TÍTULO

ANÁLISIS DE SENTIMIENTO EN TWITTER. REPUTACIÓN DE LAS PRINCIPALES AEROLÍNEAS EUROPEAS TRAS LA CRISIS SANITARIA DEL COVID-19

AUTOR

Adrián Amigo Portilla

Esta edición electrónica ha sido realizada en 2021

Tutores	Dr. D. Juan Diego Borrero Sánchez ; Dr. D. Gonzalo Antonio Aranda Corral
Instituciones	Universidad Internacional de Andalucía ; Universidad de Huelva
Curso	*Máster en Economía, Finanzas y Computación (2019/20)*
©	Adrián Amigo Portilla
©	De esta edición: Universidad Internacional de Andalucía
Fecha documento	2020

Análisis de sentimiento en Twitter. Reputación de las principales aerolíneas europeas tras la crisis sanitaria del covid-19

by

Adrián Amigo Portilla

A thesis submitted in conformity with the requirements
for the MSc in Economics, Finance and Computer Science

University of Huelva & International University of Andalusia

Septiembre 2020

Análisis de sentimiento en Twitter. Reputación de las principales aerolíneas europeas tras la crisis sanitaria del covid-19

Adrián Amigo Portilla

Máster en Economía, Finanzas y Computación

Juan Diego Borrero Sánchez
Gonzalo Antonio Aranda Corral
Universidad de Huelva y Universidad Internacional de Andalucía

2020

Abstract

The aim of this thesis is to analyse the impact of the health crisis caused by covid-19 on the reputation of the main European airlines. This reputation has been elaborated based on the comments that users leave on Twitter towards these airlines. To achieve this, a sentimental analysis of these comments has been made using Python and R, and was processed as follows. First, tweets have been extracted with Python and loaded in R for a cleaning process, creation of the corpus and analysis of the most frequent words. Then we proceeded with the sentiment analysis, in which we performed; classification of the tweet's sentiment in 10 different emotions, polarity of the words, general tweets score and general score (reputation) of the airline. Finally, a brief analysis is made of the variation in the airline's reputation after a period of one month, a comparison of overall results is made and the main conclusions obtained are discussed.

Keywords: sentiment analysis, airlines, twitter, Python, R.

Resumen

El objetivo de esta tesis es analizar el impacto que ha tenido la crisis sanitaria provocada por la covid-19 en la reputación de las principales aerolíneas europeas. Esta reputación se ha elaborado en base a los comentarios que dejan los usuarios en Twitter hacia estas aerolíneas, para lo cual, se hizo un análisis de sentimiento de estos comentarios haciendo uso de Python y R, y se procesó de la siguiente manera. Primero se realizó la extracción de tuits con Python y fue cargado en R para una limpieza del texto, creación del corpus y análisis de palabras más frecuentes. Luego se procede con el análisis de sentimiento, en el cual se realiza; clasificación del sentimiento del tuit en 10 emociones distintas, polaridad de las palabras, puntuación general de cada uno de los tuits y puntuación general (reputación) de la aerolínea. Finalmente, se hace un breve análisis de la variación de la reputación de las aerolíneas tras un periodo de un mes, se realiza una comparación de resultados generales y se abordan las principales conclusiones obtenidas.

Palabras clave: análisis de sentimiento, aerolíneas, twitter, Python, R.

Agradecimientos

De Andrés

Siempre, a mi madre, todos mis logros siempre serán gracias a ella.
A mi amigo y compañero de tesis Adrián, sin duda hacen falta más compañeros como él.

De Adrian

Primero que todo quiero agradecer a mi compañero de tesis Andrés, y más que eso amigo por su entrega, esfuerzo, y sacrificio desde comienzos del máster que hemos trabajado juntos hasta ahora. Esto fue lo que hizo posible el resultado de esta investigación.

Gracias a mi familia mi más grande tesoro en especial a mi hermana Giselle, mi padre, mi hermana Claudia, por todo el apoyo brindado desde lejos durante este año.

Gracias a los buenos amigos y compañeros que conocí durante todo este tiempo en especial a Martha, Katy, Andrés, Alex, Gerardo y Majo.

Gracias a Jeni, Claudia, Alberto, Yanelis, Jessica, Odaimys, David mis amistades de Cuba por la confianza depositada en mí, han sido un apoyo emocional muy fuerte y necesario durante este año.

De forma general los autores de esta tesis quieren dedicar un agradecimiento especial:

Al profesorado y dirección del máster de Economía Finanzas y Computación curso 2019-2020 por su entrega, capacidad de adaptación, paciencia y dedicación, que nos guiaron durante todo el curso en las circunstancias que nos tocó vivir.

Contenido

Tablas

Figuras

1 Introducción

La pandemia mundial provocada por el virus del covid-19 ha marcado un antes y un después en nuestra historia, los estragos que ha causado y las consecuencias que vendrán ya se están notando como un duro golpe en todos los sectores. En España particularmente, como se puede leer en una noticia publicada por el periódico El País (2020), el virus ha supuesto una caída trimestral del 18.5%, el mayor en comparación con otras economías referentes como Alemania (10.1%), Estados Unidos (9.5%), Francia (13.8%) o Italia (12.4%). En términos monetarios, una caída del PIB del 18.5% equivaldría a unos 300.000 millones de euros, y si añadimos que en el peor trimestre de la crisis financiera del 2008 la caída fuel 2.6%, sorprende aún más.

Otra noticia de este mismo periódico, asegura que se han destruido más de un millón de puestos de trabajo durante el confinamiento. Confinamiento que ha obligado a cancelar muchos planes turísticos, y ha supuesto enormes pérdidas para muchas compañías, como las aerolíneas. Una noticia de La Vanguardia (2020), nos dice que las aerolíneas no recuperarán el tráfico de pasajeros que tenían antes de la pandemia, hasta el año 2024.

Estas compañías aéreas no solo han tenido que afrontar enormes pérdidas, también han tenido que lidiar con miles de reclamaciones por parte de los usuarios que se han visto afectados por la cancelación de sus vuelos. La ingente cantidad de reclamaciones ha provocado un colapso que las compañías tardan en resolver, y a fecha de esta tesis aún no han resuelto muchas de ellas, podemos asegurarlo porque nos afectó personalmente, y aún estamos esperando una solución por parte de una de las aerolíneas que estudiaremos en este trabajo. La falta de solución y de respuesta por parte de estas compañías, ha hecho que los usuarios se manifiesten por diferentes medios hacia ellas. La gran mayoría de los mensajes, tenían un contexto de reclamación exigiendo una solución para su caso particular.

Todo esto ha afectado a su reputación, especialmente durante los meses del confinamiento. En esta parte es donde se hace foco con este trabajo, en la reputación de las aerolíneas en base a los comentarios de los usuarios hacia ellas en Twitter. Lamentablemente no se ha podido estudiar la reputación durante los meses del confinamiento, ya que la API de Twitter no permite extraer tuits para una fecha anterior a 7 días, por ello, se hizo el análisis para estudiar la reputación de la última semana de julio de 2020, momento en se empezó a recolectar los datos, y se volvió a

realizar para la primera semana de septiembre de 2020 y ver si en un periodo de un mes se han producido cambios significativos.

1.1 Elección de la base de datos

Las bases de datos que se han utilizado, se han extraído de las 5 principales compañías aéreas (aerolíneas) de Europa por número de pasajeros. Se ha tenido en cuenta, no solamente el número de pasajeros de estas compañías, sino también que podamos extraer de ellas una cantidad de tuits similar. En un primer momento, se intentó extraer 5000 tuits de cada una de las aerolíneas (7 al principio), no obstante, casi ninguna llegaba a esta cifra debido a la limitación de la API de Twitter, por lo que quitamos 2 aerolíneas que contaban con menos mensajes, y se disminuyeron la cantidad de tuits a extraer hasta encontrar una cifra que fuera semejante para todas.

Esta cifra fue 2000 tuits para todas ellas, con la excepción de British Airways en Julio, de la cual sólo pudimos extraer 1745. Hay que destacar también, que estos 2000 tuits son en bruto, y tras el procesamiento y limpieza, se pierden tuits, por lo que finalmente la cifra es menor. En la siguiente tabla, se pueden ver la cantidad de tuits extraídos en Julio y Septiembre y la cantidad de tuits restantes tras el preprocesamiento.

Aerolínea	Tuits extraídos (Julio)	Tuits después del procesamiento	Tuits extraídos (septiembre)	Tuits después del procesamiento
Lufthansa	2000	1415	2000	1205
British A.	1745	686	2000	1251
Easyjet	2000	1586	2000	1585
KLM	2000	1584	2000	1414
Ryanair	2000	1661	2000	1376

Tabla 1: Tuits antes y después del preprocesamiento

Fuente: Elaboración propia

En principio pensamos en excluir a British Airways, ya que la cantidad de tuits restantes tras el procesamiento es considerablemente menor que el resto. No obstante, decidimos dejarla ya que la reputación final está ponderada y es interesante ver su evolución.

En la figura 1 se puede ver el ranking de las principales aerolíneas europeas por número de pasajeros en el año 2018.

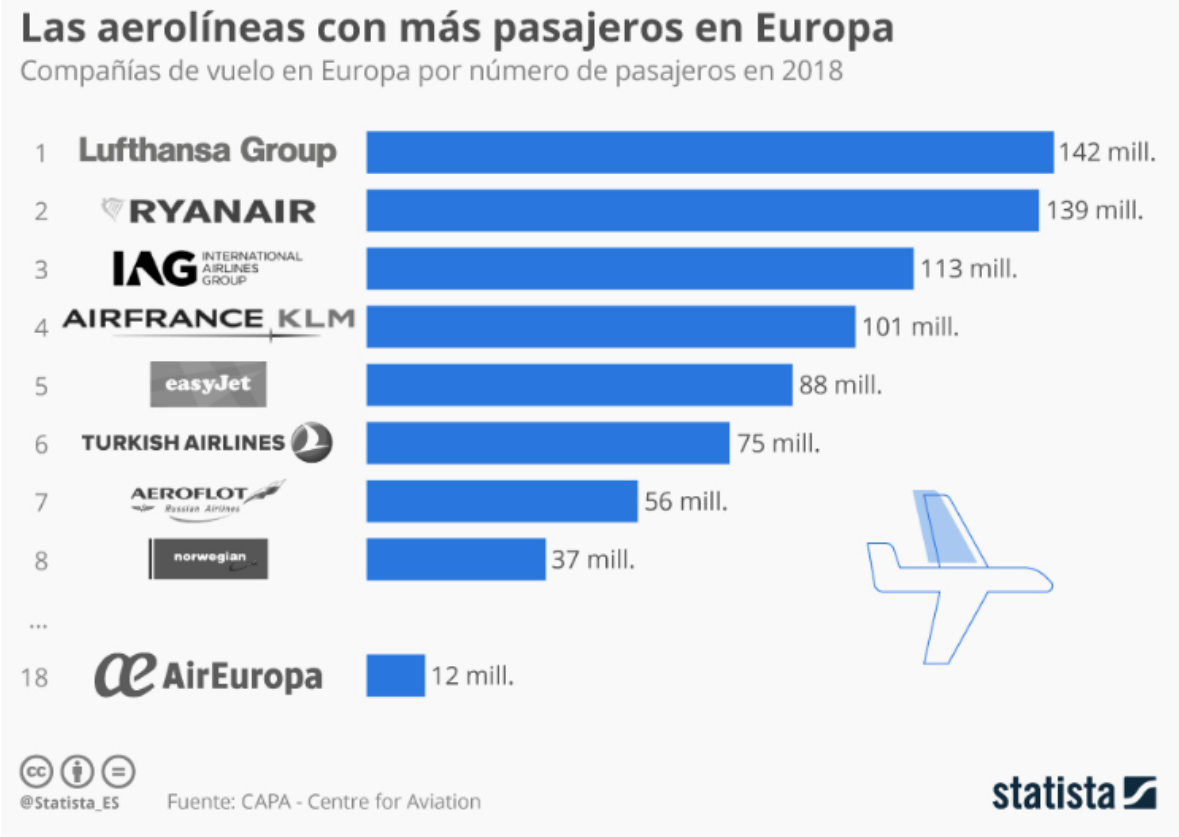

Figura 1: Aerolíneas con más pasajeros en Europa (2018)

Fuente: Statista.

Teniendo en cuenta que el grupo IAG contiene 5 aerolíneas (British Airways, Iberia, Aer Lingus, Level y Vueling) se ha seleccionado aquella, dentro de este grupo, que cumpliera con los requisitos antes descritos.

Finalmente, las compañías aéreas que son objeto de este estudio son las siguientes:

- Lufthansa
- Ryanair
- British Airways
- KLM
- Easyjet

1.2 Metodología

Lo primero que debemos tener es la fuente de los datos de donde obtendremos la información. Para ello es necesario la extracción de los tuits a través de la API de Twitter, estos se han recogido de forma periódica de modo que permita observar la variación de los datos entre dos momentos de tiempo.

Una vez que se tiene la base de datos, lo siguiente es preparar estos datos para ser procesados y obtener información. Se realiza en primer lugar la depuración del texto contenido en los tuits, se realiza una limpieza de caracteres que son de poca utilidad, eliminando símbolos extraños, links, signos de puntuación, espacios en blanco y pasando el texto a minúscula.

El próximo paso es convertir todo el texto de los tuits a un formato analizable, para poder crear una representación estructurada de los mismos. En este caso, ordenar los datos en forma de lista dentro de un corpus.

Se crea el corpus que no es más que el conjunto de documentos con el que se trabaja. También se realiza una pequeña limpieza de este corpus, se eliminan palabras vacías, es decir, aquellas que tienen ningún o poco valor para el análisis, tales como algunas preposiciones y muletillas. Se remueve los números ya que no necesitamos analizar cantidades en este caso. Por último, eliminamos los espacios vacíos excesivos, muchos de ellos introducidos por las transformaciones anteriores.

Como complemento se ha creado una nube de palabras que nos muestra los términos más usados presentes en los tuits así como generar gráficos de frecuencia con las palabras más usadas, pudiendo determinar así, cuán frecuente es un término.

Con el corpus seleccionado y estructurado, debemos reconocer los tokens (unidades gramaticales más pequeñas), lo que implica representar el texto como una lista de palabras mediante una representación vectorial. Luego de esto se procesa el texto a través de léxicos, que basándose en técnicas de PLN, permiten hacer el análisis de sentimientos.

Se utiliza el léxico NRC para establecer una relación palabra-emoción y clasificarlas por emociones y por sentimientos positivo o negativo. A su vez se utiliza el léxico Bing para de igual forma clasificar las palabras en positivas y negativas y establecer una ponderación al final de

cuan negativo o positivo puede ser el tuit según la polaridad que arroje el mismo y poder establecer una reputación final de cada aerolínea en escala [-5,5].

Se representa visualmente los datos obtenidos a través de graficas que aportan mayor comprensión e interpretación de la información que se pueda obtener del análisis de sentimientos.

1.3 Lenguajes de programación

La ciencia de datos ha posicionado a R y Python como los lenguajes de programación más usados por parte de las empresas para el manejo de los datos, la visualización de información compleja y la toma de decisiones.

Mientras el lenguaje R dispone de un enfoque matemático, Python es un lenguaje de alto nivel multipropósito cuya principal ventaja es la facilidad en la legibilidad del código. Con lo cual, si lo que se necesita, por ejemplo, es realizar un análisis estadístico complejo, la elección de R como lenguaje será más recomendada debido a la gran cantidad de librerías para este propósito de las que dispone y a la sencillez de implementación de los algoritmos necesarios en pocas líneas de código. Por otra parte, si lo que precisamos es la implementación de un código de tipo más general o no tan matemático, la elección recomendada sería Python debido a la sencillez del desarrollo con este lenguaje.

En este trabajo se ha utilizado ambos lenguajes de programación. Se utiliza Python para conectar con la API de Twitter y realizar la extracción de los tuits generados sobre las diferentes aerolíneas. Luego se cargan los tuits contenidos en un fichero csv para ser utilizados en el código de R, donde se realiza el análisis de sentimientos y el mapeo de los términos más frecuentes a través de una nube de palabras.

1.4 Objetivos

Objetivo general: Determinar la reputación de las principales aerolíneas de Europa durante la Covid 19 mediante una puntuación en escala [-5, 5]

Otros objetivos específicos que se abordan son los siguientes:

- Obtener comentarios de texto (tuits) en Twitter.
- Conocer la variación del sentimiento en un periodo de un mes.
- Clasificar los comentarios positivos o negativos.
- Clasificar los comentarios por emociones.
- Obtener las palabras más utilizadas (tópicos) y elaborar una nube de palabras
- Agrupar los comentarios por puntuación según su nivel de polaridad.
- Establecer un ranking de mayor a menor reputación de las aerolíneas estudiadas.

2 Extracción de datos (Text mining)

En este análisis se ha hecho uso de la red social Twitter para extraer los tuits que los usuarios dejan hacia las aerolíneas y utilizarlos como objeto de estudio. Para poder cumplir con este objetivo, se necesita estar registrado en esta red social y solicitar la cuenta de desarrollador, la cual se puede solicitar en el siguiente enlace https://developer.twitter.com/en/apps. Una vez aprueban la solicitud, se debe acceder a ella y crear una app, únicamente se rellenan los datos y se registra la app creada.

