Universidad de Valladolid

Facultad de Ciencias Económicas y Empresariales

Trabajo de Fin de Grado

Grado en Administración y Dirección de Empresas

Análisis de Sentimientos y engagement en la red social Twitter

Presentado por:

Juan José Hernanz Fernández

Tutelado por:

Rebeca San José Cabezudo

Valladolid, 22 de julio de 2022

RESUMEN

Las redes sociales se constituyen uno de los principales medios para compartir información, experiencias y opiniones. Estas conversaciones en muchos casos son públicas y, correctamente analizadas, pueden ser de gran valor para una empresa en términos de percepción marca. Con el objetivo de mejorar la toma de decisiones de marketing en redes, este trabajo se centra en recopilar y analizar datos de *Twitter* en el sector de festivales de música en España. Parte del desarrollo de un sistema de recopilación y procesado de *tweets*. Después se analizan las métricas principales que miden el *engagement* tratando de identificar patrones. Luego se estudia tanto el sentimiento general de los *tweets*, como el sentimiento hacia las marcas del sector y se investiga la relación que este pueda tener con el *engagement*. Finalmente, los resultados convergen en una serie de recomendaciones sobre la estrategia marketing.

PALABRAS CLAVE – CÓDIGOS JEL

Análisis de Sentimientos, Twitter, engagement, marketing – M31, O33, D91

ABSTRACT

Social networks have increasingly become one of the main means of sharing information, experiences and opinions. These conversations are in many cases public and accessible. Properly analysed, this information can be of great value to a company, in terms of brand perception or relationship with consumers. With the aim of improving marketing decision making, this work focuses on collecting and analysing Twitter data in the Spanish music festival industry. It begins with the development of a system for collecting and processing tweets. Then we investigated the main metrics that measure engagement, trying to identify patterns. After that, we studied the general sentiment of tweets, the sentiment towards the brands, and the relationship that sentiment may have with brand engagement. Finally, all the results converge in a set of recommendations to improve the marketing strategy.

KEYWORDS – JEL CODES

Sentiment Analysis, Twitter, engagement, marketing – M31, O33, D91

ÍNDICE GENERAL

Resumen *1*

Palabras clave – Códigos JEL *1*

Abstract *1*

Keywords – JEL codes *1*

índice general *2*

índice de figuras *3*

1. Introducción *4*

1.1. Motivación **4**

1.2. Objetivos y alcance **5**

1.3. Metodología **6**

1.4. Estructura del documento **7**

2. Análisis, diseño e implementación del sistema *8*

2.1. Introducción **8**

2.2. Planificación del sistema **8**

2.2.1. Recogida de datos 8

2.2.2. Procesamiento de datos. 9

2.2.3. Selección de un sector. 9

2.3. Análisis del sistema **11**

2.4. Diseño del sistema **13**

2.4.1. Tecnologías utilizadas 13

2.4.1.1. Tecnologías utilizadas para la recopilación de Tweets 13

2.4.1.2. Tecnologías utilizadas para el Análisis de Sentimiento 13

2.4.1.3. Tecnologías utilizadas para la representación de tweets 14

2.4.2. Visión del sistema 15

2.5. Implementación del sistema **16**

2.5.1. Tablas de datos. 16

2.5.2. Conexiones con las APIs. 17

2.5.3. Desarrollo de los scripts. 17

3. Análisis de engagement y sentimientos *18*

3.1. Introducción **18**

3.2. Características del conjunto de datos **19**

3.3. Patrones de *engagement* **20**

3.3.1. Métricas públicas de la totalidad de *tweets*. 21

3.3.2. Métricas públicas de *Tweets* publicados por las cuentas oficiales. 22

3.3.3. Patrones temporales. 25

3.3.4. Análisis de temáticas y entidades 26

3.4. Análisis de sentimientos **28**

3.4.1. Metodología 28

3.4.2. Análisis de sentimientos general de cada tweet. 29

3.4.3. Análisis de sentimientos de entidades 31

3.4.4. Relaciones entre el sentimiento y el *engagement* del cliente 32

Efecto del sentimiento sobre las métricas públicas 33
Efecto de los usuarios verificados 33
Efecto de los hashtags y menciones. 35

4. Conclusiones. 36
4.1.1. Recomendaciones basadas en patrones de *engagement*..... 36
4.1.2. Recomendaciones basadas en el sentimiento del cliente. 37
4.1.3. Conclusiones 38
4.1.4. Limitaciones y futuras líneas de trabajo. 39

ANEXO 1. Tablas del modelo de datos 41

ÍNDICE DE FIGURAS

Figura 1.1: Ciclo de Vida del Desarrollo de Sistemas (SDLC). (Valacich & George, 2017) 6
Tabla 2.1. Comparativa de las diferentes redes sociales y los tipos de datos que permiten extraer 9
Figura 2.1: Visión del sistema. Elaboración propia 10
Figura 2.2: Diagrama de Flujo de Datos del sistema. Elaboración propia. 12
Figura 2.3: Diagrama de diseño del sistema. 15
Figura 3.1: Métricas generales de los *Tweets* de festivales y fechas de publicación. 20
Figura 3.2: Métricas principales de *engagement* desglosadas por *keywords*. 21
Figura 3.3: Publicaciones de las cuentas oficiales de cada festival en el período de estudio. 22
Figura 3.4: Métricas principales de *engagement* desglosadas por las cuentas oficiales. 23
Figura 3.5: Análisis comparativo de las métricas públicas de 5 cuentas oficiales de festivales en el período de estudio. 23
Figura 3.6: Mapa de calor de la métrica de *engagement* desglosado por horas y días (arriba) y sólo por horas (abajo) 25
Figura 3.7: Análisis de la tasa de *engagement* por día de la semana. 26
Figura 3.8: Nubes de palabras de las temáticas de los *tweets* a través de dominios (izquierda) y entidades (derecha). 27
Figura 3.9: Frecuencia con la que se habla de diferentes músicos en las menciones al Mad Cool y al BBK. 27
Figura 3.10: Interpretación de resultados del análisis de sentimientos (Google, 2022). 29
Figura 3.11: Umbrales de para el cálculo del sentimiento. Elaboración propia. 29
Figura 3.12: Porcentaje de sentimiento por cada *keyword*. 30
Figura 3.13: sentimiento hacia los eventos detectados como entidades. 31
Figura 3.14: Filtro de *tweets* con sentimiento NEGATIVO hacia la entidad Primavera Sound. 32
Figura 3.15: Ratios de *engagement* para cada tipo de sentimiento. 33
Figura 3.16: Comparativa de los *tweets* y los seguidores de los usuarios verificados vs. los no verificados para cada *keyword*. 34
Figura 3.17: sentimiento de los *tweets* de los usuarios verificados y no verificados. Elaboración propia. 34
Figura 3.18: sentimiento de los *tweets* de los que usan *hashtags* y menciones. 35

1. INTRODUCCIÓN

1.1. Motivación

El conocimiento de los gustos, preferencias y tendencias de los clientes es clave para determinar las estrategias de marketing de una empresa (Tuten & Solomon, 2015). Gran parte de esta información se encuentra de una manera más o menos explícita en las redes sociales como consecuencia del uso que todos hacemos de ellas. Si combinamos de manera adecuada tecnologías para la **extracción** y **minería** de estos **datos**, podremos obtener información de mucho valor sobre la opinión general de los usuarios (Ravi & Ravi, 2015). Esto es clave, pues el éxito de un negocio depende de que los consumidores prefieran sus productos o servicios frente al resto que se ofrecen en el mercado.

El objetivo que nos concierne en este Trabajo de Fin de Grado (TFG) es la implementación de un sistema completo para la **extracción**, **limpieza** y **análisis** de datos procedentes de la red social *Twitter,* así como la visualización e interpretación de resultados poniendo el foco en el **Análisis de Sentimientos** de *tweets* para poder aplicarlo a nivel de marketing empresarial.

La puesta en marcha de este trabajo vino motivada por la intersección de dos de mis inquietudes más fuertes: la **analítica de datos** y el **marketing**. Como consecuencia de mis estudios conjuntos de Ingeniería de Telecomunicaciones y Administración y Dirección de Empresas, quería aprovechar mis conocimientos más técnicos de programación y análisis de datos para poder aplicarlos a la disciplina de Marketing. Adicionalmente, consideraba que era una buena oportunidad para aprender y utilizar alguna de las infinitas aplicaciones que el *Machine Learning* y la *Inteligencia Artificial* nos brinda en el presente.

Además, me parecía un proyecto muy enriquecedor tanto a nivel académico como a nivel empresarial, pues quería desarrollar un sistema pensando en que podía ser reutilizable para investigar en otros sectores, más allá del que nos centramos en este trabajo.

1.2. Objetivos y alcance

El objetivo del TFG será llevar a cabo un análisis de publicaciones en *Twitter* del sector de festivales de música. En primer lugar, se pretenden medir los niveles de *engagement* para buscar posibles patrones en el sector. Después se estudiará el sentimiento de los *tweets,* y se buscarán relaciones entre este sentimiento y el *engagement*. Todo ello será la base para sugerir acciones en las decisiones de marketing. Para conseguirlo se han de implementar un conjunto de soluciones que van desde la extracción de los tweets y su información relevante, pasando por su procesamiento hasta su visualización y análisis.

Uno de los mayores retos que ha supuesto este trabajo es acotar el alcance de este. Es tal la cantidad de información que cada día se publica en redes, y son tantas las redes sociales que actualmente se utilizan, que se vuelven casi infinitas las posibilidades de análisis y de variables a estudiar. Por lo tanto, el primer paso fue delimitar el conjunto de datos que quería recopilar:

- Se trabajará con datos procedentes únicamente de la red social *Twitter*.
- Nos centraremos en *tweets* que se han publicado únicamente en español.
- Recopilaremos *tweets* durante un período de 26 días.
- Nos centraremos en un único mercado, concretamente el de los festivales de música en España. Por eso, recopilaremos *tweets* que incluyan ciertas palabras clave referentes a ese sector.

Teniendo esto en cuenta, se definieron el conjunto de soluciones o tareas que desarrollar en el trabajo. Es más intuitivo si las enumeramos partiendo del objetivo:

- Proponer una serie de **acciones** y **decisiones de marketing**, y basadas en las interacciones de los usuarios su sentimiento hacia la marca.
- Realizar un **análisis** centrado en el **sentimiento** del consumidor y el nivel de interacciones de este con la marca.
- Para ello, se realizará una serie de **análisis de datos** de esta red social. Entre ellos se analizará información de métricas públicas, usuarios, entidades, sentimientos, *hashtags*, menciones, fechas y seguidores.

- Para realizar estos análisis, necesitamos crear una seria de **gráficas** que pueden relacionarse entre sí.
- Estas gráficas representarán diferentes datos o campos. Para organizar estos campos es preciso definir un **modelo de datos** compuesto de diferentes tablas relacionadas a través de algún campo.
- Los campos del modelo de datos procederán de **diferentes fuentes**. Algunos de ellos vendrán directamente del código que extrae los *tweets* y otra información relativa, mientras que otros serán el resultado de un algoritmo, por ejemplo, el sentimiento general del texto del *tweet.*

1.3. Metodología

El trabajo consta de una parte de desarrollo *software* en la que se ha utilizado la división cinco fases del Ciclo de Vida de Desarrollo de Sistemas, tal y como se muestra Figura 1.1. Se trata abordar la creación de un sistema comenzando con la Planificación en la que se fijan los objetivos. Le siguen las etapas de Análisis y Diseño en las que se definen los requisitos, así como los componentes del sistema. La cuarta etapa es la de Implementación y en ella se ejecutan las tareas de desarrollo software. Una vez pasados los *tests*, el sistema se despliega en producción pasando a estar operativo.

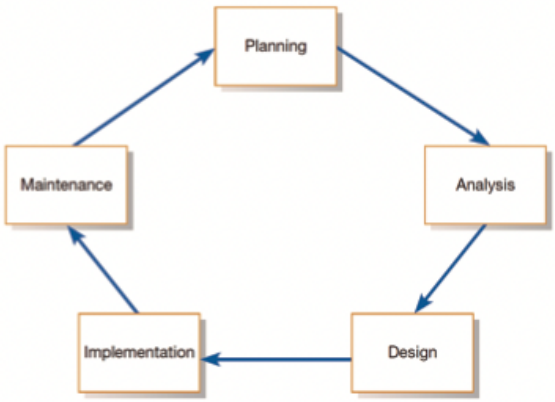

Figura 1.1: Ciclo de Vida del Desarrollo de Sistemas (SDLC). (Valacich & George, 2017)

Respecto a la metodología orientada a las decisiones de marketing, se ha optado por un enfoque orientado al dato. Es decir, basándonos en los resultados de los datos recopilados y analizados, se proponen una serie de acciones que buscan la mejora de la relación usuario-marca, mejora de la notoriedad y del sentimiento hacia la empresa.

