Recomendación de anuncios

Jorge Almonacid, Rubén Genillo, Humberto Pérez de la Blanca, Cristina Suárez
Universidad San Pablo CEU

Indice

Resumen

Este proyecto tiene como objetivo principal mejorar la efectividad de la publicidad digital mediante el análisis de datos en redes sociales, específicamente en la plataforma Twitter. Se emplearán técnicas avanzadas de inteligencia artificial y minería de datos para extraer, analizar y comprender los comentarios y opiniones de los usuarios en tiempo real.

El proyecto se divide en varias etapas clave, que incluyen la extracción de datos de Twitter, el análisis de sentimientos para comprender las actitudes de los usuarios sobre ciertos temas, el modelado de temas para identificar el foco principal del cual hablan los usuarios y la asignación de anuncios basada en los intereses y comportamientos de los usuarios.

A través de este enfoque, se busca ofrecer recomendaciones de anuncios más relevantes y personalizadas, maximizando así el impacto de la publicidad dirigida y mejorando la experiencia del usuario en línea. Además, el proyecto explorará las implicaciones estratégicas y las oportunidades potenciales que este enfoque brinda en el contexto del marketing digital y más allá.

En resumen, este proyecto representa un paso significativo hacia una publicidad más inteligente y centrada en el usuario, aprovechando el vasto conjunto de datos disponibles en las redes sociales para informar decisiones comerciales más acertadas y estratégicas.

Palabras clave: marketing - anuncios - sentiment analysis - topic modeling

0.1 Introducción

En el contexto del creciente panorama digital, la extracción y análisis de datos de redes sociales se ha convertido en un componente crucial para las estrategias de marketing y toma de decisiones empresariales. En este proyecto, nos enfocamos en la aplicación de técnicas de inteligencia artificial para la extracción, análisis y asignación de anuncios basados en datos obtenidos de la plataforma Twitter.

0.2 Brainstorming

Durante el brainstorming, se exploraron diversas áreas de interés, desde la seguridad informática con el análisis de correos maliciosos y el desarrollo de un detector de malware, hasta la aplicación de la teoría de juegos para identificar posibles trampas, lo que representa un enfoque innovador con amplias implicaciones. Además, se abordó el campo de la inteligencia artificial con la creación de un modelo de generación de texto, lo que refleja un interés en tecnologías avanzadas y su impacto potencial.

Se seleccionó una idea centrada en el ámbito del análisis de datos en redes sociales para la recomendación de anuncios. Esta decisión se llevó a cabo debido a la relevancia y actualidad de este tema en el contexto del marketing digital y la publicidad en línea. La capacidad de aprovechar la información generada por los usuarios en plataformas como Twitter para ofrecer recomendaciones de anuncios más efectivas y personalizadas es un enfoque estratégico que destaca la importancia de la inteligencia de datos en el mundo empresarial actual.

0.3 Necesidades:

Al haber seleccionado la idea centrada en el análisis de datos en redes sociales para la recomendación de anuncios, se pone de manifiesto la importancia de estar actualizados, ser precisos y comprender las respuestas de los clientes. Estas necesidades reflejan la relevancia de mantenerse al tanto de las tendencias cambiantes en línea, la precisión en la interpretación de datos y la comprensión profunda de las interacciones de los clientes en entornos digitales. Este enfoque resalta la importancia estratégica del marketing en un mundo cada vez más impulsado por la información y las interacciones en línea.

0.4 Objetivo

El objetivo de la extracción de datos y opiniones en redes sociales de usuarios para la recomendación de anuncios se fundamenta en la necesidad de acceder a información pública, en tiempo real y constantemente actualizada. Este enfoque estratégico busca aprovechar la riqueza de datos disponibles en entornos digitales públicos, garantizar la relevancia y actualidad de la información recopilada, y capturar las tendencias y opiniones más recientes. Al hacerlo, se busca informar decisiones de marketing con datos dinámicos y significativos, alineados con las demandas cambiantes del mercado y las interacciones de los usuarios en línea.

