UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Estudio de la relación entre las visitas a comercios y su reputación cultural en Twitter.

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Cano Navarro, Ignacio

Tutor/a: Ferri Ramírez, César

Cotutor/a: Hernández Orallo, José

Director/a Experimental: BALSA BARREIRO, JOSE

CURSO ACADÉMICO: 2021/2022

Agradecimientos

Quería agradecer a mis compañeros de la carrera, que han sufrido y disfrutado conmigo de estos cuatro años que sin duda ninguno olvidaremos, y en especial a Álvaro Mazcuñán, Miquel Marín, Elizaveta Gilyarovskaya y a Ángel Langdon por todos esos momentos. También a mi familia, que me ha apoyado en todo momento desde el primer día.

Por último quería agradecer a mis tutores, José Hernández Orallo, Cèsar Ferri Ramirez y José Balsa-Barreiro, porque sin ellos este trabajo no hubiera sido posible.

Ignacio Cano Navarro

Resum

El propòsit d'aquest treball és estudiar la possible relació entre el sentiment dels habitants de la ciutat de Nova York sobre països rellevants per als ciutadans nord-americans i les posteriors visites a negocis situats en la mateixa ciutat de Nova York associats a les esmentades cultures. D'aquesta manera, s'ajudarà a entendre una problemàtica apressant en el terreny de l'economia, el comportament del consumidor. Un dels principals impediments a l'hora de realitzar estudis similars és aconseguir dades fiables de visites a negocis, i més en aquest cas, ja que es tracta de negocis fàcilment identificables amb una cultura. És a dir, negocis que venen, manipulen o serveixen productes tradicionals d'uns certs països. En aquest cas, la informació prové de Safegraph, proveïdor de dades de punts d'interès (POIs). En aquest treball s'afronta també una problemàtica coneguda en el món de la sociologia, com és l'estudi del sentiment de la població. La solució proposada és una aproximació al sentiment dels habitants de Nova York mitjançant l'extracció del sentiment de tuits realitzats per novaiorquesos sobre cultures esmentades anteriorment utilitzant Transformers. Per a estudiar l'esmentada relació entre totes dues variables, s'ha proposat per a valorar si existeix, l'esmentada causa-efecte un test de causalitat de Granger, utilitzat habitualment en economia, amb una posterior interpretació d'un model OLS que permet quantificar l'error en intentar explicar les visites a negocis pel sentiment cap a la cultura que representen aquests negocis.

Paraules clau: comerç, retailing, twitter, sentiment, visites, cultures, transformers

Resumen

El propósito de este trabajo es estudiar la posible relación entre el sentimiento de los habitantes de la ciudad de Nueva York sobre países relevantes para los ciudadanos estadounidenses y las posteriores visitas a negocios ubicados en la misma ciudad de Nueva York asociados a las mencionadas culturas. De esta forma, se ayudará a entender una problemática acuciante en el terreno de la economía, el comportamiento del consumidor. Uno de los principales impedimentos a la hora de realizar estudios similares es conseguir datos fiables de visitas a negocios, y más en este caso, ya que se trata de negocios fácilmente identificables con una cultura. Es decir, negocios que venden, manipulan o sirven productos tradicionales de ciertos países. En este caso, la información proviene de Safegraph, proveedor de datos de puntos de interés (POIs). En este trabajo se afronta también una problemática conocida en el mundo de la sociología, como es el estudio del sentimiento de la población. La solución propuesta es una aproximación al sentimiento de los habitantes de Nueva York mediante la extracción del sentimiento de *tweets* realizados por neoyorquinos sobre culturas mencionadas anteriormente usando *transformers*. Para estudiar la mencionada relación entre ambas variables, se ha propuesto para valorar si existe, la mencionada causa-efecto un test de causalidad de Granger, utilizado habitualmente en economía, con una posterior interpretación de un modelo OLS que permite cuantificar el error al intentar explicar las visitas a negocios por el sentimiento hacia la cultura que representan dichos negocios.

Palabras clave: comercio, retailing, twitter, sentimiento, visitas, culturas, transformers

Abstract

The purpose of this paper is to study the possible relationship between New York City residents' feelings about countries relevant to U.S. citizens and subsequent visits to New York City businesses associated with those cultures. This will help to understand a pressing issue in the field of economics, namely consumer behavior. One of the main impediments to conducting similar studies is obtaining reliable data on visits to businesses, especially in this case, since these are businesses that are easily identifiable with a culture. That is, businesses that sell, handle or serve traditional products from certain countries. In this case, the information comes from Safegraph, a provider of POI data. This work also addresses a well-known problem in the world of sociology, which is the study of the sentiment of the population. The proposed solution is an approach to the sentiment of New Yorkers by extracting sentiment from *tweets* made by New Yorkers about aforementioned cultures using *transformers*. To study the aforementioned relationship between the two variables, a Granger causality test, commonly used in economics, has been proposed to assess whether the aforementioned cause-effect exists, with a subsequent interpretation of an OLS model that allows quantifying the error when trying to explain visits to businesses by the sentiment towards the culture represented by those businesses.

Key words: commerce, retailing, twitter, sentiment visits, cultures,transformers

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

La integración de múltiples culturas en una misma área metropolitana trae consigo inevitables desigualdades que la ciudad de Nueva York lleva tratando de paliar desde hace décadas. No solo se trata de desigualdades entre individuos, como es la desigualdad de oportunidades dependiendo de los orígenes, sino que también hay desigualdad en las decisiones de los consumidores dependiendo del país de origen, es decir, en los productos que consumen o que estarían dispuestos a ello. Para combatir los efectos de la desigualdad, se gastan al año millones de dólares en ayudar a aquellos que se están quedando atrás, incluyendo también ayudas a negocios.

En este trabajo se propone relacionar y posteriormente predecir las visitas a negocios de restauración con una clara vinculación a una cultura con el sentimiento que tienen los habitantes de Nueva York acerca de los países de origen de estos negocios. ¿Tienen los consumidores de Nueva York, por razones históricas o políticas, una tendencia a modificar las visitas de negocios pertenecientes a una cultura en función al sentimiento existente? Esto servirá, no solo a las propias compañías detrás de los negocios a entender el comportamiento de sus consumidores y a preparar sus campañas de márketing, sino también a los políticos que podrán preparar paquetes de medidas ajustadas a las comunidades que lo necesiten.

1.1 Motivación

La motivación detrás de la realización de este proyecto surgió del deseo de hacer algo innovador y que pueda sentar precedentes en el área del entendimiento del comportamiento de los consumidores. Además, este tema de investigación está fuertemente relacionado con el procesamiento del lenguaje natural, tópico muy interesante y de fuerte actualidad, del cual el autor tenía muchas ganas de profundizar sus conocimientos, en especial en el uso de *transformers* [1].

Por último, la oportunidad de trabajar con datos geolocalizados provistos por Safegraph [1] sin duda motivó a la decisión de intentar sacar provecho de ese diamante en bruto como son sus datos, que generosamente ceden a la comunidad académica sin ánimo de lucro alguno.

[1] https://www.safegraph.com/

1.2 Objetivos

El objetivo principal de este trabajo es encontrar una relación entre el sentimiento de los habitantes de la ciudad de Nueva York sobre países relevantes para los ciudadanos estadounidenses y las posteriores visitas a negocios ubicados en la misma ciudad de Nueva York asociados a las mencionadas culturas. Para conseguir esta meta habrán dos grandes fases: Una primera concentrada en la obtención de las dos series temporales que buscaremos relacionar (visitas y sentimiento) en la que destacamos el uso de *transformers* [1] para la parte de sentimiento y una segunda en la que usamos técnicas de series temporales para buscar dicha relación y una predicción mediante OLS [2].

1.3 Estructura de la memoria

En este trabajo se pueden encontrar 12 capítulos:

- **Capítulo 1, Introducción.** En este apartado se introducen los objetivos del proyecto, así como una primera toma de contacto con el problema planteado y la motivación que ha llevado a la realización de este trabajo.
- **Capítulo 2, Estado del arte.** En este capítulo se intentan aportar todos los conocimientos previos necesarios para la completa comprensión de este trabajo. También se incluye el estado actual del estado del arte respecto a las áreas que toca este trabajo, así como la propuesta.
- **Capítulo 3, Análisis del problema.** Aquí se analizan dos problemas muy concretos que deberían de tenerse en cuenta en todo proyecto, el análisis del gasto energético (relacionado con los objetivos del desarrollo sostenible), así como el análisis de riesgos, especialmente importante en cualquier proyecto con datos privados.
- **Capítulo 4, Preparación y comprensión de los datos.** Se exponen los dos principales conjuntos de datos sobre los cuales se basa este proyecto, así como su obtención y problemas asociados.
- **Capítulo 5, Preprocesamiento de los datos.** En este capítulo se desarrolla el preprocesamiento necesario para poder cumplir los objetivos del proyecto.
- **Capítulo 6, Conocimiento extraído y evaluación de modelos.** En este apartado se busca extraer de los datos los objetivos propuestos en los anteriores capítulos, con la ayuda de modelos y tests estadísticos.
- **Capítulo 7, Despliegue.** Se defiende un potencial despliegue del proyecto como producto.
- **Capítulo 8, Limitaciones encontradas.** Se definen los problemas encontrados durante la realización del trabajo.

- **Capítulo 9, Conclusiones.** Se revisan los objetivos que han podido o no cumplirse a lo largo del trabajo y se expone el legado que este trabajo ha dejado.
- **Capítulo 10, Relación del trabajo desarrollado con los estudios cursados.**
- **Capítulo 11, Trabajo futuro.** Se mencionan posibles áreas de mejora que surgen a raíz de este trabajo para proyectos futuros.
- **Capítulo 12, Relación con los Objetivos del Desarrollo Sostenible (ODS)**

CAPÍTULO 2
ESTADO DEL ARTE

2.1 Conocimiento previo

2.1.1. Aprendizaje Automático

El aprendizaje automático es una evolución relativamente moderna de los tradicionales algoritmos informáticos, diseñada para imitar la inteligencia humana mediante el aprendizaje del entorno. Con el aprendizaje automático se consigue que las máquinas mejores o obtengan procedimientos de actuación a través de la experiencia (datos). En sectores como la medicina, la investigación biomédica, transportes, seguridad y en muchos otros, las técnicas basadas en el aprendizaje automático están triunfando y no dejan de implementarse en nuevos ámbitos.

2.1.2. Aprendizaje Profundo

El aprendizaje profundo, Deep Learning en inglés [4], es un conjunto de algoritmos del aprendizaje automático que intenta modelar la realidad mediante abstracciones de alto nivel en los datos de entrada y que utiliza complejas arquitecturas integradas en su aprendizaje e inferencia. Es muy común confundir los términos Inteligencia Artificial, Aprendizaje Automático y Aprendizaje Profundo, así que se van a destacar las diferencias entre todos ellos:

El lector puede imaginar un esquema de matrioskha [1] en el cual, en la capa superior, se encontraría la inteligencia artificial, la cual engloba el resto de componentes. Justo debajo se encuentra el aprendizaje automático, el cual engloba el aprendizaje profundo y las redes neuronales.

La mayor diferencia entre el aprendizaje automático y el aprendizaje profundo se encuentra en la forma de aprender de cada uno. En el aprendizaje profundo se añaden más capas de abstracción, lo que permite que las redes produndas aprendan mayor cantidad de información de los datos de entrada. En el aprendizaje automático, los modelos que se utilizan son mucho más dependientes del conjunto de datos de entrenamiento. En cambio, en el aprendizaje profundo, se elimina esta pequeña (pero importante) restricción y permite que los datos de entrenamiento del conjunto de modelos pertenecientes al deep learning, utilicen

[1] https://es.wikipedia.org/wiki/Matrioshka

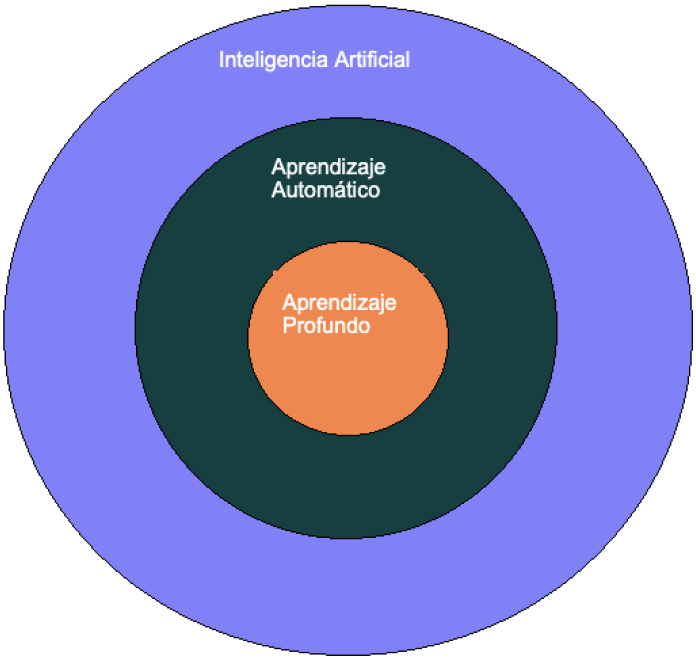


Figura 2.1: Diferencias entre Inteligencia Artificial, Aprendizaje Automático y Aprendizaje Profundo. Fuente: Elaboración Propia

datos en grandes cantidades aunque como se podrá observar en futuras secciones siempre es conveniente tener cuidado y una mínima limpieza con los datos de entrada de cualquier modelo.

2.1.3. Procesamiento del Lenguaje Natural (PLN)

Definición

El procesamiento del lenguaje natural (PLN) es la capacidad que tiene un programa informático para entender el lenguaje humano hablado y escrito, conocido como lenguaje natural. Es un componente de la inteligencia artificial (IA). A continuación se procederá a resumir los tratamientos más comunes a los datos de entrada de cualquier proyecto que involucra el uso de PLN:

El preprocesamiento de datos consiste en preparar y "limpiar"los datos de texto para que las máquinas puedan analizarlos. El preprocesamiento pone los datos en forma viable y destaca las características del texto con las que puede trabajar un algoritmo. Hay varias formas de hacerlo, entre ellas:

- Tokenización. Se trata de dividir el texto en unidades más pequeñas con las que trabajar.
- Eliminación de las palabras vacías. En este caso se eliminan las palabras comunes del texto para que queden las palabras únicas que ofrecen más información sobre el texto.

- Lematización y stemming. En este caso, las palabras se reducen a su raíz para poder procesarlas.

- Etiquetado de parte del discurso. Es cuando las palabras se marcan en función de la parte del discurso que son, como sustantivos, verbos y adjetivos.

Una vez preprocesados los datos, se desarrolla un algoritmo para procesarlos. Hay muchos algoritmos diferentes de procesamiento del lenguaje natural, pero se suelen utilizar dos tipos principales:

- Sistema basado en reglas. Este sistema utiliza reglas lingüísticas cuidadosamente diseñadas. Este enfoque se utilizó al principio del desarrollo del procesamiento del lenguaje natural, y todavía se utiliza.

