Análisis Predictivo de la Popularidad de Noticias Online con Machine Learning

David Gomez Agudelo, Santiago Trespalacios Bolivar, Juan Diego Duque Jimenez
Programa de Ingeniería de Sistemas
Universidad de Antioquia
Medellín, Colombia

***Abstract*—This project presents a predictive analysis about online news popularity based on supervised Machine Learning algorithms. The study employs the Online News Popularity dataset from Mashable, which contains over 39,000 articles and 58 features, to anticipate the number of times a news article will be shared before publication. The methodology includes data preprocessing, feature selection, and model training, aiming to support editorial decisions and enhance content visibility.**

Keywords: Machine Learning, Prediction, Popularity, News, Digital Media, Data Analysis

En la actualidad, el consumo de noticias ha cambiado completamente. Cada día se generan millones de noticias, artículos e información en noticieros, revistas y medios digitales, por ende, sobresalir a ojos de los lectores es una de las tareas más importantes y difíciles para los escritores.

Las técnicas de Machine Learning se han convertido en herramientas esenciales para lograr comprender por que una noticia o articulo es más popular o mejor valorada por los usuarios, permitiendo ademas predecir el comportamiento de las personas con respecto a las publicaciones y contribuir al éxito de las mismas. No se trata solo de escribir buenos textos, sino también de saber cómo, cuándo y dónde publicarlos para conseguir el mayor impacto en el público objetivo.

En este contexto, el presente proyecto tiene como objetivo desarrollar una solución a un problema mediante la aplicación de distintas técnicas de Machine Learning aprendidas en clase, utilizando una base de datos seleccionada como fuente principal de información y previamente autorizada para su uso. El proceso incluye la preparación y la exploración de los datos, la selección de las características mas relevantes, el entrenamiento y validación utilizando distintos modelos de aprendizaje automático.

I. Descripción del problema

I-A. Contexto del problema y utilidad de la solución

Los medios digitales enfrentan el desafío de llegar a la mayor cantidad de usuarios con sus contenidos, en un panorama donde cada lector selecciona minuciosamente que noticias consumir. Es por tal razón que, para los medios tanto de comunicación como de blogs de noticias les resulta bastante importante el no solo producir artículos que sean de calidad, sino también el maximizar su alcance, para que pueda llegar a un público mas extenso.

En este caso, se mide la popularidad de una noticia, que comúnmente se da por el numero de veces que la noticia es compartida por los medios digitales, resultando ser un indicador clave del impacto y de la viralidad del contenido.

Ante esta necesidad, el desarrollo de una solución basada en Machine Learning para predecir la popularidad de un artículo antes de la publicación del mismo tiene una utilidad práctica inmensa en el campo de los medios de comunicación y de medios digitales. Esto le permitiría a los editores realizar lo siguiente:

Optimizar el contenido: Realizar ajustes en lo que respecta al titular, la longitud del artículo o el uso de imágenes y videos para potenciar su viralidad por medio de las redes sociales.

Mejorar la estrategia de publicación: Decidir el mejor momento o canal de visualización (negocios, tecnología, entretenimiento, etc.) para publicar un artículo y así poder potenciar su visibilidad.

Asignar recursos de marketing: Identificar los artículos que podrían tener un alto potencial de popularidad para promocionarlos de una manera activa y lograr maximizar los ingresos de estos mismos por publicidad.

Llegando al punto de que, el utilizar el sistema predictivo se convierte en una herramienta que da un soporte a las personas con la decisión estratégica que se puede tomar al momento de publicar algún articulo o alguna noticia para competir en los medios digitales y tener un mayor alcance.

I-B. Composición de la base de datos

El dataset [1], es proveniente de la plataforma de noticias de internet llamada Mashable, y este contiene información sobre unos 39,644 artículos publicados durante un periodo de dos años. La base de datos está compuesta por un total de 61 atributos:

Número de muestras: 39,644.

Número de variables: 58 variables predictoras y 1 variable objetivo shares. Dos atributos adicionales url y timedelta son informativos pero no se usarán directamente en la predicción.

