Las Enfermedades Raras en grupos de Facebook: evaluación de la línea de actuación de FEDER mediante minería de textos

Natalia Reguera Terrazo
Posgrado en Inteligencia de Negocio y Análisis de Datos
Análisis de datos

Dra. Laia Subirats Maté / Dr. Manuel Armayones Ruiz
Dra. Maria Pujol Jover / Dra. Teresa Sancho Vinuesa

20 de Febrero de 2017

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Las Enfermedades Raras en grupos de Facebook: evaluación de la línea de actuación de FEDER mediante minería de textos*
Nombre del autor:	*Natalia Reguera Terrazo*
Nombre del consultor/a:	*Dra. Laia Subirats Maté y Dr. Manuel Armayones Ruiz*
Nombre del PRA:	*Dra. Maria Pujol Jover y Dra. Teresa Sancho Vinuesa*
Fecha de entrega (mm/aaaa):	02/2017
Titulación:	*Posgrado en Inteligencia de Negocio y Análisis de Datos*
Área del Trabajo Final:	*Análisis de un conjunto de datos*
Idioma del trabajo:	*Castellano*
Palabras clave	*Análisis de datos, red social, enfermedad*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El objeto de este trabajo es el estudio de grupos de Facebook que tratan las enfermedades raras y su relación con el Decálogo de Prioridades de la Federación Española de Enfermedades Raras (FEDER). Para ello se ha utilizado la herramienta Netvizz para descargar los datos y mediante el uso de diversas librerías en R se han realizado nubes de palabras. Se ha utilizado TextBlob en Python para realizar análisis de sentimientos. Además, se hace realiza el cálculo del log-likelihood para comparar las palabras obtenidas en Facebook y en el decálogo. Los resultados muestran que las fotos son el tipo de publicación que genera mayor número de 'likes', reacciones y 'engagement'. Además, podemos observar que las polaridades positivas también generan un nivel de 'engagement' mayor, y que sin embargo la subjetividad no está tan ligada al 'engagement'. Los resultados de la minería de textos muestran que los grupos de Facebook están más preocupados por las familias, las personas y las ayudas sin embargo el Decálogo está más enfocado en las discapacidades, la formación y la enfermedades. En cuanto a la influencia de las variables temporales podemos concluir que existe significación estadística entre la polaridad del mensaje y la hora de emisión de los posts. Por todo esto sería recomendable que dentro de la línea de actuación de FEDER se tomasen acciones para dar más apoyo a las familias.

Abstract (in English, 250 words or less):

The purpose of this work is the study of Facebook groups dealing with rare diseases and their relation with the Priorities Decalogue of the Spanish Federation of Rare Diseases (FEDER). For this purpouse the Netvizz tool has been used to download the data and through the use of several libraries in R wordclouds have been created. To perform sentiment analysis the TextBlob library from Python has been used. In addition, log-likelihood has been calculated to compare the word obtained on Facebook and the Decalogue. The results show that the photos are the type of publication that generate a bigger number of 'likes', 'shares' and 'engagement. Furthermore, we can see that positive polarities also generate a higher 'engagement', however subjectivity is not so closely linked to 'engagement'. The results of text mining show that Facebook groups are more concerned about families, people and help, but the Decalogue is more focused on disabilities, training and diseases. Regarding the influence of temporary variables we can conclude that there is statistical significance between the polarity of the message and the hour it has been posted. For all of this, it would be advisable that within FEDER´s activities some actions are taken to give more support to families.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

En Europa, una enfermedad se denomina rara (ER) cuando tiene una incidencia menor a 5 casos de cada 10.000 habitantes. Existen unas 7.000 enfermedades raras conocidas, que según cálculos de la Organización Mundial de la Salud (OMS) afectan al 7% de la población mundial. Es decir, que sólo en España existen más de 3 millones personas afectadas por una ER.

Veamos unos datos que nos ayudarán a hacernos una idea de la magnitud del problema:

"El 65% de estas patologías son graves e invalidantes y se caracterizan por:

- Comienzo precoz en la vida (2 de cada 3 aparecen antes de los dos años)
- Dolores crónicos (1 de cada 5 enfermos)
- El desarrollo de déficit motor, sensorial o intelectual en la mitad de los casos, que originan una discapacidad en la autonomía (1 de cada 3 casos)
- En casi la mitad de los casos el pronóstico vital está en juego, ya que a las enfermedades raras se le puede atribuir el 35% de las muertes antes de un año, del 10% entre 1 y 5 años y el 12% entre los 5 y 15 años." (Fuente: FEDER).

La Federación Española de Enfermedades Raras (FEDER) nació en 1999 con el objetivo de dar a conocer las ER, servir de nexo con la Administración y luchar por la normalización del colectivo.

Para marcar las líneas de trabajo cuentan con un decálogo de prioridades [1], que sirve para que todas las asociaciones locales trabajen sobre los mismos objetivos. Este decálogo ha sido elaborado en consonancia con las preocupaciones del colectivo; pero debe permanecer vivo para asegurar su efectividad.

Buscaremos con este trabajo analizar si el decálogo cubre las necesidades del colectivo de Enfermedades Raras.

1.2 Objetivos del Trabajo

Para asegurar la vigencia del mencionado decálogo se debe garantizar que éste refleja la realidad social. El auge de las redes sociales en los últimos años está generando un volumen de información que puede ser muy útil para conocer las preocupaciones del colectivo con ER.

El objetivo de este trabajo será realizar un estudio sobre la información que fluye en la red social Facebook en relación a las enfermedades raras y relacionarlo con las prioridades de FEDER.

1.3 Enfoque y método seguido

Para conseguir el objetivo marcado, en primer lugar, se realizará un estado del arte que ayudará a contextualizar los organismos/asociaciones relacionados con las ER y los estudios existentes que puedan ayudar a una mejor comprensión del objetivo.

Posteriormente se extraerá de Facebook un conjunto de posts emitidos por grupos en los que interaccionan pacientes con ER. Sobre ellos se analizará el contenido del post y se generará una nube de palabras. Además, se hará un análisis de sentimientos para clasificar la polaridad del mensaje. Con la información obtenida se buscarán correlaciones entre la polaridad del mensaje y el resto de variables que configuran un post (nº de likes, comentarios, reacciones, etc.). Para realizar este análisis se trabajará en R y/o Python, según las necesidades que vayan surgiendo. El código elaborado se encuentra a disposición en Github (https://github.com/natt77/UOC---TFP).

Finalmente se buscará relacionar la información obtenida con el decálogo de FEDER y se proporcionaran recomendaciones o comentarios que permitan tomar acciones destinadas a cubrir de forma eficiente las preocupaciones de los colectivos con ER.

1.4 Planificación del Trabajo

Para la elaboración del calendario partiremos de las fechas marcadas para las PECs, de forma que garanticemos que el contenido requerido en cada una de ellas esté realizado para las fechas solicitadas. Para esta planificación inicial tendremos en cuenta:

- **PEC1:**
  - Concretar trabajo
  - Elaboración de calendario
  - Redactar introducción: Ámbito, Objetivos, Metodología
- **PEC2:**
  - Definir estructura de Índice
  - Elaborar estado del arte
  - Toma de contacto con los datos, análisis de las variables
  - Creación de nube de palabras
- **PEC3:**
  - Revisión de la planificación
  - Establecer correlaciones entre variables
  - Realizar análisis de sentimientos
  - Relacionar información obtenida con decálogo de prioridades de FEDER
  - Conclusiones
  - Documentar referencias y Anexos

Elaboramos el calendario previsto, teniendo en cuenta las fiestas navideñas y estimando que la dedicación media para el trabajo será aproximadamente de 1h en días laborables (L-V) y de 3h en festivos.

Figura 1: Calendario

1.5 Breve sumario de productos obtenidos

Se han obtenido los siguientes productos, disponibles en Github:

- Librería para crear nubes de palabras
- Librería para realizar análisis de sentimientos (polaridad y subjetividad)
- Librería para realizar comparativas entre Facebook y el Decálogo
- Librería para calcular correlaciones
- Librería para calcular modelos que predicen la polaridad

Los productos del trabajo se han descrito en un artículo que se ha enviado a la siguiente conferencia científica: Reguera, N., Subirats, L., Armayones, M. Mining Facebook data of people with rare diseases. IEEE International Symposium on Computer-Based Medical Systems, June 2017. (submitted).

1.6 Breve descripción de los otros capítulos de la memoria

En primer lugar, expondremos el estado del Arte en materia de análisis de sentimientos en el ámbito de las ciencias de la salud, a continuación, realizaremos varias nubes de palabras que nos proporcionarán información sobre los temas tratados en Facebook y sobre el contenido del decálogo de prioridades de FEDER. Posteriormente realizaremos un análisis de sentimientos sobre los posts emitidos en Facebook y buscaremos correlaciones entre las distintas variables. Para finalizar realizaremos buscaremos conclusiones con los resultados obtenidos, con el objetivo de proporcionar a FEDER mejoras sobre su línea de trabajo.

2. Análisis de datos

2.1 Estado del Arte

La minería de textos es una disciplina que surgió en los años 80 y que se focaliza en la obtención de conocimiento a partir de textos. Por otro lado, el auge de las redes sociales en los últimos años está generando un volumen ingente de datos desestructurados o semi-estructurados de los que gran parte están en un formato de texto.

Además, en los últimos años se han producido muchos avances en minería de textos, y ya es posible obtener en tiempo real conocimiento de lo que está ocurriendo en las redes sociales.

En este contexto, han surgido en el mercado varias soluciones que permiten el análisis de lo que está ocurriendo en las redes sociales, de las que para el objeto de nuestro trabajo destacamos:

- Grytics: permite exportar grupos de Facebook, y realizar estadísticas básicas y tendencias. [2]
- Sociograph: diseñada específicamente para Facebook, realiza estadísticas, actividad a lo largo del tiempo e información detallada de las interacciones de los miembros del grupo. [3]
- Netvizz: integrada en Facebook, permite exportar la actividad de los grupos públicos, permitiendo la descarga entre fechas o por número de registros. [4]

La asociación FEDER, surgió con el objetivo de dar a conocer las enfermedades raras. Además de dicha asociación, de ámbito nacional, existen numerosas asociaciones internacionales preocupadas por las ER, destacamos:

- European Organization for rare diseases (EURODIS). [5]
- Alianza Iberoamericana de Enfermedades poco frecuentes (ALIBER). [6]
- National organization for rare diseases (NORD) – USA. [7]
- Raregenomics. [8]

Encontramos que existen varios trabajos que **buscan enmarcar las ER en las redes sociales**:

- Rareconnect (EURODIS). Plataforma en la que se crean comunidades por enfermedades o grupos de discusión de ER. [9]

- Crowdmed. Plataforma colaborativa en la que los pacientes exponen su caso (enfermedad no diagnosticada o síntomas crónicos) y una red de médicos aporta información para su resolución. [10]

- #LoweResearchProject. Proyecto que busca extraer conocimiento sobre la enfermedad de Lowe a partir de la información de las redes sociales. [11]

- Publicación que enseña cómo promover el uso de la RS para difundir las ER. [12].

- Como se usan las RS para reclutar pacientes para probar nuevos fármacos. [13].

- Lista con información sobre las personas más influyentes de las RS en temas relacionados con las ER y los medicamentos huérfanos. [14].

- Publicación que informa sobre cómo conseguir conectar con los pacientes en las redes sociales. [15].

Por otro lado, existen también numerosos estudios que **explotan la información** de las redes sociales, por ejemplo:

- Existen estudios en los que se confirma que el uso de las redes sociales no sólo está extendido por parte de las asociaciones dedicadas a las ER, sino que también su uso es clave para paliar el aislamiento social. [16].

- Estudio que ayuda a predecir brotes de asma basándose en información de redes sociales combinada con otros datos. [17].

- Encontramos también iniciativas para crear nuevas plataformas sociales en las que pueden interactuar médicos, pacientes y familiares relacionados con enfermedades neurológicas. [18] [19].

También existen estudios **que utilizan tecnologías de análisis de sentimientos y minería de textos** en el ámbito de las ciencias de la salud:

- Estudio para comprender el impacto de las redes sociales a través de análisis de sentimientos en relación al déficit de atención. [20].

- Estudio que, basándose en información recogida de Twitter en tiempo real, es capaz de analizar el estado de ánimo de personas con problemas mentales y entender su dispersión. Interesante conclusión que observa una correlación entre la expresión de emociones y las tasas de suicidio/carga de ansiedad. [21].

- Estudio que, extrayendo información de foros con temática de salud, realiza un análisis de sentimientos. [22].

- Estudio que analiza los mensajes intercambiados en una lista de correos orientada a pacientes con una enfermedad rara. Contrariamente a los que presuponían los investigadores la temática tratada se centra en búsqueda de información biomédica y no en expresiones emocionales. [23].