- Modelos basados en aprendizaje automático y aprendizaje profundo. Este apartado es el que se desarrollará en mayor profundidad en un futuro apartado, ya que uno de los modelos basados en aprendizaje profundo, transformers, es el elegido para la tarea que se llevará a cabo en este proyecto, el análisis de sentimiento.

Tareas en el PLN

El procesamiento del lenguaje natural se puede aplicar a muchas tareas, entre ellas se destacan:

- Traducción

- Agentes virtuales y chatbots

- Detección de spam

- Análisis del sentimiento. El Análisis de Sentimiento es un área novedosa y amplia del Procesamiento del Lenguaje Natural (PLN) cuyo objetivo es comprender los sentimientos y opiniones de las personas sobre un tema determinado. La aplicación del análisis de sentimientos puede usarse para evaluar automáticamente las reseñas de productos y servicios online, o conocer la opinión de las masas sobre una compañía o estado. Sin duda, esta herramienta es muy útil no solo para empresas sino también para políticos también. Las empresas y organizaciones gastan una gran cantidad de dinero en encontrar las opiniones y los sentimientos de los clientes, ya que esta información es útil para explotar su marketing-mix con el fin de conocer a su consumidor.

  Típicamente, cuando se realiza el análisis de sentimiento de un texto, éste se clasifica en **positivo**, **neutro** o **negativo** (como se verá más adelante, éste es el caso de este proyecto).

2.1.4. Transformers

En relación con el aprendizaje profundo, encontramos una relativamente nueva arquitectura de red, la cual ha supuesto toda una revolución en el mundo de la inteligencia artificial. Originalmente diseñada por un equipo de *Google*, esta arquitectura fue publicada en el conocido trabajo académico *attention is all you need* [1], convirtiéndose en uno de los descubrimientos más relevantes de los últimos años. La clave en el gran descubrimiento y posterior éxito de los *Transformers* es poder aplicar una capa de atención que es capaz de captar más información de los datos de entrada.

La estructura general de un *transformer* es la siguiente:

Atención

Una de las características principales que diferencian a los modelos basados en la arquitectura de los transformers del resto es la capa de atención (*self attention*). La *self attention*, traducida a veces como autoatención, es un mecanismo de atención que relaciona diferentes posiciones de una misma secuencia para calcular una representación de la misma y dotar de contexto dicha secuencia de entrada.

Cuando un ordenador recibe una frase, considera cada palabra como un token t, y cada token tiene un *word embedding* V. Pero estos *word embedding* de palabras no tienen contexto. Así que la idea es aplicar algún tipo de ponderación o similitud para obtener los *word embedding* finales Y, que tienen más contexto que los *word embedding* iniciales V. A partir de esta aproximación, calculamos los vectores de pesos W, haciendo el producto escalar entre los *embeddings* de las palabras que forman el input (típicamente una frase).

Multi-head attention, Queries, Keys y Values

El problema de la autoatención es que no se entrena nada. Pero si añadimos algunos parámetros entrenables, la red puede entonces aprender algunos patrones que dan un contexto mucho mejor. Este parámetro entrenable puede ser una matriz cuyos valores se entrenan. Así que se introdujo la idea de consulta, clave y valores que se encapsulan dentro del concepto del *Multi-head attention*.

Estos vectores creados como abstracciones son útiles para calcular la atención (más detalles sobre cada uno a continuación). Se calculan multiplicando el vector de entrada (X) por matrices de peso que se aprenden durante el entrenamiento.

- *Query vector*:

$$q = X * Wq$$

  Este vector representa una palabra de la secuencia de entrada.

- *Key vector*:

$$k = X * Wk$$

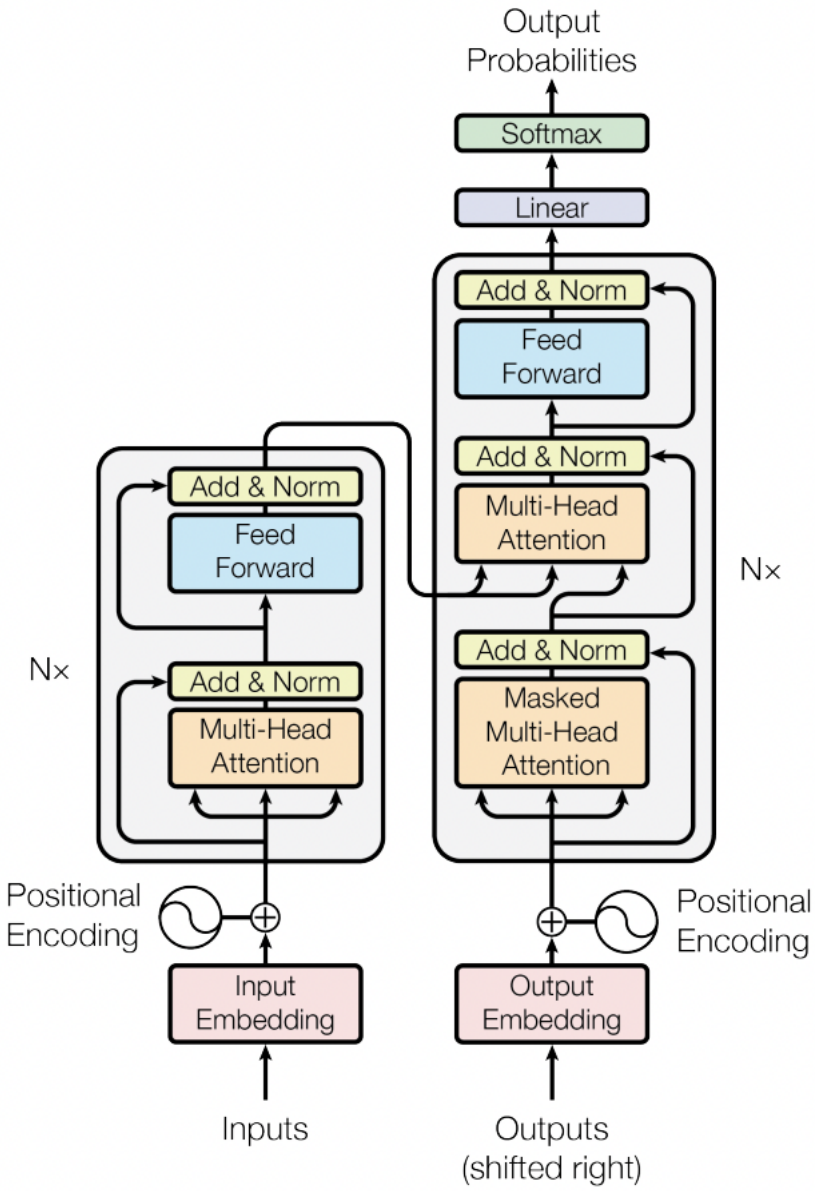


Figura 2.2: Transformer. Fuente: [1]

Este vector es el resultado de multiplicar el *query vector* por cada uno de los embeddings de la secuencia de entrada.

- *Value vector*:

$$v = X * Wv$$

Podemos representar el cálculo de la *multi-head attention* de una cadena de entrada i con la siguiente expresión:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Además, las matrices Mk, Mq y Mv pueden ahora ser entrenados con una red neuronal, y dan un contexto mucho mejor que el de la autoatención. Mejorando así el modelo de atención explicado anteriormente.

Existen tres tipos de *multi-head attention* posibles aplicables en un modelo:

- Atención en codificador-decodificador: las *queries* provienen de la capa anterior del decodificador y las *keys values* de la memoria provienen de la salida del codificador. Esto permite que cada posición del decodificador atienda a todas las posiciones de la secuencia de entrada.
- Atención en los bloques de codificación: En la capa de autoatención todas las *keys*, *values* y *queries* provienen del mismo lugar, en este caso, la salida de la capa anterior en el codificador.
- Atención propia en la secuencia de salida: Una cosa que debemos tener en cuenta aquí es que el alcance de la autoatención se limita a las palabras que ocurren antes de una palabra dada. Así se evita cualquier fuga de información durante el entrenamiento del modelo. Esto se hace enmascarando las palabras que ocurren después para cada paso. Así, para el paso 1, sólo la primera palabra de la secuencia de salida *no* está enmascarada, para el paso 2, las dos primeras palabras *no* están enmascaradas y así sucesivamente.

Gracias a esta estructura codificador-decodificador en la que las entradas de un bloque sirven como entrada del siguiente, utilizando la atención y todas las estrategias vistas en esta sección, se consigue crear uno de los modelos más dominantes actualmente en el área del NLP, y del cual han ido surgiendo variantes estos últimos años como [6] [9] [7]

2.1.5. Transfer learning (usando Hugging Face)

La mayoría de *transformers* [1] más conocidos como BERT [6], BART [9], GPT-3 [7] se han entrenado con grandes cantidades de texto en bruto de forma autosupervisada. El aprendizaje autosupervisado [2] es un tipo de entrenamiento en el

[2] https://es.wikipedia.org/wiki/Aprendizaje_no_supervisado

que la variable a predecir se calcula automáticamente a partir de las entradas del modelo. En el caso de texto, se predice el siguiente *token* a partir de los anteriores. Esto significa que no se necesitan expertos que etiqueten los datos a mano.

Este tipo de modelo desarrolla una comprensión del lenguaje sobre el que se ha entrenado, pero no resulta prácticamente útil para ejercicios concretos, como puede ser la tarea que se va a llevar a cabo en el presente trabajo, el análisis de sentimiento. Por eso, el modelo general preentrenado pasa por un proceso llamado aprendizaje transferido (más conocido como *transfer learning* [5] en inglés). Durante este proceso, el modelo se perfecciona de forma supervisada, es decir, utilizando *datasets* creados por humanos para una tarea específica como puede ser el análisis de sentimiento.

El *transfer learning* [5] es una de las técnicas más de moda de los últimos tiempos, con la que se consigue muy buenos resultados en todo tipo de tareas y sin duda uno de los motivos por los que los *transformers* [1] han conseguido tanta fama y éxito.

2.1.6. Métricas

Las métricas utilizadas son muy distintas en función del modelo utilizado así como del objetivo del proyecto. Aquí se presentan las métricas más comunes:

Modelos de regresión

Si el modelo es de regresión, las métricas que se deben utilizar tienden a minimizar la desviación entre el valor real y el predicho por el modelo. Por tanto siempre se querrá minimizar este tipo de métrica. Algunas de las métricas más importantes son:

- MAE o *"mean absolute error"*. Media de los errores en valor absoluto.

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2.1}$$

- MSE o *"mean squared error"*. Similar al MAE pero con los errores al cuadrado..

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.2}$$

- RMSE o *"root-mean-square error"*. Igual que el MSE pero con una raíz cuadrada, se penalizan los casos en los que hay mucha distancia entre el valor predicho y el real.

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{2.3}$$

- R^2 o coeficiente de determinación. Explica la variabilidad que el modelo es capaz de captar de los datos, se encuentra en un intervalo entre 0 y 1, siendo 1 una explicación perfecta de la variabilidad.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=0}^{n}(\hat{y}_i - \frac{\sum y}{n})^2}{\sum_{i=0}^{n}(y_i - \frac{\sum y}{n})^2} \tag{2.4}$$

Modelos de clasificación

Los modelos de clasificación utilizan otro tipo de métricas, muy distintas de las vistas anteriormente. Aquí solo se muestran absolutamente necesarias para este proyecto, pero hay muchas otras:

- Matriz de confusión. Es una herramienta que permite la visualización del desempeño de un modelo de clasificación. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a la clase real. [3] El objetivo principal es incrementar el número de muestras que caen en la diagonal principal de la matriz (aumento de predicciones correctas). También, nos permite ver como el modelo es capaz de discriminar entre las clases objetivo.

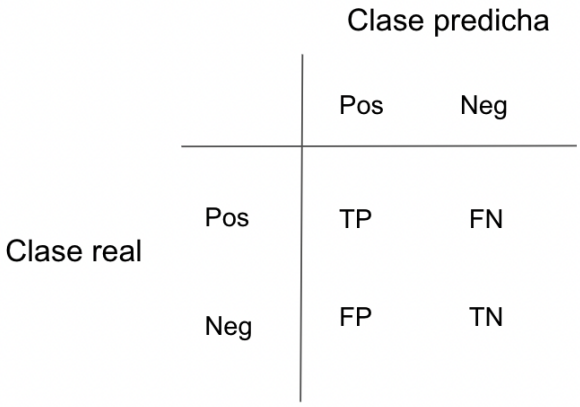


Figura 2.3: Matriz de confusión binaria. Fuente: Elaboración propia

- *Accuracy*. Muestra el porcentaje de muestras que acierta el modelo.

$$Accuracy = \frac{TP + FP}{n} = \frac{TP + FP}{TP + FP + TN + FN} \tag{2.5}$$

2.1.7. Correlación cruzada

La correlación cruzada es una medida que rastrea el decalaje entre dos series temporales. Se utiliza para comparar series temporales y determinar objetivamente, mediante coeficientes de correlación que se calculan en cada decajale, lo

[3] https://es.wikipedia.org/wiki/Matriz_de_confusi%C3%B3n

bien que coinciden entre sí y, en particular, en qué momento se produce la mejor coincidencia.

La correlación cruzada se utiliza generalmente cuando se mide la información entre dos series temporales diferentes. El rango posible para el coeficiente de correlación de los datos de las series temporales es de -1 a 1. Cuanto más se acerque el valor de la correlación cruzada a 1, más idénticas serán las series en el decalaje i.

La fórmula que generalmente se usa para calcular el coeficiente en cada paso de la correlación cruzada es:

$$c_k = \sum_n a_{n+k} * \bar{v_n} \tag{2.6}$$

Con las secuencias a y v rellenándose de ceros cuando sea necesario y $\bar{v_n}$ denotando la conjugación compleja.

2.1.8. Test de Granger

La causalidad de Granger o test de Granger [8] es una prueba econométrica utilizada para verificar la utilidad de una variable para predecir otra.

Básicamente, el test prueba añadir valores anteriores de la hipotética variable X para predecir la variable Y y si estos valores dan al modelo información estadísticamente relevante, se dice que la serie temporal X *causa grangerianamente* Y. Esto se realiza usando una serie de F-tests y t-tests [4] [5], entre otros.

2.2 Precedentes

Para encontrar inspiración y referencias, se intentó buscar en la plataforma de gestión de trabajos académicos de la Universidad Politécnica de Valencia [6], sin embargo, no se encontró ningún trabajo previo un objetivo parecido. Tampoco en *google scholar*. Sin embargo en este último se encontró el trabajo académico que inspiró este trabajo [10]. Como se puede ver, en este trabajo, se busca relacionar el sentimiento de usuarios de *Twitter* con el crecimiento de compañías de telecomunicaciones en India. A partir de esta idea, surge la posibilidad de estudiar el componente cultural y con ello los objetivos de este proyecto.

