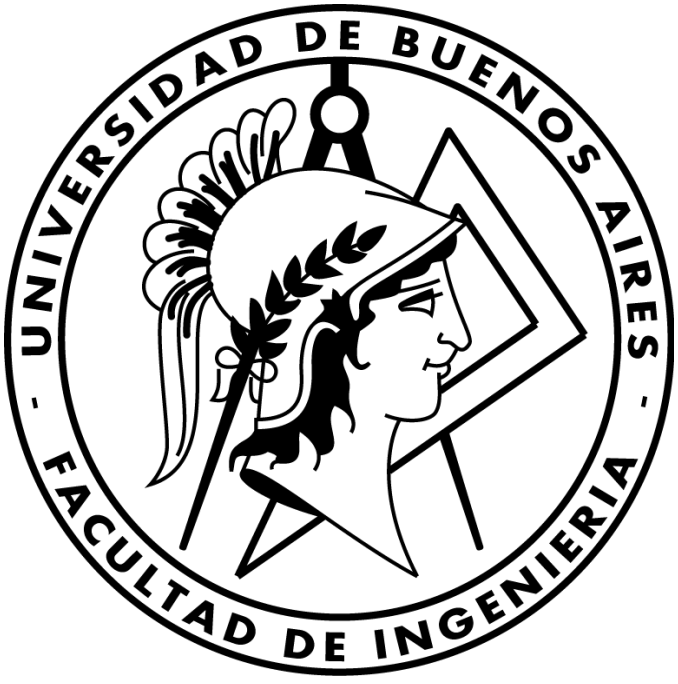
7506 - Organización de Datos: TP1

NLP with Disaster Tweets - Análisis exploratorio de datos

Grupo 1 - Alumnos:

- Cozza, Fabrizio
- López Lecube, Lucio
- Fonzalida Miguel Angel
- Gonzalez, Felipe



Facultad de Ingeniería - Universidad de Buenos Aires
1c 2020

Índice

Introducción

En el presente trabajo práctico se describen los resultados más importantes obtenidos mediante el análisis exploratorio de tweets del set de datos de la competencia https://www.kaggle.com/c/nlp-getting-started.

El archivo train.csv con el que trabajamos cuenta con 7613 registros con 5 atributos que son:

- **id**: identificador unico para cada tweet.
- **text**: el texto del tweet.
- **location**: ubicación desde donde fue enviado.
- **keyword**: un keyword para el tweet.
- **target** : indica si se trata de un desastre real (1) o no (0)

En primer lugar se trabajó sobre el análisis general del set de datos, observando con qué features se cuenta, de qué tipo, cuántos valores únicos hay por feature, visualizando las distribuciones de las distintas variables y las posibles relaciones entre ellas, qué atributos tienen valores nulos y cuántos son, para realizar luego un proceso de limpieza y depuración del set.

Una vez que se obtuvo el panorama inicial de los datos, el segundo paso consistió en buscar cosas más allá de lo que se obtiene a primera vista:

- Lenguaje de los tweets
- Información relevante dentro de los tweets cómo hashtags, etiquetas, links
- Creación de nuevos features que nos ayuden a un más profundo análisis de los tweets en función del target.

Es importante notar que en el informe se presentan los resultados que tienen algún valor de presentación. Todo lo que se fue intentando en el camino se puede leer en los notebooks en el repositorio.

Objetivos

El objetivo del presente trabajo es obtener un conocimiento lo más detallado posible de los datos, teniendo en cuenta que esto debería dejar el camino preparado al segundo trabajo práctico, cuyo objetivo será poder predecir si un tweet habla de un desastre real o no.

Herramientas utilizadas

El análisis exploratorio se realizó en Python 3 con la librería Pandas. Para la visualización se utilizaron las librerías matplotlib, wordcloud, plotly, pysankey y seaborn.

Para el control de versiones se utilizó Git; todo el tp se encuentra en GitHub en https://github.com/miguelAfonzalida/75.06-TP1-Grupo1.

Librerías para control del lenguaje de los tweets: polyglot, pycld2, pyicu.

Librerías para recuperar países a partir de las ubicaciones: geopy, pycountry.

Librerías para manejo de texto: NLTK, FlashText.

Análisis exploratorio

¿Qué datos… no se encuentran?

En primer lugar es de interés conocer con cuántos datos se cuenta. Hay 7613 registros, de los cuales dos de sus atributos tienen la siguiente cantidad de nulos:

- **keyword**: 61 nulos
- **location**: 2533 nulos

A continuación se presenta la proporción de valores faltantes por cada uno de los features que tiene al menos algún registro nulo.

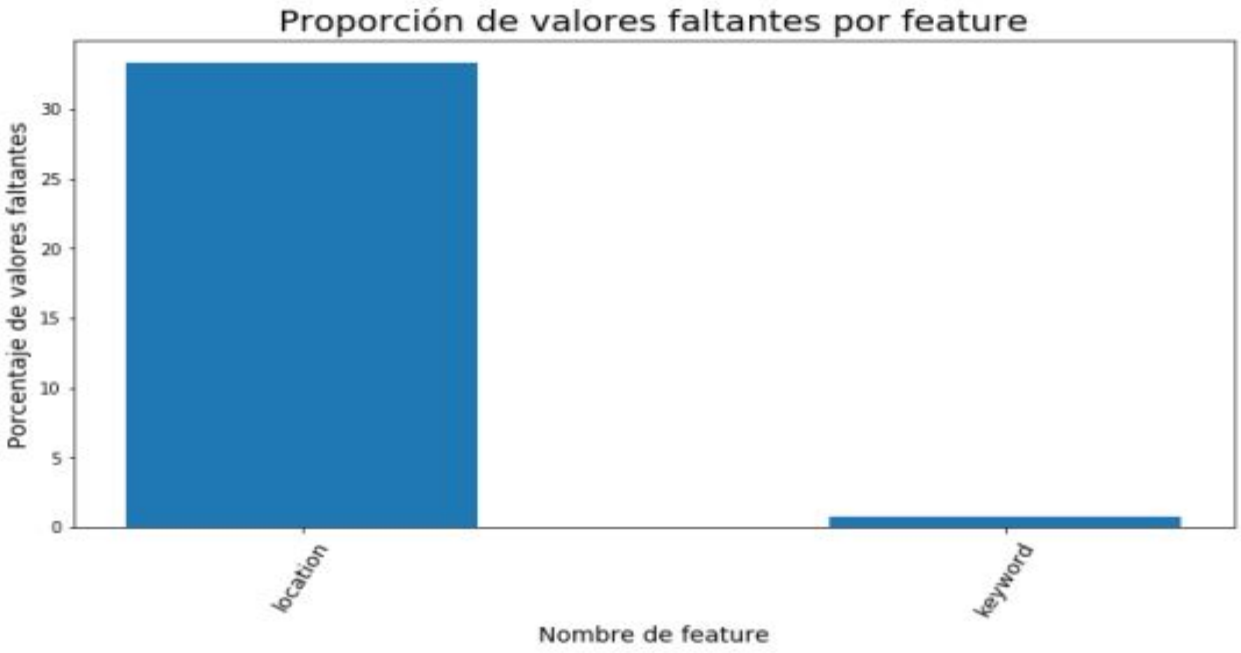


Figura 1

Variable a predecir (target)

También en un primer análisis quisimos saber cómo se distribuyen las cantidades de los tweets en relación al target al que pertenecen:

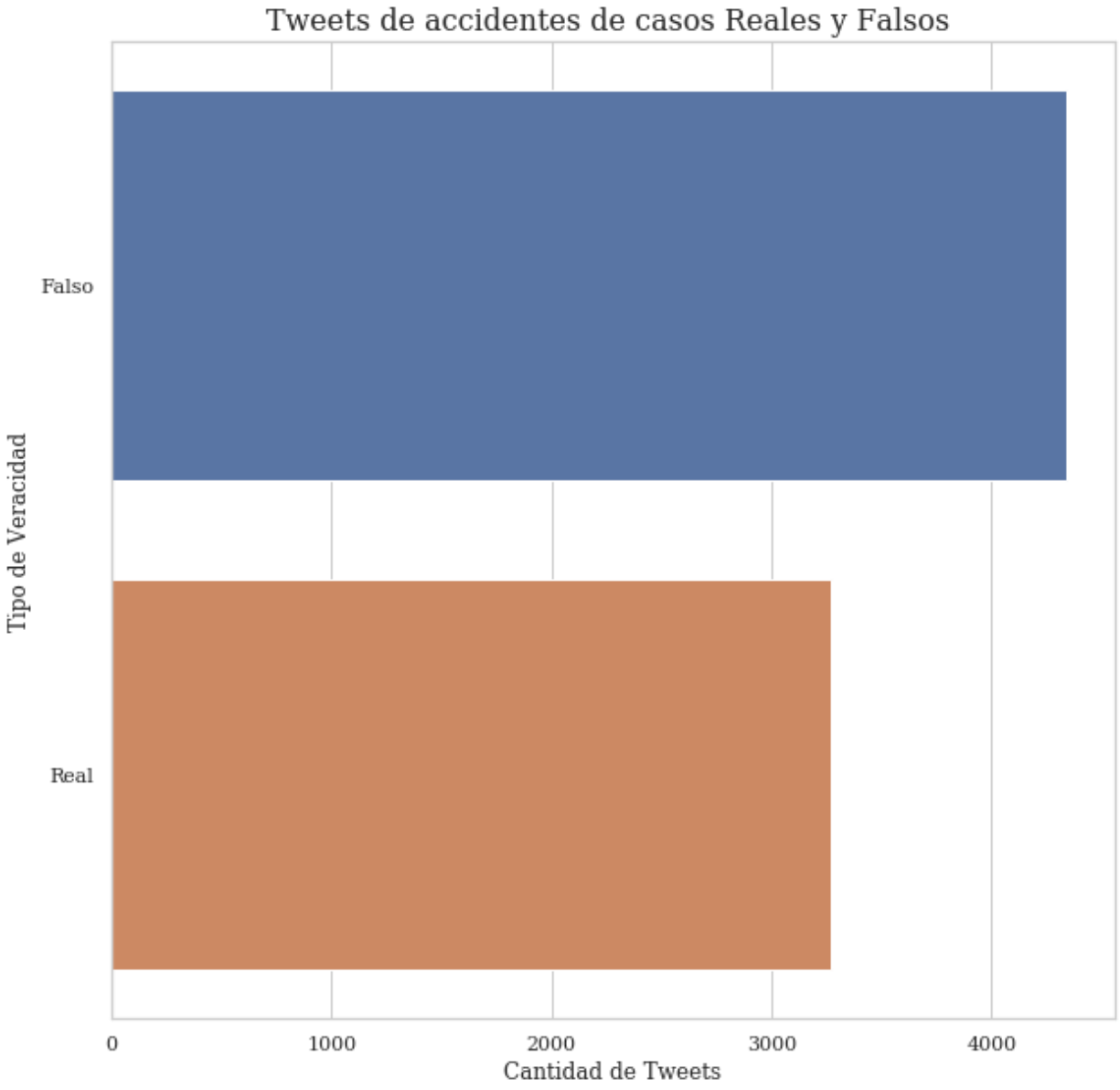


Figura 2

Podemos ver que en la variable que se va a querer predecir en un futuro se tienen más valores de desastres no reales que de reales por lo tanto siempre que se analize según los targets durante el informe se tenderá a ver una mayor cantidad de desastres no reales. En particular, disponemos 4342 para target=0 y 3271 para target=1.

Limpieza y depuración del set

Durante este proceso buscamos que los textos de los tweets sean los más limpios posible para luego realizar un análisis más certero.
También nos resultó interesante mostrar los contrastes entre realizar un análisi inicial del texto sin depuración y luego otro con depuración, primero para no perder algún detalle que pueda quedar en el camino en el proceso de depuración y por otra parte para ver si los cambios en los mismos representaban una diferencia notoria o no.

Keyword

Con este atributo lo primero que realizamos es una depuración de ciertos caracteres que mayormente se presentaban en las keyword de dos palabras, las cuales tenían unos caracteres concatenados a las mismas. Luego de esto se procedió a completar valores faltantes en las keywords.
Nos encontramos con 61 keyword nulas a las que tratamos de asignar una de las key ya existentes, en base al contenido del texto de los tweets. Se logró completar de esta manera a 38 nulos de 61 y para los 23 casos restantes tomamos la decisión de asignarles una keyword llamada “other”.
En relación al hecho de tener tantas keywords distintas (222) decidimos crear Keys Globales, a modo de concepto que trate de sectorizar a las keywords para luego realizar un análisis menos engorroso y un poco más detallado. En total se crearon 11 Keys Globales que son: “Natural Phenomenon”, “Burning”, “Emergency”, “Danger”, “Accident”, “Apocalypse”, “Security”,”Attack”, “Fatality”, “Survivor”, “Otro”. Los nombres de las mismas no son azarosos, ni tampoco la agrupación de las keywords son aleatorias sino que se trató de que por ejemplo, los tweets que hablaran de incendios forestales, incendios de edificios, etc., pertenezcan a la Key Global “Burning”, que los tornados, las tormentas de arena, los huracanes, terremotos, etc., pertenezcan a la Key Global “Natural Phenomenon” y así para los demás casos.

Text

Al texto contenido en los tweets del set de datos le aplicamos un proceso de depuración que consiste básicamente en:

- Quitar todos los links.
- Quitar todo lo que no sea alfanumérico.
- Pasamos todo el texto a minúscula.
- Quitamos todos los espacios en blancos que estén de más.

Lenguaje de los tweets: ¿está todo en inglés?

Antes de meternos en los posibles análisis sobre el texto, es conveniente saber si estamos viendo todo en el mismo lenguaje. Para esto se utilizo la libreria polyglot que detecta el lenguaje de un texto con un medidor de confianza que indica cuán confiable es el hallazgo de la librería.
Afortunadamente la librería devolvió solo 35 registros con lenguaje diferente de los cuales mirándolos en detalle, nos pudimos dar cuenta que eran todos en inglés excepto por 3 registros en particular que mezclaban portugues con inglés y español con inglés, que los dejamos marcados en el notebook por si luego se quieren traducir en un futuro análisis de sentimiento. Dicho esto, concluimos que están todos los tweets en inglés.

Análisis de la feature location

¿Ruido o datos útiles?

Veamos ahora si la columna en cuestión puede sernos útil en el futuro o contiene información meramente demostrativa. A simple vista se puede ver que la columna cuenta con información variada, desde ciudades y países reales hasta comentarios que pueden ser en forma de broma o lugares ficticios.
Lo que se hizo para simplificar el análisis fue la utilización de la librería geopy para poder extraer los países a partir de las ciudades y trabajar todo con países. Al tener todo unificado en este tipo de locación, se puede analizar el feature más globalmente y en detalle. Esto nos dio un total de 127 países con al menos 1 aparición.
Ahora, lo que la librería devolvió no era del todo exacto. Lo que se hizo fue revisar 1 a 1 los registros que tengan menos de 15 apariciones (aproximadamente 100 países) y ver si la información provista por la librería era de utilidad o no.
Varios países/ciudades eran detectados correctamente pero otros eran recuperados de texto que era puro ruido y la librería aún devolvía un valor. Lo que se hizo a continuación fue remover todo aquel registro que fuera obtenido de ruido y asignar correctamente los países si la librería fallaba en ciertos casos.
Luego de toda esta limpieza quedó una cantidad de registros de 3840 sobre los 7613 totales del dataset (50%). Si bien se eliminó una gran cantidad de ruido y se corrigieron algunos datos, estos 3840 registros aún siguen teniendo información imprecisa, ya que no se revisaron todos sino solo aquellos con menos de 15 apariciones y si se vuelve a hacer un análisis rápido se puede seguir viendo que muchos datos aún siguen teniendo ruido.
Por lo cual la conclusión a la pregunta original es que la feature de location, en el caso de que se desee utilizarla, tendrá más de un 50% de los datos que no serán correctos por lo cual en un principio no debería tener un aporte significativo en el futuro y no debería ser útil. Para este TP1 en particular, son meramente demostrativos.

Nota sobre esto: algunas apariciones “interesantes” en esta feature

- Milky way
- Narnia
- In the word of god
- Planet earth
- Wakanda

Un panorama básico

Como se mencionó en el apartado anterior, la variable location en teoría no tendrá gran relevancia para el futuro pero aún así podemos ver y darnos una idea acerca de los valores de la misma.

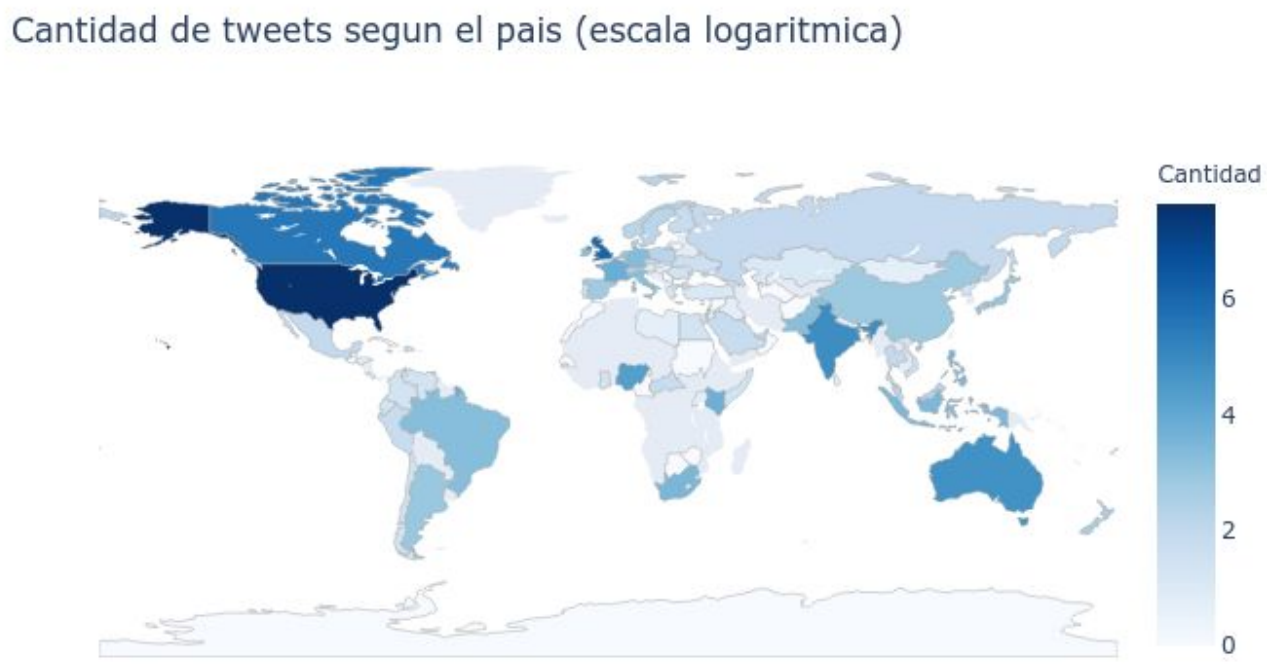


Figura 3

Como podemos ver en el choropleth, hay algunos países dominantes como EEUU, Canadá y el Reino Unido que son los que más tweetearon. Nótese la escala logarítmica para apaciguar el efecto de la gran cantidad de apariciones de EEUU.

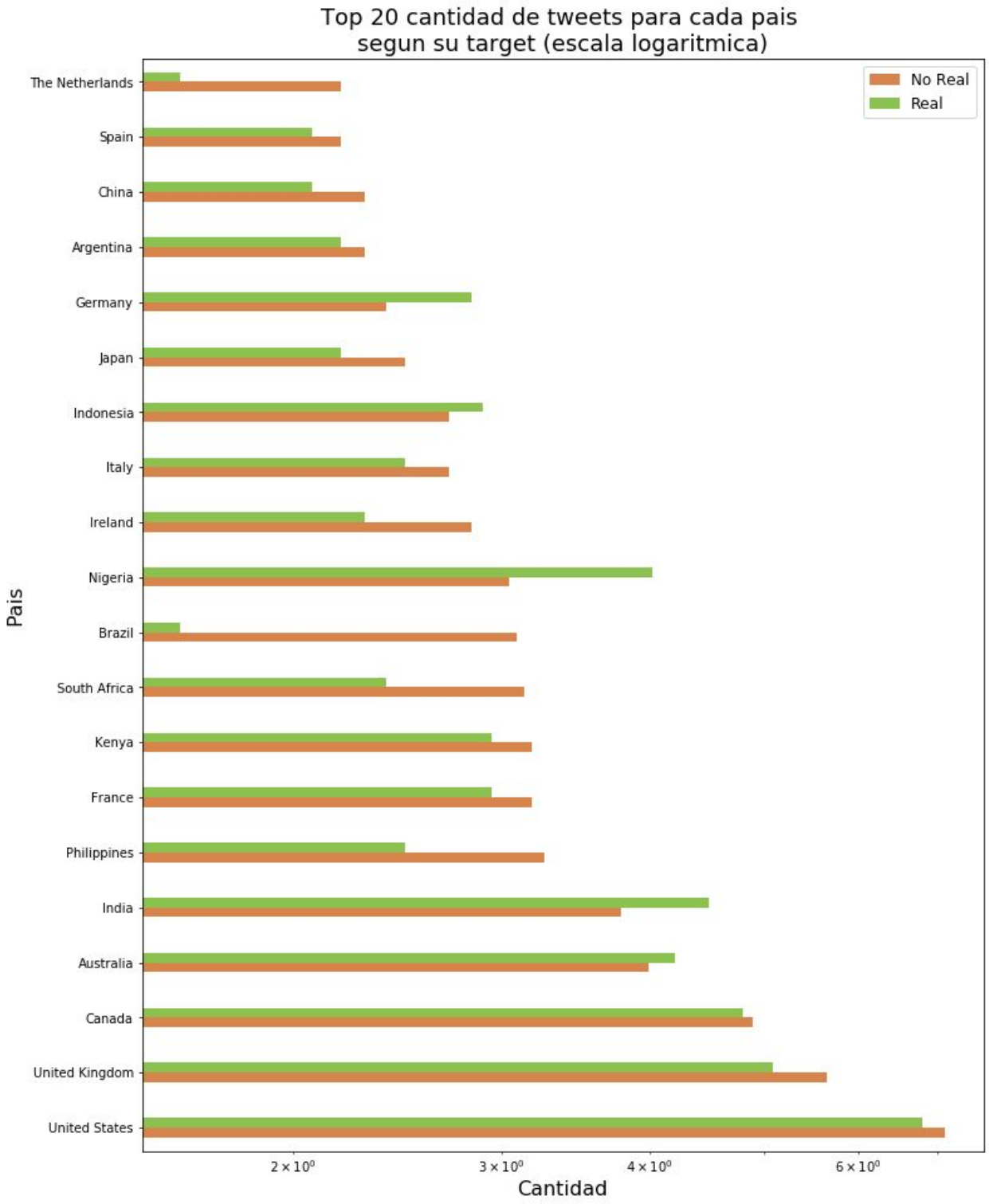


Figura 4

De aquí podemos ver que existe una mayor cantidad de tweets de Estados Unidos, seguido por el Reino Unido y en ambos casos con mayoría de tweets pertenecientes al target=0 (No Real). También se aprecia que los tweets, de nuestro set de datos, provenientes de países como

India, Australia, Nigeria, Indonesia, Alemania hay mayor cantidad de casos pertenecientes al target=1 (Reales), destacándose Nigeria como la que posee una mayor diferencia entre casos reales y no reales. Del otro lado tenemos a Brasil y The Netherlands con una proporción de casos no reales de mucho mayor que los reales.

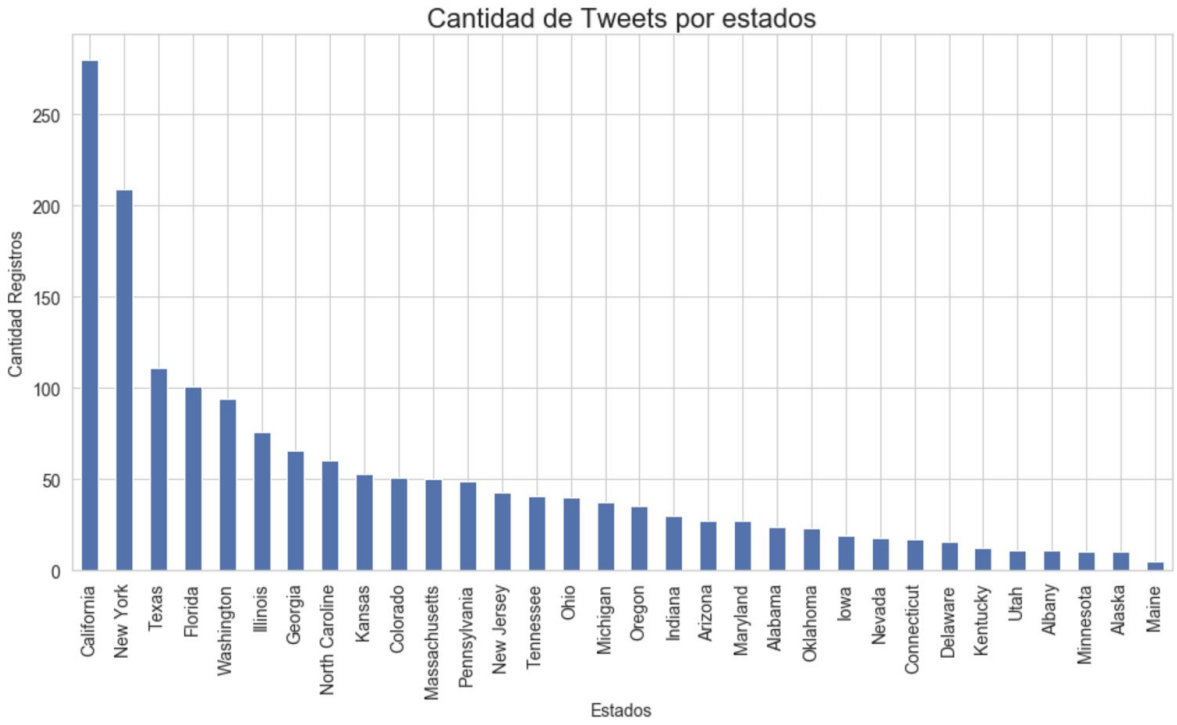


Figura 5

Dado que hemos visto que la mayoría de los tweets se encuentran en Estados Unidos, hemos realizado un análisis tratando de depurar, separando las localizaciones en estados. Aca podemos observar que la mayor cantidad de registros se concentran en el estado de California seguido de New York. Dada la cantidad de ruido del set de datos, no fue sencillo realizar esta separación.

Análisis de la feature keyword

Respecto a este feature tenemos 222 keyword distintas y luego de haber realizado el proceso de depuración de las mismas, comentado con anterioridad, en un primer análisis para tener una idea de cuales son las KeyWords más frecuentes de los tweets realizamos un Wordcloud de las mismas.