1.4. Estructura del documento

Este documento está dividido en 4 capítulos teniendo en cuenta este introductorio.

El Capítulo 2 tiene un mayor carácter técnico ya que en él se detalla el contenido relativo a la planificación, análisis, diseño e implementación del sistema de recolección y análisis de tweets. Aunque no es el foco principal, se ha considerado relevante dedicar algunas páginas a la construcción del sistema. En primer lugar, porque es una parte imprescindible para las posteriores labores de investigación y, además, ayuda a tener una visión global de todas las tareas desempeñadas, así como a ser consciente del esfuerzo que se ha invertido.

El Capítulo 3 es sin duda el más relevante en esta investigación y abarca todo el contenido relativo al análisis de los datos y presentación de resultados. Comienza con una introducción para poner en valor la importancia de los tipos de análisis que se van a desarrollar, así como la delimitación de los objetivos del análisis. A lo largo de las distintas secciones se analizan los datos recopilados poniendo el foco en las métricas de *engagement*, los sentimientos y la relación entre ambas.

Finalmente, en el Capítulo 4 se resumen las ideas más importantes de la investigación y se sugieren un conjunto de acciones basadas en los datos y orientadas a la estrategia de marketing en redes.

Las últimas páginas se corresponden con el Anexo 1 y tienen información detallada de las tablas y los campos del conjunto de datos.

2. ANÁLISIS, DISEÑO E IMPLEMENTACIÓN DEL SISTEMA

2.1. Introducción

En este Capítulo se expone todo el contenido relativo al desarrollo del sistema necesario para obtener la información adecuada que se utilizará para el análisis y exposición de resultados.

2.2. Planificación del sistema

En la etapa de **Planificación** del SDLC se realizan aquellas tareas que tienen que ver con la identificación y selección del proyecto tal y como se ha descrito en las secciones 1.1 y 1.2. No obstante, en esta sección se describirán algunos de los problemas principales encontrados durante esta etapa.

Como ya se menciona en la sección 1.2, uno de los mayores retos del trabajo fue la delimitación del alcance este. Tarea que se llevó a cabo en esta fase inicial.

La idea de partida era que queríamos trabajar con datos de redes sociales, después aplicar algún tipo de procesado basado en algoritmos de *machine learning* y, finalmente, desarrollar una estrategia de marketing basada en estos resultados. De cada una de estas actividades surgían una serie de preguntas que debían ser respondidas durante la etapa de planificación

2.2.1. Recogida de datos

Se trataba de la primera fase que había que abordar. Las primeras preguntas que surgieron fueron: **¿Qué tipos de datos podemos obtener de las distintas redes sociales y cómo puedo recogerlos?**

Tras un análisis exploratorio de las diferentes redes sociales se creó la *Tabla 2.1*. con los resultados. Como podemos observar, tanto *Twitter* como *Twitch* e *Instagram* disponen de *APIs* oficiales con la que los desarrolladores pueden obtener datos almacenados en estas herramientas. Sin embargo, sólo *Twitter* ofrece datos de otros usuarios diferentes a la cuenta propia cuenta del

desarrollador. Por esta razón la primera decisión fue utilizar datos que se podían obtener de la *API* de *Twitter*.

Tabla 2.1. Comparativa de las diferentes redes sociales y los tipos de datos que permiten extraer

	Disponibilidad de los datos	**Tipos de datos**
Twitter	Dispone de *API* oficial	• Tweets de un usuario. • Tweets de un *hashtag*. • Búsqueda de tweets basada en publicaciones recientes (últimos 7 días). • Métricas públicas de un tweet. • Conteo de Tweets y Seguidores. • Respuestas y conversaciones a partir de un tweet.
TikTok	Existen *APIs* no oficiales	• Vídeos publicados por un usuario. • Vídeos que le gustan a un usuario. • Vídeos basados en un *hashtag* • Vídeos en tendencias. • Listas de usuarios.
Twitch	*API* oficial limitada	• Únicamente datos que pertenecen a tu cuenta. No datos de otros usuarios.
Instagram	*API* oficial limitada	• Únicamente datos que pertenecen a tu cuenta. No datos de otros usuarios.

2.2.2. Procesamiento de datos.

La pregunta clave es **¿Qué queremos hacer con los datos?** La respuesta a esta pregunta nos ayudaría a acotar 2 problemas principales: ¿qué tipo de análisis de datos quiero llevar a cabo? y ¿qué variables de todas las que puedo escuchar son las que quiero medir para llevar a cabo ese análisis?

Llegar a responder estas preguntas no fue nada fácil. Tras realizar bastantes investigaciones y cursos relacionados con análisis de datos de *Twitter*, llegué a la misma conclusión del principio: eran tantas las alternativas de análisis y procesado de datos que finalmente opté centrar todos mis esfuerzos en realizar un **Análisis de Sentimiento** del texto de los tweets recopilados.

2.2.3. Selección de un sector.

Acotados todos los límites anteriores, sólo falta por decidir sobre qué industria o marcas vamos a enfocar nuestro análisis. Esta decisión tampoco es trivial, pues

necesitamos recopilar tweets que incluyan ciertas palabras clave (*keywords*) del sector teniendo en cuenta los siguientes requisitos:

- Deben ser ***keywords* únicas del sector**. Es decir que incluyan el menor ruido posible entre los resultados de los tweets que recopilamos. Por ejemplo, si buscamos *tweets* que incluyan la palabra "Protos" en referencia a la marca de vino, vamos a obtener además muchos *tweets* del sector de videojuegos, pues también es el nombre de un personaje.
- Deben ser ***keywords* que aparezcan en un número considerable de tweets a diario**. Por ejemplo, en el sector vitivinícola, había marcas que no eran mencionadas ninguna o muy pocas veces en los últimos 7 días. Esta escasez de datos hace imposible sacar unas conclusiones a nivel agregado respecto a la percepción de la marca.

Por ello, tras valorar inicialmente a la industria vitivinícola, se desestimó la idea en favor del **sector de festivales de música**. Pues tienen nombres más diferenciales y, sobre todo, acumulan más menciones a diario.

Al final de esta fase de Planificación, quedan perfectamente seleccionados los objetivos y funcionalidades que el sistema pretende cubrir. En la Figura 2.1 se ilustra una primera visión del problema.

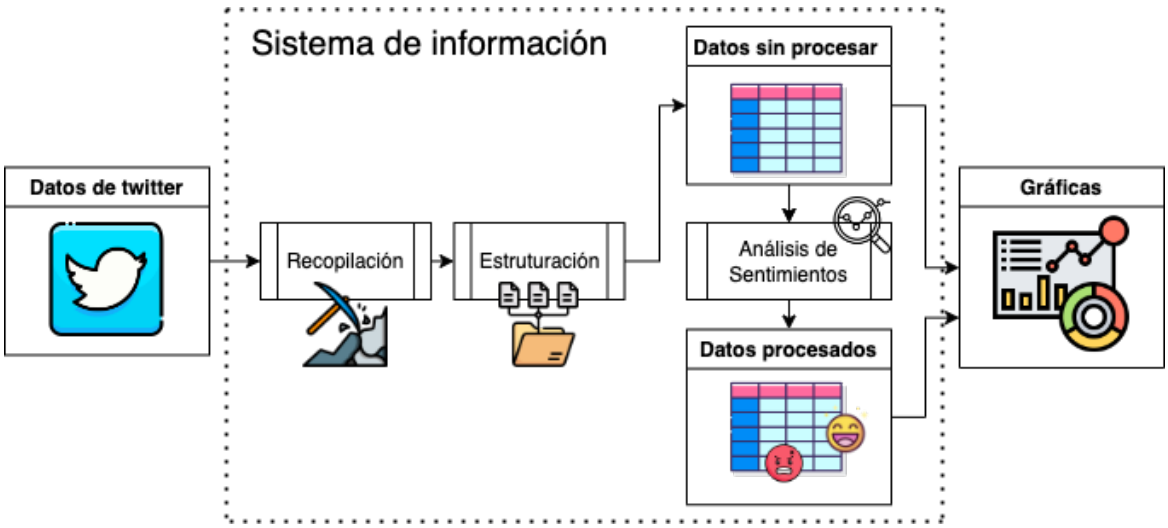

Figura 2.1: Visión del sistema. Elaboración propia

2.3. Análisis del sistema

Durante la etapa de Análisis se determina qué información va a necesitar el sistema y qué servicios de procesamiento de esta información se necesitan para apoyar los objetivos. Se trata de una fase en la que los esfuerzos se centran en definir y estructurar requisitos de acuerdo con los objetivos planificados, así como tener una primera visión de los flujos de información y los procesos internos del sistema (Valacich & George, 2017).

En esta etapa se creó el **modelo de proceso** tal y como se muestra en la Figura 2.2. utilizando los símbolos propuestos por Gane y Sarson (Sarson & Gane, 1979). Se trata de un diagrama que muestra el **flujo de datos** (flechas), los **procesos** (cuadros redondeados) entendidos como acciones que se realizan sobre los datos y los **almacenes de datos** (rectángulo sin el lado derecho) que representan datos en reposo. Este diagrama también representa con rectángulos el **origen y destino** de los datos.

En el Diagrama del Flujo de Datos de la Figura 2.2 se reflejan las 5 funciones principales que debe realizar el sistema y que se representan en **5 procesos:**

1. **Recopilar tweets a través de peticiones a la *API*.** Mediante estas peticiones que el sistema hace a la *API* de *Twitter*, obtiene respuestas con los tweets que cumplen las condiciones de esas peticiones.
2. **Formatear la respuesta y Anexar los datos a las tablas.** Las respuestas de la *API* contienen mucha información asociada a cada *tweet*. En este proceso nos quedamos sólo con aquellos campos de cada *tweet* que nos serán de utilidad. Además, estos datos se estructurarán en 3 tablas diferentes (D1, D2 y D3) en función de la naturaleza de su información.
3. **Análisis de Sentimiento general** de cada *tweet.* Se trata de un proceso que tiene como entrada el texto de cada publicación y como salida información del sentimiento general de ese texto. Esa información se almacena en la tabla D4.
4. **Análisis de Sentimiento de entidades** de cada *tweet.* La entrada de este proceso es la misma que el anterior, pero en este caso no se obtiene como salida el sentimiento general de cada *twee*t, sino que se extraen las

distintas entidades que se mencionan en el texto y el sentimiento asociado a cada una. Esa información se almacena en la tabla D5.

5. **Transformar y representar datos de los tweets**. Es el último proceso, consiste en utilizar todos los datos de las 5 tablas que se han generado a lo largo del sistema y representarlas a través de gráficas. Esas gráficas se organizan en *dashboards* que serán interpretados por el Equipo de Marketing.

De esta manera, queda totalmente representado qué forma toma la información en cada momento, y qué procesos realiza el sistema para que los datos fluyan desde Twitter (origen), hasta el equipo de Marketing (destino).

