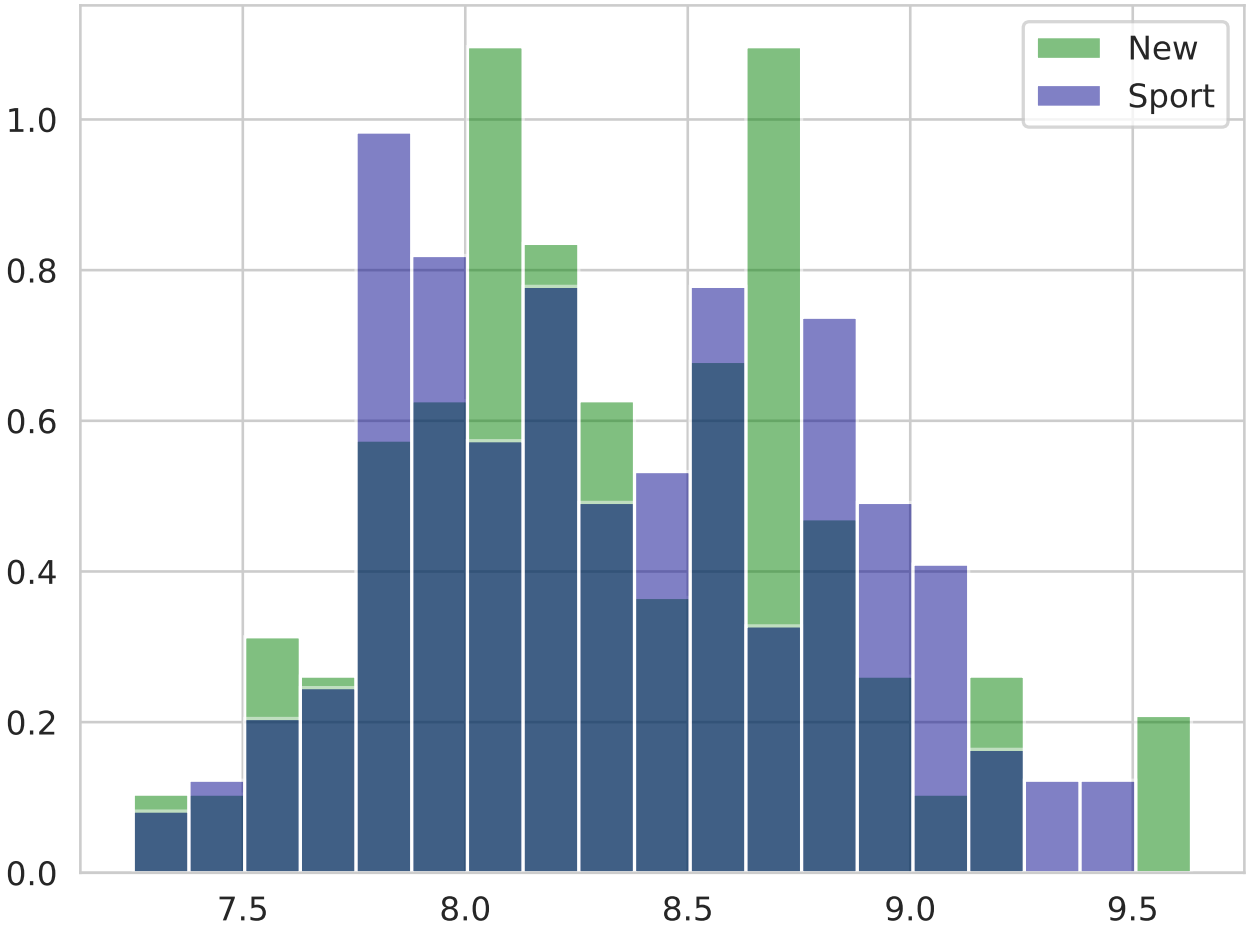

Figura 8: Densidad de palabras por sección. - Elaboración propia.

