75.06/95.58 Organización de Datos

Grupo: DATAVID-20

Trabajo práctico 1: Análisis exploratorio de datos

Nombre y apellido	Padrón	Correo electrónico
Manuel Bilbao	102732	bilbaomanuel98@gmail.com
Filyan Karagoz	101933	fkaragoz@fi.uba.ar
Darío Markarian	98684	dmarkarian@fi.uba.ar
Lautaro Stroia	100901	lstroia@fi.uba.ar

Link al repositorio de GitHub:

https://github.com/Filyank/Datavid-20/tree/master/TP1-1c2020

Fecha de entrega: 21/05/2020

Primer cuatrimestre de 2020

Índice

1. Introducción

El informe aquí mostrado posee la información que hace referencia al primer trabajo práctico de esta materia.

Para la realización de este trabajo, se utilizó un set de datos de tweets que tratan sobre posibles catástrofes naturales. Dicho set fue conseguido en Kaggle, e incluye algunos tweets redactados desde distintos puntos del mundo.

El objetivo del trabajo práctico es realizar un minucioso (o no tanto) análisis exploratorio de esos datos, en busca de reglas que nos hagan relacionar la información del set para luego obtener conclusiones al respecto.

2. Información general sobre el set de datos

2.1. Aspecto general

Antes de empezar con el análisis de los datos, debemos conocerlos.

Trabajando sobre el set provisto, observamos que:

- El archivo posee 38065 registros distribuidos en 7613 filas y en 5 columnas. No todos esos registros poseen información válida.
- Los registros tienen información relacionada a tweets que redactan información, verdadera o no, sobre desastres naturales.
- Las columnas del set son las siguientes:
  - **id:** identificador único para cada tweet
  - **keyword:** una palabra clave particular del tweet (puede estar vacío)
  - **location:** la ubicación desde donde fue enviado el tweet (puede estar vacío)
  - **text:** el texto del tweet
  - **target:** denota si un tweet es sobre un desastre real (1) o no (0)

2.2. Primeros pasos en el manejo de los datos

Empezamos analizando los tipos de datos presentes en el set para ver con qué nos podemos encontrar.

2.2.1. Tipos de datos presentes en el set

Utilizando el método *dtypes* provisto por Pandas, observamos lo siguiente:

- Las columnas **id** y **target** son del tipo *int64*, lo que indica que son valores enteros.
- Las columnas **keyword, location** y **text** son del tipo *object*, indicando que pueden ser valores del tipo string, unicode o tipos mixtos.

2.2.2. Uso de memoria del set de datos

Utilizando la función *memory_usage()* obtuvimos como respuesta que el set de datos provisto para este trabajo ocupa un total de 0.29 Megabytes.

2.2.3. Cantidad de valores nulos

Buscando la cantidad de valores nulos (o *NaN*) presentes en las columnas, nos topamos con los siguientes números:

- En las columnas **id, text** y **target** nos encontramos con que no tienen valores nulos: en **id** y en **target** era de esperar, ya que son valores únicos para cada tweet, pero no pensábamos lo mismo de la columna **text**
- La columna **keyword** posee un total de **61 valores nulos**, cantidad que parece despreciable con respecto a las 7613 filas que posee el set
- Por último, en la columna **location** nos encontramos con **2533 valores nulos**, los que no son pocos respecto a la cantidad total de filas

2.2.4. Cantidad de valores no repetidos

Aplicando la función *nunique()* al set de datos, obtuvimos los siguientes valores:

- la columna **id** no tiene valores repetidos, por lo que hay 7613 valores no repetidos
- la columna **keyword** posee 221 valores no repetidos, por lo que hay 7392 valores repetidos
- la columna **location** posee 3341 valores únicos, lo que implica una cantidad de 4272 valores repetidos
- en la columna **text** nos encontramos con 7503 valores únicos, o sea 110 repetidos
- por último, la columna target es obvio que va a tener solo 2 valores únicos: 0 y 1.

3. Análisis de los datos

3.1. Análisis de los tweets

En esta sección, vamos a hacer un análisis sobre la columna *text* de nuestro set de datos.

Primero, analizamos la longitud de los tweets y buscamos la cantidad de tweets que poseen cierta longitud. Observamos que la gran mayoría de los tweets tiene una longitud de entre 130 y 140 caracteres; el tweet más corto tiene una longitud menor a 10 caracteres (7 específicamente) y el más largo 157[Figura 1].

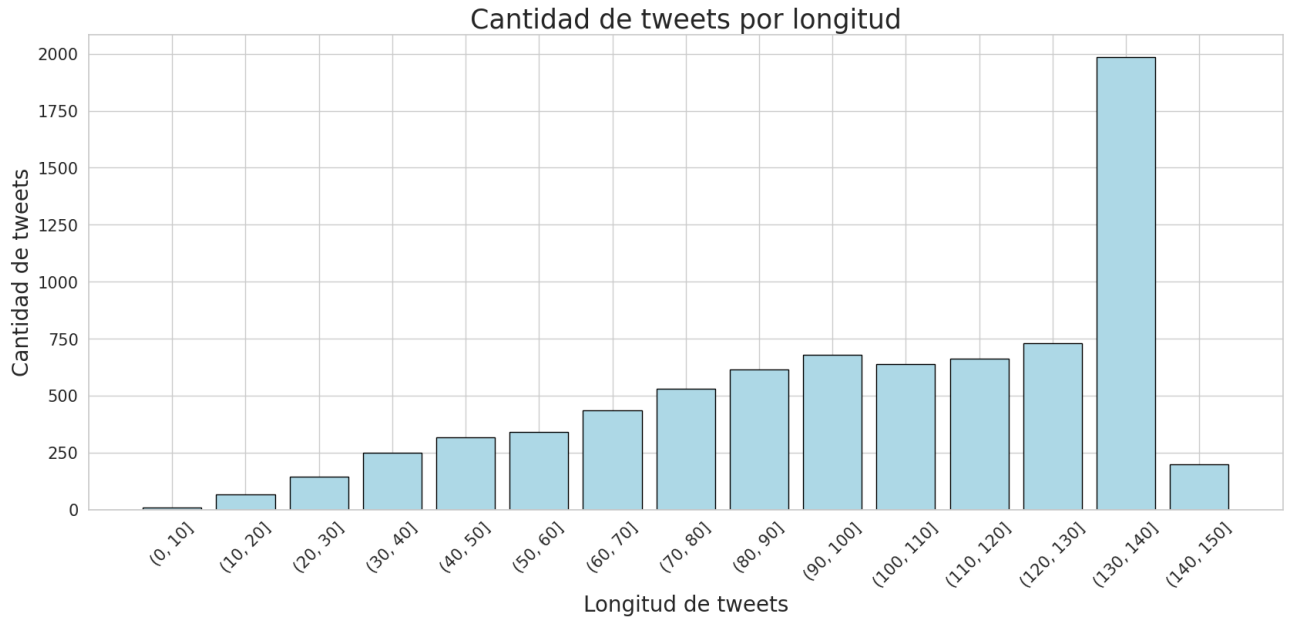

Figura 1: Cantidad de tweets por longitud

3.1.1. ¿Qué cantidad de tweets poseen emojis?

Para este análisis, usamos la librería Emoji. Los resultados obtenidos no son dignos de realizar un análisis más profundo, ya que solo 10 de 7613 tweets poseen emojis, por lo que tampoco realizamos una visualización.

3.1.2. ¿Cuáles son las 10 usuarios más mencionados en los tweets?

Analizamos aquellos tweets que poseen alguna mención hacia otro/s usuario/s de la plataforma, obteniendo los siguientes 10 como los más *arrobados* en el dataset [Figura 2].