Figura 6

Podemos ver que las KeyWords que más parecen destacar son “body bag”, “building burning”, “forest fire” entre otras. Pero resulta importante analizar cuales son las KeyWords que más destacan en los tweets, en relación a sí hacen referencia a un accidente real (target = 1) o no (target = 0). Para lo cual realizamos los siguientes gráficos:

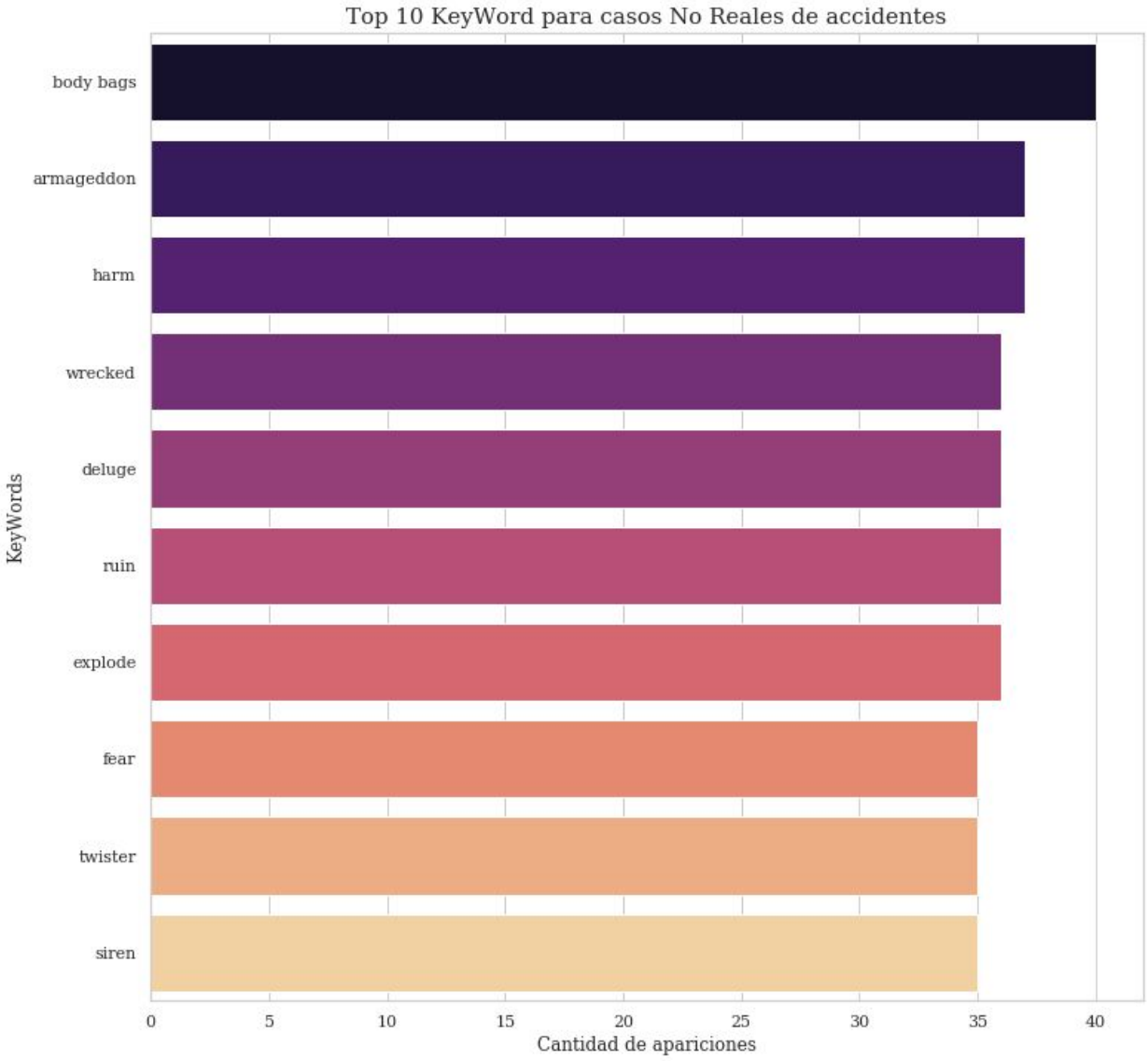


Figura 7

Vemos que las 10 keyword que más aparecen para los tweets que no hacen referencia a accidentes reales, lo hacen con una frecuencia de entre 35 y 40 veces destacándose para este target “body bags” con mayor cantidad de apariciones.

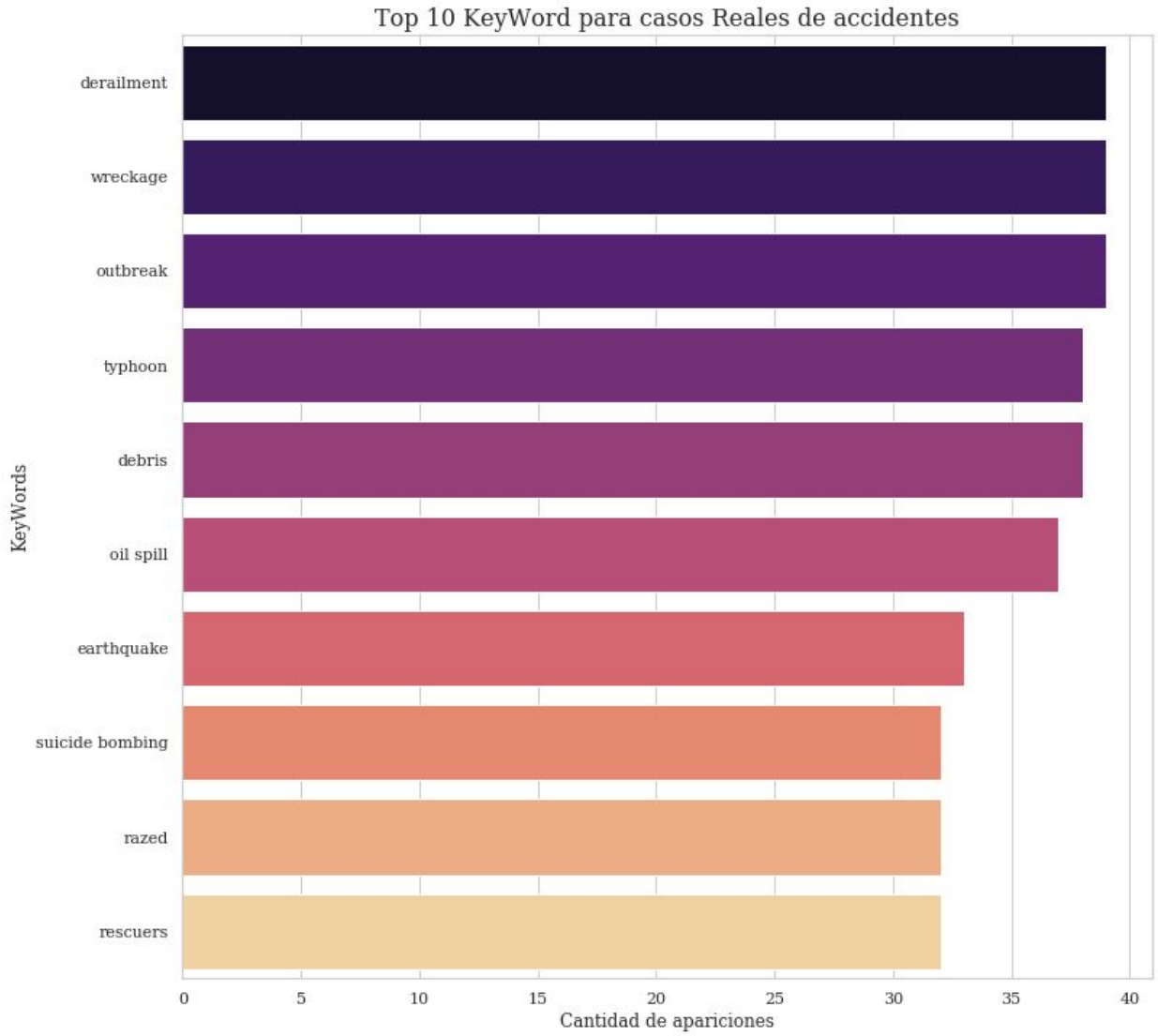


Figura 8

En cuanto a los tweets que hacen referencia a accidentes reales, las frecuencias de apariciones de las 10 keyword más populares están en un rango de un poco más que 30 y menos que 40 apariciones, siendo: “derailment”, “wreckage” y “outbreak” las más destacadas.

Como mencionamos anteriormente por el hecho de tener tantas keywords distintas decidimos crear Keys Globales, a modo de concepto que trate de sectorizar a las keywords para luego realizar un análisis más detallado. En total se crearon 11 Keys Globales que son: “Natural Phenomenon”, “Burning”, “Emergency”, “Danger”, “Accident”, “Apocalypse”, “Security”,”Attack”, “Fatality”, “Survivor”, “Otro”.

Para poder entender mejor veamos en un gráfico como quedan relacionadas, las distintas Key Globales, con los target (de si los tweets hacen referencia o no a un accidente real).

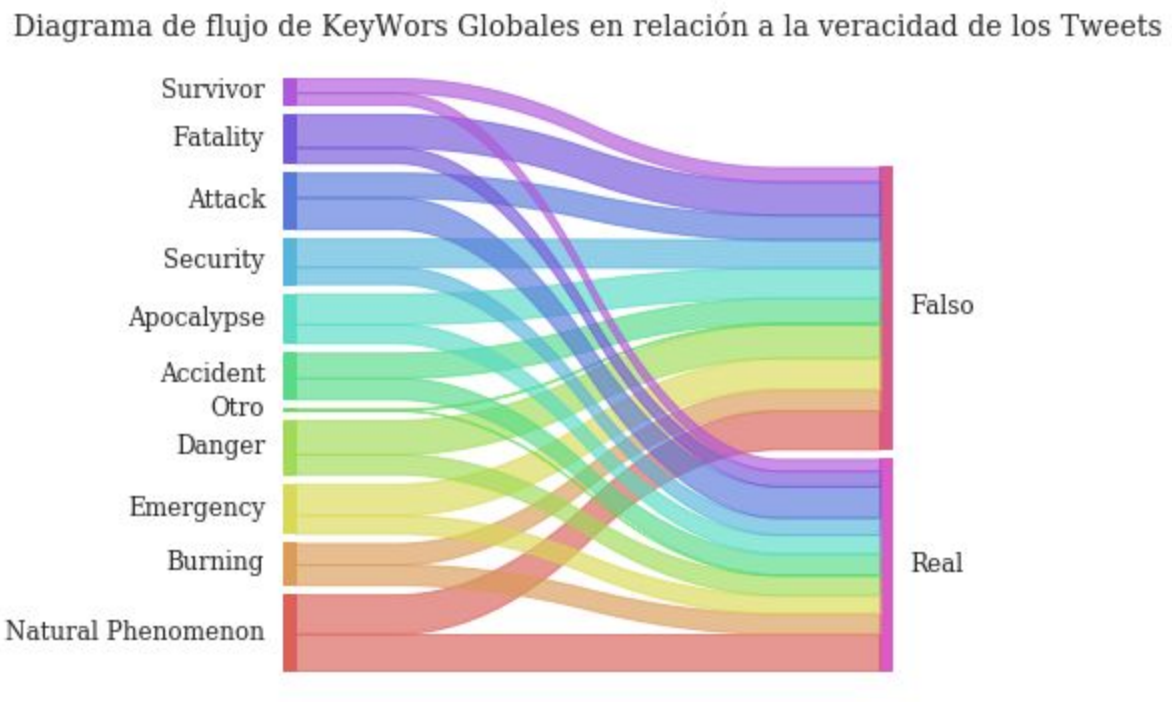


Figura 9

Si también analizamos las frecuencias de las Key Globales mediante Wordcloud tenemos:

Figura 10

Vemos que la Key Global denominada: “Natural Phenomenon” es la que predomina, osea, a lo que más se hace referencia en los tweets de nuestro set de datos, seguidos por tweets que hablan de temas relacionados a “Apocalypse”, ”Attack” y “Danger”.

Otra pregunta interesante que surge es. ¿De las Keywords que disponemos habrán algunas que son exclusivamente de un target?

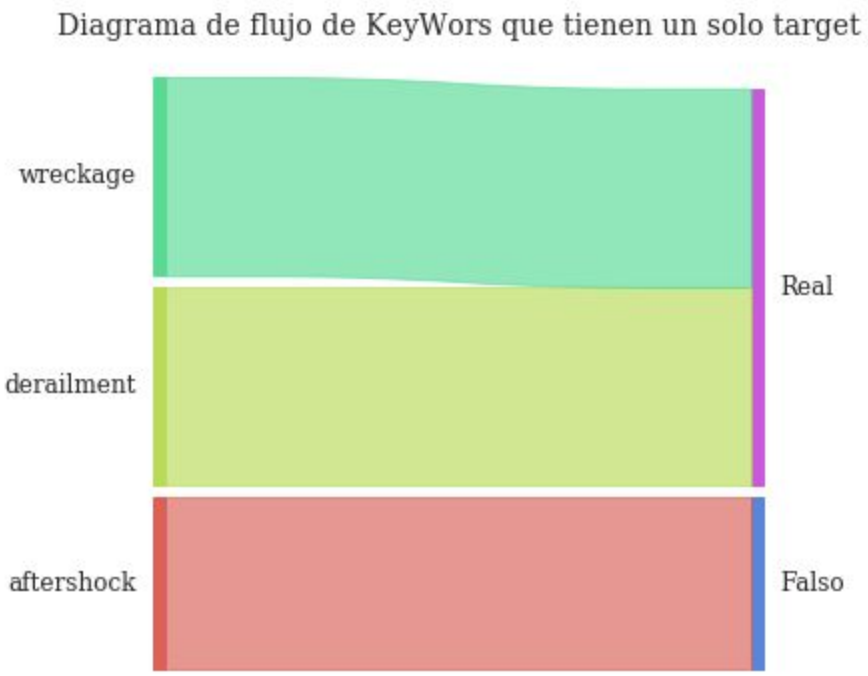


Figura 11

Vemos que existen 3 keywords exclusivas que son: “wreckage”, “derailment” y “aftershock” de las cuales las dos primeras, que corresponden a tweets accidentes reales, presentan un leve mayor grosor de flujo que la keyword exclusiva del raget “Falso” o no real, lo que indica una mayor cantidad de apariciones en nuestro set de datos.

¿Qué cantidad de tweets tenemos, reales o no, según las Key Globales?

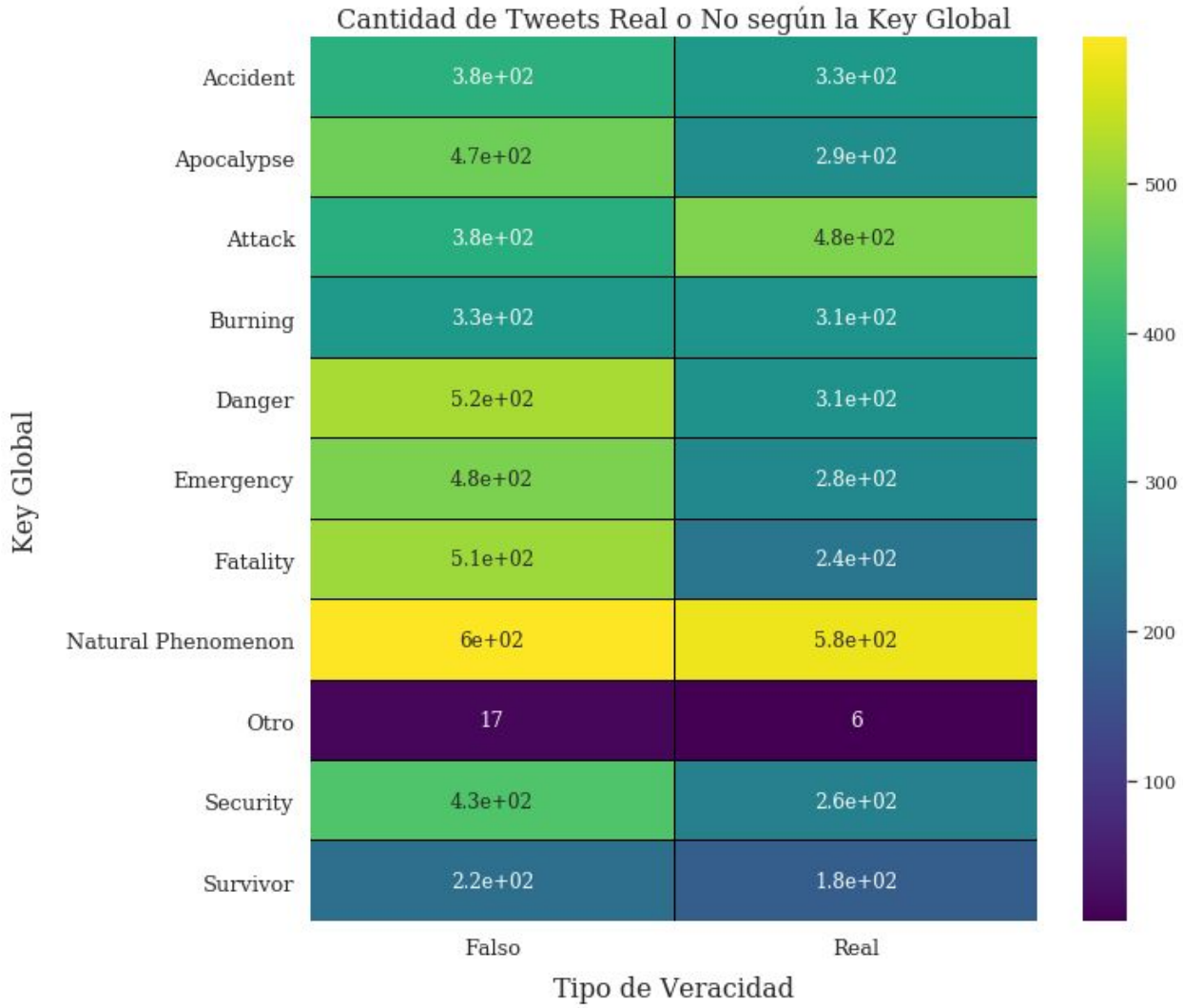


Figura 12

En este heatmap vemos en forma más explícita que la Key Global “Natural Phenomenon” es la más popular, con unas 600 apariciones en tweets relacionados con el target “Falso” y 580 apariciones con el target “Real”.

Salvo para el caso de la Key Global “Attack” (que tiene 480 apariciones para el target “Real”, mayor a las 380 que tiene para el target “Falso”), se observa que, en los demás casos, las Key Globales presentan mayor ocurrencia para el target “Falso” y esto puede deberse al motivo de que, como se mostró al principio del informe, en la figura 2, tenemos mayor cantidad de tweets relacionados con el target “Falso” (target = 0) que tweets con el target “Real” (target = 1).

Análisis de la feature text

Links, menciones y hashtags: que se puede ver dentro de los tweets comunes

Antes de desglosar y analizar features nuevos a partir del texto, se decide analizar lo intuitivo que posee un tweet en su defecto, siendo esto los hashtags, menciones y links de referencia a otros tweets.

Hashtags

Los hashtags de un tweet permiten a los usuarios aplicar un etiquetado que ayuda a otros usuarios a encontrar fácilmente mensajes con un tema o contenido específico. Es la forma más común de referirse a un tema en twitter y pueden ayudar a identificar cómo son etiquetados los desastres en este caso. En el siguiente gráfico podemos ver los hashtags más referidos en el campo de texto.

Wordcloud frecuencia de hashtags

Figura 13

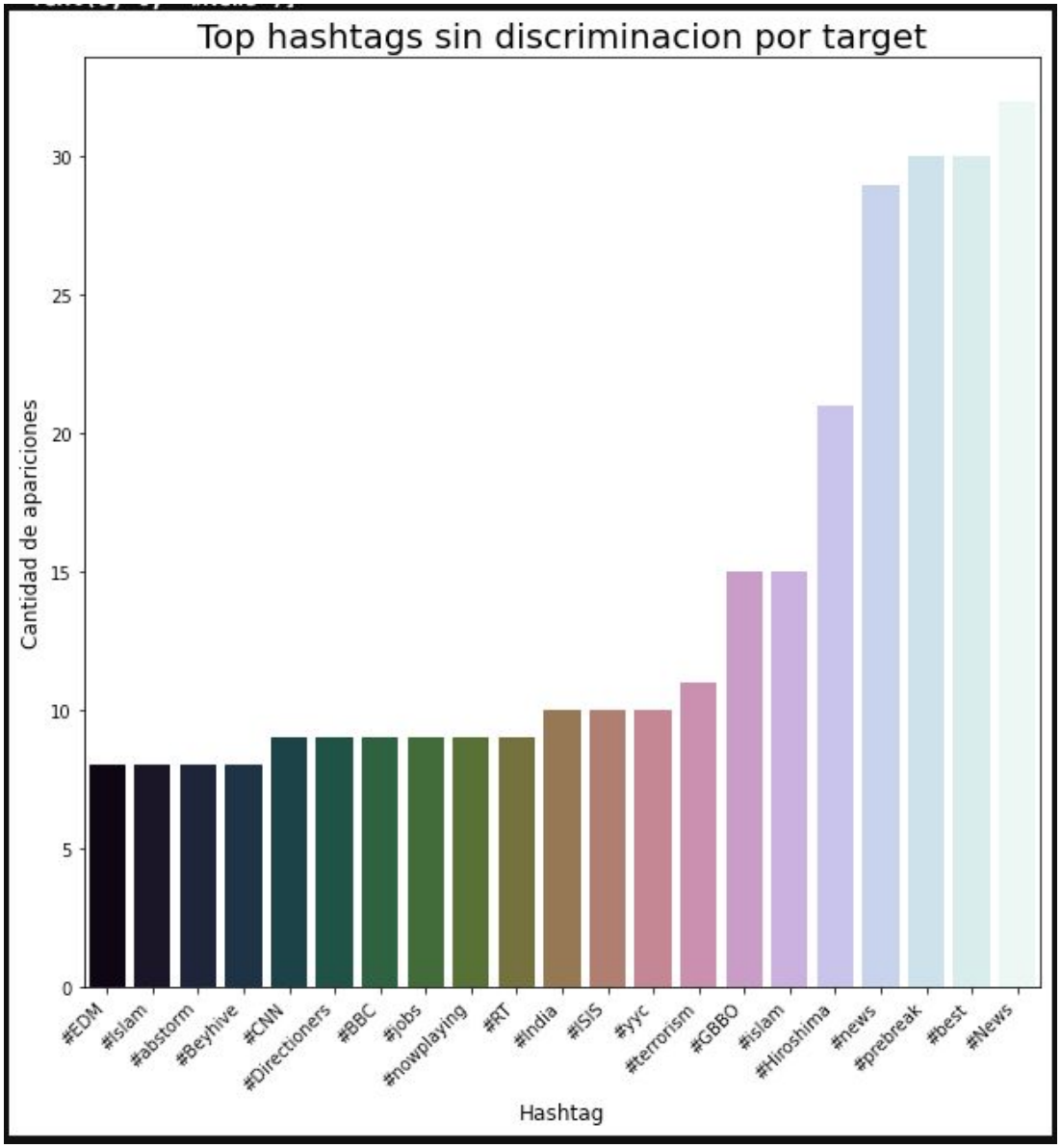


Figura 14

Pero… ¿qué pasa si los separamos por target?

- Target 0

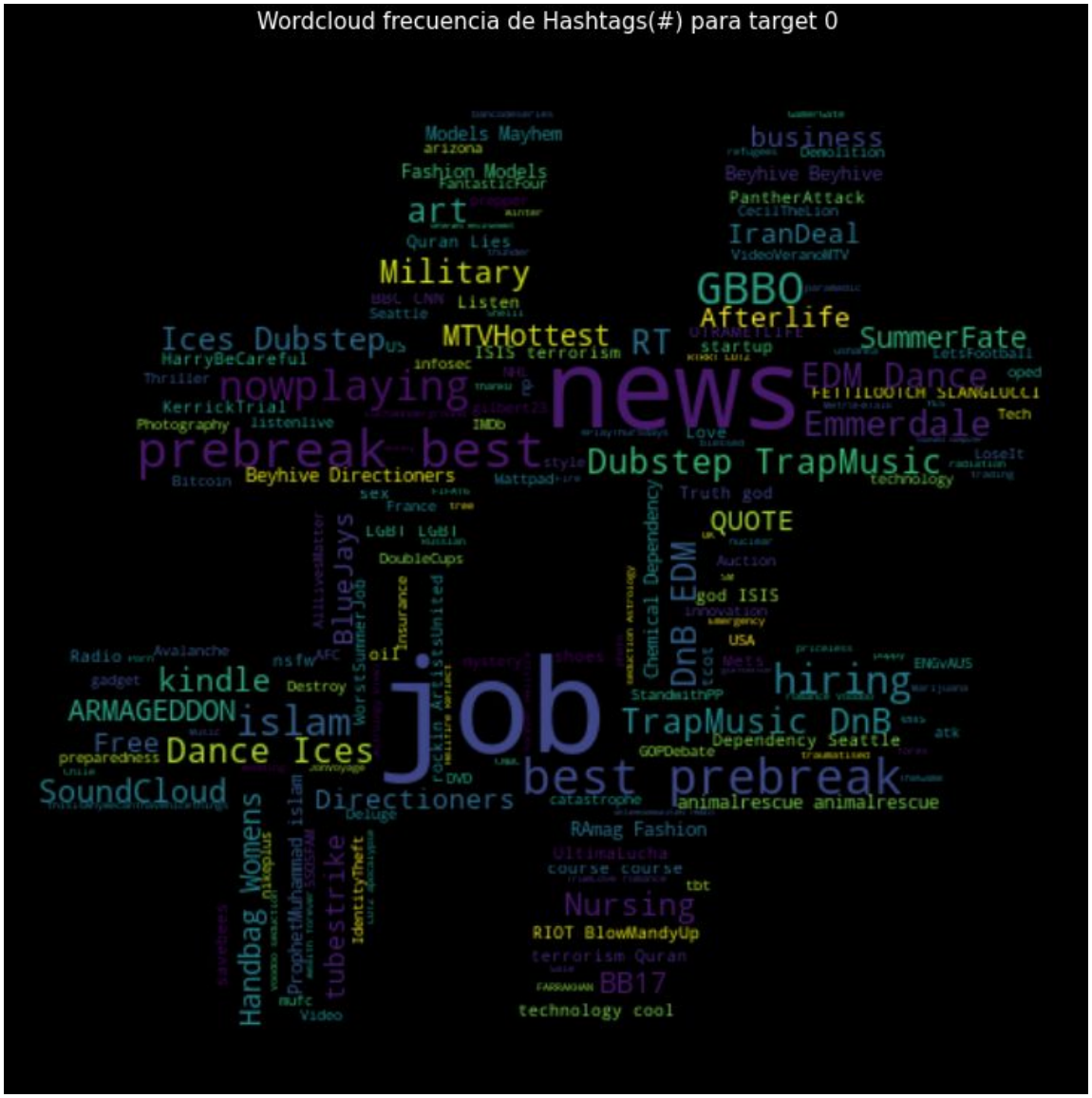


Figura 15

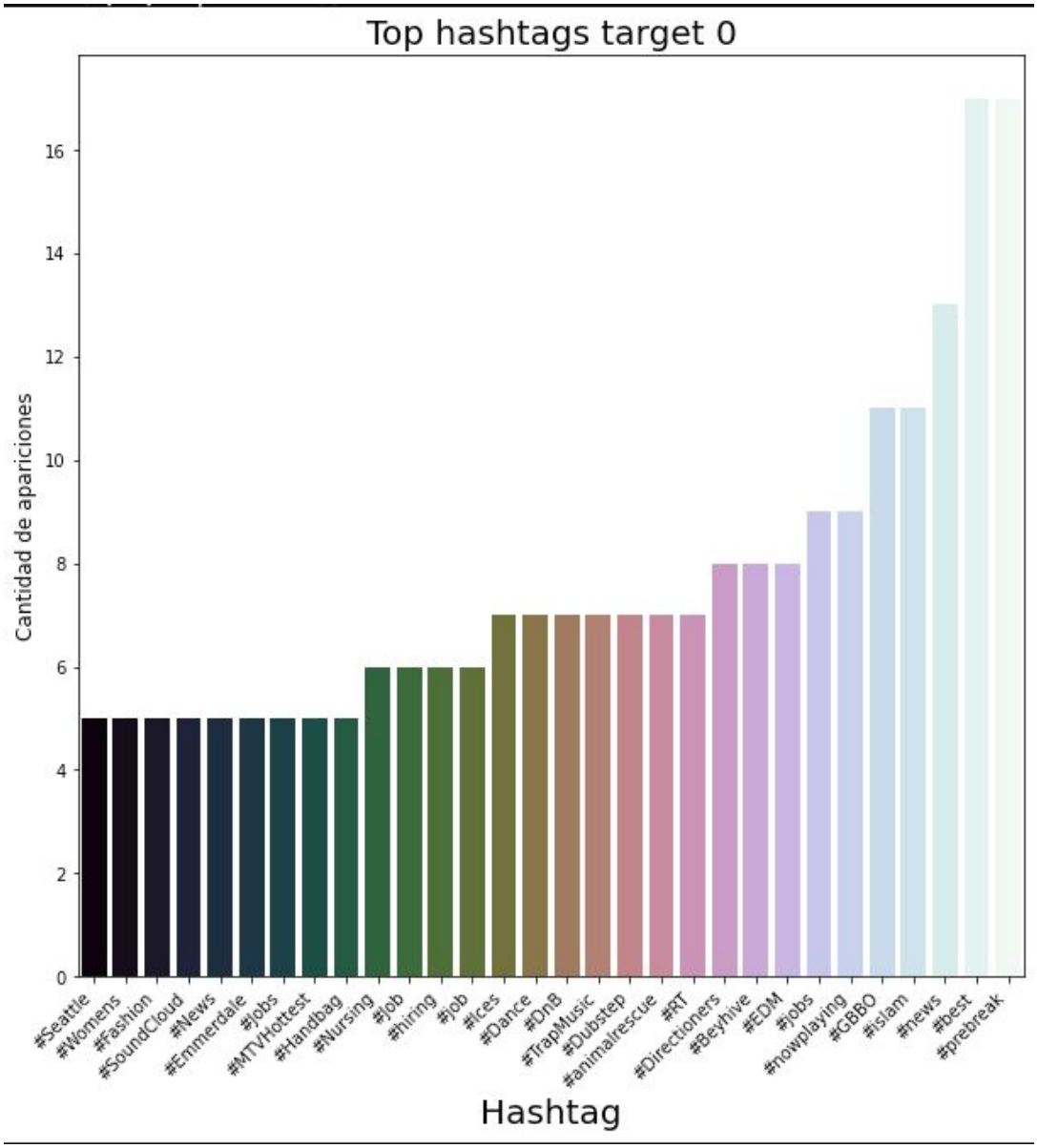

Top hashtags target 0
Cantidad de apariciones
0
2
4
6
8
10
12
14
16
#Seattle
#Womens
#Fashion
#SoundCloud
#News
#Emmerdale
#jobs
#MTVHottest
#Handbag
#Nursing
#job
#hiring
#job
#ices
#Dance
#DnB
#TrapMusic
#Dubstep
#animalrescue
#RT
#Directioners
#Beyhive
#EDM
#jobs
#nowplaying
#GBBO
#islam
#news
#best
#prebreak
Hashtag

