Alberto Bendayan, Theo Shlamovitz, Theo Stanfield

Detección automática de titulares clickbait en noticias digitales

Resumen

El objetivo de este trabajo es analizar titulares de noticias para identificar patrones asociados al fenómeno del *clickbait* y entrenar modelos de procesamiento de lenguaje natural que permitan distinguir entre titulares engañosos y titulares informativos. Para ello, trabajaremos con un dataset público que contiene aproximadamente 32.000 titulares de noticias en inglés, categorizados como clickbait o no clickbait.

Además de ser un fenómeno común en medios digitales y redes sociales, el clickbait tiene un impacto directo en la calidad de la información que consumen los usuarios y en la confianza hacia los medios de comunicación. Poder detectar automáticamente este tipo de titulares resulta valioso tanto para plataformas de noticias como para investigadores interesados en estudiar la manipulación del lenguaje en entornos digitales. Asimismo, desarrollar modelos que reconozcan patrones lingüísticos de clickbait aporta a la comprensión de cómo se diseña el contenido persuasivo y cómo este influye en el comportamiento de los lectores.

A lo largo del proyecto realizaremos un análisis exploratorio de las características léxicas y estructurales de los titulares, y luego desarrollaremos experimentos de clasificación utilizando representaciones de texto tradicionales (bag of words, TF-IDF) y modelos basados en embeddings. Nuestro objetivo es evaluar qué tipo de representación y modelo logra un mejor rendimiento en la detección de clickbait y qué patrones lingüísticos se asocian más fuertemente con esta práctica.

Datos

El dataset proviene de la plataforma Kaggle: Clickbait Dataset y contiene un total de 32.000 titulares de noticias en inglés, organizados en dos clases:

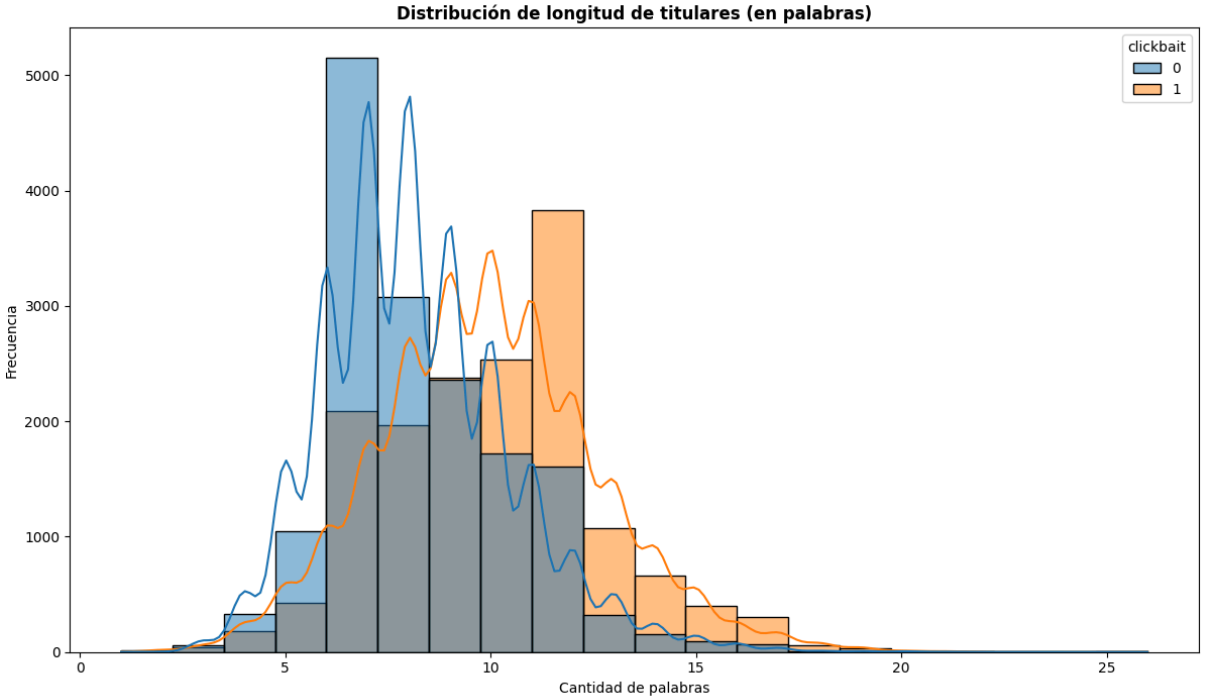
- **Clickbait**: titulares engañosos diseñados para inducir al usuario a hacer clic.
- **No clickbait**: titulares informativos o neutrales.

