Clasificación del Corpus BBC News Summary utilizando J48 en Weka

Classification of the BBC News Summary Corpus using J48 in Weka

Daniel Avelar Jaime[1], Misael López Ramírez[1], Claudia Angélica Rivera Romero[2], Rafael Guzmán Cabrera[1]
[1] Universidad de Guanajuato
[2] Universidad Autónoma de Zacatecas
d.avelarjaime@ugto.mx[1] lopez.misael@ugto.mx[2] c.a.riveraromero@uaz.edu.mx[3] guzmanc@ugto.mx[4]

Resumen

El artículo presenta una metodología para la clasificación de noticias de la BBC utilizando técnicas de aprendizaje automático y procesamiento de lenguaje natural. Se destaca la flexibilidad en el preprocesamiento, con opciones como eliminación de stopwords y ajuste de umbral. Se utiliza el algoritmo J48 para la clasificación, obteniendo una precisión del 81.79%. Además, se resalta el uso efectivo de la ganancia de información como parte del proceso para mejorar la calidad de los resultados.

Palabras clave: Clasificación de noticias, Aprendizaje automático, Preprocesamiento, Algoritmo J48, Weka

Introducción

La clasificación de noticias de la BBC, que son documentos estructurados con un formato específico, implica categorizar estas noticias en diferentes temas o secciones, como deportes, política, entretenimiento, etc. Para llevar a cabo esta tarea, se pueden utilizar diversas técnicas de aprendizaje automático, como árboles de decisión, sistemas basados en reglas o incluso redes neuronales, dependiendo de la complejidad de la clasificación requerida. La elección de la técnica dependerá de las necesidades específicas de la tarea y de la cantidad de datos disponibles.

La clasificación de noticias estructuradas es crucial en el procesamiento de datos, ya que permite automatizar la organización y etiquetado de contenido informativo, lo que resulta en una mejor gestión y acceso a la información (Yu et al., 2023). Esta automatización puede ser beneficiosa en aplicaciones empresariales y gubernamentales, ya que agiliza la búsqueda de información y facilita el seguimiento de eventos relevantes en tiempo real (Jurafsky & H. Martin, 2023).

Para resolver este problema, se utilizan técnicas de aprendizaje automático y PNL (procesamiento del lenguaje natural) (Hu et al., 2021). Los enfoques comunes incluyen el uso de algoritmos de aprendizaje supervisado, como el clasificador J48.

J48 es un algoritmo de aprendizaje automático que crea árboles de decisión, útiles para clasificar noticias. Es fácilmente interpretable, automatiza la clasificación, ofrece precisión sólida y es escalable, siendo una opción eficaz para asignar categorías a noticias (H. Witten et al., 2016).

El proceso de clasificación de noticias consta de varios pasos clave:

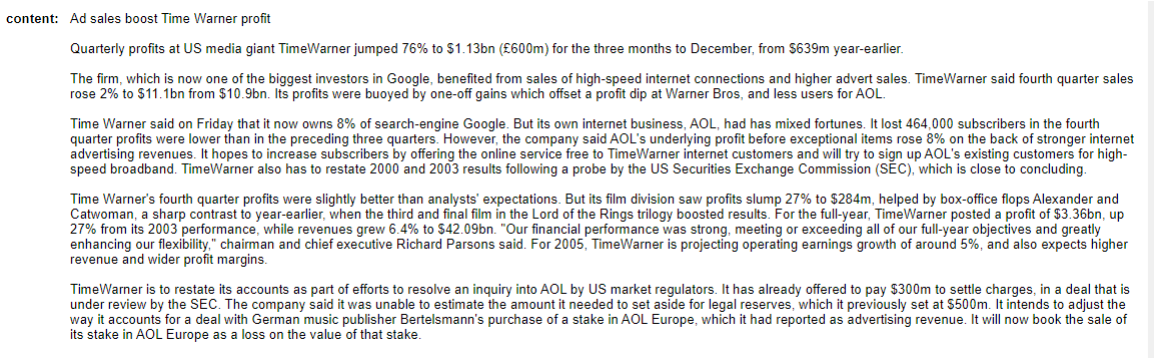
- Preprocesamiento: Esto implica limpiar y normalizar los documentos, lo que incluye la eliminación de caracteres no deseados y palabras irrelevantes.
- Extracción de características: Se extraen características numéricas relevantes de los documentos, como la frecuencia de palabras o características específicas del dominio.
- Entrenamiento de modelos: Utilizando conjuntos de datos etiquetados, se entrenan modelos de clasificación mediante algoritmos de aprendizaje automático para asignar características a etiquetas de clasificación.
- Evaluación y ajuste: El rendimiento del modelo se evalúa utilizando métricas como la precisión y la recuperación. Si es necesario, se ajustan parámetros o se prueban diferentes algoritmos para mejorar los resultados.
- Predicción: Una vez entrenado y evaluado, el modelo se usa para clasificar nuevos documentos asignándoles etiquetas basadas en el aprendizaje realizado durante el entrenamiento (Raja et al., 2022).