Además, este enfoque puede tener usos alternativos significativos que van más allá de la recomendación de anuncios. Por ejemplo, la extracción de datos y opiniones en redes sociales puede ser invaluable para la investigación de clientes, permitiendo una comprensión más profunda de sus necesidades, preferencias y comportamientos en línea. Asimismo, la identificación de *Brand Ambassadors*, es decir, usuarios influyentes que puedan promover de manera auténtica una marca, es otro beneficio clave de esta estrategia. Además, el manejo de reputación en línea se ve reforzado por la capacidad de monitorear y responder de manera proactiva a las interacciones en redes sociales, lo que puede impactar positivamente la percepción pública de una empresa o producto. Estos son solo algunos ejemplos de cómo este enfoque puede ser aprovechado para una variedad de aplicaciones estratégicas más allá de la publicidad y recomendaciones comerciales.

0.5 Tareas

La tarea se puede dividir en los siguientes temas:

0.5.0.1 Extracción de Datos:

En esta fase, se enfocará en la extracción de datos clave de la plataforma Twitter. Se buscará obtener información pública de los usuarios, como nombres, biografías, publicaciones, día y hora de publicación, comentarios, interacciones como me gusta, seguidores, seguidos y hashtags utilizados. Esta información será recopilada a través de una API que permitirá acceder a estos datos de manera estructurada. Además, se profundizará en la obtención de información personal más detallada, como edad, género, localización, ocupación e intereses de los usuarios para enriquecer el análisis.

0.5.0.2 Filtrar posts relevantes:

Una vez recopilados los datos, el siguiente paso será filtrar y clasificar los posts relevantes. Para lograr esto, se implementará un análisis de temas avanzado. Se utilizarán algoritmos especializados como el Análisis Semántico Latente (LSA) y la Asignación Latente de Dirichlet (LDA) para identificar patrones y agrupar los posts según los temas principales que abordan. Este proceso permitirá segmentar la información de manera efectiva y comprender mejor las discusiones que tienen lugar en la plataforma.

0.5.0.3 Análisis de sentimiento:

Otro aspecto fundamental de esta tarea es el análisis de sentimiento. Aquí se enfocará en determinar la polaridad de las opiniones expresadas en los posts recopilados. Para ello, se emplearán lexicons de sentimiento que ayudarán a clasificar las opiniones como positivas, negativas o neutras. Además, se explorará un método alternativo que involucra el uso de LMQL o langchain con Large Language Models para mejorar la precisión y profundidad del análisis de sentimiento.

Se estudiarán detenidamente los datos obtenidos de Twitter para llevar a cabo estas tareas con rigor y obtener información significativa que puedan impulsar estrategias efectivas en marketing digital y toma de decisiones empresariales.

0.5.0.4 Topic modeling:

Finalmente aplicamos el topic modeling para categorizar automáticamente los temas principales de los datos extraídos previamente, los cuales se guardarán en una base de datos. Una vez tenemos los temas, se procederá a la asignación de anuncios, los cuales serán clasificados mediante un resumen por el mismo algoritmo de topic modeling y serán asignados a la gente con las que coincidan los temas.

0.6 Diagramas

Tuvimos que realizar 2 diagramas, por un lado el diagrama de desarrollo, que muestra que actividades se van a realizar en el proyecto. Y por otro lado el diagrama de Gantt, que se encarga de distribuir las tareas y actividades en una línea temporal.