Realizado lo anterior, se accede a la pestaña «*keys and Access Tokens*» y se pulsa en el botón «*Generate My Access Token and Token Secret*», estas son credenciales para poder acceder a la API de Twitter y extraer los tuits que se necesitan. Véase la figura 2.

Para realizar la extracción de Tuits, se ha hecho uso de la herramienta Jupyter Notebook (Anaconda 3) utilizando Python. El primer paso con esta herramienta, es la instalación de las librerías necesarias que permiten acceder a la API de Twitter.

Primero se instala "**pip**", esta permitirá importar la librería **tweepy**, la cual es necesaria para trabajar con twitter desde Python. En la figura 3 y 4 se observa se realiza esta instalación en Jupyter Notebook.

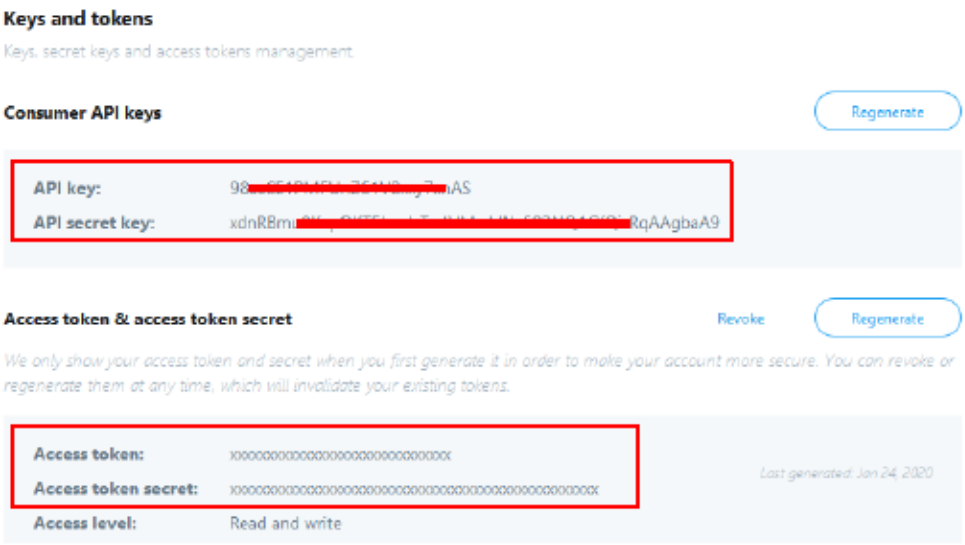

Figura 2: Credenciales de Twitter

Fuente: Elaboración propia

```
1 pip
```

```
Note: you may need to restart the kernel to use updated packages.
Usage:
  C:\Users\Andre\anaconda3\python.exe -m pip <command> [options]

Commands:
  install                     Install packages.
  download                    Download packages.
  uninstall                   Uninstall packages.
  freeze                      Output installed packages in requirements format.
  list                        List installed packages.
  show                        Show information about installed packages.
  check                       Verify installed packages have compatible dependencies.
  config                      Manage local and global configuration.
  search                      Search PyPI for packages.
  cache                       Inspect and manage pip's wheel cache.
  wheel                       Build wheels from your requirements.
  hash                        Compute hashes of package archives.
```

Figura 3: Instalación de pip en Jupyter Notebook

Fuente: Elaboración propia

```
1 # INSTALAMOS LIBRERÍA TWEEPY
2 pip install tweepy
```

```
Requirement already satisfied: tweepy in c:\users\andre\anaconda3\lib\site-pack
ages (3.8.0)
Requirement already satisfied: requests-oauthlib>=0.7.0 in c:\users\andre\anaco
nda3\lib\site-packages (from tweepy) (1.3.0)
Requirement already satisfied: six>=1.10.0 in c:\users\andre\anaconda3\lib\site
-packages (from tweepy) (1.12.0)
Requirement already satisfied: PySocks>=1.5.7 in c:\users\andre\anaconda3\lib\s
ite-packages (from tweepy) (1.7.1)
Requirement already satisfied: requests>=2.11.1 in c:\users\andre\anaconda3\lib
\site-packages (from tweepy) (2.22.0)
Requirement already satisfied: oauthlib>=3.0.0 in c:\users\andre\anaconda3\lib
\site-packages (from requests-oauthlib>=0.7.0->tweepy) (3.1.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\andre\anaconda3\l
ib\site-packages (from requests>=2.11.1->tweepy) (2019.9.11)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in c:\users\andre\anaconda
3\lib\site-packages (from requests>=2.11.1->tweepy) (3.0.4)
Requirement already satisfied: idna<2.9,>=2.5 in c:\users\andre\anaconda3\lib\s
ite-packages (from requests>=2.11.1->tweepy) (2.8)
```

Figura 4: Instalación de tweepy

Fuente: Elaboración propia

Una vez realizado este paso, ya se puede trabajar en el código, el cual se ha dividido en 4 partes principales (véase anexo I para ver el código completo): Autenticación, interfaz de usuario, extracción y limpieza.

Importamos las librerías **tweepy** y **csv**, la primera para poder acceder a la API de Twitter desde Python y la segunda para guardar los tuits en un csv que posteriormente será cargado en R. Guardamos las credenciales en variables y las autenticamos en Twitter.

```python
#Se importa la librería tweepy
import tweepy
import csv

#Credenciales del Twitter API
consumer_key = "98so[redacted]"
consumer_secret = "Xd[redacted]"
access_token = "3648[redacted]"
access_token_secret = "[redacted]"

#Se autentica en twitter
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit=True, wait_on_rate_limit_notify=True, compression=True)
```

Figura 5: Autenticación de las credenciales en Twitter desde Python.

Fuente: Elaboración propia

Una vez se ejecuta el código, este generará una interfaz en la que se le preguntará al usuario lo siguiente:

- **Palabra clave por la que desea realizar la búsqueda**: En nuestro caso, la palabra clave será el nombre de la aerolínea en cuestión, de esta forma conseguimos extraer los tuits que hagan mención a ella. Comentar que no se extraen los tuits que realiza la aerolínea, sino que se han extraído tuits que van dirigido hacia estás aerolíneas, es decir, tuits que hagan mención a estas cuentas, ya sea mencionándolas directamente, o por medio de hashtags o en el mensaje normal. No obstante, si la aerolínea en cuestión se menciona a ella misma, este mensaje si estará incluido, lo cual no suele ser común, pero si en los casos que son retuits. Los retuits y replies también se extraen, no obstante, los retuits son eliminados en el procesamiento ya que son duplicidades, y en cuanto a las replies, sólo se extrae el mensaje de respuesta si se nombra la aerolínea en cuestión, lo cual es bueno considerarlo ya que es una opinión más.
- **Número de tuits a extrae**r: Hay que tener en cuenta que la "*Search API*", la API de Twitter que permite extraer tuits, tiene dos limitaciones. La primera es que sólo permite extraer tuits con una profundidad en el tiempo máxima de 7 días. La segunda es que está limitada a 180 peticiones cada 15 minutos. Es decir, si pones un número alto de tuits, el código generará error porque la API no le deja acceder debido a esta limitación, por ende, el código se detendrá y no se podrán extraer el resto de tuits. No obstante, se ha programado el código para que se ponga en "pausa" hasta que pasen los 15 minutos y pueda seguir extrayendo, de esta forma, aunque se llegue al límite, el código no

devolverá error, sino que estará en pausa para continuar con la extracción una vez acabe el tiempo límite.

Además de esto la interfaz también le recordará que la búsqueda realizada debe ser en inglés.

Una vez introducidos los datos, el código le avisará que se están extrayendo los tuits y una vez concluido le saldrá otro mensaje informando que la extracción se produjo satisfactoriamente. En la figura 6 está representado todo este proceso de ejemplo.

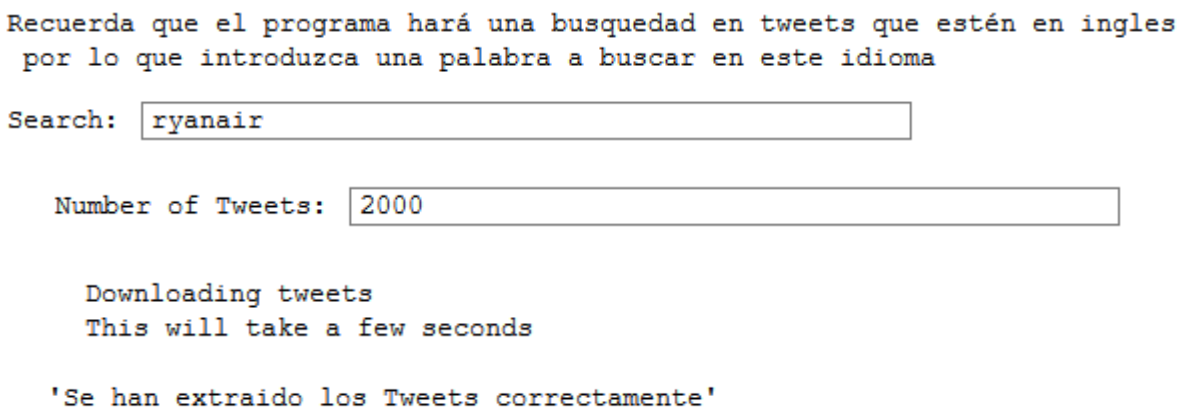

Figura 6: Interfaz para la extracción de tuits en Python

Fuente: Elaboración propia

Ya descargados los tuits, se habrá generado un csv con toda la información extraída. La figura 7 representa una muestra del csv generado al descargar los tuits dirigidos hacia *Ryanair*.

	A	B	C	D	E	F	G	H	I
1	id,"created_at","text","retweet_count","favorite_count"								
2	1287283340983144448,"2020-07-26 07:07:19","@Ryanair can me and my family fly to BCN in August ???? #ryanair","0","0"								
3	12872829269	be happy to refund. I would try them , you never knowâ€¦ https://t.co/OkQh418OmT","0","0"							
4	1287282636864593920,"2020-07-26 07:04:31","@Ryanair I think you should update your site to inform customers they will h								
5	1287282497114517504,"2020-07-26 07:03:58","RT @aimee_gilhooley: My friend has just flew back from Tenerife with @Rya								
6	1287282447516925953,"2020-07-26 07:03:46","@rossp77 @Ryanair Iâ€™m hoping they wonâ€™t for Tenerife though as the								

Figura 7: CSV generado al descargar los tuits de Ryanair.

Fuente: Elaboración propia

Una vez extraído los tuits, ya tenemos nuestra base de datos con la que vamos a trabajar. Esta base de datos, cuyos datos están en bruto, serán procesados y analizados haciendo uso de la herramienta R Studio y el lenguaje de programación R.

3 Procesamiento de los datos

Preparación de la base de datos.

Una vez creada la base de datos, y para evitar que surjan errores de procesamiento es necesario preparar los datos para su posterior análisis. En este sentido se suprime información y texto contenido en el tuit que es irrelevante para el análisis de sentimientos. Lo primero es eliminar los tuits duplicados ya que estos no aportan información relevante debido a que el texto que estos contienen ya ha sido analizado en el tuit original. Esto evita ponderar dos veces una misma polaridad negativa o positiva sobre el contenido de un tuit que ya ha sido procesado con anterioridad. En R una vía para eliminar tuits duplicados es de la siguiente forma:

tuit_text<-unique(tuit_text)

Una vez que se han eliminado los tuits duplicados se procede a la limpieza del texto en específico.

3.1 Limpieza del texto y creación del corpus

Limpieza del texto.

Es muy común que cuando procesamos información obtenida de internet y datos procedentes de otras fuentes con otros formatos, sea necesario un proceso de limpieza de texto antes de ser analizado. En este aspecto gana vital importancia la función gsub().

Esta funcionalidad permite que; dada una cadena de texto, buscar patrones similares a esta cadena de texto dentro de un vector columna, y cada ocurrencia modificarla por un nuevo texto que se especifica dentro de la función gsub().

Por ejemplo: ***tuit_text <- gsub("€", " ",tuit_text)***

Sobre el vector tuit_text, en la columna llamada tuit_text, todo texto o palabra que sea igual a: € (símbolo de euro) remplázalo por “” (espacio en blanco). Esta es una forma muy sencilla de eliminar símbolos y caracteres que no aportan validez emocional al texto contenido en el tuit. Y que luego dichos espacios en blancos pueden ser eliminados haciendo uso de la misma función. Se obtiene como resultado un tuit libre de caracteres y símbolos innecesarios.

En este caso son varias las transformaciones que son necesarias realizar y que se estimaron irrelevantes para los resultados que se desean obtener con el análisis de sentimiento. Los patrones que se quieren eliminar del texto contenido en el tuit se definen a continuación:

- **Convertir el texto a minúscula**: Se convierten todas las palabras que conforman el texto del tuit para estandarizarlo. Para esto se utiliza el comando *tolower,* una de las ventajas que aporta es eliminar los acentos a todos aquellos caracteres que lo contengan.
  tuit_text <- tolower(tuit_text)

- **Enlaces a páginas webs**: Cuando se crea un post, y o se comparte una publicación, en este caso un tuit, es muy común que se utilicen referencias a otras páginas a través de enlaces conocidos como direcciones url. En Twitter es muy habitual este comportamiento. Estos enlaces pueden generar errores de procesamiento ya que contienen muchos caracteres y símbolos diferentes que no aportan información al análisis de sentimiento.

  tuit_text <- gsub("http\\S+\\s*", "",tuit_text)

- **Nombre de Usuario (@):** En Twitter para poder registrarse es necesario tener un alias conocido como nombre de usuario. Esto permite mencionar a ese usuario en el momento de escribir un tuit utilizando el caracter @. Este símbolo puede provocar problemas en el análisis de sentimientos por lo tanto es necesario suprimirlo.
  tuit_text <- gsub("@\\w+", "", tuit_text)

- **Nombres específicos**: De la misma forma que un símbolo puede generar ruido en el análisis de sentimientos, también lo puede hacer un nombre dependiendo del tema de análisis que se desee hacer. En este caso es necesario eliminar del tuit la palabra que sugiere el nombre de la aerolínea ya que no es relevante para determinar la polaridad y por ende el sentimiento del tuit. El proceso es muy sencillo, se pueden especificar nombres que se quieran suprimir del texto a analizar como se muestra a continuación:

  #tuit_text <- gsub("british", "",tuit_text)

  #tuit_text <- gsub("airway", "",tuit_text)

- **Retuit (RT) a otros Usuarios**: Esta función que brinda Twitter, permite a los usuarios volver a publicar nuevamente un tuit. Cuando se hace un retuit la estructura del nuevo tuit que se genera sigue la siguiente forma: “RT @NombreUsuario: texto_tuit…”. Como se puede observar estas palabras no son necesarias para analizar por lo que de igual forma se suprimen

  tuit_text <- gsub("rt", "",tuit_text)

- **Hashtags**: Un hashtag es una palabra o varias palabras concatenadas precedidas del símbolo de almohadilla #. De esta forma convierte a la palabra en “clicable” y sirve para etiquetar un mensaje en la web e identificarla fácilmente. Esto conlleva a que en muchos casos un usuario pueda redactar un tuit expresando sentimientos pero de la siguiente forma: me han #estafado, #no #reembolso ejemplos como “la #peor aerolínea”. Como se puede observar contiene información relevante que es necesaria tener en cuenta en el análisis de sentimiento. Para esto se suprime el símbolo de almohadilla # y se mantiene el resto del texto.

  tuit_text <- gsub("[[:punct:]]", "", tuit_text)

- **Espacios en blancos**: Debido a todas las transformaciones que sufre en el texto contenido en el tuit durante todo el proceso de limpieza de texto, se generan espacios en blancos adicionales antes y/o después de cada palabra analizada. Estos espacios se pueden eliminar de la siguiente forma:

  Espacios en blancos al principio: ***tuit_text <- gsub("^ ", "",tuit_text)***

  Espacios en blancos al final: ***tuit_text <- gsub(" $", "",tuit_text)***

Limpieza del texto generado en el Corpus: Como se ha mencionado, el corpus es el conjunto de datos estructurado que contiene ejemplos reales lingüísticos, sirve como diccionario de términos con los que se realiza en este caso el análisis de sentimientos.

Una vez que se realiza la limpieza del texto contenido en el tuit es necesario estructurar ese texto de forma tal que pueda ser entendible por las herramientas que realizan el análisis de opiniones. En este proceso de creación del Corpus pueden surgir pequeñas modificaciones o desajustes en el texto, con lo cual a modo de seguridad se realiza una limpieza del texto contenido en el corpus.

Para esto se utiliza la librería *tm_map*: Esta función toma un vector de caracteres como el que contiene el texto de los tuits y lo transforma devolviendo como salida un vector documento de igual longitud con el texto que se le ha pasado. Aporta a su vez una serie de funcionalidades que sirven para hacer modificaciones sobre el texto del corpus, en este caso depurar la información de datos innecesarios.