En este trabajo se utilizó el Corpus “BBC News Summary” el cual se refiere a un conjunto de noticias de la BBC que se han recopilado y organizado en un corpus para su uso en investigaciones, análisis de tendencias y otros fines similares.

Este corpus contiene un total de 2225 noticias las cuales están clasificadas en 5 categorías las cuales se dividen de la siguiente manera:

- Business tiene 510 noticias
- Entertainment tiene 386 noticias
- Politics tiene 417 noticias
- Sport tiene 511 noticias
- Tech tiene 401 noticias

A continuación, en la figura 1 muestra un ejemplo de una noticia de la categoría Business.



content: Ad sales boost Time Warner profit

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier.

The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL.

Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the online service free to TimeWarner internet customers and will try to sign up AOL's existing customers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the US Securities Exchange Commission (SEC), which is close to concluding.

Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins.

TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to settle charges, in a deal that is under review by the SEC. The company said it was unable to estimate the amount it needed to set aside for legal reserves, which it previously set at $500m. It intends to adjust the way it accounts for a deal with German music publisher Bertelsmann's purchase of a stake in AOL Europe, which it had reported as advertising revenue. It will now book the sale of its stake in AOL Europe as a loss on the value of that stake.

Figura 1 Ejemplo de una noticia Business
Fuente: Elaboración propia.

Weka es un software de código abierto utilizado en minería de datos y aprendizaje automático (Jurafsky & H. Martin, 2023). Ofrece una interfaz gráfica fácil de usar y una variedad de algoritmos para análisis de datos. Para clasificar noticias de la BBC, se desarrolló un código en Python que permitía un preprocesamiento personalizado, incluyendo opciones como stopwords, lematización y umbral. Esto generó un archivo .arff que Weka pudo manejar, y se aplicó el algoritmo J48 de aprendizaje automático para la clasificación. Weka y el código en Python se utilizaron en conjunto para lograr resultados precisos en la categorización de noticias de la BBC. En el siguiente punto, se explicará con más detalle la metodología utilizada.

Existe varios trabajos en donde se clasificación noticias ya sea del mismo corpus o parecidos. En el trabajo (Krishnan et al., 2019) presentó un modelo de abstracción de texto basado en la extracción supervisada de oraciones clave de documentos fuente. Se enfoca en resaltar características esenciales para seleccionar las oraciones más relevantes. La evaluación se llevó a cabo utilizando datos de BBC News Summary y la métrica ROUGE. Además, se comparó este modelo con otros, como KNN, Bagging, Random Forest, SMO, Naive Bayes y J48. En el trabajo se resalta la importancia de la abstracción de texto como una herramienta para facilitar la comprensión de grandes volúmenes de información.

El trabajo (Patel & Patel, 2022) se centró en la utilización de la red neuronal Generative Pretrained Transformer (GPT) para resumir noticias financieras y destaca su potencia en la summarización de texto en general. Aunque se utiliza un corpus de noticias financieras en lugar del BBC News Summary Dataset, se sugiere que los resultados podrían ser relevantes para este último.

En el (Viet Thang Ho Chi et al., 2021), se presentó un concepto de resumen extractivo que utiliza la puntuación de las oraciones para generar un resumen que maximice la cobertura de la información importante y evite la redundancia. Se utiliza Programación Lineal Entera (ILP) para resolver la función objetivo de suma máxima de los puntajes de las oraciones, con la restricción habitual de longitud del resumen. Se comparó el rendimiento del enfoque ILP con el enfoque moderno utilizando BERT en dos conjuntos de datos de resumen de noticias, la BBC y CNN/DailyMail.

En el (Deokar & Shah, 2021), los autores investigaron la automatización de la síntesis de noticias utilizando modelos de aprendizaje automático como BART y T5. Utilizaron técnicas de web scraping para recopilar datos y corpus relevantes para su estudio como lo es BBC News Summary.