Figura 1: Diagrama de desarrollo

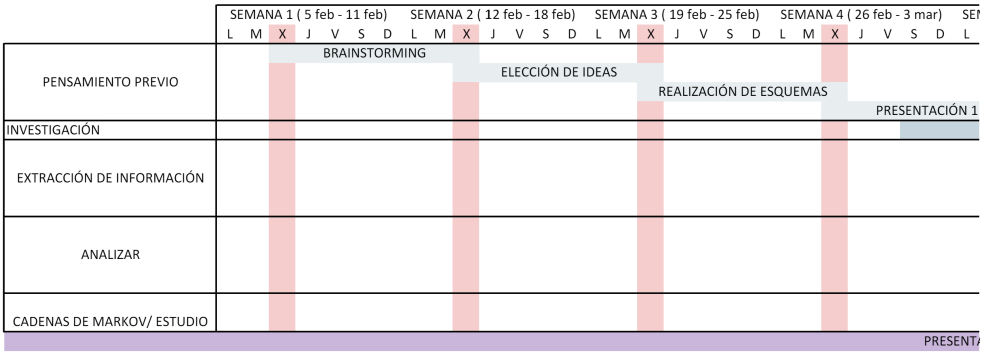

Figura 2: Gant parte 1

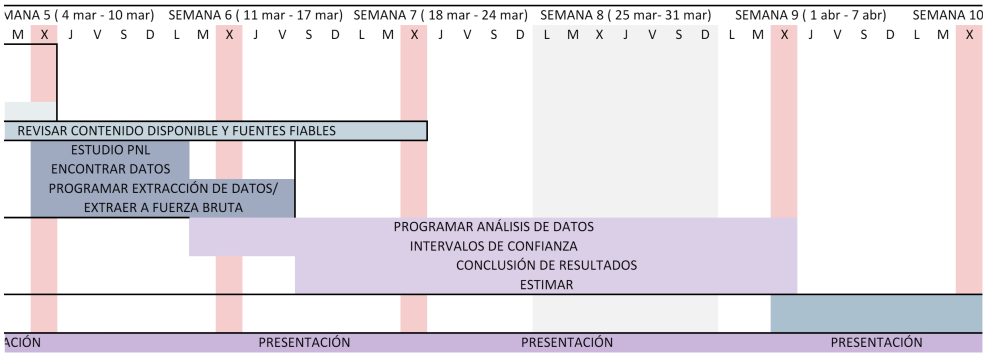

Figura 3: Gant parte 2

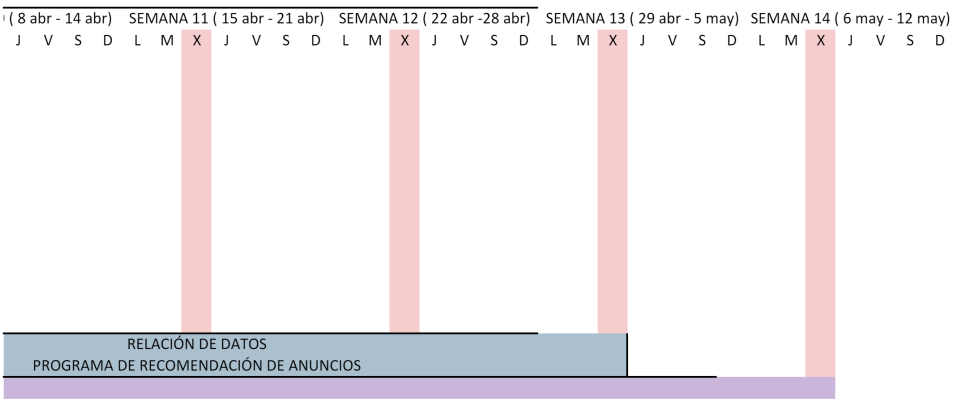

Figura 4: Gantt parte 3

0.7 Proyecto

0.7.0.1 Extracción de Datos

La extracción de datos es una parte fundamental de la minería de datos en redes sociales. Para este proyecto, se pensó en extraer datos de Twitter, una de las plataformas más populares para la interacción en línea. Originalmente, contemplamos utilizar la API de Twitter para acceder a datos como perfiles de usuario, publicaciones y mucho más. Sin embargo, cambios recientes en la política de acceso de la API nos llevaron a adoptar técnicas de web scraping como alternativa eficaz.

El web scraping implica la extracción automatizada de datos del HTML de páginas web. En nuestro caso, extraemos información pública de tweets, incluyendo el contenido completo, número de retweets, likes y los nombres de usuario. Esta información se almacena en formatos estructurados como CSV para su análisis posterior.

Es importante destacar que nuestro enfoque se centra exclusivamente en información pública, sin infringir la privacidad ni las normativas legales, asegurando que nuestra recopilación de datos se utiliza únicamente con fines investigativos y académicos.