El archivo se encuentra en formato CSV, con dos columnas principales:

- **`headline`**: texto del titular (string).
- **`clickbait`**: etiqueta binaria (1 = clickbait, 0 = no clickbait).
    - En el dataset, los titulares se etiquetan según la fuente de la que provienen. Si pertenecen a una fuente confiable, se les asigna la etiqueta 0, ya que el lector puede confiar en que la información será veraz independientemente del enlace que siga. En cambio, los titulares procedentes de fuentes no confiables suelen emplear clickbait para atraer al usuario, buscando que haga clic sin garantizar que el contenido sea verdadero o realmente relevante.

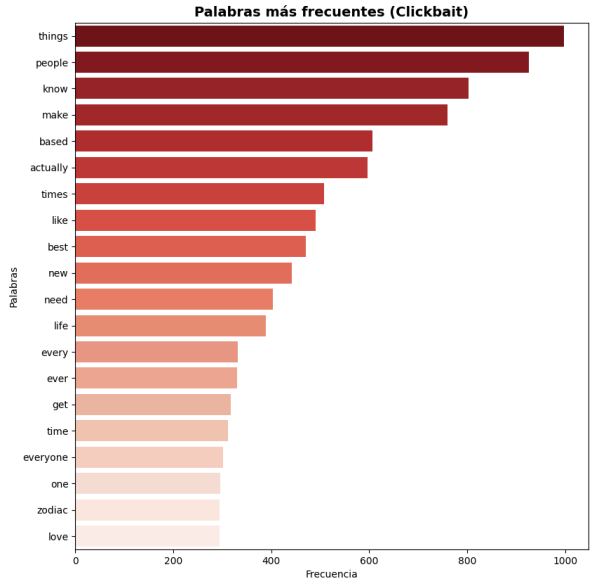
La distribución de clases es relativamente balanceada (aprox. 16k por clase), lo que facilita el entrenamiento supervisado sin necesidad de técnicas de oversampling o undersampling.

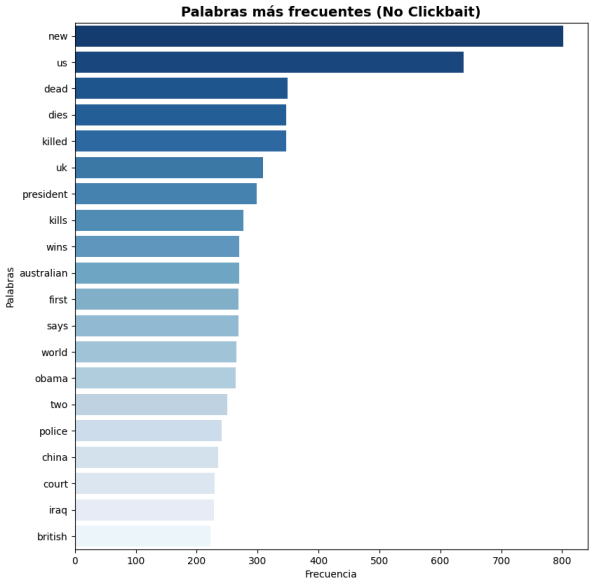
Análisis exploratorio



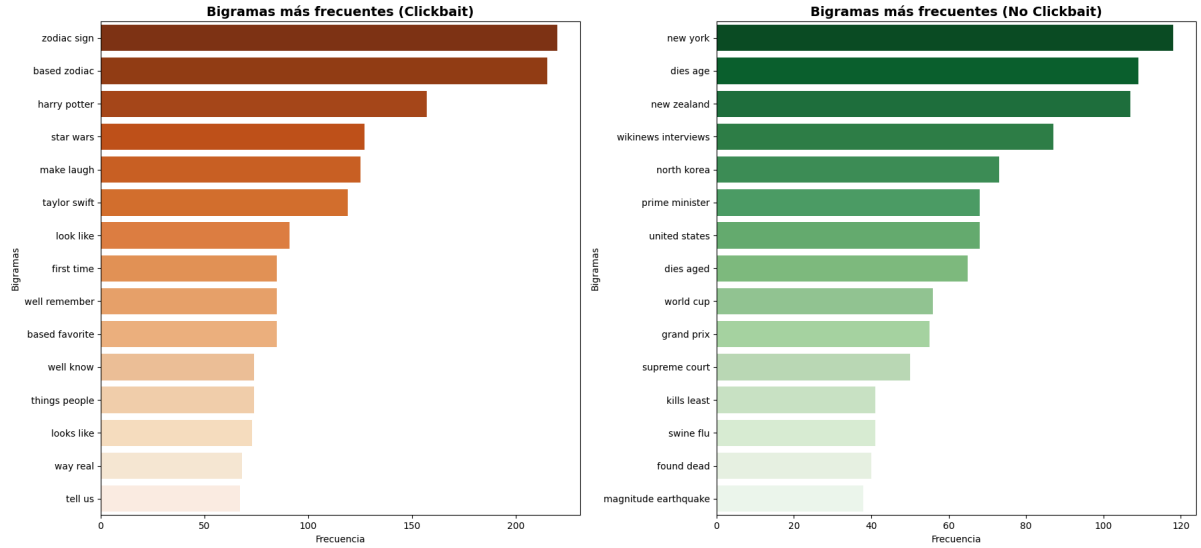
El conjunto contiene 32.000 titulares balanceados (50% clickbait, 50% no clickbait).