Marco teórico

J48 es un algoritmo de aprendizaje automático utilizado en minería de datos para crear árboles de decisión. Utiliza un conjunto de datos de entrenamiento para construir un árbol que permite tomar decisiones basadas en reglas. Cada nodo del árbol representa una característica de los datos y se ramifica según criterios que maximizan la homogeneidad de las clases. Estos árboles se emplean para clasificar nuevas instancias de datos o predecir resultados en función de sus características. J48 es una implementación específica de C4.5 en el software Weka (H. Witten et al., 2016).

El término "Weka" se refiere a una suite de software de código abierto desarrollada en la Universidad de Waikato, Nueva Zelanda, utilizada para minería de datos y aprendizaje automático. Weka ofrece una variedad de algoritmos de análisis de datos y proporciona una interfaz gráfica de usuario que facilita la carga de datos, la aplicación de algoritmos y la evaluación de modelos. Es una herramienta versátil para investigadores y profesionales que desean realizar tareas de análisis y modelado de datos sin necesidad de programación intensiva (H. Witten et al., 2016).

El Procesamiento Natural del Lenguaje (PNL) es un campo de la informática que se centra en la interacción entre las computadoras y el lenguaje humano. Se utiliza para desarrollar sistemas que comprenden y generan lenguaje humano, como chatbots y traductores automáticos, y también es esencial en la clasificación automática de noticias y textos (Jurafsky & H. Martin, 2023).

Metodología propuesta

La metodología de clasificación del corpus de la BBC News Summary se compone de varios pasos esenciales. Comienza con la selección del corpus y continúa con el preprocesamiento, que incluye la tokenización, eliminación de stopwords, filtrado por frecuencia y lematización. Además, se desarrolló un código en Python que permite a los usuarios elegir si desean usar stopwords, realizar lematización y definir un umbral para la frecuencia de palabras, generando así un archivo preprocesado.

Posteriormente, en Weka, se aplica un modelo de aprendizaje automático, como J48, y la opción de ganancia de información es opcional. Si se elige utilizarla, esta se realiza en Weka. Luego, la información generada en Weka se limpia con un macro desarrollado en Excel para prepararla nuevamente para su uso en Weka. Finalmente, se procede con la clasificación de resúmenes, proporcionando resultados.

En la Figura 2 se presenta la metodología.

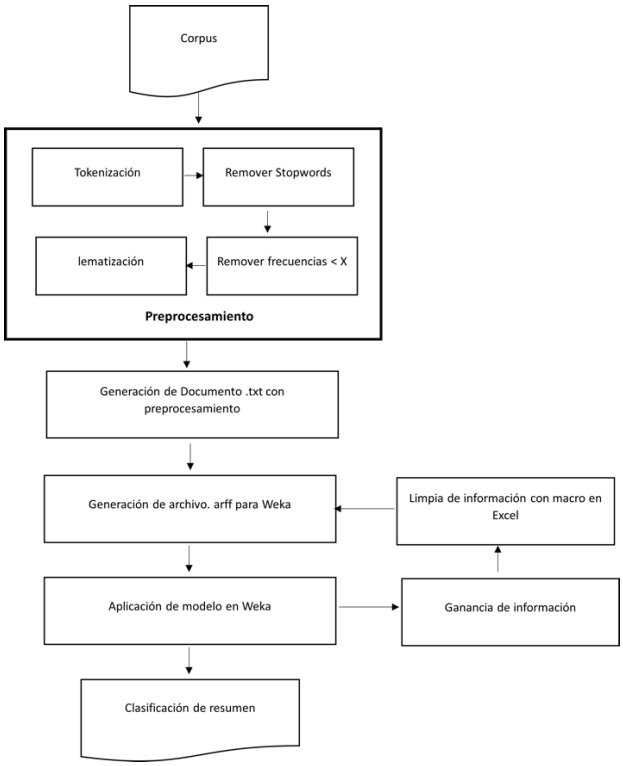


Figura 2 Metodología propuesta
Fuente: Elaboración propia.

Es importante destacar que esta metodología brinda flexibilidad en el proceso de preprocesamiento y en la elección de si se aplica o no la ganancia de información, lo que permite adaptar el enfoque a las necesidades específicas del usuario.

Resultados

Después de experimentar con la metodología presentada, se lograron resultados prometedores en la clasificación de noticias de la BBC. Utilizando el algoritmo J48, se obtuvo una precisión del 81.79%, lo que representa un rendimiento sólido en la asignación de categorías a las noticias.

