Análisis de letras de canciones en función de su país de origen

Minería de datos y el paradigma Big Data
Curso 2018–2019

Autor
Alejandro Díaz Román
Jaime Del Rey García
Paula Munoz Lago

Facultad de Informática
Universidad Complutense de Madrid

Índice

Capítulo 1

Introducción

El auge de la producción musical en los últimos años ha propiciado un aumento en el volumen de los datos, de los cuales se puede extraer una gran cantidad de información. Para el desarrollo de este proyecto nos interesará obtener el contexto social y cultural del país de procedencia del autor de cada canción en función de las letras de sus canciones. Para ello utilizaremos R, el cual se adecua a la perfección a la tarea que queremos realizar, ya que cuenta con una gran cantidad de librerías ya creadas (tanto por R como por la comunidad) que nos ayudarán en gran medida a la hora de análisis de textos, generación de gráficos y minería web. Esta última técnica será una parte importante de nuestro trabajo, ya que necesitaremos obtener información de fuentes externas como Wikipedia, para completar la información que nos aportan los datasets de los que partiremos.

En los siguientes capítulos plantearemos en detalle los objetivos propuestos al inicio del desarrollo del proyecto (Capítulo 2). A continuación expondremos la metodología utilizada en el Capítulo 3 para después explicar en detalle los pasos seguidos en la Implementación del proyecto en el Capítulo 4. Los resultados obtenidos tras implementar el sistema se explican en el Capítulo 5. Finalmente las conclusiones y el trabajo futuro se presenta en el último Capítulo, número 6.

Capítulo 2

Objetivos

En esta práctica, nuestro objetivo es extraer, además de las palabras más utilizadas de cada autor, y de cada país, el sentimiento que prima en dicho país.

Partimos de la hipótesis de que la temática principal de las letras de las canciones españolas es el amor, y lo comprobaremos como parte de esta práctica. El alcance de nuestro trabajo lo hemos establecido en función al tiempo del que disponemos, por lo que no pretendemos cubrir la totalidad de los países, sino los más significativos o de los que dispongamos la información suficiente.

Capítulo 3

Metodología

Para realizar la investigación hemos barajado las distintas metodologías y hemos seleccionado *KDD (knowledge discovery in databases)*. Este modelo está directamente relacionado con la extracción de información y conocimiento de los datos disponibles. El proceso iterativo nos ayudará a obtener, transformar e interpretar toda la información con la que vamos a trabajar.

Capítulo 4

Implementación

4.1. Desarrollo de la aplicación

Dentro del campo de procesamiento de textos, el análisis de letras de canciones no es el más popular, sin embargo existen algunos trabajos que nos han servido de guía para establecer los límites y punto de comienzo de nuestro proyecto. Sin duda el proyecto existente que más se ajusta a nuestra primera idea es *Data, data*, que hace un estudio exhaustivo de las canciones del cantante Uruguayo Jorge Drexler, tanto de la letra como de la música.

Como hemos comentado en el Capítulo 2, nuestros objetivos son obtener los sentimientos que predominan en diferentes países, y comprobar si la hipótesis que hemos establecido indicando que el tema dominante en España es el amor.

4.2. Creación del conjunto de datos

Nuestro punto de partida son los datos que encontramos en varios datasets de Kaggle [1] [2]. Cuyo formato, en el primer caso, es el que se aprecia en la Figura 4.1. Se componede 4 columnas, la primera contiene el nombre del grupo, la segunda el título de la canción, seguida en la siguiente columna de un link que unido a http://www.lyricsfreak.com/ nos redirige a una página donde encontramos más información de la canción, como el año de publicación entre otros, y por último la letra. En el caso del segundo dataset, el formato es el que se aprecia en la Figura 4.2. En él encontramos el nombre del grupo, seguido del año de publicación de la canción, cuyo nombre aparece en la tercera columna, y finalmente la letra.

Un dato relevante en el desarrollo de ésta práctica del cual no disponemos a través de Kaggle es el país de procedencia de cada artista. Para obtenerlo

[1] `https://www.kaggle.com/mousehead/songlyrics`
[2] `https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics`

ABBA	Ahe's My Kind Of Girl	/a/abba/ahes+my+kind+of+girl_20598417.html	Look at her face, it's a wonderful face And it means something special to me Look at the way that she smiles when she sees me How lucky can one fellow be? She's just my kind of girl, she ma...