Significado de las variables: Las características se pueden agrupar en varias categorías:

- Atributos de contenido: Longitud del título n_tokens_title, longitud del contenido

n_tokens_content, número de enlaces, número de imágenes, etc.
- Atributos lingüísticos: Tasa de palabras únicas rate_unique_tokens, tasa de palabras no triviales rate_non_stop_words, polaridad del sentimiento global_sentiment_polarity, etc.
- Metadatos: Canal de la noticia (entretenimiento, negocios, tecnología, etc.), día de la semana de publicación.
- Popularidad de palabras clave: Medidas promedio, mínima y máxima de la popularidad de las palabras clave asociadas al artículo.

Variable objetivo: shares, un valor entero que representa el número de veces que el artículo fue compartido. Es una variable con una distribución muy sesgada a la derecha, donde la mayoría de los artículos son pocos compartidos y unos pocos se vuelven ”virales”.

Datos faltantes: El dataset está preprocesado y no contiene valores faltantes, por lo que no se requiere una estrategia de imputación de datos.

Codificación de variables: Las variables categóricas ya han sido codificadas. Los canales de noticias data_channel_is_* y los días de la semana weekday_is_* están representados mediante codificación one-hot. La variable is_weekend es binaria. El resto de las variables son numéricas (enteras o de punto de flotante).

La profundización y la descripción detallada de estas variables se puede encontrar en la Tabla I, donde se especifican las características junto con sus respectivas descripciones, tipos y grupos funcionales que pueden servir como guía para la solución del problema.

I-C. Paradigma de Aprendizaje

Para este proyecto, se decidió abordar el problema tanto desde el enfoque de regresión como de clasificación, con un énfasis principal en la clasificación binaria. Esta elección se fundamenta en que la distribución de la variable objetivo (shares) presenta una marcada asimetría hacia la derecha, lo que dificulta la predicción precisa del número de veces que una noticia es compartida. En contraste, formular el problema como una tarea de clasificación definiendo un umbral que distinga entre noticias populares y no populares ofrece una interpretación más práctica y coherente con los objetivos del estudio. No obstante, también resulta de interés explorar el paradigma de regresión de manera complementaria, con el fin de analizar el comportamiento continuo de la variable y comparar el desempeño de ambos enfoques aunque este se toma desde un punto de vista mas exploratorio y menos central dentro del estudio. Con este enfoque se busca analizar el comportamiento de la variable shares y como esta se ve afectada por pequeños cambios dentro de las variables predictoras. Esto permite conocer la sensibilidad de los modelos ante la dispersión de los datos y determinar cual modelo ofrece una visión más detallada del problema de popularidad.

II. Estado del arte

Para comprender las soluciones existentes al problema de predecir la popularidad de noticias en línea, se ha realizado una revisión de fuentes especializadas, incluyendo trabajos que utilizan el mismo dataset.

A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News

Este es el artículo original que introduce el dataset. Los autores presentan un Sistema Inteligente de Soporte a la Decisión diseñado para maximizar la popularidad de los artículos antes de ser publicados. Ademas se plantea un sistema proactivo el cual sugiere cambios fáciles de implementar para mejorar el pronóstico de las noticias.

El Paradigma de aprendizaje se aborda como una tarea de clasificación, para determinar si un artículo será “popular” o no, utilizando un umbral de compartidos, también se explora como un problema de regresión para predecir el número exacto de compartidos, en el proceso se utilizaron cinco modelos de clasificación: Random Forest (RF), Adaptive Boosting (AdaBoost), Máquinas de Vectores de Soporte (SVM), K-Nearest Neighbors (KNN) y Naïve Bayes.

Para la metodología de validación se emplearon ventanas deslizantes (rolling windows). Este enfoque simula un escenario real donde el modelo se reentrena periódicamente con datos nuevos, para asi poder evaluar las métricas y teniendo que la métrica principal fue el Área Bajo la Curva ROC (AUC), también se reportaron Exactitud (Accuracy), Precisión y Recall. Para la tarea de regresión, se utilizó el Error Absoluto Medio (MAE), con esto se obtuvo como resultado que los modelos de ensamble como Random Forest y Gradient Boosting mostraron el mejor rendimiento. En la tarea de clasificación, Random Forest alcanzó una exactitud cercana al 67 %.[2]

Prediction & evaluation of online news popularity using machine intelligence

En este trabajo se utiliza un paradigma de aprendizaje supervisado, específicamente un problema de clasificación binaria. El objetivo es clasificar cada noticia basándose en si supera un umbral predefinido de compartidos.