También, son de actualidad los trabajos académicos que buscan estudiar el análisis de sentimiento en relación con la bolsa de valores [11] [12]. De todos estos trabajos se ha conseguido inspiración y ideas a la hora de tratar con datos de redes sociales, análisis de sentimiento, series temporales...

[4] https://en.wikipedia.org/wiki/F-test
[5] https://www.jmp.com/es_co/statistics-knowledge-portal/t-test.html
[6] https://riunet.upv.es/handle/10251/10994

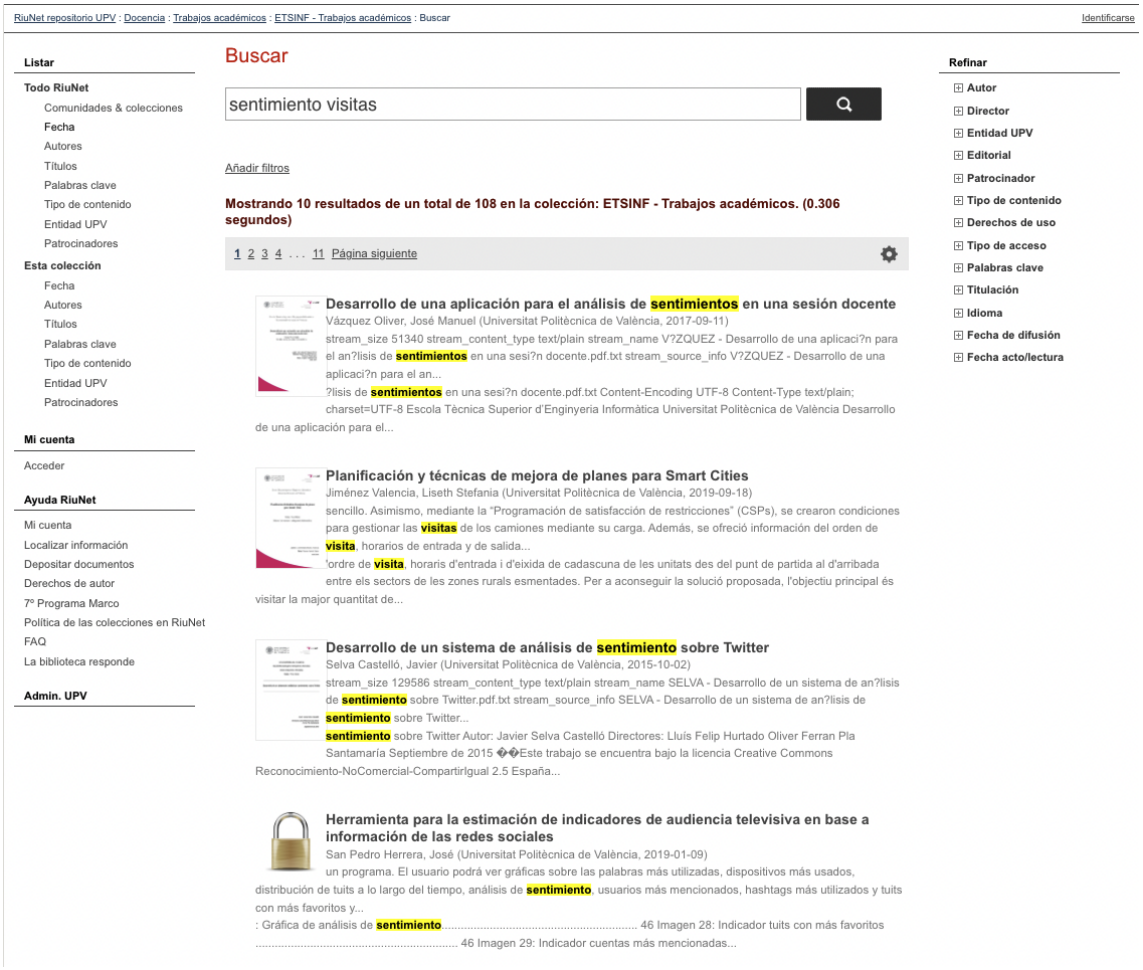


Figura 2.4: Búsqueda riunet. Fuente: Elaboración propia

Sin embargo, como podemos ver en la **Figura 2.4** y en la falta de resultados de *Google scholar* [7], el área de investigación que se pretende presentar en este proyecto, es ciertamente innovadora y que explora un área relativamente desconocida.

[7] https://scholar.google.es/

CAPÍTULO 3
Análisis del problema

3.1 Análisis energético

El análisis energético es un apartado fundamental para el medioambiente, ya que para la realización de este trabajo, han sido necesarias muchas horas de cómputo y si no se hubiera tenido en cuenta este punto para la realización del mismo, hubieran sido muchas más.

Gracias al uso de *threads* [1] en las operaciones realizadas en las llamadas a la *API* [3] de *Twitter* se ha conseguido reducir las horas medias de cómputo de 15 horas y 30 minutos a 5 horas y 45 minutos. En cuanto al procesamiento de texto con *transformers*, se ha conseguido reducir el tiempo de 6 horas y 35 minutos a 3 horas y 25 minutos.

Esto ha conseguido generar un impacto positivo para el medio ambiente, ya que reducimos el gasto de electricidad y aumentamos el tiempo de vida útil de los dispositivos electrónicos utilizados.

3.2 Análisis de riesgos

Los principales riesgos asociados a este proyecto, vienen por el uso de datos privados. Por un lado, se utilizan datos de la *API* [3] de *Twitter*, los cuales sólo son accesibles mediante acceso académico. Este acceso académico viene dado por un *token* único que identifica todas las llamadas hechas por el autor y cuya pérdida podría implicar la revocación de dicho acceso. Esto impediría la creación del dataset necesario para este trabajo. También hay que tener en cuenta que los datos accedidos mediante el uso académico no pueden ser compartidos y por tanto, se ha de tener extremo cuidado con su uso y posterior eliminación para no tener problemas en el futuro. La naturaleza de los datos no es sensible, ya que son *posts* públicos hechos por usuarios, los cuales han accedido previamente a las condiciones de uso de *Twitter*.

Por otro lado, los datos de *Safegraph* han sido cedidos también por uso exclusivo académico, ya que son datos de pago y no pueden ser filtrados en ningún caso. Cualquier filtración de estos datos supondría una pérdida económica im-

[1] https://realpython.com/intro-to-python-threading/

portante para la empresa y por tanto, se ha de tener mucho cuidado también con la manipulación, almacenaje y posterior eliminación de mencionados datos.

Por ello, los datos se han guardado siempre en disco duro y nunca se han subido a ningún tipo de almacenamiento en la nube, para minimizar el riesgo de robo de cuentas, y con ello, el robo de esta información tan preciada.

El mayor riesgo, por tanto sería la posible filtración de estos datos. Esto podría suceder, por ejemplo, tras la pérdida del ordenador portátil en el que se almacenan estos datos. Ante cualquier pérdida de esta información los pasos a seguir serían contactar con el equipo de soporte tanto de *Twitter* como *Safegraph* para poner en conocimiento esta pérdida de información tan valiosa y que puedan actuar en consecuencia.

3.3 Metodología

En el desarrollo de la metodología de este proyecto han influido dos metodologías muy destacadas, por un lado encontramos la metodolodgía *CRISP-DM* [2] y por otro lado la *Keep It Simple, Stupid* (*KISS*) [3]. Esta confluencia de metodologías es debido a que a pesar de que *CRISP-DM* es una metodología perfectamente aplicable a cualquier proyecto en ciencia de datos, en este caso hay que mantenerlo lo más simple posible y reducir ciertas de las fases definidas en la metodología, ya que este no es un proyecto comercial de ciencia de datos.

Sin embargo, se han definido tareas siguiendo la estructura de el *CRISP-DM* y las cuales forman parte de una de las siguientes partes:

- Fase I. Estudio y comprensión de los datos. Durante esta fase, se convierte el conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos. Por tanto, todas las tareas definidas tienen relación con la obtención y primera toma de contacto con los datos. Aquí encontramos tareas como la creación de llamadas automáticas a la *API* de *Twitter* o descarga y posterior filtrado de los datos de visitas de *Safegraph*.

- Fase II. Análisis de los datos y selección de características. Una vez realizada la primera toma de contacto con los datos, se busca pasar de datos en bruto a datos procesados y listos para la creación de modelos que permitan alcanzar los objetivos definidos. Ejemplos de las tareas de esta fase son el procesamiento de los datos de *Twitter* con *transformers* o la normalización de los datos de *Safegraph*

- Fase III. Modelado y evaluación. Durante esta fase, se seleccionan y aplican modelos que nos permitan alcanzar los objetivos. En este caso, las tareas correspondientes serían la búsqueda de correlación cruzada, la aplicación del test de Granger y la creación de un modelo de *OLS*.

[2] https://www.sngular.com/es/data-science-crisp-dm-metodologia/
[3] https://es.wikipedia.org/wiki/Principio_KISS

Como podemos ver, esta definición del *CRISP-DM* varía sustancialmente de la definición original. Esto es debido a que la naturaleza de este proyecto no es la de un proyecto comercial, por lo tanto hay fases que no son necesarias y no corresponden a este caso. Sin embargo, la definición de tareas es necesaria para la correcta consecución de cualquier proyecto y por ello la aplicación las metodologías *CRISP-DM* y *KISS*

CAPÍTULO 4
Preparación y comprensión de los datos

El objetivo principal de este capítulo es mostrar los conjuntos de datos que se han utilizado en este proyecto, así como su procedencia y obtención, como también la tecnología usada para todo ello.

4.1 Tecnologías usadas

Para este trabajo se ha utilizado el lenguaje de programación Python [1], uno de los lenguajes utilizados en el grado en Ciencia de Datos.

Las librerías de Python utilizadas para este proyecto se dividen en tres grupos:

- Librerías de cálculo matemático y estadístico. Aquí encontramos numpy [2], statsmodels [3], sklearn [4] y Hugging Face [5]
- Librerías de manejo de datos. Pandas [6].
- Librerías de visualización. Plotly [7] y matplotlib [8]

Por último, se destaca el uso del framework Streamlit [9], usado para la visualización en conjunto de los distintos gráficos generados a lo largo de este proyecto.

[1] https://es.wikipedia.org/wiki/Python
[2] https://numpy.org/
[3] https://www.statsmodels.org/stable/index.html
[4] https://scikit-learn.org/stable/
[5] https://huggingface.co/
[6] https://pandas.pydata.org/
[7] https://plotly.com/
[8] https://matplotlib.org/
[9] https://streamlit.io/

4.2 Introducción a los datos utilizados

Como se ha mencionado anteriormente, el objetivo de este proyecto es encontrar una relación entre el sentimiento de los habitantes de nueva york respecto a un país, el cual es extraído mediante el uso de transformers, frente a las visitas a negocios de restauración asociados a estos países. Para ello encontramos dos *datasets* claramente definidos:

- Datos de *tweets* procedentes de *Twitter*
- Datos de visitas a negocios procedentes de Safegraph

4.2.1. Datos procedentes de *Twitter*

El primer conjunto de datos procede de la conocida red social *Twitter*. Estos datos proceden directamente de su API accesible mediante diferentes niveles de acceso, tienendo en este caso el mayor de todos, el acceso académico [10]. Con este nivel de acceso, se permite acceder a todo el histórico de *tweets* almacenados en los servidores de la mencionada red social, permitiendo filtrar dichos *tweets* por palabras clave, fecha, localización del *tweet*, entre otros [11].

Para esta parte del proyecto, era necesario recoger *tweets* que hicieran referencia a un país en concreto. Con ese objetivo en mente, se decidió, primero, filtrar los tweets por aquellos que contuvieran el nombre del país o su gentilicio. Por ejemplo, para el caso de China, se buscaron tweets que contuvieran la palabra *China* o *Chinese*. Gracias a este sistema, podemos asignar a cada país una cierta cantidad de *tweets*, lo que nos permitirá conocer el sentimiento asociado.

Otra condición que se impuso a los tweets es que fueran escritos desde la ciudad de Nueva York, ya que, en este estudio, se busca centrarse en dicha ciudad.

Por último, para tener una muestra de datos representativa, se decidió escoger un periodo temporal de un año y medio, desde el 30/06/2020 hasta el 31/12/2021. La elección de este periodo viene marcada por un compromiso entre evitar el periodo marcado por el COVID19, el periodo disponible de los datos de visitas (explicados en la siguiente subsección) y la intención de minimizar el abultado tiempo de descarga de los *tweets*.

Por tanto, la obtención de este conjunto de datos se hizo mediante una serie de llamadas a la API, una llamada por día en el periodo del 30/06/2020 hasta 31/12/2021, limitando el número de tweets obtenidos por día a 500.

Limitar el número de *tweets* obtenidos por día a 500, es una decisión totalmente arbitraria, promovida por el abrumador tiempo de descarga de los datos por país. De media, para el periodo seleccionado y sin tener en cuenta el posterior proceso de análisis con *transformers*, el tiempo de descarga es de 5 horas y 43 minutos.

[10] https://developer.twitter.com/en/products/twitter-api/academic-research

[11] https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query

Entonces, teniendo en cuenta el periodo temporal de 548 días, del cual se recogen 500 tweets por día para cada uno de los países que se ha decido estudiar, hace un número de 274.000 tweets por país. Para una mayor claridad, la siguiente tabla muestra un ejemplo de como sería este *dataset*:

Tabla 4.1: Ejemplo de *tweets* relacionados con el país Italia

Text	Date
@tarrott @Marion_Geu I absolutely love italian food, especially pizza!!!	2020-06-30
@evistra stuck between wanting to spend the rest of my life in LA and wanting to run away to italy	2021-03-12
cinema in the outskirts of naples italy 1956 HTTPURL	2021-04-11
@TonyBartley968: @BigJupiter2 @Charlotte3003G Italy reclassified 90 % of their deaths as “non COVID19” about 7 weeks ago what a bunch of idiots!!	2021-06-24
@ActivistPost: Opposition to 5G is worldwide, specially in Italy. Cities and countries continue to take action to ban, delay, halt, and limit installation.	2021-01-11
@CapehartJ: A year ago tomorrow, I flew to Italy to spend the month in Rome. I always knew it was a blessing to be able to do that	2021-01-20
@honey_cna: I’ve never wanted to go to Italy so much	2021-08-02

Como paso final, y tal como se explicará en secciones futuras, sobre estos datos se aplicará un modelo de transformers que analice el sentimiento de cada *tweet* para con ello, ser capaz de comparar estos datos con los de visitas.

4.2.2. Datos procedentes de Safegraph

El segundo conjunto de datos procede de la empresa de Safegraph [12]. Los datos de visitas a comercios de Estados Unidos, se calculan a partir de la posición espacial de los dispositivos móviles que Safegraph rastrea a diario. Si un dispositivo móvil permanece más de cuatro minutos en el área asociada a un negocio, a éste se le atribuye una visita.

El conjunto de datos original tiene información de las visitas a negocios de la ciudad de Nueva York. Esto se puede ver más claramente en la siguiente tabla.