Figura 4.1

ego-remix	2009	beyonce-knowles	Pop	Oh baby, how you doing? You know I'm gonna cut right to the chase Some women were made but me, myself I like to think that I was created for a special purpose You know, what's more special than you? Y...

Figura 4.2

utilizaremos técnicas de minería web para extraer dicha información a través de Wikipedia u otras fuentes. Una vez obtenida la información (país, autor, letras de sus canciones), limpiaremos los datos escogiendo sólo las palabras más relevantes, siendo éstas los verbos, sustantivos y adjetivos, eliminando artículos, conjunciones.

4.3. Limpieza y preprocesamiento de los datos

La limpieza de las letras es un paso crucial en el desarrollo de éste proyecto, puesto que muchas letras están compuestas de palabras que no aportan información relevante a la hora de realizar el análisis, como son los artículos, conjunciones, etc. Por ello, una vez obtenidos todos los datos, el paso siguiente será eliminar las llamadas *stopwords* Figura 4.3 , tanto del inglés como del castellano. Éste conjunto de palabras está compuesto por los tipos comentados anteriormente, es decir, palabras que no aportan información al análisis. Para ello, hemos desarrollado una función en R que, dado un texto, extrae únicamente las palabras que se encuentran en el diccionario, tanto en el de inglés como en el de castellano. Con ello buscamos evitar inconsistencias además de identificar y eliminar todos aquellos datos que aportarían ruido al análisis, como espacios extra o números.

```
ASCL[,3] <- removeWords(as.character(ASCL[,3]), words = c(stopwords("english"), "oh", "ah", "eh", "uh", "ma"))
ASCL[,3] <- stripWhitespace(ASCL[,3])
ASCL[,3] <- removePunctuation(ASCL[,3])
```

Figura 4.3: *C*ódigo en el que se retiran las llamadas "stopwords"

El dataframe resultante contendrá, pues, 3 columnas indicando el grupo, la canción y las letras de cada canción. Figura 4.3

Pink Floyd	lucifer sam	lucifer sam siam cat always sitting side always side cats...
Pink Floyd	lucy leave	leave ask leave lucy please far away lucy go little girl s...
Pink Floyd	money	money get away get good job pay okay money gas grab...
Pink Floyd	octopus	trip heave ho fro word trip trip dream dragon hide wing...
Pink Floyd	on the turning away	turning away pale downtrodden words say understand ...
Pink Floyd	point me at the sky	hey eugene henry mcclean finished beautiful flying mac...

Figura 4.4: *DataFrame resultante del pre-procesamiento de datos*

Hemos tenido que eliminar los carácteres raros que no se encuentren en el diccionario en inglés (y también las palabras que los contienen) ya que no nos serán útiles para analizar el sentimiento de las canciones. Para ello utilizaremos funciones basadas en la librería *"qdapDictionaries"*. En este caso, como se puede ver en la figura 4.5, la cual devolverá si la palabra introducida como parámetro está en inglés o español.

```
is.word  <- function(x) x %in% GradyAugmented
```

Figura 4.5: *Función usada en el proceso de eliminar carácteres raros y palabras que los contienen*

Crearemos un data frame en el que tendremos disponible la cantidad de canciones de cada grupo, junto al nombre del mismo. Figura 4.6

```
artists <- cbind(as.data.frame(table(ASCL$artist), stringsAsFactors = FALSE), NA)
```

Figura 4.6: *Creación del data Frame con los artistas y la frecuencia de sus canciones*

Para facilitar el trabajo a la hora de identificar el nombre de los artistas, normalizaremos el nombre de los grupos para que tengan todos la misma forma. Figura 4.7

Como parte del preprocesamiento de los datos y con el fin de obtener una descripción más visual del set de datos obtenido, hemos realizado algunas pruebas. En primer lugar, hemos obtenido las palabras que aparecen en las canciones más conocidas del grupo Queen.