- Se suprimen las cifras numéricas contenidas en el texto del tuit puesto que para este análisis esas cantidades no aportan valor sentimental.

  docs <- tm_map(docs, removeNumbers)

- Cuando clasificamos una palabra como neutral normalmente refiere a que no aporta algún significado o en el caso de análisis de sentimiento no brindan valor sentimental. Este tipo de palabras son conocidas como palabras vacías o stopwords. Dependiendo del idioma la cantidad de stopwords existente varía. En el idioma español se calcula que varía entre un 30% y 40% en ingles la cifra aumenta puesto que es un idioma donde estas palabras no existen por si solas, sino que acompañan o modifican a otras. Por lo general son artículos, preposiciones, pronombres, algunos verbos y otras que en minería de texto pueden ser filtradas antes o después del procesamiento de los datos.

  docs <- tm_map(docs, removeWords, stopwords("en"))

  docs <- tm_map(docs, removeWords, stopwords("SMART"))

- Se eliminan signos de puntuación y espacios en blanco mayormente generados de las transformaciones que sufre el texto.

  docs <- tm_map(docs, removePunctuation)

  docs <- tm_map(docs, stripWhitespace)

3.2 Tratamiento de Bots

Cuando se menciona el termino bot, las personas lo asocian a un proceso que se realiza de forma automatizada o una cuenta que se mantiene en el anonimato. Según el blog oficial de Twitter un bot es solo una cuenta automatizada y no un comentario. Twitter se ha enfocado en resolver lo que en un principio suponía un problema, tratando de mitigar los efectos de estos bots sobre la manipulación de la red social como plataforma y no como servicio de comunicación. O sea, el uso indebido de la automatización centrando su atención en el comportamiento de los bots y no en el contenido que generan. Es decir, las técnicas que los bots puedan utilizar para manipular las conversaciones de los usuarios en Twitter y no al contenido que estos comparten.

Por lo tanto, Twitter trata de eliminar los bots que hacen:

- Uso malicioso de la automatización para socavar e interrumpir la conversación pública, tratando de hacer que algo sea tendencia.
- Amplificación artificial de conversaciones en Twitter, incluso mediante la creación de cuentas múltiples o sobrepuestas.
- Generar, solicitar o comprar interacciones falsas.
- Tuitear, seguir o interactuar con cuentas de manera masiva o exagerada.
- Usar Hashtags con fines de spam, incluyendo el uso de Hashtags no relacionados a un tuit (también conocido como "hashtag cramming").

Mensualmente la red social suspende permanentemente millones de cuentas que son reportadas o no deseadas, pero aun así no toma como spam o comportamiento indebido ni siquiera a un usuario que postea más de 100 veces al día bajo un mismo hashtag. Para identificar un bot tienen en cuenta muchos otros aspectos como el nombre de la cuenta, el nivel de interacción en el servicio, la ubicación en la biografía, los Hashtags utilizados, etc.

En el estudio que se realiza para las aerolíneas en cuestión, la presencia de bots es muy poco probable. Para el tamaño de la muestra que se está analizando cómo se hacía alusión al principio, es poco probable que exista alguno y en caso de existir no altera la finalidad de los resultados porque su contribución al sentimiento podría ser nula o insignificante. Teniendo en cuenta todo el poder tecnológico que realiza Twitter para identificar y eliminar de manera proactiva los bots, no es necesario hacer un tratamiento de ellos de forma particular en este trabajo.

4 Análisis

Una vez se ha realizado la limpieza del texto, se puede proceder a realizar el análisis del texto. Durante este epígrafe se muestran los distintos análisis sobre los cuales se basa este estudio como el análisis de frecuencia y el análisis de sentimientos. Se hará uso de los léxicos Bing y NRC que explicaremos luego y se irá procediendo según la metodología y objetivos planteados al inicio de este trabajo.

4.1 Análisis de frecuencia - Nube de palabras

Antes de realizar la nube de palabras, se ha hecho uso de *TermDocumentMatrix* del paquete de minería de texto para elaborar, primero, una matriz de términos que contiene las 10 palabras más frecuentes, así como un plot para visualizar el histograma de frecuencias de los mismos. Esto es importante para conocer exactamente la frecuencia de las palabras más empleadas por los usuarios, ya que nos ayudará a interpretar mejor la nube de palabras.

La librería *Wordcloud* es la que permite generar una nube de palabras en la que se muestran los términos en diferentes tamaños según la frecuencia en la que aparecen. Cuanto más grande se vea la palabra en la nube, quiere decir que su frecuencia en el texto es mayor que otras palabras cuyo tamaño es más pequeño en la nube.

La principal ventaja de una nube de palabra es la posibilidad de obtener una representación visual de las palabras más frecuentes o más importantes, según se mire, utilizadas en un párrafo o fragmento de texto. Los resultados obtenidos para cada una de las aerolíneas se muestran a continuación.

British Airways

Palabra	Frecuencia
staff	83
babetrayal	54
travel	45
fireandrehire	44
flight	43
amp	31
support	29
loyal	27
stop	24
voucher	24

Tabla 2: Matriz de frecuencia de palabras (British Airways)

Fuente: Elaboración propia

Figura 8: Histograma de frecuencia de palabras (British Airways)

Fuente: Elaboración propia

Figura 9: Nube de palabras (British Airways)

Fuente: Elaboración propia

Easyjet

Palabra	Frecuencia
flight	251
refund	211
cancelled	147
travel	126
fly	108
voucher	98
change	96
quarantine	90
cancel	81
ive	77

Tabla 3: Matriz de frecuencia de palabras (Easyjet)

Fuente: Elaboración propia

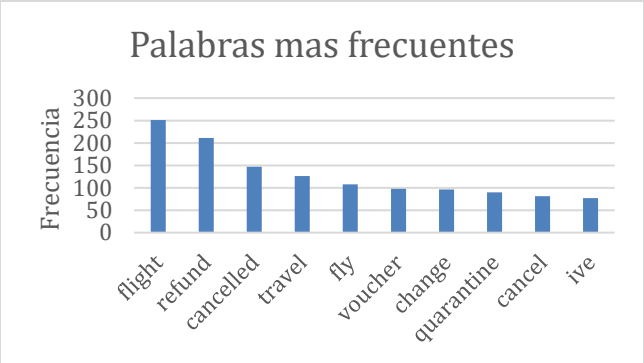

Figura 10: Histograma de frecuencia de palabras (Easyjet)

Fuente: Elaboración propia

Figura 11: Nube de palabras (Easyjet)

Fuente: Elaboración propia

KLM

Palabra	Frecuencia
flight	69
lisa	58
love	49
bts	48
today	42
man	40
good	40
refund	40
airline	37
woman	37

Tabla 4: Matriz de frecuencia de palabras (KLM)

Fuente: Elaboración propia

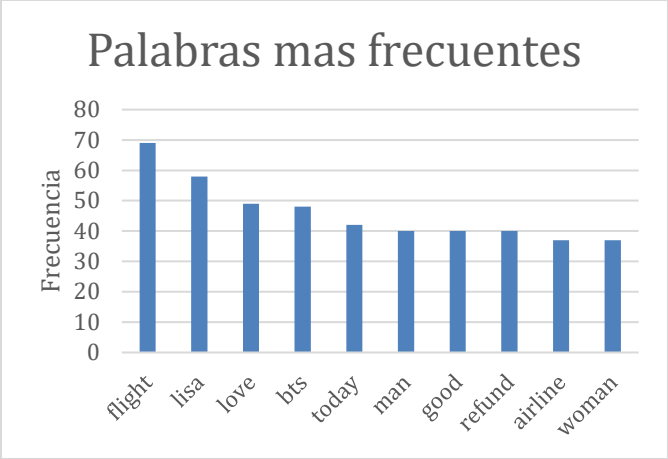

Figura 12: Histograma de frecuencia de palabras (KLM)

Fuente: Elaboración propia

Figura 13: Nube de palabras (KLM)

Fuente: Elaboración propia

Lufthansa

Palabra	Frecuencia
flight	174
refund	152
travel	88
india	84
cancelled	75
frankfurt	65
airlines	58
waiting	56
air	52
customer	47

Tabla 5: Matriz de frecuencia de palabras (Lufthansa)

Fuente: Elaboración propia

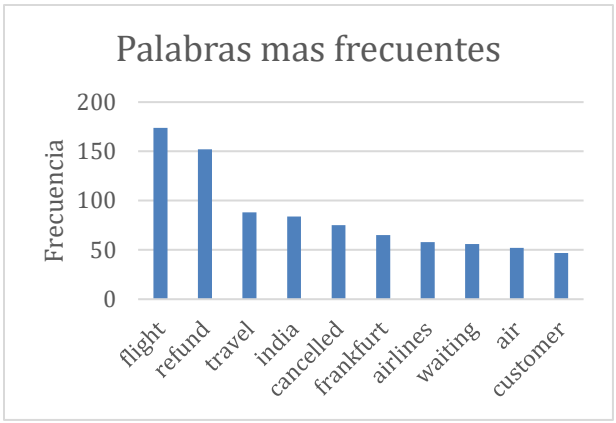

Figura 14: Histograma de frecuencia de palabras (lufthansa)

Fuente: Elaboración propia

Figura 15: Nube de palabras (lufthansa)

Fuente: Elaboración propia

Ryanair

Palabra	Frecuencia
refund	304
flight	235
waiting	119
cancelled	108
money	100
change	93
back	89
fly	82
refunds	76
booked	74

Tabla 6: Matriz de frecuencia de palabras (Ryanair)

Fuente: Elaboración propia

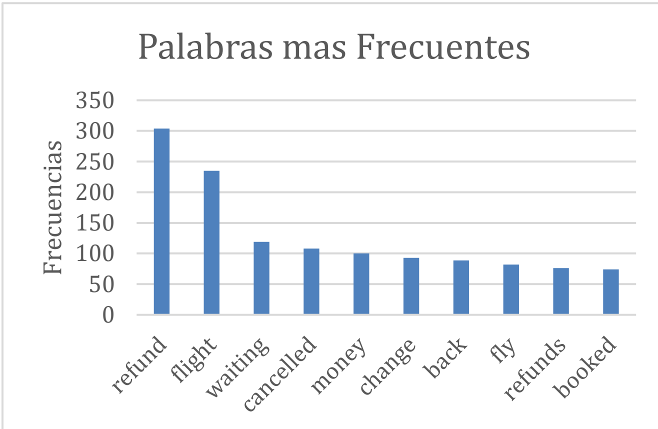

Figura 16: Histograma de frecuencia de palabras (Ryanair)

Fuente: Elaboración propia

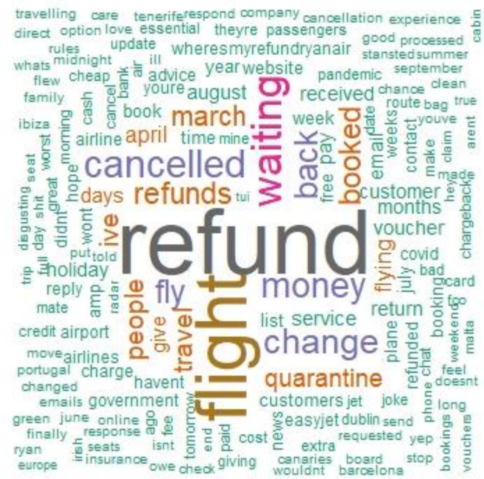

Figura 17: Nube de palabras (Ryanair)

Fuente: Elaboración propia

Se puede apreciar que hay palabras en cada aerolínea que son comunes en todas ellas, alguna quizá no aparezca en el top 10, pero si se encuentra muy cerca. En la siguiente tabla se ha recogido las 6 palabras más comunes en todas las aerolíneas y en la figura 18 se puede ver su frecuencia en cada una de estas compañías.

Palabra	Frecuencia
flight	1578
refund	965
cancel	599
book	436
travel	427
voucher	246

Tabla 7: Palabras más comunes en las 5 aerolíneas

Fuente: Elaboración propia

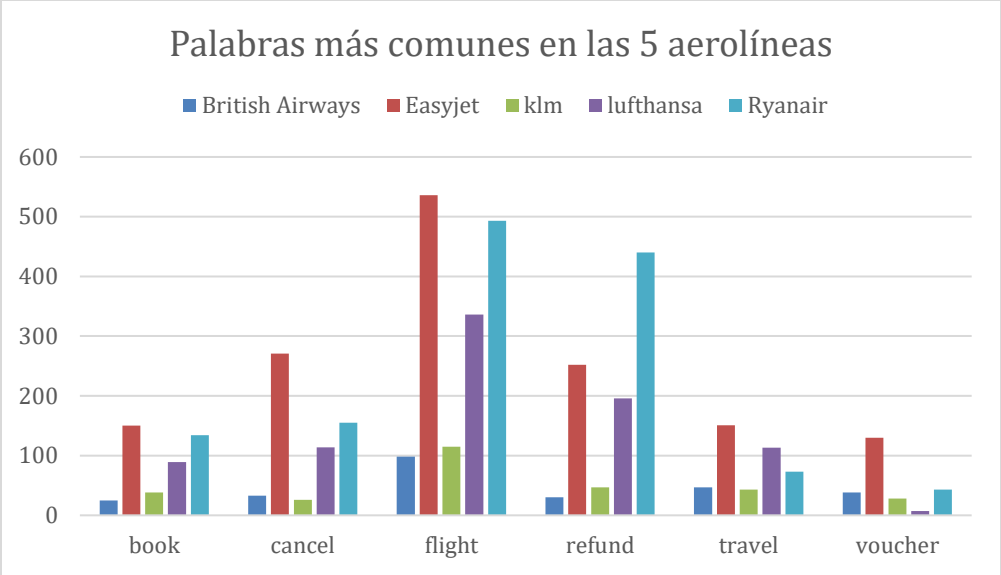

Figura 18: Frecuencia por cada aerolínea de las palabras más comunes.

Fuente: Elaboración propia

Tiene bastante sentido que entre las 5 palabras más comunes que han empleado los usuarios de Twitter en sus mensajes hacia las aerolíneas, se encuentre la palabra “refund”, “voucher" y “cancel”. Debido a la cancelación de los vuelos por el virus del covid-19 la gran mayoría de los usuarios se dirigen hacia las aerolíneas solicitando una solución para su vuelo cancelado, solución que suele ser o bien reembolso del dinero o bien un vale en su cuenta de la aerolínea respectiva.

4.2 Análisis de sentimiento

El análisis de sentimientos en la red social Twitter nos permite tener una sensación general de lo que se está tuiteando sobre un tema en particular. Parte de la aplicación de técnicas de procesamiento del lenguaje natural PNL, que nos permiten identificar y obtener información subjetiva sobre un fragmento de texto que a priori puede o no proyectar un sentimiento de forma clara.

4.2.1 Léxicos

Antes de tratar el análisis de sentimientos, es importante comentar los léxicos o diccionarios específicos utilizados, de donde parte la base del procesamiento de texto que se realiza en este estudio.

La definición que da Wikipedia de la palabra "léxico", es la siguiente *"Léxico es el conjunto de palabras que conforma un determinado lecto y, por extensión, también se denomina así a los diccionarios que los recogen".* R como lenguaje de programación aporta varias librerías y paquetes que contienen léxicos que permiten evaluar las emociones predominantes en los textos contenidos en uno o más tuits.

En el análisis de sentimientos que se desarrolla en este estudio se utilizan los léxicos de propósito general Bing y NRC. Estos léxicos, son unigramas que contienen palabras únicas en inglés. Los tuits que se recogen son sobre aerolíneas que operan con usuarios en su mayoría, de habla inglesa, por todo ello, este análisis se enfoca en tuits en idioma inglés. Estos léxicos, para cada una de estas palabras que incorporan en su diccionario, las procesarán de una forma u otra según el tipo de léxico.

NRC Emotion Lexicon

NRC Word-Emotion Association Lexicon (originalmente llamado EmoLex y actualmente llamado NRC Emotion Lex) es el primer léxico usado, este léxico fue desarrollado en el año 2010 por Saif Mohammad y Peter Turney. Su función consiste en clasificar las palabras según sus emociones, es decir, analiza cada una de las palabras del texto contenido en los tuits, y las asocia, con dos sentimientos, positivo y negativo, y con una de sus 8 emociones siguientes: ira, confianza, alegría, sorpresa, disgusto, miedo, tristeza y anticipación.

En la figura 19 y tabla 7 se puede ver más información sobre esté léxico y una muestra de su diccionario.