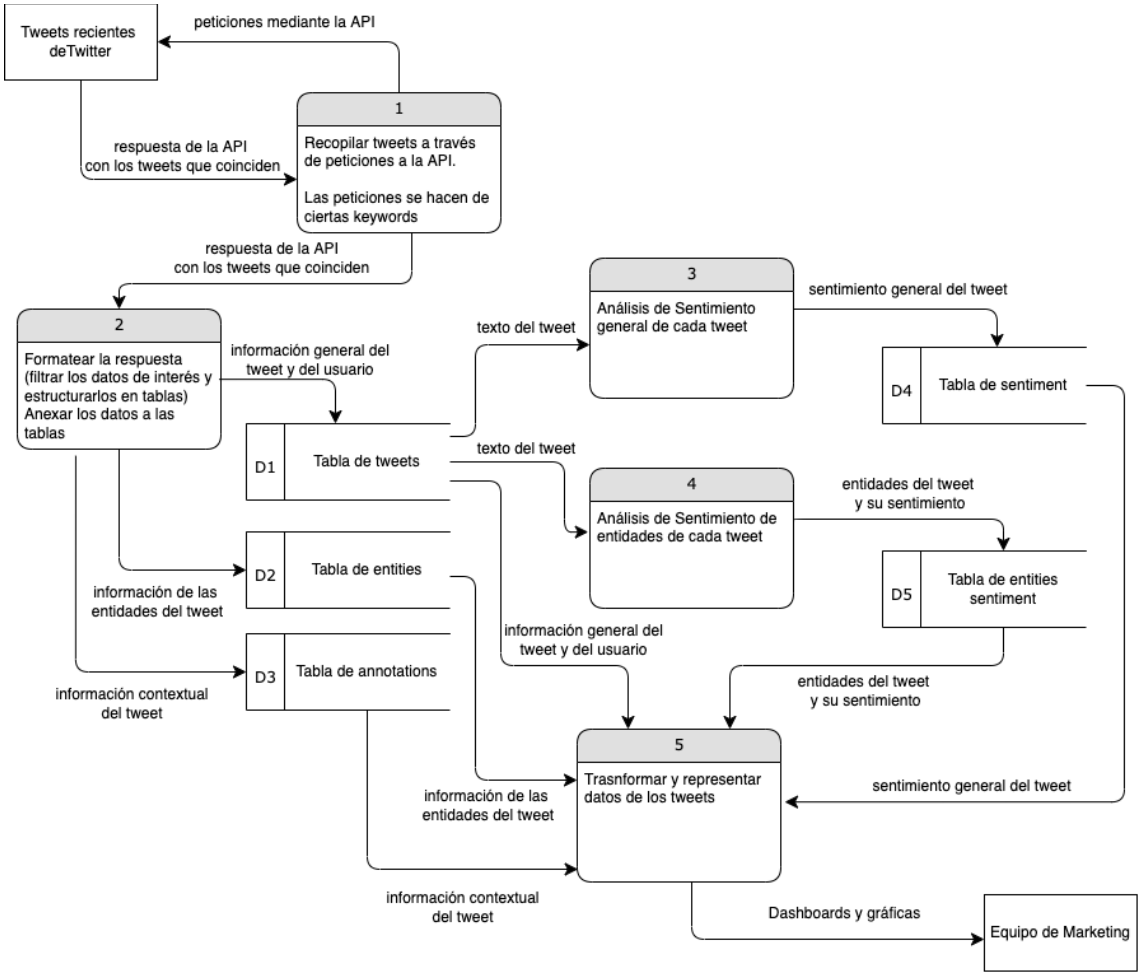

Figura 2.2: Diagrama de Flujo de Datos del sistema. Elaboración propia.

2.4. Diseño del sistema

Se trata de la fase previa al desarrollo y en ella se define cómo va a operar el sistema planteando la solución definitiva que se va a desarrollar. Para ello, se ha de tener en cuenta tanto los requisitos como los flujos y procesos de información de las fases anteriores.

2.4.1. Tecnologías utilizadas

Se procede a enumerar y describir brevemente las **tecnologías utilizadas.**

2.4.1.1. Tecnologías utilizadas para la recopilación de Tweets

- ***API de Twitter:*** se trata de la tecnología utilizada como intermediario entre la persona encargada de recopilar los tweets y la base de datos donde *Twitter* almacena toda su información. A través de esta plataforma, *Twitter* pone a disposición del público, ciertos datos y métricas asociados a los *tweets* y usuarios entre otros (Twitter, 2022).
- ***Python 3*****:** es el lenguaje de alto nivel de programación (Python, 2022) empleado en el desarrollo de todas las funcionalidades. Algunas de sus librerías fundamentales para la consecución de este TFG son:
    - ***tweepy*****:** librería que pone a nuestra disposición las herramientas necesarias para el acceso a la *API de Twitter* (Tweepy, 2022). Se ha utilizado para obtener los tweets y su información asociada.
    - ***pandas*****:** se trata de una de las herramientas más potentes para análisis y manipulación de datos (Pandas, 2022). Se ha utilizado principalmente en los procesos de estructuración de datos.
- ***CSV*** (del inglés comma-separated values). No es un tipo de tecnología como tal, sino el formato de archivo de texto utilizado para almacenar las tablas de datos. Se decidió almacenar los datos en CSVs en lugar de en bases de datos por su sencillez, reducido tamaño y compatibilidad con el resto de las tecnologías.

2.4.1.2. Tecnologías utilizadas para el Análisis de Sentimiento

El Análisis de Sentimiento de un texto a través de técnicas de Procesamiento del Lenguaje Natural (NLP) ha sido objetivo de muchos desarrolladores y grandes

empresas en los últimos años. Para resolver este problema se emplea la Inteligencia Artificial para crear modelos que, tras ser entrenados para resolver el problema, son capaces de obtener soluciones con mayor o menor acierto en función de lo bueno que sea el modelo. Tras probar distintas soluciones, la tecnología que mejores resultados proporcionaba, sobre todo en español, es la ***API de Natural Language de Google*** (Google, 2022). Se trata de una herramienta que utiliza modelos entrenados previamente para procesar textos como entrada y ofrecer resultados sobre esos textos. Se han utilizado dos tipos de servicios:

- **Análisis de Opiniones:** identifica la actitud general del autor del texto de entrada como positiva, negativa, neutral o mixta. También ofrece el sentimiento de cada frase que compone el texto.
- **Análisis de Opiniones de entidades**: procesa el texto para extraer las entidades sobre las que se habla en el texto (por ejemplo, si habla de una persona, un bien de consumo…) y determina la opinión predominante hacia ellas.

Para el desarrollo del código encargado de realizar las peticiones y obtener las respuestas de estos servicios también se ha utilizado **Python 3** utilizando el paquete de software **language_v1**.

2.4.1.3. Tecnologías utilizadas para la representación de tweets

A partir de los datos obtenidos tanto de la *API de Twitter* como de la *API de Natural Language de Google,* y que están almacenados en tablas en formato *CSV*, se han utilizado 2 herramientas adicionales para la interpretación de estos datos:

- ***Tableau Prep Builder*:** que es un programa que proporciona una interfaz para preparar, combinar y transformar datos previamente a su análisis.
- ***Tableau*:** que es la plataforma utilizada para la visualización de los datos de manera que podamos realizar todos los análisis propuestos (Tableau, 2022).

2.4.2. Visión del sistema

El último paso previo al desarrollo del sistema es la definición de los elementos que lo constituyen, concretando las tecnologías que se van a utilizar y cómo van a fluir los datos entre ellos. En la Figura 2.3. se representa el diseño del sistema.

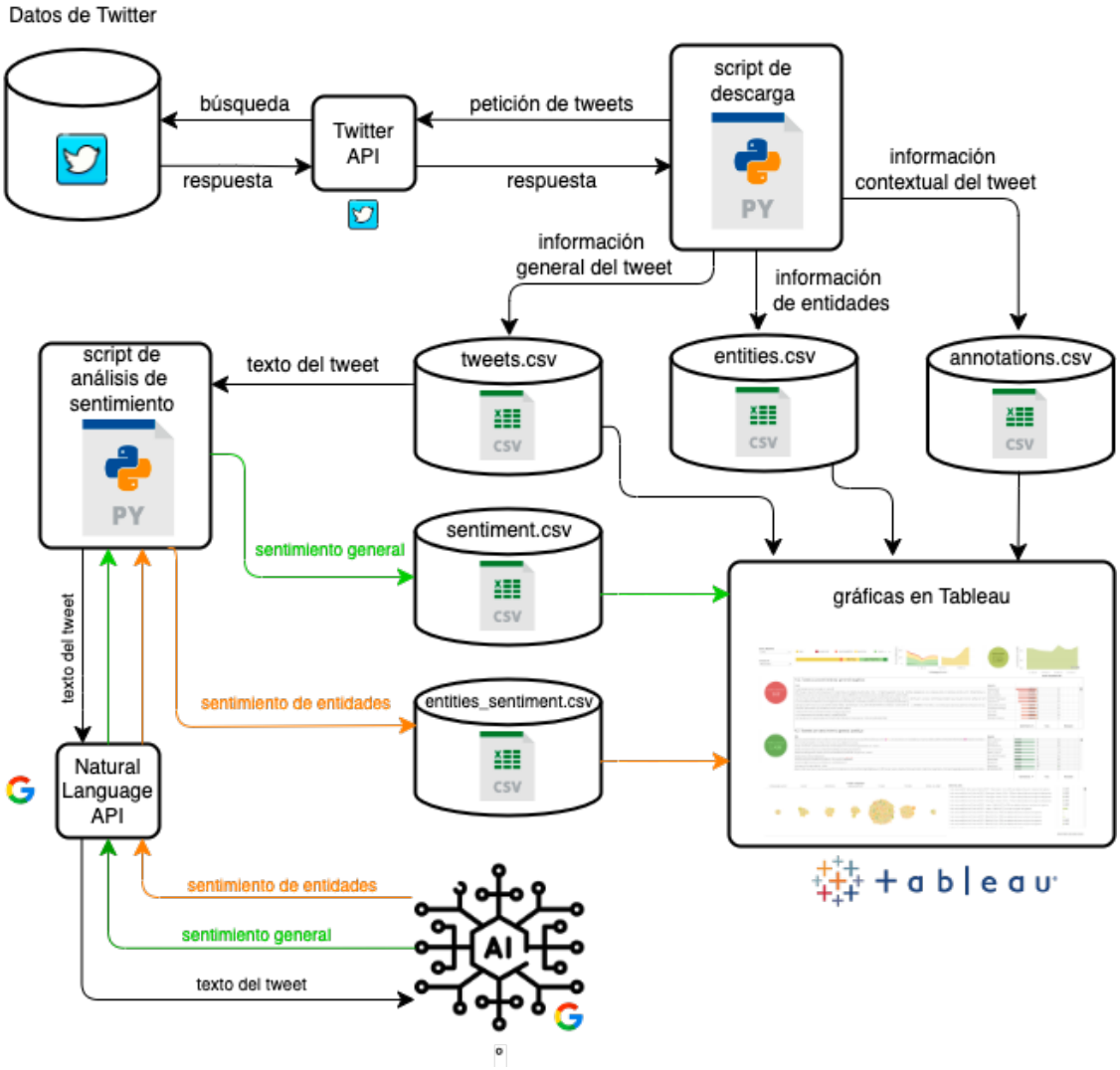

Figura 2.3: Diagrama de diseño del sistema.

De entre todos los elementos que componen el sistema en la Figura 2.3, los dos principales son los *scripts* de código *python*. Ambos scripts junto con la creación de gráficas son elementos cuya creación ha acaparado casi la totalidad de los esfuerzos de implementación. El ***script* de descarga** es un fichero de código que se encarga de recopilar los tweets de interés a través de la comunicación con la ***Twitter API*.** Además, realiza la estructuración de los datos de cada tweet y los anexa a sus **ficheros csv** correspondientes (*tweets, entities o annotations*).

Por otro lado, con cada ejecución del ***script* de análisis de sentimiento**, se leen los textos de cada *tweet* desde la tabla de *tweets.csv* y se procesan por la **Inteligencia Artificial** a través de la ***Natural Language API de Google***, para obtener los resultados tanto de sentimiento general como de sentimiento de entidades. Esos resultados son organizados por el *script* y se anexan a sus tablas correspondientes (*sentiment* y *entities_sentiment*).

Las cinco tablas en formato csv, se utilizan como fuentes de datos para la creación de las visualizaciones que utilizaremos para construir nuestra estrategia de marketing.

2.5. Implementación del sistema

La fase de **Implementación** abarca todas aquellas tareas que tengan que ver con el desarrollo software y construcción de todas las piezas del sistema. Se trata de la fase más técnica y costosa del SDLC. Al no ser objetivo final del presente TFG, simplemente se harán enumeran las configuraciones que se han considerado más relevantes.