Figura 16

- Target 1

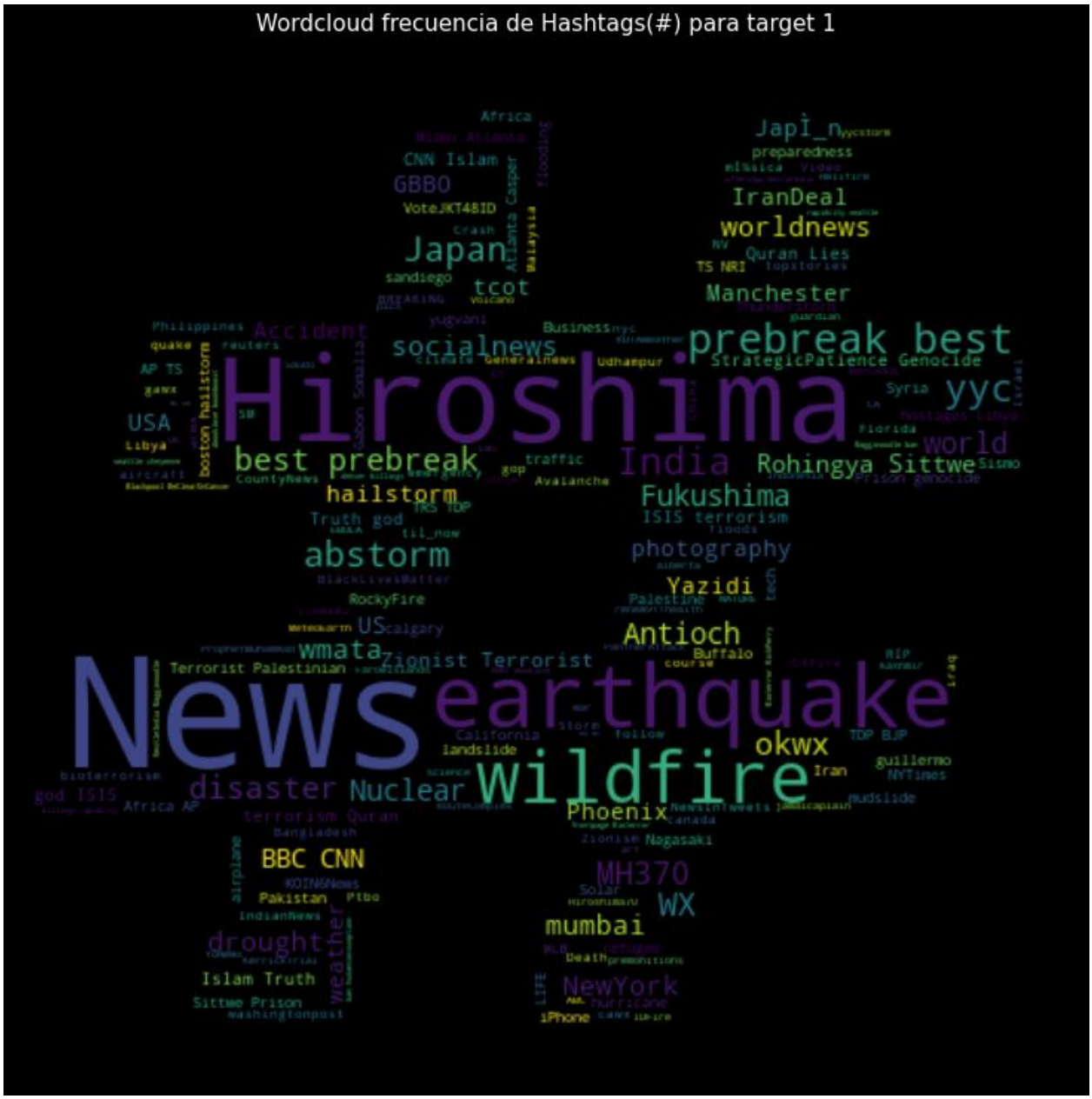


Figura 17

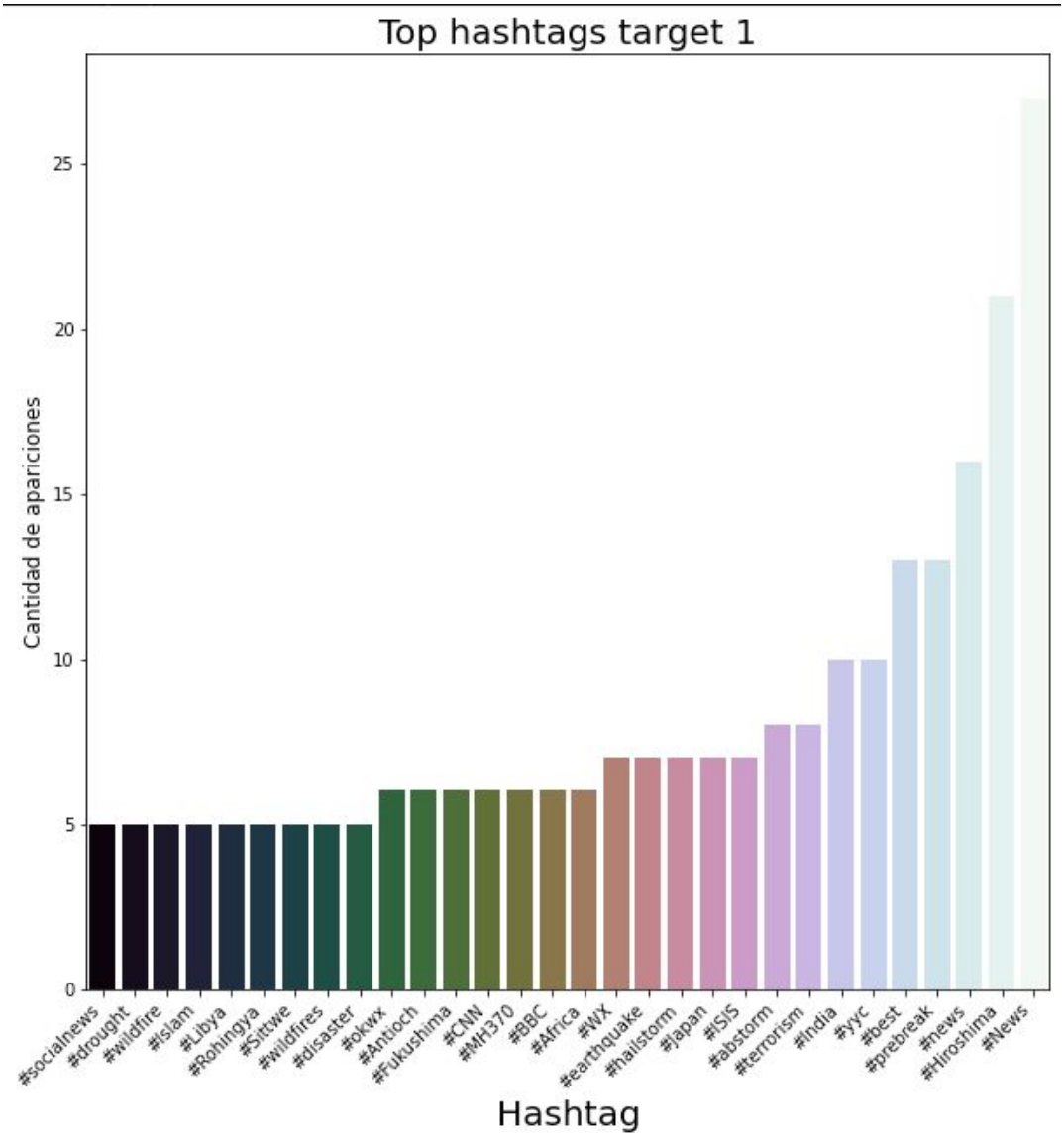


Figura 18

Al separar los targets podemos ver algunos hashtags en común y algunos otros que se separan según corresponda.
Por ejemplo tanto “news” como ”prebreak” son importantes en ambos targets pero para el target 0 podemos destacar por ejemplo “jobs”, “islam” y “GBBO” mientras que para el target 1 podemos destacar por ejemplo “Hiroshima” y “earthquake”.
Mirándolo de forma global no existe una gran diferenciación, ya que para ambos targets los hashtags tienen relación con el tema desastres.

De los hashtags más utilizados (count>9) ¿qué porcentaje de desastre o no desastre poseen?

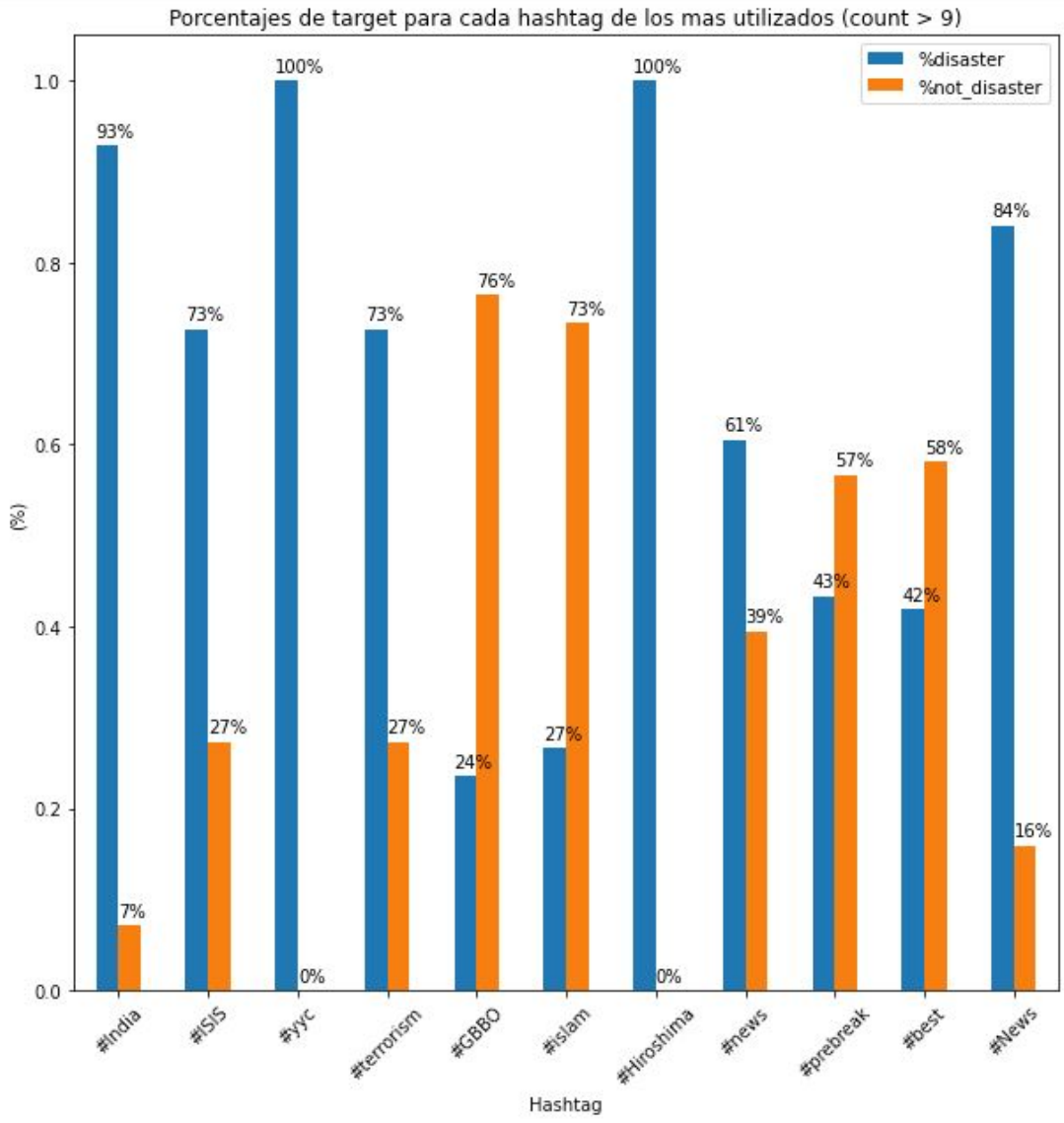


Figura 19

Podemos ver hashtags con un alto grado de posibilidad de que sean utilizados en tweets acerca de desastres cómo #india (región muy afectada por desastres naturales), #ISIS, #Hiroshima, #terrorism.

Por el otro lado tenemos hashtags con bajo grado de posibilidad de que sean utilizados para desastres cómo #GBBO (The Great British Bake Off) e #islam.

Podemos observar que #GBBO a pesar de que se trata de una tendencia sobre un programa

televisivo cuenta con casi un 25% de tweets que son desastres… muy alto en base a lo antedicho, esto es debido a que el tamaño de la muestra es de solo 17 registros.

Etiquetas

Las etiquetas tienen el mismo concepto que los hashtags pero suelen referir a personas, organizaciones o entidades. En el siguiente gráfico podemos ver las etiquetas más referidas en el campo de texto.

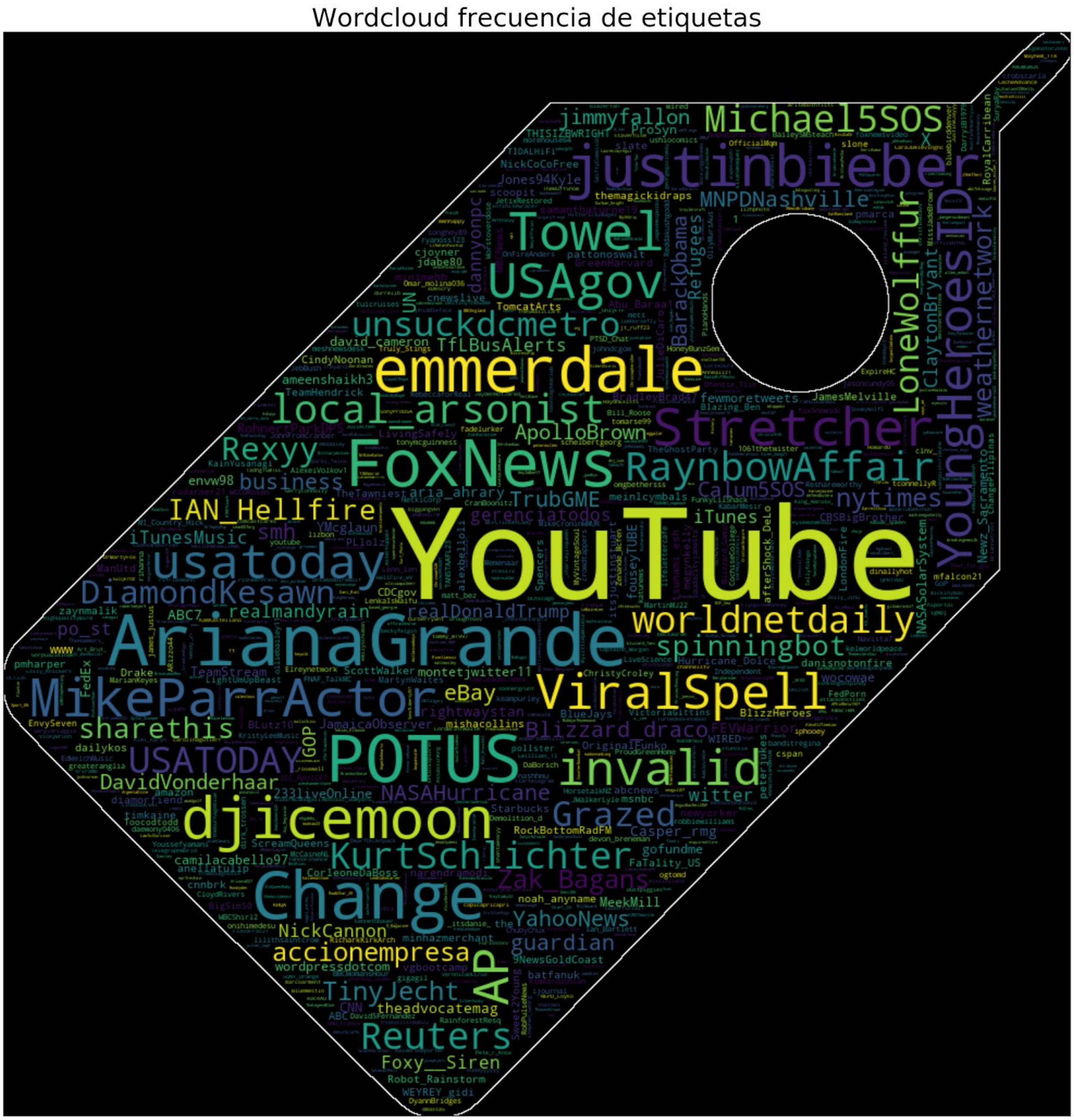


Figura 20

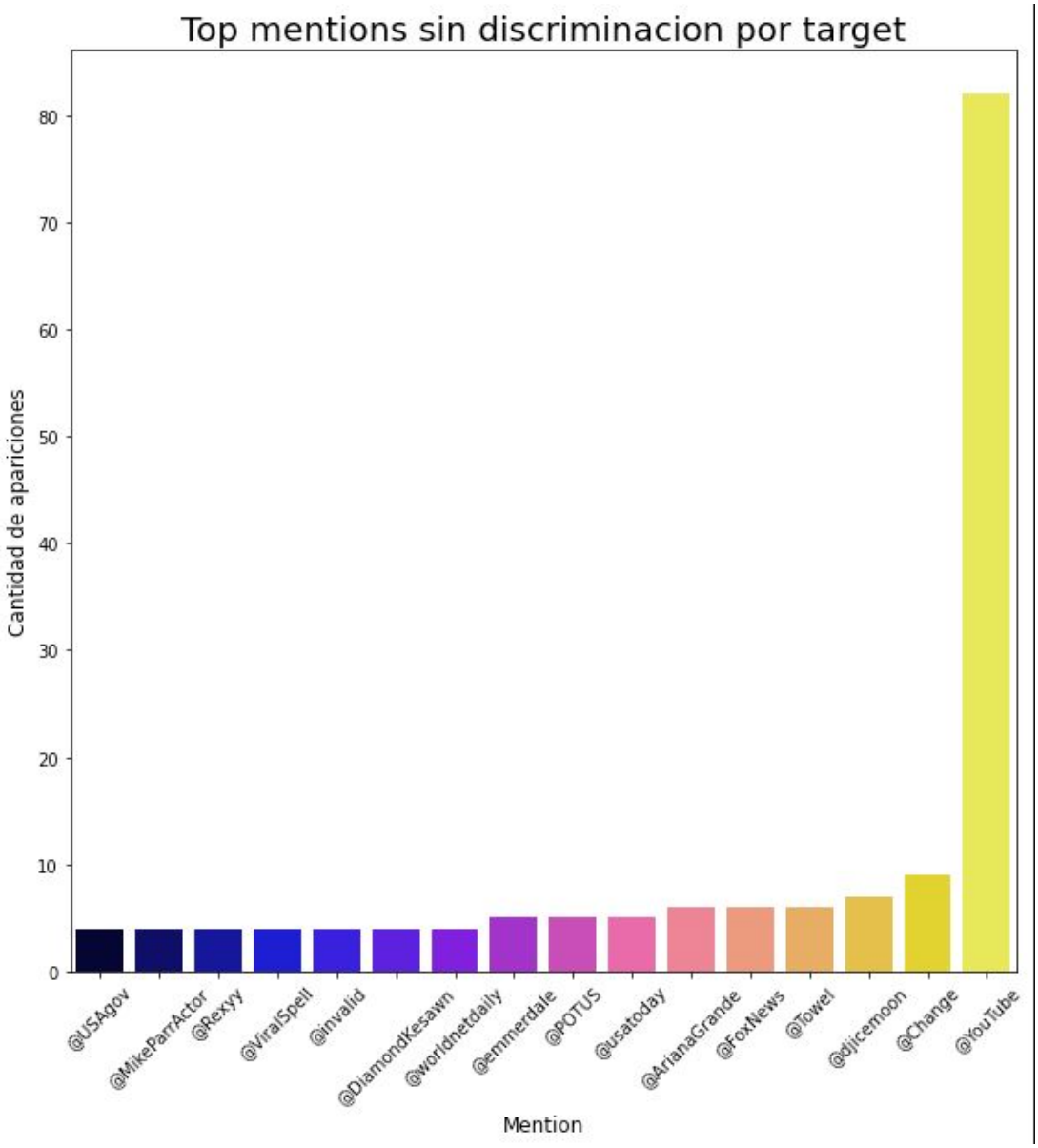


Figura 21

Pero… ¿qué pasa si los separamos por target?

- Target 0

Figura 22

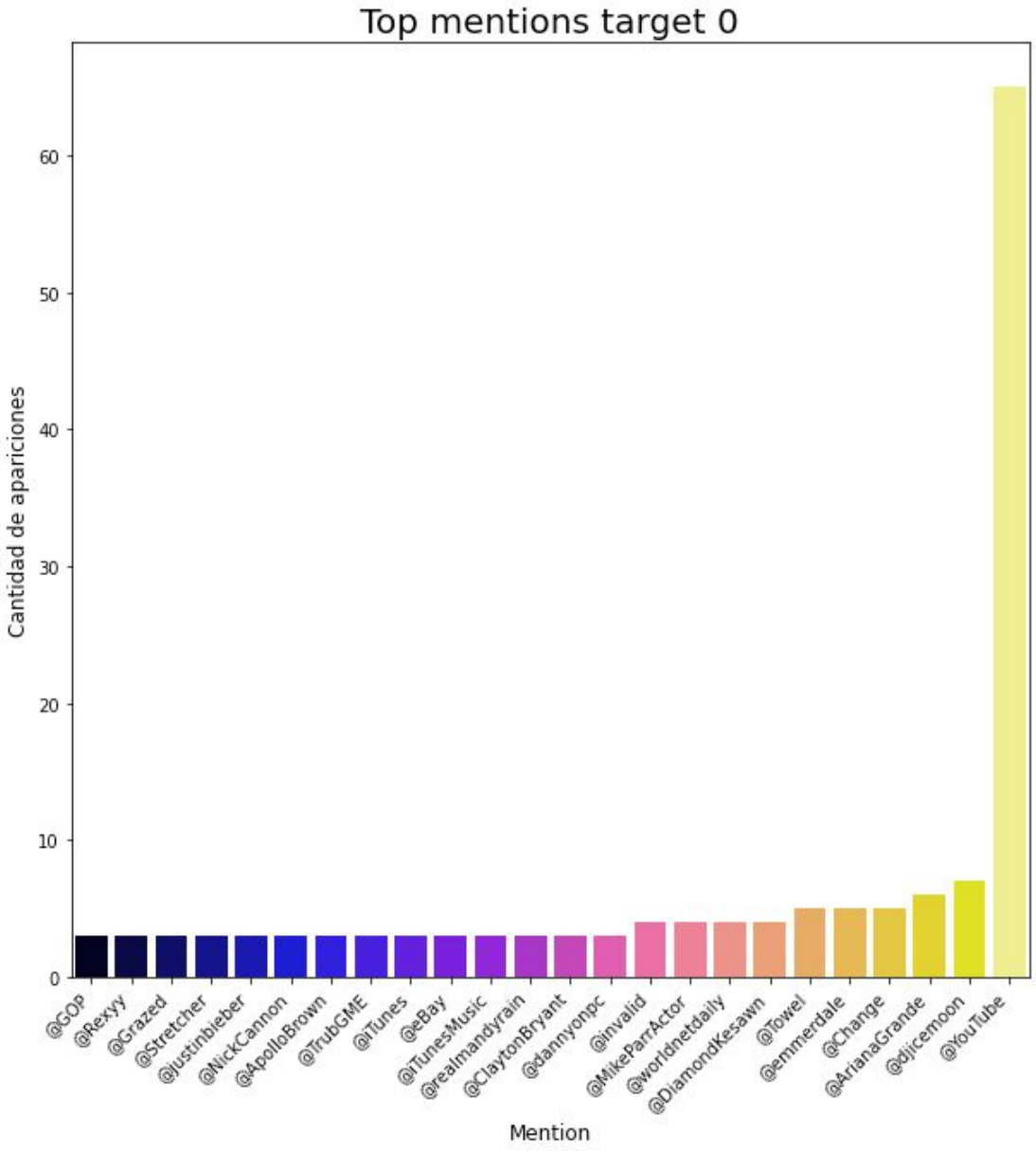

Top mentions target 0
Cantidad de apariciones
0
10
20
30
40
50
60
@GOP
@Rexyy
@Grazed
@Stretcher
@justinbieber
@NickCannon
@ApolloBrown
@TrubGME
@iTunes
@eBay
@iTunesMusic
@realmandyrain
@ClaytonBryant
@dannyonpc
@invalid
@MikeParrActor
@worldnetdaily
@DiamondKesawn
@Towel
@emmerdale
@Change
@ArianaGrande
@djicemoon
@YouTube
Mention