Léxico de la Asociación	Versión	# de términos	Categorías	Puntuaciones de asociación	Método de creación	Documentos
Léxico de asociación palabra-emoción y palabra-sentimiento						
Léxico de NRC Word-Emotion Association (también llamado EmoLex) README	0,92 (2010)	14,182 unigramos (palabras) ~ 25.000 sentidos *	**sentimientos: emociones** negativas, positivas : ira, anticipación, disgusto, miedo, alegría, tristeza, sorpresa, confianza	0 (no asociado) o 1 (asociado) no asociada, débil, moderada o fuertemente asociada	**Manual:** mediante crowdsourcing en Mechanical Turk. **Dominio:** General	**Crowdsourcing a Word-Emotion Association Lexicon** , Saif Mohammad y Peter Turney, *Computational Intelligence,* 29 (3), 436-465, 2013. Documento (pdf) **Emociones de BibTeX evocadas por palabras y frases comunes: uso de Mechanical Turk para crear un léxico de emociones** , Saif Mohammad y Peter Turney , en *actas del taller NAACL-HLT 2010 sobre enfoques computacionales para el análisis y la generación de emociones en texto,* junio de 2010, LA, California. Presentación del trabajo de resumen (pdf)

Figura 19: Detalles resumidos del Léxico de la emoción de NRC

Fuente: saifmohammad

English (en)	Positive	Negative	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
aback	0	0	0	0	0	0	0	0	0	0
abacus	0	0	0	0	0	0	0	0	0	1
abandon	0	1	0	0	0	1	0	1	0	0
abandoned	0	1	1	0	0	1	0	1	0	0
abandonment	0	1	1	0	0	1	0	1	1	0
abate	0	0	0	0	0	0	0	0	0	0
abatement	0	0	0	0	0	0	0	0	0	0
abba	1	0	0	0	0	0	0	0	0	0
abbot	0	0	0	0	0	0	0	0	0	1
abbreviate	0	0	0	0	0	0	0	0	0	0

Tabla 8: Muestra del NRC Emotion Lexicon, 10 primeros ejemplos.

Fuente: Elaboración propia

Bing Lexicon

El léxico Bing que lleva el nombre de su creador Bing Liu, es un léxico que contiene una lista de adjetivos añadidos manualmente a los cuáles se les agrupa según si son positivos o negativos. Este léxico contiene un total de 6786 palabras, de las cuales 2005 son positivas y 4781 son negativas. En la tabla 8 y 3 podemos ver una muestra de esta lista.

Word	Sentiment	Word	Sentiment
2-faces	Negative	Abound	Positive
Abnormal	Negative	Abounds	Positive
Abolish	Negative	Abundance	Positive
Abominable	Negative	Abundant	Positive
Abominably	Negative	Accessable	positive
Abominate	Negative	Accesible	Positive
Abomination	Negative	Acclaim	Positive
Abort	Negative	Acclaimed	Positive
Abrade	Negativa	Acclamation	Positive
Abrasive	Negative	Accolade	Positive

Tabla 9: Muestra léxico Bing, 10 primeros ejemplos negativos y positivos.

Fuente: Elaboración propia

Este léxico no solamente se ha utilizado para clasificar las palabras en positivo o negativo, también se ha empleado para calcular una puntuación general del tuit. Es decir, en base a la cantidad de palabras positivas y negativas que haya en el tuit, se ha establecido una puntuación para cada uno de los tuits, lo cual, permitió luego establecer una puntuación general para la aerolínea.

4.2.2 Identificación de las distintas emociones

Más allá del afecto positivo-negativo, identificar las emociones que están presentes en las opiniones que expresan los usuarios y/o clientes, permite comprender sus comportamientos y respuestas ante un determinado evento. Permite identificar relaciones ya existentes o nuevas y facilitar el trabajo de las empresas para la toma de decisiones.

A pesar de las diferencias culturales se ha demostrado que la mayoría de las reglas o normas afectivas son estables en todos los idiomas. El Léxico NRC fue el primer léxico en establecer

relaciones palabra-emoción y presenta versiones para más de 40 idiomas. Sobre la versión en idioma inglés, se hace uso de este léxico para clasificar las palabras contenidas en el texto de los tuits; que connotan emociones y así obtener una medida cuantificada de cuanto está contribuyendo cada emoción a la polaridad de ese tuit.

NRC clasifica las palabras en 8 emociones básicas: enfado, disgusto, anticipación, miedo, alegría, tristeza, sorpresa, confianza, así como sentimientos positivo y negativo. Para clasificar las palabras se toma como entrada a un vector de texto que contiene las palabras de los tuits. Esta función devuelve un dataframe donde cada columna representa cada una de las 8 emociones, así como el puntaje positivo y negativo; y cada fila representa un tuit.

tuit_sentiment <- get_nrc_sentiment((tuit_text))

Básicamente se lee como: función get_nrc sentiment itera sobre el vector de texto *tuit_text*, y crea un dataframe con las emociones y sentimientos presentes en *tuit_text* para cada tuit.

Las gráficas a continuación brindan una representación visual de cuál es la ponderación de cada una de estas emociones en los mensajes analizados para cada aerolínea.

KLM

Figura 20: Emociones de los mensajes (KLM)

Fuente: Elaboración propia

Ryanair

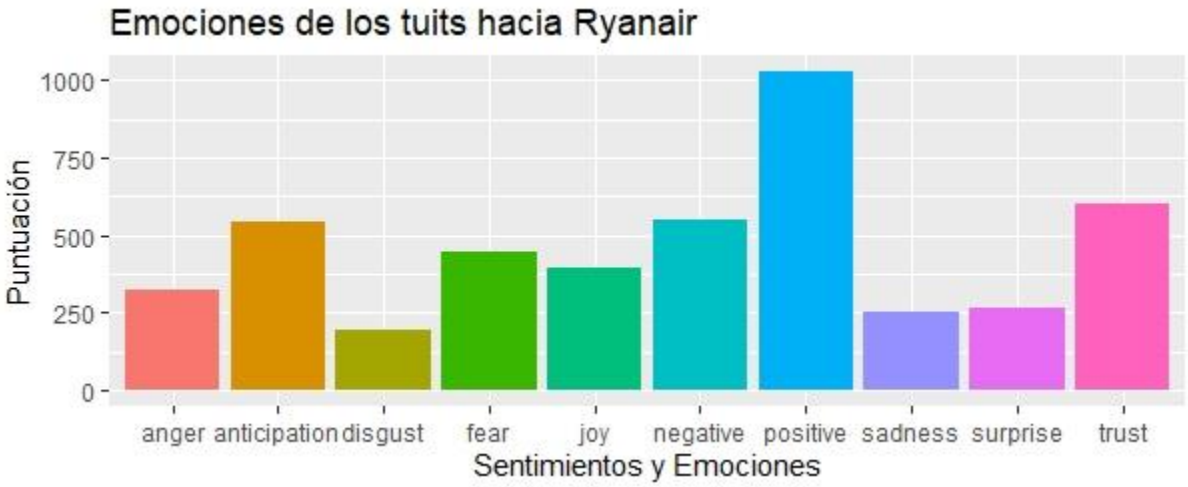

Figura 21: Emociones de los mensajes (Ryanair)

Fuente: Elaboración propia

Lufthansa

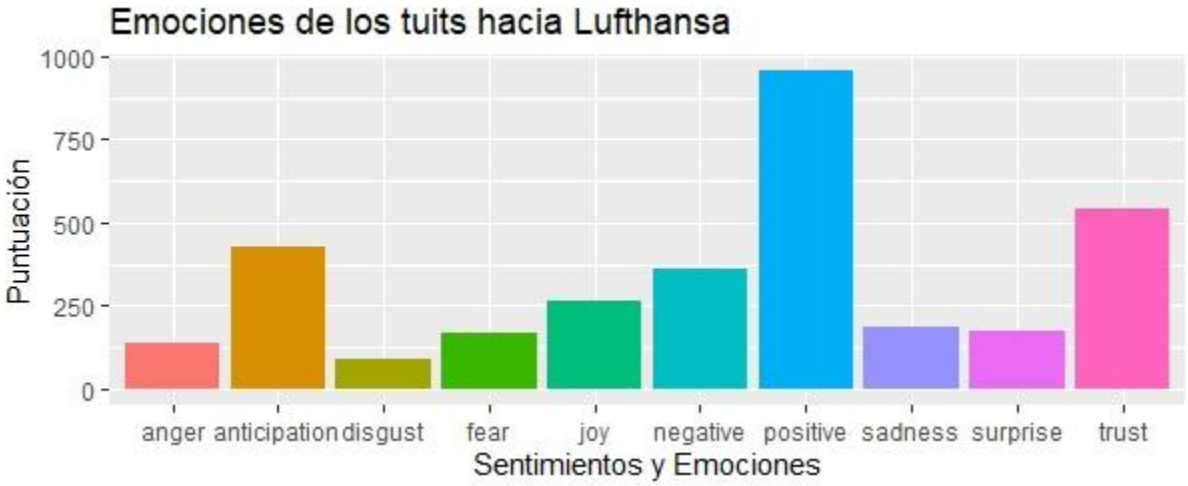

Figura 22: Emociones de los mensajes (Lufthansa)

Fuente: Elaboración propia

British Airways

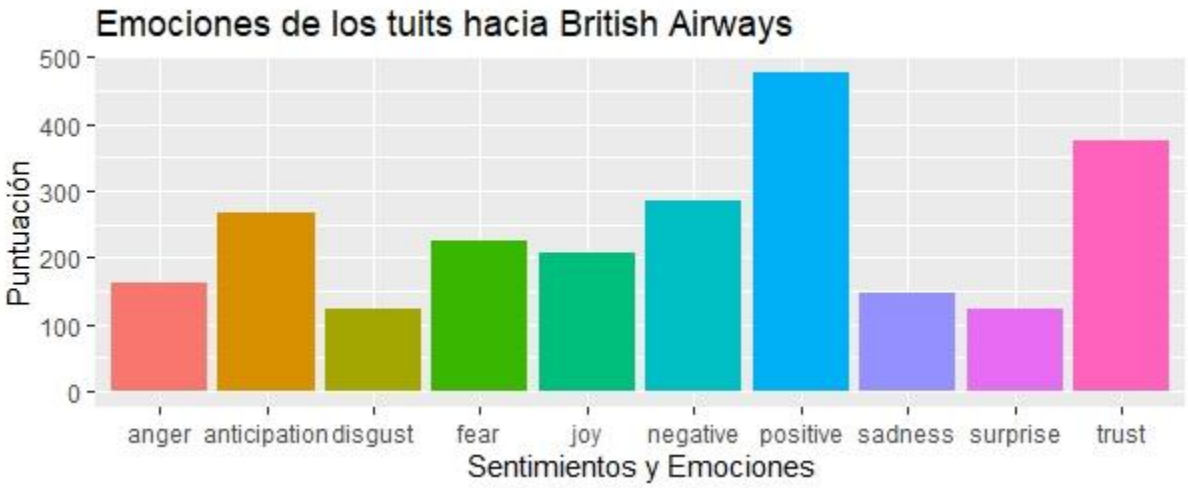

Figura 23: Emociones de los mensajes (British Airways)

Fuente: Elaboración propia

Easyjet

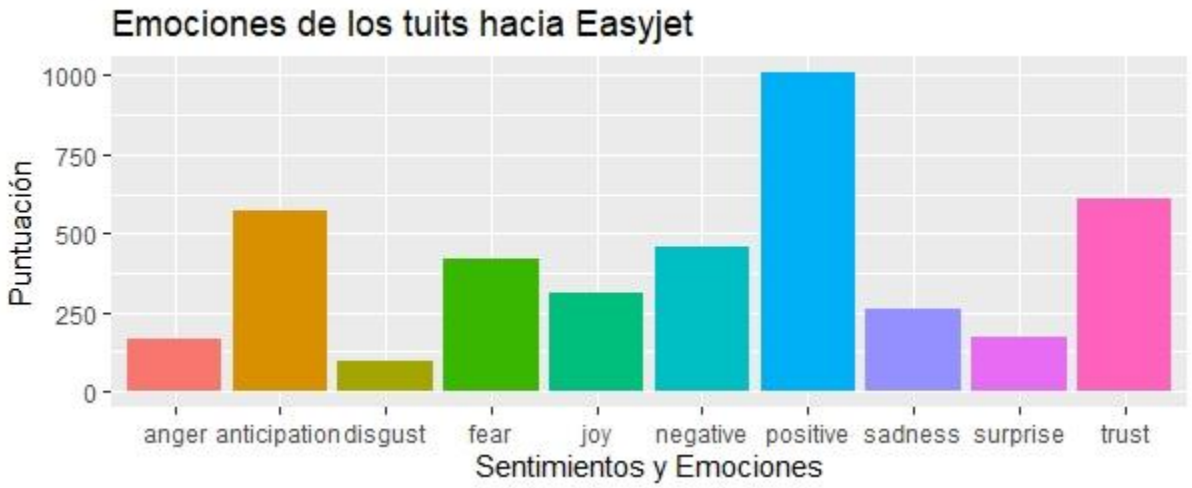

Figura 24: Emociones de los mensajes (Easyjet)

Fuente: Elaboración propia

A partir de las puntuaciones obtenidas, se observa que en todas las aerolíneas predominan dos emociones que son anticipación y confianza y el sentimiento positivo. Esto tiene toda la lógica si se analiza la situación desde diferentes perspectivas.

Si bien es cierto que muchos han sido los clientes afectados por vuelos cancelados, también es verdad que durante los meses posteriores luego de haber iniciado la pandemia que se vive durante el año 2020, las aerolíneas se han encargado de concientizar a sus usuarios sobre las consecuencias de la situación generada. Han cambiado fechas de vuelos, lanzado cupones de descuentos, realizado reembolsos pero también lanzando ofertas. Informando en todo momento de las medidas que tomarían para afectar lo menos posible a sus clientes y demostrar cómo empresas su capacidad de ser resolutivas. Lo que trajo consigo sentimientos de confianza para muchos, y opiniones en su mayoría positivas.

Por otra parte, el no saber cuándo va a terminar la pandemia y desconocer a ciencia cierta cuáles serán los resultados económicos y sociales de haber vivido bajo esta situación durante casi todo un año ha generado mucha incertidumbre a la hora de tomar decisiones. Sobre el sector aeronáutico, los medios de comunicación han generado mucha información y desinformación; noticias a cada momento de cada decisión que iban tomando las aerolíneas para enfrentarse a la Covid19. Esto ha generado que predominen en este caso de estudio; los tuits informativos y de noticias sobre las aerolíneas tratadas y por ende comentarios positivos y/o neutrales la gran mayoría y pocos negativos para ese entonces.

El resultado que experimentan de toda esta situación los usuarios más que opiniones negativas, son opiniones que expresan miedo en su momento a no poder volver a viajar pronto, anticipación ante medidas que fueran a tomar las aerolíneas principalmente con los reembolsos de los vuelos cancelados, desesperación por demora. Lo cual por parte de los clientes se entiende dado que la gran mayoría de estas medidas son impuestas a las aerolíneas bajo una necesidad del país y no medidas tomadas por decisión propia.

4.2.3 Palabras positivas y negativas más comunes

Otra de las aplicaciones ventajosas que podemos obtener de usar las funcionalidades del léxico Bing es hacer un recuento de las palabras en este caso, positivas y negativas según el marco de datos entre sentimiento-palabra que contiene el léxico.

Este análisis resulta de mucha utilidad porque ayuda a crear una percepción global de la polaridad de las opiniones que rodean a una determinada aerolínea y permite a la vez identificar indicios de hacia dónde converge, quizá, la reputación general de dichas aerolíneas.

British Airways

Palabras únicas negativas	Palabras únicas positivas
116	98

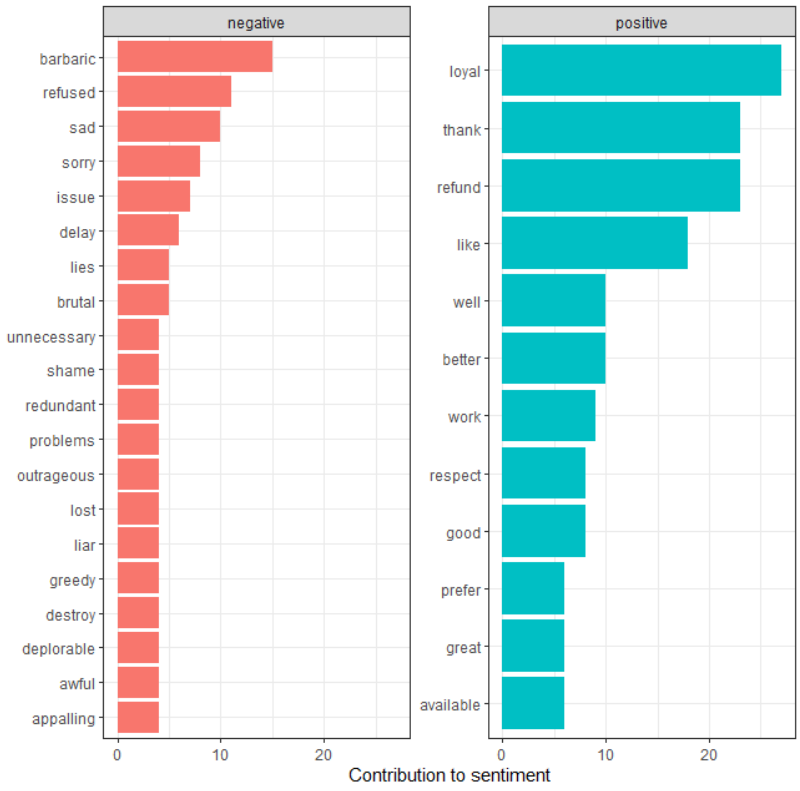

Figura 25: Palabras positivas y negativas (Easyjet)

Fuente: Elaboración propia

Se tomará como ejemplo la aerolínea British Airways. La figura 25, tenemos primero una tabla que nos muestra la cantidad de palabras únicas tanto negativas como positivas recogidas en el total de tuits analizados hasta el momento. El gráfico de esta misma figura nos muestra las palabras negativas y positivas más utilizadas para referirse a esa aerolínea, así como su frecuencia de aparición en los tuits.

