NEOLAND

MEMORIA

PROYECTO FINAL

PREDICCIÓN DE TENDENCIAS EN TWITTER

BOOTCAMP DATA SCIENCE

NEOLAND, DICIEMBRE 2020

AUTORES: CARLOS RUIZ RODADO, DANIEL W. TÜMMLER ROSALES, ISMAEL MADALENA ARANGUREN Y PILAR LORENTE CONESA

ÍNDICE

1. DESCRIPCIÓN

1.1 OBJETIVOS

Las tendencias de Twitter son las palabras o *hashtags* más utilizadas en la red social durante un rango de tiempo determinado. El objeto principal de este trabajo es predecir la probabilidad de que un *hashtag* o un conjunto de palabras clave se convierta en tendencia en un futuro cercano, concretamente en las próximas 6 horas.

De forma secundaria y como parte el análisis de los datos, se pretende conocer qué hashtags son pagados- es decir, son tendencia porque alguien ha pagado para que lo sea- y qué hashtags o palabras están posicionados como *trends* de forma totalmente orgánica o por la mera interacción de los usuarios.

1.2 JUSTIFICACIÓN

El marketing y la publicidad tradicionales murieron en el momento en el que las redes sociales pasaron a ser un escaparate global. Así, las campañas en estas plataformas representan un punto fundamental en las estrategias de comunicación de las organizaciones, ya que una inversión correcta que proporcione un rendimiento óptimo, puede prever grandes ventas.

En nuestro trabajo, se ha decidido el estudio de Twitter por la propia naturaleza de la misma. A diferencia de Instagram o Facebook, la publicidad de un producto en Twitter no depende solo de la persona que lo anuncie o de una imagen bonita, sino también del texto. Twitter es una red eminentemente textual, por lo que aquí el poder pasa de la imagen a las letras. En un sector donde la imagen es fundamental, invertir en texto es más difícil y arriesgado. Por eso, con el algoritmo, se pretende brindar certeza en la inversión a la hora de lanzar hashtags con proyección de convertirse en tendencia, así como el impacto futuro.

1.3 METODOLOGÍA

El proyecto se ha basado en una investigación de los profesores Das et al., (2015), donde se hace un análisis de los hashtags, pero sin embargo se obvian las palabras. Además, la configuración de Twitter ha cambiado desde 2015 hasta ahora, por lo que se ofrece un enfoque renovado y mejorado.

Los conjuntos de datos elegidos se han obtenido mediante elaboración propia. En primer lugar, el periodo escogido ha sido de dos días consecutivos aleatorios, en este caso los llamaremos D1 y D2. De esta forma, los datasets necesarios para modelar son dos: dataset D1 con los tweets que contienen hashtags o palabras tendencia y hashtags o palabras que, a pesar de su interacción y número de tweets, no lograron ser tendencia; y el mismo conjunto

de datos, pero en el D2. La extracción de los tweets se ha ejecutado con la librería Twint[1], ya que trabaja de forma independiente a la API de Twitter y, por tanto, las limitaciones también son menores.

Por otro lado, también era necesario extraer las tendencias por hora de ambos días, para así conocer qué tweets se han posteado mencionando la tendencia y que tweets no. Así, se ha programado un web scraping de la web https://getdaytrends.com/spain/, con el que obtuvimos las listas de tendencias por hora de D1 Y D2.

Out[29]:

	Dia	N. Tweets	Usuarios	Tendencias
0	24	1399539	425182	290
1	25	1632046	454611	321

Figura 1. Número de tweets extraídos. Fuente: elaboración propia.

El preprocesamiento también se bifurca, distinguiendo entre tendencias y no tendencias. Llamando T1 al conjunto de tweets que contienen tendencias en D1, necesitamos el mismo conjunto, pero de no tendencias, que llamaremos NT1. En este punto, necesitamos establecer ciertos criterios para buscar esos hashtags o palabras no tendencia, que en nuestro caso han sido: número de tweets e interacción de los tweets. Así, hemos obtenido un conjunto de palabras y hashtags que nunca llegaron a ser tendencia, pero cuentan con características parecidas a las tendencias en cuanto a criterios.[2]