- Este otro estudio se centra en monitorizar la actividad de varios grupos de Facebook relacionados con anomalías congénitas y concluye que para mejorar la atención que reciben las familias y pacientes es necesario que el personal médico se involucre en la red social. [24].

- En este estudio se hace uso de análisis de sentimientos para evaluar la experiencia de los usuarios del sistema público de salud inglés. Los usuarios a través de una página web tienen la posibilidad de evaluar el sistema, tanto de forma cuantitativa como escribiendo un texto. Se concluye que el análisis de sentimientos es capaz de predecir correctamente la polaridad de los mensajes, comparándolo con la evaluación cualitativa de los usuarios. [25].

Encontramos también los siguientes estudios que nos ayudan a comprender el estado del arte en relación al trabajo que realizaremos con **corpus y nubes de palabras**:

- Existe varias medidas para comparar las diferencias entre los distintos corpus. Destacamos los siguientes trabajos:
    - Comparativas de uso entre el log-likelihood (LL) y la razón de oportunidades (Odds ratio – OR). [26].
    - Se establece el cómputo del LL como método para comparar la frecuencia de aparición de términos entre 2 corpus. [27].
    - Se compara el uso del LL frente al ranking por diferencia de frecuencias (%%DIFF). [28].
    - Uso del Bayes factor (BIC). [29].
    - El efecto del tamaño en el LL (ELL). [30].
- Este estudio proporciona un método para comparar dos nubes de palabras y da como recomendación en uso de 500 palabras para visualizar en una pantalla de ordenador estándar (1150 x 550 pixeles). [31].
- Usando los datos de twitter, se analiza si existe un mal uso de los antibióticos en caso como 'gripe' o 'resfriado'. Se hace uso de una nube de palabras (150 términos) para mostrar los resultados. [32]

2.2 Obtención de los datos

El punto de partida de nuestro análisis se centra en la obtención de los datos. En base a lo que nos indican desde FEDER el criterio de selección que se realizó fue el siguiente:

- Se realiza desde el buscador de Facebook una preselección de grupos, que contengan las palabras ‘Síndrome de’ o ‘Asociación de’
- Se preseleccionan los 10 primeros de cada búsqueda
- De entre el primer grupo de búsqueda se seleccionaron 5, que debían cumplir las siguientes condiciones:
    - Asociación Española
    - Con presencia en FEDER (es decir, que representen una o varias enfermedades raras)
    - Más de 300 participantes
    - Que haya como mínimo 10 personas distintas que participen
    - Grupos en castellano

    - Que afecte a niños
- Entre los grupos preseleccionados con el término 'Asociación de' no se selecciona ninguno, por no cumplirse las condiciones mencionadas anteriormente.

Tenemos por tanto 5 grupos seleccionados:

ID Facebook	Nombre del Grupo	Síndrome/Enfermedad
106839068915	Asociación Española para la Investigación y Apoyo al Síndrome de Worfam	Síndrome de Worfam
55698954650	Síndrome de Moebius España	Síndrome de Moebius
338568431089	Asociación Española del Síndrome del Vómito Cíclico	SVC
333386441351	Asociación Española de Síndrome de Marfan	Síndrome de Marfan
82783734050	Síndrome 5P o CRIT du Chat	Síndrome 5P

Tabla 1: Grupos de Facebook

A partir de aquí y usando la herramienta de Facebook Netvizz se procede a la descarga de datos, obteniéndose un total de 3917 registros desglosados de la siguiente manera. Se indica también fecha desde la que se recogen datos:

ID Facebook	Volumen	Fecha Desde
106839068915	456	05/2009
55698954650	894	03/2009
338568431089	319	03/2010
333386441351	1547	02/2010
82783734050	701	05/2009

Tabla 2: Volumen datos Facebook

El total de los 3917 registros se encuentran recogidos en el archivo "ArchivoAsociaciones20161128ANSII.txt".

Por otro lado, se accede a la página de FEDER y se elabora un fichero "decalogodeprioridades.txt" a partir del decálogo de prioridades detallado en la página.

2.3 Comparativa entre datos

Una vez obtenidos los datos de trabajo procedemos a realizar un análisis en R. Vamos a realizar una comparativa entre los datos de Facebook y el decálogo de prioridades.

En primer lugar, extraemos de los datos de Facebook la variable post_message que contiene el texto de las publicaciones. Creamos un Corpus y lo simplificamos quitando la puntuación, números, minúsculas. Además, eliminamos una serie de palabras que consideramos no son útiles para nuestro análisis (adverbios, palabras específicas de las enfermedades) y lo reducimos hasta tener un tamaño manejable.

Dibujamos un histograma que nos represente la frecuencia de aparición de cada palabra:

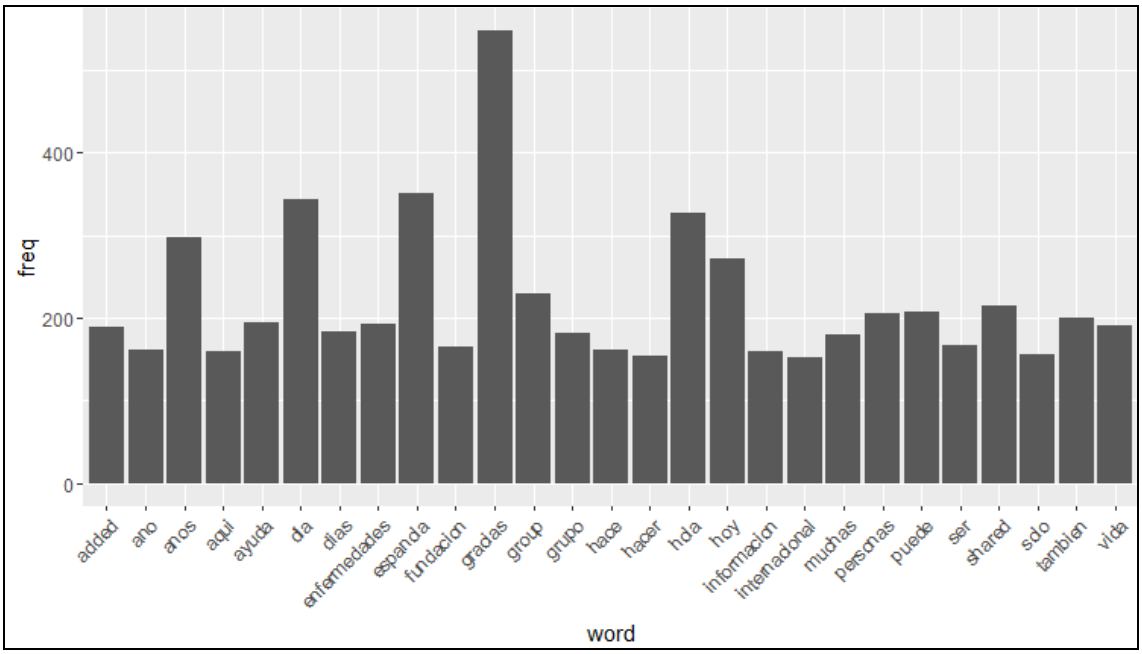

Figura 2: Frecuencias de palabras - Facebook

Y una nube de palabras que visualmente lo representa de forma más clara:

Figura 3: Nube de palabras - Facebook

De forma análoga hacemos lo mismo con el decálogo de prioridades, creando un corpus y mostrando un histograma y una nube de palabras:

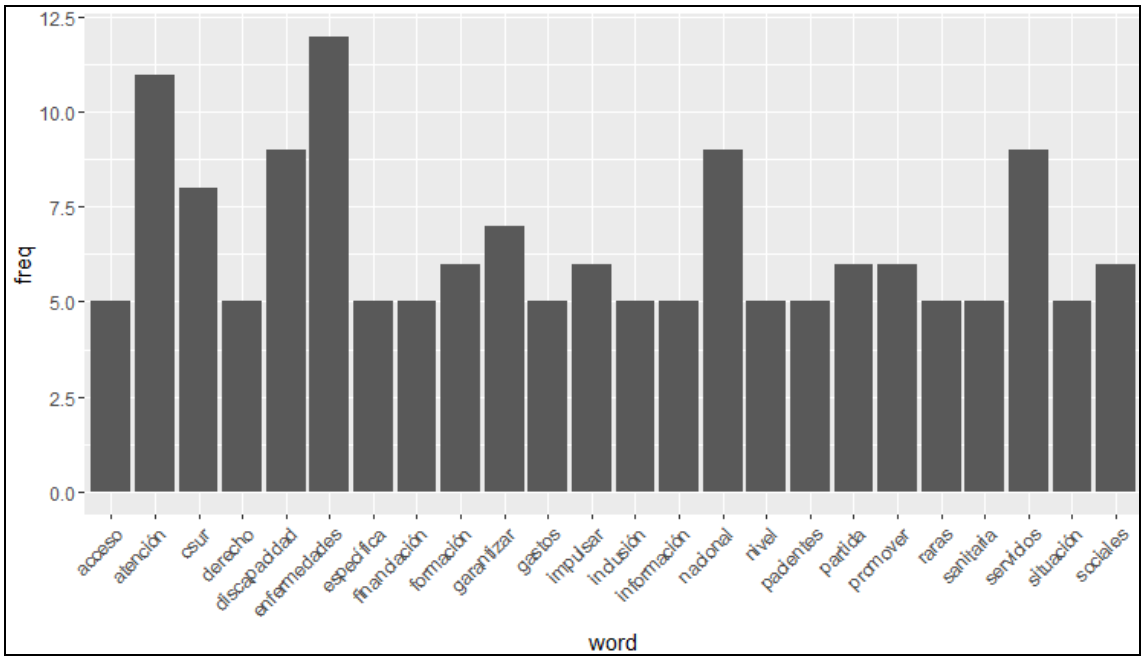

Figura 4: Frecuencias de palabras - Decálogo

Figura 5: Nube de palabras - Decálogo

Para poder realizar una comparativa entre ambos utilizamos el log-likelihood mencionado en el estado del arte, que nos compara las frecuencias relativas entre dos corpus. Con esta información podemos ver cuáles son las palabras que tienen un LL más alto, que nos indicará que su frecuencia es muy dispar entre los dos corpus.

El cálculo del LL se basa en los siguientes datos:

	Facebook	Decálogo	Total
Frecuencia de la palabra	a	b	a+b
Frecuencia del resto de palabras	c-a	d-b	c+d-a-b
Total	c	d	c+d

Tabla 3: Tabla de contingencia para el cálculo del LL

Y se calcula en base a las siguientes fórmulas:

E1 = c*(a+b)/(c+d)
E2 = d*(a+b)/(c+d)

LL = 2 * (a*log(a/E1) + b*log(b/E2))

Los resultados obtenidos son los siguientes:

Palabra	Frec - Facebook	Frec - Decálogo	LL
nacional	32	9	20,178392
discapacidad	53	9	13,399595
nivel	28	5	7,80963
ayuda	195	1	7,514368
profesionales	23	4	6,093768
referencia	28	4	4,972844
Vida	190	2	4,283898
Frecuentes	35	4	3,769416
Enfermedades	192	12	3,331146
ser	167	2	3,231924
personas	206	3	2,985333
hijo	112	1	2,971193
dos	107	1	2,721853
hacer	153	2	2,627232
persona	46	4	2,43415

Tabla 4: Log-likelihood - Valores altos

Palabra	Frec - Facebook	Frec - Decálogo	LL
causa	48	2	6,25E-02
social	36	1	5,05E-02
difundir	23	1	4,65E-02
medio	24	1	3,12E-02
experiencias	34	1	2,71E-02
forma	52	2	2,07E-02
general	32	1	1,05E-02
cuanto	26	1	1,04E-02
todas	91	3	6,99E-03
dice	29	1	1,35E-05

Tabla 5: Log-likelihood - Valores bajos

Para poder interpretar correctamente los resultados con un LL alto es necesario saber en cuál de los corpus ha aparecido con más frecuencia la palabra. Para ellos hacemos uso de una herramienta online disponible en [33]. Mediante dicha herramienta podemos observar por ejemplo que 'discapacidad' es más frecuente en el Decálogo, y sin embargo 'ayuda' es más frecuente en los grupos de Facebook.

Para finalizar realizamos una nube de palabras doble, que contiene tanto los términos de Facebook como los del Decálogo. Esta nube muestra cuales son las palabras con mayor y menor coincidencia:

Figura 6: Nube de palabras: Decálogo vs Facebook

2.4 Estadísticas

En esta segunda parte hemos calculado unas estadísticas básicas para las variables numéricas que representan los distintos contadores: mínimo, máximo, media, mediana y desviación típica. Calculamos también el nº de ocurrencias de cada factor en el caso de la variable categórica que representa el tipo de contenido. Además, calculamos las medias de las variables numéricas en función de la variable categórica.