Finalmente, tras estudiar los gráficos obtenidos y ver que tiene sentido que las palabras más utilizadas en dichas canciones son las que se muestran, procedemos a generalizar un poco más nuestra exploración, para ver cuales son las palabras más utilizadas por el grupo británico. Como se muestra en la figura 4.13, la palabra *amor* aparece más de 400 veces en las canciones de las que disponemos de dicha banda.

```
for (i in 1:length(artists$Var1)) {
  artists[i, 1] <- paste(unlist(firstup(split_words(artists[i, 1]))), collapse = "_")
}
```

Figura 4.7: Estandarizamos el nombre de los artistas

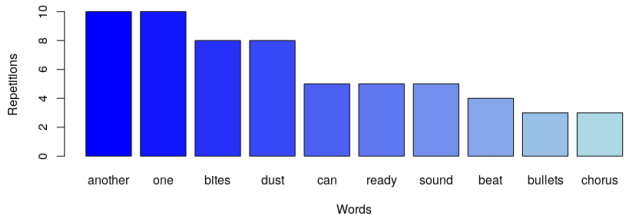

Figura 4.8: *Another One Bites the Dust* - Queen

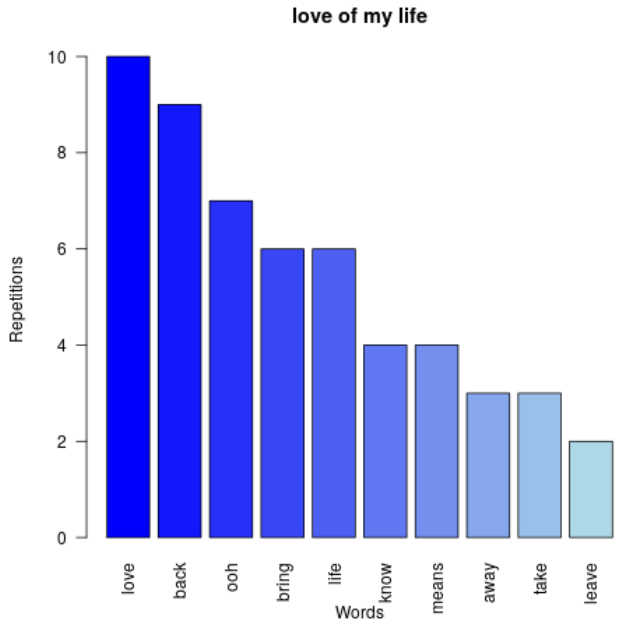

Figura 4.9: *Love Of My life* - Queen

4.4. Transformación de los datos

Para poder trabajar de manera eficiente y general con los datos no es suficiente con una estructura que los contenga a todos, ya que en R concretamente podemos encontrar algunas que trabajan con tipos de datos heterogéneos. Si bien los datos en bruto comparten el tipo character, el conjunto