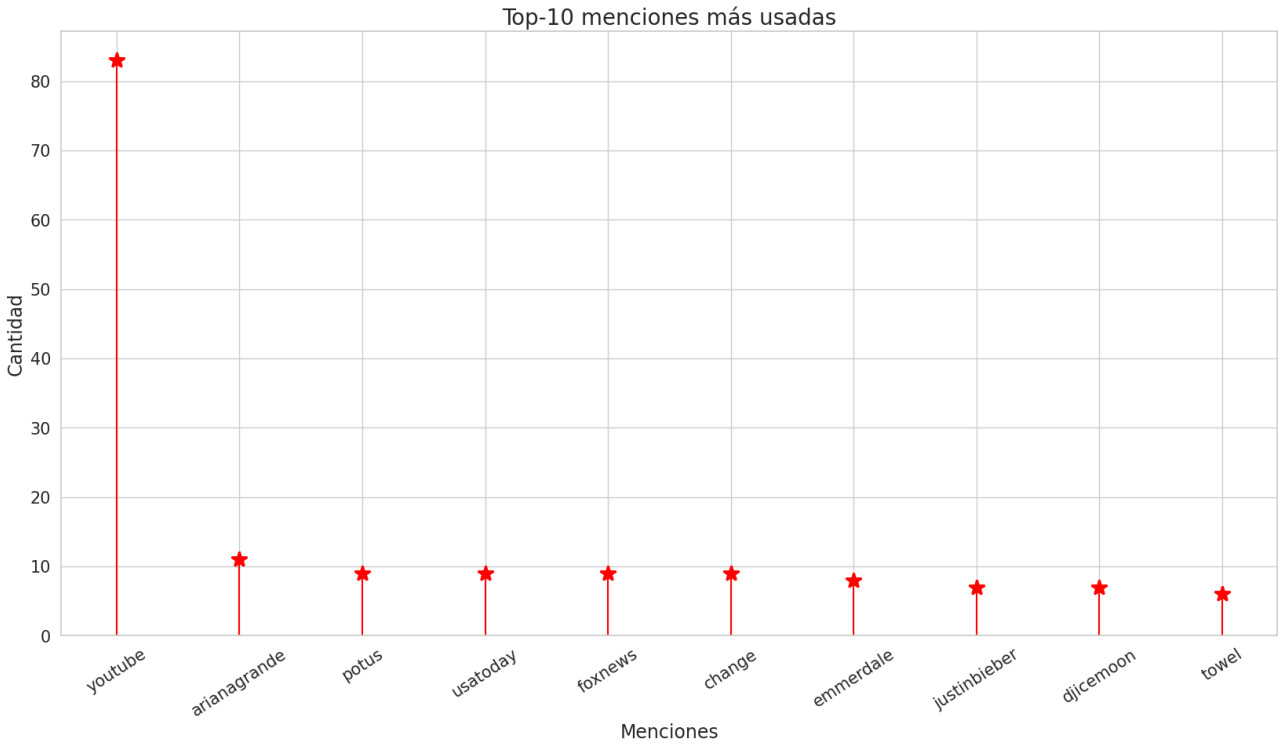

Figura 2: Los 10 usuarios más mencionados

Podemos observar que el usuario de Twitter más mencionado es **YouTube**, lo que nos podría dar un indicio de que esos tweets no aportan información sobre desastres reales. De todos esos usuarios más mencionados, los tweets que podrían contener información real son los que mencionan a **Fox News**, cuenta de un canal de noticias estadounidense.

3.1.3. ¿Cuáles son los 10 hashtags más usados en los tweets?

Similar al análisis anterior, esta vez nos enfocamos en aquellos tweets que poseen hashtags (#), ya que podrían ser indicio de algún desastre natural real (o no)[figura 3].

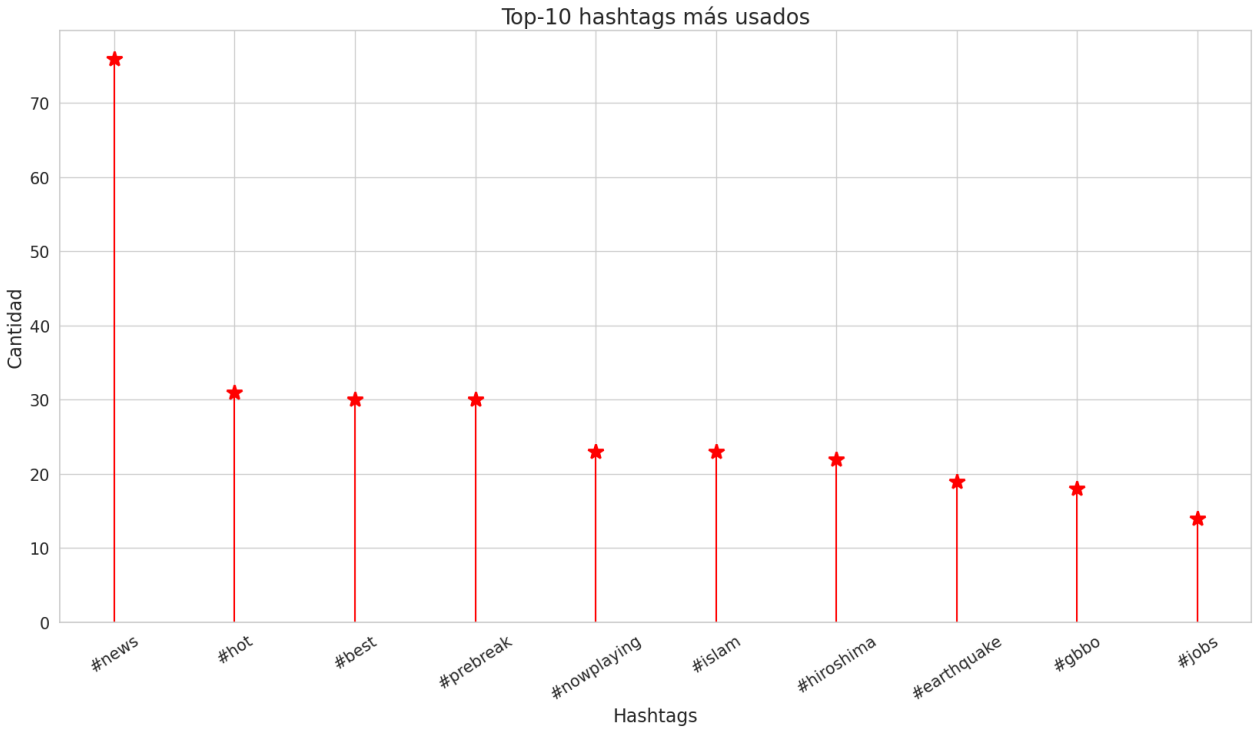

Figura 3: Los 10 hashtags más utilizados

3.1.4. ¿Cuántos tweets poseen URLs?

Buscamos la cantidad de tweets que, en su contenido, poseen URLs. Es un análisis sencillo, el cual nos arrojo los siguientes resultados:

- En el dataset hay **3971** tweets que incluyen **URLs** en su contenido
- En total, el dataset posee **4723 URLs**

3.1.5. ¿Cuántos tweets poseen palabras mayúsculas?

Analizamos el dataset en busca de aquellos tweets que contienen palabras en mayúsculas, considerando palabra a todo conjunto con un mínimo de dos letras, y obtuvimos que:

- En el dataset hay **4312** tweets que poseen palabras en mayúsculas
- En total, el dataset posee **10351** palabras en mayúsculas

También buscamos cuáles son las 10 palabras con mayúsculas más frecuentes en el dataset [Figura 4].