Tabla 4.2: Ejemplo de datos obtenidos de Safegraph

Date_range_start	Date_range_end	Location_name	City
2020-07-01	2020-08-01	Enterprise Parking Systems	New York
2020-09-01	2020-10-01	Alfredo’s	New York
2021-01-01	2021-01-02	Papa’s Panini	New York
2020-10-01	2020-11-01	Juancho’s	New York
2021-09-01	2021-10-01	RGR Family Thrift Store	New York
2020-08-01	2020-09-01	Jerez Jewelry	New York
2021-03-01	2021-04-01	House of Nails By Natty	New York
2021-06-01	2021-07-01	DSW Designer Shoe	New York

[12] https://www.safegraph.com/

Tabla 4.3: Continuación del ejemplo de datos obtenidos de Safegraph

Category_tags	Visits_by_day
Electronics	[0,0,1,1,0,1,1,0,1,0,1,0,1,0,0,1,2... (hasta fin de mes)]
Italian Food	[14,17,18,11,15,18,16,12,11,19,13,12... (hasta fin de mes)]
Italian Food	49,30,41,52,50,31,20,39,45,41,28,30... (hasta fin de mes)]
Mexican Food	14,17,18,11,15,18,16,12,11,19,13,12... (hasta fin de mes)]
Family Clothing Stores	11,14,11,17,21,38,16,10,9,12,23,14... (hasta fin de mes)]
Jewelry Stores	12,13,12,10,9,15,11,13,8,16,17,13... (hasta fin de mes)]
Nail Salons	18,11,14,15,12,10,10,8,10,9,6,8... (hasta fin de mes)]
Shoe Stores	20,18,19,21,15,21,26,22,18,29,23,26... (hasta fin de mes)]

CAPÍTULO 5
Preprocesamiento de los datos

A la hora de seleccionar los datos necesarios para las siguientes secciones de este proyecto, el primer filtrado que debía de realizarse era la selección de los países por los cuales se filtrarían tanto los conjuntos de datos de visitas como los de sentimiento.

Para ello, debían elegirse países que fueran interesantes no sólo desde el punto de vista de sentimiento en twitter, es decir, que tuvieran un cierto interés para los estadounidenses sino que también desde el punto de vista de la restauración, es decir, que la gastronomía fuera interesante para los neoyorquinos.

Con esta motivación, se eligieron los siguientes países:

- China
- Italia
- Mexico
- Rusia

5.1 Preprocesamiento de los datos de Twitter

Para poder aprovechar los datos obtenidos de la *API* de *Twitter* en anteriores secciones, era necesario pasar el *dataset* por un modelo basado en *transformers* para poder comparar el sentimiento con las visitas. Para ello, había que tener en cuenta que el conjunto de datos utilizado no estaba etiquetado, con lo cual se abrían dos opciones:

- Etiquetar a mano un número relevante de *tweets* para poder aplicar transfer learning y evaluar modelos basados en *transformers*.
- Utilizar modelos basados en *transformers* ya preentrenados en conjuntos de datos similares al presente, y evaluarlos en función a un dataset ejemplo.

Esta última opción fue la elegida y es posible gracias a que el conjunto de datos al uso (tweets en inglés) junto con la tarea de lenguaje natural que se quiere aplicar (análisis del sentimiento), es muy común y por tanto, hay numerosos *datasets* y *transformers* para elegir.

El dataset fue elegido en la página web Kaggle [1] y los *transformer* candidatos en Hugging Face [2]:

- Dataset: *Twitter-tweets-sentiment-dataset* [3]
- Transformer 1: bert-base-multilingual-uncased-sentiment [4]
- Transformer 2: bertweet-base-sentiment-analysis [5]
- Transformer 3: twitter-roberta-base-sentiment-latest [6]

El dataset en cuestión, cuenta con casi 28000 *tweets*, los cuales pueden tener un sentimiento positivo, negativo o neutro y con una distrubución de clases del 40, 31 y 29 porciento respectivamente.

Con el mencionado *dataset* y los modelos seleccionados, se decidió aplicar cada uno de los *transformers* sobre el conjunto de datos y calcular para cada modelo el *accuracy* a partir de la matriz de confusión resultante de la inferencia. Con esta métrica explicada en secciones anteriores, se decidió que modelo usar para el preprocesamiento.

Tabla 5.1: Resultados experimentación *transformers*

Transformer	Accuracy sentimiento positivo	Acc negativo	Acc neutro	Acc medio
bert-base-multilingual-uncased-sentiment	48.3	70.1	65.2	65.2
bertweet-base-sentiment-analysis	63.5	70.9	71.1	68.4
twitter-roberta-base-sentiment-latest	75.1	69.1	67.5	70.7

Como se puede observar, el modelo que mejor ha realizado la inferencia sobre el conjunto de datos (sin realizar *transfer learning* sobre él) es *twitter-roberta-base-sentiment-latest*. Para tomar la decisión se ha tenido en cuenta la media del accuracy por clase, teniendo en cuenta que en el problema de clasificación actual se tenía tres clases objetivo, positiva, negativa y neutra. Si alguno de los valores hubiera sido muy alto en una clase pero muy bajo en otra, se hubiera detectado un problema de inferencia en nuestro modelo. En nuestro caso, los tres modelos probados han conseguido discriminar perfectamente entre las tres clases

Sin embargo, antes de proceder a aplicar el modelo elegido al dataset creado de *tweets*, falta realizar una limpieza previa, común en tareas relacionadas con texto y el procesamiento del lenguaje natural. Siguiendo las recomendaciones de [13], se preprocesó los *tweets* de nuestro conjunto de datos.

Una vez realizado este paso, se procedió a realizar la inferencia sobre todo el conjunto de datos. Para poder comparar dos series temporales, era necesario transformar los resultados del modelo escogido.

Como se ha visto en secciones anteriores, los resultados podían ser **positivo**, **negativo** o **neutro**. Para convertir esto en datos numéricos, se transformó de la siguiente manera:

[1] https://www.kaggle.com/
[2] https://huggingface.co/
[3] https://www.kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset
[4] https://huggingface.co/bert-base-multilingual-uncased
[5] https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis
[6] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

Tabla 5.2: Transformación resultados *Transformer*

Transformer resultado	Transformación
Positivo	1
Negativo	-1
Neutro	0

Por último, como los datos de visitas son diarios, era necesario agrupar estos datos procedentes de *tweets* por días también. Para ello, se crearon tres variables:

- **Positive Sentiment**. En esta variable se guarda la suma de los datos de sentimiento positivos por día.
- **Negative Sentiment**. En esta variable se guarda la suma de los datos de sentimiento negativos por día.
- **Mean Sentiment**. En esta variable se guarda la suma de los datos de sentimiento por día.

Gracias a todos estos pasos, el dataset de sentimiento finalmente tiene la siguiente forma:

Tabla 5.3: Ejemplo conjunto de datos de sentimiento

Date	Mean sentiment	Positive sentiment	Negative sentiment	Country
2020-07-01	-168	74	-242	Italy
2020-07-02	-111	61	-172	Italy
2020-07-03	87	177	-90	Italy
2020-07-04	-93	125	-121	Italy
2020-07-05	21	142	-132	Italy
2020-07-06	-29	103	-136	Italy
2020-07-07	-33	103	-138	Italy
2020-07-08	-34	104	-141	Italy
2020-07-09	-9	132	-106	Italy
2020-07-10	100	178	-78	Italy
2020-07-11	33	143	-110	Italy
2020-07-12	14	141	-127	Italy

5.2 Preprocesamiento de los datos de Safegraph

Con el objetivo de este proyecto en mente, se decidió escoger los datos de negocios de restauración que tuvieran relación con el país que se decidiera analizar el sentimiento en *Twitter*, filtrando por la variable *category tags*.

Para ello, una vez filtrado por *category tags*, se agrupó por día para con ello tener las visitas diarias a negocios asociados a un país.

Teniendo en mente las consecuencias del COVID19 en los datos [15], se decidió normalizar los datos de visitas a negocios asociados a un país diviendo este

Tabla 5.4: Ejemplo datos de visitas después del preprocesamiento

Date	Visits	Country	Normalized Visits
2021-01-01	921	Italy	4.96
2021-01-02	627	Italy	4.11
2021-01-03	1100	Italy	3.92
2021-01-04	744	Italy	3.97
2021-01-05	1128	Italy	3.75
2021-01-06	1062	Italy	4.38
2021-01-07	832	Italy	4.2
2021-01-08	1104	Italy	3.78
2021-01-09	1140	Italy	3.71
2021-01-10	1177	Italy	3.74
2021-01-11	1316	Italy	3.98
2021-01-12	1272	Italy	3.91
2021-01-13	1206	Italy	4.28
2021-01-14	1096	Italy	4.51
2021-01-15	996	Italy	4.1
2021-01-16	1154	Italy	3.61
2021-01-17	1161	Italy	3.67

dato entre las visitas a los negocios de restauración de todos los países que registra Safegraph.

Con esto, se obtuvo el siguiente conjunto de datos:

CAPÍTULO 6
Conocimiento extraído y evaluación de modelos

Una vez vista la obtención de los datos de este proyecto, así como el preprocesamiento utilizado tanto en el conjunto de datos de visitas como en el de sentimiento, en esta sección se procede a explicar el análisis y experimentación realizado con estos datos.

Esta sección se dividirá en los siguientes apartados, los cuales aportarán una mayor claridad a los pasos seguidos:

- Búsqueda de correlación cruzada entre series temporales
- Test de Granger y *OLS* con el decalaje encontrado y estudio de métricas

6.1 Búsqueda de correlación cruzada entre series temporales

Primero, se realizarán pruebas de correlación desplazando las series en búsqueda de una mayor correlación que pueda indicar que una varibale está influenciada por otra. Una vez se encuentre el decalaje que maximice el valor de correlación cruzada, habrá que tener en cuenta que esta correlación cruzada no implica que una serie esté influyendo en la otra, por tanto, se realizará un test de causalidad de Granger con el decalaje elegido, y se estudiarán los resultados del mismo.

Desde el primer momento, antes de realizar esta búsqueda de correlación cruzada, se pudo observar como al comparar visualmente los datos de sentimiento de un país (cualquiera de los escogidos) con los datos de visitas, especialmente en éstos últimos, había un comportamiento inusual, el cual se puede ver más claramente en la **Figura 6.1**:

En los los datos de la **Figura 6.1**, así como en el del resto de países (disponibles en el **Apéndice C**), se aprecian claramente dos problemas. Por un lado, una alta variabilidad en ambas series en cortos periodos de tiempo, con claros picos en días concretos. Por otro, se aprecia la recuperación ocurrida durante el periodo de verano y otoño del año 2020, las posteriores restricciones que se produjeron

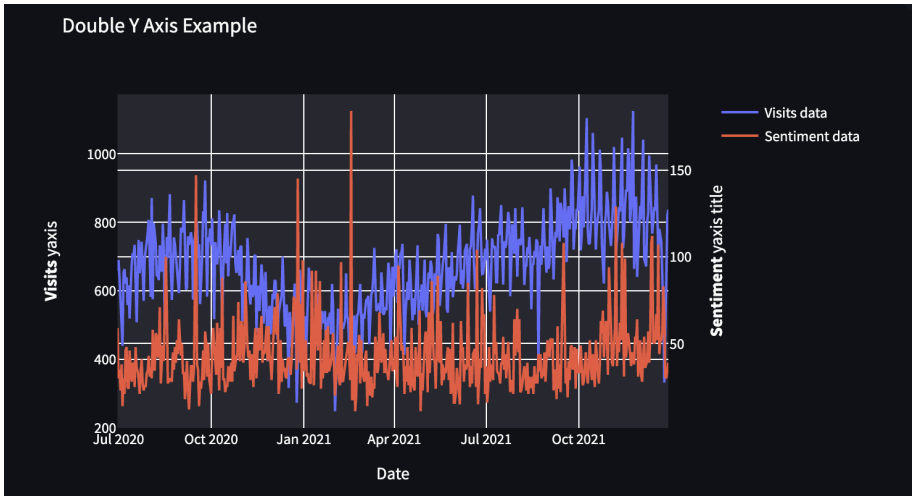


Figura 6.1: Datos de sentimiento y visitas correspondientes con México. Fuente: Elaboración propia

durante el periodo de inicios de 2021 y la final recuperación de las visitas durante todo el resto de 2021 a causa de la vacunación contra el COVID19 [15].

Identificados estos problemas, se propusieron las siguientes soluciones:

- Para la alta variabilidad, se aplicaron medias móviles [1] con ventanas de 7 días para los datos de visitas como los de sentimiento.
- Para reducir los efectos del COVID-19 y como se anticipó en el apartado de preproceso, se decidió utilizar datos de visitas normalizados por los valores del resto de países.

Las mismas dos series, con los cambios mencionados, tienen mejor aspecto en la siguiente gráfica:

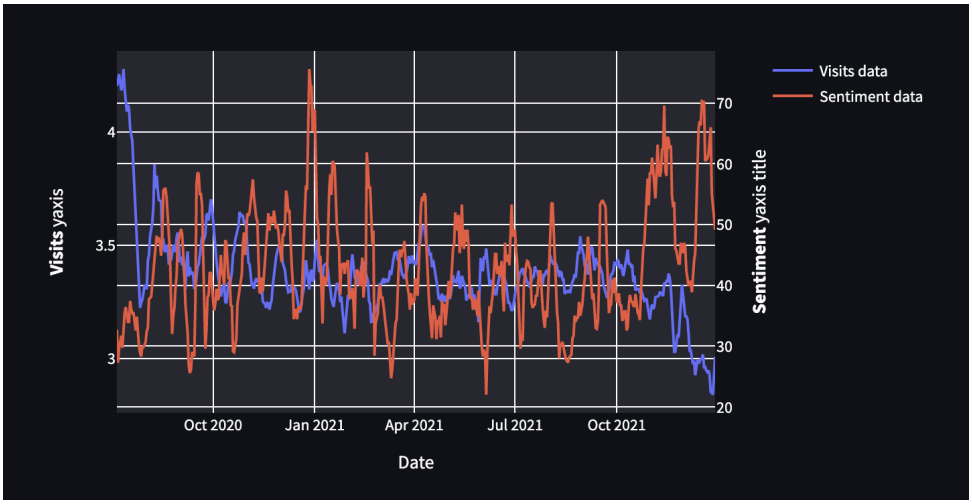


Figura 6.2: Datos de sentimiento y visitas suavizados correspondientes con México. Fuente: Elaboración propia

[1] https://es.wikipedia.org/wiki/Media_m%C3%B3vil

En este momento, se decidió estudiar tambien otras agrupaciones temporales que podrían ser de interés, como las agrupaciones por semana, quincenales y mensuales, las cuales se pueden apreciar en las siguientes gráficas: (solo se muestran de un país por motivos de espacio)