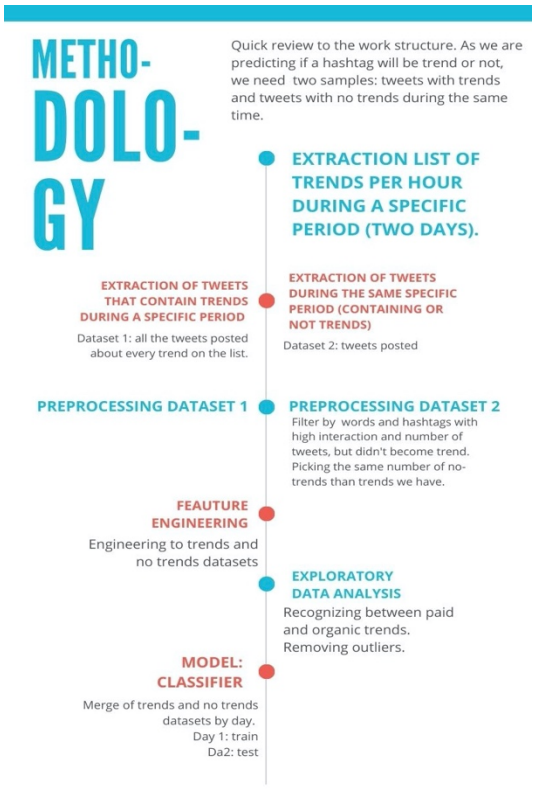

Figura 2. Methodology structure. Fuente: elaboración propia

Una vez obtenidos los datasets T1 U NT1 y T2 U NT2, aplicamos ingeniería de variables (apartado 1.3.1 VARIABLES) estudiando las últimas seis horas de la tendencia o no tendencia, análisis exploratorio de los datos y, por último, modelado.

1.3.1 VARIABLES

La ingeniería de variables propuesta consta de variables agregadas y variables dependientes del tiempo por cada una de los hashtags o palabras (término) que componen T1 U NT1 y T2 U NT2.

Variables agregadas

- Total tweets: número total de tweets posteados que contienen el término.
- Total_hashtags: número total de hashtags utilizados en los tweets que contienen el término (excluyendo el término en el caso de que sea un hashtag).

[1] Twint Project, https://github.com/twintproject

[2] ¿Cómo se determinan las tendencias?, https://help.twitter.com/es/using-twitter/twitter-trending-faqs

- Total_mentions: número total de menciones en los tweets que contienen el término.
- Total_reply_to: número total de respuestas de los tweets que contienen el término.
- Total_url: número total de tweets que contienen links por término.
- Total_photo: número tweets que contienen fotos por término.
- Total_interaction: suma del total de likes, replies y retweets de los tweets que contienen el término.

Variables dependientes del tiempo

En este caso, el estudio se ha hecho sobre las últimas seis horas de vida antes de que el término explote como tendencia o, en el caso de los términos que son no tendencia, las últimas seis horas antes de la hora en la que aparece por última vez.

- Tweet_count: número de tweets de cada término por hora estudiada.
- User_count: número de usuarios que postean el término por hora estudiada.
- Velocity: diferencia del conteo de tweets entre la hora t y t-1.
- Acceleration: diferencia de las velocidades entre la hora t y t-1

1.4 CONTEXTO: QUÉ ES TWITTER

Twitter se puede definir como "un servicio online de microblogging que reúne las características de blogs, redes sociales y mensajería instantánea, permitiendo a sus usuarios estar en contacto a tiempo real a través de *tweets*" (Fernández, 2020). Millones de usuarios de todo el mundo pueden postear pequeños textos de no más de 240 caracteres, incluyendo imágenes, vídeos, links e incluso clips de audio. Además, permite que los usuarios interactúen entre ellos. De esta forma, Twitter genera un abundante y continúo flujo de usuarios, lo que se traduce en un escaparate gratuito para organizaciones y empresas y un canal de comunicación directo con los clientes.

Las tendencias de Twitter tienen un espacio reservado dentro de la plataforma, destacando estos hashtags o palabras sobre el resto. Entrar en esta lista durante unas horas (o incluso minutos) aporta visibilidad e interacción.

El algoritmo de Twitter para establecer qué es tendencia y qué no tiene en cuenta diferentes aspectos, entre los que se encuentra el número de tweets, la interacción que obtienen esos tweets [3] y la naturaleza poco común de la palabra (Twitter España).

2. RESULTADOS

2.1 ANÁLISIS EXPLORATORIO DE LOS DATOS

El análisis exploratorio de los datos o EDA nos ha permitido distinguir entre aquellas tendencias que son pagadas para que se posicionen como tendencia y las que lo son por el mero tráfico de usuarios.