2.5.1. Tablas de datos.

En este sistema, el dato lo es todo. Necesitamos datos para tomar las decisiones de marketing y estos datos deben estar disponibles con la calidad y en formato apropiado para representarse en las gráficas. Por todo ello, la primera parte de la implementación consistió en crear un **modelo de datos.** Gracias a ese modelo de datos tendríamos la información completa de qué campos se van a incluir en cada una de esas cinco tablas, qué tipo de datos tiene cada campo y cómo se relacionan las tablas entre sí. En el **Anexo 1** se adjuntan las 5 tablas con esos detalles. A pesar de haber separado la información en 5 tablas, todas ellas están relacionadas mediante el **campo *ID***, que es el identificador único de cada tweet. Por ejemplo, si queremos saber el sentimiento general de un *tweet* cuyo texto tenemos guardado en *tweets.csv*, es suficiente con mirar en esa misma tabla su *ID* e ir a consultar el campo *document_sentiment_score* asociado a ese *ID* en la tabla de *sentiment.csv*

2.5.2. Conexiones con las APIs.

Los servicios que ofrecen las APIs son de gran valor, pero sus desarrolladores se han encargado de limitar su uso para no hacer una explotación abusiva de ellas. Esto supone una serie de obstáculos que hay que superar previamente a desarrollar el código para utilizar sus servicios.

Por ejemplo, la ***API de Twitter*** tiene varias versiones más o menos limitadas. La que se ha utilizado en este TFG es la versión estándar que, si bien es de uso gratuito, entre sus límites están que no se pueden realizar más de 450 peticiones cada 15 minutos, cada petición devuelve un máximo de 100 tweets y sólo se pueden recopilar tweets que se hayan publicado hace menos de 7 días.

Por otro lado, la ***API de Natural Language de Google*** puede usarse de manera gratuita mientras no se sobrepase el límite de 5000 peticiones mensuales. A partir de entonces, comienzan a cobrarte en función del uso que hagas de ella.

Conseguir amoldar el código a estas restricciones es un objetivo clave para tener unos datos de calidad, a bajo coste y sin desperdiciar recursos.

2.5.3. Desarrollo de los scripts.

Los scripts no son más que eso, líneas de código que realizan ciertas funcionalidades necesarias para el funcionamiento del sistema. No vamos a entrar más en profundidad, pero no está de más aclarar un par de conceptos. Los scripts los crea un desarrollador (yo en este caso) y se ejecutan de principio a fin cuando el desarrollador quiere. Por lo tanto, el sistema no funciona si los scripts no se ejecutan. Dicho esto, el ***script de descarga***, por ejemplo, está programado para solicitar *tweets* a la *API de Twitter* según ciertos parámetros de configuración. Concretamente, el script se ha ejecutado cada día durante 26 días. En cada ejecuión recopila los tweets que cumplen las siguientes condiciones: que estén en español, que se hubiesen publicado hace 6 días, que no hayan sido *retweet* y, lo más importante, que incluyan los nombres de 6 de los festivales más importantes de España: “Arenal Sound”, “BBK”, “FIB”, “Mad Cool”, “Primavera Sound” y “Sonorama”.

3. ANÁLISIS DE ENGAGEMENT Y SENTIMIENTOS

3.1. Introducción

Ya tenemos los datos, ahora han de transformarse en conocimiento y utilizarse para la toma de decisiones de marketing en el seno de la empresa. La publicación de *tweets* puede afectar de manera significativa a la percepción de una marca por el gran número de conexiones que existen y su gran capacidad de influencia (Ibrahim, Wang, & Bourne, 2017). A su vez, correctamente analizada, supone una importante fuente de información acerca de la opinión general hacia tu marca y la de tus competidores. Poseer esta información de inteligencia de marketing, resulta muy relevante para la mejora del rendimiento de una empresa (He, 2015). En este trabajo nos centraremos en medir el *engagement* y el sentimiento de los usuarios.

Al estado de estar comprometido, conectado, involucrado o interesado en algo se le conoce como ***engagement*** (Ibrahim, Wang, & Bourne, 2017). Se trata de uno de los conceptos más valiosos para una organización, pues simboliza la relación a largo plazo entre empresa y consumidor. En el contexto de las redes sociales, hace referencia a la comunicación o interacción entre los usuarios y marcas. Más concretamente en *Twitter*, lo más común es medir este *engagement* a través de métricas de interacciones como *favs*, *retweets,* respuestas o citas (SproutSocial, 2022).

Para obtener los sentimientos, emociones, opiniones u actitudes del texto escrito por un usuario, se utilizan técnicas de procesado de ese texto conocidas como **análisis de sentimiento**. En esencia consiste en asignar una emoción positiva, negativa o neutra a una frase o conjunto de ellas. La importancia empresarial de este tipo de técnicas ha quedado demostrada, por ejemplo, al estudiar la consistente correlación que existe entre el sentimiento en *Twitter* y el rendimiento de los mercados de valores (He, 2015). Por ello, se ha escogido como fuente de información para esta investigación.

La elección del análisis del sector de festivales se debe a que se trata de una industria centrada en el entretenimiento y de un público generalmente joven. Todo esto contribuye a que las opiniones que mencionan este tipo de eventos tengan una mayor carga emocional que en otros sectores, y un mayor volumen de tráfico, lo que enriquecerá el análisis y la calidad de los datos. Según (Gilstrap, 2021), el 56% de los asistentes a festivales comparten sus experiencias durante los espectáculos en directo.

Teniendo todo lo anterior en cuenta, El objetivo de este estudio se ha centrado en obtener respuesta a las siguientes preguntas a través de los datos:

1) ¿Qué patrones de *engagement* podemos identificar entre los *tweets* del estudio?
2) ¿Cuál es el sentimiento general y las tendencias de los usuarios de Twitter hacia los festivales de música en España?
3) ¿Cómo se relacionan el *engagement* y el sentimiento de los usuarios en *Twitter*?

Finalmente, las respuestas a estas preguntas serán utilizadas para proponer recomendaciones y acciones de marketing a las empresas del sector.

3.2. Características del conjunto de datos

Como punto de partida del análisis, vamos a describir las características principales del **conjunto de datos** con el que vamos a trabajar:

- Todos los *tweets* contienen, al menos, alguna de estas *keywords*: “Arenal Sound”, “BBK”, “FIB”, “Mad Cool”, “Primavera Sound”, “Sonorama”. Además, están escritos en español y se excluyen *retweets*.
- Ventana de tiempo de **26 días**: del 14 de junio al 9 de julio de 2022.
- **11929 *tweets* únicos** con todas sus métricas asociadas. Las más importantes serían los *retweets*, *favs*, datos de usuario y de sentimiento.
- En la parte superior de la Figura 3.1 se reflejan las métricas principales del conjunto de datos del que disponemos.

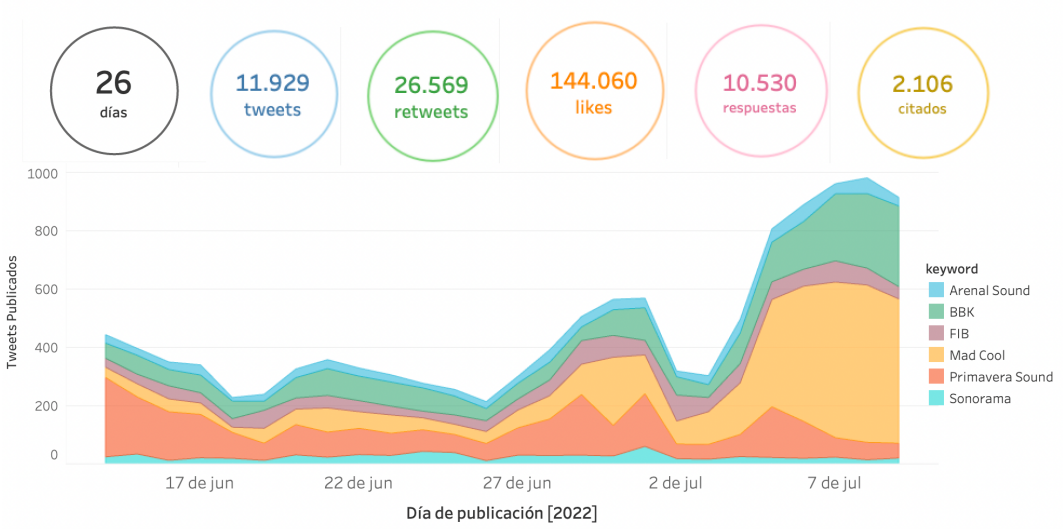

Figura 3.1: Métricas generales de los *Tweets* de festivales y fechas de publicación.

En la gráfica de áreas de la Figura 3.1, vemos la evolución de las menciones de cada festival (representado por cada *keyword*). Con esta gráfica ya podemos sacar algunas conclusiones. Por ejemplo, vemos como en los primeros días (del 14 al 19 de junio) el Primavera Sound es el festival que más tráfico genera entre los usuarios, aunque es decreciente, lo cual tiene sentido porque el festival tuvo lugar entre el 1 y el 12 de junio de 2022. La mayor parte de del tráfico se acumula hacia el final del período de estudio a medida que se acercan el resto de los festivales. Destacan sobre todo los aumentos del Mad Cool y del BBK, que comenzaban los días 6 y 7 de julio respectivamente.

3.3. Patrones de *engagement*

A través de las métricas públicas que *Twitter* nos pone a disposición, podemos analizar el nivel de implicación que los usuarios tienen respecto a las marcas o temáticas a través de las interacciones (*retweets*, respuestas, etc.). En este apartado, mediante el análisis de estas métricas, vamos a estudiar si existen comportamientos comunes o patrones. Más tarde en el capítulo 4, utilizaremos los resultados a nuestro favor en la toma decisiones de marketing en este sector.

Inicialmente en la sección 3.3.1 se analizarán las métricas públicas para la totalidad de los *tweets*. A continuación, haremos lo mismo en la sección 3.3.2,

pero únicamente para los *tweets* publicados por las cuentas oficiales de las empresas. La sección 3.3.3. se dedica a analizar si existen momentos mejores que otros para *twittear*. Finalmente, en la sección 3.3.4. se propone un análisis de temáticas y tendencias del sector de festivales de música.

3.3.1. Métricas públicas de la totalidad de *tweets*.

Al desglosar las métricas principales para cada festival, como se muestra en la Figura 3.2, tenemos una primera aproximación al nivel de *engagement* de cada festival a través de las interacciones de cada *keyword* (en el período de estudio). Destaca el *Mad Cool* como festival del que más se habla con 3986 *tweets*. Le siguen con cifras importantes el Primavera Sound y el BBK.

query_keywo..	Número de Tweets	Retweets	Favs	Respuestas	Citaciones
Mad Cool	3.986	11.345	47.582	5.292	1.169
Primavera Sound	2.853	1.474	21.827	1.634	464
BBK	2.479	9.640	63.848	1.659	233
FIB	1.270	1.468	7.402	595	115
Arenal Sound	821	1.782	897	798	39
Sonorama	637	860	2.504	552	86
Total general	11.929	26.569	144.060	10.530	2.106

Figura 3.2: Métricas principales de *engagement* desglosadas por *keywords*.

Según (SproutSocial, 2022), una métrica será más relevante para medir el *engagement* cuanto más esfuerzo suponga al usuario llevarla a cabo. Esto se observa claramente en la Figura 3.2, en la que vemos que los ***favs*** son las interacciones mayoritarias seguidas de los ***retweets***, pues son las acciones que menos esfuerzo requieren al usuario. En el otro extremo están las **respuestas**, que se producen la mitad de las veces que los *retweets,* seguidas de las **citas**, que acumulan el menor número de interacciones, pero que son de gran valor para medir ese *engagement*, pues tienen una gran implicación.

Analizando cada festival, vemos como el número de *tweets*, las citas y las respuestas siguen distribuciones similares. En general, a mayor número de tweets, mayor número de respuestas y de citas. Sin embargo, vemos que esto no ocurre con el número de *retweets* y de *favs*. De esta forma, podemos afirmar

que los niveles de *engagement* del BBK, medidos a través de estas dos métricas, son mucho mejores que los de sus competidores, pues acumulan más interacciones con niveles de tráfico bastante menores.

3.3.2. Métricas públicas de *Tweets* publicados por las cuentas oficiales.

Análogamente, podemos realizar el mismo análisis de las métricas públicas, pero restringiendo los datos únicamente para aquellas publicaciones realizadas por las cuentas oficiales de cada festival. Si bien el análisis anterior nos daba una visión general del tráfico que genera cada festival en la totalidad de los usuarios, con este análisis tendremos una visión de las interacciones de las cuentas oficiales de cada festival y nos permitirá comparar si el nivel de *engagement* es similar a través de estos canales oficiales.

Figura 3.3: Publicaciones de las cuentas oficiales de cada festival en el período de estudio.