Campo	Min	1er cuartil	Mediana	Media	3er cuartil	Max	Desviación típica
likes_count_fb	0	1	4	6,38	8	86	8,58
comments_count_fb	0	0	0	2,09	2	80	4,87
reactions_count_fb	0	1	4	6,46	8	86	8,73
shares_count_fb	0	0	0	0,30	0	86	2,71
engagement_fb	0	2	5	8,86	11	119	11,85
polarity	-1	0	0	0,13	0	1	0,24

Tabla 6: Estadísticas por variable

Campo	event	link	note	photo	status	video
likes_count_fb	3,8 (4,3)	5,3 (6,1)	2,3 (3,2)	12 (12,8)	4,6 (6,5)	6,1 (6,4)
comments_count_fb	0,3 (0,6)	0,8 (1,9)	0,4 (1,0)	2,6 (5,3)	2,9 (5,9)	0,8 (1,8)
reactions_count_fb	3,9 (4,3)	5,3 (6,3)	2,3 (3,2)	12,2 (13,1)	4,7 (6,6)	6,1 (6,5)
shares_count_fb	4,1 (9,7)	0 (0,4)	0 (0)	0,9 (4,9)	0,1 (1,4)	0,4 (3,0)
engagement_fb	8,3 (9,5)	6,2 (6,9)	2,8 (3,6)	15,6 (17,4)	7,7 (10,3)	7,3 (9,4)
Frecuencia	34	1063	9	792	1787	232

Tabla 7: Estadísticas por tipo de contenido

A continuación, realizamos otro tipo de análisis; extraemos de nuestros datos el tipo de contenido publicado en Facebook y los contadores que nos indican el nº de likes, comentarios, engagement, etc. que tiene cada una de ellas. Realizamos un análisis factorizado por el tipo de contenido, para ver si existen diferencias significativas. Encontramos que el tipo de contenido 'photo' es el que más reacciones provoca, seguido de cerca por 'Status', que contiene textos. Visualizamos mediante gráficos los resultados.

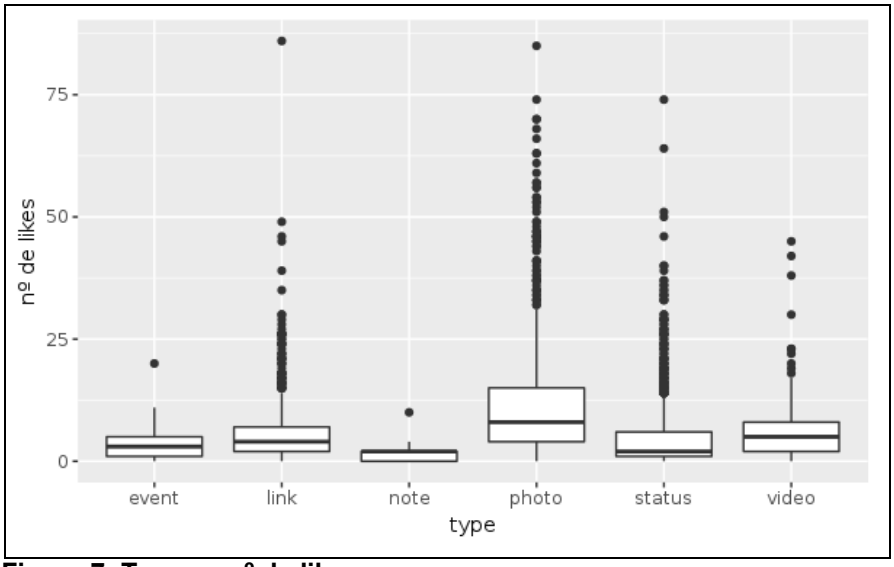

Figura 7: Type vs nº de likes

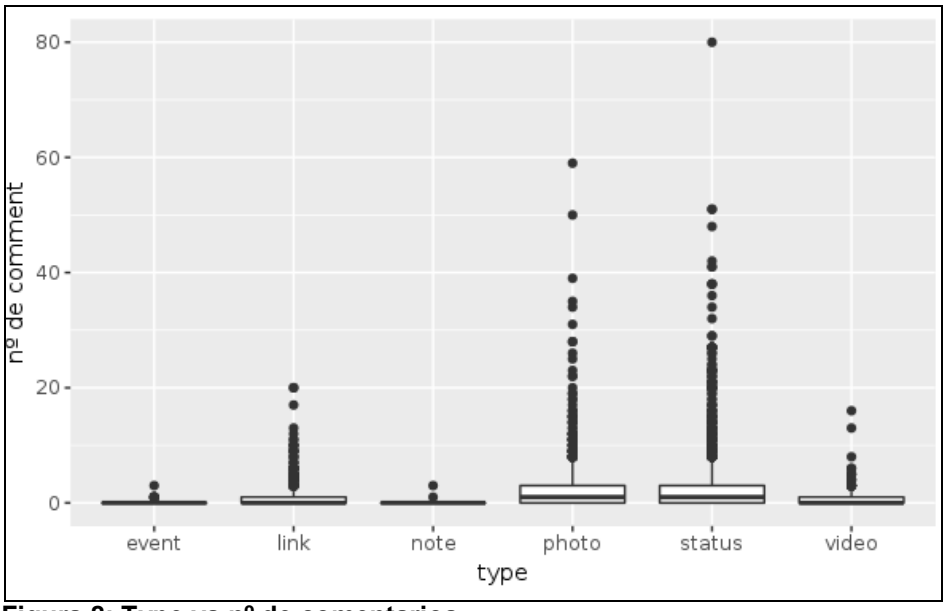

Figura 8: Type vs nº de comentarios

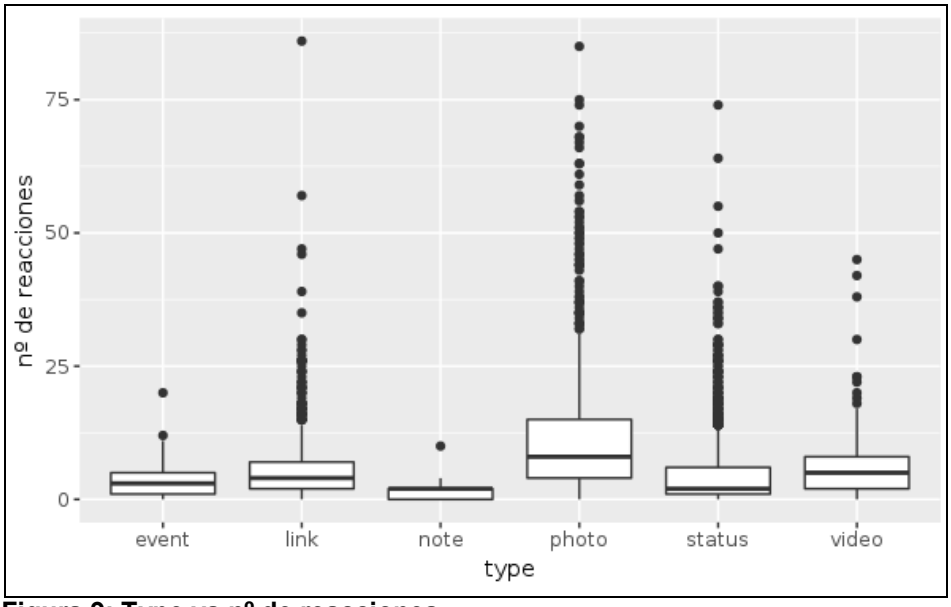

Figura 9: Type vs nº de reacciones

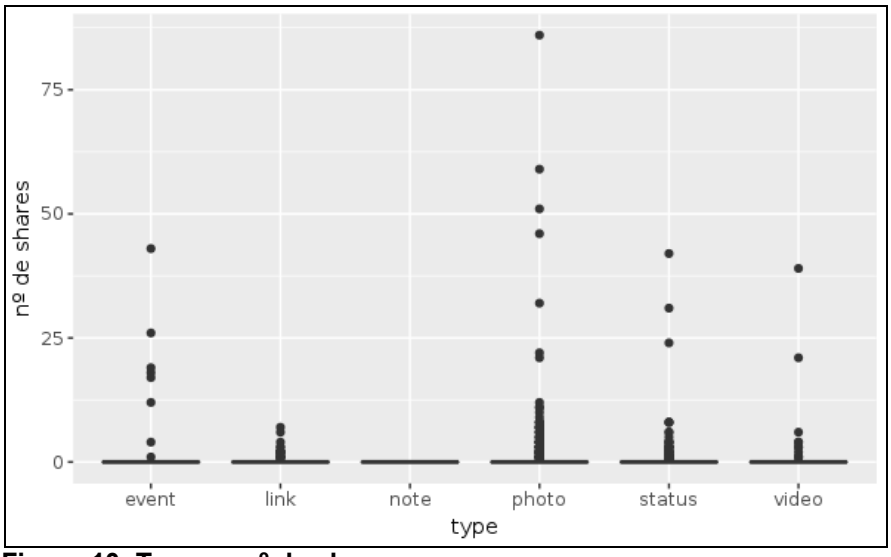

Figura 10: Type vs nº de shares

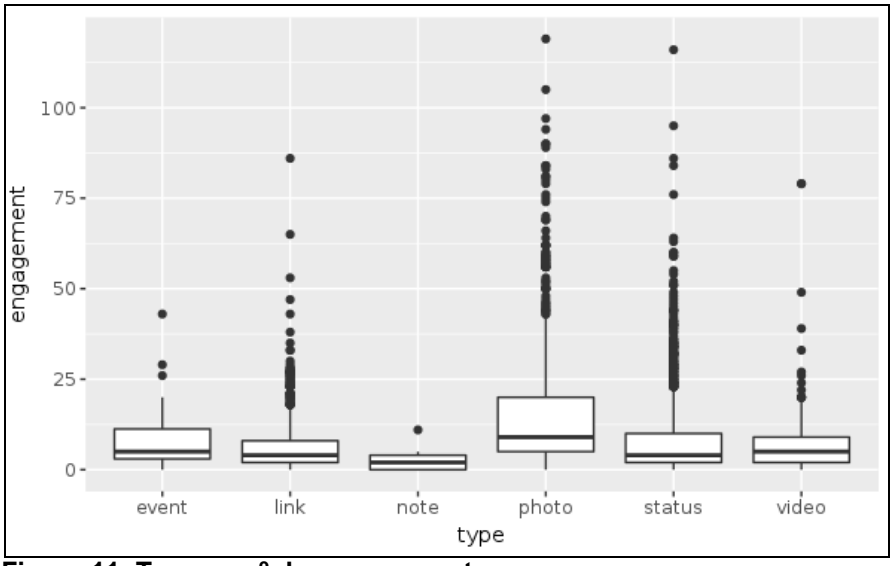

Figura 11: Type vs nº de engagement

Para el tipo de contenido 'Status' realizamos un análisis más a fondo. Para ello dividimos nuestros datos en 3 grupos según el valor del engagement de cada publicación. Nuestro objetivo es ver si somos capaces de establecer cuáles son las palabras más usadas dentro del grupo con un engagement alto y si estas se diferencian de las que tienen un engagement bajo. Realizamos un histograma que muestre la frecuencia de las palabras diferenciando por nuestros 3 grupos.

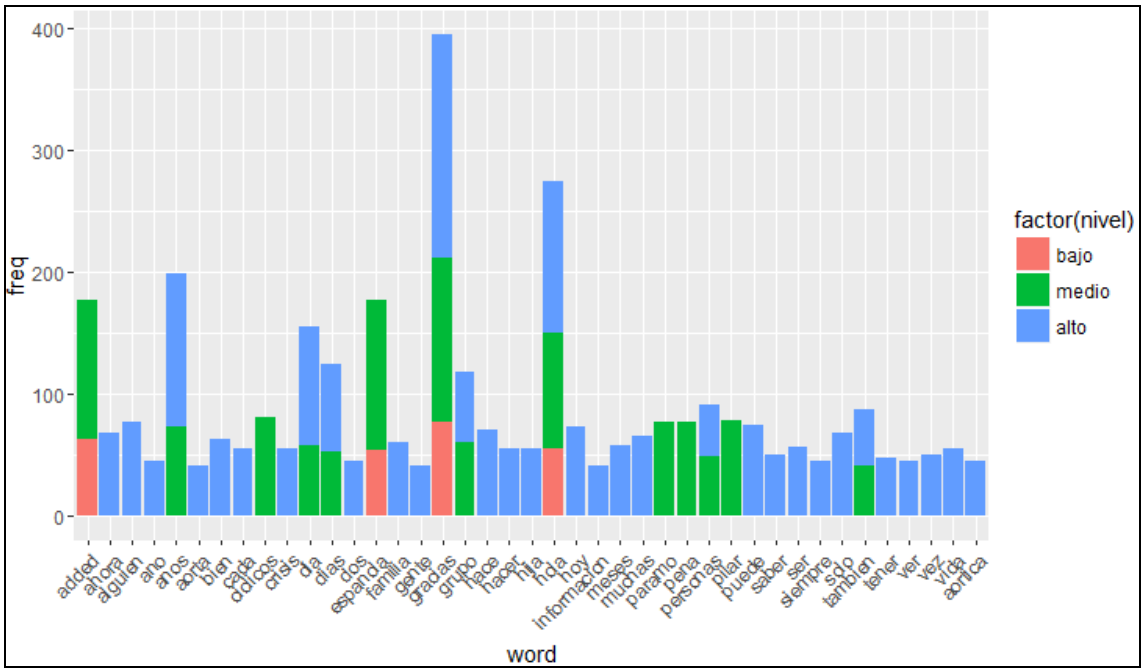

Figura 12: Frecuencia de palabras por nivel

Para completar el análisis mostramos 3 nubes de palabras, una para cada tipo de engagement: alto, medio y bajo.