Cuando se analiza la información que brinda el gráfico vemos que en una escala de 0 a 20, los términos positivos tienen mayor frecuencia de uso respecto a los términos negativos. Al contrastar esta información con los datos de la tabla puede confundir esta conclusión dado que existen más palabras negativas que positivas.

Pero esto solo significa que, aunque British Airways cuenta con más términos negativos únicos utilizados por los usuarios en los tuits, los términos positivos se utilizan con mayor frecuencia al momento de los usuarios expresar sus opiniones.

Se observa también que en el gráfico existen más términos negativos que positivos lo cual tiene sentido según la información que brinda la tabla. Se observa que los tres términos positivos más utilizados tienen mayor frecuencia de uso en la escala por encima de 20, que los tres términos negativos más utilizados donde estos se encuentran por debajo de 20 más cercanos a 10.

Un comportamiento similar se observa en las figuras y tablas a continuación que recogen los datos del resto de las aerolíneas analizadas en este estudio.

Easyjet

Palabras únicas negativas	Palabras únicas positivas
171	100

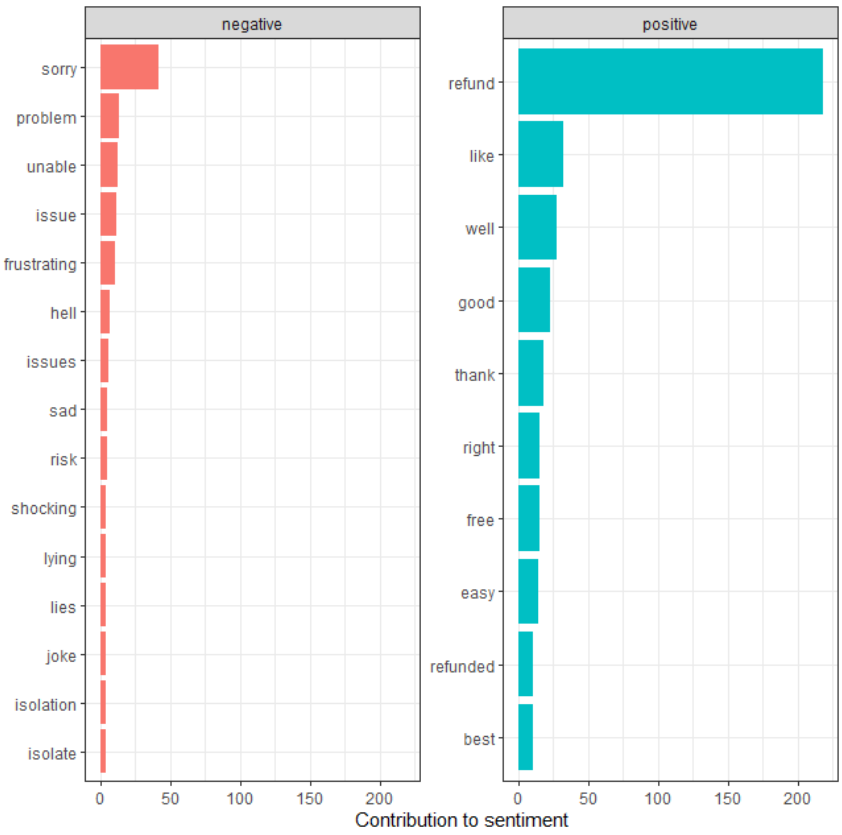

Figura 26: Palabras positivas y negativas (Easyjet)

Fuente: Elaboración propia

KLM

Palabras únicas negativas	Palabras únicas positivas
202	215

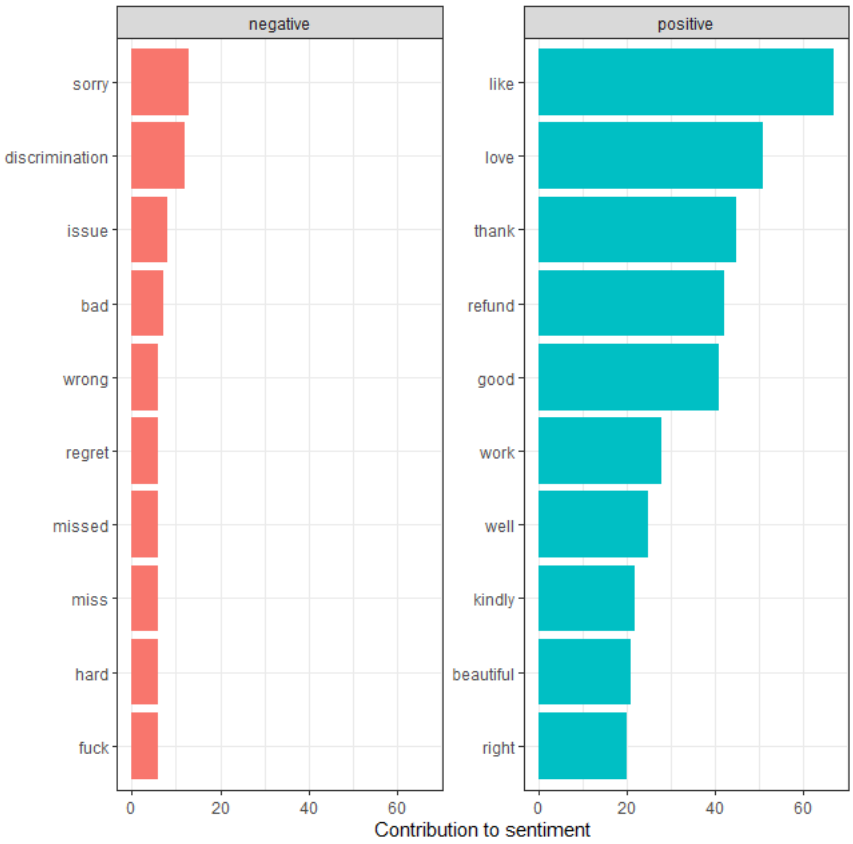

Figura 27: Palabras positivas y negativas (KLM)

Fuente: Elaboración propia

Lufthansa

Palabras únicas negativas	Palabras únicas positivas
157	128

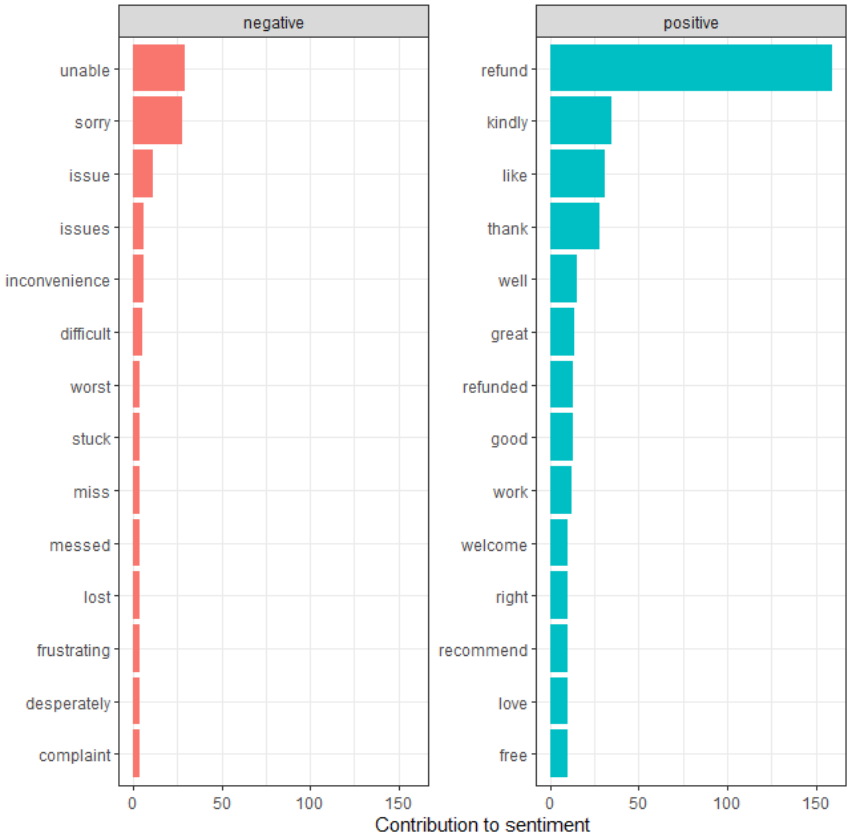

Figura 28: Palabras positivas y negativas (Lufthansa)

Fuente: Elaboración propia

Ryanair

Palabras únicas negativas	Palabras únicas positivas
225	134

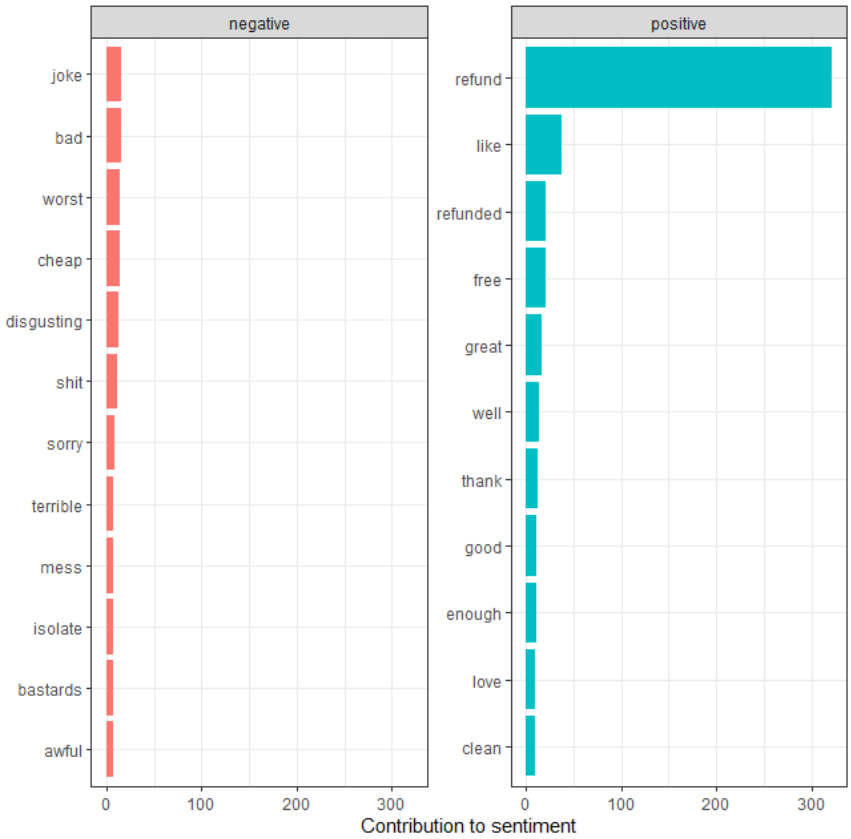

Figura 29: Palabras positivas y negativas (Ryanair)

Fuente: Elaboración propia

4.2.4 Obtención de la puntuación

Para obtener la puntuación del tuit se puede hacer uso directamente de otro léxico, concretamente, del léxico *AFINN*, ya que este léxico asigna directamente una puntuación positiva o negativa en función del sentimiento de dicha palabra en el léxico. No obstante, como se ha hecho uso del léxico Bing para agrupar cada una de las palabras según su polaridad, se ha decidido utilizar este mismo léxico para generar "manualmente" la puntuación de cada uno de los tuits que nos permitirá luego establecer la puntuación general de cada una de las aerolíneas y conocer así la reputación de estas.

4.2.4.1 Puntuación de cada uno de los tuits

Para obtener la puntuación de cada uno de los tuits, se ha programado un código en R para que analice cada una de las palabras que contiene el tuit, y según su polaridad dentro del léxico Bing, se le otorgará una puntuación en base a la cantidad de palabras positivas y negativas que este contenga.

Más detalladamente, se crea una función que genera una *tibble*[1] a partir de este léxico y añade una columna nueva llamada "score", esta columna tendrá un valor "-1" si la palabra viene recogida en el léxico como negativa, o un valor de "+1" si se trata de una palabra positiva.

Tras crear esta nueva columna, la función analiza cada una de las palabras que contiene el tuit y asigna un valor global para dicho tuit que se calcula de la siguiente manera:

- +1 punto para cada palabra positiva del tuit
- -1 punto para cada palabra negativa del tuit
- 0 puntos si la palabra no se encuentra en el léxico Bing
- 0 puntos si el tuit no contiene palabras

Se suma el resultado total y se obtiene la puntuación de cada uno de los tuits. La polaridad de cada tuit, teniendo en cuenta el resultado, se puede clasificar generalmente de la siguiente forma:

- **Negativa**: puntuación < 0
- **Neutral**: puntuación = 0
- **Positiva**: puntuación > 0

Para facilitar la interpretación, el código nos devuelve un histograma que recoge agrupados cada uno de los tuits en base a su puntuación. A continuación, podemos ver los resultados para cada una de las 5 aerolíneas.

[1] Simple data frames (no cambian ni nombres de variables ni tipos y no hacen emparejamiento parcial) https://tibble.tidyverse.org/