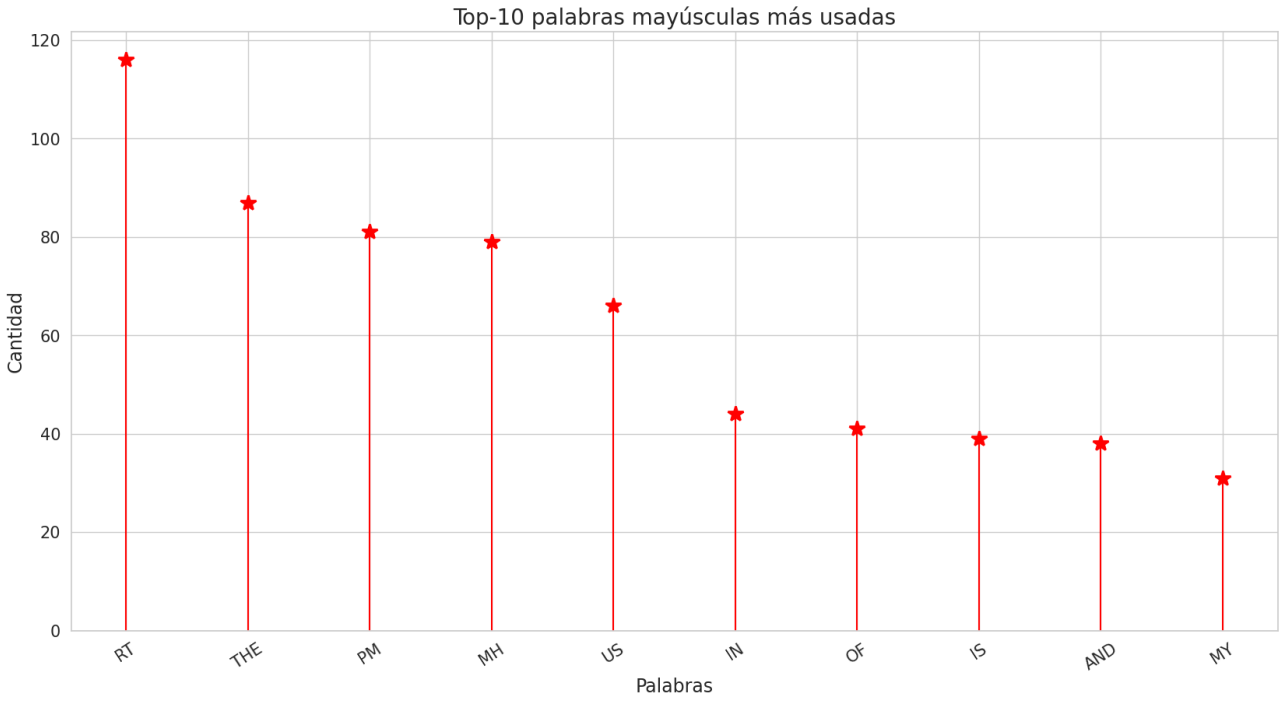

Figura 4: TOP-10 palabras con mayúsculas

3.2. Análisis de tweets según su target

Sabemos que, según el valor del *target*, un tweet puede contener información sobre **desastres reales (si su target es 1)** ó sobre **desastres falsos o información sin sentido (si su target es 0)**.

Haciendo manejo de esta información, vemos que en el dataset hay **4342 tweets falsos** y **3271 tweets reales**.

3.2.1. ¿Hay alguna relación entre la longitud de los tweets y la veracidad de los mismos

Para responder esta pregunta, observemos el siguiente plot [Figura 5].

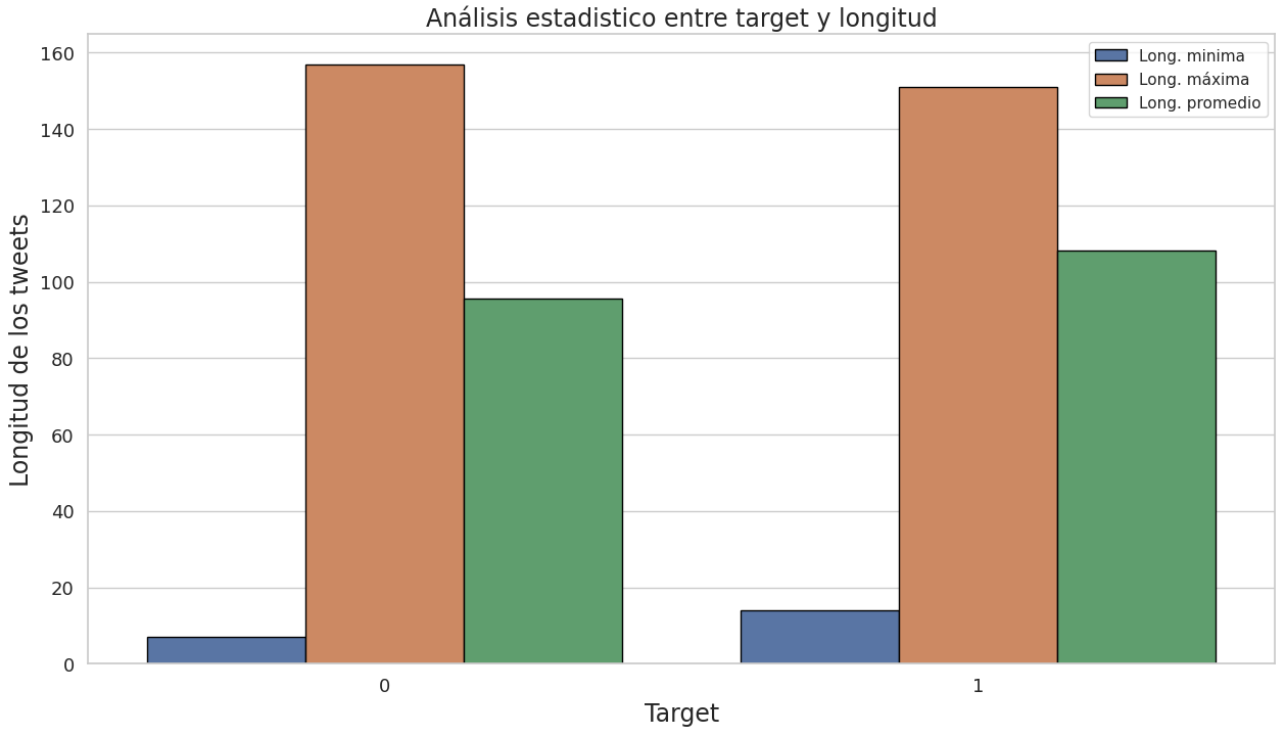

Figura 5: Longitud de los tweets en función de su target

- Tweet real más corto: **14** caracteres
- Tweet falso más corto: **7** caracteres
- Tweet real más largo: **151** caracteres
- Tweet falso más largo: **157** caracteres
- Longitud promedio de tweets reales: **108.11** caracteres
- Longitud promedio de tweets falsos: **95.71** caracteres.

La relación que pudimos sacar entre la veracidad de los tweets y su longitud es que tanto el tweet más corto como el más largo, son falsos. Además, en promedio, los tweets verídicos son más largos, ya que podrían contener información algo detallada sobre alguna emergencia o catástrofe de la que sus redactores fueron testigos. En cambio, los tweets de menor longitud (falsos), tienen un contenido que nada tiene que ver con emergencias.

3.2.2. Cantidad de menciones, hashtags y URLs según target

Analizando los tweets reales, llegamos al resultado de que hay **656** que poseen menciones hacia otros usuarios en su contenido, **858** que contienen hashtags, y **2172** con URLs.

Ahora, analizando aquellos que son falsos, obtuvimos **1338** tweets en cuyo texto hay menciones hacia otros usuarios, **885** que incluyen hashtags, y otros **1799** con URLs.