Figura 23

- Target 1

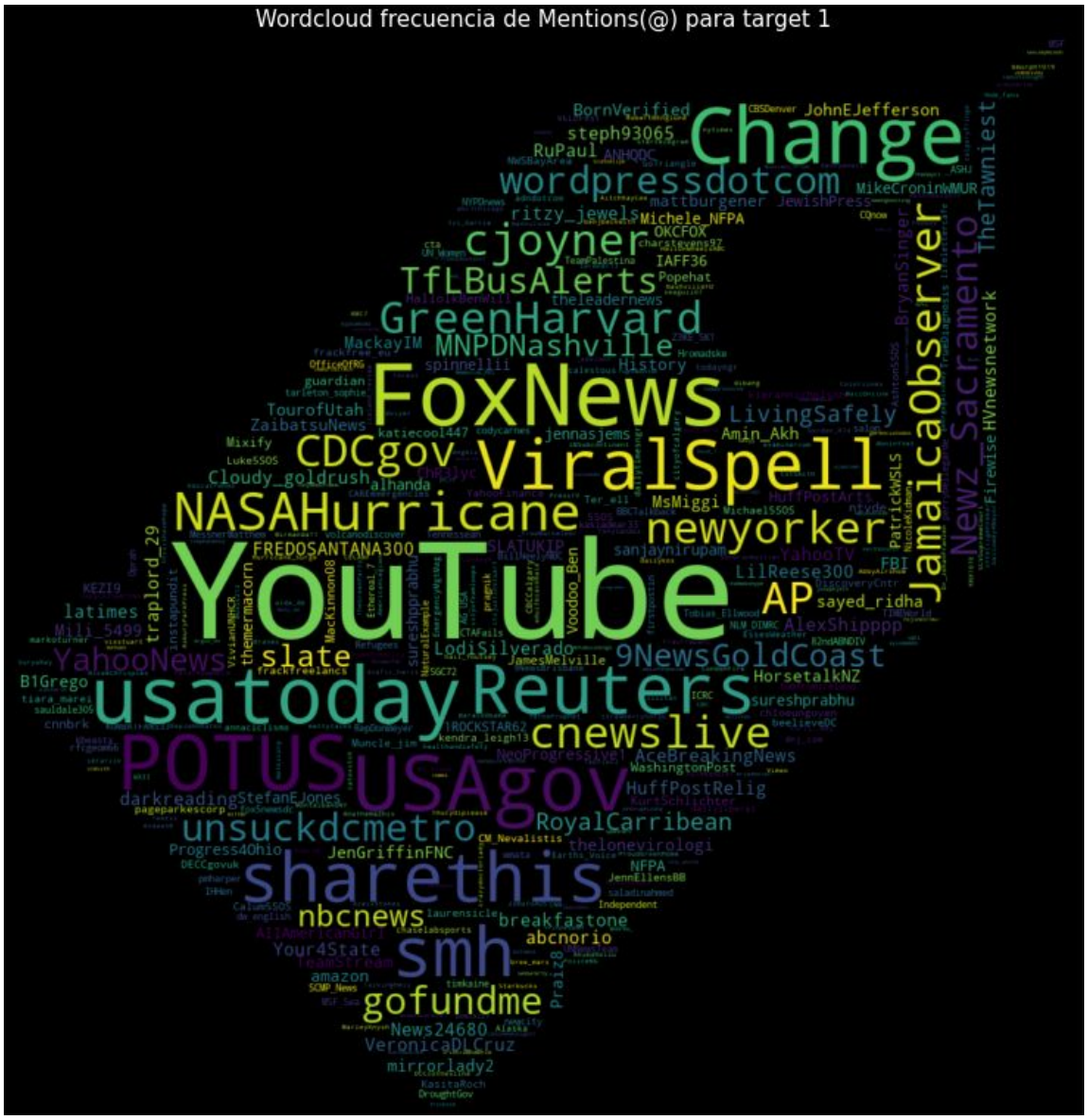


Figura 24

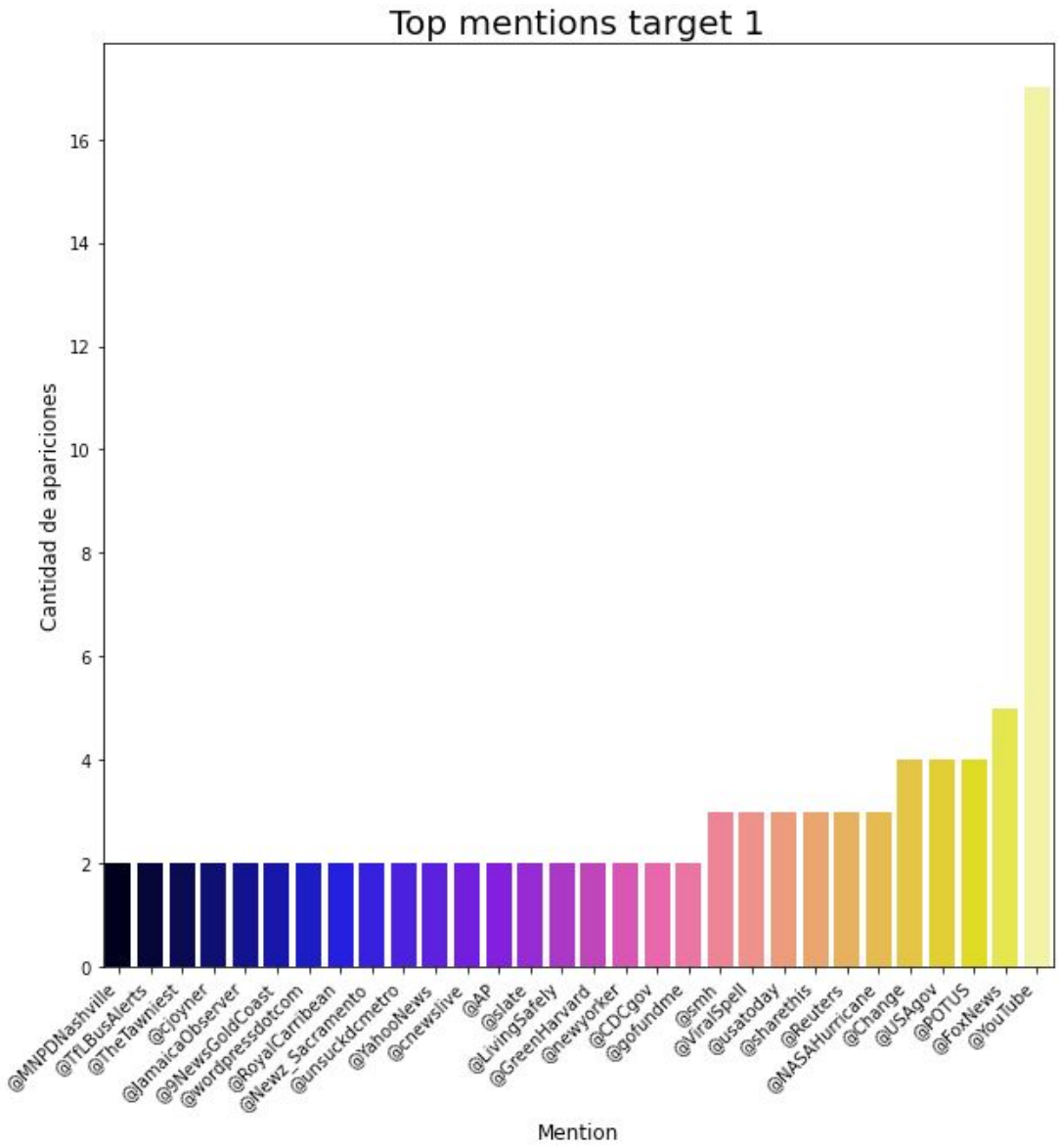


Figura 25

Entonces al separar los targets podemos notar que “Youtube” es clave para ambos y que para el target 0 podemos destacar por ejemplo “ArianaGrande” y “djicemoon” mientras que para el target 1 podemos destacar por ejemplo “FoxNews” y “POTUS”.
En este caso sí podemos ver una diferenciación ya que por un lado se ven celebridades y por el otro se habla de noticias y el presidente de los EEUU, posiblemente enfocado entonces a desastres en este último caso.

Links

Las referencias a otros links se utilizan generalmente para ampliar sobre el tema en una página exterior al tweet por lo cual veamos cuáles fueron los más referidos en este caso.

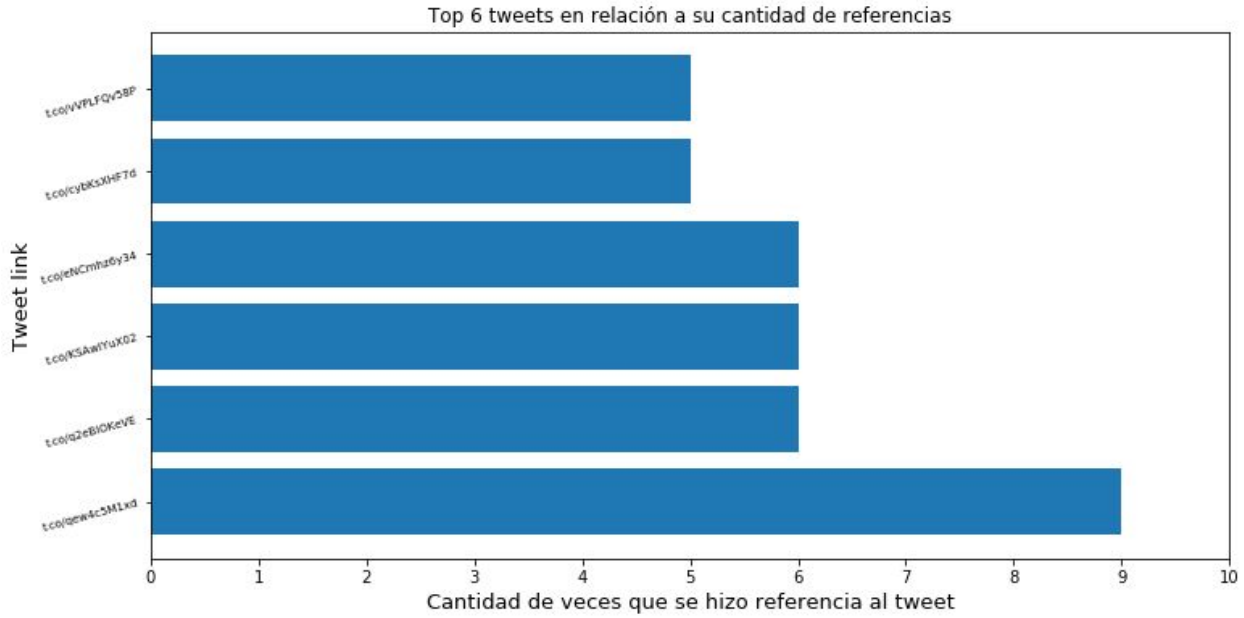


Figura 26

A continuación se muestran los links que aún están disponibles, son interesantes y están relacionados con los desastres:

Link más referido: Cómo prepararse para un tornado

Video de tsunamis y terremotos en youtube: Tsunamis y terremotos

Offtopic: Fotos de busqueda del llamado Escuadrón 731 en coreano

Nuevos features a partir de la variable texto

Se buscó crear nuevos features con el objetivo de analizar la variable texto en mayor profundidad.

Caracteres

En este apartado analizamos la longitud de caracteres de cada campo de texto.

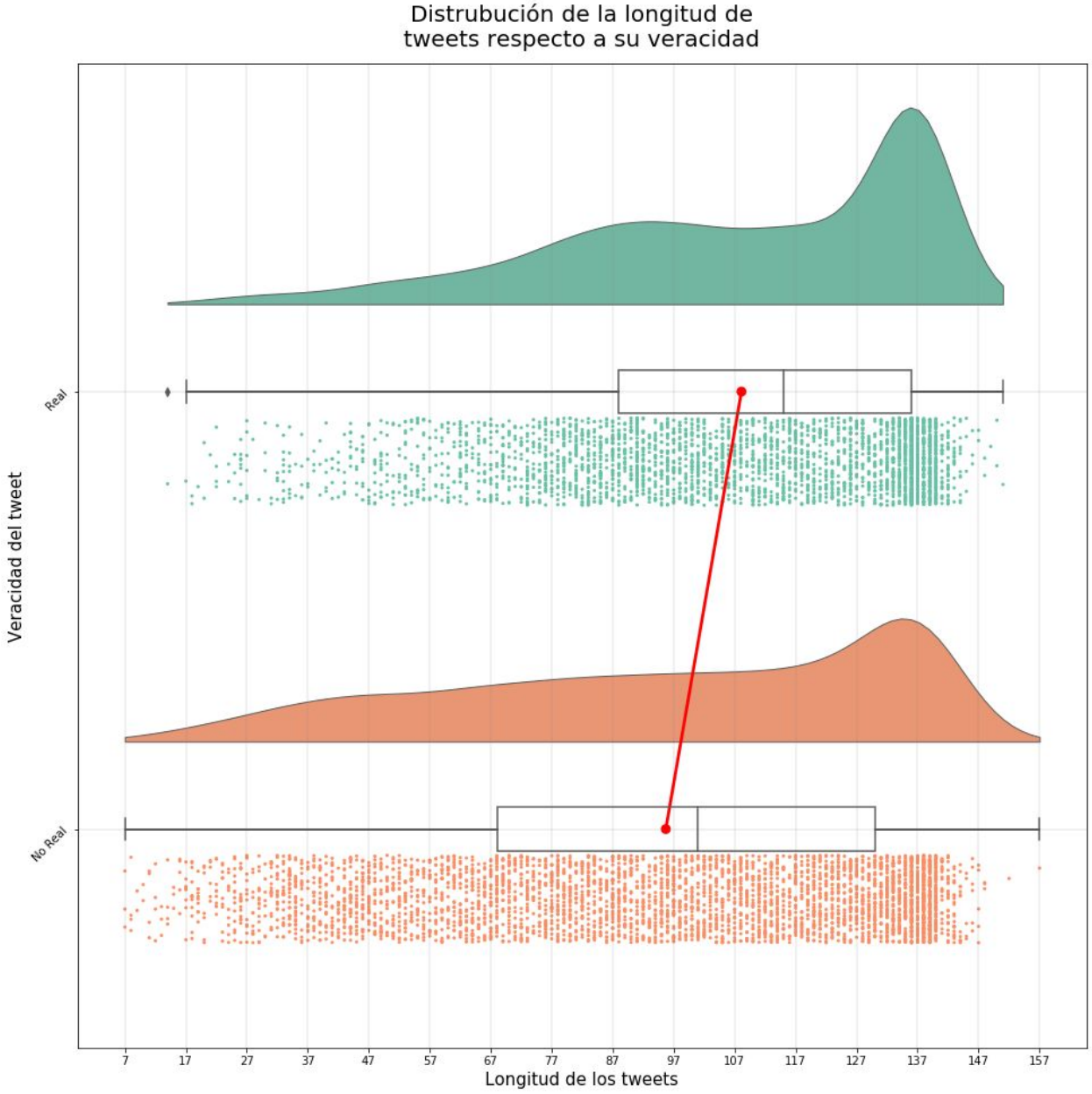


Figura 27

Podemos ver que la longitud de los textos se distribuye levemente hacia la izquierda lo que quiere decir que en general la mayoría de los tweets son largos y hay una menor cantidad de tweets cortos. El promedio de la longitud los tweets anda alrededor de 90 y 100 según el target.

Palabras

En este apartado analizamos la cantidad de apariciones de cada palabra para ver cuales son aquellas que resaltan en los diferentes tweets.

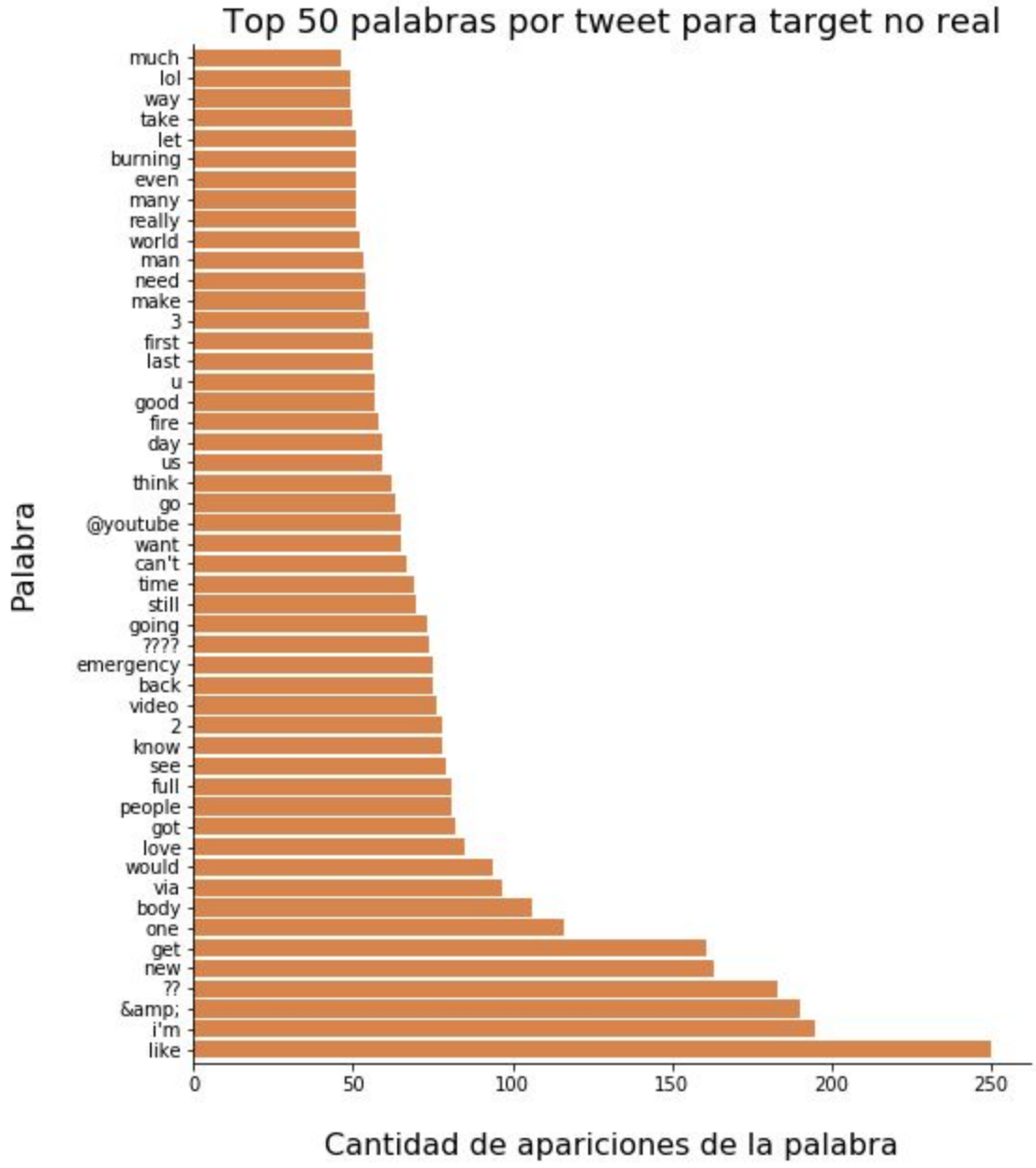


Figura 28

Para el caso de los desastres no reales destacamos la palabra “like” como aquella que tiene más apariciones pero las importantes son aquellas relacionadas con el tema como por ejemplo “emergency”, “body” o “fire”. También se puede ver la aparición de palabras que se pueden mejorar para mayor claridad como por ejemplo reemplazar “??” por “?” o remover/reemplazar “&” qué es el símbolo & para HTML.

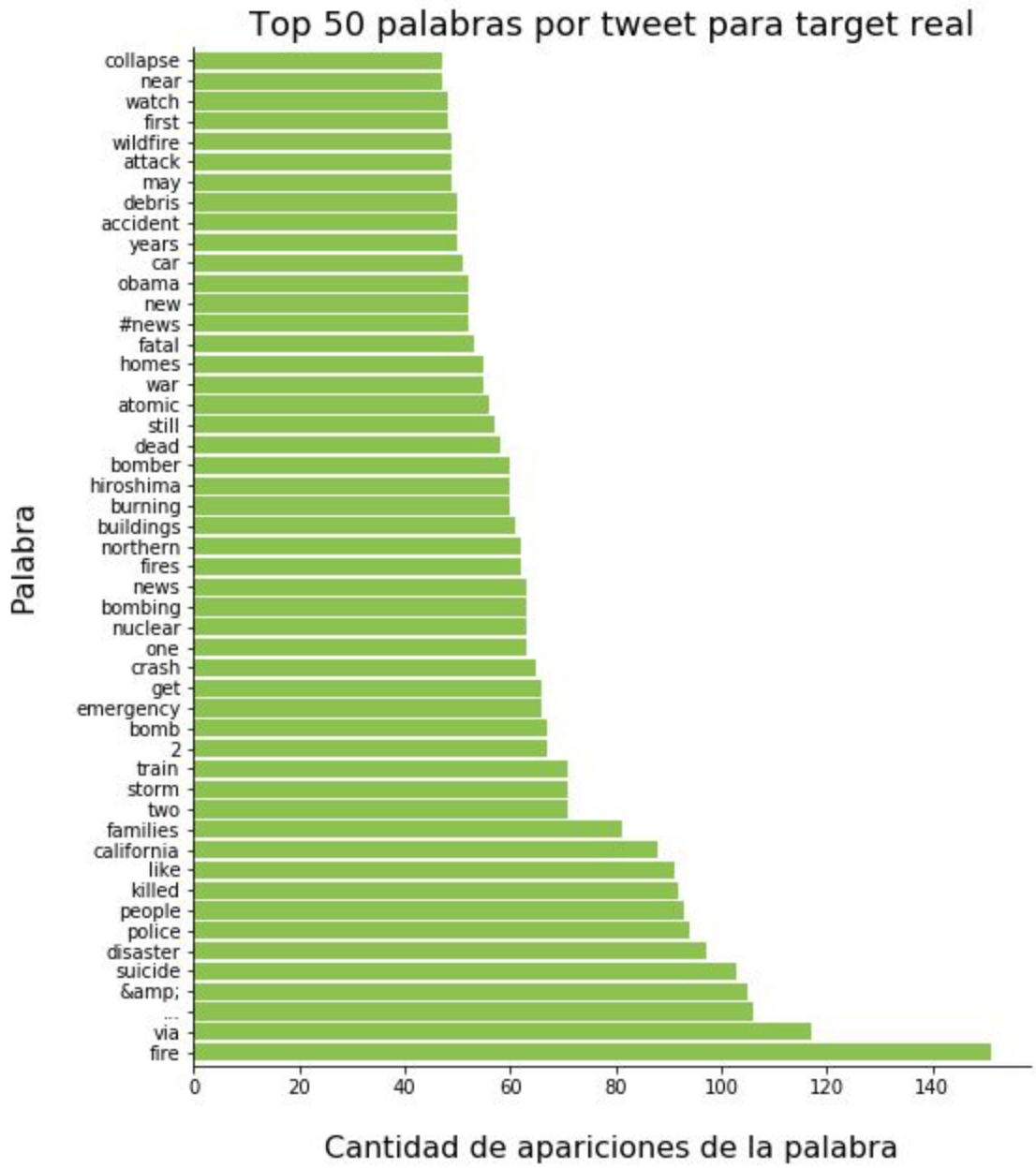


Figura 29

Para el caso de los desastres reales se destaca la palabra “fire” como aquella con mayor apariciones como también podemos ver caracteres que se pueden remover/reemplazar.
Lo importante aquí es ver que hay una mayor cantidad de palabras relacionadas con desastres lo cual es consistente con lo que se esperaba ver.
Habrá que tener cuidado entonces con aquellas palabras relacionadas al tema que aparezcan en ambos targets al momento de predecir.

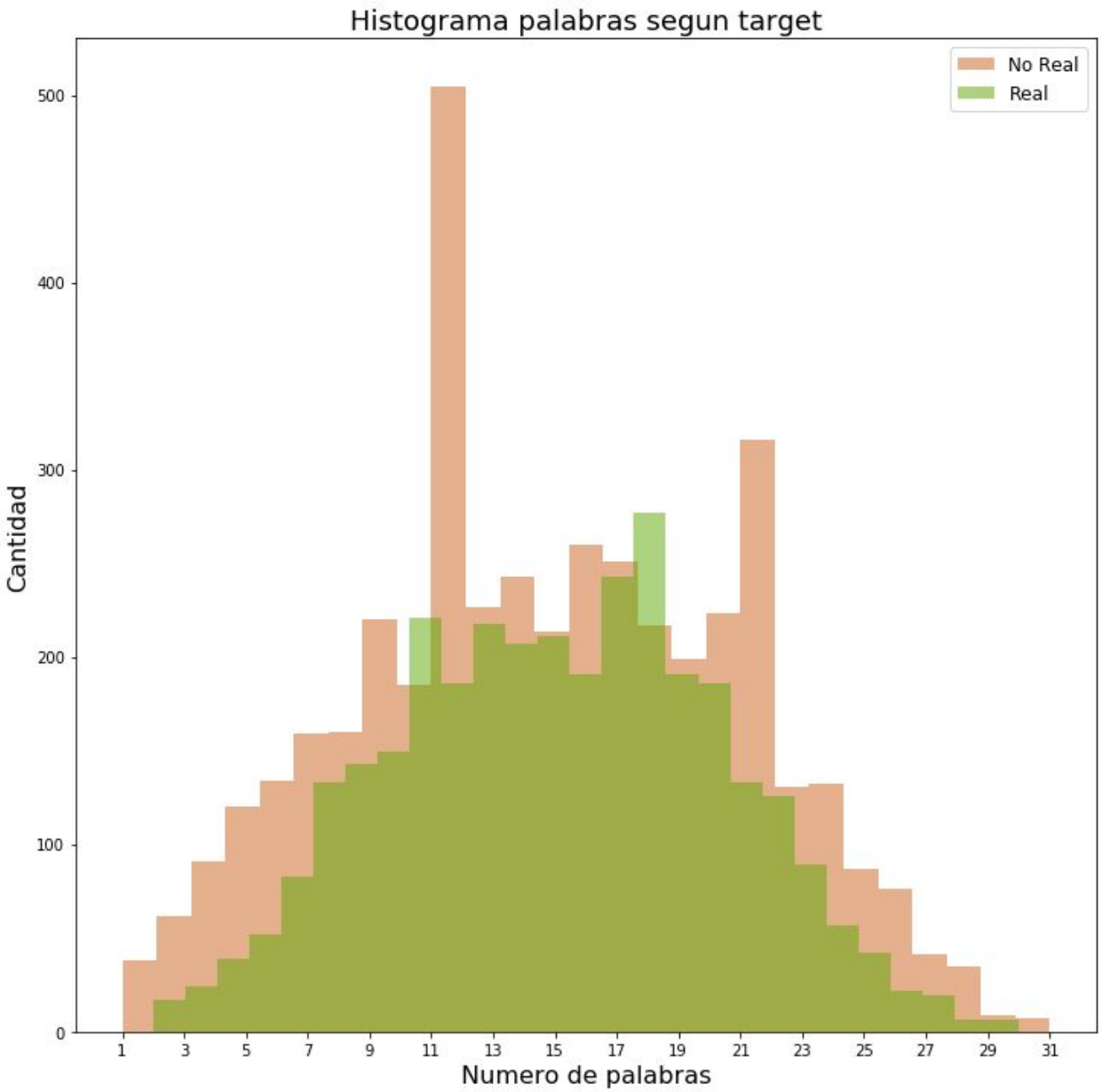


Figura 30

En cuanto a la distribución podemos ver que se distribuye normalmente aunque hay algunas irregularidades en el caso del target no real por la aparición de valores atípicos, principalmente en 11 palabras.

Stopwords

Las llamadas stopwords refieren generalmente a las palabras más comunes en un idioma, en este caso en inglés. Lo que vemos a continuación es cuáles fueron las stopwords más frecuentes en todos los tweets y su cantidad según si era real o no.

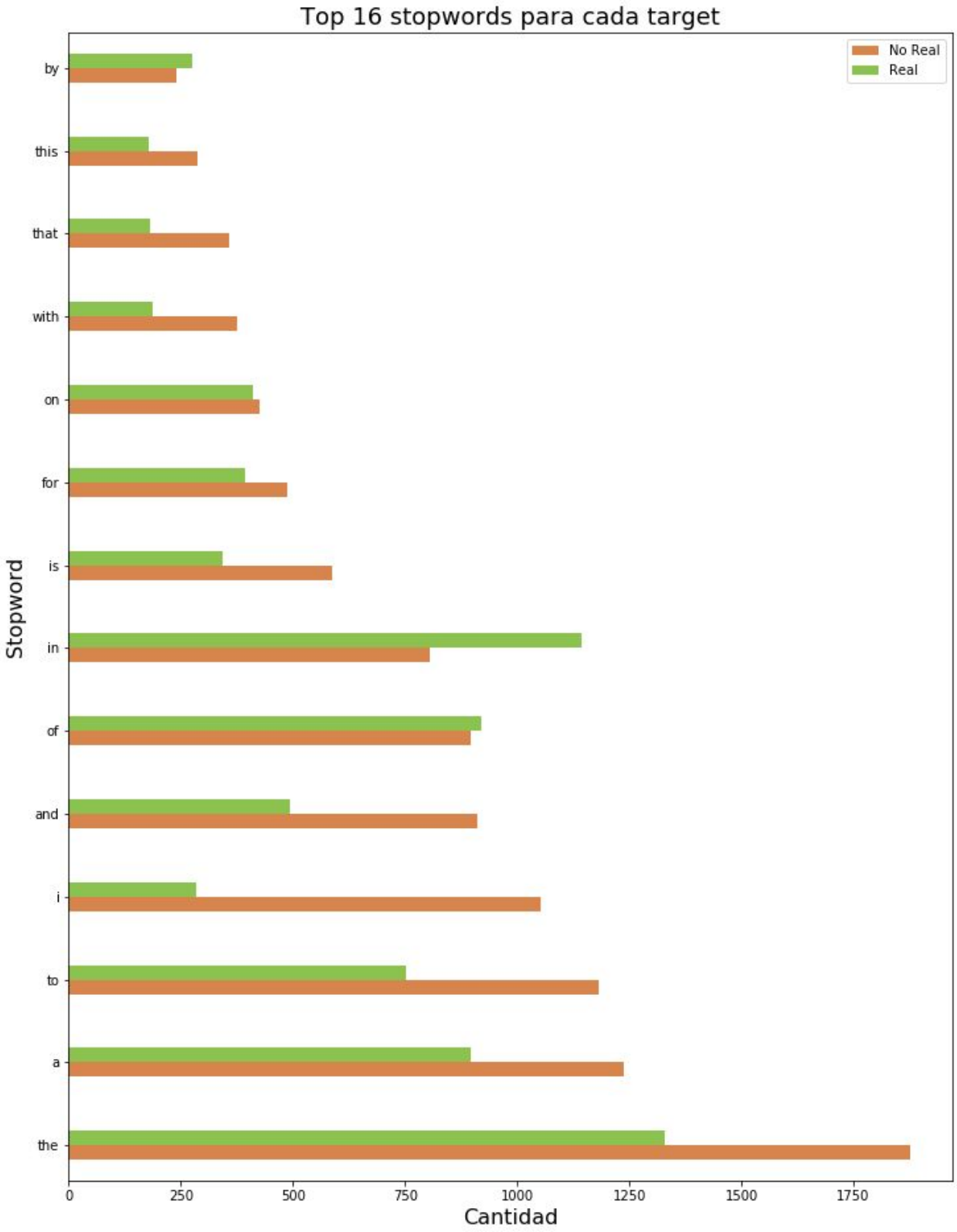

Top 16 stopwords para cada target
No Real
Real
Stopword
by
this
that
with
on
for
is
in
of
and
i
to
a
the
0
250
500
750
1000
1250
1500
1750
Cantidad

Figura 31

En este gráfico podemos ver que las stopwords son más predominantes cuando el desastre no es real ya que en general hay una mayor cantidad de ellas, como también podemos ver cual son las más frecuentes como por ejemplo “the” y “to”

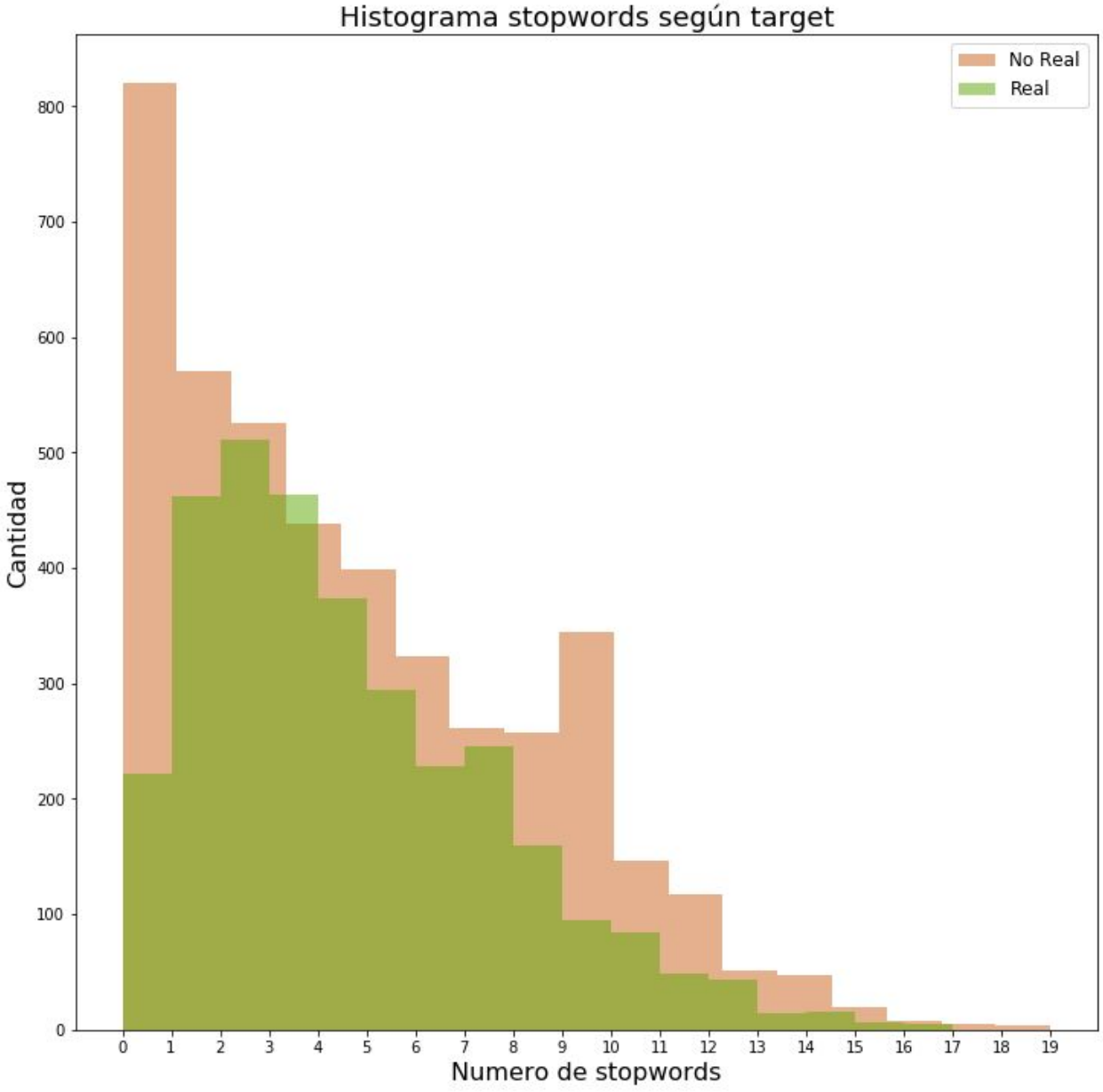


Figura 32

Aquí podemos ver que tanto para desastres reales como no reales la distribución de las stopwords presenta right-skewing, los datos sesgados a la derecha tienen algunos valores grandes que impulsan la media hacia arriba, pero no afectan el lugar exacto de los datos.
Es lógico que se distribuya de esta manera ya que en un texto uno tiende a utilizar algunas stopwords pero no son el foco del texto, por lo cual hay una gran cantidad de valores con pocas apariciones de stopwords en los tweets.