El Paradigma de aprendizaje se enmarca como un problema de clasificación, haciendo uso de una estrategia central que consiste en aplicar Análisis Discriminante Lineal y posteriormente poner un énfasis especial en los algoritmos de boosting como AdaBoost y LPBoost, realizando una comparación con el popular Random Forest. Además utiliza una metodología de validación conocida como Hold-out para validar el modelo, evaluando su sensibilidad a la partición de los datos mediante siete diferentes proporciones de división entre entrenamiento y prueba.

A pesar de que se reportaron métricas estándar (Exactitud, Precisión, Recall), el F-measure es la métrica clave donde el modelo propuesto demuestra su superioridad. Dando como resultado que la combinación de LDA con AdaBoost superó

al resto de modelos, alcanzando un F-measure del 73 % y una exactitud del 69 % [3]

Hyper Parameter Optimization using Genetic Algorithm on Machine Learning Methods for Online News Popularity Prediction

En este articulo, se propone el mejorar la predicción de la popularidad de las noticias mediante la optimización de los modelos de Machine Learning utilizados, esto por medio de algoritmos genéticos en lugar del Grid Search, reduciendo el tiempo y manteniendo la misma precisión.

Fue abordado como problema de clasificación evaluando los modelos: Support Vector Machine (SVM), Random Forest, Adaptive Boosting, K-Nearest Neighbors y Naive Bayes. Ademas se optimizaron los hiperparámetros usando Genetic Algorithm. Por otro parte, para validar se hizo una división del dataset, una parte de entrenamiento equivalente a un (70 %) y otra de prueba equivalente a un (30 %).

El resultado es evaluado usando la exactitud (accuracy) y el tiempo requerido para buscar los hiperparametros correctos (tiempo computacional), siendo este el ultimo el mas importante. Logrando que la utilización del Algoritmo Genético redujera de una manera significativa el tiempo de cómputo manteniendo una precisión similar a la del Grid Search. En este caso, el mejor modelo fue Random Forest, con una exactitud de un 67 %.[4]

Predicting the Popularity of Online News Using Classification Methods with Feature Filtering Techniques

En este articulo, se analizan diferentes métodos de clasificación para predecir la popularidad de las noticias en linea, buscando identificar cual es el modelo mas preciso y ademas las características que son mas influyentes.

Se aborda como problema de clasificación utilizando los siguientes modelos: Random Forest (RF), Bayes Net (BN), Logistic Function, SimpleCart y C4.5. También se aplicaron técnicas de filtrado de características (InfoGain, Chi-Squared, Correlation, Gain Ratio y OneR) para sacar las 10 mejores. Ademas de aplicarse un Fold Cross validation de 10 pliegues para todos los modelos evaluados.

Con el fin de medir su efectividad se utilizo: Accuracy, Precision, Recall, F1 y valores AUC, con los cuales se logro elegir el mejor modelo para solucionar el problema. Dando como resultado que el modelo mas efectivo para el problema fue Random Forest, alcanzando los mejores resultados en cada una de las métricas de evaluación. [5]

En conclusión, la literatura sugiere que los modelos basados en ensambles de árboles de decisión (Random Forest, Gradient Boosting, XGBoost, ADAboost) son los más prometedores para este dataset. Además, la selección de características y la transformación del problema a clasificación son estrategias comunes y efectivas.

REFERENCIAS

[1] K. Fernandes, P. Vinagre, and P. Cortez, "Online News Popularity Data Set," UCI Machine Learning Repository, 2015, doi: 10.24432/C5G01C. [Online]. Available: https://archive.ics.uci.edu/dataset/332/online+news+popularity

[2] K. Fernandes, P. Vinagre, and P. Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News," *University of Minho, Department of Information Systems*, 2015. [Online]. Available: https://repositorium.uminho.pt/server/api/core/bitstreams/06251279-8d1a-46f8-bbb3-92697a2b3eae/content

[3] D. Deshpande, "Prediction & evaluation of online news popularity using machine intelligence," in *Proc. 2017 Int. Conf. on Computing, Communication, Control and Automation (ICCUBEA)*, 2017, pp. 1–6, doi: 10.1109/ICCUBEA.2017.8463790. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8463790"'