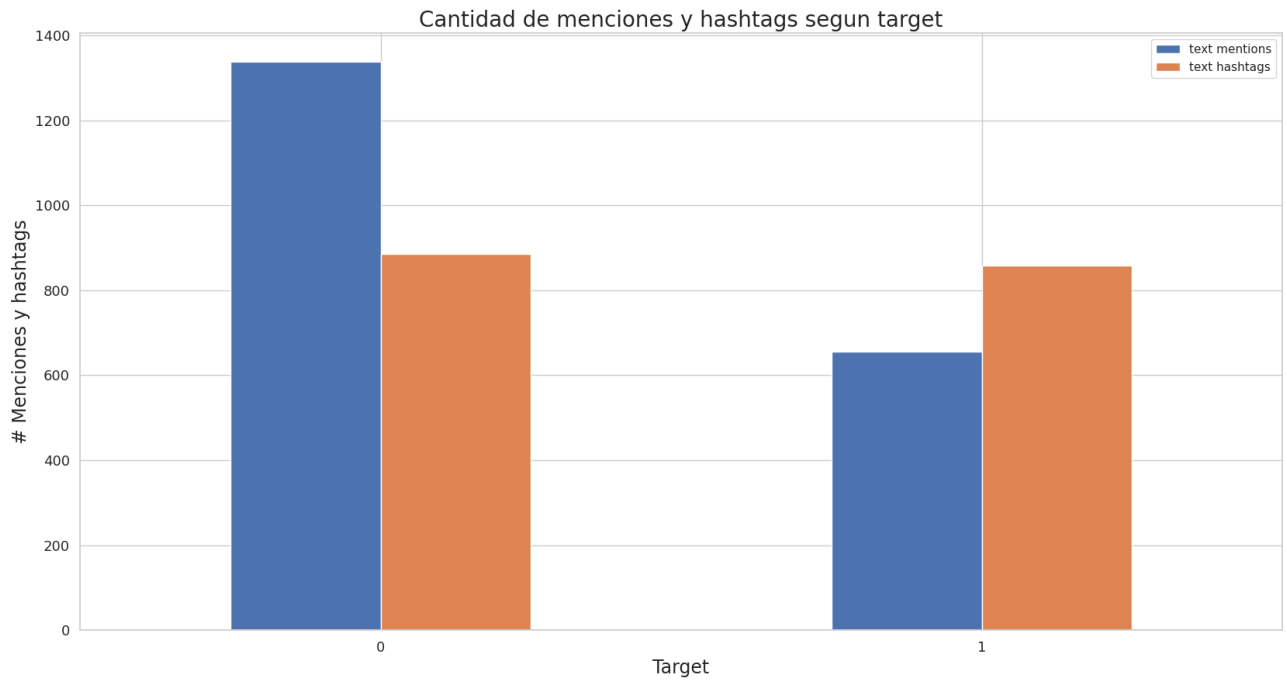

Figura 6: Cantidad de menciones y hashtags

Dentro de éste análisis, también analizamos que sucede con el hashtag mas utilizado: **#news**, observando que aparece más en tweets reales que en falsos [Figura 7].

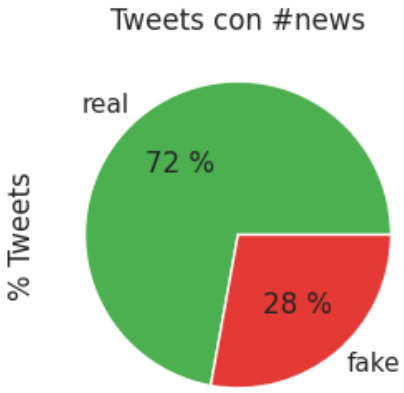

Figura 7: Porcentaje de aparición del hashtag más frecuente

Repetimos el análisis para el usuario de twitter más mencionado: **YouTube**, viendo que es mencionado mayormente en tweets falsos [Figura 8].

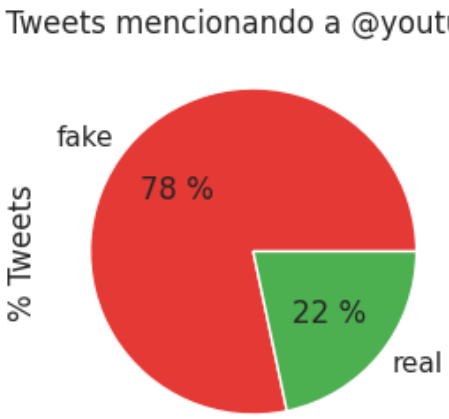

Figura 8: Porcentaje de aparición del usuario más mencionado

Si analizamos como se comportan los tweets que mencionan a **YouTube**, vemos que cuando son los únicos mencionados, la proporción de reales contra falsos crece levemente. En cambio, cuando además se menciona a otro usuario, los tweets son casi en su totalidad falsos.

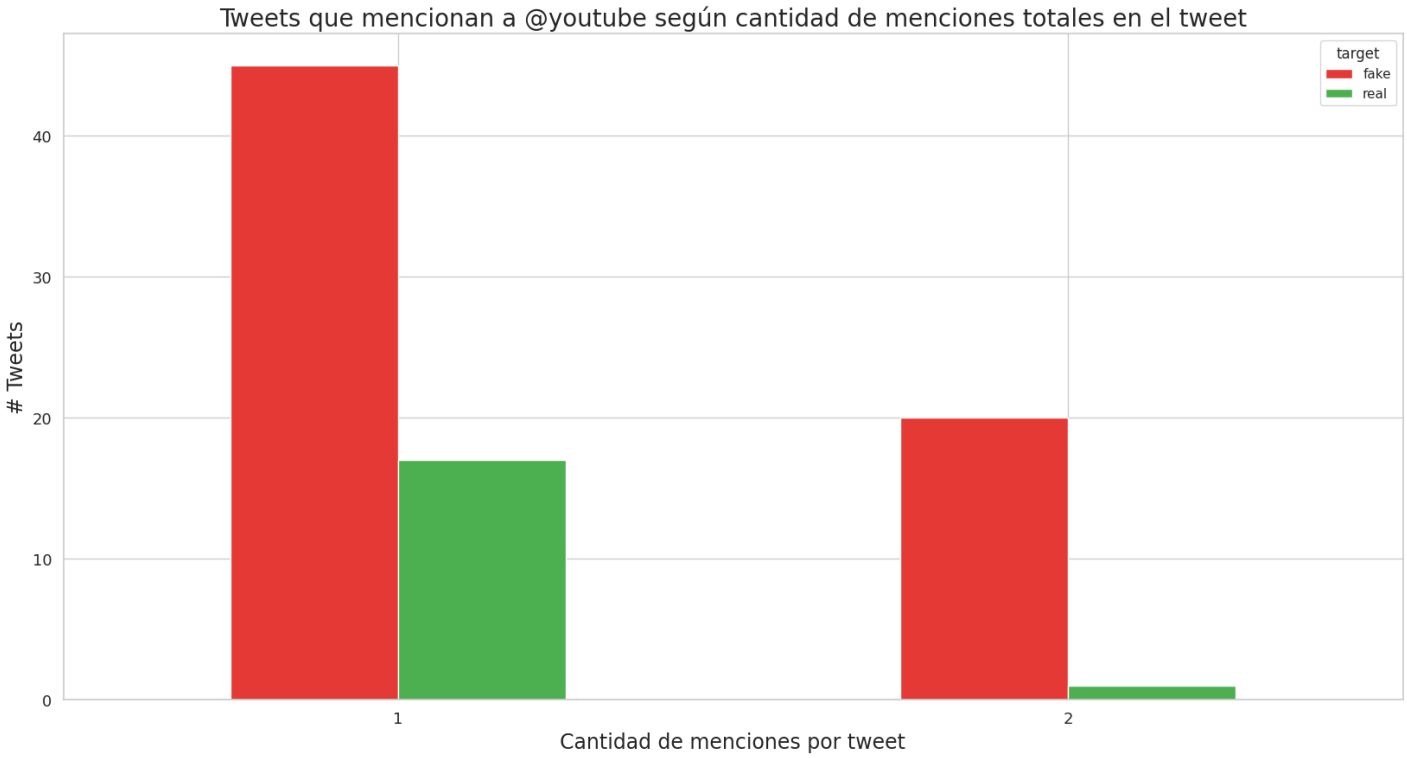

Figura 9

Analizaremos también la veracidad de los tweets que contienen la palabra **RT**, la más utilizada de las mayúsculas. De esta forma vemos que su utilización es pareja.