Símbolos de puntuación

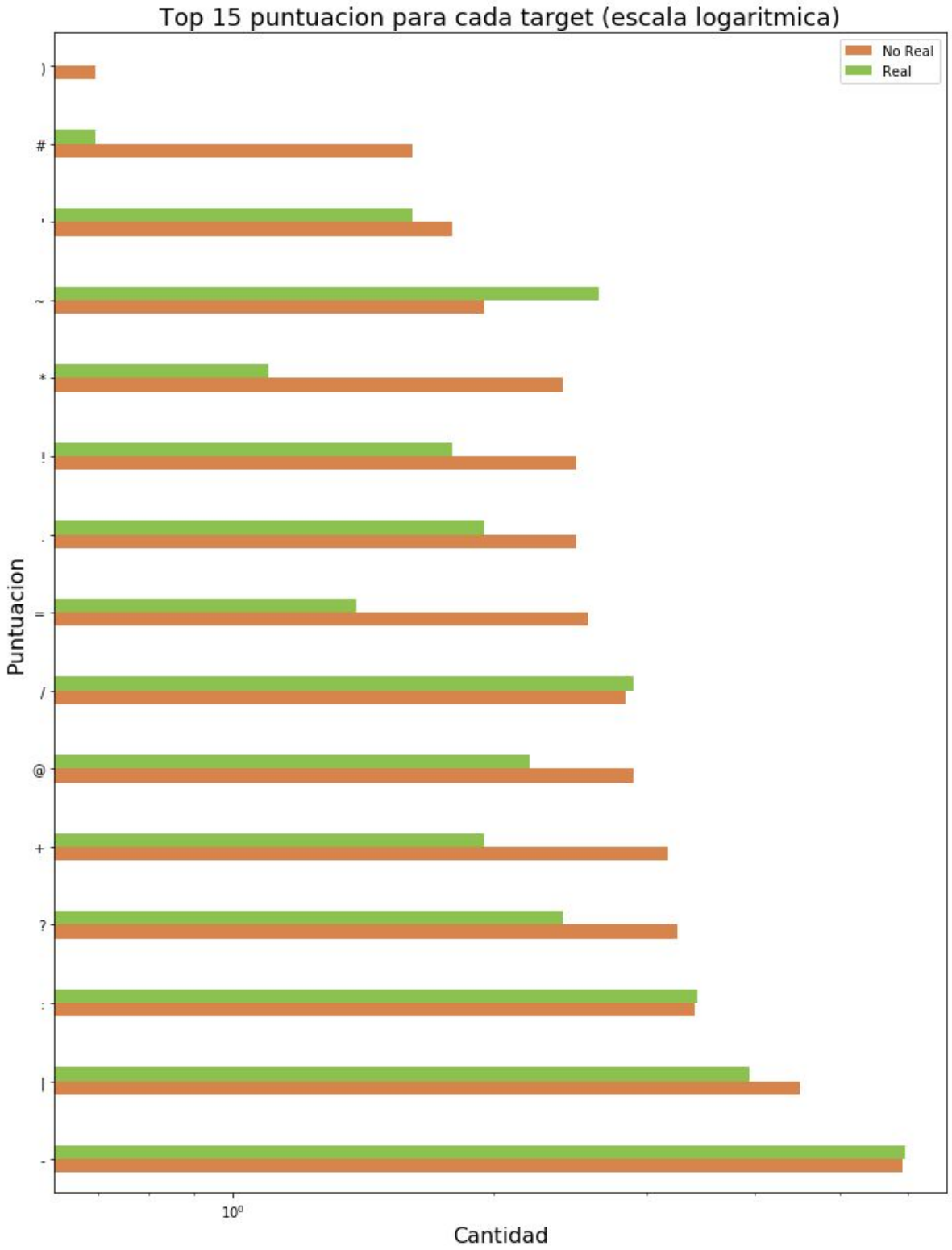


Figura 33

En este gráfico no hay una tendencia general en relación al target, pero sí nótese la escala logarítmica aplicada para poder mostrar algunos valores ya que por ejemplo el símbolo “-” aparece predominante ante los demás y si no se aplicase la escala el gráfico no se apreciaría.

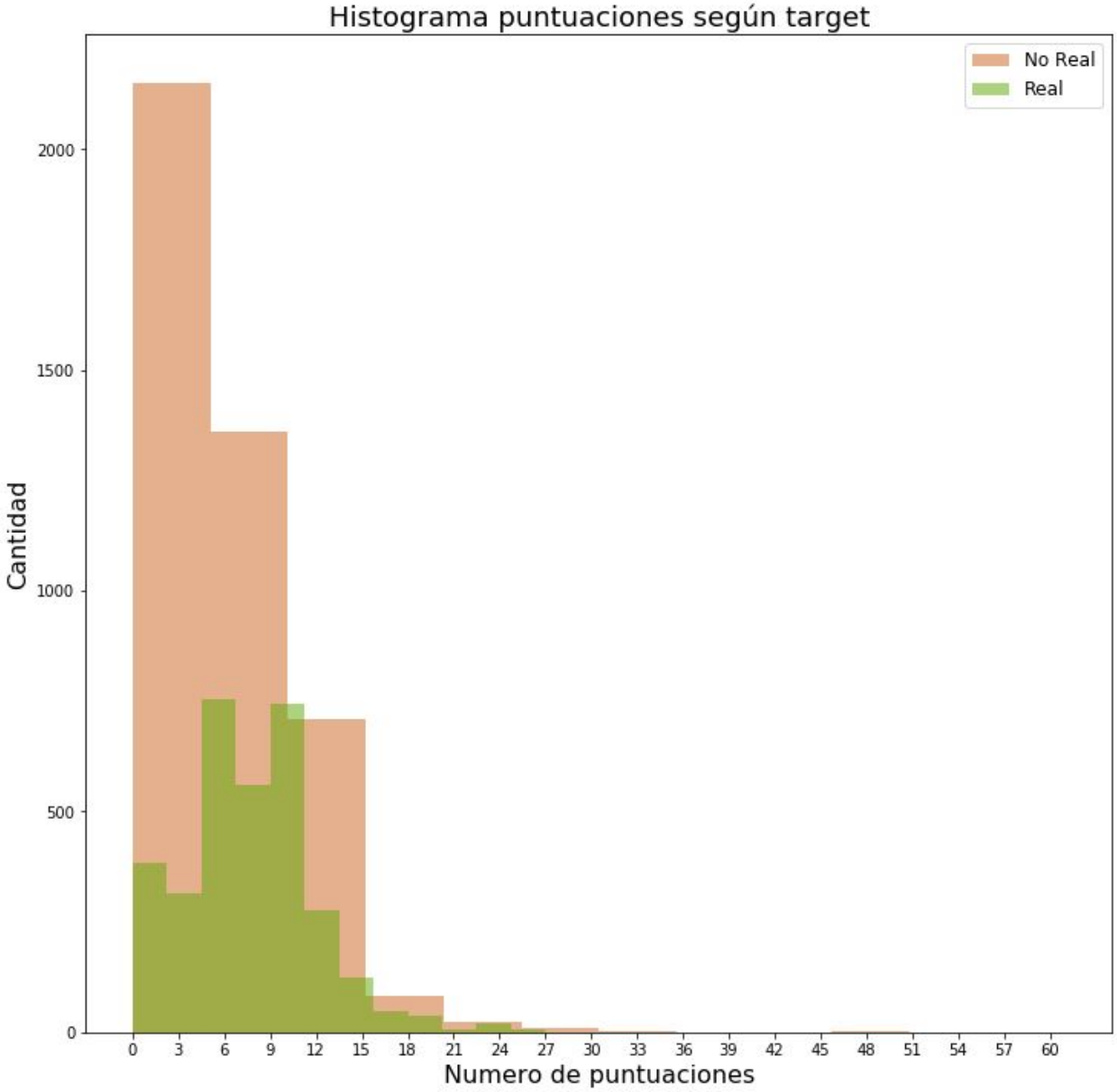


Figura 34

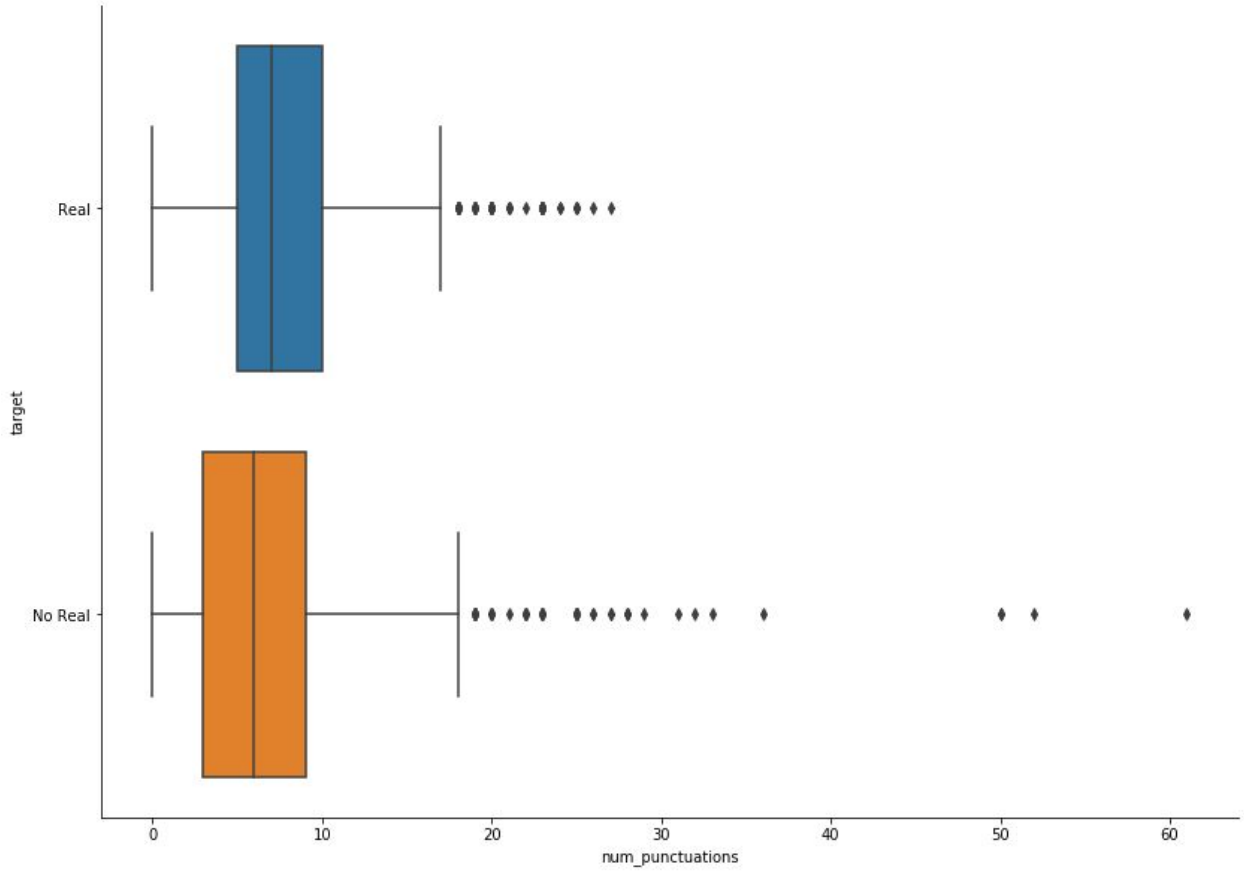


Figura 35

En este caso el right skewing es aún más notorio como también se nota que hay una gran diferencia entre la cantidad de apariciones entre los dos target, ya que en los desastres no reales hay una cantidad mucho mayor de apariciones de símbolos de puntuación.
El análisis con respecto a la cantidad de puntuaciones es similar al hecho en las stopwords. En particular, se eligió añadir un boxplot ya que se puede ver que posee algunos outliers en este caso.

Palabras mayúsculas

En este apartado analizamos la cantidad de apariciones de cada palabra mayúscula para ver cuales son aquellas que resaltan en los diferentes tweets.

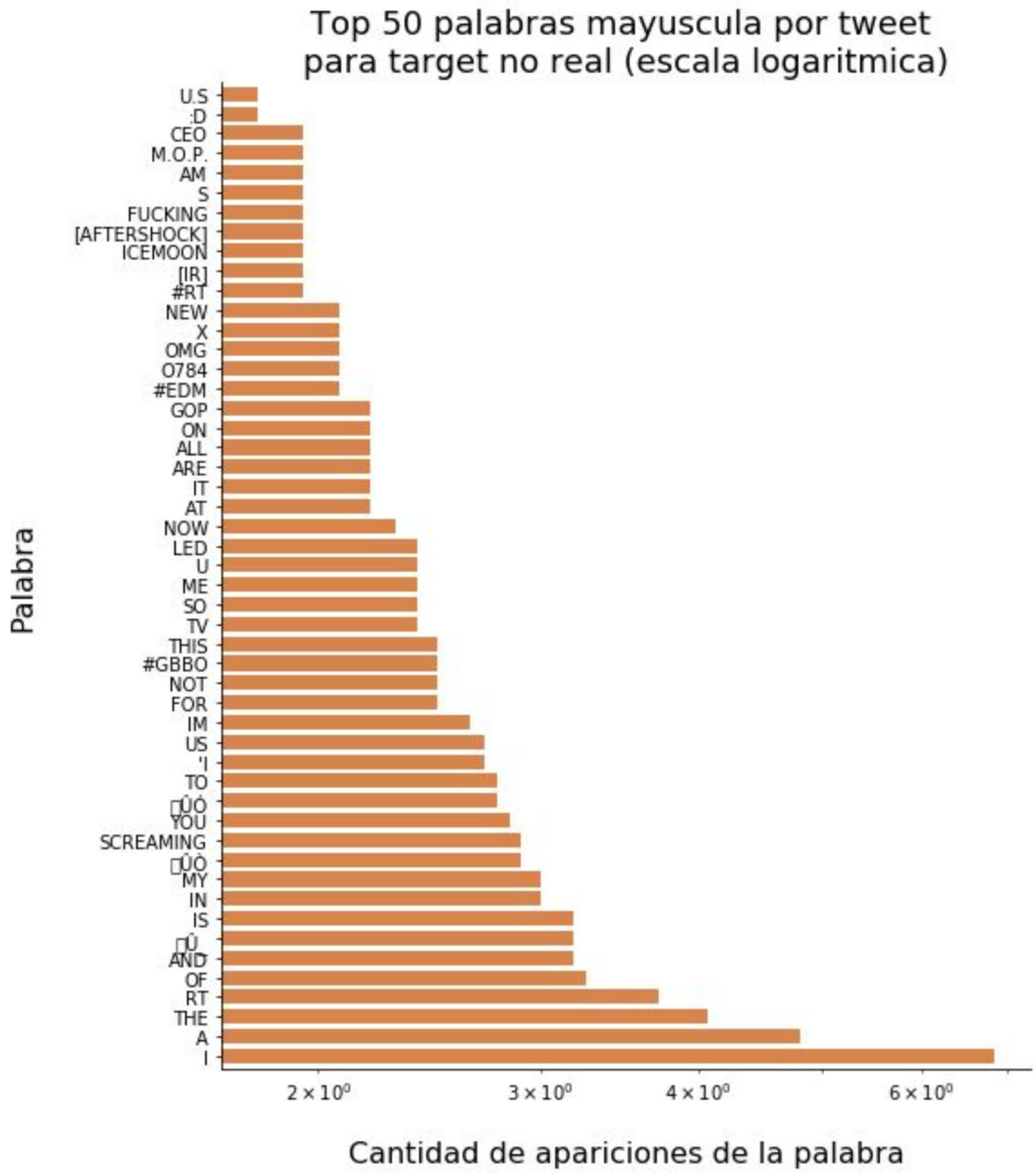

Top 50 palabras mayuscula por tweet
para target no real (escala logaritmica)
Palabra
U.S
:D
CEO
M.O.P.
AM
S
FUCKING
[AFTERSHOCK]
ICEMOON
[IR]
#RT
NEW
X
OMG
O784
#EDM
GOP
ON
ALL
ARE
IT
AT
NOW
LED
U
ME
SO
TV
THIS
#GBBO
NOT
FOR
IM
US
'I
TO
YOU
SCREAMING
MY
IN
IS
AND
OF
RT
THE
A
I
2 × 10^0
3 × 10^0
4 × 10^0
6 × 10^0
Cantidad de apariciones de la palabra

Figura 36

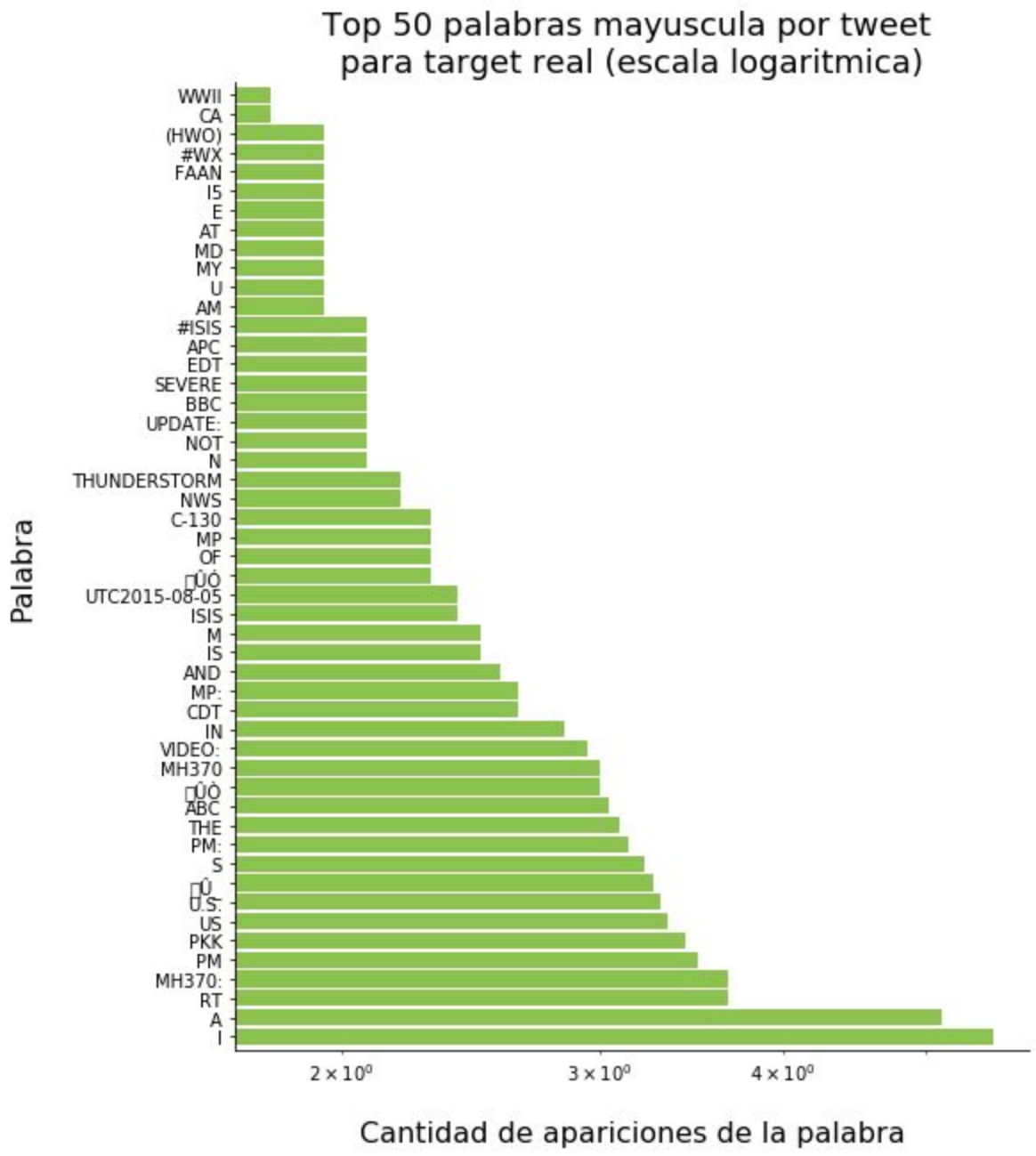


Figura 37

Para ambos target tenemos que se destacan las palabras “I”, “A”, “RT”. En el caso de la primera tiene sentido ya que es común hablar de uno mismo y en particular el “yo” en ingles no se puede escribir en minúsculas por regla, por eso le damos más importancia en este caso a las otras palabras. También destacamos la aparición de algunos hashtags/menciones que refieren a siglas de empresas/entidades

Luego podemos también ver que hay basura para eliminar/reemplazar en ambos casos.

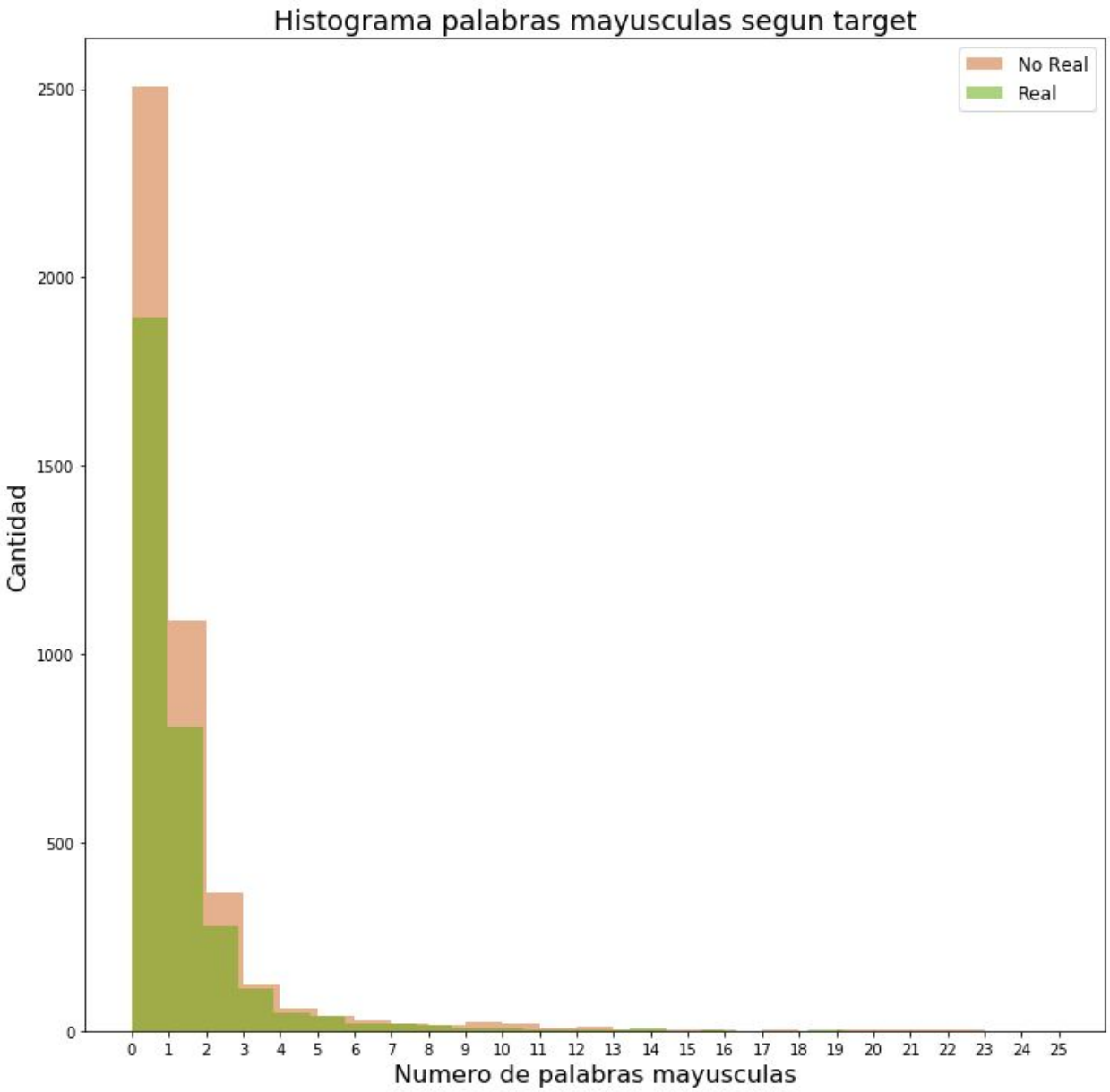

Histograma palabras mayusculas segun target
No Real
Real
Cantidad
2500
2000
1500
1000
500
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Numero de palabras mayusculas

Figura 38

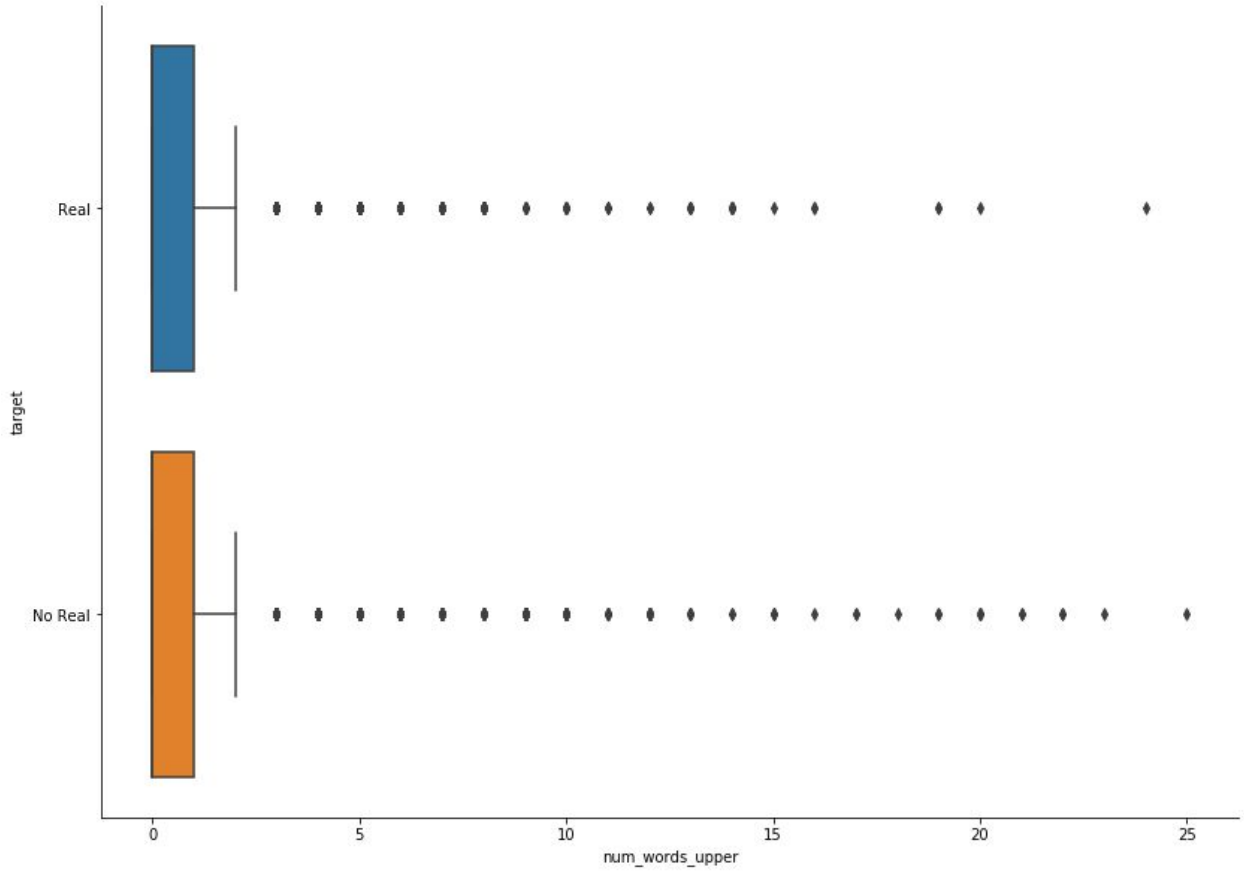


Figura 39

De todos los vistos hasta ahora las palabras mayúsculas deben ser de las que más presentan distribución right-skewed lo cual debe ser así porque no es muy común escribir en palabras mayúsculas y en el boxplot hay una gran cantidad de outliers que modifican a la distribución. Probablemente sean estos las menciones a entidades/organizaciones/países que tienen siglas y aparecen mayormente en mayúsculas.