Lo primero que llama la atención de la Figura 3.3 es que la cuenta oficial del Festival Internacional de Benicàssim (@fiberfib) no aparece, al no haber publicado ningún tweet bajo las condiciones de búsqueda en el período de estudio. Si comparamos el tráfico generado por todos los usuarios de la Figura 3.1 con el generado por las cuentas oficiales de la Figura 3.3 es claro que el grueso del uso de *keywords* proviene de cuentas no oficiales, pero también es interesante ver cómo ambas gráficas siguen unas distribuciones similares; en ambas se observa que el tráfico va disminuyendo desde los primeros días de estudio hasta aproximadamente el 26 de junio, y luego aumenta con dos picos claros a finales de junio y el primer fin de semana de junio.

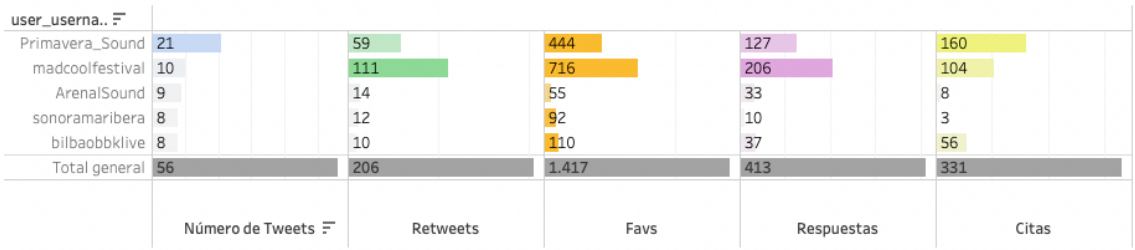

Figura 3.4: Métricas principales de *engagement* desglosadas por las cuentas oficiales.

En la Figura 3.4, se muestra el análisis de las métricas públicas de los *tweets* publicados por las cuentas oficiales de los festivales durante el período de análisis. Se arrojan algunos resultados interesantes, sobre todo si lo comparamos con los mismos datos para todos los usuarios de la Figura 3.2. Si bien en la Figura 3.4 se sigue cumpliendo que los *favs* es la interacción más utilizada, observamos cómo las citas y sobre todo las respuestas cobran mucha más importancia, si lo comparamos con los *retweets*, por ejemplo. Esto implica que las cuentas oficiales deben tener muy presentes estas métricas si quieren medir el *engagement* con el resto de los usuarios, pues tienen mucho más peso.

Figura 3.5: Análisis comparativo de las métricas públicas de 5 cuentas oficiales de festivales en el período de estudio.

Observando la Figura 3.5, podemos realizar un análisis comparativo entre las 5 cuentas oficiales para las que tenemos datos. Lo primero que observamos es que casi todas han publicado en torno a 9 *tweets*, excepto @Primavera_Sound que ha *tweeteado* más del doble (21 *tweets*). Además, se adjunta una tabla con el promedio de seguidores de cada cuenta en el período de estudio, pues es importante para contextualizar los números ya que es esperable que, a más seguidores, más interacciones. Sin embargo, se pone de manifiesto que la cuenta @madcoolfestival es la que goza de mejores niveles de *engagement* a través de estas métricas. Podemos afirmar esto porque, a pesar de ser la segunda cuenta con menos seguidores y habiendo *tweeteado* de manera similar a sus competidores, es la clara vencedora en 3 de las cuatro métricas representadas en términos absolutos. Siguiendo con las gráficas de sectores, destaca negativamente el bajo nivel de interacciones que acumula la cuenta de @ArenalSound teniendo en cuenta la gran base de seguidores con la que cuenta. Seguramente sea explicable por la lejanía del festival (2 de agosto).

En la parte inferior de la Figura 3.5, se han calculado las ratios métricas/*tweet*, para eliminar las diferencias entre el número de publicaciones de cada cuenta. En estas gráficas se ve aún más claro como @madcoolfestival es la que cosecha mejores resultados también en términos relativos, estando siempre en valores superiores a 10 interacciones por cada tweet en las 4 métricas. En segundo lugar, está siempre @Primavera_Sound que también tiene buenos resultados gracias a tener la mayor base de seguidores. En tercer lugar, está @bilbaobbklive cuyos valores son aceptables si tenemos en cuenta que cuenta con menos de la mitad de los seguidores que el @Primavera_Sound, pero bastante mejorables respecto al @madcoolfestival ya que cuenta con más seguidores. Tanto @sonoramaribera como @ArenalSound tiene los niveles más bajos. Sabiendo que son festivales que se celebran en agosto, pone de manifiesto cómo las interacciones aumentan con la cercanía de los eventos. Sin embargo, es cierto que los números del @ArenalSound deberían mejorar si tenemos en cuenta que tiene más de 110.000 seguidores.

3.3.3. Patrones temporales.

En lo referente al tiempo, se han evaluado 2 tipos de patrones de *engagement*: el **patrón de horas** y el de **días de la semana**. El objetivo es utilizar los datos para ver qué horas y días de la semana son más adecuadas para *twittear*. Para ello, vamos a utilizar una tasa de *engagement* y la representaremos por horas. Esa tasa se calcula como la agregación de todas las métricas públicas divididas entre el número de *tweets* que las generan. Es decir, la suma de *retweets*, *favs*, respuestas y citas que recibe en promedio cada *tweet*. En la Figura 3.6, se utiliza un mapa de calor para representar esa métrica por horas para los 26 días del estudio. En la parte inferior de la Figura, tenemos el análisis únicamente por horas, en el que queda claro que las horas que menos interacciones acumulan son entre las 4:00 y las 9:00 de la mañana. Por el contrario, las mejores franjas son entre las 13:00-16:00, las 19:00-23:00 y las 00:00-3:00.

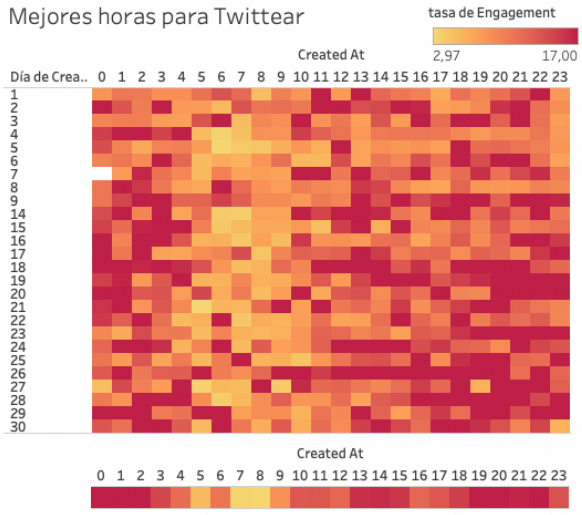

Figura 3.6: Mapa de calor de la métrica de *engagement* desglosado por horas y días (arriba) y sólo por horas (abajo)

Para saber si existe algún día de la semana mejor que otro para obtener más interacciones se han creado las gráficas de la Figura 3.7. En la parte izquierda de la Figura se muestra el diagrama de caja y bigotes del total de interacciones. Las cajas contienen el 50% de los datos, y están divididas por la mediana. Los bigotes representan los valores máximos y mínimos. Si nos fijamos en las cajas, podemos intuir que los lunes sábados y domingos son los días con mejores tasas de *engagement*, aunque con mucha dispersión en el caso de los domingos y

lunes. Esto se confirma claramente en la gráfica de la derecha, en la que se representan el 60% y 80% del promedio. Gracias a ella, es posible visualizar claramente que los días con mejores niveles de *engagement* son los fines de semana. Y entre semana destacan positivamente el lunes y el jueves.

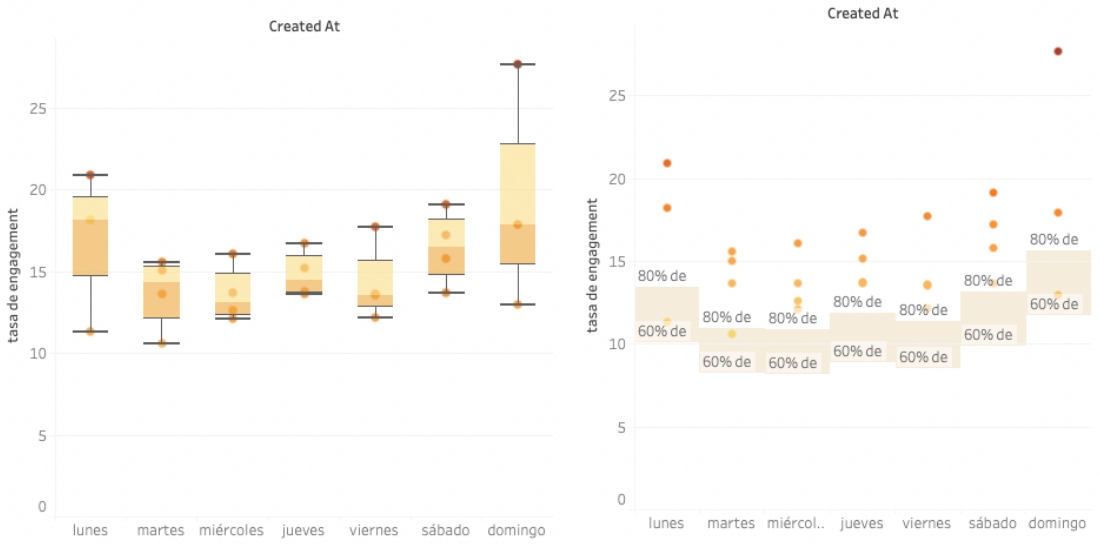

Figura 3.7: Análisis de la tasa de *engagement* por día de la semana.

3.3.4. Análisis de temáticas y entidades

Gracias a los datos recopilados en la *Tabla de annotations* que se puede consultar en el Anexo 1, tenemos información que la propia *API* de *Twitter*, proporciona sobre el contexto de cada tweet. El campo *context_annotations_entity_name* nos da información sobre las **temáticas** de las que trata el *tweet*. En un nivel superior, estas temáticas se agrupan en **dominios**, guardados en el campo *context_annotations_domain_name*. Debemos tener en cuenta que ambos campos están en inglés.

Las posibilidades de análisis de esta información combinada con otras métricas ya analizadas en el trabajo son casi infinitas. Por ejemplo, podemos crear nubes de palabras para ver qué temáticas son más frecuentes entre los tweets que mencionan a los festivales. A la izquierda de la Figura 3.8 se muestra el resultado de los 20 dominios más comunes entre todos los *tweets* del estudio. Ambas nubes de palabras son interactivas, lo que quiere decir que, si seleccionamos un elemento de una de ellas, esa selección actúa como filtro sobre la otra. Como en

la Figura 3.8 tenemos seleccionado el dominio de géneros musicales (“music genre”), a la derecha se muestran más grandes aquellos géneros sobre los que más se habla en nuestro conjunto de datos.

Figura 3.8: Nubes de palabras de las temáticas de los *tweets* a través de dominios (izquierda) y entidades (derecha).

Un ejercicio muy interesante es analizar los músicos (del dominio “musicians”) sobre los que más se habla entre las menciones de dos festivales diferentes. Los resultados que se muestran en la Figura 3.9 comparan los músicos más mencionados entre los *tweets* del Mad Cool y del BBK. Es una información muy valiosa, para saber cuáles son las tendencias o grupos más deseados entre tu audiencia. Así pues, aunque muchos son compartidos entre ambos festivales, *Metallica* o *Imagine Dragons* aparecen mucho más frecuentemente en las conversaciones del Mad Cool, y lo mismo ocurre con *Phoebe Bridgers* o *LCD Soundsystem* en las del BBK.