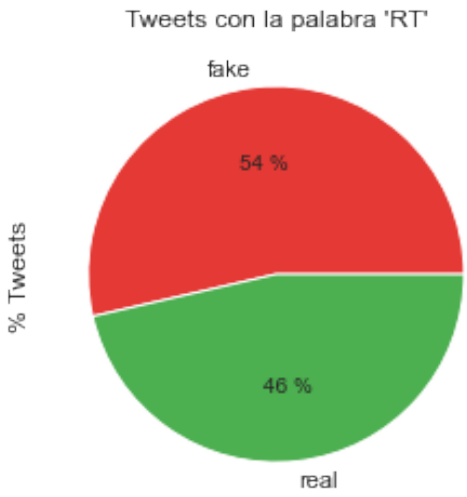

Figura 10: Porcentaje de aparición de la palabra **RT**

3.3. Análisis de palabras claves (*keywords*)

3.3.1. Limpieza de datos

Al analizar esta información de nuestro dataset, nos dimos cuenta que la gran mayoría poseían el símbolo " **%20**", que hace referencia a un código que representa espacios en blancos. Para hacer un uso correcto de la información, procedimos a eliminarlos con el uso de expresiones regulares de la librería **re**.

3.3.2. ¿Cuáles son las keywords más usadas en tweets reales?

Un análisis que nos pareció interesante, fue el de buscar cuáles son las 10 palabras claves que más se repiten en los tweets reales, lo que nos proporcionaría una información de qué desastres reales ocurrieron. Esto se ve en la siguiente figura: [Figura 11]

Figura 11: Las 10 keywords más utilizadas en tweets reales

Observando la figura, vemos que las keywords hacen referencia a escombros, suicidios, brotes, evacuados, bombas, rescatistas, descarrilamientos, destrucción, etc, lo que indica que son accidentes/desastres reales.

3.3.3. ¿Cuáles son las keywords más usadas en tweets falsos?

Mismo análisis que en la subsección anterior, pero esta vez enfocándonos en aquellos tweets falsos [Figura 12].

Figura 12: Las 10 keywords más utilizadas en tweets falsos

Cómo vemos, también se tratan de palabras que indican desastres/accidentes, pero su target nos dice que corresponden a tweets falsos.

Ahora, vamos a ver si hay alguna relación entre aquellas keywords que aparecen en los tweets reales, y aquellas que aparecen en los falsos. Nos vamos a centralizar en ver qué tantas palabras, de las más repetidas en los tweets reales, aparecen en los tweets falsos, y viceversa. Lo vamos a ver en un el siguiente plot [Figura 13].

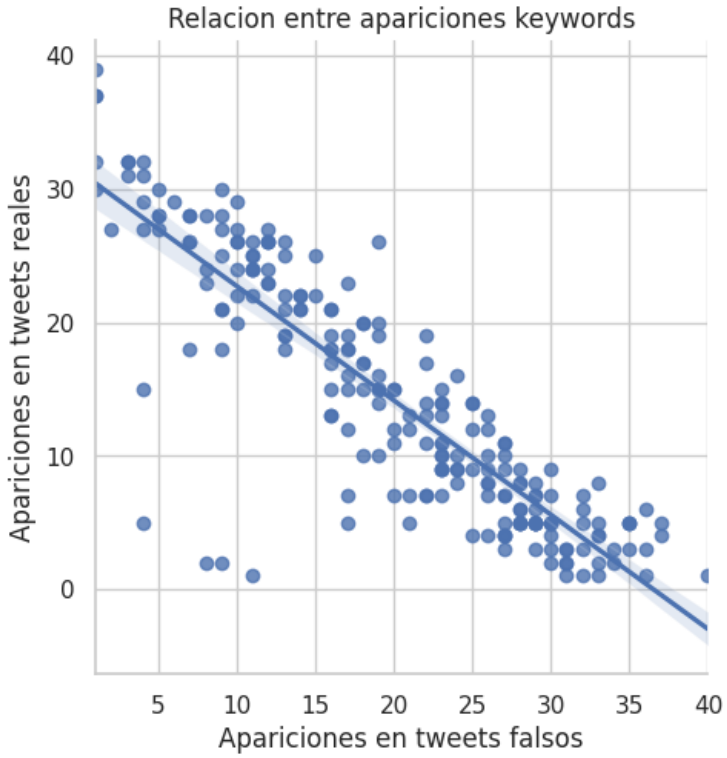

Figura 13: Relación entre apariciones de las keywords

Para explicar un poco el gráfico: los puntos en la parte superior izquierda indican que hay keywords que aparecen muchas veces en tweets reales pero pocas veces en tweets falsos; en la parte inferior derecha indican que hay keywords que aparecen muchas veces en tweets falsos, pero pocas en tweets reales. La linea representa una regresión lineal.

3.3.4. Raíces de las keywords

En los análisis anteriores sobre keywords, vimos que hay muchas de ellas que están relacionadas con muchas otras. Por ejemplo: ablaze, ablazing. Por ello, procedemos a buscar las raíces de esas palabras, utilizando la función *SnowballStemmer()* de la librería **nltk**.

En las siguientes figuras (14 y 15), podemos ver las raíces de las keywords más usadas en tweets reales y falsos respectivamente.

Figura 14: Raíces de keywords

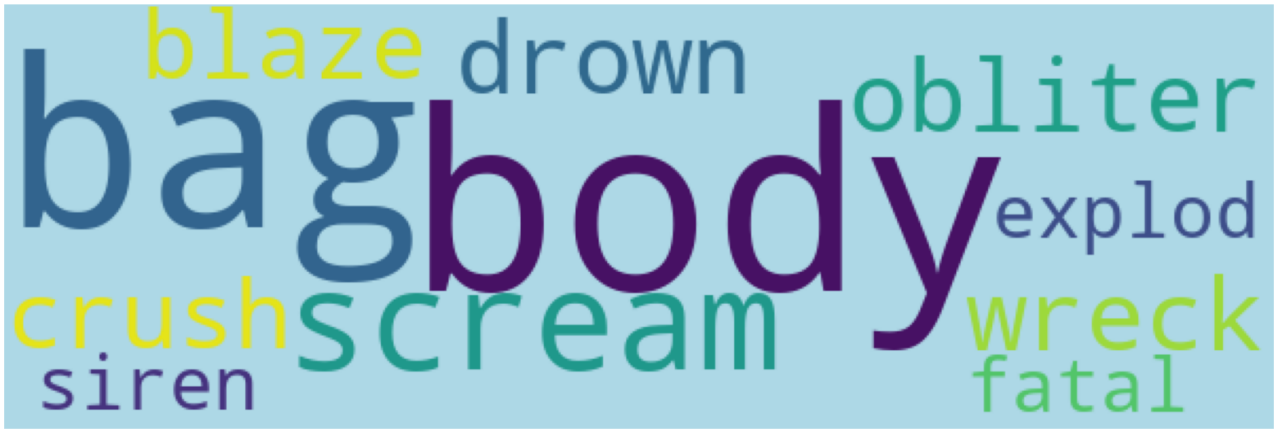

Figura 15: Raíces de keywords

3.4. Análisis de ubicación (location)

En esta sección, vamos a analizar todo lo relacionado con la columna *location*.