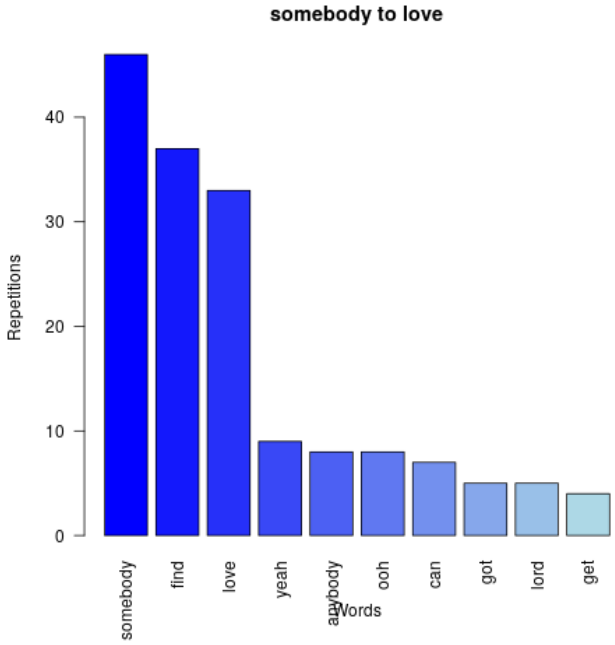

Figura 4.10: *Somebody To Love* - Queen

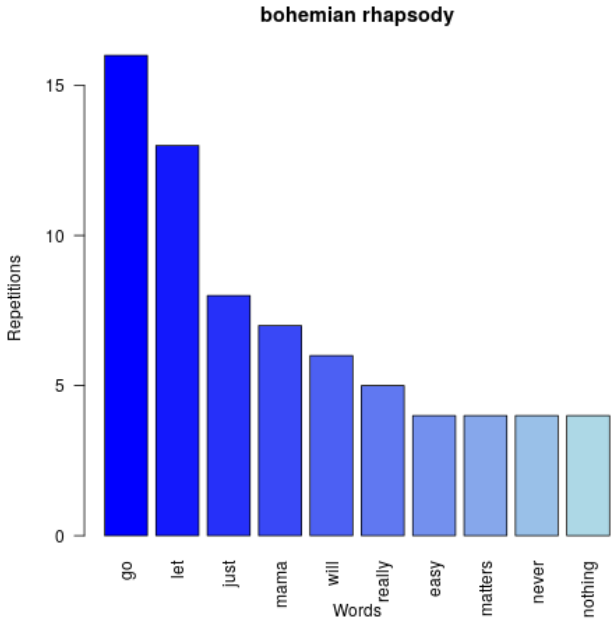

Figura 4.11: *Bohemian Rhapsody* - Queen

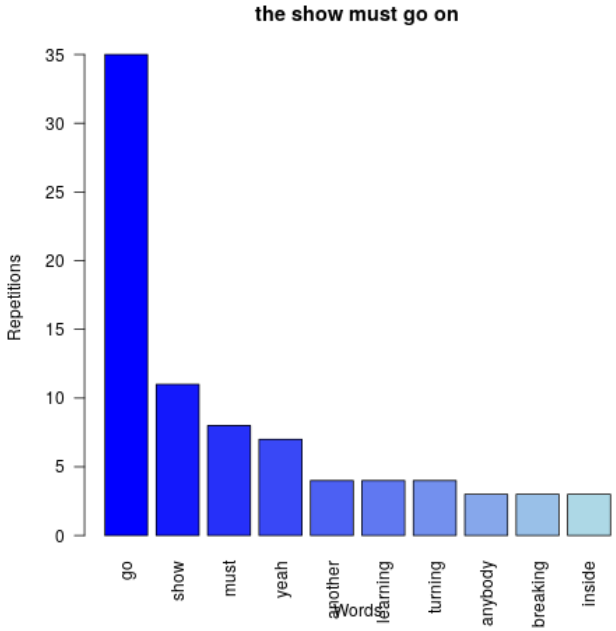

Figura 4.12: *The Show must go on* - Queen

cuenta con caracteres específicos no comunes en el lenguaje habitual además de variar entre mayúsculas y minúsculas. Este último detalle es vital ya que para el reconocimiento de palabras, tanto para su análisis como su limpieza, la homogeneidad de los datos será fundamental.

Tras trabajar con diferentes características homogéneas, concluimos que las letras de las canciones y sus títulos han de transformarse a minúsculas y el nombre de los artistas mantenerlo según la fuente para poder usarlo en búsquedas posteriores con técnicas de minería web.

4.5. Elección de la tarea de minado adecuada

Para conseguir extraer el sentimiento de las canciones utilizaremos diccionarios de sentimientos Figura 4.14, en los cuales diferentes palabras tienen asignados uno o más sentimientos. Para ello elaboraremos diferentes funciones para determinar tanto el sentimiento de cada palabra, lo cual nos permitirá extraer el sentimiento general tanto del artista como del país al que pertenece.

También necesitaremos extraer el pais de origen de cada artista, para ello utilizaremos técnicas de minería web para extraer dicha información de Wikipedia.

