Predecir la popularidad de noticias en línea utilizando modelos de clasificación y/o regresión

Calderon, S., Arenas, K., Torres, J.

1. Resumen

2. Introducción

El estudio de la popularidad de las noticias en línea ha cobrado gran relevancia en la última década, ya que los medios compiten por la atención en un entorno digital saturado. La capacidad de predecir qué noticias serán populares es esencial para optimizar estrategias de contenido y aumentar la participación de la audiencia. Para abordar este desafío, se han desarrollado modelos basados en datos históricos y aprendizaje automático que consideran una variedad de características, como el contenido de la noticia, el comportamiento del usuario y las dinámicas de las plataformas de distribución. Los usuarios, quienes son los consumidores de estos medios usualmente comparten, se suscriben e interactúan con la información. La noción de popularidad suele expresarse investigando el número de interacciones en la web y las redes sociales, por ejemplo, la tasa de clics, el número de compartidos, me gusta y retweets [1]. El número de veces que se comparte una noticia es un buen indicador de su popularidad potencial. La predicción de la popularidad de un artículo tiene múltiples aplicaciones en el mundo real, desde el punto de vista de [**?**], hay dos caminos de predicción más conocidas: el primero relacionado con las características que se conocen luego de la publicación y los que no utilizan dichas características. El primero es el más conocido y más común [2, 3, 4, 5, 6]. De acuerdo a [7], la popularidad de un artículo candidato puede ser estimado utilizando un módulo de predicción y luego un módulo de optimización, según el autor sugiere cambios en el contenido y la estructura del artículo, con el fin de maximizar su popularidad esperada. Existen varios estudios que sugieren, la predicción puede ser alta utilizando los metadatos posteriores a la publicación, dándonos ventaja en la atención recibidas de la información luego de ser recibida [4, 5]. Por el contrario [7], menciona que al utilizar las características de metadatos antes de la publicación de los contenidos resulta ser difícil, aunque la precisión de la predicción esperada es comparativamente baja en el método de antes de la publicación, ya que estamos utilizando sólo características de metadatos en lugar del contenido original de la noticia.

El objetivo de este artículo es desarrollar un modelo de aprendizaje automático que logre predecir la popularidad de los artículos de noticias en línea antes de su publicación. La sección de estado del arte presenta trabajos previos en la materia, incluyendo la metodología y resultados. La sección Diseño del experimento se centra en explicar la metodología de este artículo, incluyendo la descripción del conjunto de datos y los modelos a ser entrenados.

3. Estado del arte

Esta sección proporciona un resumen cohesivo de estudios de investigación enfocados en medir y predecir la popularidad de los artículos de noticias en línea. Los dos primeros estudios abordan los desafíos de predecir la popularidad de las noticias en el momento de su publicación, mientras que los últimos tres profundizan en la aplicación de técnicas de aprendizaje automático en un único conjunto de datos para mejorar las capacidades predictivas de la popularidad de los artículos de noticias.

[8] investiga las dificultades asociadas con predecir la popularidad de los artículos de noticias inmediatamente después de su publicación. Utilizando un conjunto de datos de 13,319 artículos de Yahoo News, los autores critican investigaciones previas, que afirmaban una alta precisión en la predicción de

*Corresponding author

la popularidad de las noticias utilizando características basadas en el contenido. Se argumenta que la tarea es más compleja de lo que se pensaba, principalmente debido a la alta asimetría en la distribución de popularidad. Sus experimentos revelan que los métodos de clasificación y regresión existentes están sesgados hacia la predicción de artículos impopulares, fallando en identificar con precisión los populares, que son cruciales para la identificación temprana. El estudio concluye que los métodos actuales son inadecuados para predecir la popularidad de las noticias desde el inicio utilizando solo características basadas en el contenido, destacando la necesidad de enfoques más robustos.