Además, optamos por analizar los textos en inglés para facilitar la integración con las redes neuronales empleadas en el análisis de sentimientos y el topic modeling.

```
# Importamos las librerías necesarias
from twikit import Client
import pandas as pd

# Creamos un cliente de Twitter con las credenciales necesarias
client = Client('en-US')

## Esta parte solo se runea la primera vez para guardar las cookies, después la comentamos o elim
# client.login(
#     auth_info_1='_____________',
```

```
#     password='______',
# )
# client.save_cookies('cookies.json');

# Cargamos las cookies para no tener que iniciar sesión cada vez
client.load_cookies(path='cookies.json');

# Obtenemos los tweets de un usuario específico (esto se modifica a mano para cada persona)
user = client.get_user_by_screen_name('ElonMusk')
tweets = user.get_tweets('Tweets')

# Creamos una lista vacía para almacenar los tweets
tweets_to_store = [];

# Iteramos sobre los tweets para extraer la información relevante
iter = 0
Itmax = 2

while iter < Itmax:
    for tweet in tweets:
        tweets_to_store.append({
            'created_at': tweet.created_at,
            'favorite_count': tweet.favorite_count,
            'full_text': tweet.text,
            'user': tweet.user.screen_name,
        })
    tweets = tweets.next()
    iter += 1
    print(len(tweets_to_store))

# Convertimos la lista de tweets en un marco de datos de pandas
df = pd.DataFrame(tweets_to_store)
# Guardamos los tweets en un archivo CSV
df.to_csv('tweets.csv', index=False)
```

Una vez hemos obtenido los datos, podemos proceder a la siguiente fase del proyecto, el análisis de sentimientos.

0.7.0.2 IA para análisis de sentimiento

El análisis de sentimiento implica la aplicación de técnicas avanzadas de inteligencia artificial para comprender la carga emocional de los mensajes expresados en las redes sociales.

El objetivo es determinar la polaridad de las opiniones y emociones de los textos, en nuestro caso empleamos *sentiment lexicons* basados en unigramas. Estos nos proporcionan las emociones y

sentimientos asociados a las palabras individuales de un texto, tomando la emoción y sentimiento de un texto como el mayor sentimiento en este.

El uso de sentiment lexicons frente a otras técnicas de sentiment analysis ofrece la principal ventaja de ser más rápido, lo que es necesario para hacer el cálculo rápido de opinión de muchos individuos. Cabe aclarar que lo que conseguimos mediante los sentiments lexicons realmente son la positvidad/neutralidad/negatividad de un texto, lo cual no siempre significa siempre sacar una opinión final, ya que si el texto habla muy negativamente sobre un tema y finalmente da un veredicto positivo se puede ver muy influenciado por una mayor cantidad de comentarios negativos y obtener que dicho texto habla negativamente. O como, por ejemplo, citar el comentario negativo de otra persona en un texto puede alterar negativamente el resultado positivo de un texto.

Iniciando nuestro análisis de sentimiento probamos distintos sentiment lexicons con opiniones polares, ya que existen otros como *ncr*[1] que son más concretos con el tipo de emoción, que fueron AFINN y bing.

- AFINN asigna a ciertas palabras una puntuación de entre -5 y 5, siendo una puntuación negativa un sentimiento negativo y una puntuación positiva un sentimiento positivo.
- Bing directamente asigna el valor positivo o negativo a cada palabra.

A la hora de realizar test con texto extraído de twitter y comparar la positividad del texto en si surgían varios inconvenientes. El primero era que, por ejemplo, al analizar las opiniones en twitter del tráiler de la última película de Deadpool se penalizaban palabras como "Strange", del nombre de un personaje de Marvel Dr.Strange, o "KILLED", que se refiera a la expresión "they KILLED IT!!!" siendo esta positiva, lo cual se puede llegar a corregir (o incluso prever) al analizar las palabras con una emoción asignada que más aparecen e ignorarlas. Otro problema que surge es no capturar matices de su dominio como tomar en cuenta calificativos delante de una palabra como "no es bueno" o "no es cierto", sarcasmo, o en general ideas que dependan del contexto. Dados los problemas surge una alternativa, VADER.