[3] Retweets, likes y replies.

En este sentido y acorde al objetivo principal del proyecto, es necesario eliminar las tendencias pagadas, ya que nuestro modelo pretende predecir una tendencia en función de la interacción de los usuarios, sin ningún tipo de inversión.

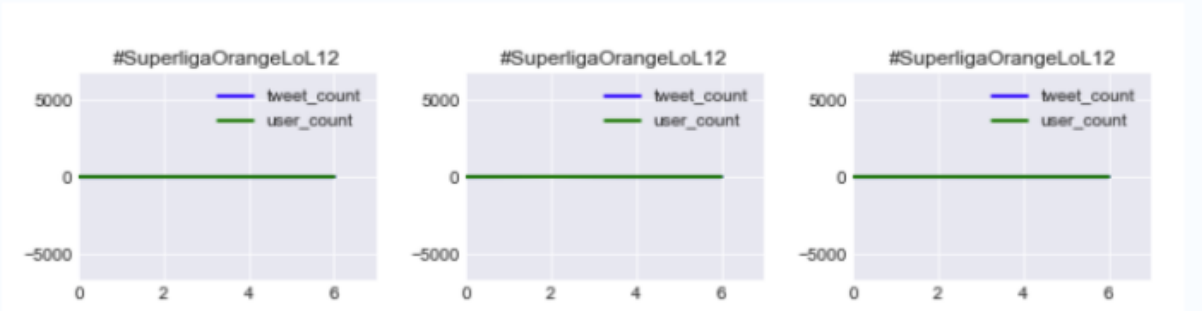

Figura 3. Tendencia pagada. Fuente: elaboración propia.

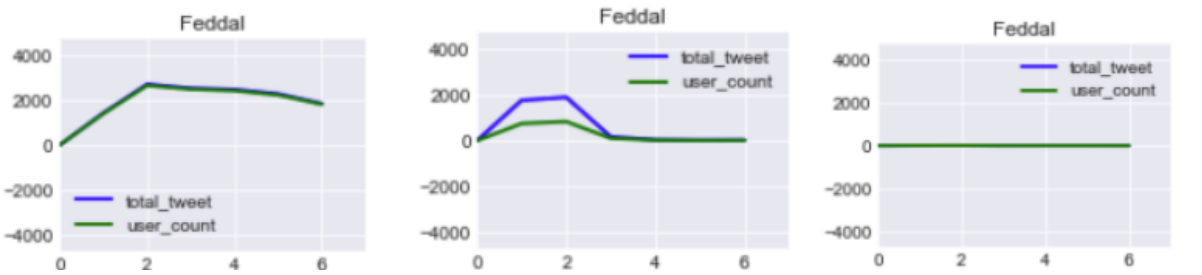

Figura 4. Tendencia orgánica. Fuente: elaboración propia.

En las figuras 3 y 4 se observa la diferencia en la evolución de una tendencia pagada y una orgánica, ambas posicionadas durante 3 horas como tendencia en Twitter España. Mientras que en la figura 3 no existe movimiento, es decir, no hay nadie escribiendo sobre ello, la figura 4 tiene una progresión de subida y de bajada en el volumen de tweets y usuarios.

De este análisis también se desprenden ideas en cuanto a las divergencias entre los términos que consiguen ser tendencia y los que no. Ambos tienen un punto de "explosión", sin embargo, las no tendencias no presentan la misma temporalidad que las tendencias (Figura 5), ni siquiera las mismas características gramaticales (palabras comunes).