Palabras con la primera letra en mayúscula

En este apartado analizamos la cantidad de apariciones de cada palabra con la primer letra en mayúscula para ver cuales son aquellas que resaltan en los diferentes tweets.

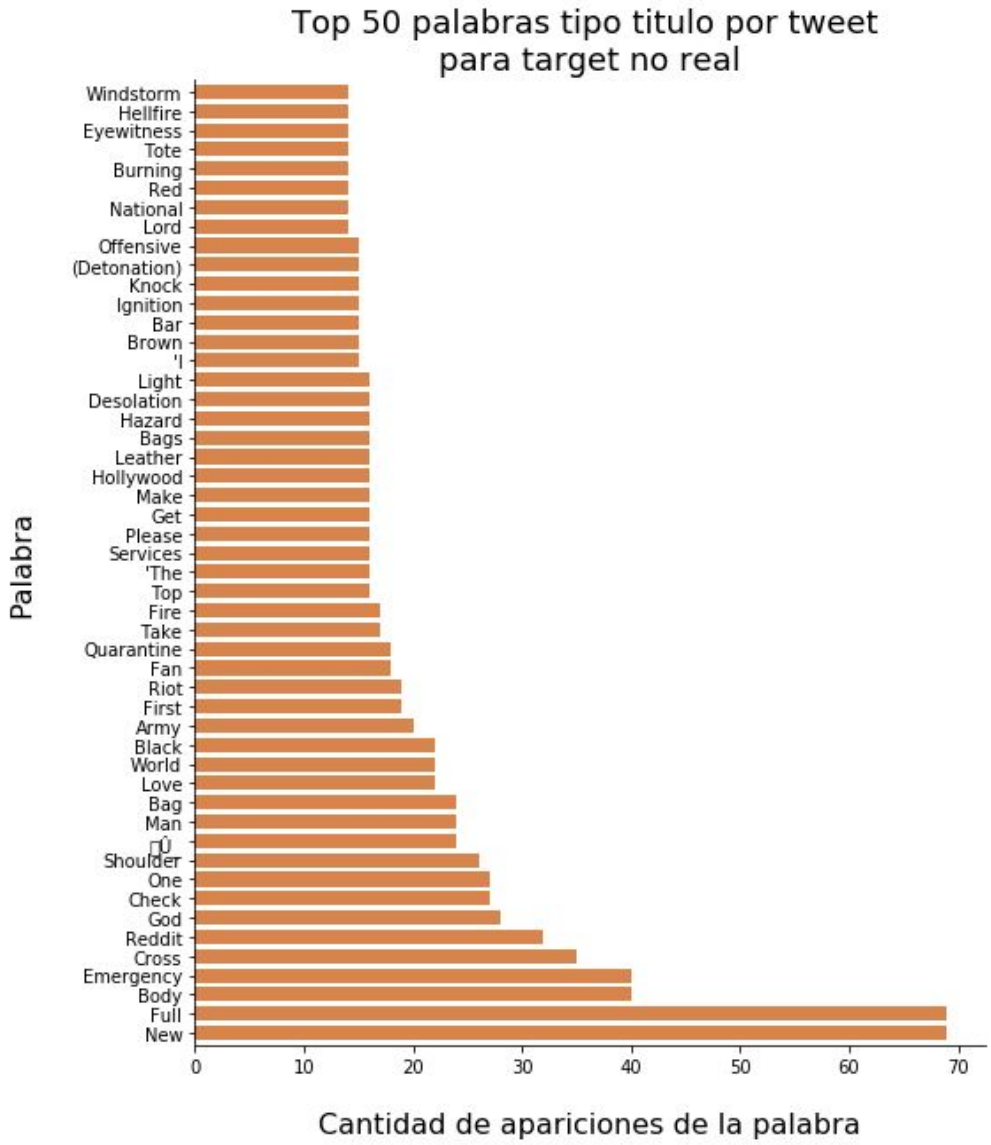

Top 50 palabras tipo titulo por tweet
para target no real
Palabra
Windstorm
Hellfire
Eyewitness
Tote
Burning
Red
National
Lord
Offensive
(Detonation)
Knock
Ignition
Bar
Brown
'I
Light
Desolation
Hazard
Bags
Leather
Hollywood
Make
Get
Please
Services
'The
Top
Fire
Take
Quarantine
Fan
Riot
First
Army
Black
World
Love
Bag
Man
Shoulder
One
Check
God
Reddit
Cross
Emergency
Body
Full
New
0
10
20
30
40
50
60
70
Cantidad de apariciones de la palabra

Figura 40

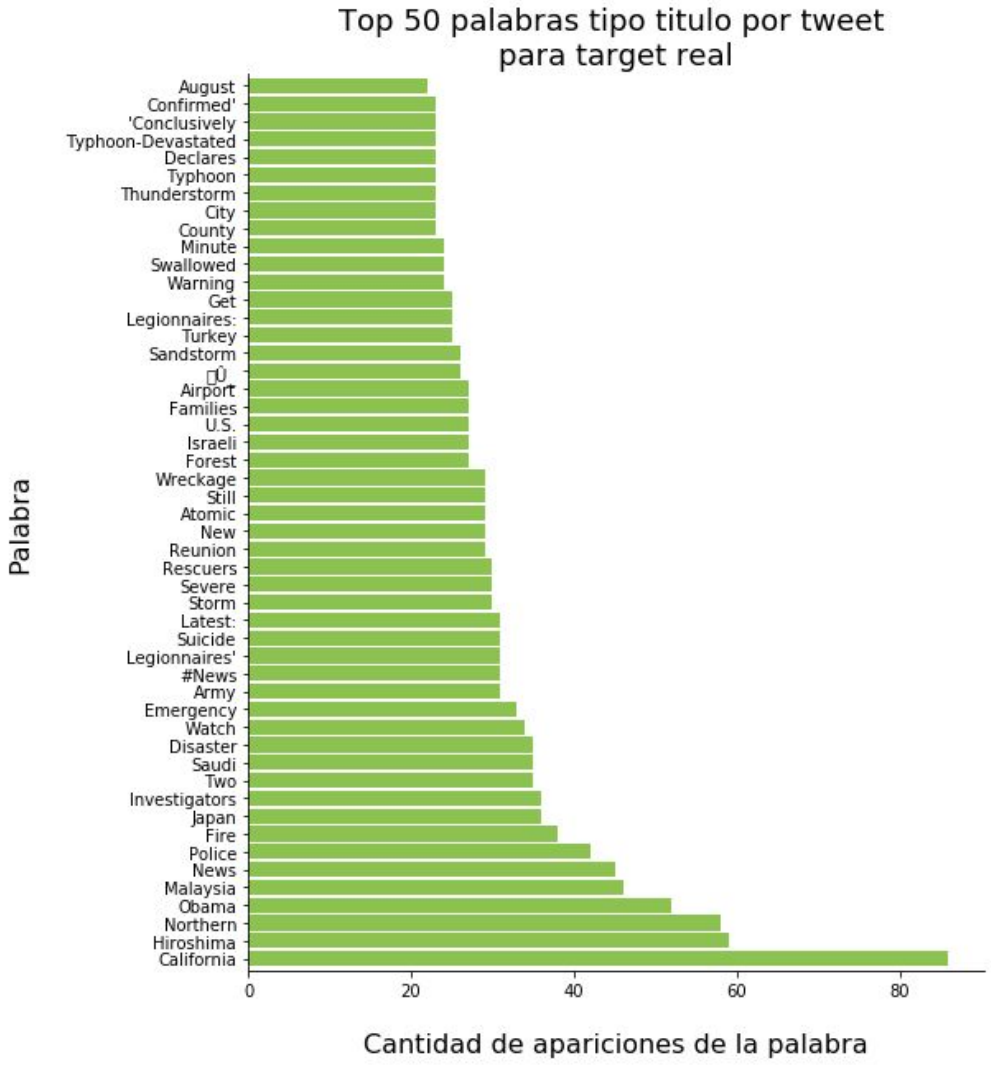


Figura 41

Podemos ver en el caso de los desastres reales apariciones de palabras que deben llevar mayúsculas por defecto como nombres de lugares o nombres propios (en especial el de Obama). Más allá de esto, a simple vista no hay algo que resalte pero los resultados mostrados son en general más claros que en los otros mostrados.

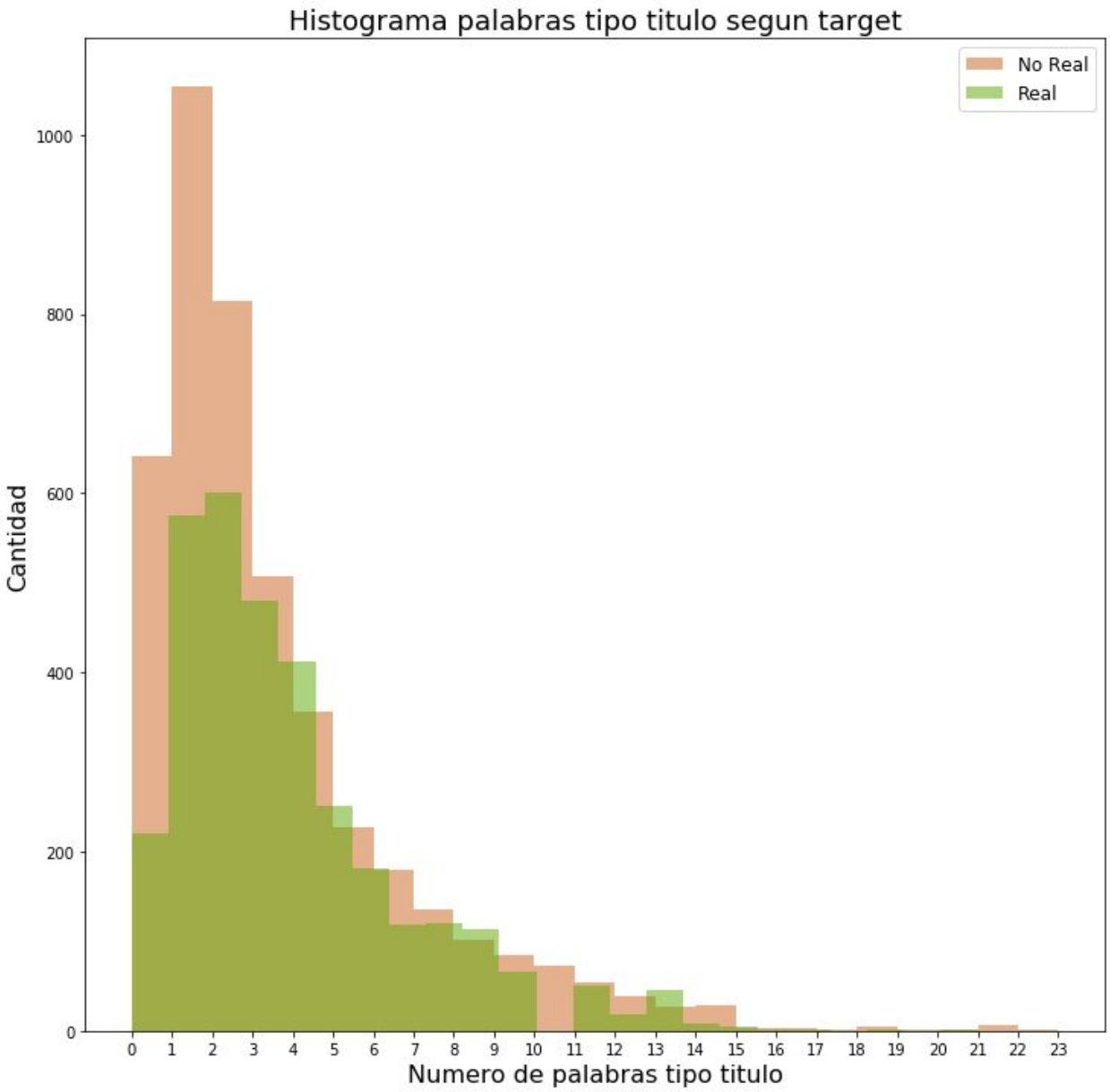


Figura 42

Podría decirse que el análisis de la distribución en este caso es similar al de las stopwords.

Longitud promedio de cada palabra

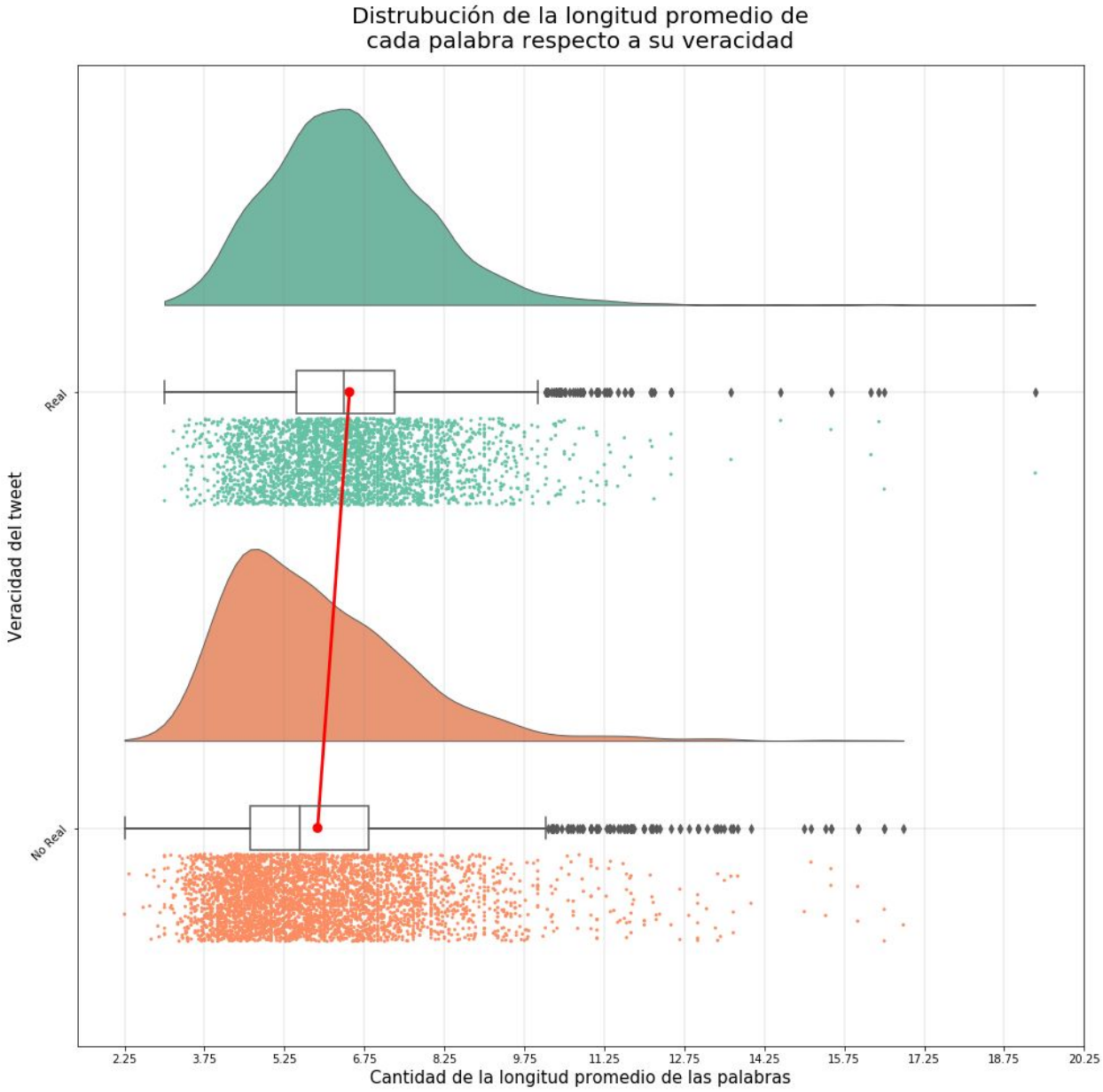


Figura 43

Podemos ver que la longitud promedio de cada palabra se distribuye levemente hacia la derecha por la aparición de varios outliers, aunque si no estuvieran, su distribución sería normal. Podemos ver que la longitud de cada palabra está, en promedio, entre 5 y 7 caracteres.

Depuración de Tweets

Al texto contenido en los tweets del set de datos le aplicamos un proceso de depuración que se detalló anteriormente y luego de ello proseguimos con los siguientes análisis:

Luego de este proceso: ¿Como quedan las longitudes de los tweets, sean Reales o No , antes y después de ser depurados?

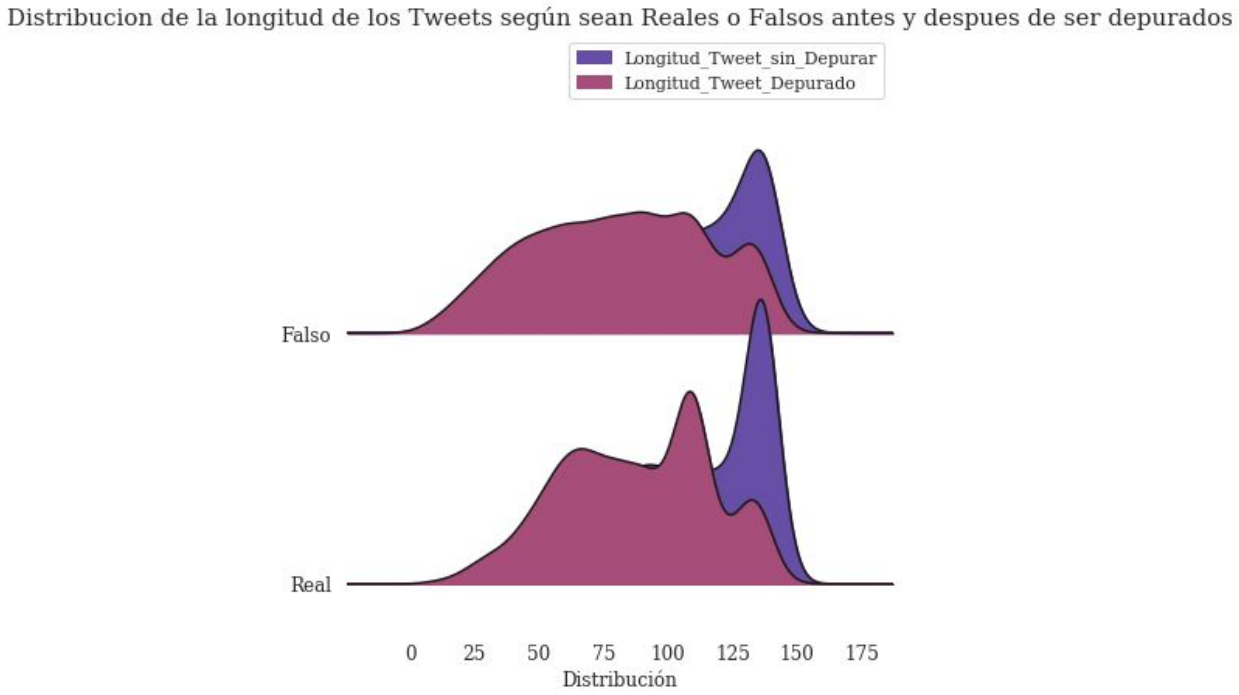


Figura 44

A primera vista pareciera ser un gráfico simple que muestra cómo se distribuyen longitudes de textos, pero en realidad podemos sacar mucha información del mismo. Se observa que en los tweets sin depurar hay mayores picos, con longitudes de varios tweets entre 125 y 150 caracteres y esto es muy interesante.
Recordar que Twitter, que es una plataforma de microblogging, dispuso inicialmente que la longitud de mensajes de sus usuarios fuera como máximo de 140 caracteres. Pero ¿Cómo es que hay casos de tweets, (mayormente sin depurar) que según el gráfico, tienen 150 caracteres?
Una primera respuesta podría ser que esos tweets sin depurar, que tienen más de 140 caracteres contienen links.
Investigando un poco (https://help.twitter.com/es/using-twitter/how-to-tweet-a-link) vemos que Twitter convierte cualquier URL que se comparta en un mensaje a 23 caracteres (incluso si el vinculo tiene un menor tamaño). Por lo tanto, puede ser que esos tweets contengan en realidad

un link mayor a 23 caracteres, que al ser pasado a un archivo para ser analizados, como en este trabajo práctico, se pase el texto real del tweet con la cantidad total de caracteres del link.
También podemos realizar suposiciones acerca del período al que corresponden los tweets ya que el 26 de septiembre de 2017 Twitter dejó que algunos miles de sus usuarios puedan escribir textos del doble de longitud, 280 caracteres, meses más tarde de este testeo inicial, la empresa decidió hacer disponible los 280 caracteres para todos sus usuarios (de repente las ocurrencias ingeniosas ya no tenían que caber en una sola frase). Este gráfico valida de alguna manera la decisión de extender la cantidad de caracteres ya que, al menos en este set de datos, se muestra claramente que la mayoría de sus usuarios ocupaba casi la totalidad de los mismos. Una prueba no concluyente de que los tweets del set de datos es de por lo menos antes de septiembre de 2017 podemos verlo en el tweet de nuestro set correspondiente al “id = 53” (https://twitter.com/AnyOtherAnnaK/status/629195955506708480) que data de agosto de 2015.
Continuando con el análisis del gráfico vemos que para los casos de tweets con el target=0 (Falso), existe un achatamiento de la curva luego de la depuración, pasando de tener inicialmente una mayor cantidad de tamaños entre 125 y 150 caracteres a una mayoría de tweets con tamaños entre 50 y 125 caracteres. Para el target=1 (Real) el comportamiento es similar, con un descenso más brusco en el pico de longitudes entre 125 y 150 caracteres, con un comportamiento no tan plano como el tenido en el caso anterior y con un pico de longitud de tweets entre 100 y 125 caracteres.

¿Cómo se distribuyen las longitudes de los tweets según la Key Global, antes y después de ser depurados?

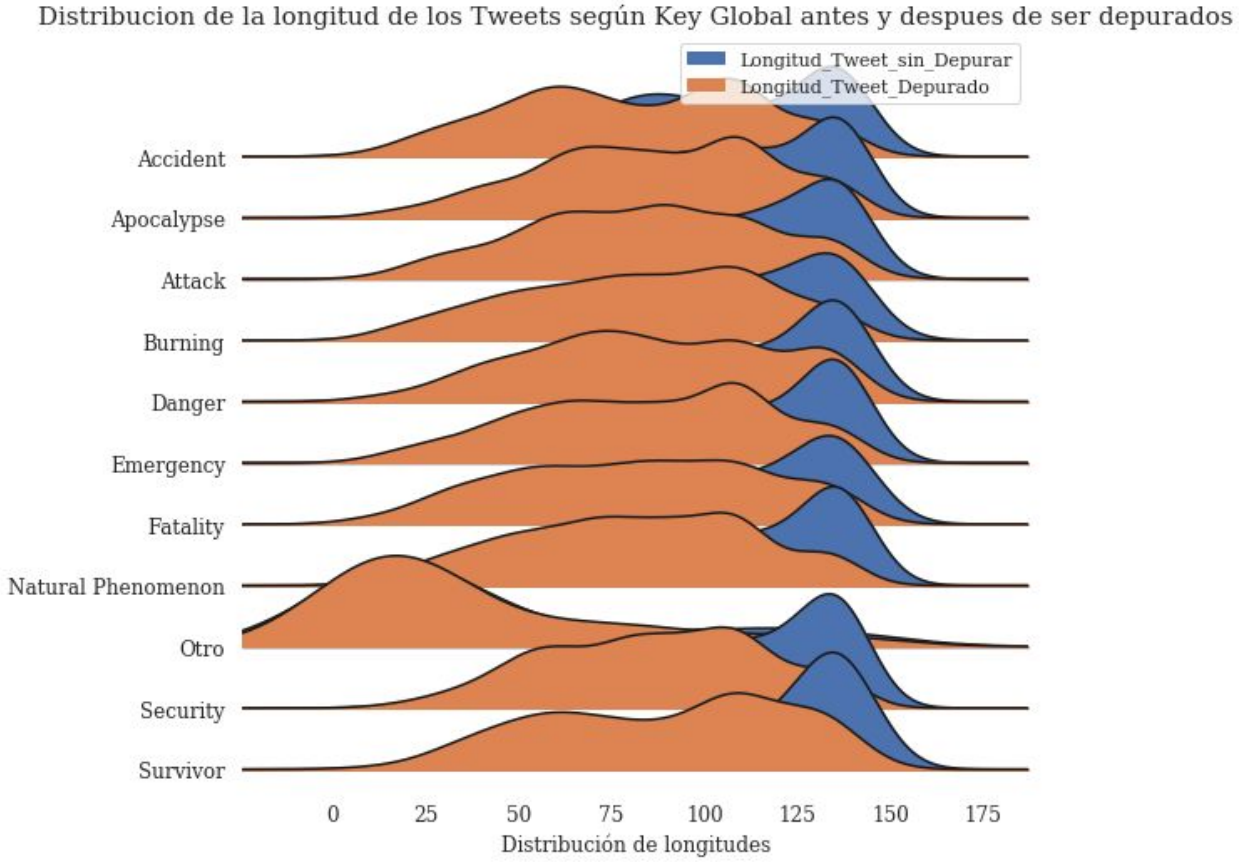


Figura 45

En este gráfico observamos un comportamiento bastante similar en las longitudes de los tweets, antes y después de la depuración, para todas la Key Globales salvo a la que definimos como “Otro”, en el que las curvas están solapadas y su comportamiento es prácticamente el mismo antes y después de depurar los tweets. Recordad que esta última Key Global (“Otro”) es la que agrupa a todas las keyword denominadas “other” en el proceso de depuración de keywords, que son casos en los que a las keywords nulas, que inicialmente eran 61, no le pudimos encontrar y asignar una de las key ya existentes, en base al contenido del texto de los tweets. Pudimos encontrar una key para 38 de las 61 casos nulos y a los 23 restantes se les asignó como keyword “other” y como Key Global “Otro”, que como vimos en (figura 12) estos 23 casos se reparten en 17 para target=0 y 6 para target=1.