VADER *(Valence Aware Dictionary and sEntiment Reasoner)* es un sentiment lexicon basado en reglas gramaticales diseñado específicamente para analizar sentimientos expresados en redes sociales. VADER es capaz de analizar la positidad/neutralidad/negatividad un texto y dar un puntuaje ente -1 y 1, y al estar basado en reglas gramaticales soluciona la mayoría de los principales problemas como la negación o el uso de exclamaciones, e incluso tomar la intensidad de palabras en mayúsculas o tomar sentimientos de emojis hechos con caracteres como " :(".

Para probar el rendimietno de VADER se utilizaron principalmente 2 datasets pertenecientes a kaggle, IMDB Dataset of 50K Movie Reviews y Massive Rotten Tomatoes Movies & Reviews. Ambos datsets de 50.000 y 1444963 filas respectivamente, contienen reviews de paginas como IMDB y Rotten Tomatoes junto con un sentimiento positivo y negativo asociado. Tras comparar los resultados se obtiene que VADER tuvo un 69,226% de aciertos con el dataset de IMDB y un 58,278% de aciertos, que no esta nada mal comparado con el resto de lexicons, sin embargo, la tolerancia entre falsos sentimientos positivos o falsos sentimientos negativos se pude ir ajustando dependiendo de las restricciones que tenga el análisis.

[1]El NRC Emotion Lexicon es una lista de palabras en inglés y sus asociaciones con ocho emociones básicas (ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y disgusto) y dos sentimientos (negativo y positivo).

0.7.0.3 IA para topic modeling (clasificar por temas)

El topic modeling es una técnica poderosa en el análisis de datos que permite identificar y categorizar automáticamente los temas principales discutidos en grandes conjuntos de documentos, como los mensajes de redes sociales. En este proyecto, se ha empleado inteligencia artificial para llevar a cabo esta tarea de manera eficiente y precisa. Además de utilizar algoritmos estándar como LSA (Análisis Semántico Latente) y LDA (Asignación Latente de Dirichlet), se exploran métodos emergentes que mejoren la segmentación y comprensión de las discusiones en la plataforma. Esto permitirá una visión más profunda de las tendencias y preocupaciones de los usuarios, proporcionando así una base sólida para la asignación estratégica de anuncios.

La *asignación latente de Dirichlet* es uno de los algoritmos más comunes para el modelado de temas. Sin entrar en los detalles matemáticos del modelo, podemos entender que está guiado por dos principios.

- Cada documento es una mezcla de temas. Visualizamos que cada documento puede contener palabras de varios temas en proporciones específicas. Por ejemplo, en un modelo de dos temas podríamos decir "El Documento 1 consiste en un 90% del tema A y un 10% del tema B, mientras que el Documento 2 está compuesto por un 30% del tema A y un 70% del tema B".
- Cada tema es una mezcla de palabras. Por ejemplo, podríamos imaginar un modelo de dos temas sobre noticias estadounidenses, con un tema para "política" y otro para "entretenimiento". Las palabras más comunes en el tema de política podrían ser "Presidente", "Congreso" y "gobierno", mientras que el tema de entretenimiento podría incluir palabras como "películas", "televisión" y "actor". Es importante destacar que las palabras pueden compartirse entre temas; una palabra como "presupuesto" podría aparecer igualmente en ambos.

LDA es un método matemático usado para estimar los documentos y los temas a la vez. Se trata de encontrar la mezcla de palabras que está asociado a cada tema, y la mezcla de temas que está asociado a cada documento

- Clasificación:

Para clasificar los temas usaremos python. Para ello usaremos la librería bertopic, que es una implementación de LDA que utiliza BERT, otro lenguaje de programación orientado al procesamiento natural del lenguaje, para la clasificación de los mensajes.

```
# Importamos las librerías necesarias
import pandas as pd
from bertopic import BERTopic

# Cargamos el modelo preentrenado
topic_model = BERTopic.load("MaartenGr/BERTopic_ArXiv")

# Leemos los datos
```