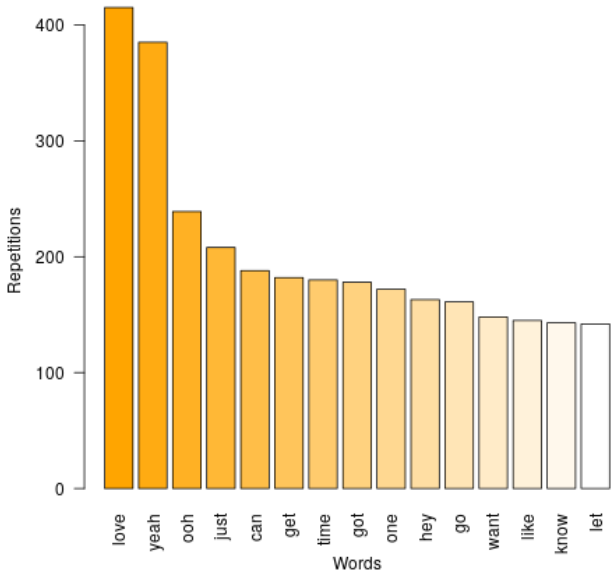

Figura 4.13: Most used words by Queen

214	adorable	joy
215	adorable	positive
216	adoration	joy
217	adoration	positive
218	adoration	trust
219	adore	anticipation
220	adore	joy
221	adore	positive
222	adore	trust

Figura 4.14: Diccioanrio con los sentimientos

4.6. Elección del algoritmo de minería de datos

A la hora de decidir cómo afrontar la clasificación de sentimientos una de las opciones que barajamos fue usar el clustering para saber si podíamos agrupar el conjunto de datos en base al sentimieno predominante.

Para ello aplicamos el método "K-Means"sobre el dataframe que veremos más adelante que agrupa los sentimientos de los países. Con tres clusters no nos aportó resultados relevantes. Después probamos con "DIANA"para dejar al algoritmo decidir, pero tampoco conseguimos gran cosa. Por ello decidimos afrontar el problema con la mentalidad de abarcar todo el dataset, para ello, en la siguiente sección explicaremos todos los pasos ejecutados.

4.7. Aplicación del algoritmo elegido

Para poder utilizar el diccionario de sentimientos primero necesitaremos descomponer las palabras que conforman las letras de las canciones, para ello utilizaremos la función de la Figura 4.15.

```
get_existing_words <- function(x){
  lyric <- list()
  all_words <- unlist(stri_extract_all_words(x))
  for (w in all_words) {
    if (is.word(w)){
      lyric <- c(lyric, w)
    }
  }
  return(unlist(lyric))
}
```

Figura 4.15: Función para extraer las palabras de las letras y existan en el diccionario

Esta función iterará por todas las letras de las canciones y aquellas que estén presentes en el diccionario de sentimientos, las guardará. Como tiene que procesar cada palabra del total de 419887 canciones tiene un coste computacional muy elevado. Esto ha sido uno de los principales retos con los que nos hemos encontrado, ya que no podíamos ejecutar el algoritmo en cada iteración ya que tardaba muchas horas en realizarlo, para ello fuimos cogiendo pequeñas partes del data set a modo de pruebas para comprobar que los resultados eran satisfactorios.

El diccionario que usaremos está indicado en la Figura 4.16.

```
words_sentiments <- get_sentiments(lexicon = "nrc")
```

Figura 4.16: Diccionario que usaremos para la asignación de sentimientos

Ahora organizaremos las 419887 entradas por autores y tendremos un data frame resultante indicado en la figura 4.17

Con toda la información ordenada ya podemos asignar el sentimiento principal de cada artista con la función de la Figura 4.18. Esta función de-

David Bowie	list [165 x 2] (S3: data.frame	A data.frame with 165 rows and 2 columns
artist	factor	Factor with 18874 levels: "'n Sync", "ABBA", "Ace Of Base..
lyrics	character [165]	'someday let now must agree times telling changing free ..
David Guetta	list [63 x 2] (S3: data.frame)	A data.frame with 63 rows and 2 columns
artist	factor	Factor with 18874 levels: "'n Sync", "ABBA", "Ace Of Base..
lyrics	character [63]	'hands stretch got give best turn cause always impress al..

Figura 4.17: Data frame resultante con los artistas agrupados con sus letras de las canciones

volverá una lista de palabras, que serán los sentimientos, con la cual elaboraremos el data frame que asigna a cada artista el sentimiento principal tal y como se ve en la Figura 4.19.