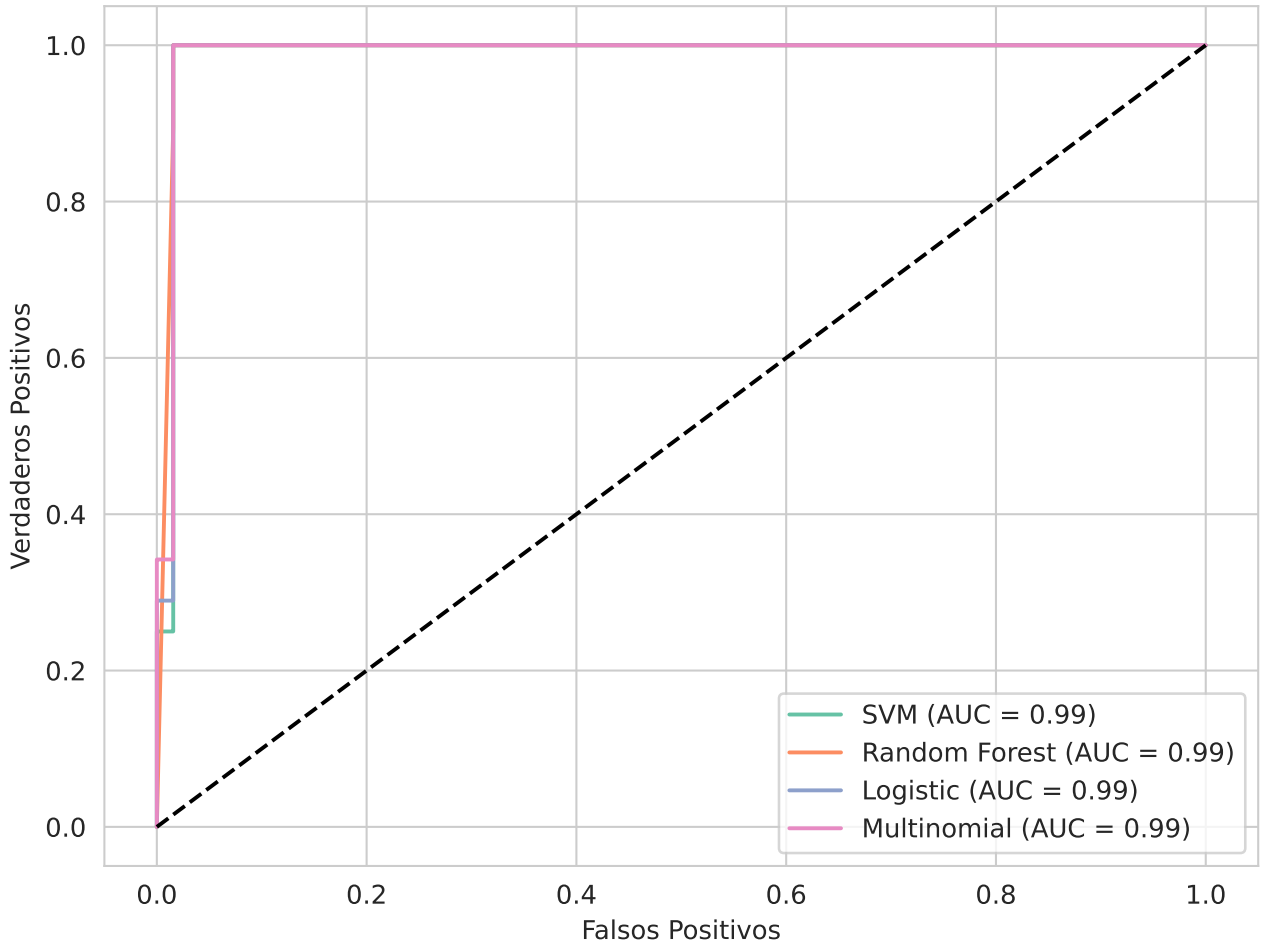

Figura 9: Área bajo la curva por modelo. - Elaboración propia.