Buscaremos desde qué ubicación fueron redactados la mayoría de los tweets, tanto reales como falsos.

3.4.1. ¿Desde qué ciudades fueron escritos la mayoría de los tweets?

En las siguientes figuras (16 y 17), mostramos el resultado que obtuvimos al consultar desde qué ubicaciones fueron twitteados la mayoría de los tweets, tanto reales como falsos.

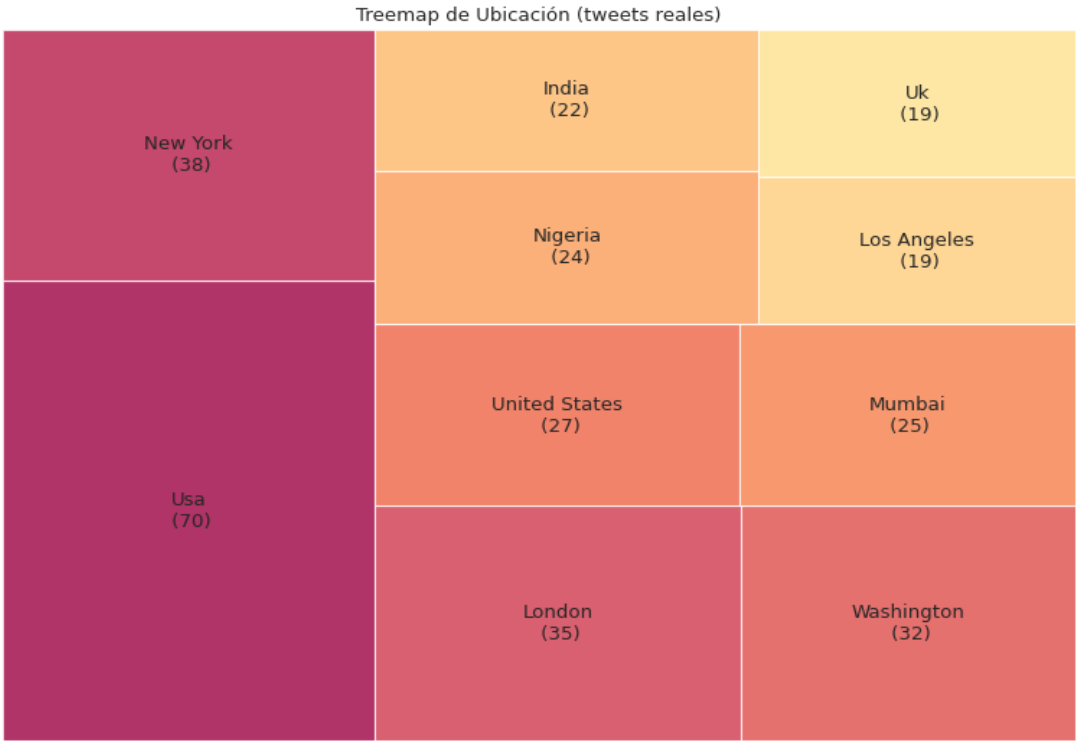

Figura 16: Top-10 de ubicaciones de tweets reales

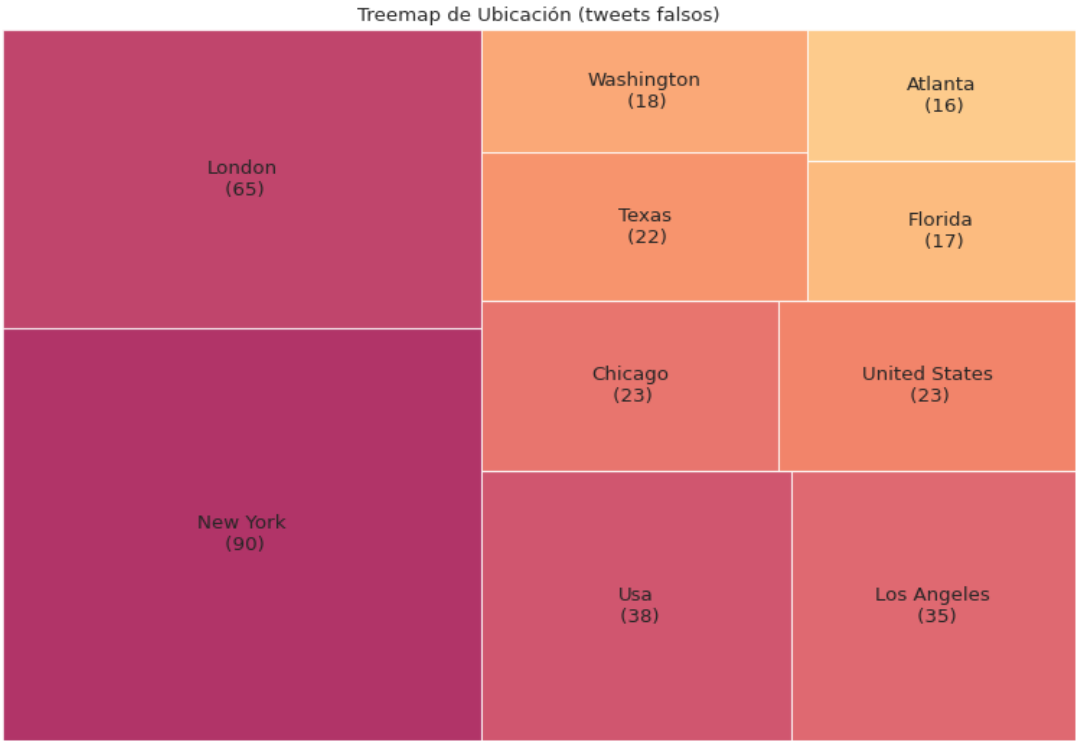

Figura 17: Top-10 de ubicaciones de tweets falsos

Como podemos ver, la mayoría de los tweets fueron escritos desde ciudades o estados pertenecientes a Estados Unidos. Esto nos lleva a pensar en que podría ser interesante hacer un análisis exploratorio, similar al que venimos haciendo, pero de todos los datos provenientes solamente de Estados Unidos.

3.5. Análisis exploratorio Estados Unidos

En esta sección, vamos a realizar un análisis profundo sobre los tweets emitidos desde Estados Unidos.

3.5.1. Limpieza de los datos

Luego de analizar los datos de la columna *location* de nuestro set de datos, se pudo observar que un gran porcentaje de los tweets en estudio fueron emitidos desde Estados Unidos. A pesar de ello esta columna presentaba muchos datos erróneos. Se procedió entonces a filtrar los datos para poder identificar los tweets que realmente fueron emitidos desde el país en estudio. Con éste objetivo se aplicaron 4 filtros.

1. Se eliminaron todos aquellos tweets que no poseían información en la columna *location* pues no hay manera de extraer información sobre su origen sin este dato. A partir de la siguiente etapa se comenzará a reconocer ubicaciones correspondientes a lugares de Estados Unidos. Para ello se utilizó un archivo auxiliar que contiene el nombre de todas las ciudades de Estados Unidos y su correspondiente estado asociado.

2. Se reconocen tweets en donde el texto que indica su ubicación mencione el nombre de alguna ciudad de Estados Unidos.

3. En esta etapa no quedan, en nuestro set de datos, tweets que contengan el nombre de alguna ciudad de USA. Sin embargo, puede haber tweets en donde su ubicación mencione algún estado de ese país. En esta etapa se localizaron dichos tweets.

4. Como ya fueron filtrados aquellos tweets cuya ubicación mencione el nombre de alguna ciudad o algún estado estadounidense, solo quedan en el set de datos aquellos que mencionen el nombre del país en cualquiera de sus variantes. El objetivo de esta etapa es detectar dichos tweets.