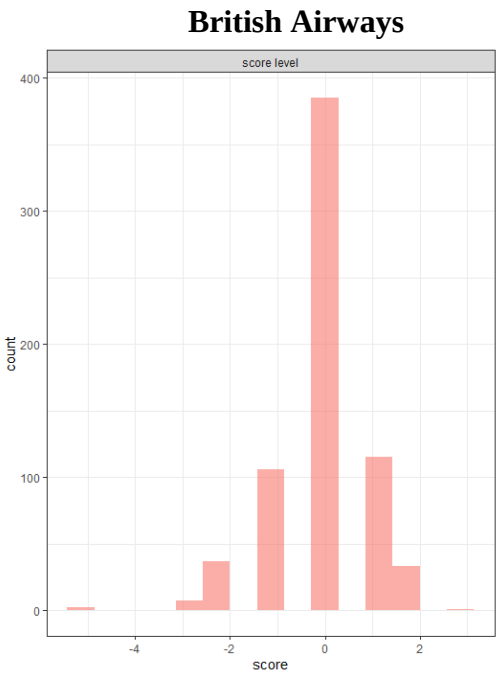

Figura 30: Puntuación de los tuits de British Airways

Fuente: Elaboración propia

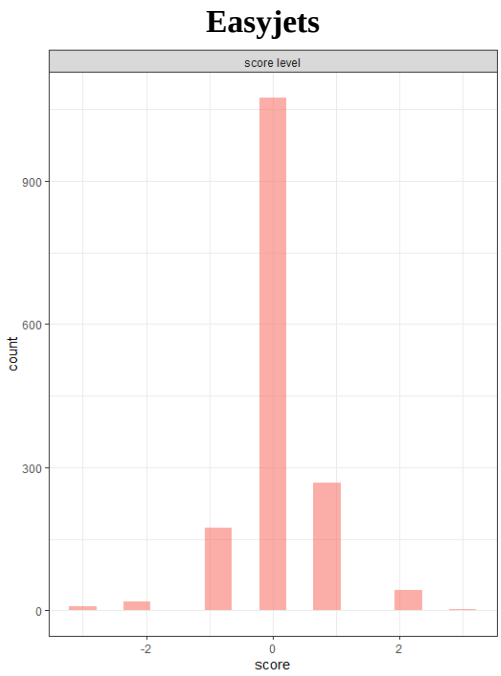

Figura 31: Puntuación de los tuits de Easyjet

Fuente: Elaboración propia

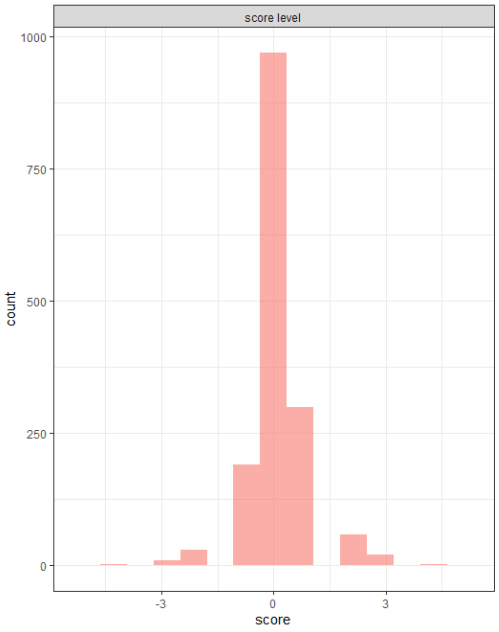

Figura 32: Puntuación de los tuits de KLM

Fuente: Elaboración propia

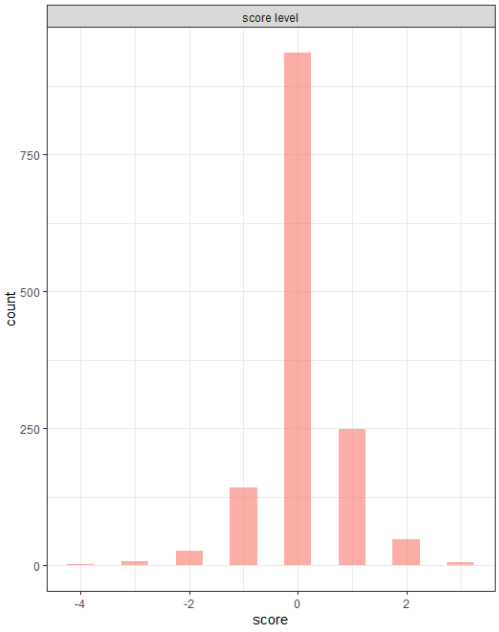

Figura 33: Puntuación de los tuits de Lufthansa

Fuente: Elaboración propia

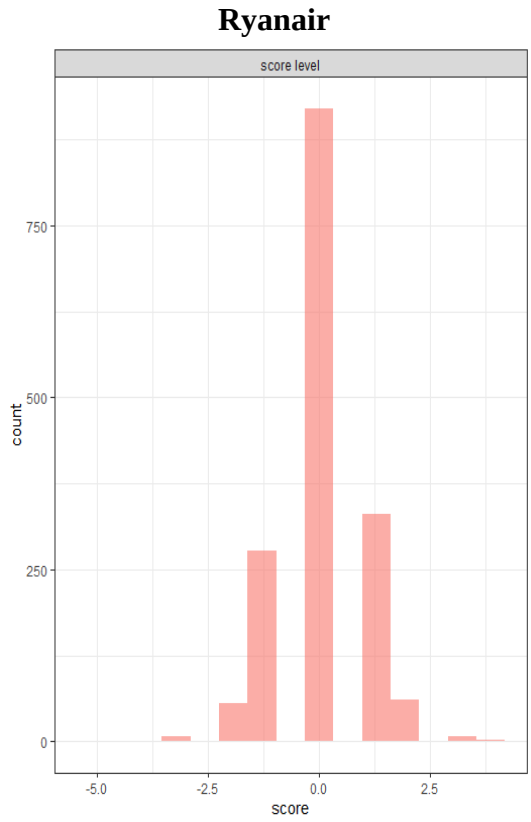

Figura 34: Puntuación de los tuits de Ryanair

Fuente: Elaboración propia

En todas las aerolíneas predominan los tuits con valor neutro, es decir, la mayoría de los tuits tiene una puntuación de "0", además, todas ellas tienen una distribución más o menos parecida, los tuits neutros predominan formando un gran "muro" que separa los tuits con polaridad positiva de los de polaridad negativa, ambos lados de este "muro", a simple vista, parece que tienen una cantidad más o menos similar de tuits por lo que cabría esperar que en la puntuación que obtendrán las aerolíneas serán mas o menos parecida, ya que todas siguen una distribución semejante.

Si agrupamos los tuits en tres clases, neutro (=0), positivos (>0) y negativos (<0), la distribución de cada una de las aerolíneas sería las siguientes.

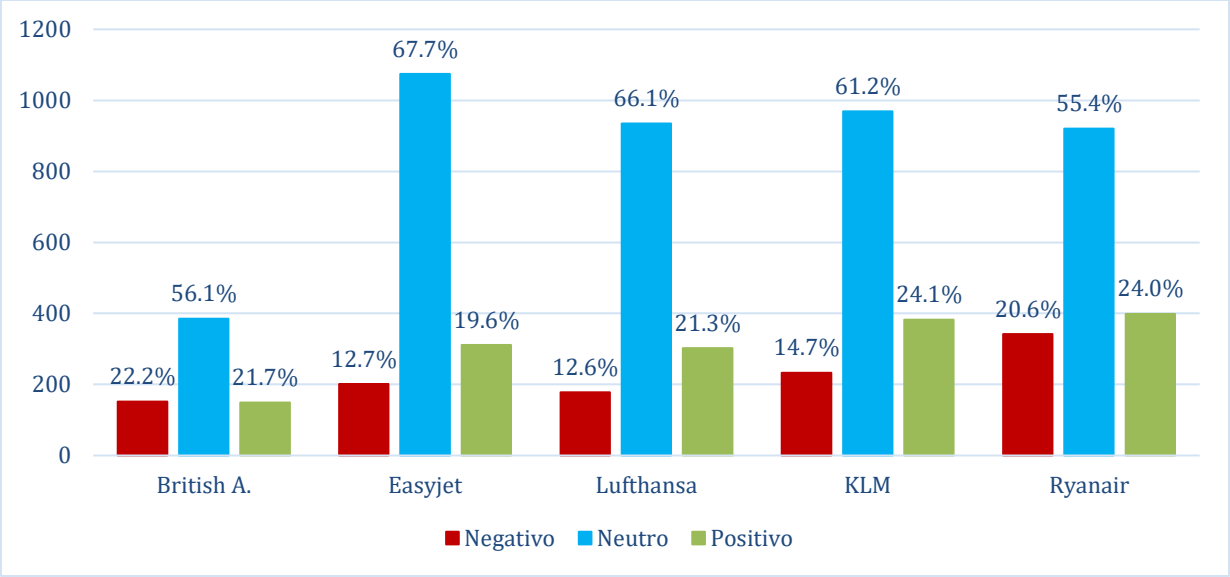

Figura 35: Polaridad de los tuits en 3 clases por aerolínea en Julio

Fuente: Elaboración propia

Sin embargo, al conocer la puntuación de cada uno de los tuits, esto nos ofrece la oportunidad de agruparlos de diferentes maneras, en las figuras anteriores se ha agrupado según la polaridad, positivo, negativo o neutro. No obstante, podemos realizar más agrupaciones, a continuación, se ha realizado una agrupación de los tuits en 5 clases:

1. **Muy negativo**: Todos aquellos tuits con puntuación inferior a -2
2. **Negativo**: Todos aquellos tuits con puntuación entre [-2,-1]
3. **Neutro**: Todos aquellos tuits con puntuación igual a 0
4. **Positivo**: Todos aquellos tuits con puntuación entre [1,2]
5. **Muy positivo**: Todos aquellos tuits con puntuación superior a 2

Los resultados son los siguientes.

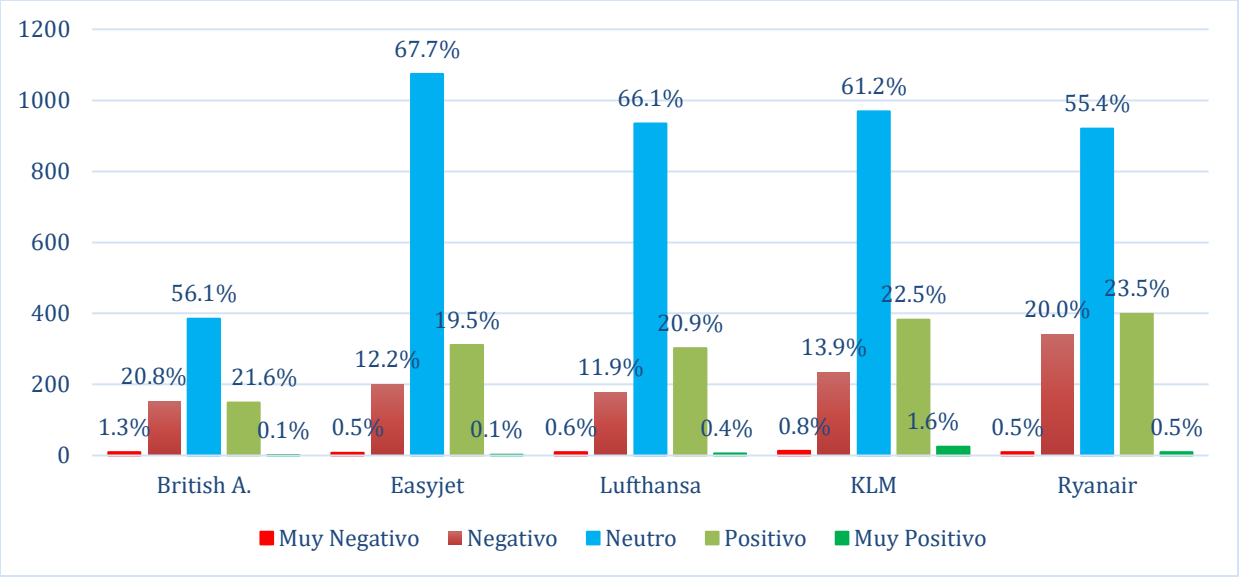

Figura 36: Polaridad de los tuits en 5 clases por aerolínea en Julio

Fuente: Elaboración propia

Como se puede ver, son muy pocos los tuits que se sitúan en los extremos (muy positivo y muy negativo) y es curioso que Easyjet y Lufthansa presentan una distribución muy parecida, esto se reflejará luego en la puntuación general en la que se verá que las dos tendrán una muy similar. También se puede ver en estos gráficos, que la aerolínea que peor puntuación presentará será British Airways, seguido muy de cerca por Ryanair, ya que ambas cuentan con una mayor cantidad de tuits negativos, es de esperar, a priori, que British Airways sea la única aerolínea que cuente con una puntuación ligeramente negativa.

Más del 55% de los tuits de todas las aerolíneas tienen una polaridad neutra, y los tuis negativos apenas llegan al 15%, exceptuando British Airways y Ryanair, las cuales superan ligeramente el 20%, teniendo la primera una mayor proporción de tuits negativos que positivos. Por su parte, Easyjet, KLM y Lufthansa presentan el caso contrario, la cantidad de tuits de polaridad neutra no solamente es más superior que las de las otras dos aerolíneas, sino que, además, reciben una mayor cantidad de tuits positivos que negativos.

Es curioso, que tanto en el caso de Ryanair como el de British Airways, cuando se clasificó las palabras según su polaridad, vimos que se producía una mayor repetición de palabras positivas, sin embargo, la cantidad de palabras negativas es mayor, es decir, los usuarios tienden a usar palabras positivas similares, cómo “refund” por ejemplo, pero usan una mayor cantidad de

palabras negativas diferentes, esto se pudo ver claramente en las tablas de palabras positivas y negativas únicas hacia cada aerolínea. Es por eso que, en resultados generales, estas dos aerolíneas cuentan con una cantidad mayor de tuits negativos.

Algo similar ocurre con Easyjet y Lufthansa, las cuales también tienen más palabras negativas únicas que positivas, no obstante, las palabras positivas se repiten con mucha más frecuencia que cantidad de palabras negativas diferentes se emplean, por ello se produce lo contrario que en el caso anterior y estas aerolíneas presentan una mayor cantidad de tuits positivos que negativos.

KLM, por su parte, es la única de las 5 aerolíneas que no solamente presenta una mayor proporción de palabras similares positivas, sino que también las palabras positivas únicas son ligeramente superiores a las negativas, por lo que es más evidente que, por lo general, tenga una mayor proporción de tuits positivos que negativos.

Aunque ya podemos hacernos una idea de la reputación que tendrá cada una de las aerolíneas a partir de la observación de estos gráficos, vamos a confirmarla y a establecer un número concreto que mida dicha reputación.

4.2.4.2 Puntuación general (reputación)

Para el cálculo de la reputación general, se ha empleado una simple fórmula en la que se ha sumado la multiplicación de cada tuit por su respectiva puntuación en base al léxico Bing, es decir, se ha sumado la polaridad de cada tuit, y el resultado se ha dividido por el número total de tuits de la aerolínea en cuestión.

Esto nos da como resultado, una puntuación en una escala de máximo [-5,5], siendo 0 una puntuación neutra. En las siguientes tablas tenemos las puntuaciones y cantidad de tuits de cada una de las aerolíneas.

$$\text{Puntuación} = \frac{\sum \text{polaridad del tuit}}{\text{Total de tuits de la aerolínea}}$$

British Airways

Puntuación	Cantidad
-5	2
-4	0
-3	7
-2	37
-1	106
0	385
1	115
2	33
3	1

Tuits = 686

$\sum$ **polaridad del tuit** = (2*-5) + (0*-4) + (7*-3) + (37*-2) + (106*-1) + (385*0) + (115*1) + (33*2) + (1*3) = **-27**

Puntuación = - 27 / 686 = - 0.0393

Tabla 10: Puntuación de cada uno de los tuits (British Airways)

Fuente: Elaboración propia

Easyjet

Puntuación	Cantidad
-3	8
-2	19
-1	174
0	1074
1	267
2	42
3	2

Tuits = 1586

∑ polaridad del tuit = (8*-3) + (19*-2) + (174*-1) + (1074*0) + (267*1) + (42*2) + (2*3) = **121**

Puntuación = 121 / 1586 = 0.0762

Tabla 11: Puntuación de cada uno de los tuits (Easyjet)

Fuente: Elaboración propia

KLM

Puntuación	Cantidad
-5	1
-4	3
-3	9
-2	30
-1	190
0	969
1	299
2	58
3	21
4	3
5	1

Tuits = 1584

∑ polaridad del tuit = (1*-5) + (3*-4) + (9*-3) + (30*-2) + (190*-1) + (969*0) + (299*1) + (58*2) + (21*3) + (3*4) +(1*5)= **196**

Puntuación = 196 / 1584 = 0.1237

Tabla 12: Puntuación de cada uno de los tuits (KLM)

Fuente: Elaboración propia

Lufthansa

Puntuación	Cantidad
-4	2
-3	7
-2	27
-1	142
0	935
1	249
2	47
3	6
4	0

Tuits = 1415

$\sum$ **polaridad del tuit** = (2*-4) + (7*-3) + (27*-2) + (142*-1) + (935*0) + (249*1) + (47*2) + (6*3) = **136**

Puntuación = 136 / 2000= 0.0961

Tabla 13: Puntuación de cada uno de los tuits (Lufthansa)

Fuente: Elaboración propia

Ryanair

Puntuación	Cantidad
-5	1
-4	0
-3	8
-2	56
-1	277
0	920
1	330
2	60
3	7
4	2

Tuits = 1661

$\sum$ **polaridad del tuit** = (1*-5) + (0*-4) + (8*-3) + (56*-2) + (277*-1) + (920*0) + (330*1) + (60*2) + (7*3) + (2*4) = **61**

Puntuación = 61 / 1661= 0.0367

Tabla 14: Puntuación de cada uno de los tuits (Ryanair)

Fuente: Elaboración propia

Lista ordenada de mayor a menor:

- **KLM, 0.1237**
- **Lufthansa, 0.0961**
- **Easyjet, 0.0762**
- **Ryanair, 0.0367**
- **British Airways, - 0.0393**

Los resultados es la materialización de lo que ya nos adelantaban los gráficos anteriores. British Airways era la única que recibía una mayor cantidad de tuits negativos que positivos, no obstante, la proporción de tuits negativos no se aleja mucho de los positivos (22.2% frente a 21.7%) por lo que esto hace que se compensé mucho y su reputación apena este unas décimas por debajo de cero.

Caso parecido ocurre con Ryanair, no obstante, en este son los positivos los que difieren un poco por encima de los negativos (24% frente a 20.6%), por lo que ocurre algo similar que British Airways, pero con símbolo opuesto.

Easyjet y Lufthansa presentan resultados muy similares, la polaridad de sus tuits apenas se aleja varios puntos porcentuales de una aerolínea a otra, siguiendo el orden de; negativo, neutro, positvo; para Easyjet es de 12.7%, 67.7%, 19.6% frente a Lufthansa de 12.6%, 66.1%, 21.3%. La poca diferencia que hace que Lufthansa tenga una puntuación ligeramente superior a Easyjet, es que la primera cuenta con una mayor cantidad de tuits "muy positivos" (0.4% frente a 0.1%) y con menos palabras únicas negativas y más palabras únicas positivas.

La polaridad de KLM en cuanto a positivos y negativos es la que más difiere, hay casi 10 puntos porcentuales (9.4) que separan los tuits positivos de los negativos, esto hace que sea la aerolínea que presenta una mayor puntuación, lo cual es lógico siguiendo lo que nos mostraba la frecuencia de palabras, ya que es la única que cuenta no solamente con una mayor frecuencia de palabras positivas repetidas que negativas, sino también con una mayor cantidad de palabras únicas positivas frente a las negativas.

5 Comparación de resultados

Una vez se ha establecido la reputación de las aerolíneas para la última semana de julio, se ha vuelto a realizar el mismo análisis para la primera semana de septiembre, y ver la variación en los datos que han experimentado las aerolíneas tras el periodo del mes de agosto. Dado que el objetivo específico de este trabajo es establecer la reputación actual (Julio de 2020) de cada aerolínea, la cual se realizó en el apartado anterior, y estudiar la variación de dicha reputación en un periodo de un mes, no se va a proceder a estudiar la variación total (emociones, nubes de palabras, etc) de cada una de las aerolíneas, sino que se mostrará únicamente el caso más relevante. Si se mostrará en su totalidad, como se ha indicado en los objetivos, la variación de esta reputación.