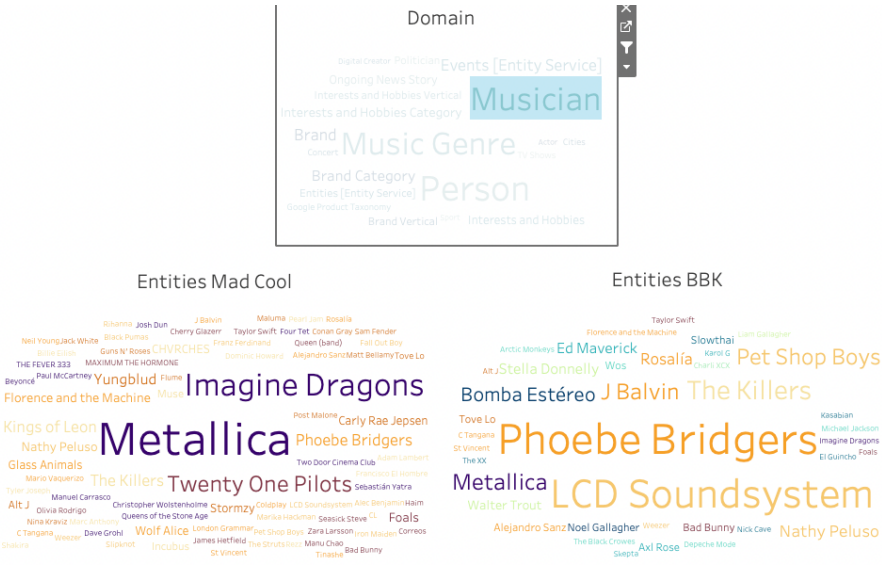

Figura 3.9: Frecuencia con la que se habla de diferentes músicos en las menciones al Mad Cool y al BBK.

3.4. Análisis de sentimientos

El análisis de sentimientos es un área creciente dentro del Procesamiento del Lenguaje Natural (NLP) (Esuli & Sebastiani, 2010). Consiste en procesar el contenido textual de un texto para obtener un resultado del sentimiento o la opinión predominante. Al aplicarlo a los *tweets*, podemos obtener unos resultados muy valiosos para una empresa en términos de percepción de la marca, satisfacción de los consumidores y relación consumidor-empresa.

En este apartado, inicialmente se explica la metodología utilizada para llevar a cabo el análisis. En la sección 3.4.2. se exponen los resultados del sentimiento general de los usuarios que mencionan a los festivales. A continuación, en la sección 3.4.3. se profundiza un poco más en el análisis, a través de los resultados de opiniones sobre las entidades detectadas en el texto y, más concretamente, en los festivales. Finalmente, en la sección 3.4.4. se analizan las relaciones entre *engagement* y sentimiento.

3.4.1. Metodología

Para este trabajo se ha utilizado la *API* de *Natual Language* de *Google* para obtener estos resultados. Esta *API,* procesa los textos con modelos de Inteligencia Artificial pre-entrenados y devuelve unos resultados numéricos (Google, 2022). En la Sección 2.4.1.2 ya se adelantó que se iban a utilizar dos servicios distintos para este análisis. El primero es el de **análisis de opiniones** para obtener la a actitud del *tweet* en general y de las frases que lo componen en particular (para más información consultar los campos de la *Tabla de sentiment* del Anexo 1). El segundo es el de **análisis de opiniones de entidades** que es un análisis más profundo ya que extrae las entidades que menciona el autor en el texto y el sentimiento que éste tiene hacia ellas (más información de los campos en la *Tabla de entities_sentiment* del Anexo 1).

Antes de ir con los resultados, es importante tener en cuenta algunos conceptos básicos. El sentimiento o *sentiment_score* (ya sea de todo el *tweet* o de sus entidades) toma valores numéricos donde **+1** es un **sentimiento positivo** y **-1** es un **sentimiento negativo**. Los valores intermedios representan **sentimientos neutros** si no tienen actitudes positivas ni negativas, o **mixtos** si tienen ambos

tipos de opiniones y se compensan en el resultado. Para discernir entre ambas, entra en juego la **magnitud** o *sentiment_magnitude*, que no está normalizada (puede tomar valores absolutos mayores que 1) e indica la intensidad de la emoción. Para una mejor compresión se adjunta en la Figura 3.10 las posibles interpretaciones del sentimiento a la combinar ambas métricas.

Opinión	Valores de muestra
Claramente positiva*	`"score"`: 0.8, `"magnitude"`: 3.0
Claramente negativa*	`"score"`: -0.6, `"magnitude"`: 4.0
Neutral	`"score"`: 0.1, `"magnitude"`: 0.0
Mixto	`"score"`: 0.0, `"magnitude"`: 4.0

Figura 3.10: Interpretación de resultados del análisis de sentimientos (Google, 2022).

El último concepto importante es el de la **saliencia** o *salience_score*, que se utiliza en el análisis de opiniones de entidades e indica la relevancia de cada entidad dentro del conjunto de todo el texto.

3.4.2. Análisis de sentimientos general de cada tweet.

En este trabajo se han procesado todos los *tweets* en los que se mencionaba al menos uno de los 6 festivales bajo estudio. Para poder interpretar los resultados numéricos se han fijado ciertos umbrales del campo *sentiment_score* para definir la métrica que he denominado **sentimiento**. Esto queda reflejado en la Figura 3.11, en la que también se representa que aquellos textos con *sentiment_score* entre -0.2 y +0.2 se catalogarán con sentimiento **NEUTRAL** siempre que la magnitud sea menor o igual que 1; en caso de que sea mayor que 1, será **MIXTO**.

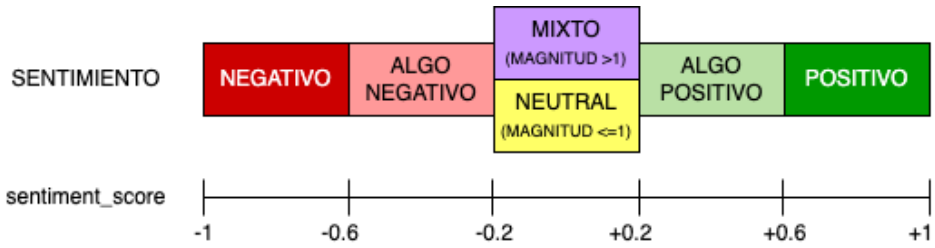

Figura 3.11: Umbrales de para el cálculo del sentimiento. Elaboración propia.

De los 11.929 *tweets* recopilados en los que se mencionaban los 6 festivales, se ha realizado el análisis del sentimiento general del texto para cada uno de ellos. Como ya hemos visto, no todos los festivales (o *keywords*) tienen la misma cantidad de tráfico, por lo que es mejor comparar cada tipo de sentimiento en términos porcentuales tal y como se representa en la Figura 3.12.

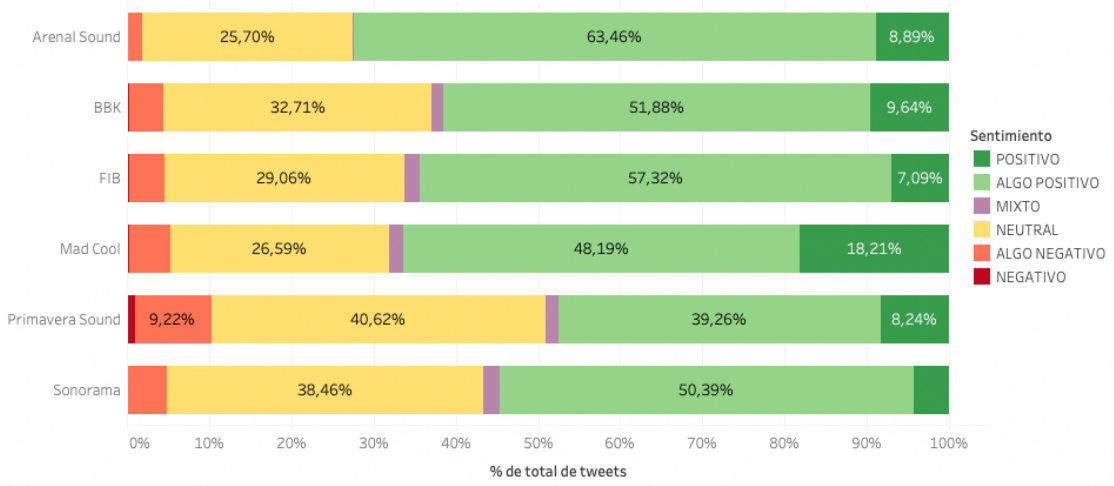

Figura 3.12: Porcentaje de sentimiento por cada *keyword*.

Como se muestra en la Figura 3.12, en el acumulado de *tweets* "positivos" o "algo positivos" es mucho mayor que sus análogos negativos (los cuáles no suelen pasar del 10%). Si nos centramos en los sentimientos negativos, destaca el Primavera Sound por sus altos niveles, mientras que en el Arenal Sound estos valores no llegan al 2%. Del lado positivo, es interesante ver cómo también es el Arenal Sound el que sobrepasa el 70% de tweets positivos. Además, es importante comentar cómo sobresale positivamente el Mad Cool con ese más de 18% de *tweets* muy positivos mencionando el evento.

La distribución de los diferentes sentimientos en el Primavera Sound demuestra que es el festival con opiniones más diversas. Los promotores del evento deberían poner especial interés en estos datos, pues se trata de tweets publicados tras su finalización y seguro que todos esos comentarios, sobre todo los más extremos, guardan valiosa información sobre lo que ha ido bien y ha ido mal.

Por último, hay que comentar que son pocos los tweets con sentimiento “mixto”. Esto demuestra que cuando un usuario publica en la red social, lo hace con un sentimiento definido y no es habitual que se mezclen actitudes contrarias.

3.4.3. Análisis de sentimientos de entidades

Mientras que en la sección anterior la información del sentimiento se refería a todo el contenido del *tweet* en general, gracias a los datos de la *Tabla de entities sentiment* (descritos en el Anexo 1) podemos obtener una información de mayor calidad respecto a la opinión directa del autor del *tweet* referida a cada entidad que menciona.

Para obtener los resultados de este análisis se han seleccionado, entre todas aquellas entidades detectadas por la API de Google, aquellas que se referían a cada uno de los 6 festivales por su nombre. Después se ha creado el campo del sentimiento de la entidad con los mismos umbrales escogidos de la Figura 3.11 y se ha representado en la Figura 3.13 la distribución del sentimiento de cada evento. Así pues, mientras en la Figura 3.12 se analizaba el sentimiento general del tweet en el que se mencionaba cierta *keyword*, en la Figura 3.13 se representa el sentimiento hacia esa *keyword* concreta. Lo primero que llama la atención al comparar ambas Figuras es que, cuando el análisis de sentimiento va referido a entidades concretas y no al tweet en general, disminuyen enormemente los sentimientos positivos en favor del aumento de los neutrales y los negativos.

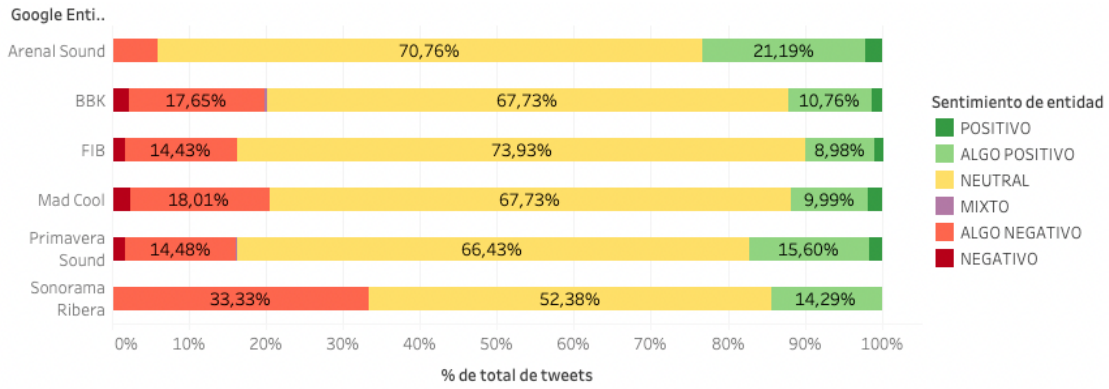

Figura 3.13: Sentimiento hacia los eventos detectados como entidades.

Este análisis resulta tremendamente útil para las marcas en concreto, pues eliminan el “ruido” que pudiera existir en el contenido del *tweet* al calcular el sentimiento, obteniendo una medida mucho más directa y acertada de la opinión de los usuarios concretamente hacia su marca.

Si combinamos esta gráfica, por ejemplo, con una tabla con información de los *tweets* coincidentes, podemos prestar especial atención a los usuarios que expresan unas emociones más fuertes, así como detectar cuáles requieren una atención más inminente. A modo ilustrativo en la Figura 3.14, se observa cómo al filtrar por los *tweets* con sentimiento negativo del Primavera Sound, podemos ver exactamente qué dicen esos *tweets*, quién los ha escrito y cuántos seguidores tiene. Además, podemos atender antes aquellos que están teniendo más *retweets* o *favs* intentando que no se difunda mala propaganda.