5. Se realizan algunas correcciones manuales que se detectaron a la hora de observar nuestros resultados.

Ya tenemos nuestro set de datos listo para poder comenzar a trabajar.

3.5.2. Primera mirada

Veamos una primera mirada de los datos y calculemos a nivel país cuál es el porcentaje de tweets verdaderos y falsos que fueron emitidos desde el territorio estadounidense.

Se observa que la cantidad de tweets emitidos por cada categoría es muy pareja, encontrando una diferencia porcentual de tan solo 5 puntos. Se observa que se mantiene la tendencia encontrada en el análisis general donde la cantidad de tweets falsos supera a la cantidad de tweets verdaderos pero disminuyéndose la brecha entre ambas.

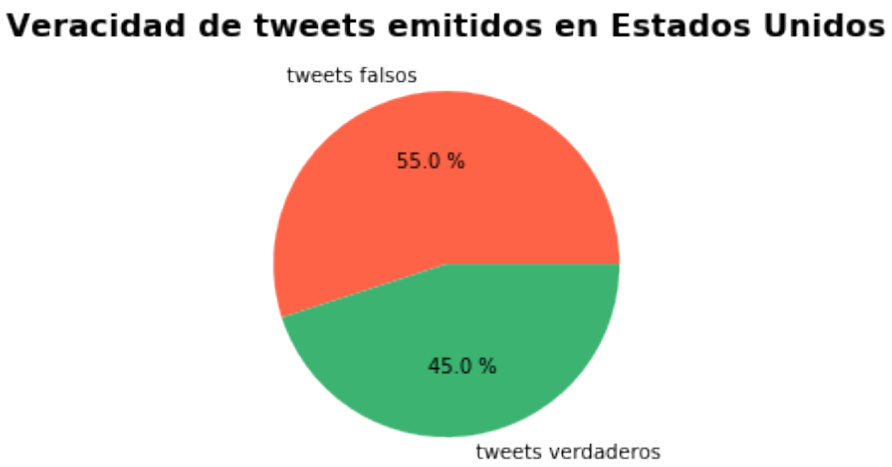

Figura 18: Porcentaje de tweets reales y falsos en EEUU

3.5.3. Tweets verdaderos y falsos por estado

Resulta interesante entonces analizar un poco más en profundidad y ver cómo se distribuyen estos tweets entre los distintos estados. Lamentablemente el análisis no podrá ser del todo preciso, ya que para una cierta cantidad de tweets solo pudimos detectar su país de origen pero no el estado ni la ciudad al que pertenecen. Por otra parte los tweets para los que se reconoció su ciudad y no su estado, sí forman parte de este análisis, ya que conocida la ciudad fue posible ubicar el estado en el que se encuentra dicha ciudad.

La siguiente visualización nos permite apreciar la diferencia entre la cantidad de tweets falsos y la cantidad de tweets verdaderos emitidos en cada estado (verdaderos - falsos). Se destacan cuatro estados clave: **Texas, Florida, Tennessee y New York**.

En éste último es donde se registra la mayor diferencia, siendo 116 la cantidad de tweets falsos emitidos, 50 la cantidad de tweets verdaderos emitidos y una diferencia de 66 en valor absoluto. Los siguientes dos países en donde encontramos una diferencia a favor de los tweets falsos, menor pero no menos importante, son **Texas y Florida**. Para Texas se obtuvo 77 tweets falsos, 56 verdaderos y una diferencia de 21 (valor absoluto), mientras que para Florida 54 tweets falsos, 30 verdaderos y una diferencia de 24. En el resto de estados en donde la cantidad de tweets falsos fue superior, la situación estuvo más pareja, no habiendo en ningún caso una diferencia superior a 20 tweets.

Si nos centramos ahora en los tweets verdaderos, sí encontramos un dato muy curioso pues en casi ningún estado la cantidad supero a la de los falsos en más de 10 unidades, solo en el estado de **Tennessee**, el cuál además es el estado en el que los tweets verdaderos obtuvieron la mayor ventaja y superaron a los falsos en 12 unidades, notando lo leve que es si la comparamos con la supremacía de tweets falsos encontrada en los tres estados anteriores.

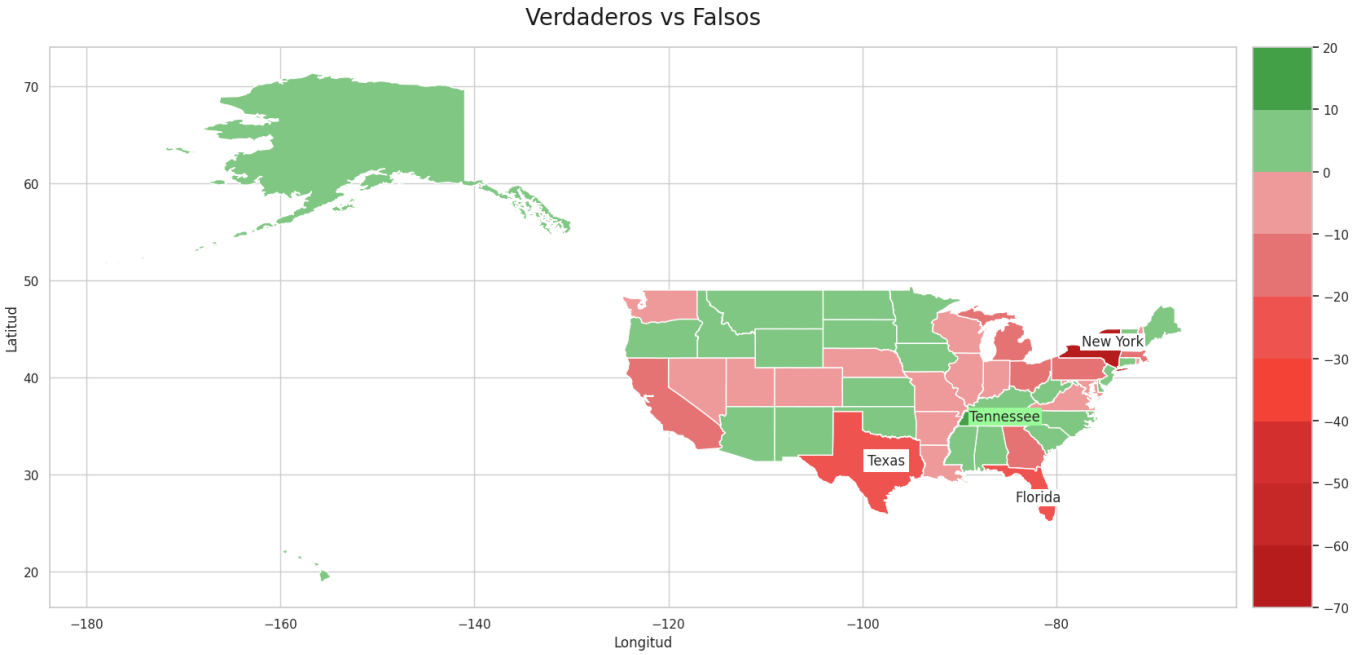

Figura 19: Distribución de tweets en los diferente estados

3.5.4. Análisis de la longitud de los tweets

Otro análisis interesante es revisar si hay alguna relación entre la longitud de los tweets y su veracidad.

Se analizó para cada target cuál fue la mayor longitud de un tweet, cual la menor y cuál la longitud promedio. Se observa que los tweet verdaderos son en promedio más largos que los tweet falsos. Sin embargo tanto el tweet más corto como más largo son falsos.

3.5.5. Analizando ciudades

Enfoquémonos ahora en las diferentes ciudades de **USA**. Encontremos cuales de ellas son las principales emisoras de tweets y en cuales longitud de los tweets falsos y verdaderos se encuentran por encima y por debajo de los respectivos promedios. Nuevamente resulta importante aclarar que los resultados no son exactos, ya que solo para algunos de los tweets emitidos desde USA fue posible encontrar su ciudad.