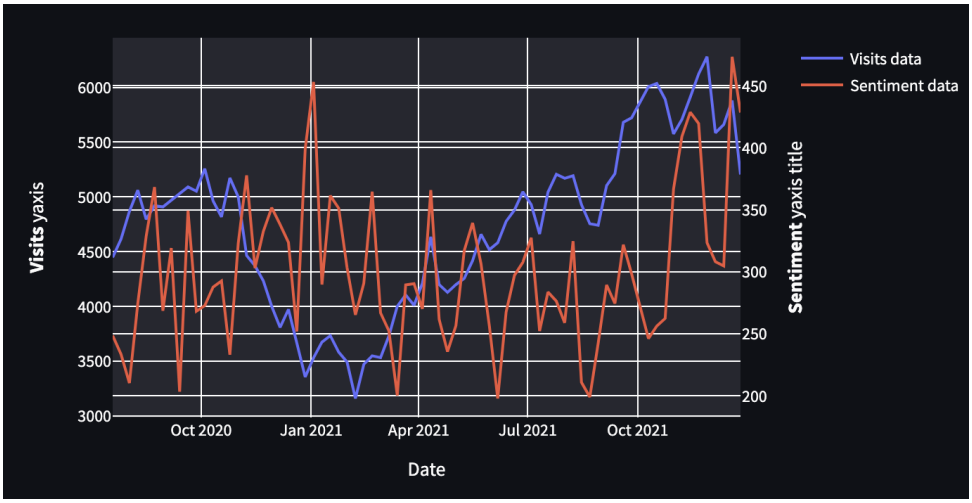


Figura 6.3: Datos de sentimiento y visitas semanales correspondientes con México. Fuente: Elaboración propia

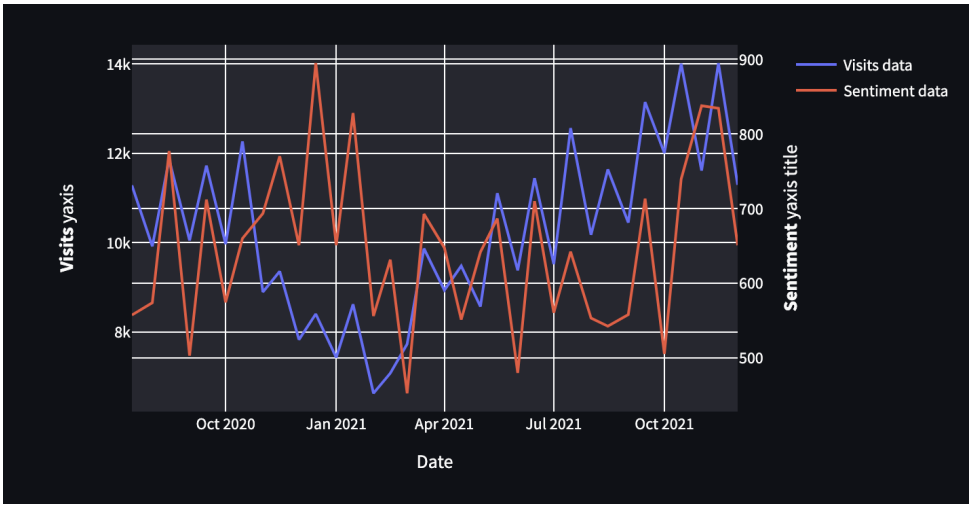


Figura 6.4: Datos de sentimiento y visitas quincenales correspondientes con México. Fuente: Elaboración propia

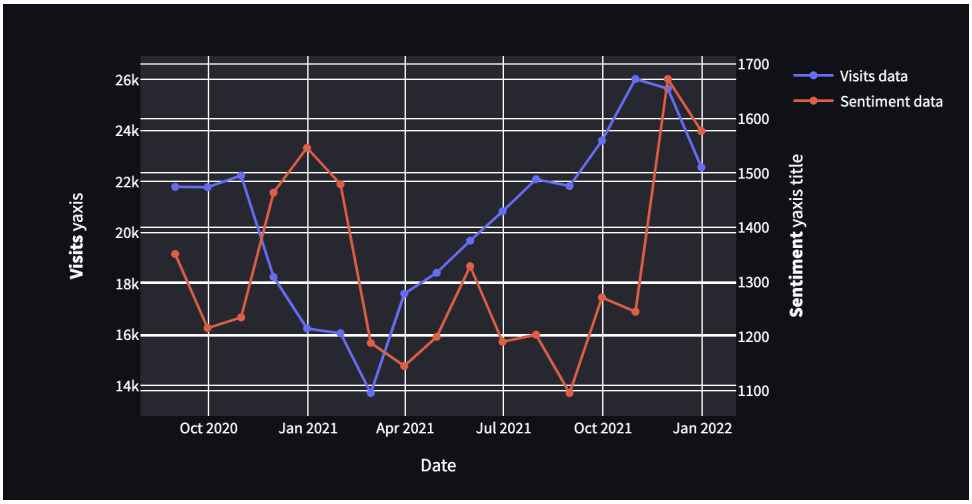


Figura 6.5: Datos de sentimiento y visitas mensuales correspondientes con México. Fuente: Elaboración propia

Con esto, se decidió observar los resultados de buscar la correlación cruzada entre todas las combinaciones de sentimiento (éste podía ser positivo, negativo o medio), países (China, Italia, Mexico, Rusia)y agrupaciones temporales (diaria, semanal, quincenal y mensual). También hay que tener en cuenta que se busca relacionar los datos de sentimiento con los de visitas, desplazando positivamente estos últimos para estudiar si la correlación aumenta.

El número de desplazamientos se escogió en función de la agrupación temporal de la siguiente forma:

- En la agrupación diaria, se decidió probar un decalaje de máximo 30 días
- En la agrupación semanal, se decidió probar un decalaje de máximo 4 semanas
- En la agrupación quincenal, se decidió probar un decalaje de máximo 2 quincenas
- En la agrupación mensual, se decidió probar un decalaje de máximo 1 mes

El objetivo de utilizar la correlación cruzada es encontrar el desplazamiento (*lag*) que maximice el coeficiente de correlación. Los resultados de este experimento se presentan en la siguientes tablas:

Tabla 6.1: Coeficiente de correlación con los datos de visitas normalizados y sentimiento positivo con diferentes agrupaciones con decalaje 0

País	Diario	Semanal	Quincenal	Mensual
China	-0.30	-0.31	-0.40	-0.43
Italia	0.14	0.17	0.45	0.31
Mexico	-0.30	-0.31	-0.37	-0.40
Rusia	0.10	0.11	0.18	0.12

Tabla 6.2: Coeficiente de correlación máximos con decalaje a partir del cálculo de la correlación cruzada con los datos de visitas normalizados y sentimiento positivo con diferentes agrupaciones

País	Diario	Semanal	Quincenal	Mensual
China	-0.34	-0.34	-0.50	-0.54
Italia	0.23	0.48	0.61	0.37
Mexico	-0.33	-0.36	-0.56	-0.63
Rusia	0.16	0.15	0.22	0.15

Tabla 6.3: Decalaje óptimo a partir del cálculo de la correlación cruzada con los datos de visitas normalizados y sentimiento positivo con diferentes agrupaciones

País	Diario	Semanal	Quincenal	Mensual
China	7	0	1	0
Italia	21	2	0	0
México	13	4	1	1
Rusia	9	1	2	0

Solo se presentan los resultados de la correlación cruzada en la que se usan los datos de sentimiento positivo, ya que son los más interesantes para el análisis y los que mejor resultado han dado. Para la elección del número de desplazamientos óptimo se ha hecho con la ayuda de gráficos como el siguiente:

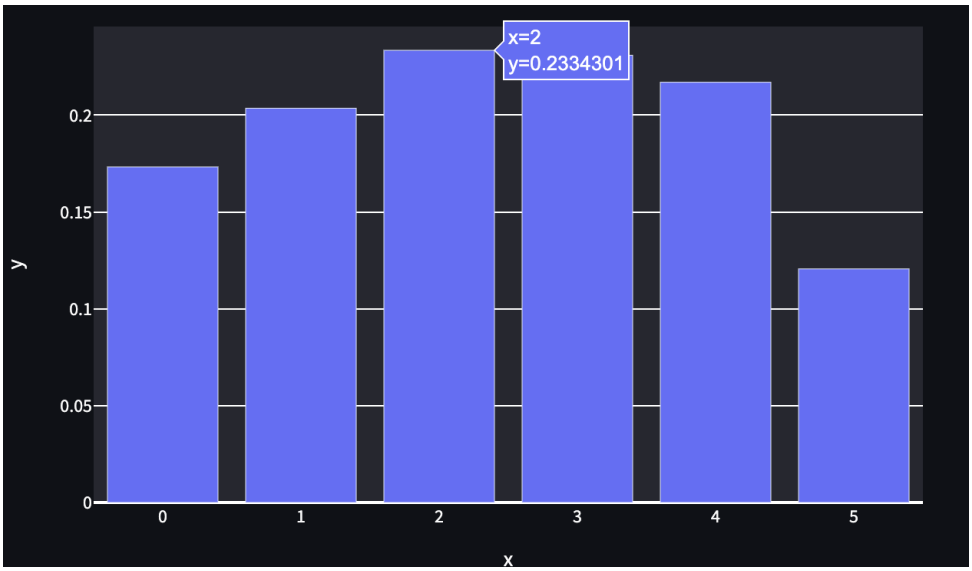


Figura 6.6: Resultados correlación de Italia con datos semanales con decalaje positivo. Fuente: Elaboración propia

Los resultados, como se puede observar, no son los esperados. Con la excepción de Italia, el cual destaca especialmente en la agrupación quincenal, en la que se consigue un 0.61 de correlación con decalaje 0. Incluso en este caso, en el que se obtiene un coeficiente de correlación razonable, se observa que el decalaje es 0, lo que lleva a la inevitable conclusión de que no se pueden usar valores previos de sentimiento para predecir las visitas. Aunque sí que se podría afirmar, al menos

en este caso concreto, que la dirección del sentimiento y de las visitas es al menos, similar (ya que solo tiene un 0.61 de correlación).

Por otro lado, los resultados de los demás países ponen en relevancia la falta de correlación entre el sentimiento positivo de los usuarios de *Twitter* y las visitas a los negocios de dichos países. Y aún peor, son los casos de México y China, en los cuales claramente se ve como existe una correlación negativa con los datos de sentimiento positivo, lo que nos llevaría a concluir en esos casos, que a mayor *tweets* con carga positiva respecto a estos países, menor número de visitas a negocios asociados.

Otra evidencia que estos resultados demuestran, es la gran diferencia que hay entre los datos de los distintos países, lo cual se ve reflejado tanto en las tablas 6.1 y 6.2, como en los resultados del decalaje de la tabla 6.3, en los que parece que estemos trabajando con variables totalmente distintas y en las que es imposible encontrar algún patrón común, como hubiera sido el resultado óptimo de este proyecto.

A pesar de lo que suponen estos resultados, y de los cuales se discutirá más en detalle en el apartado de conclusiones, se decidió realizar el test de Granger y si procede, el modelo de mínimos cuadrados ordinarios sobre el mejor resultado obtenido (con decalaje distinto a 0), para ejemplificar lo que se quería haber hecho en el caso de que los resultados hubieran sido mejores. Para ello, se escoge el segundo mejor resultado obtenido, que es el **0.48** de correlación obtenido con los datos semanales de Italia con sentimiento positivo y decalaje 2.

6.2 Test de Granger y *OLS* con el decalaje encontrado y estudio de métricas

La idea de utilizar el test de Granger proviene de la necesidad de no caer en la falacia *Post hoc ergo propter hoc* (correlación no implica causalidad), como se ha mencionado antes. Para ello se introduce el concepto de test de Granger.

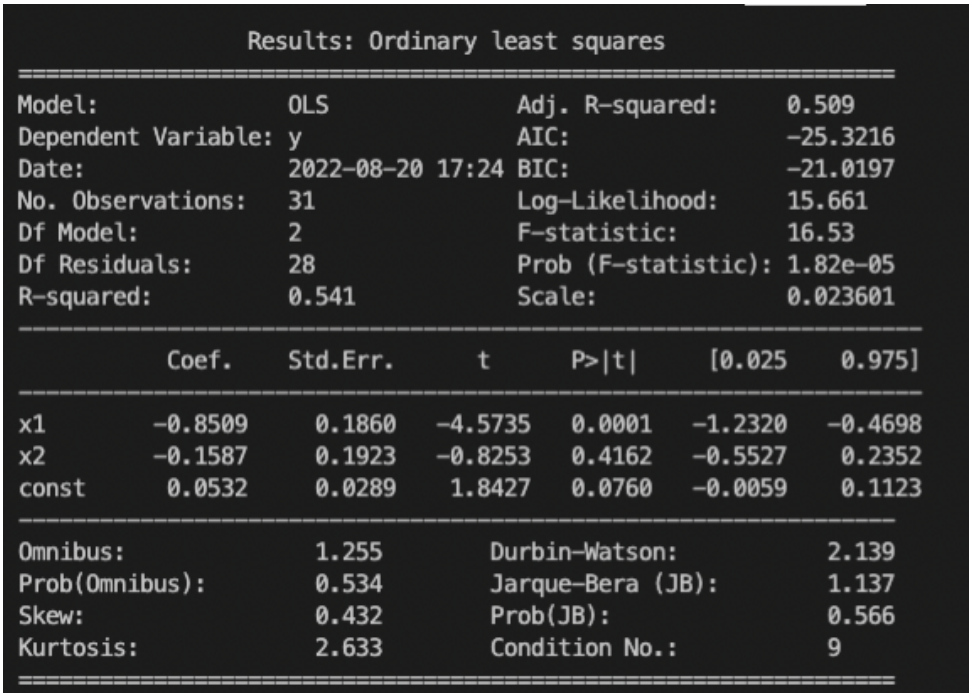
Con el caso escogido en la sección anterior de Italia, se realiza un test de causalidad de Granger con decalaje 2, respetando los resultados de la **tabla 6.3**. Los resultados se pueden ver en la siguiente figura:

```
Granger Causality
number of lags (no zero) 2
ssr based F test:         F=4.4099  , p=0.0224  , df_denom=26, df_num=2
ssr based chi2 test:   chi2=10.5159 , p=0.0052  , df=2
likelihood ratio test: chi2=9.0548  , p=0.0108  , df=2
parameter F test:         F=4.4099  , p=0.0224  , df_denom=26, df_num=2
```

Figura 6.7: Resultado test de Granger. Fuente: Elaboración propia

En la implementación que se ha utilizado del test de Granger, implementada en la librería *statsmodels* [2], se utilizan cuatro tests estadísticos para comprobar si es posible o no rechazar la hipótesis nula (que el conjunto de datos de sentimientos no tiene relación con el conjunto de datos de visitas). Como se puede observar, los cuatro tests tienen un *p-valor* menor a 0.05, con lo que se puede afirmar, con una confianza del 95 porciento, que una si que causa la otra.