En [9], los autores abordan el desafío de predecir qué artículos de noticias aparecerán en una lista de "más leídos". Proponen que la popularidad es un concepto relativo influenciado por el atractivo de los artículos publicados simultáneamente. Empleando una función lineal en una representación de bolsa de palabras, utilizan Máquinas de Soporte Vectorial de Ranking (Ranking SVMs) entrenadas en pares de artículos de la misma fecha y medio. El modelo, que utiliza información mínima como contenido textual, fechas de publicación y estado de popularidad, predice con éxito los artículos populares e identifica palabras clave influyentes. Utilizando un conjunto de datos de diez medios importantes en inglés durante un año, el estudio logra precisiones que a menudo superan el 60%, superando los métodos de clasificación binaria. Esta investigación destaca la naturaleza dinámica del atractivo de las noticias y proporciona un marco robusto para predecir la popularidad de las noticias.

En [**?**], se introduce un Sistema Proactivo de Soporte de Decisiones Inteligente (IDSS) está diseñado para predecir la popularidad de los artículos de noticias en línea antes de su publicación. Utilizando un conjunto de datos de 39,000 artículos del sitio web Mashable, el IDSS aprovecha diversas características, incluyendo contenido digital, popularidad de noticias referenciadas, participación de palabras clave y análisis de sentimientos. El sistema emplea modelos de aprendizaje automático como Bosques Aleatorios (Random Forest), Adaptive Boosting y Máquinas de Soporte Vectorial (SVM) para una tarea de clasificación binaria. El modelo Random Forest alcanza un poder de discriminación del 73%, mientras que un módulo de optimización que utiliza búsqueda local por escalada estocástica mejora la probabilidad de popularidad estimada en 15 puntos porcentuales. El IDSS no solo predice, sino que también sugiere modificaciones en el contenido y la estructura del artículo para mejorar la popularidad esperada, demostrando ser una herramienta valiosa para los autores de noticias en línea.

Usando el mismo conjunto de datos, [10] evalúa diversas técnicas de aprendizaje automático para predecir la popularidad de los artículos de noticias en línea. Se obtuvieron conocimientos iniciales utilizando regresión lineal y logística, logrando la regresión logística una precisión del 66% al categorizar la variable objetivo en categorías binarias. Usando SVM, los autores enfrentaron inicialmente problemas de alto sesgo, pero mostraron mejoras marginales con núcleos más complejos, alcanzando una precisión del 55%. Sin embargo, el modelo de Random Forest emergió como el más efectivo, logrando una precisión del 70% con parámetros óptimos. Al aprovechar múltiples árboles de decisión y subconjuntos de características, Random Forest mitigó eficazmente la varianza, proporcionando las predicciones más precisas.

Finalmente, [11] explora la predicción de la popularidad de artículos de noticias utilizando el conjunto de datos de Mashable mediante diversas técnicas de aprendizaje automático. Se aplicaron métodos de selección de características como la selección univariada, la eliminación recursiva de características y el análisis de componentes principales para identificar las características más relevantes que influyen en la popularidad de los artículos. Evaluaron once modelos de clasificación, incluidos Naïve Bayes, regresión logística, árboles de decisión, redes neuronales, bosques aleatorios y máquinas de vectores de soporte. Entre estos, el método de potenciación del gradiente (gradient boosting) surgió como el modelo más eficaz, logrando una precisión del 79.7%. El estudio concluyó que los métodos en ensamble, en particular gradient boosting, son los que mejor funcionan para predecir la popularidad de los artículos de noticias.

Esta colección de investigaciones destaca las complejidades y avances en la predicción de la popularidad de las noticias en línea. La predicción temprana desde el inicio sigue siendo un desafío debido a las distribuciones de popularidad sesgadas y los sesgos de los modelos. Sin embargo, las técnicas sofisticadas de aprendizaje automático y las estrategias exhaustivas de preprocesamiento muestran prome-

sas para mejorar la precisión de la predicción. La investigación y desarrollo continuos de metodologías robustas son cruciales para futuros avances en este dominio.

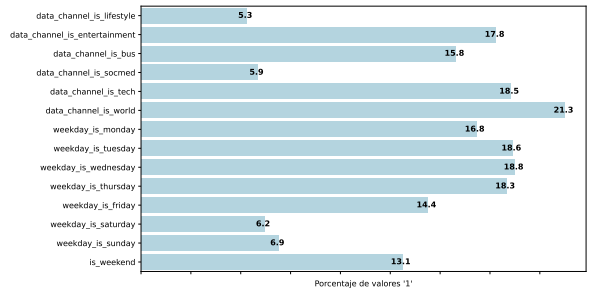
4. Diseño del experimento

Para este estudio, se utilizará el conjunto de datos elaborado en [**?**]. Cuenta con 59 características y una columna target. esta última contiene información de las veces que el artículo fue compartido (numérica), siendo esta la medición de popularidad. En general, el conjunto de datos permite un análisis exhaustivo de las características del contenido, la participación del usuario y la efectividad de varios tipos de contenido a lo largo de diferentes canales y en el tiempo. A continuación, se presenta una descripción más detallada.