Figura 13: Nube de palabras: nivel alto

Figura 14: Nube de palabras: nivel medio

Figura 15: Nube de palabras: nivel bajo

2.5 Análisis de Sentimientos y Correlaciones

Para realizar el análisis de sentimientos hemos partido de nuestro fichero con los datos obtenidos de Facebook y hemos utilizado Python para calcular la polaridad y la subjetividad.

En primer lugar, dado que la mayoría de los posts son en castellano hemos hecho uso de 'Google translator' para convertir nuestros datos al idioma inglés. Después hemos usado la librería TextBlob para obtener la polaridad y la subjetividad de cada uno de nuestros posts. El rango de cada una de estas variables es: polaridad [-1,1] y subjetividad [0,1]. Es importante observar que en los casos en los que no existe texto tanto la polaridad como la subjetividad tomarán el valor cero.

A continuación, hemos calculado la correlación entre la polaridad y las distintas variables numéricas disponibles: nº de likes, nº de comentarios, nº de reacciones, nº de comparticiones y nivel de engagement. Para no desvirtuar el análisis hemos eliminado de nuestros datos todos los registros que no eran un texto, ya que en esos casos tanto la polaridad como la subjetividad serán cero. Se han obtenido valores cercanos a cero para todos ellos.

	Likes	Comentarios	Reacciones	Shares	Engagement
Polarity	**0,063**	-0,015	**0,063**	**-0,046**	0,029
Subjectivity	0,038	**0,103**	0,039	-0,009	**0,070**

Tabla 8: Correlación entre polaridad/subjetividad y contadores

Hemos comprobado los p-valores obtenidos para los distintos modelos que nos predicen la polaridad en función de cada una de las variables obteniendo que son estadísticamente significativos los relacionados con el nº de likes, el nº de reacciones y el nº de comparticiones.

Análogamente hemos realizado el mismo análisis para la subjetividad encontrando que son estadísticamente significativos los relacionados con el nº de comentarios y el engagement.

A continuación, hemos realizado unos gráficos para mostrar la relación entre la polaridad y cada una de las variables.