5.1 Variación de la reputación

El procedimiento es el mismo que en el apartado anterior. Una vez extraídos los tuits y realizado el mismo proceso, obtenemos la siguiente puntuación. Recordemos la fórmula empleada.

$$\text{Puntuación} = \frac{\sum \text{polaridad del tuit}}{\text{Total de tuits de la aerolínea}}$$

British Airways

Puntuación	Cantidad
-3	10
-2	45
-1	173
0	724
1	244
2	51
3	4

Tuits = 1251

$\sum$ **polaridad del tuit** = (10*-3) + (45*-2) + (173*-1) + (724*0) + (244*1) + (51*2) + (4*3) = **65**

Puntuación Septiembre = 65 / 1251 = 0.0519 ⬆
Puntuación Julio = - 27 / 686 = - 0.0393

Tabla 15: Puntuación de cada uno de los tuits (British Airways)

Fuente: Elaboración propia

Easyjet

Puntuación	Cantidad
-3	9
-2	55
-1	226
0	954
1	293
2	43
3	5

Tuits = 1585

∑ polaridad del tuit = (9*-3) + (55*-2) + (226*-1) + (954*0) + (293*1) + (43*2) + (5*3) = **31**

Puntuación Septiembre = 31 / 1585 = 0.0195 ⬇

Puntuación Julio = 121 / 1586 = 0.0762

Tabla 16: Puntuación de cada uno de los tuits (Easyjet)

Fuente: Elaboración propia

KLM

Puntuación	Cantidad
-3	7
-2	41
-1	190
0	903
1	211
2	44
3	16
4	2

Tuits = 1414

∑ polaridad del tuit = (7*-3) + (41*-2) + (190*-1) + (903*0) + (211*1) + (44*2) + (16*3) + (2*4) = **62**

Puntuación Septiembre = 62 / 1414= 0.0438 ⬇

Puntuación Julio = 196 / 1584 = 0.1237

Tabla 17: Puntuación de cada uno de los tuits (KLM)

Fuente: Elaboración propia

Lufthansa

Puntuación	Cantidad
-3	2
-2	21
-1	152
0	816
1	167
2	37
3	7
4	3

Tuits = 1205

∑ polaridad del tuit = (2*-3) + (21*-2) + (152*-1) + (816*0) + (167*1) + (37*2) + (7*3) + (3*4) = **74**

Puntuación Septiembre = 74/ 1205 = 0.0614 ⬇

Puntuación Julio = 136 / 1415= 0.0961

Tabla 18: Puntuación de cada uno de los tuits (Lufthansa)

Fuente: Elaboración propia

Ryanair

Puntuación	Cantidad
-3	2
-2	35
-1	223
0	706
1	351
2	55
3	3
5	1

Tuits = 1376

$\sum$ **polaridad del tuit** = (2*-3) + (35*-2) + (223*-1) + (706*0) + (351*1) + (55*2) + (3*3) + (1*5) = **176**

Puntuación Septiembre = 176 / 1376 = 0.1279

Puntuación Julio = 61 / 1661 = 0.0367

Tabla 19: Puntuación de cada uno de los tuits (Ryanair)

Fuente: Elaboración propia

En la tabla 19 se recoge esta información, así como la variación porcentual de cada aerolínea. La columna de la derecha es independiente de la tabla y únicamente muestra el nuevo orden (de mayor a menor reputación) de las aerolíneas tras la puntuación de septiembre.

Aerolínea	Puntuación Julio	Puntuación Septiembre	% Variación	Variación	Nuevo Orden
KLM	0.1237	0.0438	-65%	-0.0799	**Ryanair**
Lufthansa	0.0961	0.0614	-36%	-0.0347	**Lufthansa**
Easyjet	0.0762	0.0195	-74%	-0.0567	**British A.**
Ryanair	0.0367	0.1279	249%	0.0912	**KLM**
British A.	-0.0393	0.0519	232%	0.0912	**Easyjet**

Tabla 20: Variación de la puntuación Julio – Septiembre

Fuente: Elaboración propia

Sin embargo, está tabla se puede ver mejor en la siguiente figura.

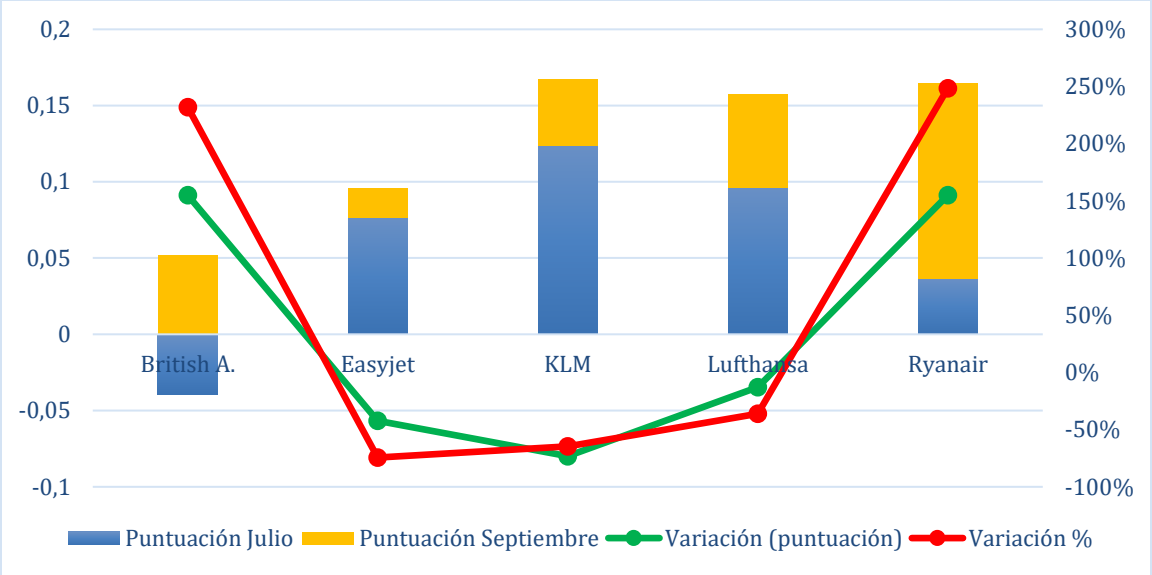

Figura 37: Variación de la reputación

Fuente: Elaboración propia.

Vemos que, aunque el periodo de la variación es relativamente corto, un mes, se han producido cambios interesantes en la puntuación y se ha alterado el ranking de Julio.

Esta variación se explica mejor con la siguiente gráfica.

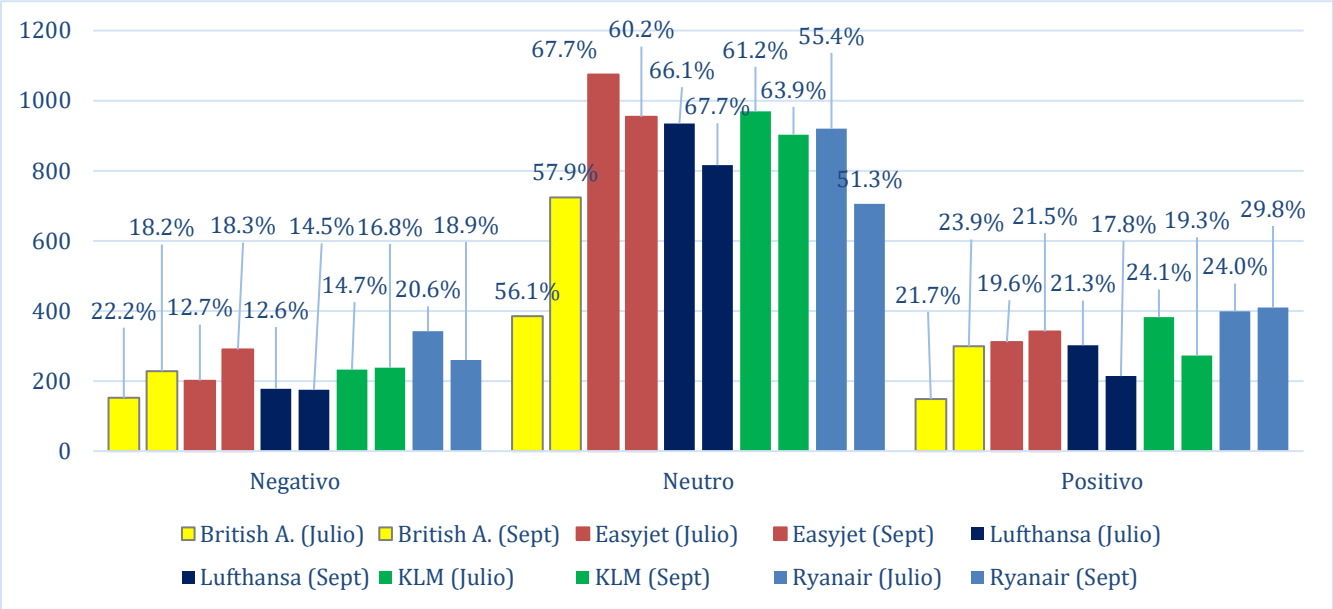

Figura 38: Variación de la reputación

Fuente: Elaboración propia.

Cada color representa a una aerolínea y la barra de la izquierda corresponde a Julio y la de la derecha a septiembre.

KLM, que era la aerolínea que tenía una mayor reputación, ha perdido un 65% (-0.0799 puntos, la que más en términos de puntuación) que la han llevado a la posición número 4 del ranking. Su caída se debe a una disminución de los tuits positivos (-4.8%) que han ido a favorecer la cantidad de tuits negativos (+2.1%) y neutros (+2.7%).

Lufthansa ha variado en un -36%, es la que menos lo ha hecho tanto en términos porcentuales como en puntuación (-0.0347). Este ligero cambio negativo se debe a lo mismo ocurrido con KLM, sólo que en menor medida. Lufthansa sufrió un descenso de sus tuits positivos (-3.6%) en favor de los neutros y los negativos (+1.9), por eso el cambio no es tan significante, de hecho, sigue conservando la misma posición en la tabla.

Easyjet sigue la misma línea que en los dos casos anteriores, y es la que más ha perdido en términos porcentuales (-74%) que se traducen en -0.0567 puntos. La principal diferencia con respecto a las dos aerolíneas anteriores, es que el cambio drástico de Easyjet viene dado en la variación de sus tuits neutros, los cuales han disminuido en 7.5 puntos porcentuales que han ido a parar, en su mayoría, a los tuits negativos (5.6%).

Ryanair y British Airways son los casos más interesantes, ya que han experimentado la mayor variación en cuanto a porcentaje, 249% y 232% respectivamente, que se traduce, exactamente, en la misma cantidad en cuanto a puntuación para ambas aerolíneas, 0.0912.

British Airways ha disminuido sus tuits negativos en un 3.93%, que han ido a parar en 1.75% a los neutros y 2.18% a los positivos. Ryanair, por su parte, ha visto como ha disminuido tanto sus tuits negativos (-1.69%) y sus neutros (-4.08%), disminución en favor de los tuits con polaridad positiva (+5.77%). Este gran cambio ha llevado a Ryanair a la primera posición del ranking, y ha cambiado de signo la reputación de British Airways, que ahora ocupa la posición número tres justo por detrás de Lufthansa.

En definitiva, las variaciones no han sido muy grandes, pero si hemos visto que se han producido tanto variaciones positivas como negativas en un periodo de 1 mes, y dado los resultados, se

puede decir que Ryanair y British Airways, de momento, son las que están gestionando mejor, la atención a sus usuarios.

5.2 Variaciones interesantes

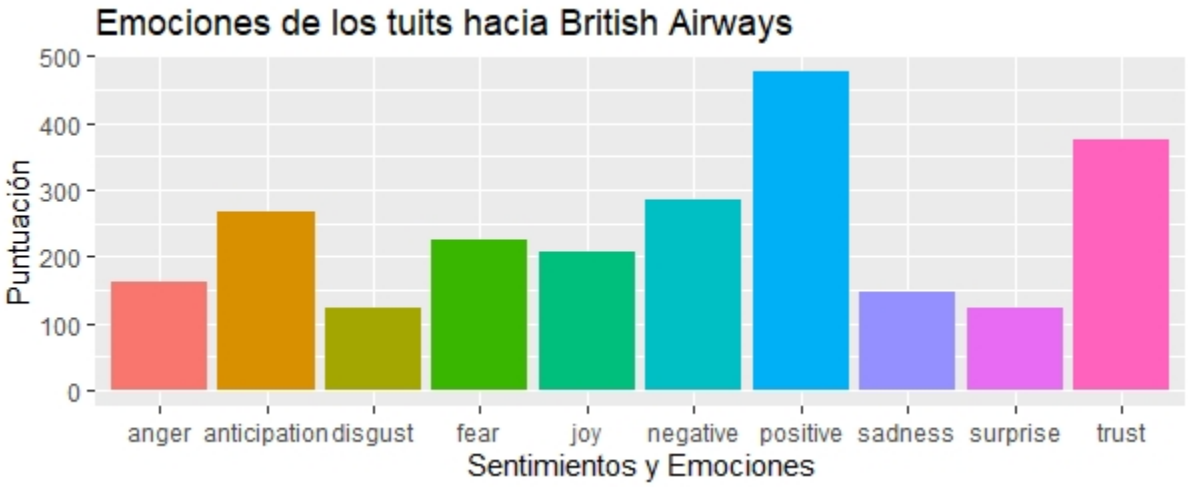

Figura 39: Emociones de los tuits hacia British Airways (Julio)

Fuente: Elaboración propia

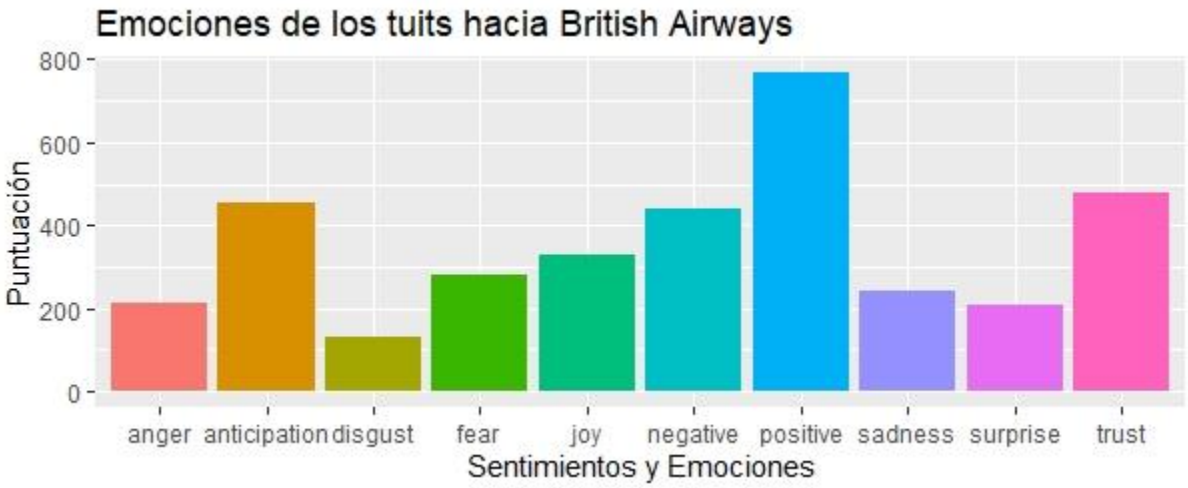

Figura 40: Emociones de los tuits hacia British Airways (Septiembre)

Fuente: Elaboración propia

Siguiendo la línea de análisis de la variación de la reputación en el apartado anterior, la aerolínea British Airways fue una de las más favorecidas. Tomándola de ejemplo observemos como han cambiado los datos en septiembre para esta aerolínea. Recordemos que la figura 39 da una

muestra cuantificada de las emociones y sentimientos que expresan los términos contenidos en el texto de los tuits para la última semana de Julio. La figura 40 muestra la misma información para la primera semana de septiembre, es decir, un mes después. Se ve como aumento considerablemente el número de tuits que expresaban sentimiento positivo, pasando de 469 que equivale a un (19.9%) respecto a un total de 2362 emociones contenidas en los tuits a 768 (21.7%) respecto a un total de 3533. Es decir, aumentó el número de tuits positivos, y aumento su proporción total. Esto significa que los tuits que expresaban emociones aumentaron en aproximadamente un 49.57 %.

Si observamos las nubes de palabras para esta aerolínea de julio y septiembre respectivamente, vemos que han cambiado de forma general. Términos como refund, waiting, cancelled, ganaron relevancia y aumentaron su frecuencia de aparición respecto a los datos de julio.

Figura 41: Nube de palabras British Airways (Julio)

Fuente: Elaboración propia

Figura 42: Nube de palabras British Airways (Septiembre)

Fuente: Elaboración propia

Aquí ocurre un contraste de los datos bastante interesante porque las nubes de palabras, aunque son diferentes, muestran en ambos momentos de tiempo términos diferentes pero que a priori pueden estar siendo empleados en comentarios de una misma polaridad. Es decir, tuits negativos que expresen quizás emociones de enojo, decepción, inconformidad o desespero entre otras. Pero por otro lado ocurre que el análisis de emociones en comparación al análisis de frecuencia, muestra resultados hasta cierto punto opuestos.