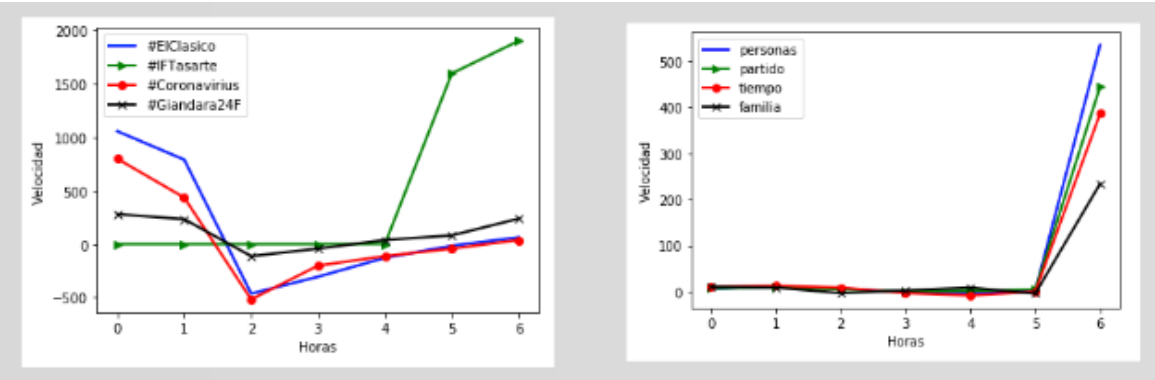

Figura 5. Diferencia entre trend y no trend. Fuente: elaboración propia

Por último, es importante destacar los también diferentes métodos para eliminar outliers. La distribución de estos no se puede tratar por igual en las tendencias que en las no tendencias. Por ejemplo, una tendencia con 0 interacción y 3 tweets es un outlier. Sin

embargo, una no tendencia con 0 interacción y 3 tweets no tiene porqué serlo, ya que precisamente es eso: no tendencia. Por lo tanto, mientras que en las tendencias se han eliminado los outliers por encima del cuartil superior(Q3) con un aumento del rango intercuartílico (IQR) y por debajo del cuartil inferior(Q1), en las no tendencias solamente se han obviado los que quedan por encima del cuartil superior. (Figura 6)

$$IQR = Q3 - Q1$$

TENDENCIAS {interaction >= Q1 | interaction<= Q3 + 2.5 *IQR}

NO TENDENCIAS { interaction <= Q3}

Figura 6. Fórmula para eliminar outliers por cuartiles. Fuente: elaboración propia.

2.2 MODELADO

El problema se ha afrontado aplicando algoritmos de clasificación de aprendizaje supervisado (SML), en concreto, de la librería scikit-learn. Además, ya que estamos tratando con datos de clasificación binaria (trend o no trend), etiquetamos a nuestros patrones como 0 = no tendencia y 1 = tendencia.

Respecto a los hiperparámetros de nuestros modelos, se ha decidido usar *GridSearchCV*, que a partir de una búsqueda exhaustiva de los parámetros elegidos por nosotros para cada *grid,* nos ha permitido obtener los mejores resultados posibles.

Los algoritmos de clasificación usados han sido:

- RandomForest Clasiffier
- Logistic Regression
- GaussianNB
- KNeighbors Classifier
- DecisionTree Classifier
- SVM Classifier

Y las métricas:

- Accuracy
- Precision
- Recall
- F1 – Score
- AUC

A partir de estos algoritmos y métricas, hemos obtenido los resultados que se muestran en la figura 7. Como podemos observar en la tabla, en base a los mejores resultados obtenidos, dada la métrica AUC comprobamos que el mejor algoritmo de clasificación ha sido el *SVM Classifier* con un resultado de 0.895. En cuanto a la exhaustividad (*Recall*), también comprobamos que el algoritmo que más se aproxima a nuestro objetivo vuelve a ser *SVM Classifier* con 0.92.

	Model	Accuracy	Precision	Recall	F1-Score	AUC
0	SVM (Classifier)	0.823	0.696	0.923	0.794	0.897
1	RandomForestClassifier	0.812	0.697	0.871	0.774	0.881
2	KNeighborsClassifier	0.817	0.698	0.889	0.782	0.859
3	GaussianNB	0.742	0.627	0.753	0.684	0.820
4	LogisticRegression	0.827	0.705	0.918	0.797	0.816
5	DecisionTreeClassifier	0.772	0.645	0.853	0.735	0.770

Figura 7. Tabla de métricas. Fuente: elaboración propia.

Si observamos la matriz de confusión (figura 8), se puede ver un ligero descenso del recall con respecto a la predicción de las no tendencias, que puede deberse a los múltiples criterios que Twitter toma en cuenta para hacer un término tendencia, entre los que se encuentra el estudio de los usuarios, algo que por falta de tiempo no ha sido posible de realizar. Independientemente de esto, podemos afirmar que diferencia de forma muy clara lo que se convertirá en tendencia.