4.1. Descripción del conjunto de datos (falta completar)

El conjunto de datos proporciona una visión general exhaustiva del contenido en línea, incorporando varias métricas y características relacionadas con atributos textuales, elementos multimedia e indicadores de participación. Captura atributos textuales clave, como el número de tokens en los títulos y el contenido, la proporción de tokens únicos y la longitud promedio de los tokens, ofreciendo información sobre la complejidad y estructura del contenido. El conjunto de datos también incluye métricas de palabras clave, que destacan la relevancia e importancia de las palabras clave en el contenido. Además, registra el número de hipervínculos y elementos multimedia como imágenes y videos, reflejando la riqueza y los aspectos multimedia de los artículos o publicaciones.

Además, el conjunto de datos clasifica el contenido en diferentes canales, como estilo de vida, entretenimiento, negocios, redes sociales, tecnología y noticias mundiales, proporcionando una clara clasificación de los tipos de contenido. Se incluye información temporal a través de variables que indican el día de la semana en que se publicó el contenido y si se publicó en un fin de semana. En estos casos, las variables han sido incluidas como binarias.



El conjunto de datos también presenta componentes de análisis semántico latente y puntuaciones de sentimiento, capturando el tono emocional y la objetividad del contenido. Las métricas de participación, como el número de veces que se comparte y las referencias a sí mismo, son cruciales para comprender el alcance e impacto del contenido. Todas estas son variables continuas.

4.2. Metodología

El presente trabajo se prioriza en el procesamiento de los datos y esencialmente antes del entrenamiento de los datos. Por tanto, se seguirá los siguientes pasos:

#1) Preparación de datos, este punto realizaremos la limpieza de datos, el manejo de valores faltantes, nulos y detección de los outlier.

2) Exploración y análisis de datos: en este paso realizaremos la reducción de la dimensionalidad utilizando la herramienta de análisis de principales componentes (PCA) y análisis de correlación de datos, ambos métodos nos ayudaran a visualizar y entender las relaciones y patrones de las variables.

PCA, toma encuenta la varianza ya incluye el analis de correlacion

3) Selección y entrenamiento del modelo; los datos clasificacion: 'RandomForest': RandomForestClassifier(), 'AdaBoost': AdaBoostClassifier(), 'LogisticRegression': LogisticRegression() GridSearchCV gradienBoots Redes neuronales (opcional)

considerando que los datos se acomodan para aplicar modelos de regresión, priorizaremos la metodología de , el cual es una técnica que se utiliza la mejor combinación de hiperparámetros. En este punto, además, optimizaremos los parámetros del modelo para mejorar su rendimiento.

4) En seguida realizaremos la validación y evaluación del modelo: en otras palabras, medir el rendimiento del modelo utilizando datos de prueba y métricas de evaluación. Aquiursasi y el area bajo la curva
5) Finalmente se describirá los resultados, interpretación y discusiones: en esta parte nos apoyaremos mediante figuras y gráficos estadísticos y correlacionaremos con anteriores resultados.

5. Experimentación y resultados

6. Referencias

7. Quarto

Quarto enables you to weave together content and executable code into a finished document. To learn more about Quarto see https://quarto.org.

8. Running Code

When you click the **Render** button a document will be generated that includes both content and the output of embedded code. You can embed code like this:

```
2
```

You can add options to executable code like this

```
4
```

The `echo: false` option disables the printing of code (only output is displayed). This is the default behavior.

When you use `echo: true` you will show both the code and the output.

```
print("Hello")
```

```
Hello
```

9. Using figures and equations

Using `fig-cap` allows you to specifiy a caption for a figure. Use it in code blocks where the last line prints a plot.