Se realizaron pruebas que involucraron lematización, cambios en el umbral de frecuencia (eliminación de palabras poco comunes), eliminación de stopwords y el uso de ganancia de información para evaluar su impacto en la precisión de la clasificación de noticias. Estos ajustes se llevaron a cabo con el propósito de determinar si podían mejorar el rendimiento del modelo. Es importante destacar que los valores de precisión pueden variar según la configuración de los algoritmos, por lo que la optimización adicional podría ofrecer mejoras. Para obtener detalles específicos sobre los resultados, se presentan en la Figura 3.

```
Time taken to build model: 8.51 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1824               81.7937 %
Incorrectly Classified Instances       406               18.2063 %
Kappa statistic                          0.7712
Mean absolute error                      0.0838
Root mean squared error                  0.2569
Relative absolute error                 26.2952 %
Root relative squared error             64.3476 %
Total Number of Instances             2230

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.924    0.054    0.837      0.924    0.878      0.842    0.954     0.865     deporte
                 0.767    0.027    0.858      0.767    0.810      0.775    0.911     0.775     entretenimiento
                 0.751    0.077    0.744      0.751    0.748      0.672    0.862     0.673     negocios
                 0.787    0.042    0.810      0.787    0.799      0.753    0.890     0.724     politica
                 0.848    0.031    0.859      0.848    0.854      0.822    0.922     0.790     tecnologia
Weighted Avg.    0.818    0.048    0.818      0.818    0.817      0.771    0.908     0.765

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 473   8  19  10   2 |   a = deporte
  29 297  41   9  11 |   b = entretenimiento
  30  20 384  45  32 |   c = negocios
  21  10  47 329  11 |   d = politica
  12  11  25  13 341 |   e = tecnologia
```

Figura 3. Resultados en Weka
Fuente: Elaboración propia.

En la figura 3, se pueden apreciar resultados significativos. Durante el proceso de preprocesamiento, se implementó la eliminación de stopwords, mientras que la lematización se excluyó debido a que, en general, disminuía la precisión del modelo. Se estableció un umbral de 6 en el proceso y se empleó la ganancia de información como parte de la estrategia, lo que contribuyó al resultado satisfactorio que se presenta.

Con esto concluimos este trabajo, y queda claro que la flexibilidad en el proceso de preprocesamiento desempeña un papel crucial en la obtención de mejores resultados. Aunque llevar a cabo múltiples configuraciones de preprocesamiento puede ser un proceso que consume tiempo, es evidente que puede marcar la diferencia en la precisión del modelo. Además, es importante destacar que aún se están explorando diversas técnicas de aprendizaje automático para mejorar aún más los resultados en futuros trabajos.

Referencias

Deokar, V., & Shah, K. (2021). Automated Text Summarization of News Articles. *International Research Journal of Engineering and Technology*. https://www.irjet.net/archives/V9/i2/IRJET-V9I2192.pdf

H. Witten, I., Frank, E., A. Hall, M., & J. Pal, C. (2016). *Data mining: practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann. https://www.google.com.mx/books/edition/Data_Mining/1SylCgAAQBAJ?hl=es-419&gbpv=0

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2021). *Membership Inference Attacks on Machine Learning: A Survey*. http://arxiv.org/abs/2103.07853

Jurafsky, D., & H. Martin, J. (2023). *Speech and Language Processing* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

Krishnan, D., Bharathy, P., Anagha, & Venugopalan, M. (2019). *A Supervised Approach For Extractive Text Summarization Using Minimal Robust Features*. IEEE Xplore. https://doi.org/10.1109/ICCS45141.2019.9065651

Patel, P. M., & Patel, P. M. (2022). *Financial News Summarisation using Transformer Neural Network*. https://doi.org/10.21203/rs.3.rs-2132871/v1

Raja, R., Kumar Nagwanshi, K., Kumar, S., & Ramya Laxmi, K. (2022). *Data mining and machine learning applications*. Wiley. https://doi.org/10.1002/9781119792529

Viet Thang Ho Chi, L., Huynh Nhat Ho Chi, H., Dat Ho Chi, L., Viet Thang, L., Nhat Hao, H., & Phan Thanh Dat, L. (2021). Extractive Summarization with Integer Linear Programming. *Researchgate*. https://doi.org/10.13140/RG.2.2.14847.84640

Yu, J., Yin, H., Xia, X., Chen, T., Li, J., & Huang, Z. (2023). Self-Supervised Learning for Recommender Systems: A Survey. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2023.3282907