Los titulares clickbait presentan una longitud promedio de entre 11 y 13 palabras, mientras que los no clickbait promedian más bajo entre 5 y 7 palabras.





Entre las palabras más frecuentes del clickbait aparecen *things, people, know*, mientras que en los titulares no clickbait aparecen como palabras más frecuentes *new, us, dead, dies,* etc.



En el análisis de bigramas, se puede ver que en los titulares clickbait, la frecuencia más alta es para bigramas como *zodiac sign*, *based zodiac* y *harry potter*. Mientras que en los no clickbait, frecuentan mayormente bigramas como *new york*, *dies age* y *new zealand*.

Propuesta de análisis

A partir del análisis exploratorio, proponemos los siguientes experimentos y objetivos:

1. **Clasificación binaria de titulares:**
    - Entrenar modelo clásico de Regresión Logística con BoW o TF-IDF.
    - Probar embeddings más avanzados (Word2Vec, GloVe, BERT).
    - Evaluar rendimiento de ambos modelos con `accuracy`, `precision`, `recall` y `F1-score`.
    - Comparar representaciones para determinar la más efectiva.
2. **Interpretabilidad de los modelos:**
    - Extraer los features más importantes de modelos interpretables y relacionarlos con patrones lingüísticos del análisis exploratorio.
    - Esto permite explicar qué hace que un titular sea clickbait.
3. **Reflexión sobre el fenómeno clickbait:**
    - Usar los resultados para generar hipótesis sobre estrategias comunes de clickbait: uso de pronombres en segunda persona, verbos imperativos, titulares largos, etc.
    - Contribuye a comprender cómo los titulares buscan atraer la atención del lector.

Experimentos

En esta sección se detalla la metodología experimental diseñada para cumplir con los objetivos del proyecto. Se definen tres experimentos principales: el primero establece una línea base utilizando un modelo de clasificación tradicional, el segundo explora el rendimiento de modelos avanzados basados en embeddings, y el tercero se enfoca en un análisis lingüístico para identificar los patrones característicos del clickbait.

Preparación del Entorno y Preprocesamiento de Datos

Antes de la fase de modelado, es fundamental preparar un entorno de trabajo robusto y estandarizar el conjunto de datos para garantizar la reproducibilidad y la validez de los resultados.

Se realizará un preprocesamiento del texto de los 32.000 titulares.

Limpieza y Normalización: Se aplicará un pipeline de preprocesamiento estándar a cada titular, que incluirá:

1. Conversión a minúsculas para unificar el texto.
2. Tokenización para dividir los titulares en palabras individuales.
3. Eliminación de stopwords (palabras comunes como "the", "is", "in") utilizando la lista del paquete NLTK en inglés, ya que estas no suelen aportar valor semántico para la clasificación.
4. Se evaluará la aplicación de técnicas de stemming o lematización para reducir las palabras a su raíz y consolidar el vocabulario.

Partición de Datos: El dataset se dividirá en tres conjuntos: 70% para entrenamiento, 15% para validación y 15% para prueba (test). Se utilizará una partición estratificada para mantener la proporción 50/50 de titulares clickbait y no clickbait en cada conjunto, asegurando que los modelos se entrenen y evalúen sobre distribuciones de datos representativas. El conjunto de validación se usará para el ajuste de hiperparámetros, mientras que el de prueba se reservará exclusivamente para la evaluación final del rendimiento de los modelos.

Un preprocesamiento es crucial para reducir el ruido y la dimensionalidad del vocabulario, lo que permite que los modelos se centren en las características más informativas. Una partición de datos bien definida previene el overfitting y garantiza que la evaluación del rendimiento sea una estimación fiable de cómo se comportarán los modelos ante datos nuevos y no vistos.