You can write Latex equiations:

$$E = mc^2$$

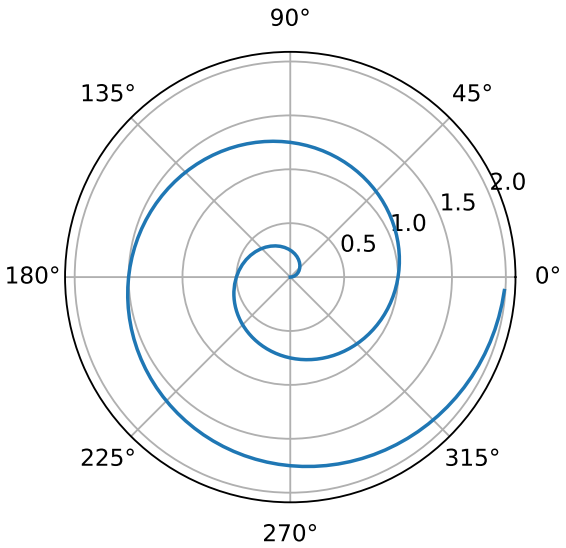


Figure 1: Radial plot

If the label attribute of a code chunk starts with `fig-` when it renders a plot, you can reference it. For example, here we reference Figure 1.

Be attentive as some things might need manual tweaking.

10. Using other formating

Multiple formatting options

1. You can use lists
2. They will numerate by themselves
3. No need to worry about counting

You are not restricted to numbered lists.

- first
- second
- third

10.1. Use of sub headers

Organize your document as you please. How much organization do you really need?

10.2. Another subheader

Does this looks nice to you? This document follows Markdown conventions.

10.2.1. A deeper level

However, is best that you don't go too deep because lower levels might not be fully supported in this template.

11. Citations

You can add citationsfor something some said at some point. Your references should be inside the `references.bib` file. Some people might have already researched this field [see 12, pp. 1-2].

References

[1] M. T. Uddin, M. J. A. Patwary, T. Ahsan, M. S. Alam, Predicting the popularity of online news from content metadata (2016) 1–5.

[2] A. Kaltenbrunner, V. Gómez, V. López, Description and prediction of slashdot activity (2007) 57–66.

[3] G. Szabo, B. A. Huberman, Predicting the popularity of online content, Communications of the ACM 53 (8) (2010) 80–88.

[4] J. G. Lee, S. Moon, K. Salamatian, Modeling and predicting the popularity of online contents with cox proportional hazard regression model, Neurocomputing 76 (1) (2012) 134–145.

[5] M. Ahmed, S. Spagna, F. Huici, S. Niccolini, A peek into the future: Predicting the evolution of popularity in user generated content, in: Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 607–616.

[6] A. Tatar, M. D. De Amorim, S. Fdida, P. Antoniadis, A survey on predicting the popularity of web content, Journal of Internet Services and Applications 5 (2014) 1–20.

[7] K. Fernandes, P. Vinagre, P. Cortez, A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News, Springer International Publishing, 2015, pp. 535–546. doi:10.1007/978-3-319-23485-4_53.
URL http://dx.doi.org/10.1007/978-3-319-23485-4_53

[8] I. Arapakis, B. B. Cambazoglu, M. Lalmas, On the Feasibility of Predicting News Popularity at Cold Start, Springer International Publishing, 2014, pp. 290–299. doi:10.1007/978-3-319-13734-6_21.
URL http://dx.doi.org/10.1007/978-3-319-13734-6_21

[9] E. Hensinger, I. Flaounas, N. Cristianini, Modelling and predicting news popularity, Pattern Analysis and Applications 16 (4) (2012) 623–635. doi:10.1007/s10044-012-0314-6.
URL http://dx.doi.org/10.1007/s10044-012-0314-6

[10] H. Ren, Q. Yang, Predicting and evaluating the popularity of online news, 2015.
URL https://api.semanticscholar.org/CorpusID:7149545

[11] A. Khan, G. Worah, M. Kothari, Y. H. Jadhav, A. V. Nimkar, News popularity prediction with ensemble methods of classification, 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (07 2018). doi:10.1109/icccnt.2018.8494095.
URL http://dx.doi.org/10.1109/ICCCNT.2018.8494095

[12] K. Fernandes, P. Vinagre, P. Cortez, P. Sernadela, Online News Popularity, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5NS3V (2015).