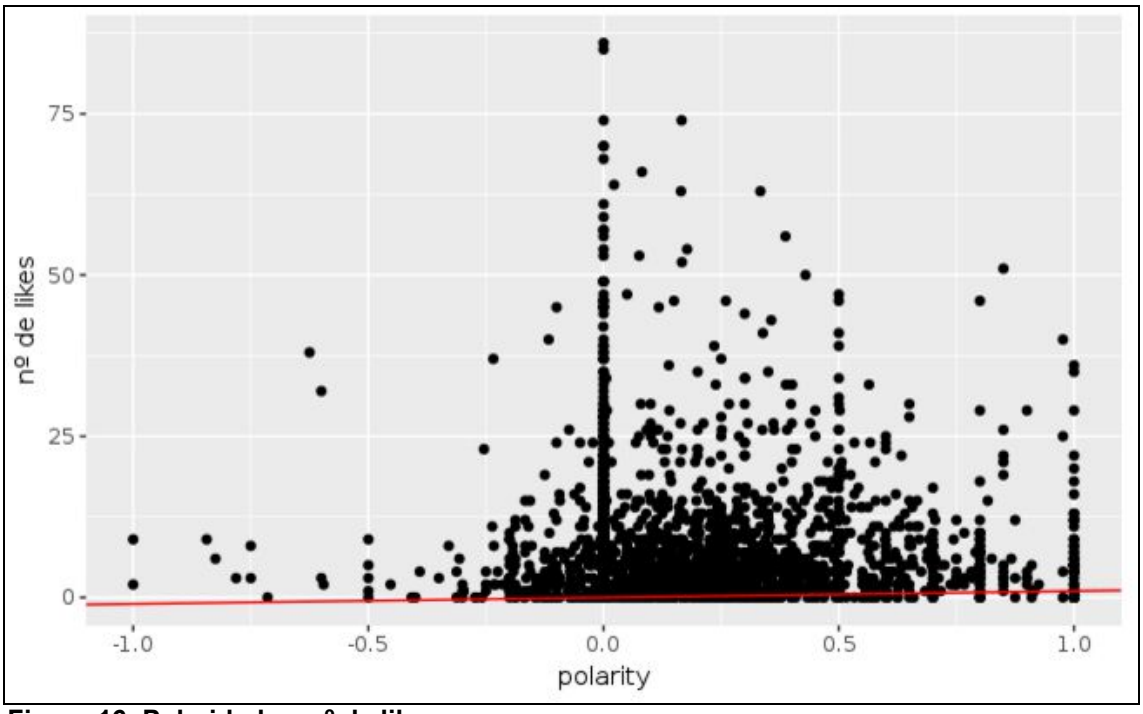

Figura 16: Polaridad vs nº de likes

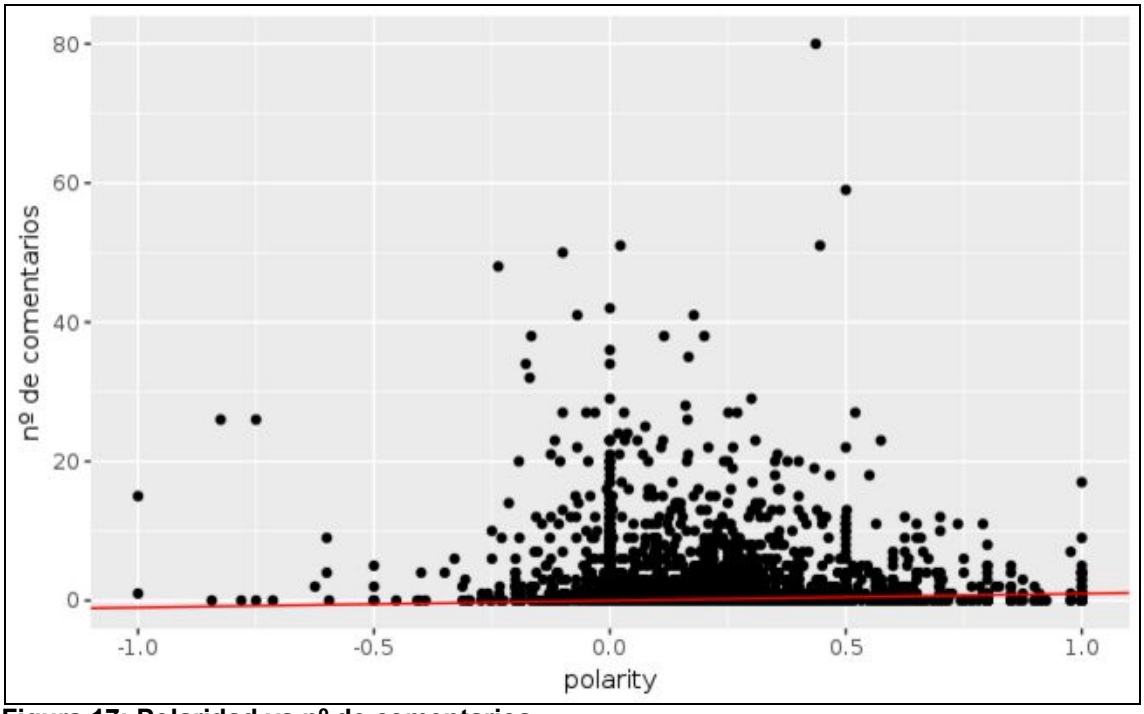

Figura 17: Polaridad vs nº de comentarios

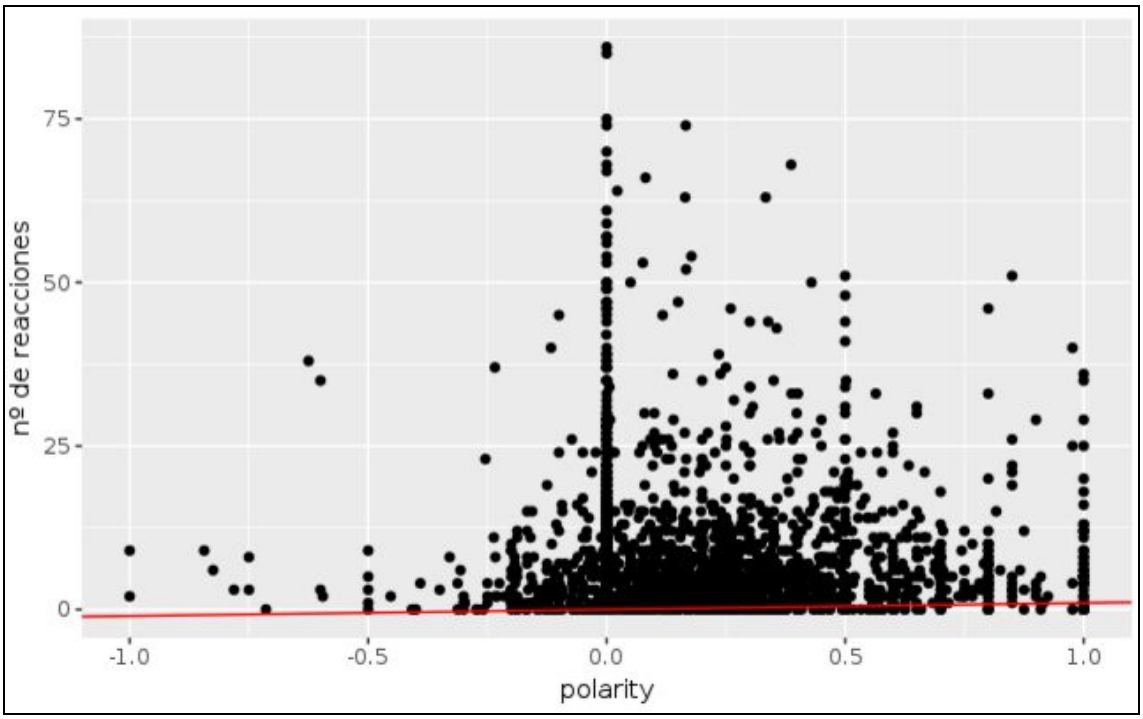

Figura 18: Polaridad vs nº de reacciones

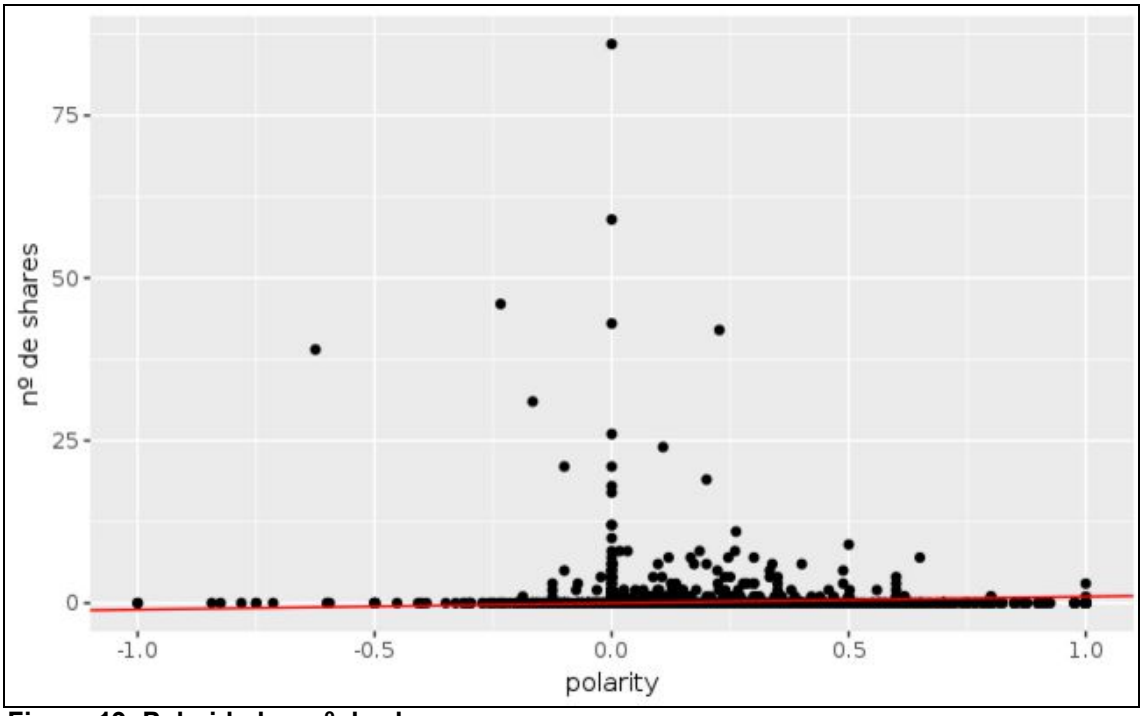

Figura 19: Polaridad vs nº de shares

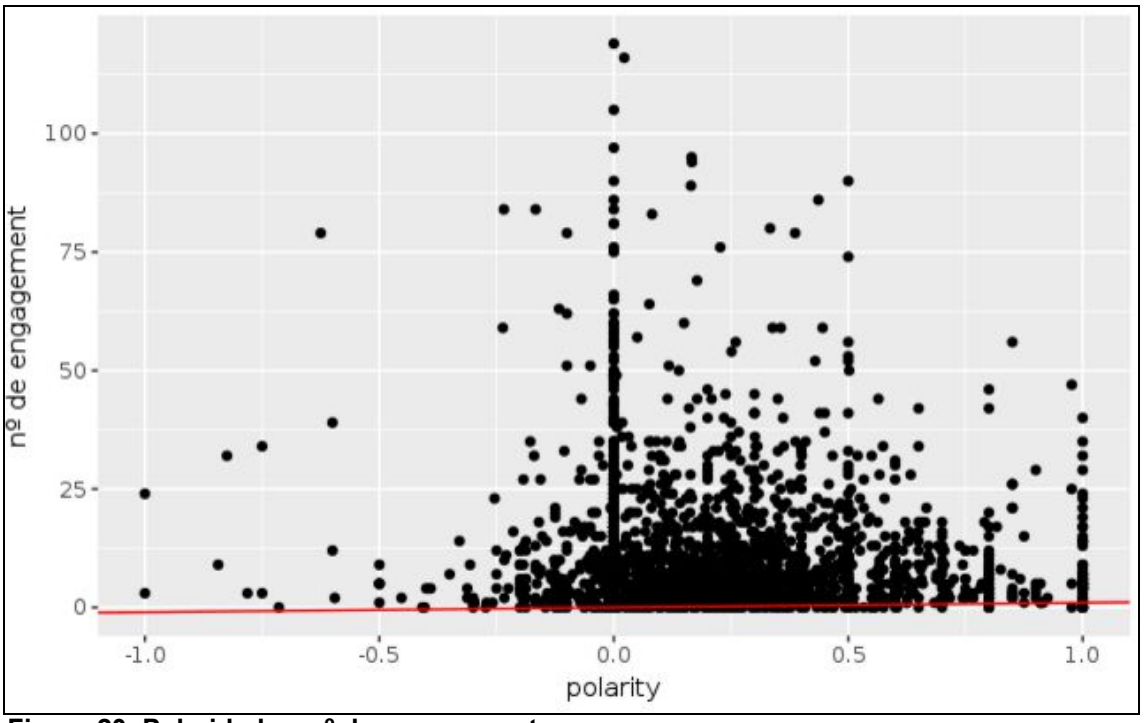

Figura 20: Polaridad vs nº de engagement

En el apartado anterior habíamos realizado un estudio de los registros con type=”status”, clasificándolos en 3 grupos en función del engagement. Vamos a completar dicho análisis incluyendo la polaridad y la subjetividad. Para ello queremos ver si es posible crear un modelo que prediga dichas variables en función del nivel de engagement. Para el caso de la polaridad vemos que el modelo lineal obtiene un p-valor menor a 0.05, por lo que concluimos que son estadísticamente significativas. No ocurre lo mismo en el caso de la subjetividad, ya que obtenemos un p-valor de 0.09. En ambos casos representamos un boxplot que nos muestra visualmente los resultados obtenidos.

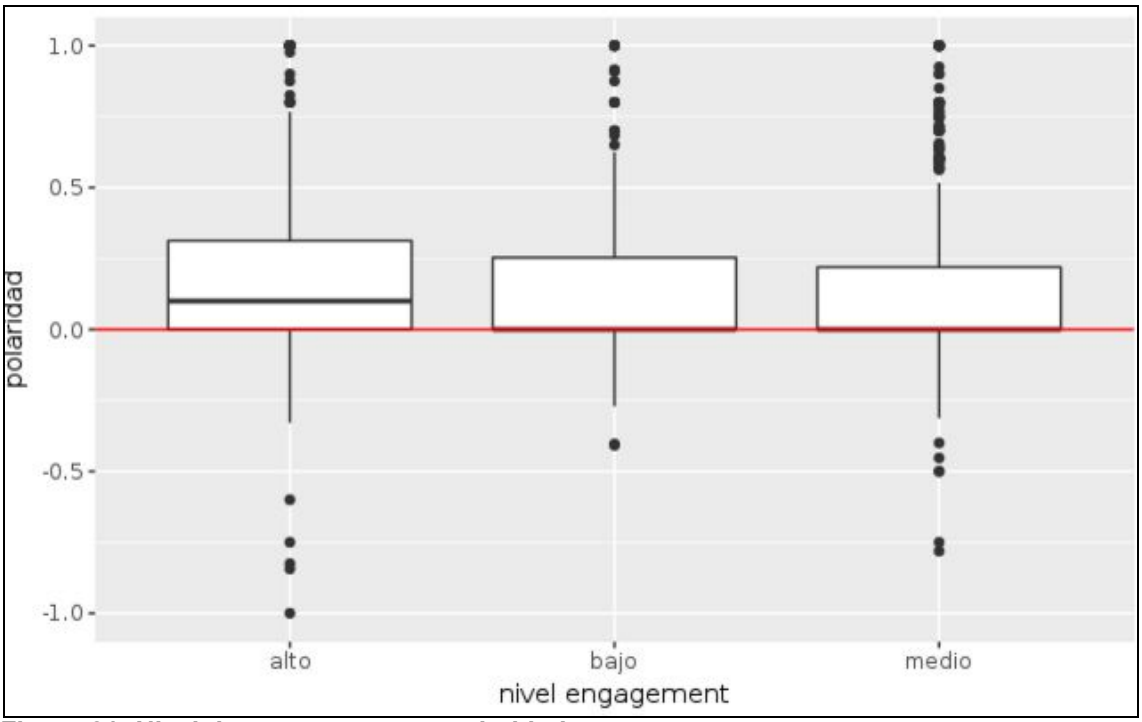

Figura 21: Nivel de engagement vs polaridad

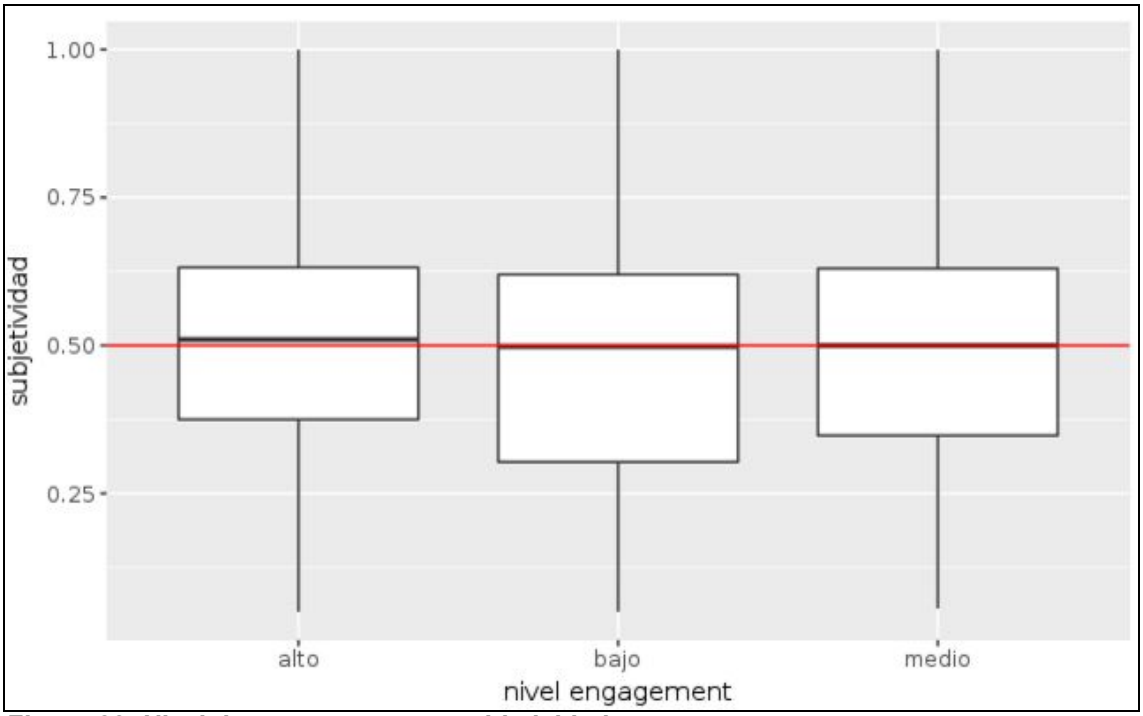

Figura 22: Nivel de engagement vs subjetividad

2.6 Análisis variables temporales

Para finalizar nuestro análisis vamos a hacer un estudio de las variables temporales. Para ello extraeremos el timestamp en el que se emitió el post y lo dividiremos para obtener el día de la semana y la hora. Con estos datos buscaremos establecer si existe alguna relación entre ellos y la polaridad o el engagement.

En primer lugar realizamos unas gráficas que nos muestran la relación entre el engagement y los días/horas de la semana:

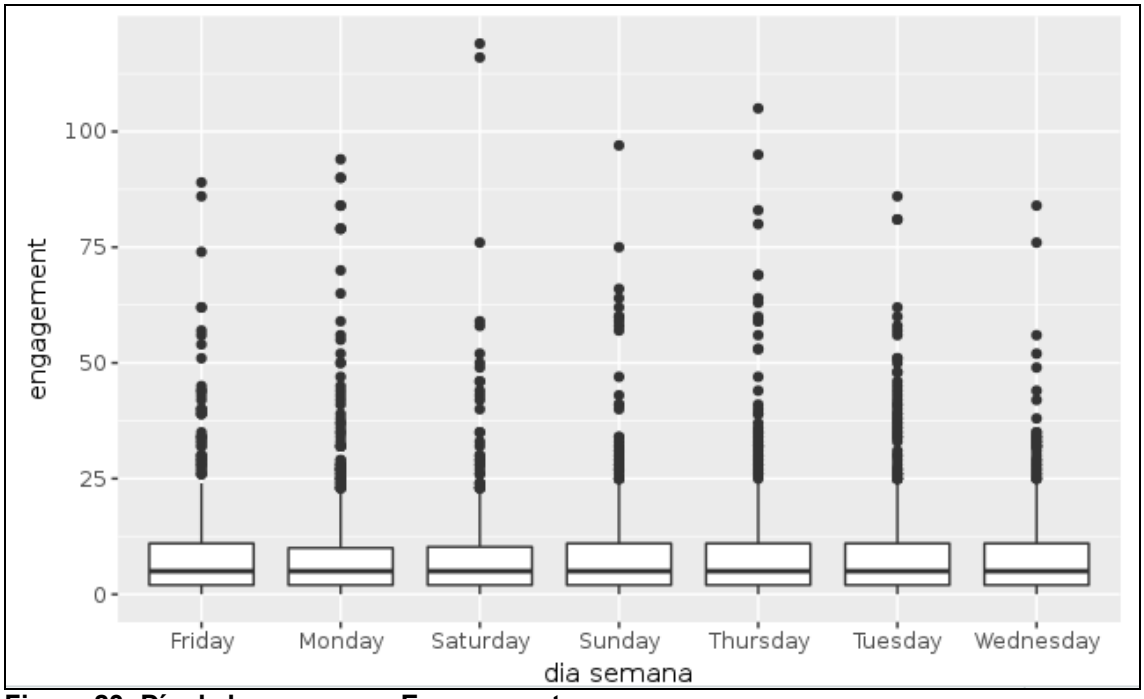

Figura 23: Día de la semana vs Engagement

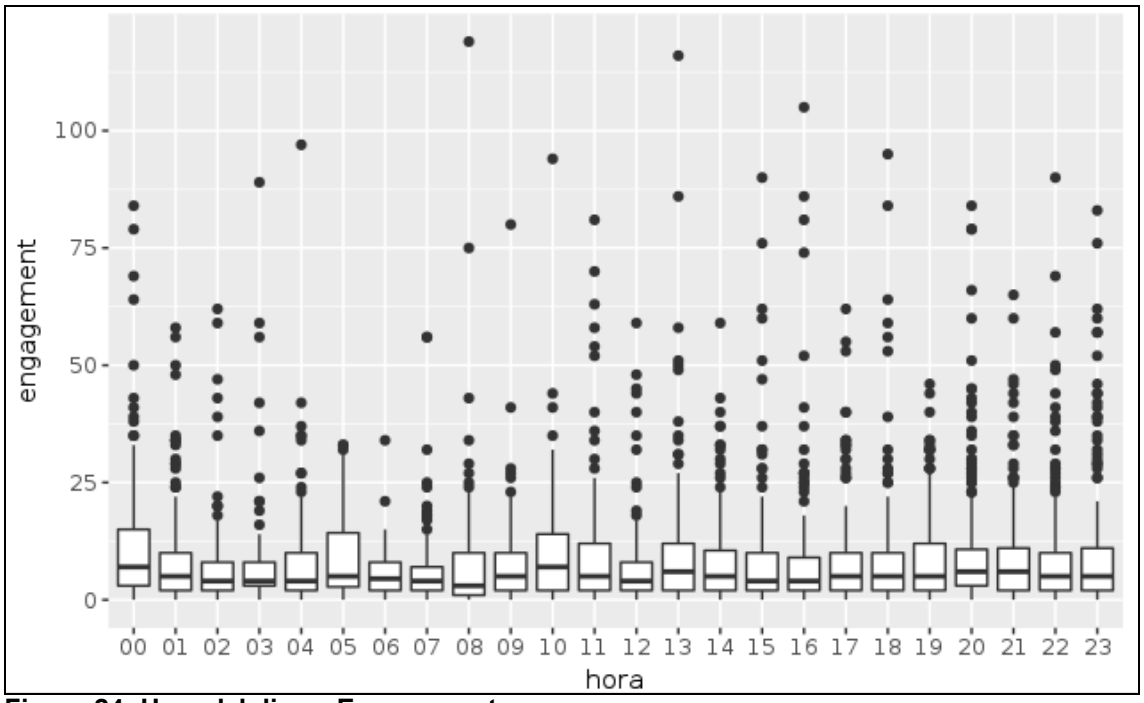

Figura 24: Hora del dia vs Engagement

A través de estos gráficos no obtenemos demasiada información, así que antes de continuar analizando el engagement y la polaridad vamos a dibujar otros gráficos que nos muestren la frecuencia de aparición de los posts por día/hora de la semana:

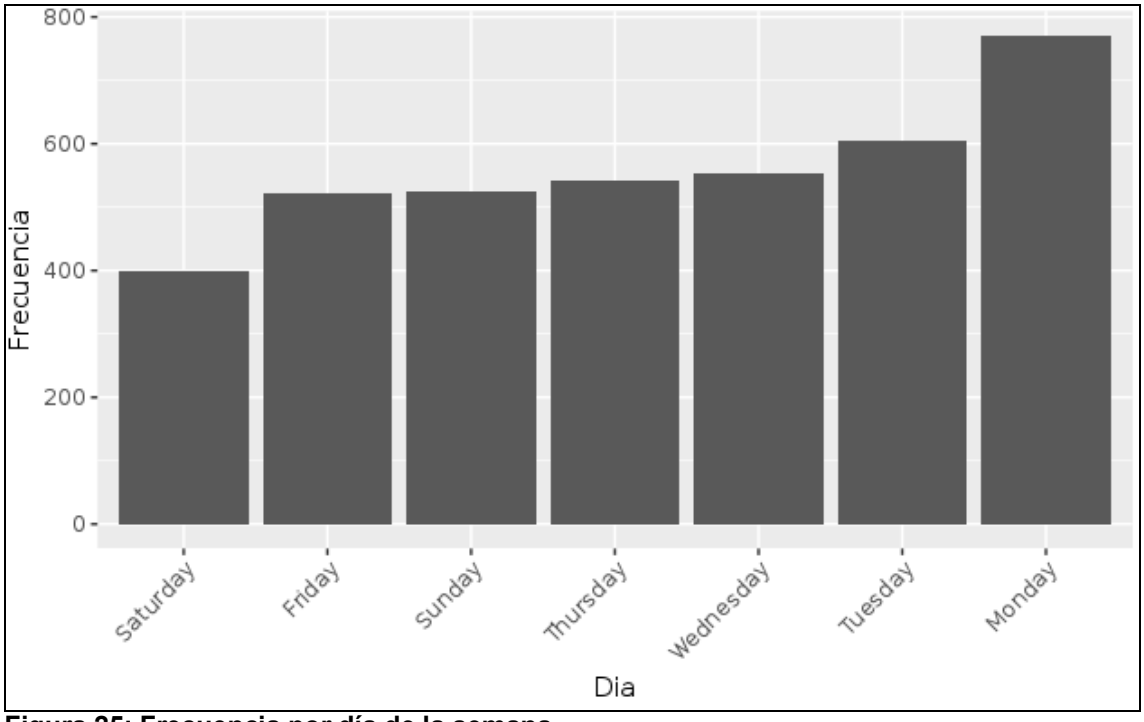

Figura 25: Frecuencia por día de la semana

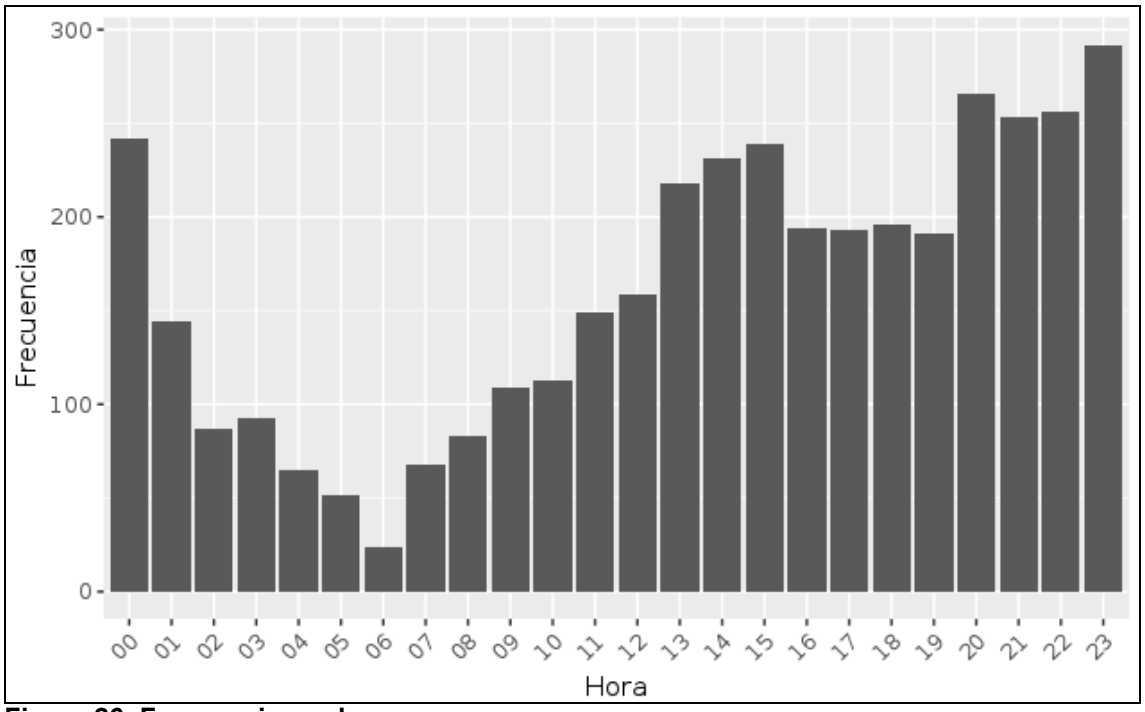

Figura 26: Frecuencia por hora

Observamos información interesante; respecto al día de la semana claramente el lunes es el día con más posts, y a medida que avanza la semana estos van decreciendo. En cuanto a la hora de emisión de los mensajes existen dos franjas en las que se emiten más, entre las 13hs-15hs y las 20hs-00hs. Estas horas coinciden con las de mayor tiempo libre y justifican una mayor actividad. También destaca como a medida que avanza la noche el nº de posts va decreciendo, alcanzando su mínimo a las 6hs.

Realizamos una visualización más para ver si el tipo de post también varía a lo largo de la semana:

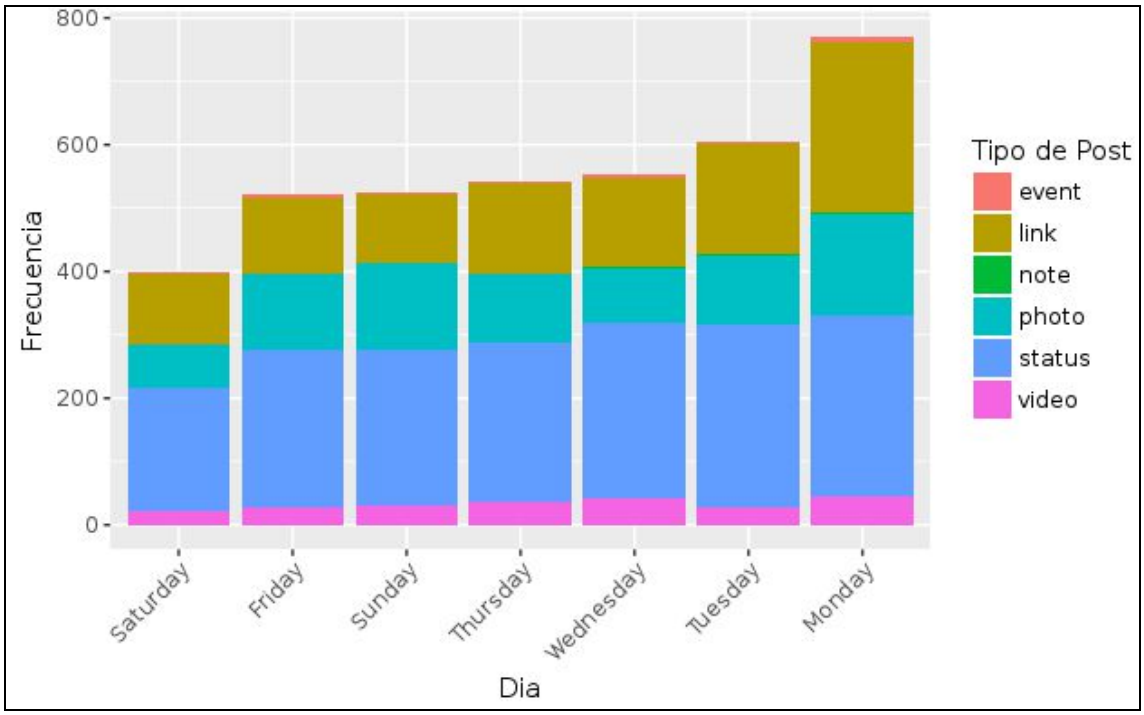

Figura 27: Frecuencia por día y tipo de post

Observamos que la mayor parte de la diferencia entre el lunes y el resto de días de la semana se debe a los post de tipo 'link'.

Vamos a buscar si existe alguna relación entre la polaridad/engagement y el día/hora en el que se emiten los mensajes. Trabajaremos con las horas agrupadas en 4 franjas horarias: Mañana (6-12), Tarde (12-18), Noche (18-24) y Madrugada (24-6).

Realizamos un análisis de la varianza (ANOVA) obteniendo los siguientes valores del p-valor:

	Día	**Hora**
Polaridad	0.158	**0.002**
Engagement	0.418	0.828

Tabla 9: P-valor con ANOVA

Observamos que el p-valor nos indica que sólo la relación polaridad-hora es estadísticamente significativa.

Veamos visualmente la relación entre la polaridad/engagement y los días/horas. Además en el caso de las horas lo mostramos también agrupado por franjas horarias.