Para confirmar estos resultados, veamos el modelo de mínimos cuadrados ordinarios [2] realizado después de los positivos resultados del test de Granger:



Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.509
Dependent Variable:	y	AIC:	-25.3216
Date:	2022-08-20 17:24	BIC:	-21.0197
No. Observations:	31	Log-Likelihood:	15.661
Df Model:	2	F-statistic:	16.53
Df Residuals:	28	Prob (F-statistic):	1.82e-05
R-squared:	0.541	Scale:	0.023601

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
x1	-0.8509	0.1860	-4.5735	0.0001	-1.2320	-0.4698
x2	-0.1587	0.1923	-0.8253	0.4162	-0.5527	0.2352
const	0.0532	0.0289	1.8427	0.0760	-0.0059	0.1123

Omnibus:	1.255	Durbin-Watson:	2.139
Prob(Omnibus):	0.534	Jarque-Bera (JB):	1.137
Skew:	0.432	Prob(JB):	0.566
Kurtosis:	2.633	Condition No.:	9

Figura 6.8: Resultado test del modelo de cuadrados mínimos ordinarios. Fuente: Elaboración propia

El modelo creado tendría la siguiente estructura:

$$y_t = \beta_0 + \beta_1 * x_{t-1} + \beta_2 * x_{t-2}$$

Siendo y_t las visitas dado un momento t (en este caso una semana t) y x_{t-1}, x_{t-2} los valores de las dos semanas anteriores de sentimiento.

Los resultados del modelo de mínimos cuadrados ordinarios provistos por la librería *statsmodels* son sorprendentemente buenos, como podemos observar en las métricas *R-squared* y *Adjusted R-squared*, las cuales se pueden interpretar como que el modelo es capaz de capturar aproximadamente el 50 porciento de la variabilidad de los datos de visitas. En cuanto a los parámetros del modelo, vemos como solo el coeficiente de x_{t-1} y β_0 parecen ser relevantes para el modelo de mínimos cuadrados ordinarios.

Esta experimentación es solo un pequeño ejemplo práctico de lo que se hubiera podido realizar si los resultados de la correlación cruzada hubieran sido

[2] https://www.statsmodels.org/stable/index.html

razonables. Hay que recordar que debido a los resultados de las tablas *6.1*, *6.2* y *6.3*, no hay margen para realizar más experimentación, al menos no con los datos utilizados. Este experimento con los datos de Italia, permite vislumbrar lo que se hubiera tenido que realizar si los resultados hubieran sido los esperados.

CAPÍTULO 7
Conclusiones

En este trabajo se ha propuesto relacionar y posteriormente predecir las visitas a negocios con una clara vinculación a una cultura con el sentimiento que tienen los habitantes de Nueva York acerca de los países de origen de estos negocios. Se considera que aunque no se haya conseguido demostrar dicha relación, ni encontrar patrones similares entre los datos de los distintos países, se ha intentado alcanzar los objetivos de forma razonable y además, se ha conseguido mediante este proyecto, sentar las bases para futuros proyectos que quieran recoger el testigo y aprender de los errores y limitaciones que han llevado a los resultados de este trabajo.

Sin embargo, a lo largo de este proyecto se han completado satisfactoriamente diversas tareas:

- Crear un conjunto de datos de sentimiento a partir del acceso académico de la red social *Twitter*, con lo que ello ha conllevado, descarga, limpieza y transformación de datos descargados de la *API* de *Twitter*, investigación sobre el procesamiento del lenguaje natural, *aprendizaje profundo* y *transformers*.
- Uso del lenguaje de programación *Python* para el desarrollo del trabajo, así como el aprendizaje de las librerías relacionadas necesarias en las fases de desarrollo del proyecto.
- Investigación sobre el tratamiento de series temporales, así como la posible existencia de relación entre ellas y la creación de modelos predictivos de unas series a partir de otras.
- Se ha intentado relacionar series temporales, utilizando las técnicas apropiadas, tanto de preproceso como estadísticas, como la correlación cruzada, test de Granger y por último la predicción con mínimos cuadrados ordinarios.

Aunque los resultados en definitiva no han sido los esperados, se considera que este trabajo no ha sido un fracaso en absoluto. Ya que, en este proyecto, se ha intentado abrir una línea de investigación interesante, en la que se ha intentado ir más allá de la relación consumidor-empresa, sino que se ha buscado intentar descubrir una motivación más allá por parte del consumidor y ver, si a partir del

sentimiento general hacia una cultura en un núcleo urbano podía llegar a afectar a las visitas de los negocios percibidos como pertenecientes a dicha cultura. Sin duda, esta revelación podría haber aportado mucha información tanto a empresas, como a investigadores, como a políticos. Por tanto, en este trabajo se ha intentado aportar un granito de arena en un tema tan interesante como complejo como es el estudio del comportamiento del consumidor, trabajo que se espera pueda aportar luz y guía para futuras investigaciones relacionadas o que intenten prosperar donde este trabajo no pudo.

Por último, para todo aquel que tenga interés en el desarollo e implementación de lo expuesto en este trabajo, puede encontrar el código desarrollado en [1]. Por desgracia, la replicabilidad de los resultados de este proyecto es complicada ya que tanto los datos procedentes de *Twitter* como los de *Safegraph* son de índole privada, lo que implica que solo se puede compartir el código desarrollado, habiendo que eliminar todos los datos utilizados en este trabajo.

7.1 Limitaciones encontradas

Debido a la naturaleza experimental de este proyecto, han habido numerosas limitaciones que se detallarán a continuación: La primera limitación, y sin duda, la que ha llevado a los malos resultados expuestos en el **capítulo 6**, es la calidad de los datos obtenidos. Todos los proyectos de ciencia de datos dependen de la calidad de los conjuntos de datos utilizados, y por más técnicas de preprocesado que se puedan aplicar a los datos, si estos no son buenos, es complicado (y en muchas ocasiones directamente imposible) solucionarlo.

En este caso, se han encontrado limitaciones en los dos conjuntos de datos utilizados, pero debido a la difícil obtención de datos de esta índole, no ha habido otra opción que usarlos e intentar extraer lo mejor de ellos. En el conjunto de datos de *Twitter*, debido a la limitación de datos de tiempo, ha sido imposible poder filtrar los datos para eliminar los mensajes escritos por *bots* [2] o que contengan contenido *spam*. Este es un tema de actualidad, tratado en diversos trabajos académicos como [14]. Esto sin duda, empeora sustancialmente la calidad del conjunto de datos. Según el repositorio de datos de la universidad de Harvard [3], "Aproximadamente el 2 porciento de los tweets tienen geolocalización". Esto implica que a la hora de generar el conjunto de datos de *Twitter*, se ha tenido que descartar desde un primer momento el 98 porciento de los *tweets*, a causa de esta limitación. Esto también ha afectado a la calidad de las *queries* que se han podido lanzar sobre la *API* de *Twitter*, ya que contando con este hándicap, resultaba imposible afinar más.

También en relación con los datos procedentes de *Twitter*, encontramos otra limitación más. Esta limitación proviene del hecho de que los datos se han obtenido directamente de *Twitter*, es decir, no han sido etiquetados por ninguna persona. Esto, como se mencionó en el **capítulo 4**, supone un problema, ya que si se hubiera utilizado *transfer learning*, los resultados de aplicar modelos basados

[1] https://github.com/n4choo/tfg-upv
[2] https://www.xataka.com/empresas-y-economia/twitter-no-sabe-cuantos-bots-hay-sus-61-millones-u
[3] https://dataverse.harvard.edu/dataverse/geo-tweets

en *transformers* hubieran sido previsiblemente mucho mejores. Sin embargo, debido a la falta de recursos de este proyecto, era totalmente inviable etiquetar un número de *tweets* superior al millón.

En los datos provistos por *Safegraph*, hay también una limitación a destacar. Tal y como se comentó en apartados anteriores, estos datos, aunque intentan parecerse lo máximo posible a los datos reales de las visitas de los negocios, no son exactamente iguales. Los datos se calculan a partir de la geolocalización de móviles y por tanto, aunque después se utilicen modelos para aproximar el número real de visitas a negocios, el margen de error puede seguir siendo importante. Otra limitación es el número de negocios que la empresa *Safegraph* tiene en su base de datos, que aunque es numeroso, no cuenta con todos los negocios de la ciudad de Nueva York, lo que aleja estos datos un poco más de la realidad de los datos de visitas a negocios de esta ciudad.

La última limitación viene de la mano de la falta de recursos económicos para poder descargar más datos de *Twitter* y utilizar *Transformers* con mayor número de parámetros. Con un equipo más potente o servidor en la nube en el que poder lanzar tareas sin tener la preocupación de tener el dispositivo en local encendido, hubiera sido posible dedicar más tiempo a tener unos *datasets* más grandes y de mejor calidad.

7.2 Relación del trabajo desarrollado con los estudios cursados

La completa realización ha sido posible a todos los conocimientos aprendidos a lo largo de los cursos académicos cursados del grado en ciencia de datos. En más detalle, la relación del trabajo desarrollado con los estudios cursados es:

- La descarga de los datos de *Twitter*, mediante llamadas a la *API*, utilizando multiprocesos tiene relación con las asignaturas "Adquisición y transmisión de datos" y "Programación".
- Todo el desarrollo del código en el lenguaje de programación *Python*, así como las estrategias usadas para la reducción del tiempo de cómputo mencionadas en el **capítulo 3**, está vinculado con las asignaturas "Programación", "Estructuras de Datos y Algoritmos" y "Algorítmica"
- El procesado de los datos de texto, así como la idea de utilizar el análisis de sentimiento mediante modelos basados en transformers, tienen relación con los contenidos de las asignaturas "Procesamiento del Lenguaje Natural" y "Técnicas Escalables en Aprendizaje Automático".
- El análisis de riesgos tiene relación con la asignatura "Marco profesional, legal y deontológico "
- La parte de experimentación con series temporales, así como su preproceso y tratamiento, está relacionada con las asignaturas "Evaluación, Despliegue y Monitorización de Modelos", "Comportamiento Económico y Social" y Modelos Estadísticos para la Toma de Decisiones I y II

- El despliegue tiene relación con los contenidos de la asignatura "Infraestructuras para el Procesamiento de Datos"

7.3 Trabajo futuro

Las propuestas de trabajo futuro que surjen a raíz de este proyecto, están relacionadas con las limitaciones encontradas en el **capítulo 8**. Para poder cumplir el objetivo propuesto en este trabajo y por tanto, poder relacionar de forma efectiva datos de sentimiento de los habitantes de una ciudad sobre un país, con los datos de visitas a negocios relacionados con dicho país, se propone un cambio en los datos de sentimiento, ya que seguramente, estos datos han sido los que con mayor seguridad han podido comprometer los objetivos de este trabajo.

Por tanto, se propone utilizar otras vías de obtención de los datos de sentimiento, ya sea mediante encuestas a pie de calle o de forma online, datos procedentes de otra red social, o datos provenientes de la misma red social pero con otra forma de obtener la geolocalización de los usuarios. Ya que con la forma actual, se descartaban aproximadamente un 98 porciento de los *tweets* disponibles.

Otra posibilidad, surgiría con el uso de modelos de análisis de sentimiento mejores al utilizado, principalmente utilizando *transfer learning* con un *dataset* apropiadamente etiquetado y con el *fine tuning* [4] correspondiente.

Por otro lado, se anima a explorar la posibilidad de que los países escogidos no hayan sido los más apropiados y quizás la elección de otros países sea más adecuada, o incluso, agrupar estos países en culturas ya que es posible que se haya afinado demasiado intentando tratar a nivel de país. Además de utilizar a ser posible datos más recientes, por un lado para evitar los efectos colaterales del COVID-19, y por otro para poder estudiar los posibles efectos de la actual situación Rusia-Ucrania en los datos.

Por último, se destaca la posibilidad de utilizar otros tests para demostrar la causalidad entre dos series, ya que este es tópico complejo, y el test de Granger utilizado no es la única manera de probar causalidad. Sin embargo, este test si que es muy utilizado en el área de la econometría, de ahí su elección.

[4] https://d2l.ai/chapter_computer-vision/fine-tuning.html

Bibliografía

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L·L Polosukhin, I. Attention is all you need. *Advances In Neural Information Processing Systems*. **30** (2017)

[2] Dismuke, C. & Lindrooth, R. Ordinary least squares. *Methods And Designs For Outcomes Research*. **93** pp. 93-104 (2006)

[3] API (Interfaz de programación de aplicaciones), https://es.wikipedia.org/wiki/Interfaz_de_programaci%C3%B3n_de_aplicaciones.

[4] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*. **61** pp. 85-117 (2015)

[5] Bozinovski, S. & Fulgosi, A. The influence of pattern similarity and transfer learning upon training of a base perceptron b2. *Proceedings Of Symposium Informatica*. **3** pp. 121-126 (1976)

[6] Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. último acceso: 15 de junio de 2022. Consultado en https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

[7] OpenAI GPT-3: Everything You Need to Know. último acceso: 21 de agosto de 2022. Consultado en https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/

[8] Granger Causality Test último acceso: 21 de agosto de 2022. Consultado en https://en.wikipedia.org/wiki/Granger_causality

[9] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv Preprint ArXiv:1910.13461*. (2019)

[10] Ranjan, S., Sood, S. & Verma, V. Twitter sentiment analysis of real-time customer experience feedback for predicting growth of Indian telecom companies. *2018 4th International Conference On Computing Sciences (ICCS)*. pp. 166-174 (2018)

[11] Nguyen, T., Shirai, K. & Velcin, J. Sentiment analysis on social media for stock movement prediction. *Expert Systems With Applications*. **42**, 9603-9611 (2015)

[12] Li, X., Xie, H., Chen, L., Wang, J. & Deng, X. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*. **69** pp. 14-23 (2014)

[13] Hemalatha, I., Varma, G. & Govardhan, A. Preprocessing the informal text for efficient sentiment analysis. *International Journal Of Emerging Trends Technology In Computer Science (IJETTCS)*. **1**, 58-61 (2012)

[14] Alothali, E., Zaki, N., Mohamed, E. & Alashwal, H. Detecting social bots on twitter: a literature review. *2018 International Conference On Innovations In Information Technology (IIT)*. pp. 175-180 (2018)

[15] Bobeica, E. & Hartwig, B. The COVID-19 shock and challenges for time series models. (ECB Working Paper,2021)

APÉNDICE A
Objetivos del desarrollo sostenible (ODS)

ANEXO

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenible	**Alto**	**Medio**	**Bajo**	**No procede**
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**				X
ODS 4. **Educación de calidad.**				X
ODS 5. **Igualdad de género.**				X
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**			X	
ODS 9. **Industria, innovación e infraestructuras.**				X
ODS 10. **Reducción de las desigualdades.**	X			
ODS 11. **Ciudades y comunidades sostenibles.**				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**			X	
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**				X

ETS Enginyeria Informàtica
Camí de Vera, s/n, 46022, València
T +34 963 877 210
F +34 963 877 219
etsinf@upvnet.upv.es - www.inf.upv.es

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

Este proyecto guarda estrecha relación con los objetivos 8, 9, 10 y 13 (trabajo decente y crecimiento económico, industria, innovación e infraestructuras, reducción de las desigualdades y acción por el clima, respectivamente) de los Objetivos de Desarrollo Sostenible de la Organización de las Naciones Unidas.
Esto es debido a que en este proyecto, se ha llegado a intentar predecir los cambios en las visitas que podrían afectar a negocios asociados a un colectivo. Con esta información, podría combatirse posibles efectos económicos adversos en estos negocios y luchar por tanto contra la desigualdad. Esto tendría relación también con el objetivo 8, ya que se estaría también trabajando en el crecimiento económico igualitario entre negocios.
Por último, como se ha mencionado en el **capítulo 3**, se ha intentado concienciar del gasto de electricidad y la reducción de vida útil que pueden sufrir los dispositivos si no se tiene especial cuidado en la forma en la que se llevan a cabo estos proyectos, lo cual tiene especial relevancia con el objetivo 13, acción por el clima.