En el apartado anterior se observa que la reputación obtenida por British Airways es positiva, lo cual puede demostrar que las medidas tomadas por la aerolínea para enfrentar los efectos de la crisis y mitigar los daños a sus clientes han sido buenas. Esto se evidencia al ver como aumentó el sentimiento positivo a gran escala, aunque también el negativo, pero este último lo hizo en una proporción muy pequeña. Se redujo la emoción de ira de un 6.9% en julio a un 6% en septiembre. Disminuyo el disgusto de un 5.1% a un 3.7%. Aumento el sentimiento de sorpresa de un 5.1% a un 5.9%, aumento la emoción de alegría de un 8.6% a 9.3%.

Por lo tanto, aunque las nubes de palabras entre julio y septiembre eran diferentes en cuanto a términos, el sentido de lo que expresaban estos términos era prácticamente el mismo. A su vez el análisis de emociones demuestra que lo que se percibe de forma visual no es lo que está ocurriendo al analizar las emociones contenidas en los tuits y por ende la polaridad.

Esto demuestra que ambos análisis no deben ir por separados sino como complementos entre ellos, y así poder hacer un juicio lo más certero posible entre lo que se percibe a través de una nube de palabras y lo que se está expresando a través de los sentimientos y emociones contenidas en los tuits.

6 Conclusiones

La investigación desarrollada y los resultados obtenidos permiten a los autores arribar a las siguientes conclusiones finales:

Gracias a los tuits de los usuarios hacia las aerolíneas se han podido cumplir con los objetivos planteados en este trabajo. Todo el pre procesamiento permitió obtener datos limpios y lograr unos resultados con el menor ruido posible.

Se identificó que las palabras más comunes en todas las aerolíneas son «book, cancelled, flight, refund, travel, waiting, y voucher», es decir, el principal tema que preocupa a los usuarios es lograr una solución (reembolso o cupón) para el vuelo cancelado por motivos del covid-19.

El análisis de emociones permitió conocer los sentimientos predominantes en los tuits analizados en cada periodo de tiempo, revelando así, como se sentían los usuarios y hacia donde convergía la reputación de las aerolíneas estudiadas.

En términos generales, tanto durante el análisis de Julio como en el de Septiembre, la polaridad de la mayoría de los mensajes es neutra, en torno a un 60%, esto se debe, en parte, a que muchos de los tuits corresponden con noticias o algún tipo de publicidad.

Cabe destacar, que puede que a priori se espere que con el paso del tiempo la reputación de las aerolíneas mejore poco a poco, pero dada la situación de incertidumbre actual, y tal y como se ha visto en la reputación, en un periodo de un mes, algunas aerolíneas vieron perjudicada su reputación mientras que otras mejoraron considerablemente. No obstante, esta reputación en todas ellas es próxima a 0, es decir, es muy cercana a polaridad neutra, y en una escala [-5,5] apenas se alejan unas décimas en positivo o negativo. Esto se debe, a parte de lo comentado en el párrafo anterior, a que la proporción de tuits negativos y positivos es similar en todas las aerolíneas, por lo que finalmente se obtiene una puntuación cercana a 0. La única que se aleja un poco más es Ryanair tras los resultados de septiembre.

Sería bueno, como posible línea de investigación futura a modo de ampliación de este trabajo, incluir un análisis en tiempo real, algo que sería de gran potencial para las empresas ya que les permitiría una mejor toma de decisiones en el momento oportuno.

7 Bibliografía

Arnold, Taylor B. (2016). cleanNLP: A Tidy Data Model for Natural Language Processing, consultado en Agosto de 2020 en https://cran.r-project.org/package=cleanNLP.

Arnold, Taylor, y Tilton, L. (2016). coreNLP: Wrappers Around Stanford Corenlp Tools, consultado en Agosto de 2020 en https://cran.r-project.org/package=coreNLP.

Amat Rodrigo, J. (2019). Text mining con R: ejemplo práctico Twitter, *Rpubs by RStudio*, consultado en Agosto de 2020 en https://rpubs.com/Joaquin_AR/334526

Feinerer, I. (2019). Introduction to the tm Package Text Mining in R, consultado en Agosto de 2020 en https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

LZP, J. (2020). Sentiment Analysis With Twitter Hashtags In R. *Gitconnected*. Consultado en Agosto de 2020 en https://levelup.gitconnected.com/sentiment-analysis-with-twitter-hashtags-in-r-af02655f2113

Maqueda, A. (2020). El virus asesta a la economía española un golpe sin precedentes con una caída trimestral del 18,5%, *crisis del coronavirus*, El Pais, consultado en Agosto de 2020 en https://elpais.com/economia/2020-07-31/el-virus-asesta-a-la-economia-espanola-un-golpe-sin-precedentes-con-una-caida-trimestral-del-185.html

Rinker, Tyler W. (2017). Sentiment: Calculate Text Polarity Sentiment. Buffalo, New York: University at Buffalo/SUNY, consultado en Agosto de 2020 en http://github.com/trinker/sentimentr.

Reuters. (2020). Las aerolíneas no recuperarán el tráfico de pasajeros de antes de la covid hasta 2024, *efectos de la covid*, La Vanguardia consultado en Agosto de 2020 en https://www.lavanguardia.com/economia/20200728/482573674081/trafico-aereo-iata-recuperacion-2024.html

Saif M, y Turney, P. (2013). Crowdsourcing a Word–Emotion Association Lexicon, *Institute for Information Technology, National Research Council Canada. Ottawa, Ontario, Canada, K1A 0R6*. Consultado en Agosto de 2020 en https://arxiv.org/pdf/1308.6297.pdf

Saif M, y Turney, P. (2013). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon, *Institute for Information Technology, National Research Council Canada. Ottawa, Ontario, Canada, K1A 0R6*. Consultado en Agosto de 2020 en http://saifmohammad.com/WebDocs/Mohammad-Turney-NAACL10-EmotionWorkshop.pdf

Saif M, y Turney, P. (2010). NRC Word-Emotion Association Lexicon, consultado en Agosto de 2020 en http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Rdrr (s.f): Stopwords in tm: Text Mining Package, consultado en Agosto de 2020 en https://rdrr.io/rforge/tm/man/stopwords.html

Sipra, V. (2020). Twitter Sentiment Analysis and Visualization using R, *How to gauge what twitter users are feeling about any particular topic*, Towards data science, consultado en Agosto de 2020 en https://towardsdatascience.com/twitter-sentiment-analysis-and-visualization-using-r-22e1f70f6967

Statista. (2019). Las compañías aéreas que transportan a más pasajeros en Europa, *Statista, el portal de estadística*, consultado en Agosto de 2020 en https://es.statista.com/grafico/14762/companias-de-vuelo-en-europa-por-numero-de-pasajeros-en-2018/

STHDA (s.f). Text mining and word cloud fundamentals in R : 5 simple steps you should know. *STHDA Statistical Tools for high-throuhput data analysis*, consultado en Agosto de 2020 en http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know

Tidytext (s.f). S Sentiment analysis with tidy data. *Tidytextmining*, consultado en Agosto de 2020 en https://www.tidytextmining.com/sentiment.html

Wikipedia, J. (2020). Léxico, *Wikipedia*, la enciclopedia libre, Wikipedia Foundation, San Francisco, CA. Consultado en Agosto de 2020 en https://es.wikipedia.org/wiki/L%C3%A9xico

8 Anexo I

Código empleado para la extracción de tuits en Python

```
########################
#### AUTENTICACIÓN ####
########################
#Se importa la librería tweepy
import tweepy
import csv
#Credenciales del Twitter API
consumer_key = "98soSE1PMFUn"
consumer_secret = "xdnRBmu8KepOKT9"
access_token = "364887747-M9AllOSJK6iBLvyAk3dt4"
access_token_secret = "RIMT3GeZOMUjnihTOkKN8FVKJBsVfZ"
#Se autentica en twitter
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
```

```
api = tweepy.API(auth, wait_on_rate_limit=True, wait_on_rate_limit_notify=True,
compression=True)
######################
##### INTERFAZ ######
######################
#mensaje
print('Recuerda que el programa hará una busquedad en tweets que estén en ingles\n por lo que introduzca una palabra a buscar en este idioma')
#Se pregunta por la palabra a buscar
palabra = input("Search: ")
#Se define la cantida de tweets a capturar
numero_de_Tweets = int(input(u"Number of Tweets: "))
#Se define el idioma de los tweets a analizar
lenguaje = 'en'
######################
###### LIMPIEZA ######
######################
#Esta funcion nos permite depurar el codigo, limpiando caracteres que luego nos permitiran un mejor manejo de los datos.
#Remover los caracteres no imprimibles y los saltos de línea del texto del tweet
def strip_undesired_chars(tweet):
    stripped_tweet = tweet.replace('\n', ' ').replace('\r', '')
    char_list = [stripped_tweet[j] for j in range(len(stripped_tweet)) if ord(stripped_tweet[j]) in range(65536)]
    stripped_tweet=''
    for j in char_list:
        stripped_tweet=stripped_tweet+j
    return stripped_tweet
######################
#### EXTRACCIÓN ####
######################
def ObtenerTweets(palabra="",times=100,leguanje=""):
```

```
    alltweets = []
    print("Downloading tweets")
    print("This will take a few seconds")
    for tweet in tweepy.Cursor(api.search, palabra, lang=lenguaje).items(numero_de_Tweets):
        try:
            alltweets.append(tweet)
        except tweepy.TweepError as e:
            print(e.reason)
        except StopIteration:
            break
    #transformar los tweets descargados con tweepy en un arreglo 2D array que llenará el csv
    outtweets = [(tweet.id_str, tweet.created_at,
strip_undesired_chars(tweet.text),tweet.retweet_count,str(tweet.favorite_count)+") for tweet in
alltweets]
    #escribir el csv
    with open('%s_tweets.csv' % palabra, "w", newline='', encoding="utf-8") as f:
        writer = csv.writer(f, quoting=csv.QUOTE_ALL)
        writer.writerow(['id','created_at','text','retweet_count','favorite_count"'])
        writer.writerows(outtweets)
    pass
    return 'Se han extraido los Tweets correctamente'
ObtenerTweets(palabra,numero_de_Tweets,lenguaje)
#Breve explicación
‘’’ Desde que se ejecuta el código, se empiezan a extraer los tuits (la cantidad que se le haya
indicado) hasta una profundidad máxima de 7 días, empezando a contar desde ese momento, es
decir, si ponemos 2000 tuits, y resulta que se han generado esos 2000 tuits en 1 día, pues
obtendrá todos los tuits con esa profundidad, pero si en 7 días sólo se han generado, por ejemplo,
500, esto es lo que extraerá, 500, aunque le hayamos indicado 2000, pero como en 7 días sólo
hay 500, no nos dará más.
```

Los tuits que se van extrayendo se van guardando en una lista vacía, que se va autocompletando, el código va extrayendo tuits y cuando se topa con la restricción de twitter, se pausa, y tras la pausa, continua justo por donde se quedó y sigue añadiendo los tuits restantes a la lista. Una vez ha terminado, esa lista se exporta en un csv. ‘’’

9 Anexo II

Código empleado en R

```
#LIMPIAMOS EL ENTORNO DE R. BORRAMOS LO QUE TUVIÉSEMOS EN MEMORIA
rm(list=ls())

##NEEDED LIBRARIES
library(tm)
library(syuzhet)
library(SnowballC)
library(wordcloud)
library(wordcloud2)
library(RColorBrewer)
library(ggplot2)

library(dplyr)
library(tidytext)
library(tidyverse)
library(textdata)
library(magrittr)
library(purrr)

###### UPLOADING TWEETS
airline <- read.csv("lufthansa2.csv")
#Visualization of the csv
View(airline)
#Keeping the text and viewing itgetwd()
tweet_text <- airline$text
View(tweet_text)

####### DATA CLEANING
#Convert all text to lower case
tweet_text <- gsub("RT", "",tweet_text)
tweet_text <- tolower(tweet_text)
#Replace urls
tweet_text <- gsub("http\\S+\\s*", "",tweet_text)
#Replacing twitter @ username handle
tweet_text <- gsub("@\\w+", "", tweet_text)
#Remove punctuations
tweet_text <- gsub("[[:punct:]]", "", tweet_text)
```

```
#Remove blank spaces at the beginnng
tweet_text <- gsub("^ ", "",tweet_text)
#Remove blank spaces at the end
tweet_text <- gsub(" $", "",tweet_text)
#Remove other character
tweet_text <- gsub("â€", "",tweet_text)
tweet_text <- gsub("€", "",tweet_text)
tweet_text <- gsub("™", "",tweet_text)
tweet_text <- removeWords(tweet_text, words = stopwords("english"))
tweet_text <- stripWhitespace(tweet_text)
#Eliminamos tweets repétidos
tweet_text<-unique(tweet_text)

#resultado
#View(tweet_text)

## CREATE AND CLEAN THE CORPUS
docs <- Corpus(VectorSource(tweet_text))
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, removeWords, stopwords("en"))
docs <- tm_map(docs, removeWords, stopwords("SMART"))
# Remove other specify stopwords
docs <- tm_map(docs, removeWords,c("due", "dont", "flights", "lufthansa", "britishairways",
"british", "airways", "klm", "ryanair", "easyjet" ))
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, stripWhitespace)
#Text stemming
#docs <- tm_map(docs, stemDocument, language = "english")

#########################################################
############### MOST FREQUENT WORDS ####################
#########################################################

### BUILDING A TERM-DOCUMENT MATRIX
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)


### GENERATING WORD CLOUD
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
        max.words=200, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))
```

```
### PLOT WOROD FREQUENCIES
#The frequency of the first 10 frequent words are plotted
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,
        col ="lightblue", main ="Palabras mas Frecuentes",
        ylab = "Frecuencia")

##########################################################
############### SENTYMENT ANALYSIS ###################
##########################################################

####### USING NRC LEXICON

tweet_sentiment <- get_nrc_sentiment((tweet_text))
#Calculation the total score for each sentiment
sentiment_score <- data.frame(colSums(tweet_sentiment[,]))
view(sentiment_score)
names(sentiment_score)<-"Score"
sentiment_score<-cbind("sentiment"=rownames(sentiment_score),sentiment_score)
rownames(sentiment_score)<-NULL

#Visualization
ggplot(data=sentiment_score,aes(x=sentiment,y=Score))+geom_bar(aes(fill=sentiment),stat =
"identity")+theme(legend.position="none")+xlab("Sentimientos y
Emociones")+ylab("Puntuación")+ggtitle("Emociones de los tuits hacia Lufthansa")

######### USING BING LEXICON

#SPLIT THE TEXT IN WORDS-TOKEN
#3.5.2 http://www.aic.uva.es/cuentapalabras/primer-analisis-de-texto.html#introduccion-2
mensaje <- tibble(text = tweet_text)
#Split in token (3.5.3 of the link above)
word_token <- mensaje %>%
  unnest_tokens(word, text)
#View(word_token)

#Sentiment analysis using bing
airline_sentiment=word_token %>% inner_join(get_sentiments("bing")) %>% count(word,
sentiment, sort = TRUE) %>% ungroup()

airline_sentiment %>% group_by(sentiment) %>% top_n(10) %>% ungroup() %>%
mutate(word = reorder(word, n)) %>% ggplot(aes(word, n, fill = sentiment)) +
geom_col(show.legend = FALSE) + facet_wrap(~sentiment, scales="free_y") + labs(title="", y =
"Contribution to sentiment",x=NULL) + coord_flip() + theme_bw()
View(airline_sentiment)
#View(get_sentiments("nrc"))
#version("nrc")
```

```
# Get the sentiment score
bing_score = function(twt){
  #text cleaning
  twt_tbl = tibble(text=twt) %>%
    mutate(
      stripped_text = gsub("http\\S+","",text)
    )%>%
    unnest_tokens(word,stripped_text)%>%
    anti_join(stop_words)%>%
    inner_join(get_sentiments("bing"))%>%
    count(word,sentiment,sort=TRUE)%>%
    ungroup()%>%
    #create column score
    mutate(
      score=case_when(
        sentiment == 'negative'~n*(-1),
        sentiment == 'positive'~n*1)
    )
  ## calculate total score
  sent.score = case_when(
    nrow(twt_tbl)==0~0, #if there are no words, score is 0
    nrow(twt_tbl)>0~sum(twt_tbl$score)#otherwise, sum the positive and negatves
  )
  ## to keep track of which tweets containet no words at all from the bing list
  zero.type = case_when(
    nrow(twt_tbl)==0~"Type 1",
    nrow(twt_tbl)>0~"Type 2"
  )
  list(score = sent.score, type=zero.type, twt_tbl = twt_tbl)
}

#Apply the function to set of tweets
#this will return a list of all the sentiment score, types, and tables of the tweets
airline_sent = lapply(airline$text,function(x){bing_score(x)})

#Creating a tibble that specifies the airline
airline_score=bind_rows(
  tibble(
    legend = "score level",
    score = unlist(map(airline_sent,'score')),
    type = unlist(map(airline_sent,'type'))
  )
)
View(airline_score)ggplot(airline_score,aes(x=score, fill = legend))+
geom_histogram(bins=15,alpha=.6)+facet_grid(~legend)+theme_bw()
```