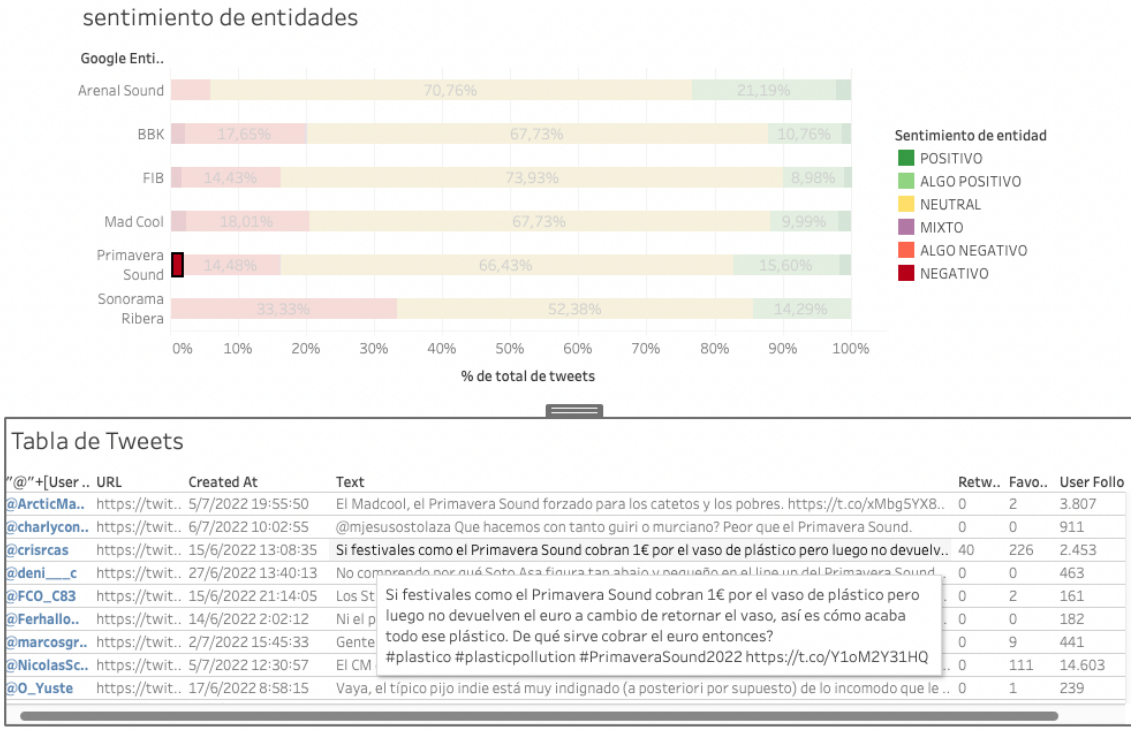

Figura 3.14: Filtro de *tweets* con sentimiento NEGATIVO hacia la entidad Primavera Sound.

3.4.4. Relaciones entre el sentimiento y el *engagement* del cliente

En esta sección, se han utilizado los datos de análisis de sentimiento para evaluar si existen relaciones entre el sentimiento general de un *tweet* y diferentes

factores de *engagement* como las métricas públicas, las menciones, los *hashtags,* o los usuarios verificados.

Efecto del sentimiento sobre las métricas públicas

Ya hemos visto que las diferentes interacciones de los usuarios con los *tweets* se pueden utilizar para medir el *engagement*. Por ello, el primer análisis entre la relación sentimiento-*engagement* se realiza a través de estas métricas públicas.

En la Figura 3.15 se representa cuál es el nivel de *engagement* que tiene cada *tweet* en función del sentimiento general de su contenido textual. Lo primero que hay que resaltar es que los *tweets* catalogados como “algo positivos”, son los que mejores niveles de *engagement* tienen en todas las métricas. Es decir, según las gráficas, los *tweets* con este tipo de opinión reciben un promedio de 13,97 favoritos, 2,68 *retweets*, 1,06 respuestas y 0,2 citas. Lo siguiente a recalcar es que las distribuciones de *retweets* y favoritos según el sentimiento, tienen la misma forma. De ello extraemos que, en media, un tweet “algo positivo” va a tener 6 veces más *retweets* que uno “algo negativo” y 3 veces más favoritos. Si observamos las respuestas podemos afirmar que los textos con actitudes negativas reciben en media menos respuestas que el resto de las emociones, pero las diferencias son menores que en otras métricas. De las citas, podemos destacar cómo los *tweets* negativos ganan terreno respecto al resto de métricas.

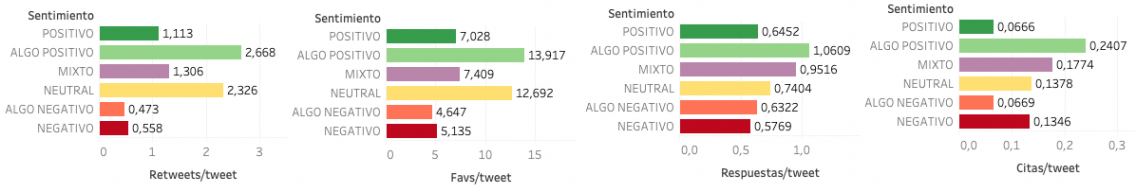

Figura 3.15: Ratios de *engagement* para cada tipo de sentimiento.

Efecto de los usuarios verificados

Twitter tiene un sistema de verificación de usuarios famosos o influyentes de su red social, que los distingue del resto de usuarios. Bajo la hipótesis de que este tipo de usuarios impactará sobre un mayor número de personas, la opinión que los usuarios verificados emitan sobre tu marca debería tener mayor peso que la del resto.

Una muestra clara de que los mensajes de los usuarios verificados llegan a más usuarios es la tabla de la Figura 3.16. En la última columna se representa el alcance de cada *tweet* calculado como la ratio entre el número de seguidores y el número de *tweets.* Este alcance es claramente superior entre los usuarios verificados superando siempre audiencias de 230.000 seguidores por cada *tweet*.

query_keyword	user_verified	Número de Tweets	Seguidores	Alcance de cada tweet
Arenal Sound	False	818	1.764.821	2.157
	True	3	82.017	27.339
BBK	False	2.283	10.558.801	4.625
	True	196	46.417.567	236.824
FIB	False	1.214	3.682.999	3.034
	True	56	25.226.463	450.473
Mad Cool	False	3.689	9.987.357	2.707
	True	297	149.092.147	501.994
Primavera Sound	False	2.750	8.462.240	3.077
	True	103	58.148.558	564.549
Sonorama	False	608	20.472.538	33.672
	True	29	8.010.971	276.240

Figura 3.16: Comparativa de los *tweets* y los seguidores de los usuarios verificados vs. los no verificados para cada *keyword*.

Una vez visto que sus comentarios son más relevantes, vamos a comparar qué actitud suelen tener este tipo de usuarios respecto al resto. Es muy revelador ver cómo, para todos los festivales de la Figura 3.17, la proporción de comentarios positivos aumenta entre los usuarios más influyentes y las actitudes negativas se reducen llegando casi a anularse. Llama la atención el 100% de conversaciones positivas sobre el Arenal Sound.

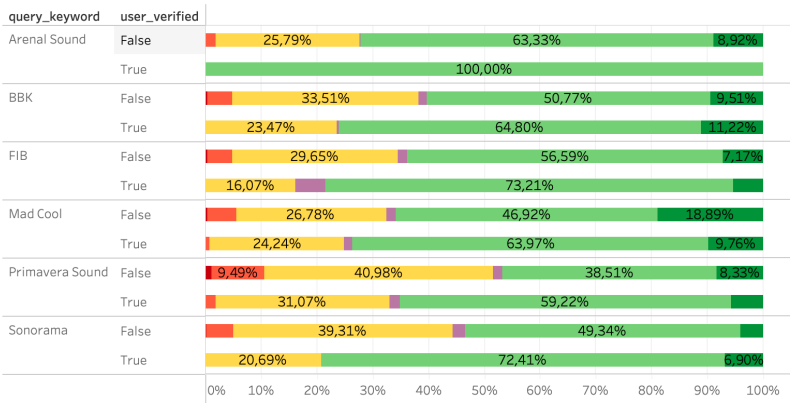

Figura 3.17: sentimiento de los *tweets* de los usuarios verificados y no verificados. Elaboración propia.

Efecto de los hashtags y menciones.

Los usuarios, al publicar contenido en Twitter, hacen uso de funcionalidades como lo ***hashtags*** para etiquetar o clasificar sus mensajes o de las **menciones** para dirigirse a otros usuarios. Con la información de la *Tabla de entities,* que está detallada en el Anexo 1, podemos evaluar el efecto este tipo de funcionalidades sobre el sentimiento. Como se muestra en la Figura 3.18, hay un patrón claro en todas las *keywords* y es que cuando los usuarios utilizan las **menciones**, lo hacen para comunicar opiniones mucho más dispersas y extremas (ya sea negativas o positivas) que cuando usan ***hashtags***. Es importante para las marcas tener esto en cuenta, pues cuando un usuario te menciona, es probable que tenga una mayor conexión emocional con tu marca que si utiliza un *hashtag*.

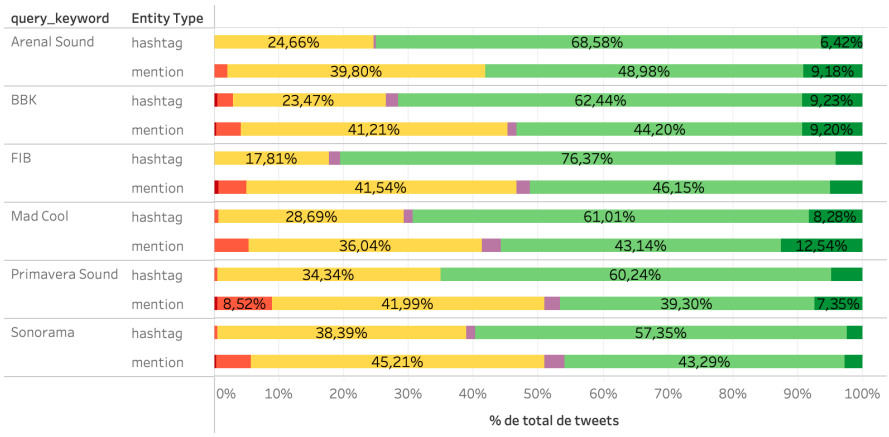

Figura 3.18: Sentimiento de los *tweets* de los que usan *hashtags* y menciones.

4. CONCLUSIONES.

En los últimos años, aquellas empresas que están invirtiendo en inteligencia de marketing a través de la recopilación y análisis de datos de diversas fuentes, se encuentran en una posición favorable respecto a su competencia a la hora de identificar áreas específicas de acción en las que los negocios están sobresaliendo, o errando, para mejorar la experiencia de los consumidores y la notoriedad de la marca. Con este objetivo y basándonos en los resultados de los análisis del capítulo anterior, se enumeran en las siguientes subsecciones algunas acciones orientadas a la estrategia de marketing en Twitter.

4.1.1. Recomendaciones basadas en patrones de *engagement*.

A la luz de los resultados expuestos en la sección 3.3, se proponen una serie de recomendaciones de marketing en *Twitter*, que pueden ser útiles para una empresa del sector de festivales:

i) **Monitoriza las métricas públicas de tu cuenta** y plantéate objetivos de crecimiento medibles y reales. Tener un gran número de seguidores, es importante para mejorar niveles de *engagement*, pero no lo es todo. Ejemplos como la cuenta del Mad Cool, demuestran tener muchas interacciones sin tener muchos seguidores.

ii) Impulsa la participación de los usuarios a través de **citas y respuestas**. Son las interacciones más difíciles de conseguir, pero las que te ayudarán a tener una mayor presencia en las redes.

iii) **Twittea** para mejorar la relación con tu audiencia. Tus seguidores más fieles están pendientes de tus publicaciones. Si no publicas nada, como el caso del FIB, estarás contribuyendo a un deterioro de esta relación a la vez que pierdes la oportunidad de llegar a nuevos clientes.

iv) **Programa tus *tweets***. Obtendrás más impacto entre los usuarios si evitas *twittear* en las horas de la madrugada hasta el mediodía. Además, en el sector de festivales, los mejores días en términos de *engagement* son los domingos, lunes, sábados y miércoles. Ten en cuenta que tu marca tendrá más presencia en el sector cuanto más cercana sea la fecha del festival. Puedes aprovechar esto para lanzar campañas, mejoras de imagen o captación de nuevos clientes.

v) **Observa las tendencias del sector** y sobre todo de tus **consumidores**. El análisis de los músicos más presentes entre las conversaciones de tus espectadores es un gran indicativo sobre la expectación que estos generan. Además, es posible que descubras artistas que no tienes en el cartel, pero que encajan con los gustos de tus audiencias y serían candidatos perfectos para próximas ediciones.

vi) **Observa a tus competidores**. Esto es aplicable a todos los datos anteriores. Aprende qué hacen bien las cuentas con mejores resultados y evita caer en los mismos errores que las que tienen peor rendimiento.