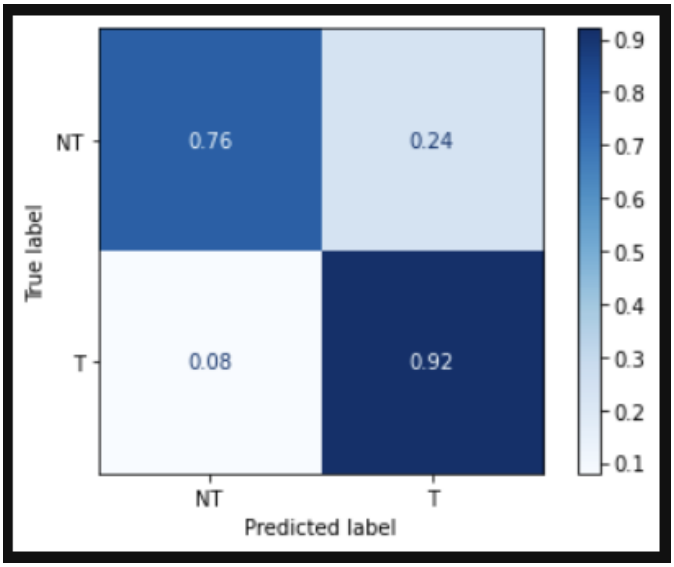

Figura 8. Matriz de confusión del algoritmo SVM.
Fuente:elaboración propia.

3. FASES Y TIEMPOS

Se pueden definir cuatro fases claras en las que se ha estructurado el trabajo: extracción de datos, preprocesamiento e ingeniería de variables, análisis y modelado.

Desde el punto de vista de las fortalezas, conocer al milímetro los datos ha sido fundamental para agilizar el trabajo. Sin embargo, este punto también se ha convertido en

una debilidad durante todo el proceso, ya que Twitter no lo pone fácil para obtener todos los datos que se desean. Por este motivo, la extracción tuvo que dividirse a lo largo de 7 días, lo que consumió un 50% del tiempo previsto.

La segunda y tercera fase se desarrollaron a lo largo 6 días, lo que representa un 40% del tiempo. Al trabajar con una librería externa (Twint) los datos venían dados con numerosas variables inservibles. Además, tuvimos que etiquetar cada uno de los tweets en función de si contenían tendencias o no tendencias, siendo la etiqueta el término en sí. Esto se traduce en una labor de búsqueda y comparación con la lista de tendencias previamente obtenida bastante pesada. Sin duda, el etiquetado de las no tendencias, es decir, la búsqueda de términos no comunes que más se repiten y satisfagan criterios de volumen e interacción ha sido uno de los principales retos a batir en esta fase. A todo lo anterior, se suman cuestiones de NLP y limpieza de texto (tokenización) que, debido a la cantidad de datos, tardaba horas en ejecutar.

El 10% de tiempo restante se invirtió enteramente en el análisis y modelado.

4. CONCLUSIONES

Los algoritmos de las redes sociales son extremadamente complejos en su funcionamiento y, por tanto, complejos en entendimiento. El algoritmo de Twitter para establecer qué es tendencia y qué no tiene en cuenta multitud de factores y esto implica un estudio arduo de sus características para descifrarlo y poder adelantarnos al comportamiento de los usuarios. Aunque a pesar de esto, el proyecto ha logrado obtener un buen resultado teniendo en cuenta solamente dos variables como características principales: el número de tweets y la interacción.

Con un 0.90 bajo la curva ROC, el algoritmo propuesto define con facilidad qué se convertirá en tendencia en las próximas horas. Conocer esto puede suponer una gran ventaja para equipos de marketing y comunicación que estén interesados en tener presencia en Twitter.

Por otro lado, la predicción de las no tendencias es algo más baja con respecto a las tendencias, aunque estamos convencidos de que con un estudio mucho más profundo y prolongado, teniendo en cuenta el perfil de los usuarios que interactúan con los términos (número de seguidores, si está verificado o no, etc), así como las relaciones entre usuarios, se puede mejorar el modelo.

Sin duda, estas plataformas contienen una cantidad ingente de datos que se generan cada segundo, algo que puede ser visto como una potencial fuente de información para elaborar algoritmos que contribuyan positivamente desde un punto de vista económico, pero también y no menos importante, desde un punto de vista social.

REFERENCIAS

Das, A. et. Alt. (2015). Predicting Trends in the Twitter Social Network: A Machine Learning Approach. Recuperado de: https://www.researchgate.net/publication/294482813_Predicting_Trends_in_the_Twitter_Social_Network_A_Machine_Learning_Approach/citations#fullTextFileContent

Fernández, R. (2020). Twitter en España - Datos estadísticos. Recuperado de: https://es.statista.com/temas/3595/twitter-en-espana/