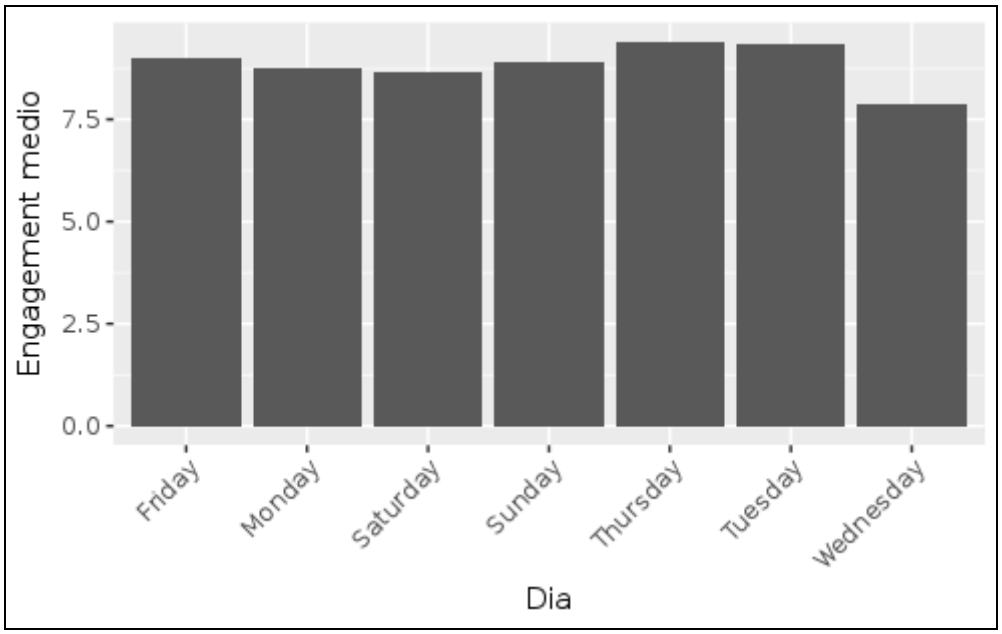

Figura 28: Engagement medio por día

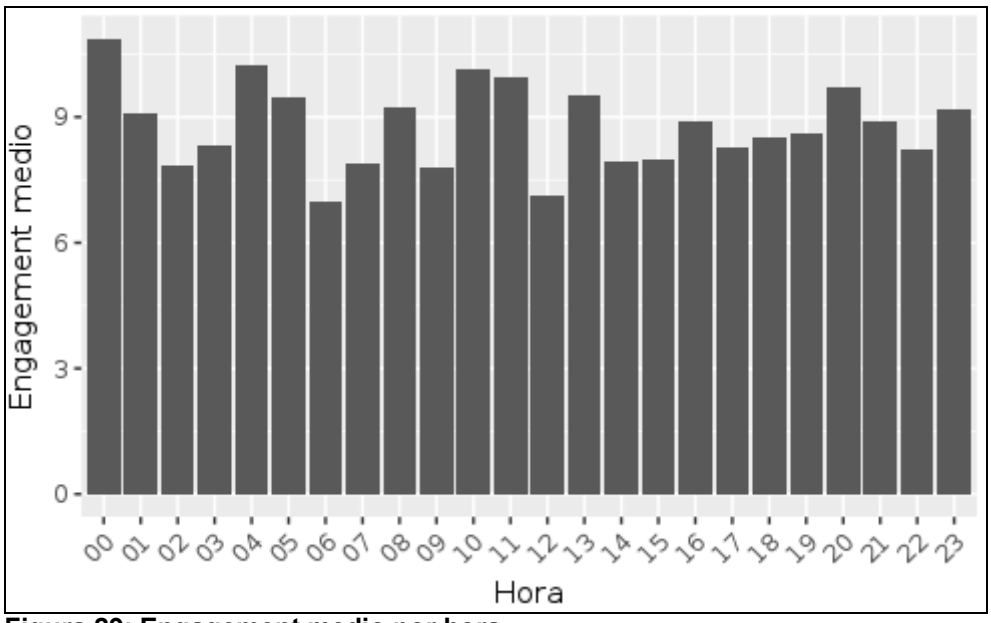

Figura 29: Engagement medio por hora

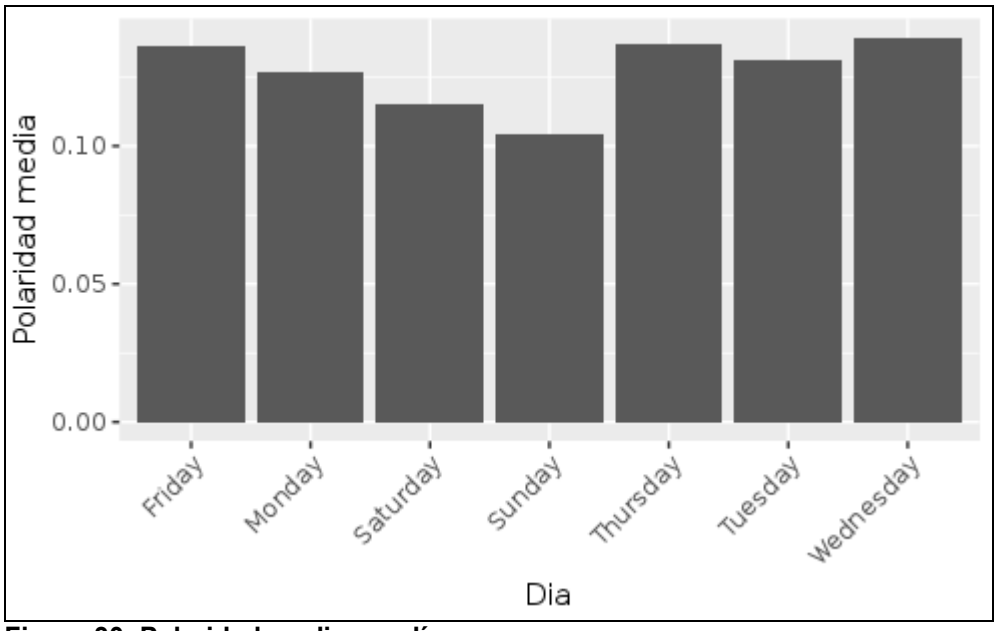

Figura 30: Polaridad media por día

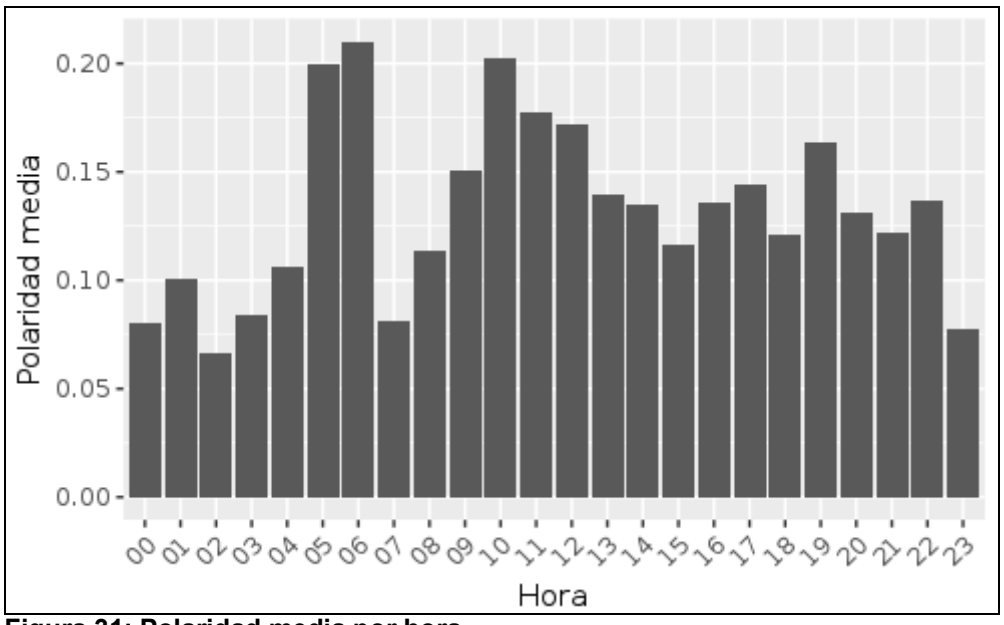

Figura 31: Polaridad media por hora

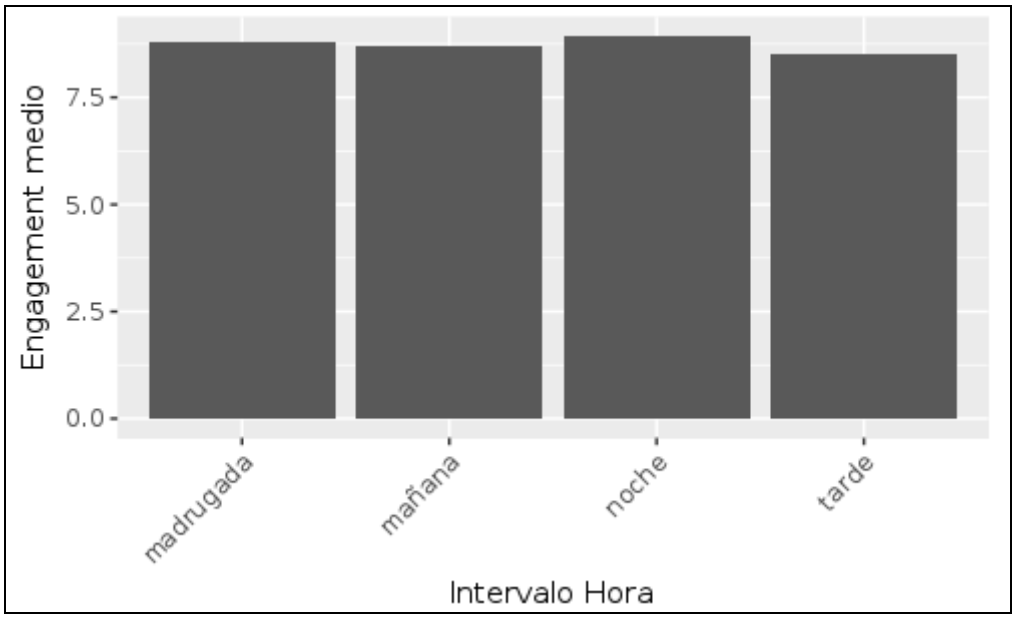

Figura 32: Engagement medio por intervalo de hora

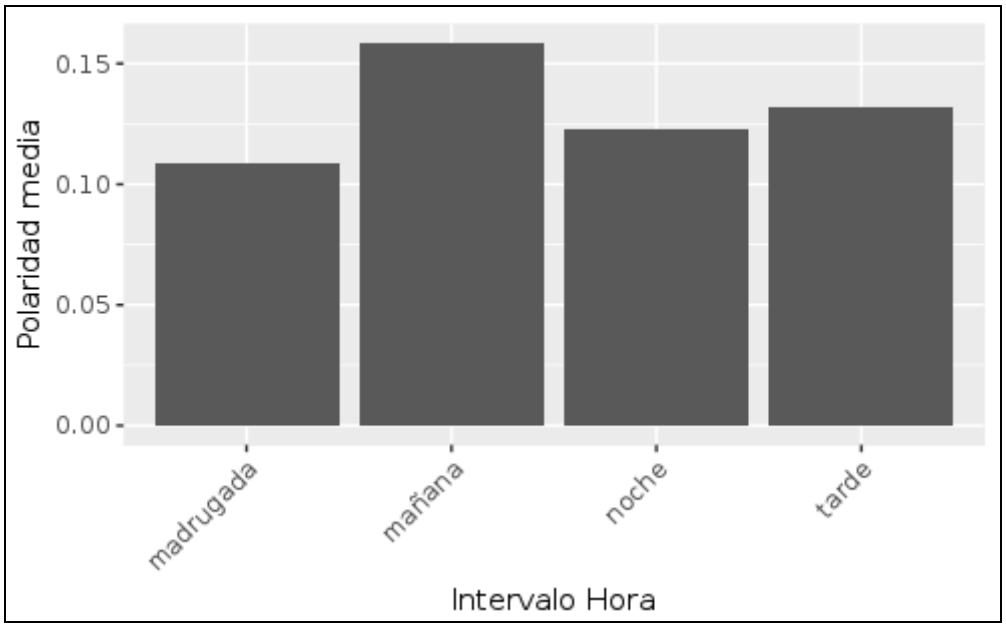

Figura 33: Polaridad media por intervalo de hora

Mediante estas visualizaciones podemos ver claramente cómo se relacionan las variables temporales con la polaridad y el engagement.

3. Conclusiones

En base a los resultados obtenidos observamos lo siguiente:

Respecto a las nubes de palabras y las palabras más frecuentes en cada una de ellos podemos ver la siguiente tendencia:

- Los temas de conversación de Facebook dan importancia a las personas, la vida, la familia, la ayuda y dan muestras de agradecimiento.
- El Decálogo está más centrado en las enfermedades, la atención, las discapacidades y los servicios.