Luego de la depuración del texto de los Tweets, ¿Cómo se distribuyen las longitudes de los Tweets según la Key Global para casos Reales y No Reales de accidentes?

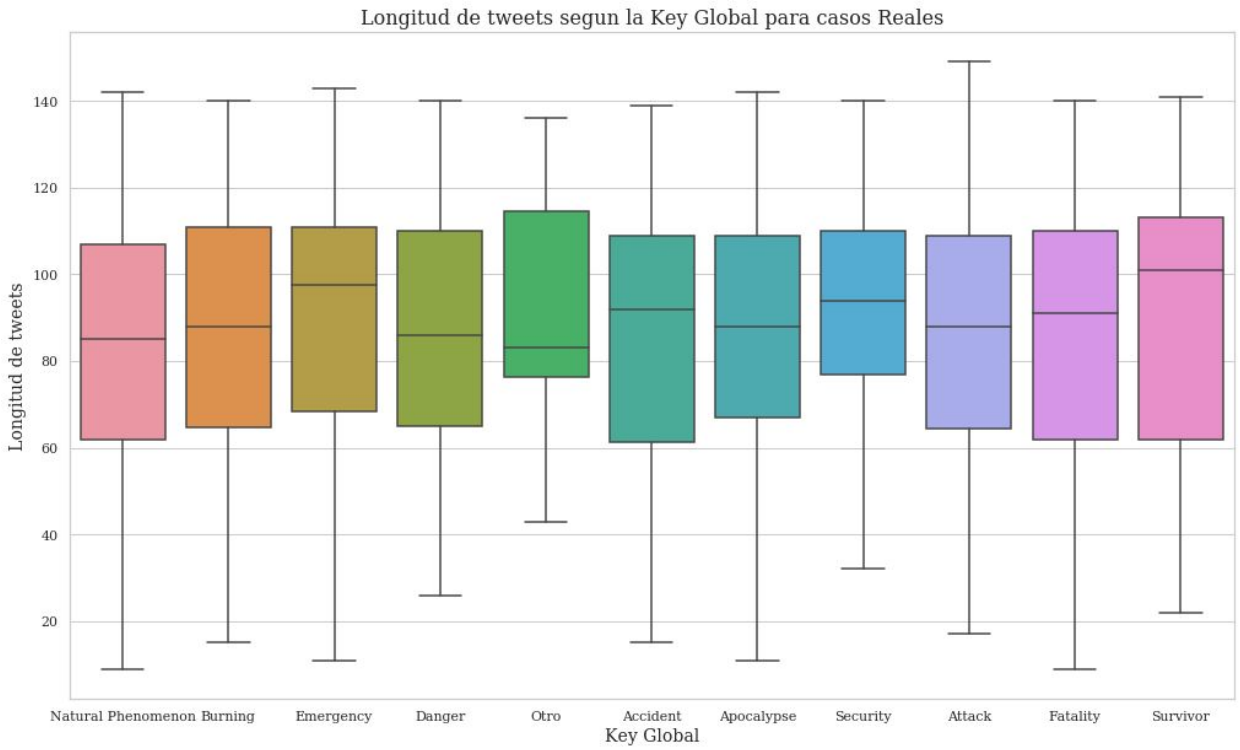


Figura 46

En este gráfico de boxplot, para casos de target =1, observamos que para todas las Key Globales la distancia entre el mínimo y el máximo es bastante grande lo que denota gran variabilidad en la longitud de los tweets y el 75% de tweets para las distintas Key Globales no supera los 120 caracteres . La mediana nos indica el valor medio de las longitudes de los tweets y la Key Global “Survivor” es la que presenta el mayor valor, apenas por encima de los 100 caracteres, también presenta una mayor variabilidad en la longitud de los mismos entre el primer cuantil y la mediana que entre la mediana y el tercer cuantil, se observa también en los tweets, que un 25% de los mismos están entre un poco más 20 y 60 caracteres, un 50% están entre 20 y 100 caracteres.

La Key Global “Emergency” tiene un comportamiento similar a la anteriormente descrita pero con un tamaño de boxplot menor lo que denota una distribución menos “amplia” que la anterior pero con valores de longitudes mínimos y máximos más elevados.

El caso de la Key Global “Otro” muestra que el 50% de los tweets con esta key está entre 40 y algo más que 80 caracteres, teniendo además menores valores en longitudes mínimas y máximas de tweets, cabe resaltar que esta key es la que agrupa la menor cantidad de tweets como se explicó anteriormente. Se destaca también la Key Global “Attack” como la key que contiene los tweets más largos.

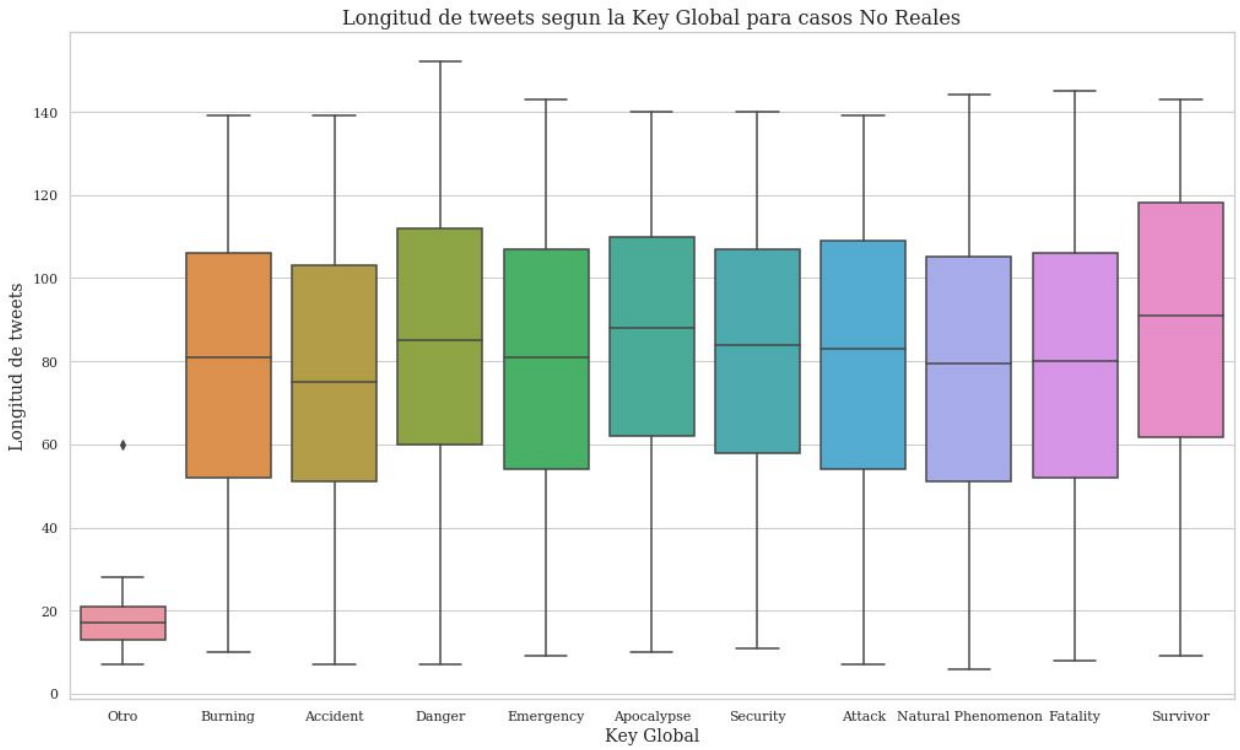


Figura 47

En el caso de los no reales de igual modo el 75% de los tweets no supera los 120 caracteres, destacándose nuevamente la Key Global “Survivor” como la que presenta una mayor mediana cercana a los 90 caracteres (menor a caso del target=1, donde su mediana era de algo más de 100 caracteres). Vemos además que la Key Global “Otro” es la que presenta una caja mucho más compacta a las otras con valores de tweets entre 10 y 30 caracteres, con un outlier de un tweet con 60 caracteres. En este caso la Key Global “Danger” es la que tiene los tweets de mayor longitud.

¿Cuáles son las cantidades de vocales y consonantes utilizadas en los Tweets, ya sea para casos Reales o Falsos?

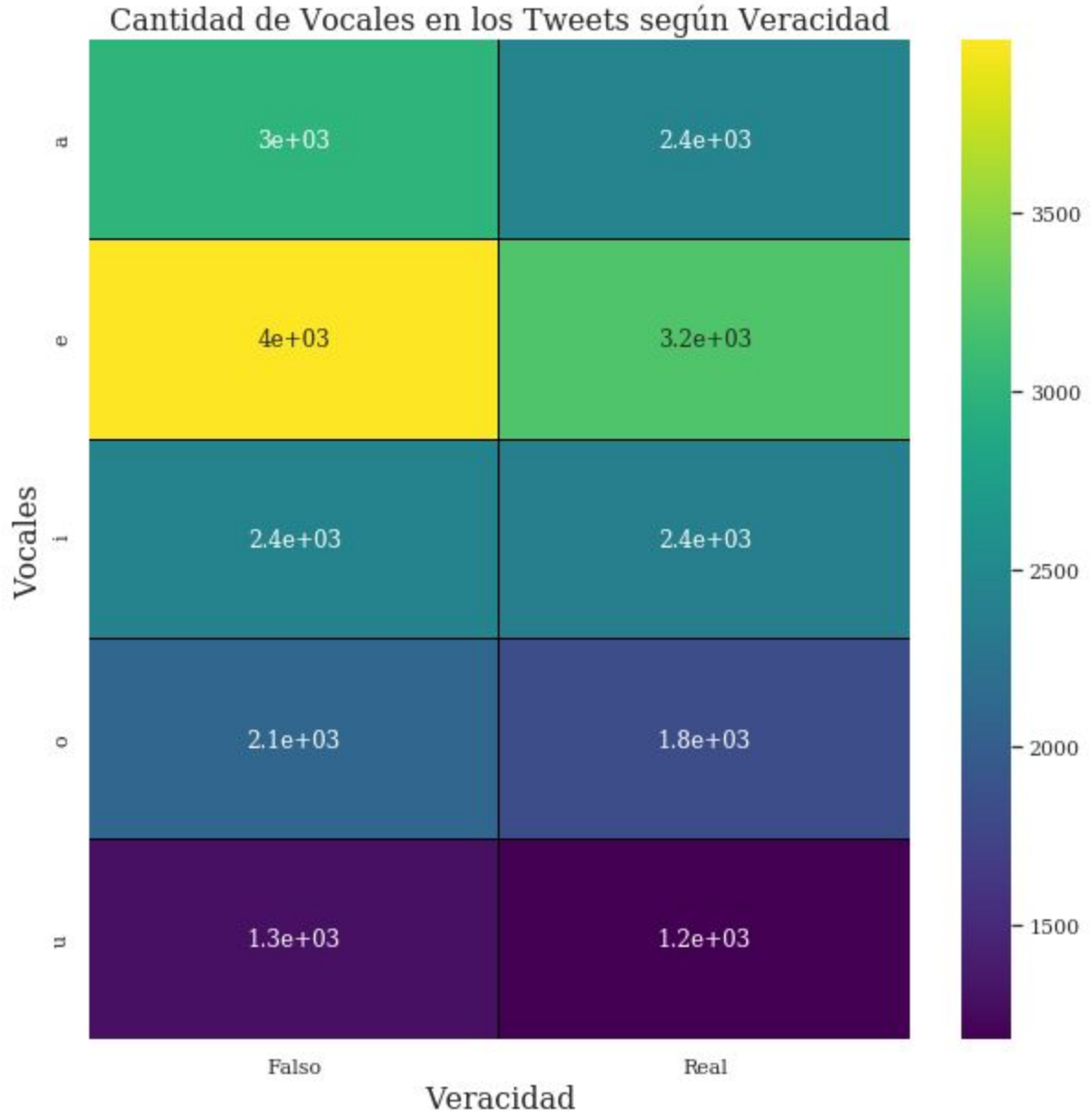


Figura 48

En este heatmap vemos que para casi todas las vocales, salvo para la letra “i”, la cantidad de vocales utilizadas en los tweets es mayor para los casos de target=0, lo cual es bastante lógico ya que en el set de datos tenemos más tweets referidos a este target. La vocal más popular es la “e” con una frecuencia de aparición de 4000 y 3200, para el target=0 y target=1 respectivamente. La vocal menos popular es la “u” con 1300 y 1200 apariciones según el target sea 0 o 1.

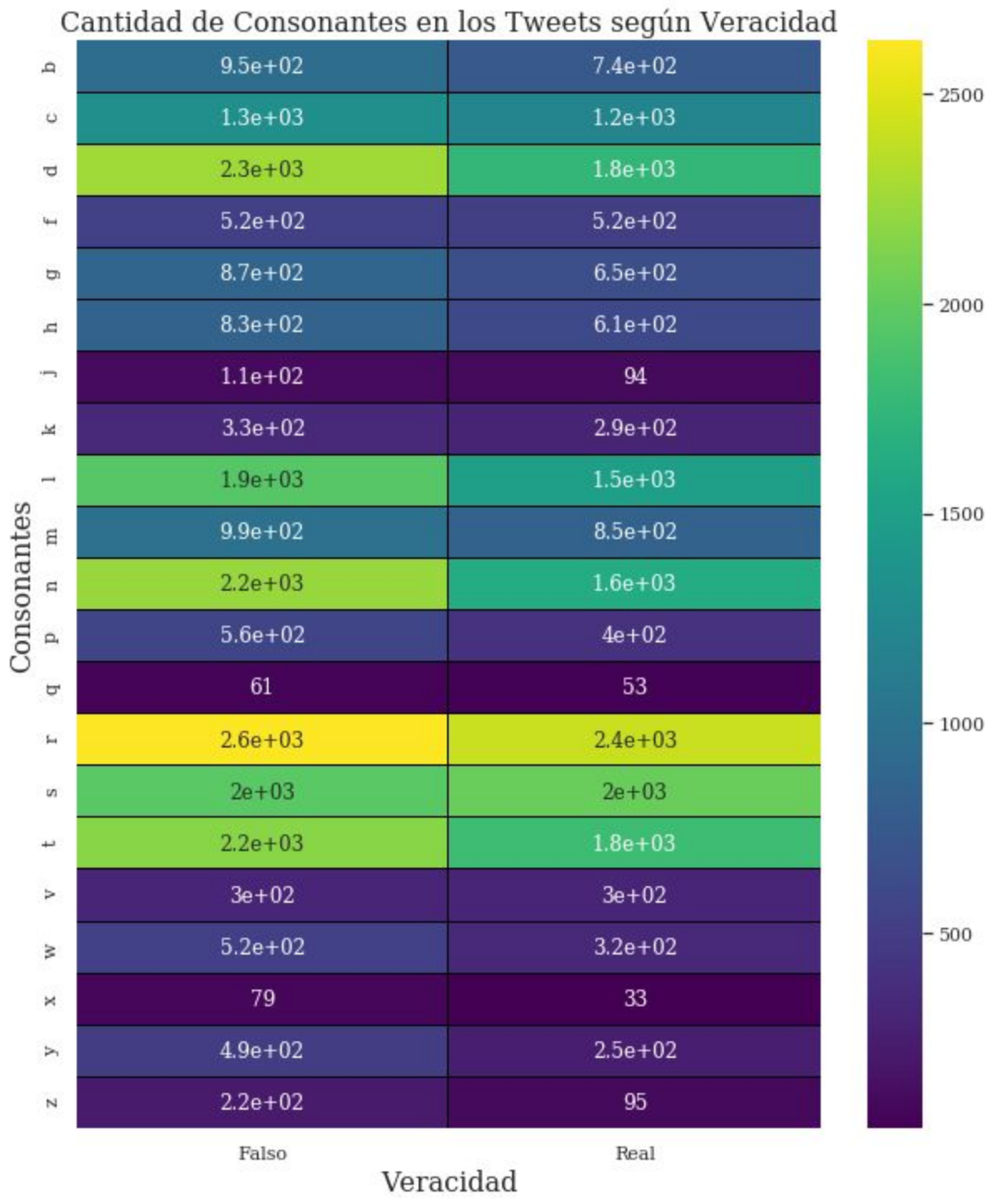


Figura 49

Al igual que para el caso de las vocales, en los tweets correspondientes al target=0 tenemos una mayor cantidad de consonantes, (que como ya se explicó, en el set de datos tenemos más tweets referidos a este target). Se destaca la consonante “r” como la más utilizada con una frecuencia de aparición de 2600 y 2400, para el target=0 y target=1 respectivamente. Así

mismo la consonante menos popular para los casos de target=1 es la “x” con una frecuencia de 33 apariciones y para los casos de target=0 es la “q” con una frecuencia de 61 apariciones.

N-Gramas

Un N-grama es una secuencia contigua de n elementos de una muestra dada de texto o discurso.
Respecto a los n-gramas, primeramente realizamos un doble filtrado que consistió:

- Quitamos todas las StopWords
- Filtramos aquellos términos que tengan una longitud menor a 3

Esto nos ayudó a obtener términos mucho más interesantes que fueron:

Unigramas

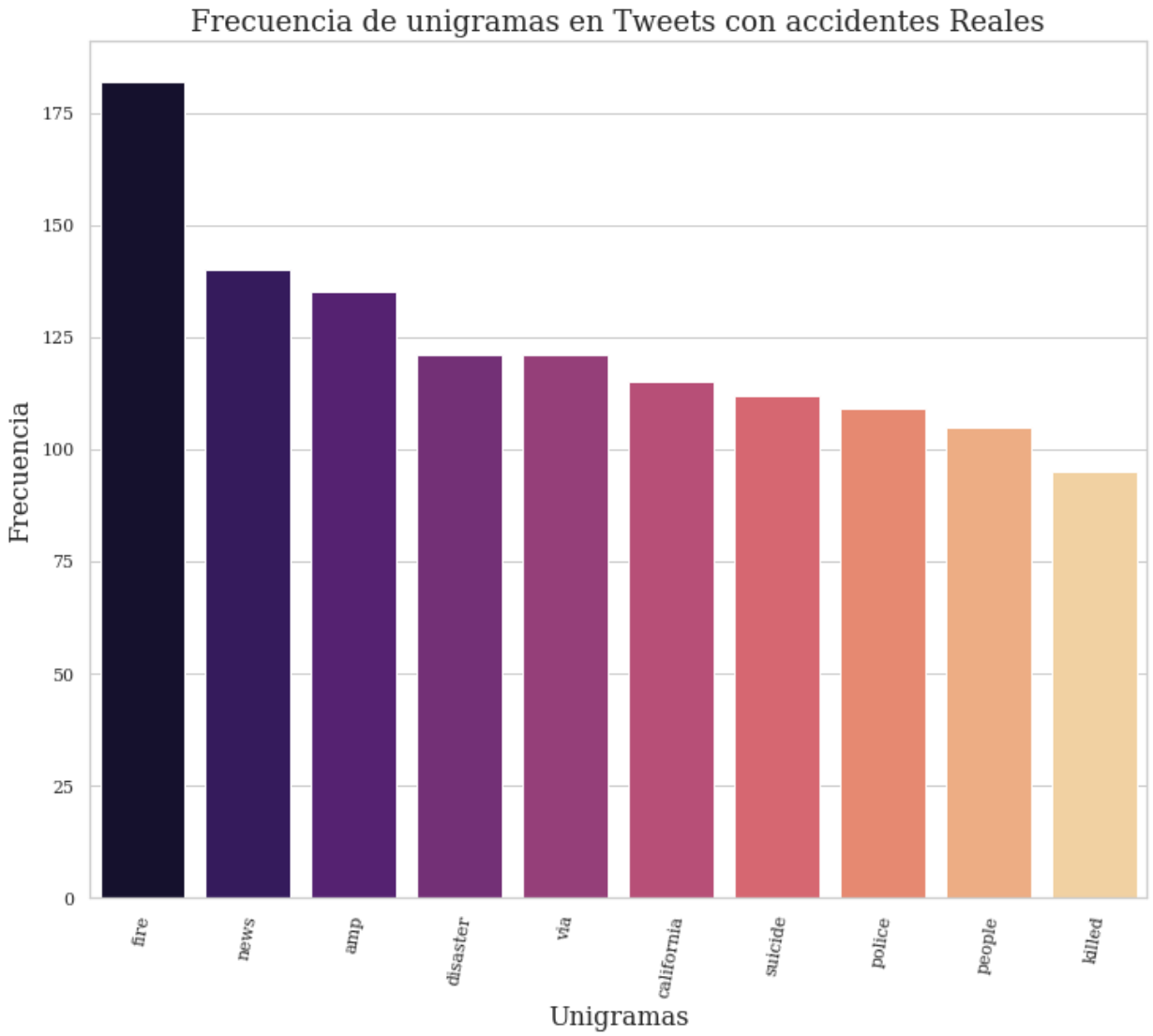


Figura 50

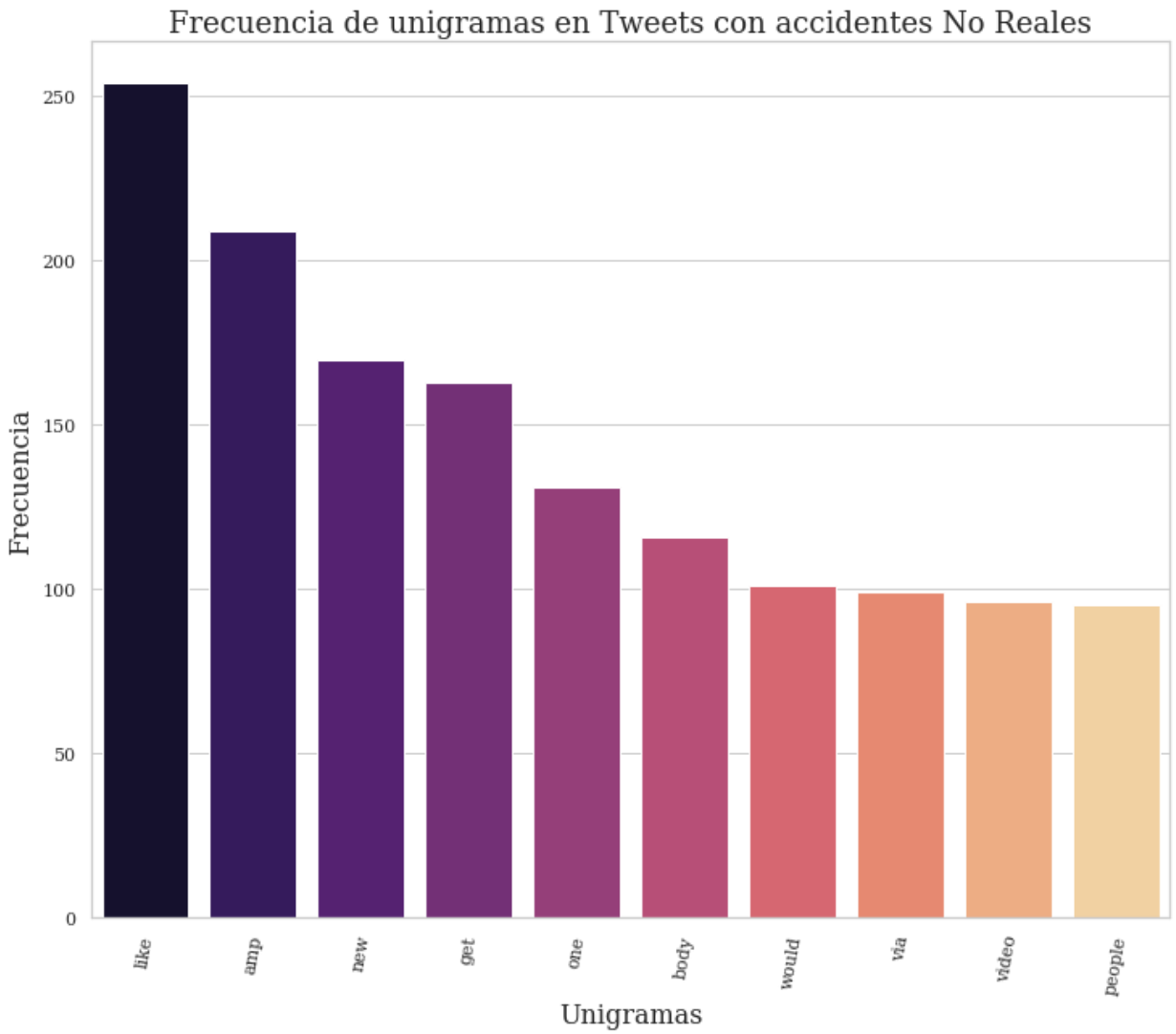


Figura 51

El análisis es similar a aquel de las frecuencias de palabras sin depurar (Figura 28 y 29) aunque en este caso por el filtrado realizado podemos ver menos basura.

En el target “real” tanto los topics de desastres como también “California” predominan, y la palabra “amp” (referente a &) habrá que tenerla en cuenta sin duda para hacer más claro al campo text.

En el target “no real” se destaca “like” y nuevamente “amp” como también “body”.

Bigramas

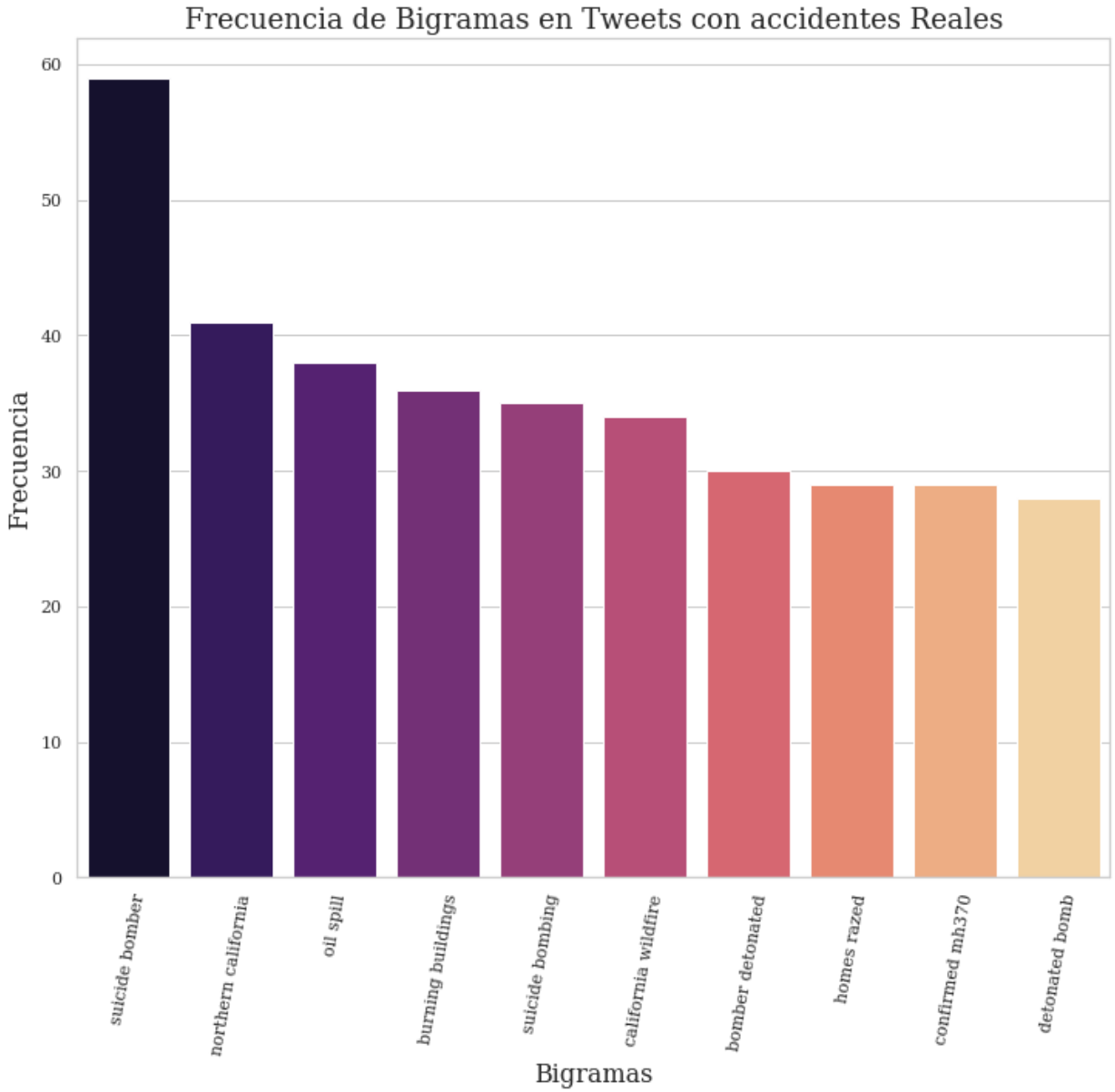


Figura 52

Podemos ver que en su mayoría los bigramas muestran información interesante acerca de los desastres reales, en particular es importante destacar a los “suicide bomber” ya que aparece múltiples veces, con ciertos cambios en las palabras, pero con el mismo significado.