APÉNDICE B
Estructura del código creado

El código desarrollado en *Github* [1] tiene la siguiente estructura:

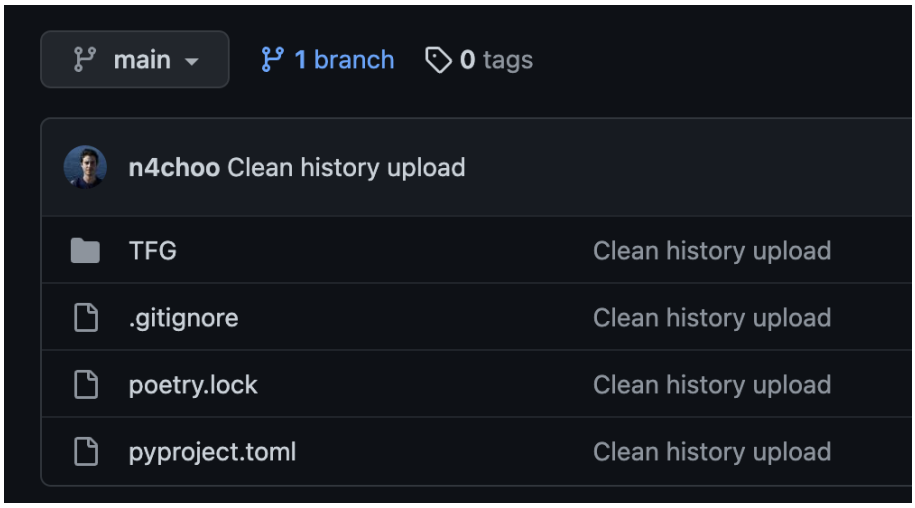


Figura B.1: Estructura del código en *Github*. Fuente: Elaboración propia

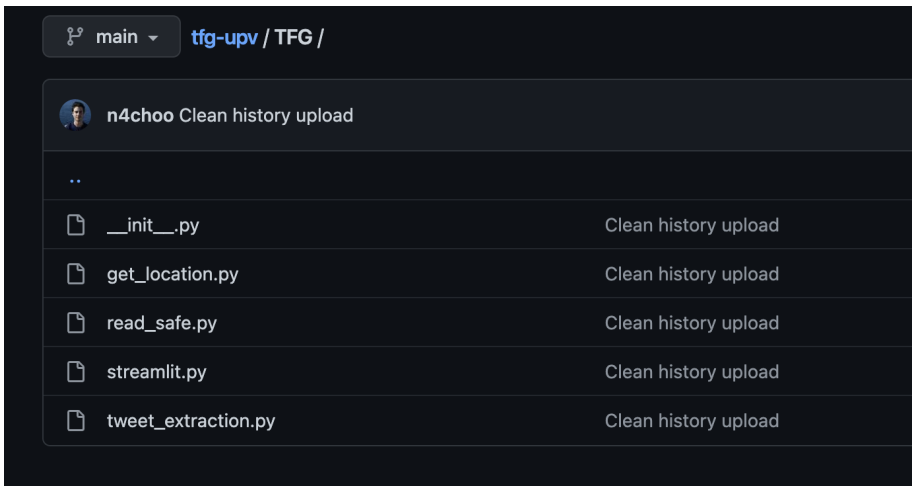


Figura B.2: Estructura del código en *Github* parte 2. Fuente: Elaboración propia

[1] https://github.com/n4choo/tfg-upv

- Replicabilidad del entorno. Para la replicabilidad del entorno se ha utilizado poetry [2], esta herramienta nos permite tener en un entorno controlado todas las librerías y paquetes necesarios para el desarrollo de código en *Python* así como las dependencias que surgen entre ellos. A este apartado corresponden los archivos *poetry.lock* y *pyproject.toml*. Con el código descargado y con *poetry* instalado, con el comando *poetry install* se creará una réplica del entorno del autor en el ordenador del lector.

- Descarga y procesado de los datos de *Twitter* y *Safegraph* En los archivos *read_safe.py* y *tweet_extraction.py* podemos encontrar la obtención y el proceso de inferencia y limpieza de los datos utilizados en este proyecto.

- *Streamlit*. En el archivo *streamlit.py* se puede encontrar el código necesario para la visualización interactiva de los datos de *Twitter* y *Safegraph* utilizada a lo largo del proyecto. Para levantar el servicio, deberá de utilizarse el comando *poetry run streamlit run TFG/streamlit.py*. En este archivo también se agregan los datos de series temporales, se realiza el test de Granger, el *OLS* y la correlación cruzada.

[2] https://python-poetry.org/

APÉNDICE C
Figuras adicionales

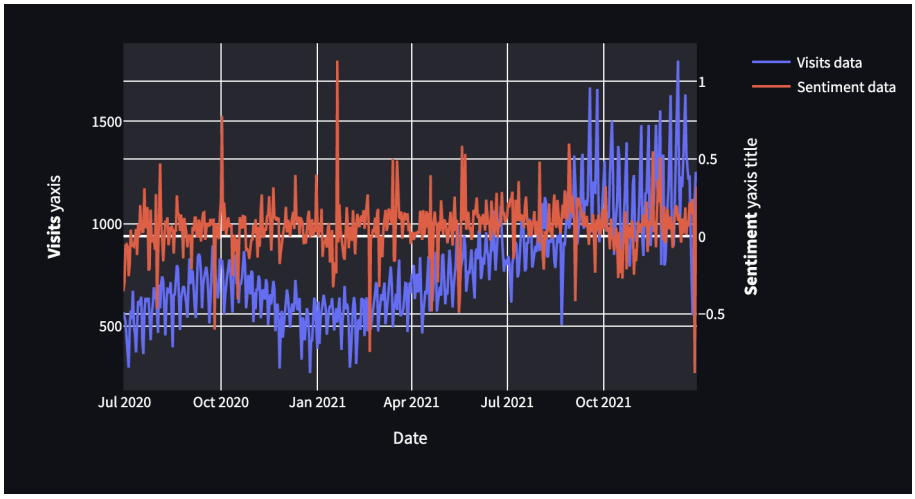


Figura C.1: Datos de sentimiento y visitas correspondientes con Italia. Fuente: Elaboración propia

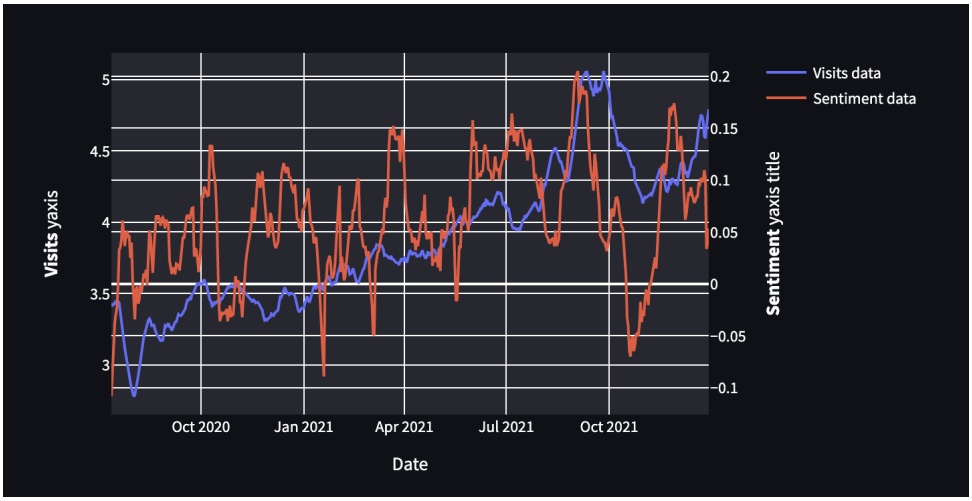


Figura C.2: Datos de sentimiento y visitas suavizados correspondientes con Italia. Fuente: Elaboración propia

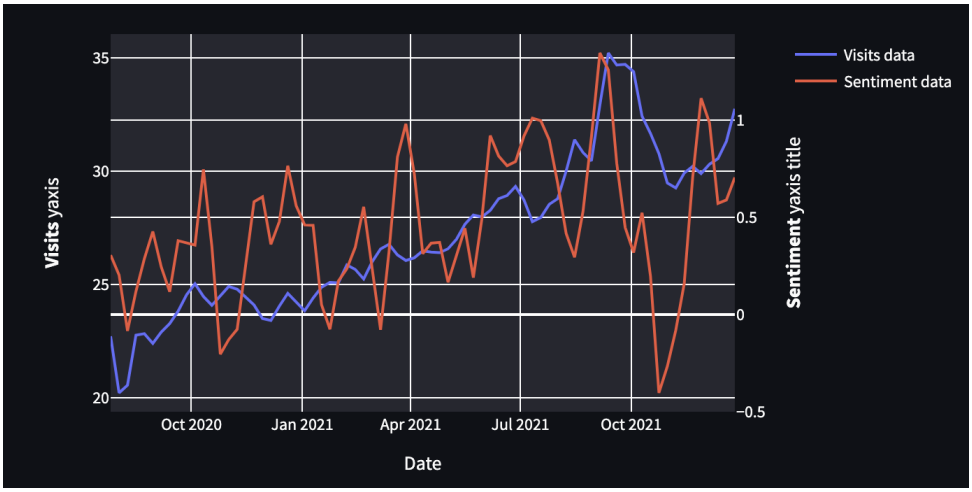


Figura C.3: Datos de sentimiento y visitas semanales correspondientes con Italia. Fuente: Elaboración propia

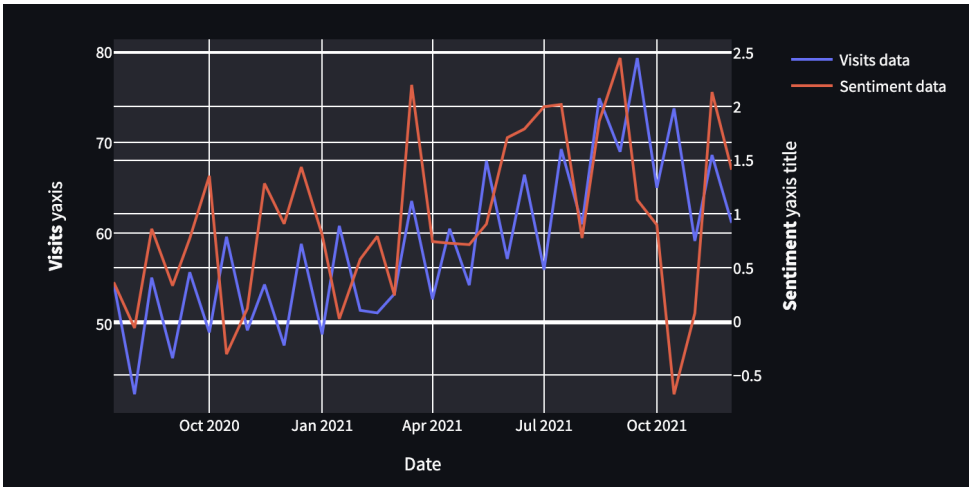


Figura C.4: Datos de sentimiento y visitas quincenales correspondientes con Italia. Fuente: Elaboración propia

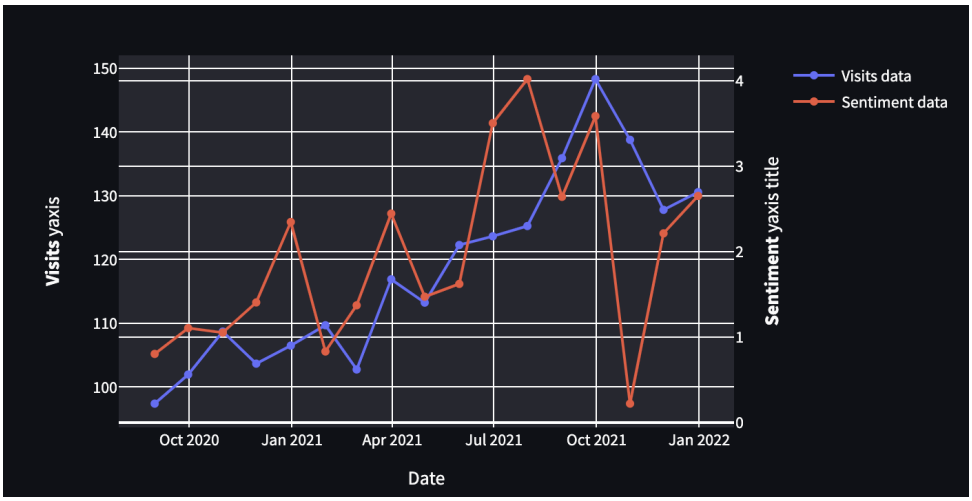


Figura C.5: Datos de sentimiento y visitas mensuales correspondientes con Italia. Fuente: Elaboración propia

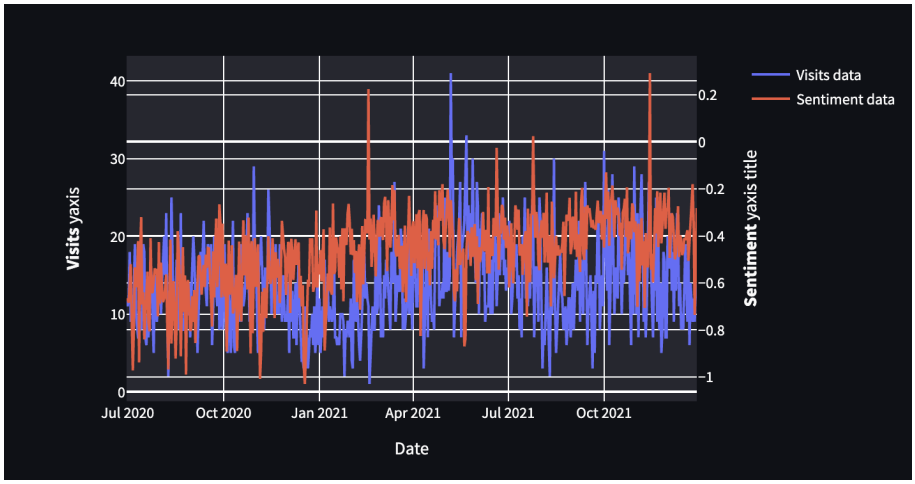


Figura C.6: Datos de sentimiento y visitas correspondientes con México. Rusia: Elaboración propia