Experimento 1: Clasificación con Modelo Clásico (Línea Base)

Este primer experimento tiene como objetivo establecer un punto de referencia (baseline) del rendimiento utilizando un algoritmo de aprendizaje automático tradicional y representaciones de texto clásicas.

¿Qué experimento voy a hacer?

En este experimento, se entrenarán y evaluarán un modelo de Regresión Logística de clasificación supervisada

¿Cómo lo voy a hacer?

Pasos a realizar:

1. Representación de Texto: Se utilizarán dos esquemas de vectorización:
    a. Bag of Words (BoW): Representará los titulares basándose en el recuento de la frecuencia de cada palabra.
    b. TF-IDF (Term Frequency-Inverse Document Frequency): Ponderará la importancia de las palabras no solo por su frecuencia en un titular, sino también por su rareza en todo el corpus.
2. Entrenamiento y Optimización: El modelo se entrenará con ambas representaciones (BoW y TF-IDF). Se utilizará el conjunto de validación para optimizar los hiperparámetros clave, como el parámetro de regularización C para la Regresión Logística
3. Evaluación: Las métricas de evaluación serán: Accuracy, Precision, Recall y F1-Score. Además, se generará una matriz de confusión para el modelo, lo que permitirá un análisis detallado de los errores (falsos positivos y falsos negativos).

¿Por qué lo voy a hacer?

El objetivo de este experimento es cuantificar la eficacia del enfoque tradicional, que es computacionalmente eficiente y altamente interpretable. Los resultados servirán como una sólida línea base para determinar si los modelos más complejos y computacionalmente costosos del Experimento 2 ofrecen una mejora significativa que justifique su uso.

Experimento 2: Clasificación con Modelos Basados en Embeddings

El objetivo de este experimento es evaluar si las representaciones vectoriales que capturan el significado semántico de las palabras pueden mejorar el rendimiento de la clasificación, validando la hipótesis de que los embeddings superarán a las representaciones tradicionales.

¿Qué experimento voy a hacer?

Se implementarán y evaluarán modelos que utilicen embeddings de palabras preentrenados y modelos de lenguaje basados en Transformers.

¿Cómo lo voy a hacer?

1. Embeddings Preentrenados (GloVe):
    a. Se cargarán vectores GloVe preentrenados.
    b. Cada titular se convertirá en un único vector promediando los vectores de embedding de todas las palabras que lo componen.
    c. Estos vectores resultantes se utilizarán como entrada para un clasificador simple, como la Regresión Logística, que será entrenado y evaluado con el mismo protocolo (partición de datos y métricas) que en el Experimento 1.
2. Modelo Basado en Transformers (BERT):
    a. Se utilizará un modelo preentrenado como bert-base-uncased de la librería Hugging Face.

b. Se realizará un proceso de ajuste fino (fine-tuning), donde se añadirá una capa de clasificación sobre la arquitectura de BERT y se reentrenará el modelo completo sobre nuestro conjunto de datos de titulares. Este enfoque permite que el modelo adapte su comprensión contextual del lenguaje a la tarea específica de detección de clickbait.
c. El modelo ajustado se evaluará en el conjunto de prueba utilizando las mismas métricas (Accuracy, Precision, Recall, F1-Score).

¿Por qué lo voy a hacer?

A diferencia de BoW y TF-IDF, los embeddings como GloVe capturan relaciones semánticas (por ejemplo, "rey" está cerca de "reina"). BERT va un paso más allá, interpretando las palabras en su contexto específico dentro de una oración. Este experimento es clave para determinar si esta comprensión semántica y contextual del lenguaje es decisiva para distinguir la sutileza del lenguaje persuasivo del clickbait.

Experimento 3: Análisis Lingüístico de Patrones de Clickbait

Más allá de la predicción, este experimento busca comprender qué características lingüísticas hacen que un titular sea clasificado como clickbait.

¿Qué experimento voy a hacer?

Se realizará un análisis cuantitativo de las características lingüísticas y se utilizarán técnicas de interpretabilidad de modelos.

¿Cómo lo voy a hacer?

- Se extraerán los coeficientes del modelo de Regresión Logística entrenado en el Experimento 1. Las palabras con los coeficientes positivos más altos serán identificadas como los predictores más fuertes de clickbait, mientras que aquellas con los coeficientes negativos más altos serán los indicadores de titulares informativos. Esto permitirá validar cuantitativamente los hallazgos del análisis exploratorio inicial.

¿Por qué lo voy a hacer?

Este análisis es fundamental para ir más allá del modelo de clasificación. Al identificar patrones lingüísticos concretos, no solo validaremos nuestras hipótesis, sino que también generaremos conocimiento sobre cómo se construye el lenguaje persuasivo y engañoso en los medios digitales.