Una vez tenemos listo el diccionario podemos asignar a cada palabra de las letras de cada artista los sentimientos que tiene asignados. Toda esta información la tendremos en un dataframe intermedio, el cual usaremos para obetener la frecuencia de todos los sentimientos y el que predomine lo asignaremos junto al autor en data frame principal Figura 4.19.

```
function(token_list){
  words_sentiments <- get_sentiments(lexicon = "nrc")
  sentiments <- list()
  for (token in token_list){
    if (any(words_sentiments$word == token)){
      sentiments <- c(sentiments, words_sentiments[which(words_sentiments$word == token), 2])
    }
  }

  return(sentiments)
}
```

Figura 4.18: Función para obtener los sentimientos de las palabras

IGGY POP	positive
IL DIVO	positive
IMAGINE DRAGONS	negative
IMAGO	positive
IMPERIALS	positive
INCOGNITO	positive
INCUBUS	negative

Figura 4.19: Tabla con el artista junto a su sentimiento predominante

La parte final será pues, obtener el país correspondiente de cada artista, para ello hemos utilizado técnicas de minería web. Vamos a utilizar Wikipedia como fuente para obtener el pais correspondiente, para ello, como la URL puede estar en diferentes formatos tenemos que obtener las distintas

formas tal y como se indica en la Figura 4.20.

```
pwebs <- c(paste("https://es.wikipedia.org/wiki/", artists[i, 1], sep=""),
           paste("https://es.wikipedia.org/wiki/", artists[i, 1], "_(banda)", sep=""),
           paste("https://es.wikipedia.org/wiki/", artists[i, 1], "_(cantante)", sep=""),
           paste("https://en.wikipedia.org/wiki/", artists[i, 1], sep=""),
           paste("https://en.wikipedia.org/wiki/", artists[i, 1], "_(singer)", sep=""),
           paste("https://en.wikipedia.org/wiki/", artists[i, 1], "_(band)", sep=""))
```

Figura 4.20: Diferentes formatos para obtener la URL

Después, con la función de la Figura 4.21 podremos saber cual de los links que hemos generado es el válido.

```
exists <- sapply(pwebs, url_exists)
```

Figura 4.21: Comprueba cual de los links es válido

Ese link, lo usaremos para finalmente obtener el artista, pero comprobaremos tanto la fuente en inglés como en español por si algún artista no tiene página en wikipedia en el otro idioma, de la forma mostrada en la Figura 4.22.

```
country <- extract_country_es(a, existing_countries_es, world.cities)
country <- extract_country_en(a, existing_countries_en, world.cities)
```

Figura 4.22: Extraer el pais adecuado de Wikipedia

Con esta nueva información generamos dos data frames, el primero de la Figura 4.23 donde tendremos el artista junto con su cantidad de canciones y el pais correspondiente y el data set final de la Figura 4.24 donde tenemos toda la información que necesitamos para inferir cual es el sentimiento general de cada país.

COOLIO	103	Estados Unidos
COQUE MALLA	14	España
COR SCORPII	8	Noruega
CORALIE CLEMENT	12	Francia
CORBIN BLEU	31	Estados Unidos
COREY HART	9	Canadá
COREY TAYLOR	1	Estados Unidos
CORINNE BAILEY RAE	45	Reino Unido

Figura 4.23: Data frame con el artista, el número de canciones y el país

	Suecia	Estados.Unidos	Reino.Unido	Australia	Filipinas	Italia
anger	29	159	73	4	2	14
negative	1	35	11	1	0	0
positive	6	218	84	13	6	1

Figura 4.24: Data frame con los sentimientos de las canciones de cada país

Capítulo 5

Resultados y Evaluación

Finalmente tranformando la información del data frame final en una gráfica con los sentimientos predominantes (Figura 5.1) podemos observar que aunque si que haya asignado sentimientos a las letras de las canciones es un poco raro que haya tal cantidad de enfado en ellas. Para ello veamos más detenidamente los resultados de otros países.