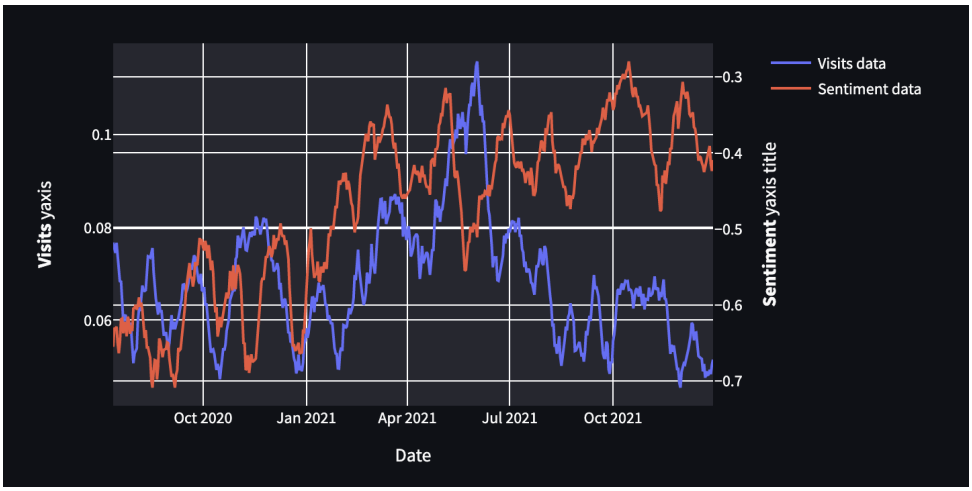


Figura C.7: Datos de sentimiento y visitas suavizados correspondientes con Rusia. Fuente: Elaboración propia

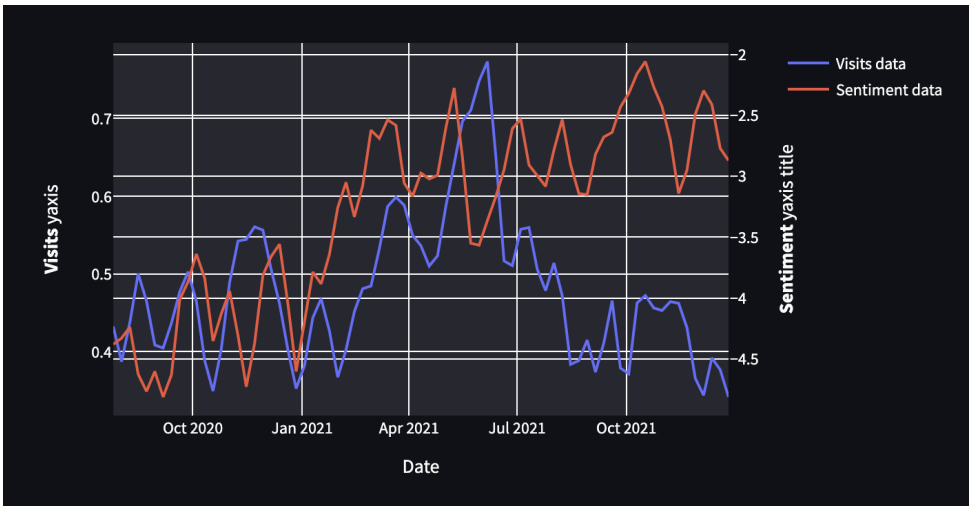


Figura C.8: Datos de sentimiento y visitas semanales correspondientes con Rusia. Fuente: Elaboración propia

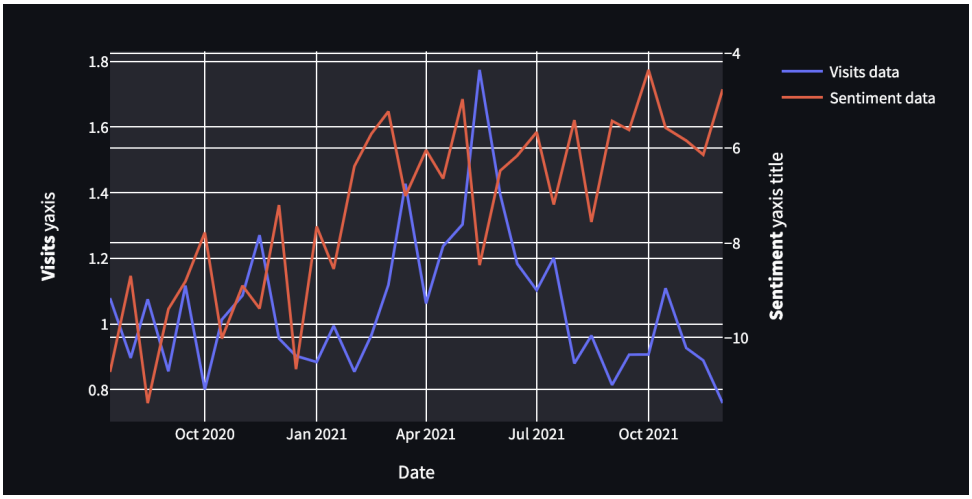


Figura C.9: Datos de sentimiento y visitas quincenales correspondientes con Rusia. Fuente: Elaboración propia

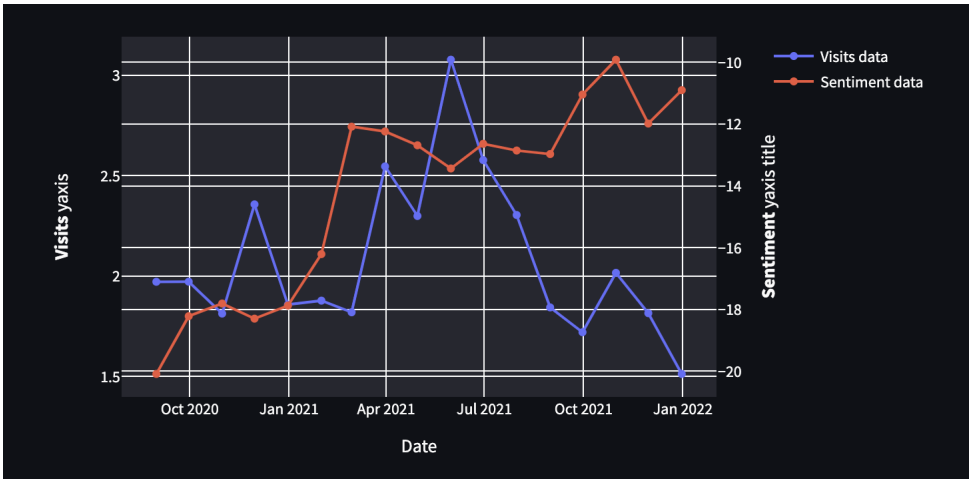


Figura C.10: Datos de sentimiento y visitas mensuales correspondientes con Rusia. Fuente: Elaboración propia

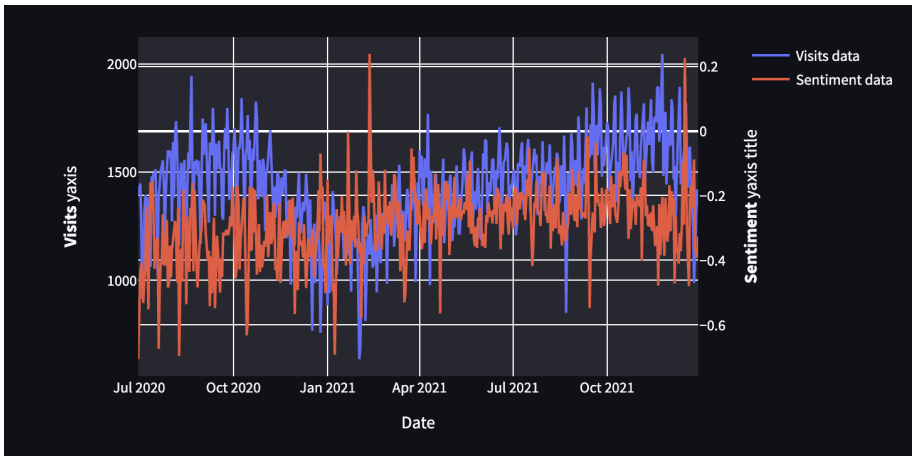


Figura C.11: Datos de sentimiento y visitas correspondientes con China. Fuente: Elaboración propia

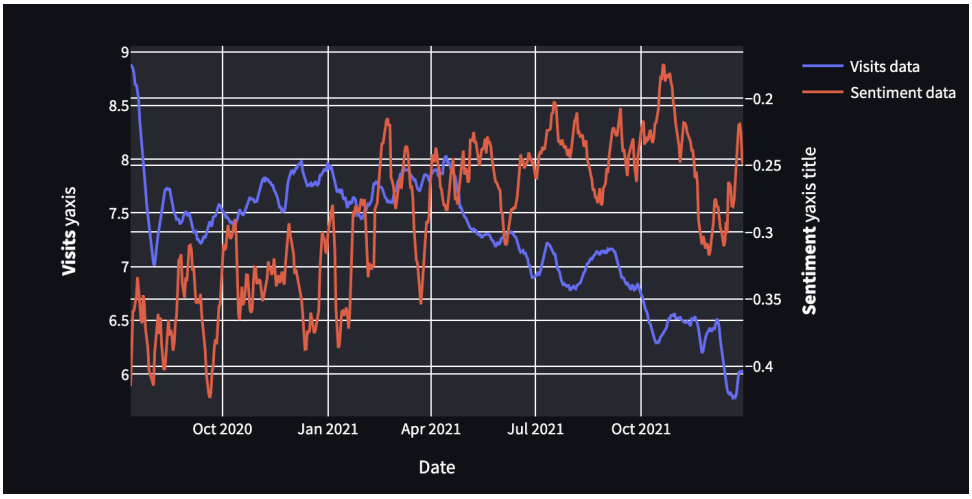


Figura C.12: Datos de sentimiento y visitas suavizados correspondientes con China. Fuente: Elaboración propia

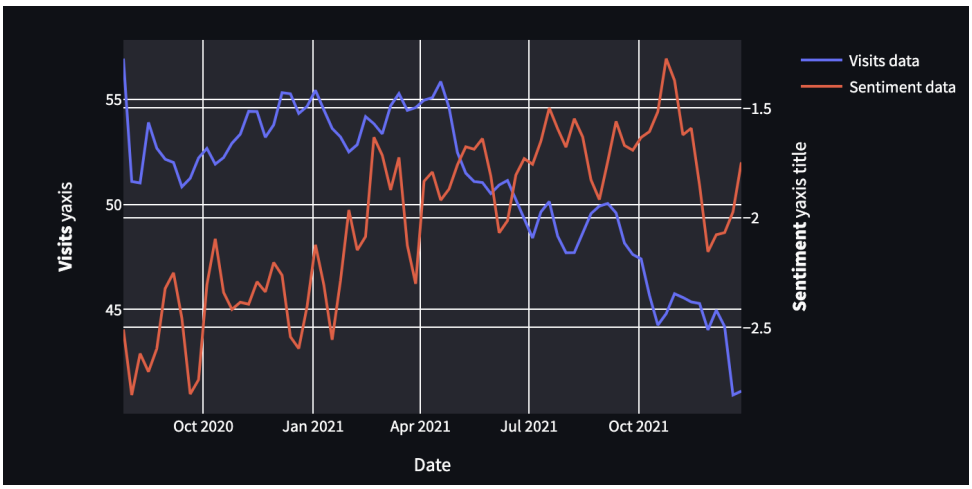


Figura C.13: Datos de sentimiento y visitas semanales correspondientes con China. Fuente: Elaboración propia

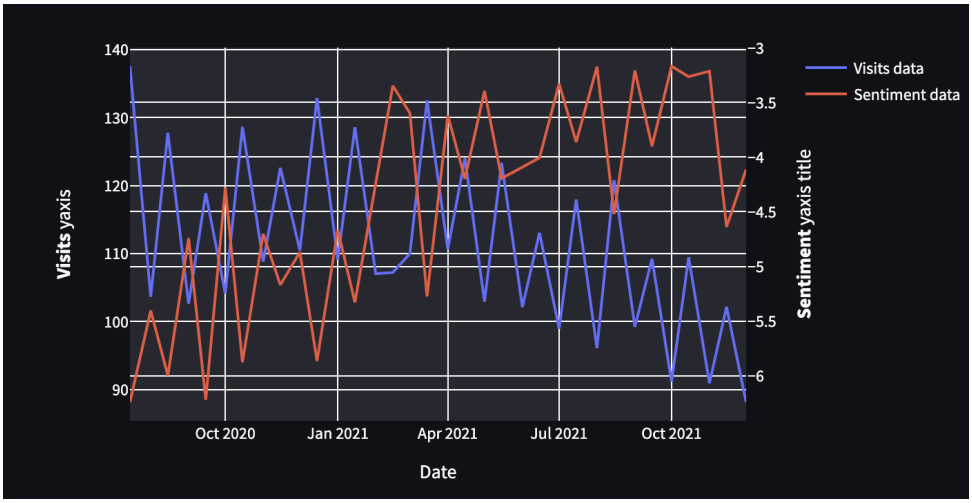


Figura C.14: Datos de sentimiento y visitas quincenales correspondientes con China. Fuente: Elaboración propia

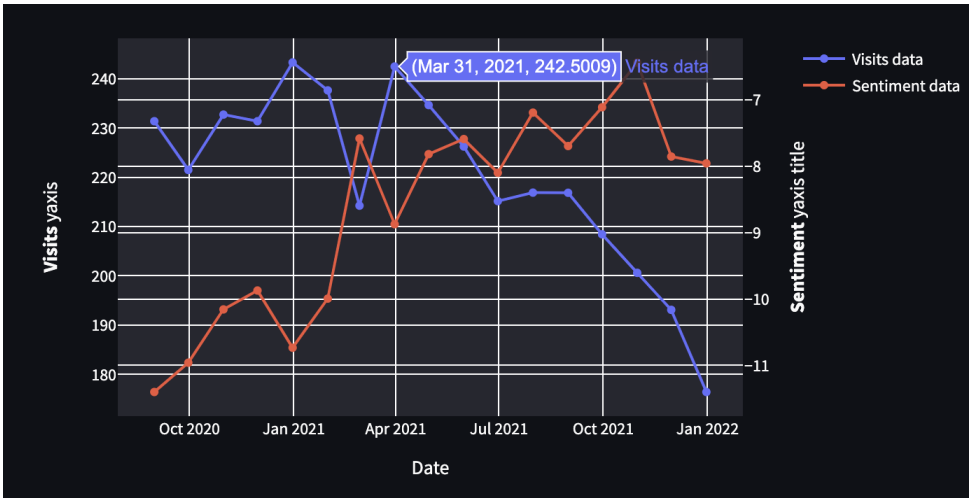


Figura C.15: Datos de sentimiento y visitas mensuales correspondientes con China. Fuente: Elaboración propia