[4] A. S. Wicaksono and A. A. Supianto, "Hyper Parameter Optimization using Genetic Algorithm on Machine Learning Methods for Online News Popularity Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 12, pp. 263–267, 2018, doi: 10.14569/IJACSA.2018.091238. Online]. Available: https://www.proquest.com/openview/23c86ce1492c42f05d5f554186843ef7/1?pq-origsite=gscholar&cbl=5444811

[5] R. Obiedat, "Predicting the Popularity of Online News Using Classification Methods with Feature Filtering Techniques," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 8, pp. 1163–1172, Apr. 2020. [Online]. Available: https://www.jatit.org/volumes/Vol98No8/4Vol98No8.pdf

Grupo Funcional	Característica	Descripción	Tipo
Contenido y Estructura	n_tokens_title	Número de palabras en el titular.	Numérica (Discreta)
	n_tokens_content	Número de palabras en el cuerpo del artículo.	Numérica (Discreta)
	num_hrefs	Número total de enlaces (hipervínculos).	Numérica (Discreta)
	num_self_hrefs	Número de enlaces internos a otros artículos de Mashable.	Numérica (Discreta)
	num_imgs	Número de imágenes.	Numérica (Discreta)
	num_videos	Número de videos.	Numérica (Discreta)
	average_token_length	Longitud promedio de las palabras en el contenido.	Numérica (Continua)
	num_keywords	Número de palabras clave en los metadatos del artículo.	Numérica (Discreta)
Lingüísticos	n_unique_tokens	Número de palabras únicas en el texto.	Numérica (Continua)
	n_non_stop_words	Número de palabras con contenido (no ”stop words”).	Numérica (Continua)
	n_non_stop_unique_tokens	Número de palabras únicas con contenido.	Numérica (Continua)
	rate_positive_words	Tasa de palabras positivas entre palabras no neutrales.	Numérica (Continua)
	rate_negative_words	Tasa de palabras negativas entre palabras no neutrales.	Numérica (Continua)
Metadatos Categóricos	data_channel_is_*	6 variables binarias que indican el canal del artículo (lifestyle, tech, business, social media, entertainment y world).	Binaria
	weekday_is_*	7 variables binarias que indican el día de la semana de publicación. (monday, etc.)	Binaria
	is_weekend	1 si el artículo fue publicado en fin de semana, 0 si no.	Binaria
Popularidad de KW	kw_min_min, kw_max_min, kw_avg_min	Popularidad (mínima, máxima y promedio) de la palabra clave menos popular.	Numérica (Continua)
	kw_min_max, kw_max_max, kw_avg_max	Popularidad (mínima, máxima y promedio) de la palabra clave más popular.	Numérica (Continua)
	kw_min_avg, kw_max_avg, kw_avg_avg	Popularidad promedio (mínima, máxima y promedio) de todas las palabras clave.	Numérica (Continua)
Análisis de Sentimiento	global_subjectivity	Subjetividad del texto (0: objetivo, 1: subjetivo).	Numérica (Continua)
	global_sentiment_polarity	Polaridad del sentimiento del texto (-1: negativo, 1: positivo).	Numérica (Continua)
	global_rate_positive_words	Tasa de palabras positivas en el texto.	Numérica (Continua)
	global_rate_negative_words	Tasa de palabras negativas en el texto.	Numérica (Continua)
	avg_positive_polarity, min_positive_polarity, max_positive_polarity	Polaridad de las palabras positivas (promedio, mínima y máxima).	Numérica (Continua)
	avg_negative_polarity, min_negative_polarity, max_negative_polarity	Polaridad de las palabras negativas (promedio, mínima y máxima).	Numérica (Continua)
	title_subjectivity	Subjetividad del titular.	Numérica (Continua)
	title_sentiment_polarity	Polaridad del sentimiento del titular.	Numérica (Continua)
	abs_title_subjectivity	Distancia a la neutralidad de la subjetividad del titular.	Numérica (Continua)
	abs_title_sentiment_polarity	Distancia a la neutralidad de la polaridad del titular.	Numérica (Continua)
Tópicos LDA	LDA_00 - LDA_04	5 variables que indican la proporción de pertenencia a un tópico generado por LDA.	Numérica (Continua)
Auto-Referencia	self_reference_min_shares, self_reference_max_shares, self_reference_avg_shares	Número de veces que se compartieron los artículos de Mashable referenciados (mínimo, máximo y promedio).	Numérica (Continua)

Tabla I: Descripción Detallada de las Características Predictoras.