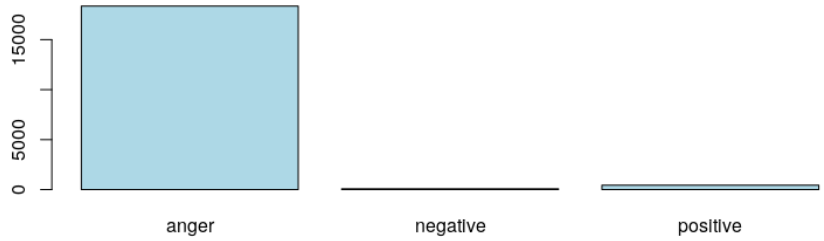

Figura 5.1: Principales sentimientos

Como podemos ver en las Figuras 5.2, 5.3 y 5.4, en países de habla inglesa como Estados Unidos, Reino Unido o Australia, el reparto entre sentimientos positivos y de enfado es bastante equitativo. Sin embargo podemos observar que países en los que las letras no son en inglés o solo una pequeña parte son en inglés automáticamente el sentimiento es clasificado como de enfado. Se puede ver claramente en países como Colombia, Noruega, Tayikistán, etc...

Por último si miramos la distribución en España en la Figura 5.6, vemos que absolutamente todos los artistas han sido clasificados con un sentimiento de enfado. Una de los posibles fallos es que el diccionario de sentimientos en cuanto detecta otra palabra en otro idioma la clasifica automáticamente como dicho sentimiento, lo cual hace que esté tan presente.

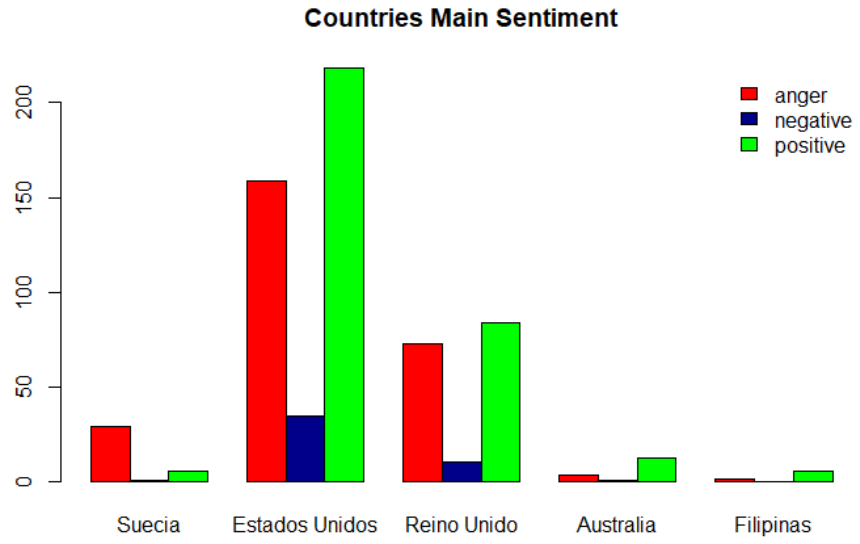

Figura 5.2: Sentimientos detallados por paises

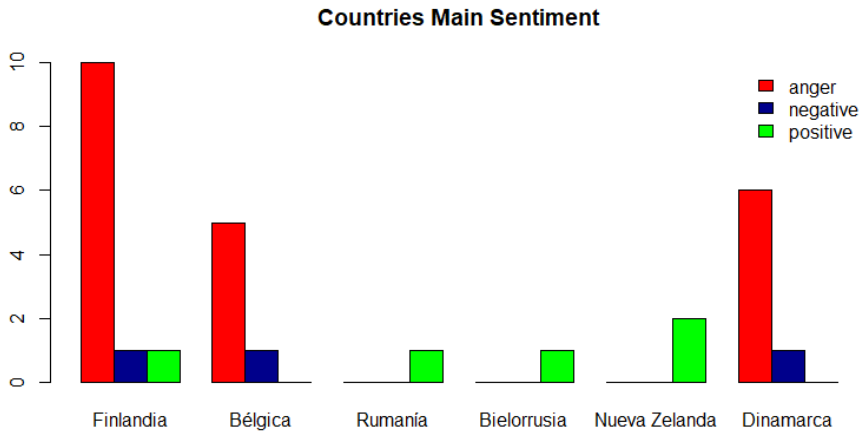

Figura 5.3: Sentimientos detallados por paises

Los resultados observados en la figura 5.6 indican de que otro de los mayores retos con los que nos hemos encontrado es la barrera del idioma. Dicho problema nos abre las puertas a continuar este trabajo y ampliarlo con nuevas técnicas de traducción e intrepretación de textos, lo cual lo veremos en el próximo Capítulo.