Si observamos la nube de palabras conjunta observamos que dentro de la de Facebook existen términos muy alejados con el Decálogo como tratamiento, amigos, hijos, y que por otro lado en el Decálogo se incide en la formación y esto apenas tiene presencia en Facebook.

Respecto a los valores obtenidos mediante el LL observamos que las siguientes palabras tiene mucha presencia en el Decálogo y poca en Facebook: discapacidad, profesionales, enfermedades. Análogamente estas son las más presentes en Facebook sin apenas representación en el decálogo: ayuda, vida, personas, hijos.

En análisis por tipo de contenido nos indica que la mejor forma de conseguir compromiso (engagement) en la red social es a través de fotos. Los mensajes de texto también son importantes, pero su contenido también influirá en el nivel de compromiso que se consigue. Destaca en este sentido el uso de palabras como familia e hijos frente a las relacionadas con información o grupos.

Mediante nuestro análisis también hemos podido comprobar que la polaridad del mensaje también influencia el nivel de compromiso, mayores polaridades generan más compromiso. Sin embargo, hemos observado que la subjetividad del mensaje no afecta al nivel de compromiso.

El análisis de las variables temporales nos ha mostrado que existe significación estadística entre la polaridad y la hora del día, que no se observa ni con el engagement ni con el día de la semana.

Como lecciones aprendidas del trabajo podemos destacar:

- La línea de FEDER debería centrarse más en las personas, sus familias, en prestarles ayuda y apoyo. Actualmente las acciones que se dedican a las discapacidades, la formación y las enfermedades tienen mucha menor presencia en Facebook.
- El uso de nubes de palabras es una herramienta muy útil para visualizar los temas tratados en los textos y en combinación con el uso de Log-Likelihood(LL), que nos compara la frecuencia relativa de aparición de las palabras en 2 corpus, nos proporciona información clara sobre la frecuencia de aparición de las palabras en ambos corpus.
- La forma más eficaz de generar engagement es a través de fotos.
- La relación entre la polaridad de los posts y la hora del día que son emitidos nos muestra significación estadística.

Respecto a los objetivos planteados inicialmente hemos sido capaces de relacionar la temática tratada en Facebook frente al Decálogo de FEDER, proporcionando recomendaciones para que el Decálogo pueda ajustarse más a las necesidades sociales mostradas en Facebook. Hubiese sido interesante poder extraer más conclusiones de las temáticas tratadas, quizá buscando asociaciones entre las palabras para darles más significado o intentar conseguir un Decálogo más detallado que permitiese un análisis más profundo de las prioridades de FEDER.

Respecto a la planificación y metodología aplicada; el contenido establecido en cada una de las entregas ha sido respetado, y en algún caso se ha ampliado en la siguiente entrega. Debido a una ampliación en el plazo final de entrega del trabajo se ha dispuesto de más tiempo del planificado originalmente lo que ha permitido incluir un análisis de las variables temporales.

En la creación del Corpus se han encontrado problemas para aplicar las funciones relativas a StemCompletion que nos permitiría establecer una única palabra raíz en lugar de todos sus derivados. Por una lado el tiempo de memoria requerido era muy grande y cuando finalmente se ejecutaba se observaba que las palabras no se habían completado correctamente. Por ejemplo, tras la ejecución se observaba que había un número muy elevado de apariciones de la palabra ‘sábado’, un análisis detallado permitía ver que el motivo se debía a que se habían agrupado bajo ella todas las que tenían raíz ‘sab-‘ lo que ocasionaba que se incluyesen ‘saber’ y ‘sabía’. Para evitar distorsionar los resultados finales se decidió no hacer uso de esta función. Por otro lado en las nubes de palabras aparecen palabras sin valor analítico, por ejemplo ‘added’

Finalmente, de cara a las líneas de trabajo futuro podría hacerse un estudio de la polaridad de todos los tipos de mensaje (fotos, videos, links), ya que sólo se ha realizado para mensajes de texto. El estudio de la polaridad se ha realizado utilizando librerías de Python que necesitaban que el texto estuviese en inglés, para lo que se ha utilizado en traductor de Google. Podría en un futuro estudiarse la posibilidad de usar un diccionario en castellano que podría dar resultados más exactos sobre la polaridad.

Otra idea que podría realizarse en el futuro es explorar grupos de Facebook en otros idiomas y realizar una comparativa con lo obtenido en castellano para ver si se comparten temas, o establecer comparativas de preocupaciones por idioma/país. También puede servir para crear una nube con los nombres de las ER o los síntomas y estudiar cómo es la distribución de las ER por países. Asimismo, se podría realizar este estudio con datos de otras plataformas como Rareconnect para obtener más información de las preocupaciones y necesidades de los enfermos y familias con enfermedades raras.

4. Glosario

ANOVA: Modelo estadístico que permite el análisis de la Varianza

Boxplot: Diagrama de cajas basado en los cuartiles estadísticos.

Corpus: Conjunto amplio y estructurado de usos de la lengua. En nuestro caso conjunto de palabras.

Correlación: Valor entre [-1,1] que indica la fuerza y dirección de relación lineal entre dos variables estadísticas.

ER: Enfermedades raras.

Engagement: Valor que nos indica el nº de interacciones entre los usuarios y un mensaje. Es la suma del nº de likes, comparticiones, reacciones y comentarios.

Facebook: Red social que permite compartir contenido entre sus usuarios.

FEDER: Federación Española de Enfermedades Raras.

LL: Log-Likelihood.

Polaridad: Clasificación que permite establecer la positividad/negatividad lingüística de un mensaje

Python: Lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible.

P-valor: Medida de significación estadística.

R: Lenguaje de programación con un enfoque en el análisis estadístico.

Subjetividad: Clasificación que permite establecer la objetividad/subjetividad de un mensaje.

StemCompletion: Método que sirve para completar la raíz de una palabra. Asociado al Stemming, que consiste en primero establecer la raíz de las palabras.

5. Bibliografía

[1] FEDER. (2015). Decalogue of priorities. Disponible online: http://www.enfermedades-raras.org/images/diamundial/pdf /Decalogoprioridades.pdf.(accedido el 22-diciembre-2016).

[2] Grytics. (2016). Analíticas y Marketing de Grupos de Facebook. Disponible online: https://grytics.com/es/(accedido el 22-diciembre-2016).

[3] Sociograph (2015). Analytics for Facebook Groups and Pages. Disponible online: https://grytics.com/es/ (accedido el 22-diciembre-2016)

[4] Netvizz. (2016). Extracción de datos de Facebook. Disponible online: https://apps.facebook.com/netvizz/ (accedido el 22-diciembre-2016)

[5] EURODIS (2016). European Organization for rare diseases. Disponible online: http://www.eurordis.org/ (accedido el 22-diciembre-2016)

[6] ALIBER (2016). Alianza Iberoamericana de Enfermedades poco frecuentes. Disponible online: http://aliber.org/web/(accedido el 22-diciembre-2016)

[7] NORD (2016). National organization for rare diseases-USA. Disponible online:(accedido el 22-diciembre-2016)

[8] Raregenomics (2016). Disponible online: http://www.raregenomics.org/ (accedido el 22-diciembre-2016)

[9] Rareconnect (2016). Conectando globalmente pacientes con enfermedades raras. Disponible online: https://www.rareconnect.org/ (accedido el 22-diciembre-2016)

[10] CrowMed (2016). CrowdMed gets patients on the right path to a cure. Disponible online:https://www.crowdmed.com/ (accedido el 22-diciembre-2016)

[11]#LoweResearchProject. Trabajo colaborativo entre investigadores, padres y clínicos para afrontar la enfermedad de Lowe. Disponible online: http://www.enfermedades-raras.org/index.php?option=com_content&view=article&id=1628%3Ala-investigacion-la-unica-opcion-de-enfermedades-como-lowe&catid=114%3Acategoriaprueba1 (accedido el 22-diciembre-2016)

[12] Rooney, E. (2016). How to promote your rare disease story through social media. Disponible online: https://globalgenes.org/wp-content/uploads/2016/02/GGtoolkitsocial-mediastorytelling021016WEB.pdf (accedido el 9-diciembre-2016).

[13] Chempetitive Group (2016). Rare Diseases: The Role of Social Media in Patient Recruitment. Disponible online: https://chempetitive.com/chemunity/rare-diseases−role-social-media-patient-recruitment (accedido el 9-diciembre-2016).

[14] Smale, F. (2015). Social media and the rare disease community. Disponible online: http://www.orphan-drugs.org/2015/01/29/social-media-rare-disease-community (accedido el 9-diciembre-2016).

[15] Li, C., Bernoff, J. (2016). How to Engage with Rare Disease Patients Using Social Media. Disponible online: http://www.desminopathy.info/pdf /siren-interactive-whitepape.pdf (accedido el 9-diciembre-2016).

[16] Armayones, M., Requena, S., Gomez-Zúñiga, B., Pousada, M., Bañon, A.M. (2014). El uso de Facebook en asociaciones espanolas de enfermedades raras: como y para qué lo utilizan? Gaceta Sanitaria, 29(5): 335-340.

[17] Ram,S., Zhang, W., Williams, M., Pengetnze, Y. (2015). Predicting Asthma-Related Emergency Department Visits Using Big Data. IEEE Journal of Biomedical and Health Informatics, 19(4): 1216-1223.

[18] Subirats, L., Ceccaroni, L., Lopez-Blazquez, R., Miralles, F., García Rudolph, A., Tormos, J.M. (2013). Circles of Health: towards an advanced social network about disabilities of neurological origin. Journal of Biomedical Informatics, 46(6): 1006-1029.

[19] Subirats, L., Lopez-Blazquez, R., Ceccaroni, L., Gifre, M., Miralles, F.,García-Rudolph, A., Tormos, J.M. (2015). Monitoring and prognosis system based on the ICF for people with traumatic brain injury. International Journal of Environmental Research and Public Health, 12(8): 9832-9847.

[20] Palomino, M., Taylor, T., Gker,A., Isaacs, J., Warber, S. (2016). The Online Dissemination of Nature Health Concepts: Lessons from Sentiment Analysis of Social Media Relating to Nature-Deficit Disorder. Int. J. Environ. Res. Public Health, 13(1): 142.

[21] Larsen, M.E., Boonstra, T.W., Batterham, P.J., ODea, B., Paris, C.,Christensen, H. (2015). We Feel: Mapping Emotion on Twitter. IEEE Journal of Biomedical and Health Informatics, 19(4): 1246-1252.

[22] Yang, F.C., Lee, A.J., Kuo, S.C. (2016). Mining Health Social Media with Sentiment Analysis. J Med Syst, 40(11):236.

[23] Lasker, J.N., Sogolow, E.D., Sharim, R.R. (2005). The Role of an Online Community for People with a Rare Disease: Content Analysis of Messages Posted on a Primary Biliary Cirrhosis Mailinglist. J. Med. Internet Res., 7(1): e10.

[24] Jacobs, R., Boyd, L., Brennan, K., Sinha, C.K., Giuliani, S. (2016). The importance of social media for patients and families affected by congenital anomalies: A Facebook cross-sectional analysis and user survey information platform. Journal of Pediatric Surgery, 51(11): 1766-1771.

[25] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., Donaldson, L. (2013). Use of Sentiment Analysis for Capturing Patient experience From Free-Text Comments Posted Online. J. Med. Internet Res., 15: e239.

[26] Pojanapunya, P., Watson Todd, R. (2016). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. Corpus Linguistics and Linguistic Theory.

[27] Rayson, P., Garside, R. (2000). Comparing Corpora using Frequency Profiling. In The workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), 1-6.

[28] Gabrielatos, C., Marchi, A. (2012) Keyness: Appropriate metrics and practical issues. CADS International Conference. Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing?

[29] Wilson, A. (2013) Embracing Bayes factors for key item analysis in corpus linguistics. New approaches to the study of linguistic variability. Language Competence and Language Awareness in Europe. 3-11.

[30] Johnston, J.E., Berry, K.J., Mielke, P.W. (2006). Measures of effect size for chi-squared and likelihood-ratio goodness-of-fit tests. Perceptual and Motor Skills: 103:412-414.

[31] Diakopoulos, N., Elgesem, D., Salway, A., Zhang, A., Hofland, K. (2010). Compare Clouds: Visualizing Text Corpora to Compare Media Frames. In Proc. of IUI Workshop on Visual Text Analytics.

[32] Scanfeld, D., Scanfeld, V., Larson, E.L. (2009). Dissemination of health information through social networks: Twitter and antibiotics. American Journal of Infection Control, 38, 182-188.

[33] Log-likelihood and effect size calculator (2016). Disponible online: http://ucrel.lancs.ac.uk/llwizard.html (accedido el 07-Enero-2017).