4.1.2. Recomendaciones basadas en el sentimiento del cliente.

Teniendo en cuenta los resultados referentes al análisis de sentimientos de la sección 3.4, se sugieren a continuación una serie de recomendaciones orientadas a mejorar la presencia de una marca de un festival en *Twitter* y su relación con los usuarios:

i) **Monitoriza los sentimientos de tu marca**. Se ha demostrado que son una fuente muy valiosa de la percepción de la marca entre los consumidores.

ii) **Interactúa** y haz sentir valiosas a aquellas personas que tienen un **sentimiento positivo** hacia tu festival, pues eso hará que se cree una relación duradera año tras año.

iii) **Atiende** a aquellos que emiten **opiniones negativas**. Esas opiniones pueden influir negativamente en tu imagen. Una atención privada y personalizada a través de mensajes directos puede terminar con esos comentarios.

iv) **Escucha a tu audiencia**, sobre todo **los días del festival y los posteriores**, pues proporcionan un *feedback* muy valioso sobre los errores y aciertos de las decisiones de la empresa.

v) **Presta especial atención a los usuarios verificados** porque el alcance de sus seguidores hace que tengan mucha más influencia y pueden ser la puerta a muchos nuevos clientes o detractores.

vi) **Revisa los *tweets* con menciones** antes que los *hashtags*, pues suelen contener una mayor diversidad de emociones.

vii) Por último, trata de ser **algo positivo** en tus publicaciones. Ya no sólo por tu imagen de marca, sino porque es la emoción que tiene mejores tasas de *engagement*.

4.1.3. Conclusiones

Me gustaría concluir respondiendo a las preguntas que nos fijamos en los objetivos del estudio.

En primer lugar, se han detectado diferentes patrones para medir el *engagement* para el conjunto de datos recogidos. Por ejemplo, que es esperable, que el número de *retweets* siga una correlación directa con los *favs*, pero no con las citas o las respuestas. También se pone de manifiesto que, existen unas franjas horarios mejores que otras para Twitter o que podemos encontrar conversaciones de temáticas parecidas entre el público de un festival concreto y que no tienen por qué coincidir con los de otro festival.

También se han obtenido resultados satisfactorios respecto al análisis de sentimientos de los usuarios. Esto resulta increíblemente útil para, inicialmente, diagnosticar la emoción general entre las conversaciones que mencionan a una marca y, posteriormente, identificar y atender de una manera más eficaz y personalizada a aquellos usuarios con opiniones más extremas.

Respecto a la relación de los sentimientos con el *engagement*, se han realizado algunos análisis que han demostrado que el volumen de interacciones que un usuario cosecha en la plataforma tiene relación con su sentimiento. Al igual que el sentimiento está relacionado con otras variables como la condición de ser usuario verificado o el uso de *hashtags* y menciones.

A nivel personal, este trabajo me ha resultado tan complejo como satisfactorio. Complejo por la cantidad de obstáculos a superar. Comenzando por delimitar el alcance del trabajo, luego desarrollando el sistema y finalmente identificando resultados relevantes para la empresa. Satisfactorio por haber superado todas

esas dificultades, obteniendo unos resultados que considero relevantes tanto a nivel académico como empresarial.

4.1.4. Limitaciones y futuras líneas de trabajo.

Sin lugar a duda, el mayor reto que ha supuesto este trabajo es fijar el alcance de este tanto en el desarrollo cómo en el análisis. Esto se ha conseguido mediante un proceso iterativo en el que, semanalmente, había que reevaluar hasta dónde se quería llegar en la investigación y redefinir los límites para conseguir una solución valiosa, pero también efectiva y factible. Esto tiene de manera inherente ciertas limitaciones, que hay que tener presentes. Por ejemplo, el sentimiento de *Twitter* es sólo una parte de todo el sentimiento que expresan los usuarios en todas las redes sociales. Al no ocurrir todos los festivales a la vez, puede existir un sesgo en la muestra debido a la cercanía o lejanía con las fechas de festival. Por último, como resultado de la delimitación del alcance, se han dejado fuera del análisis muchas variables sobre las que se podría investigar la relación con el *engagement* y/o sentimiento de los usuarios. En este sentido, surgen unas líneas de trabajo futuras, principalmente relacionadas con la mayor explotación del conjunto de datos obtenidos. Por ejemplo, analizando en profundidad los datos de la tabla de sentimiento de entidades, o estudiando otras dimensiones de los usuarios de las que disponemos para obtener una foto más nítida de nuestros consumidores y sus características.

BIBLIOGRAFÍA

Esuli, A., & Sebastiani, F. (2010). Machines that learn how to code open-ended survey data. *International Journal of Market Research.*, 775-800.

Gilstrap, C. T. (2021). Social music festival brandscapes: A lexical analysis of music festival social conversations. *Journal of Destination Marketing & Management*, 100567.

Google. (julio de 2022). *Google Cloud*. Obtenido de Conceptos básicos de la API de Natural Language: https://cloud.google.com/natural-language/docs/basics#interpreting_sentiment_analysis_values

He, W. W. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, 801-812.

Ibrahim, N. F., Wang, X., & Bourne, H. (2017). Exploring the effect of user engagement in online brand communities: Evidence from Twitter. *Computers in Human Behaviour*, 321-338.

Kumar, V., Chattaraman, V., Neghina, C., Skiera, B., Aksoy, L., Buoye, A., & Joerg, H. (2013). Data-driven services marketing in a connected world. *ournal of Service Management*, 330.

Pandas. (julio de 2022). Obtenido de Pandas Documentation: https://pandas.pydata.org/docs/

Python. (julio de 2022). Obtenido de Python Documentation: https://www.python.org/doc/

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*.

Sarson, T., & Gane, C. (1979). *Structured Systems Analysis.* Prentice Hall.

SproutSocial. (julio de 2022). Obtenido de sproutsocial.com: https://sproutsocial.com/es/twitter-marketing

Tableau. (julio de 2022). Obtenido de ¿Qué es Tableau?: https://www.tableau.com/es-es/why-tableau/what-is-tableau

Tuten, T. L., & Solomon, M. R. (2015). *Social Media Marketing.* SAGE Publications.

Tweepy. (julio de 2022). *Tweepy Documentation*. Obtenido de https://docs.tweepy.org/en/stable/

Twitter. (julio de 2022). *Developer Platform*. Obtenido de https://developer.twitter.com/en/docs

Valacich, J. S., & George, J. F. (2017). Modern Systems Analysis and Design. *Pearson*.

ANEXO 1. TABLAS DEL MODELO DE DATOS

Tabla de tweets: Contiene toda la información relativa a los tweets y al usuario		
campo	**descripción**	**tipo de dato**
ID	Identificador único del tweet	cadena
text	Texto del tweet en UTF-8	cadena
URL	URL del tweet	cadena
created_at	Fecha y hora de creación	fecha y hora
query	Cadena de petición a la API	cadena
query_keyword	Keyword de la petición	cadena
author_id	Identificador único del usuario	cadena
searched_at	Fecha y hora de petición	fecha y hora
source	App desde la que se publicó	cadena
in_reply_to_user_id	Si es una respuesta, representa el author ID del tweet original	cadena
retweet_count	Número de retweets	entero
favorite_count	Número de favoritos	entero
reply_count	Número de respuestas	entero
quote_count	Número de citaciones	entero
lang	idioma	cadena
user_name	Nombre del perfil del usuario	cadena
user_username	el usuario único pero mutable	cadena
user_location	Localización especificada en el perfil del usuario	cadena
user_profile_image_url	Url de la imagen de perfil	cadena
user_protected	True si sus tweets son privados	booleano
user_followers_count	Número de seguidores	entero
user_following_count	Número de seguidos	entero
user_tweet_count	Número de tweets	entero
user_listed_count	Número de listas públicas de las que el usuario es miembro	entero
user_url	Url del perfil.	cadena
user_verified	True si el usuario está verificado.	booleano

Tabla de annotations: proporcionan información contextual del tweet derivada del análisis del texto e incluye un emparejamiento entre dominio y entidad que se puede utilizar para analizar las temáticas de los tweets.
Los tweets se analizan y como resultado se infieren domains y/o etiquetas de entity

campo	descripción	tipo de dato
context_ID	Identificador único del tweet	cadena
context_annotations_domain_id	Número del dominio del contexto del tweet	cadena
context_annotations_domain_name	Nombre del dominio del contexto del tweet	cadena
context_annotations_entity_id	Identificador único de la entidad	cadena
context_annotations_entity_name	Nombre de la entidad asociada al dominio	cadena

Tabla de entities: proporciona información adicional del tweet como hashtags, cashtags, urls o menciones. Esta información se obtiene tras un parseo del texto. Lo que más juego da son las annotations que proporcionan información contextual de entidades (personas, lugares, productos, y organizaciones). Se extraen basadas en los que se menciona explícitamente en el tweet. *Documentación*

campo	descripción	tipo de dato
entity_tweet_ID	Identificador único del tweet	cadena
entity_type	Tipo de entidad extraída: annotation, url, hashtag, mention o cashtag	cadena
entity_name	nombre de la entidad. Depend del tipo	cadena
entity_start	Indice inclusivo donde comienza la entidad	entero
entity_end	indice exclusico donde termina la entidad. Excepto en annotations que es inclusivo	entero
entity_annotation_probability	Si *entity_type=annotation*, se refiere a la puntuación de confianza de que esa entidad sea correcta	decimal
entity_annotation_type	Si *entity_type=annotation*, indica si es de tipo persona, lugar, producto, organización u otro.	cadena
entity_mention_id	Si *entity_type=mention*, el author id del usuario que se menciona	cadena
entity_url_expanded_url	Si *entity_type=url*, indica la url expandida	cadena
entity_url_media_key	Si *entity_type=url*, indica el identificador único del contenido multimedia expandido	cadena

Tabla de sentiment: proporciona el sentimiento y su magnitud de todo el texto del tweet en general. Además, para cada tweet habrá tantas filas como oraciones se detecten. Para cada una de estas oraciones, tendremos también un score y una magnitud.		
campo	**descripción**	**tipo de dato**
sentiment_tweet_ID	Identificador único del tweet	cadena
document_sentiment_score	sentimiento general del tweet proporcionado por la API de sentiment de Google	decimal
document_sentiment_magnitude	intensidad general de la emoción proporcionada por la API de sentiment de Google	decimal
sentence_text	texto de la frase detectada asociada al ID del Tweet	cadena
sentence_sentiment_score	sentimiento de la frase detectada del tweet proporcionado por la API de sentiment de Google	decimal
sentence_sentiment_magnitude	intensidad de la emoción de la frase detectada proporcionada por la API de sentiment de Google	decimal

Tabla de entities_sentiment: cada fila representa una entidad extraída a través del análisis de entidades de google. Para cada entidad tenemos el tweet ID asociado y métricas de sentimento.		
campo	**descripción**	**tipo de dato**
google_entity_tweet_ID	Identificador único del tweet	cadena
google_entity_name	el nombre de la entidad detectada	cadena
google_entity_type	tipo de esta entidad (por ejemplo, si la entidad es una persona, una ubicación, un bien de consumo, etc.)	cadena
google_entity_salience_score	indica la importancia o relevancia de esta entidad para todo el texto del documento.	decimal
google_entity_sentiment_score	sentimiento general del tweet proporcionado por la API de entities de Google	decimal
google_entity_sentiment_magnitude	intensidad general de la emoción proporcionada por la API de entities de Google	decimal
google_entity_mention_type	indica si la mención ha sido de una entidad de nombre común o propio.	cadena