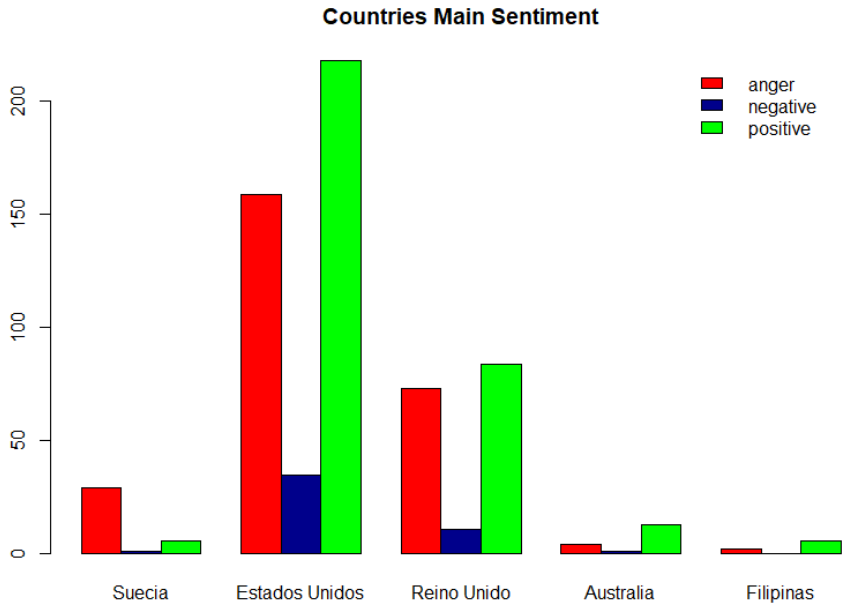

Figura 5.4: Sentimientos detallados por paises

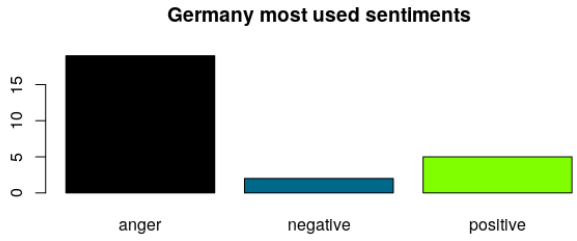

Figura 5.5: Sentimientos detallados en Alemania

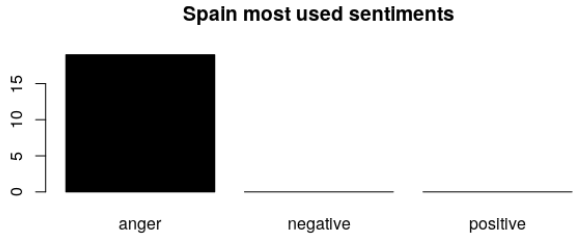

Figura 5.6: Sentimientos detallados en España

Capítulo 6

Conclusiones y Trabajo Futuro

Dado el tiempo disponible para el desarrollo de este proyecto, no hemos podido comprobar nuestra hipótesis inicial de que en España la mayoría de canciones iban a tener un sentimiento, por lo general alegre y positivo. Al haber visto que aquellas canciones que están en inglés resultan mejor clasificadas, sabemos que tendremos que incluir un proceso de traducción al inglés o bien traducir el diccionario de sentimientos a otros idiomas. Además, cabe destacar la pérdida de contexto a la hora de analizar las palabras, ya que al tratar las palabras individualmente no podemos saber el significado total de una frase, perdiendo entonces el significado contextual de la palabra y ciñiendonos al significado común. Para ello también tendremos que implementar técnicas de procesamiento de textos mucho más avanzadas y que requieren de mucho más tiempo, esfuerzo e investigación.

Futuras aplicaciones de este trabajo de investigación pueden ser, por ejemplo buscar momentos a lo largo de la historia en los que el sentimiento de las canciones cambibie e intentar encontrar causas históricas a las que pueda estar asociado, por ejemplo guerras, crisis o cuando se gana un evento deportivo.

Si encontrasemos alguna relación entre el sentimiento que desprende la música y eventos históricos tendríamos que plantearnos si es el arte el que imita a la vida o es la vida la que imita al arte.