También podemos notar cierta consistencia en relación a la Figura 5 ya que aparecen referencias al estado de California, que es uno de los que más tweets poseen.

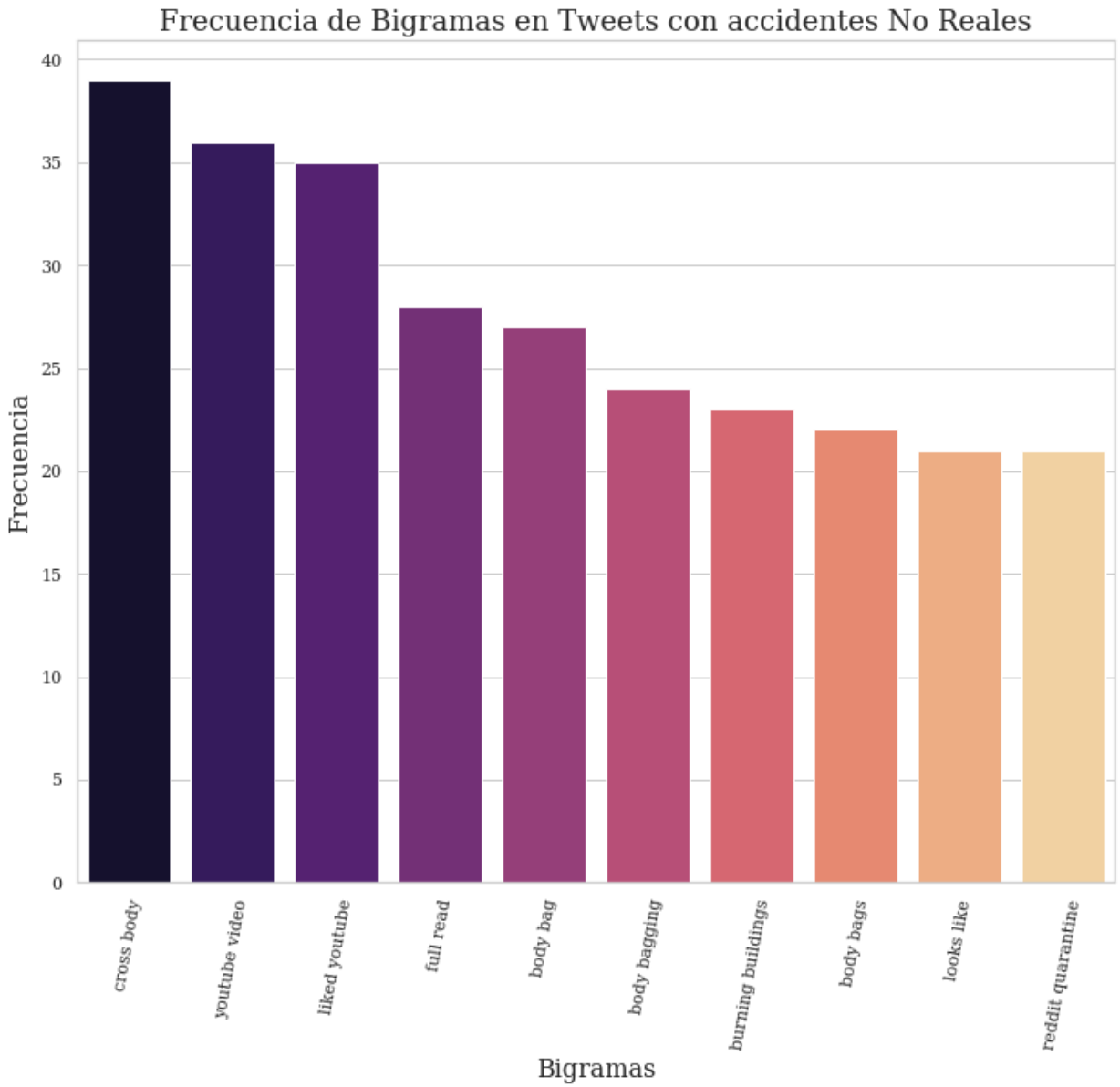


Figura 53

En el caso de lo bigramas con relación a desastres no reales, podemos volver a notar la importancia de “body bag” y “youtube” que aparecieron anteriormente en otras secciones con altas frecuencias.

Trigramas

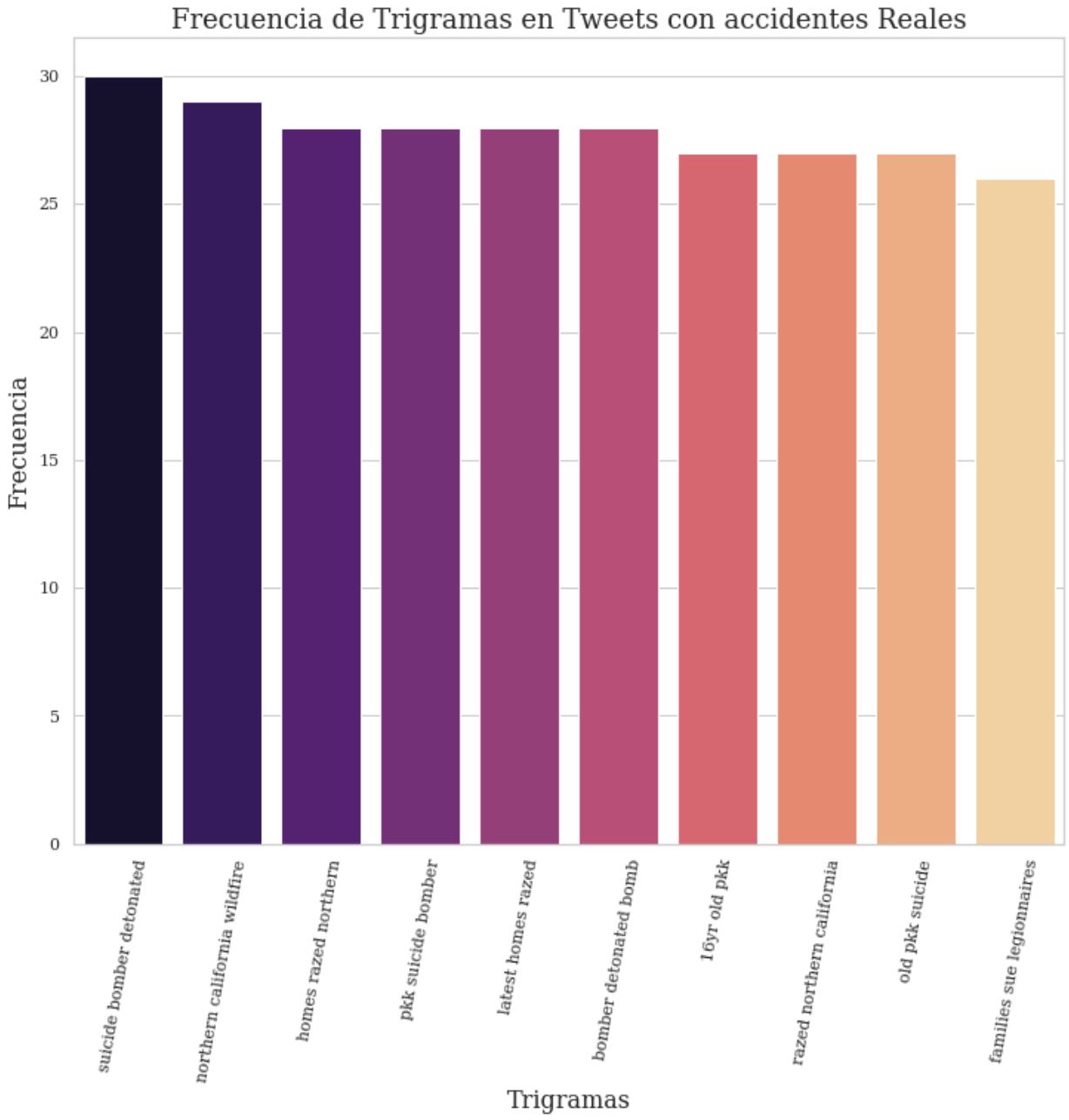


Figura 54

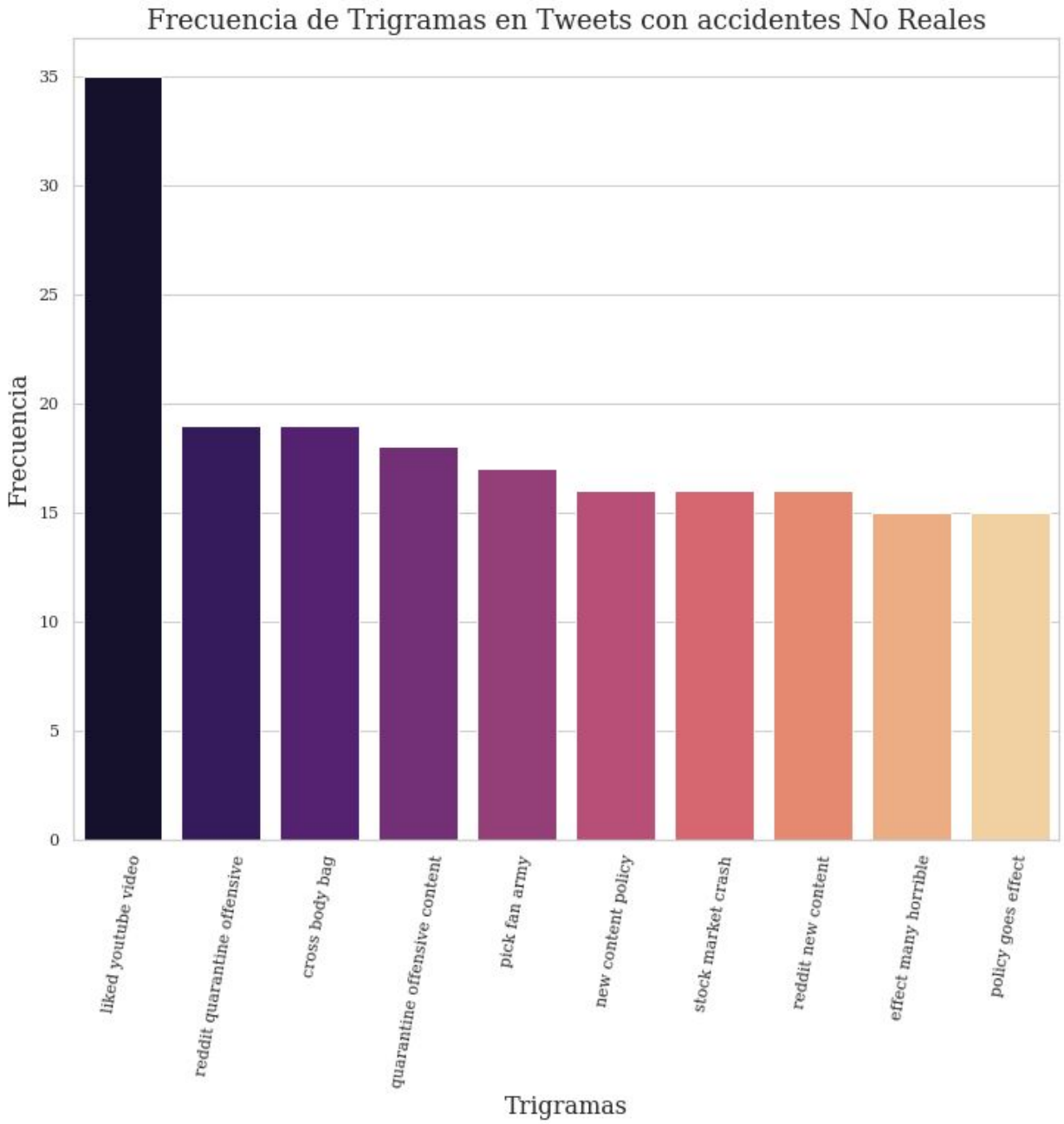


Figura 55

Tanto en los trigramas de desastres reales como no reales, el análisis es análogo a los bigramas. Podemos hacer una mención de honor a la aparición de “stock market crash” y el tema “quarantine” en desastres no reales

ReTweets

¿Existen Retweets en nuestro set de datos? Si existen, ¿Cuántos Tweets tengo que son ReTweets en mi set de datos?

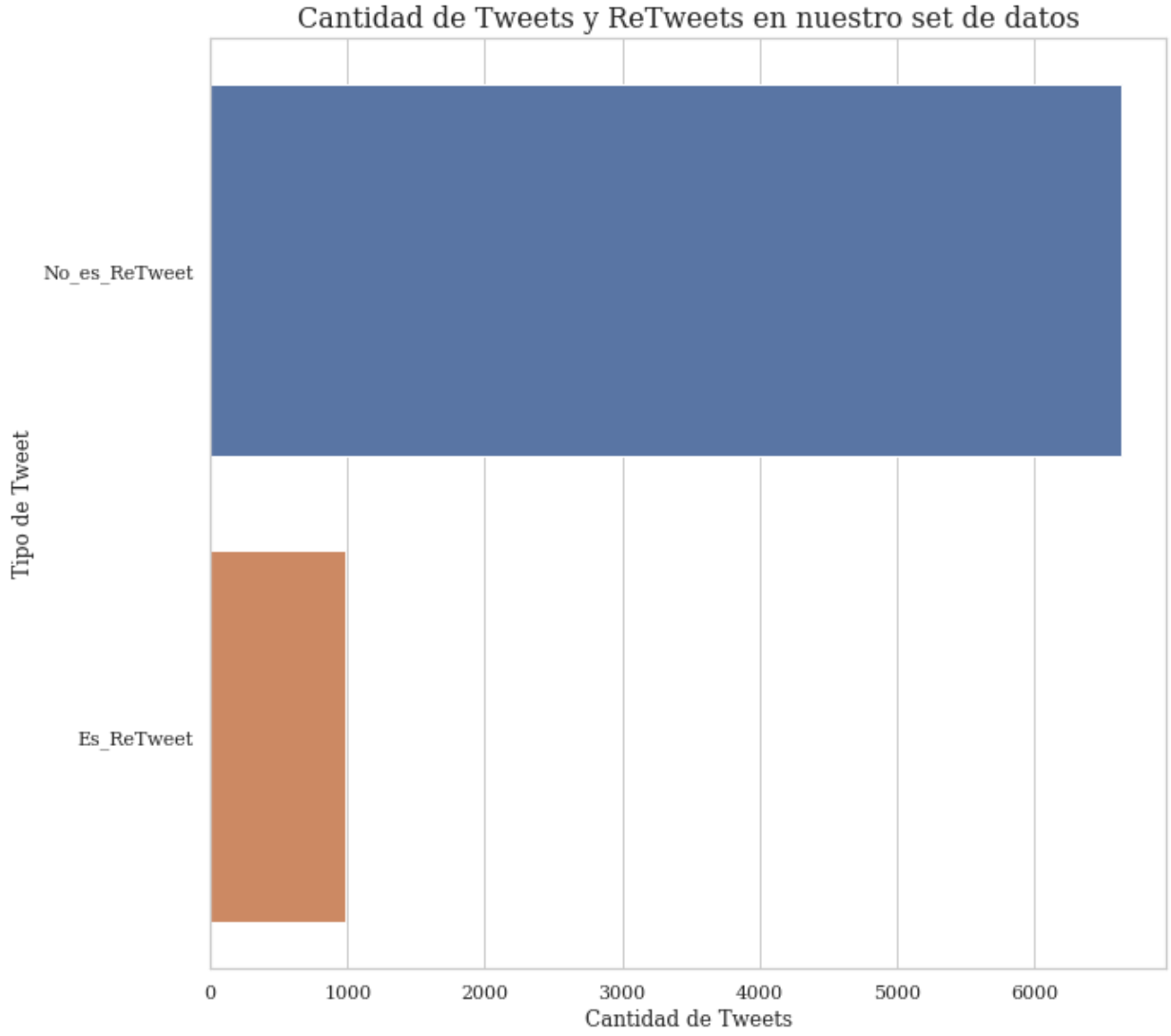


Figura 56

Podemos observar que en nuestro set de datos contamos con Tweets que son en realidad Retweets en una cantidad de casi 1000 apariciones y más de 6500 Tweets que no lo son.

¿Cómo es el flujo los Tweets que son Retweets y los que no lo son, en relación al target?

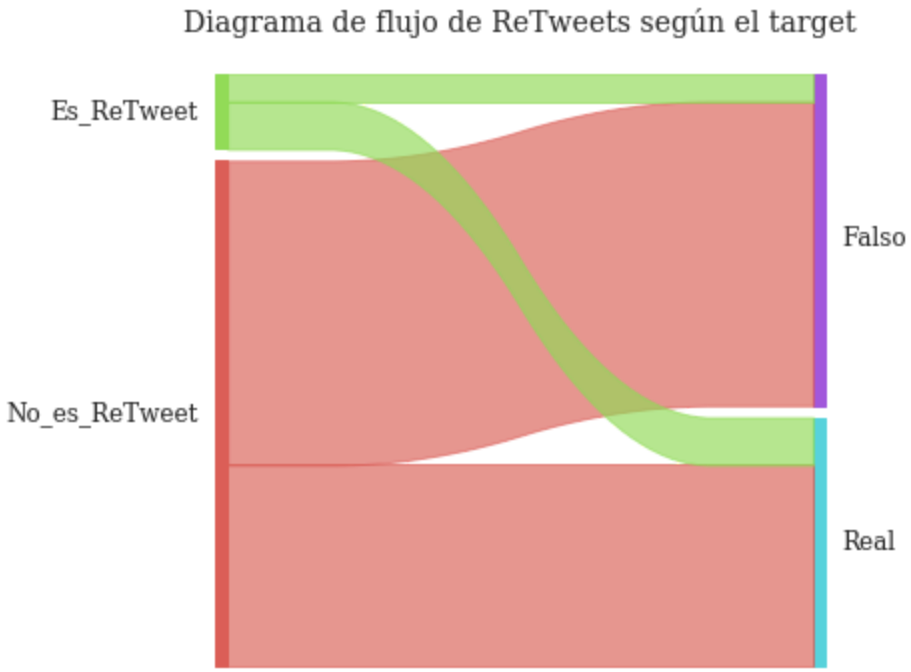


Figura 57

Observamos algo interesante, los tweets que no son Retweets, en su mayoría corresponden al target=0, recordar que en el set tenemos mayor cantidad de tweets con target=0, lo cuál es más que lógico. Lo interesante es que si un Tweet es Retweet vemos un flujo algo más ancho que se dirige hacia el target=1, en comparación al el flujo que se dirige al target=0.

Es interesante también ver sobre qué temas (Key Global) tenemos más Retweets:

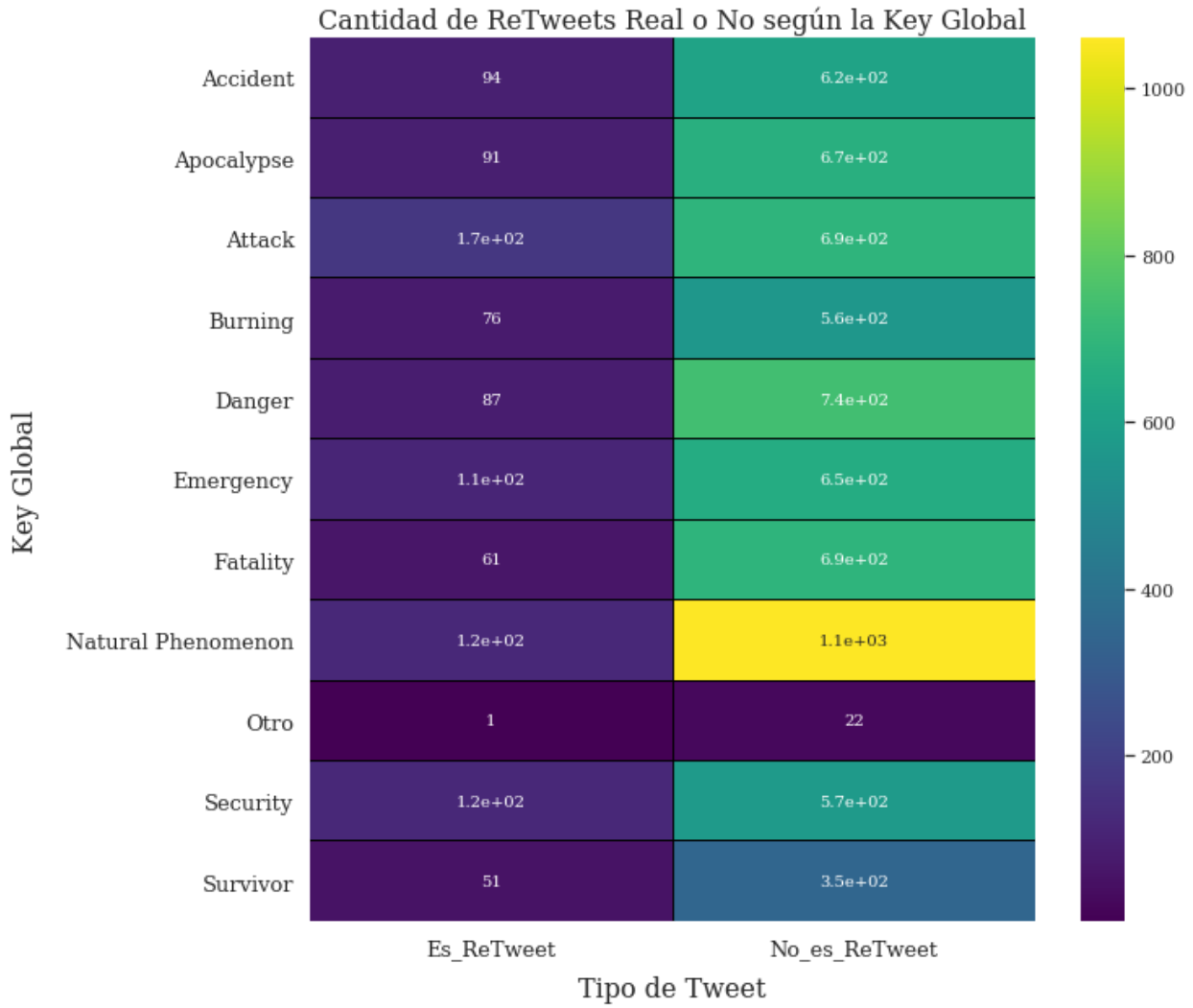


Figura 58

En el gráfico vemos que la Key Global “Natural Phenomenon” es la que más tiene tweets que no son Retweets, osea es de lo que más se habla en nuestro set de datos, pero no es la Key Global que más Retweets tiene, sino que eso le corresponde a la Key Global “Attack” con una cantidad de 170. Como era de esperar la Key Global “Otro” es la que menos Retweets tiene con solo un caso. Otro dato a destacar es que todas las Key Globales contienen ReTweets.

Correlación entre algunas features

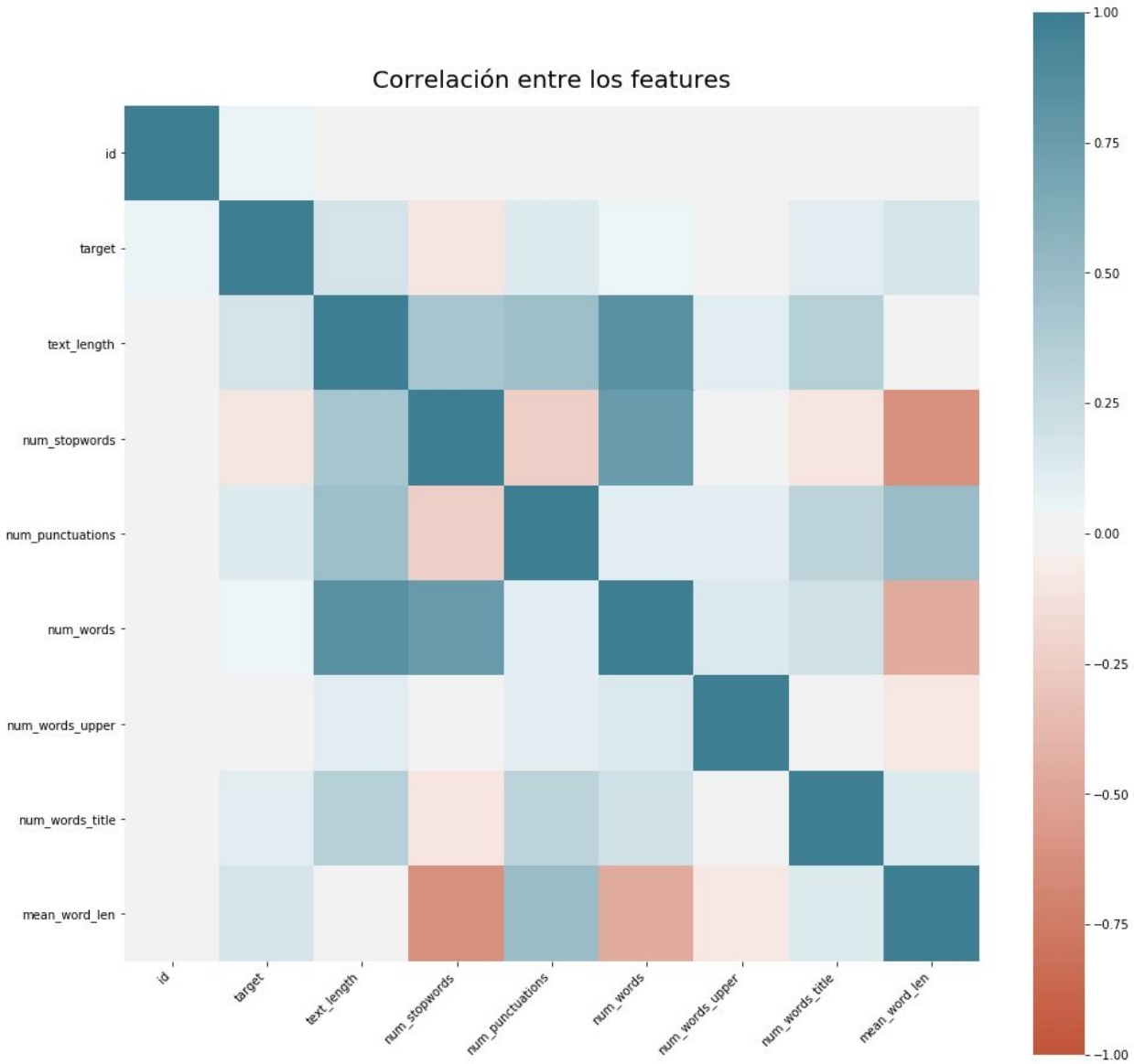


Figura 59

Como era de esperarse, algunas features realizadas a partir del texto no muestran mucha correlación con el target lo cual es bueno porque podrían en el futuro ayudar a aprender a predecir.

Por otro lado podemos ver correlaciones positivas como también negativas entre las nuevas features creadas lo cual podría llevar a tener que implementar algún tipo de feature selection para el tp2.

Conclusiones

Una vez analizado el set de datos realizando una análisis minucioso de cada feature dispuesto llegamos a las siguientes conclusiones:

- En el feature de **location** tenemos 2 problemas principales, por un lado la cantidad de registros faltantes por defecto (un 33%) y la cantidad de ruido en los registros existentes, que luego del análisis nos lleva a tener más de un 50% de registros faltantes. Por lo tanto consideramos que la ubicación en general carece de información y solo sirve para tener un paneo respecto a los otros 2 features y por lo tanto no será algo fundamental en el próximo trabajo práctico.

- En el feature de **keyword** conseguimos agrupar ciertos términos que nos ayudaron a entender cómo se relacionan las palabras respecto a si un tweet era real o no. El tema de poder agrupar temáticas nos lleva a pensar por ejemplo que en el futuro podemos asociar algún tipo de problema de clustering para este campo en particular. Teniendo esto en cuenta concluimos que este feature puede sernos útil en la siguiente etapa.

- En el feature de **text** vemos que la información desglosada y los nuevos features extraídos a partir de este nos ayudan a entender cuales son las palabras claves y relevantes como también nos da una idea de que deberíamos remover o modificar para poder aprovechar lo más posible la información provista por el campo. Estimamos que es una de las características más importantes para poder determinar la clasificación de los tweets.

Dicho esto, creemos haber generado un buen panorama acerca de lo que se debe tener en cuenta para el siguiente TP, por lo que el grupo se siente bien encaminado y con una base sólida acerca de los datos, sus relaciones y el impacto de las distintas variables sobre la variable dependiente, target.