Poedemos dividir a la ciudades en categorias (A, B ,C y D) según si la longitud de los tweets verdaderos y falsos superan los promedios obtenidos a nivel pais:

- **A**: Supera longitud promedio para tweets falsos y para tweets verdaderos
- **B**: Supera longitud promedio para tweets verdaderos pero no para tweets falsos
- **C**: Supera longitud promedio para tweets falsos pero no para tweets verdaderos
- **D**: No supera la longitud promedio para tweets falsos ni tampoco para tweets verdaderos

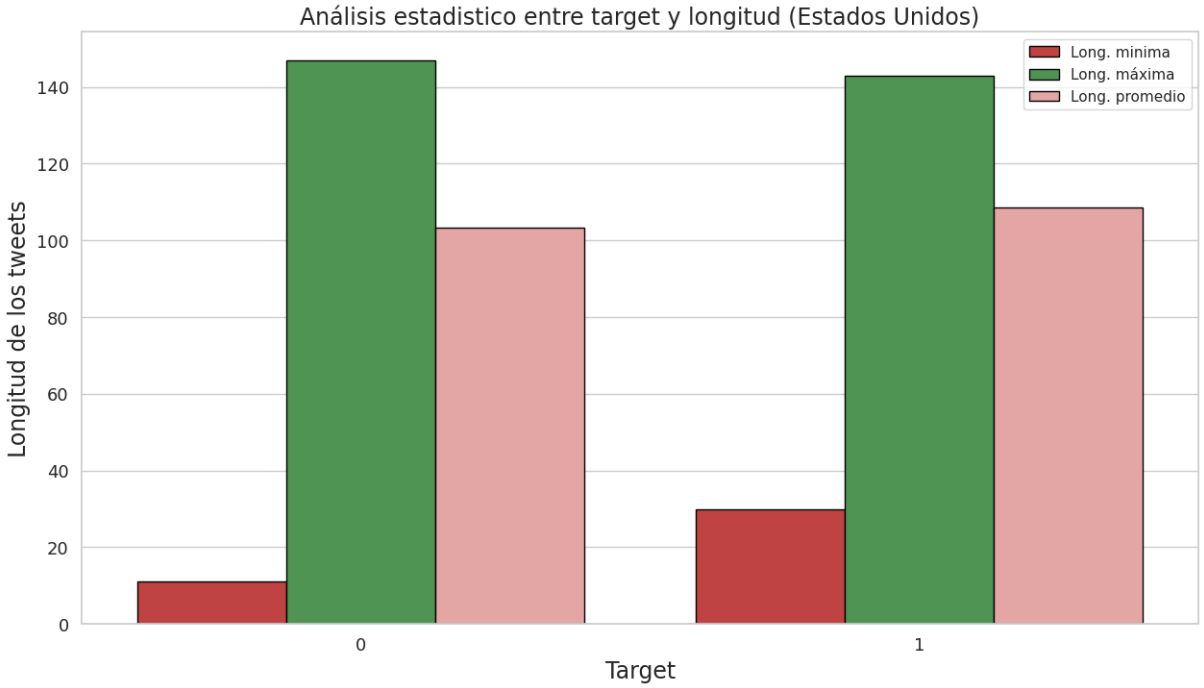

Figura 20: Longitudes de los diferentes tweets emitidos desde Estados Unidos

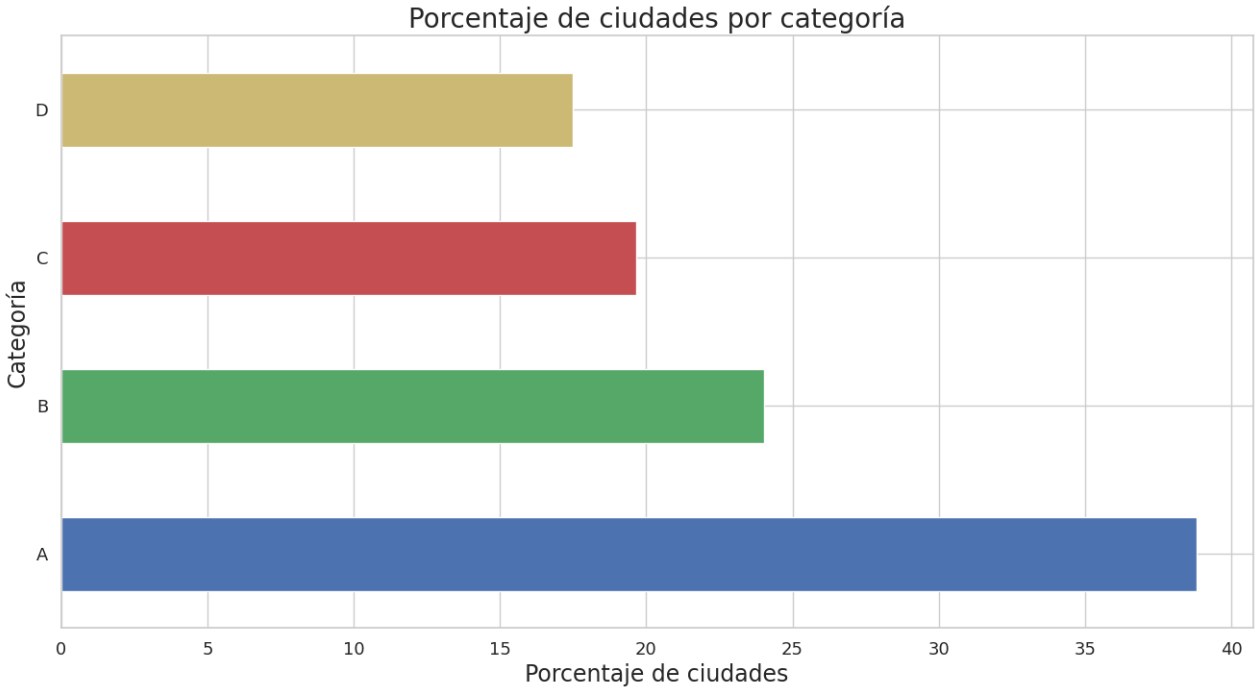

Figura 21: Porcentaje de ciudades por categoría

Detengámonos en aquellas ciudades desde las que se emitieron 30 o más tweets (indistintamente de su target), ya que son aquellas de las que se puede hacer un análisis un poco más concluyente.

Se encontró que hay solo 5 ciudades que cumplen con el requisito puesto: **Atlanta, Chicago, Los Ángeles, New York y San Francisco**. Los resultados obtenidos se muestran en la siguiente tabla:

Ciudad	Cantidad target 0	Longitud promedio target 0	Excede promedio target 0?	Cantidad target 1	Promedio target 1	Excede promedio target 1?	Cantidad total de tweets
Atlanta	22.00	102.50	NO	11.00	112.09	SI	33.00
Chicago	27.00	100.48	NO	22.00	110.14	SI	49.00
Los Angeles	35.00	110.29	SI	20.00	118.80	SI	55.00
New York	111.00	115.95	SI	47.00	118.87	SI	158.00
San Francisco	15.00	98.80	NO	17.00	99.24	NO	32.00

Figura 22: Principales ciudades emisoras

De las 5 ciudades desde las que se emitieron más tweets, en 4 de ellas la cantidad de tweets falsos emitidos supera a la cantidad de tweets verdaderos. La excepción es San Francisco, aunque la diferencia es de solo 2 tweets. En contraposición al caso de San Francisco, en las restantes cuatro ciudades, la cantidad de tweets falsos supera a la de tweets verdaderos en por lo menos 5 tweets en el caso más leve.

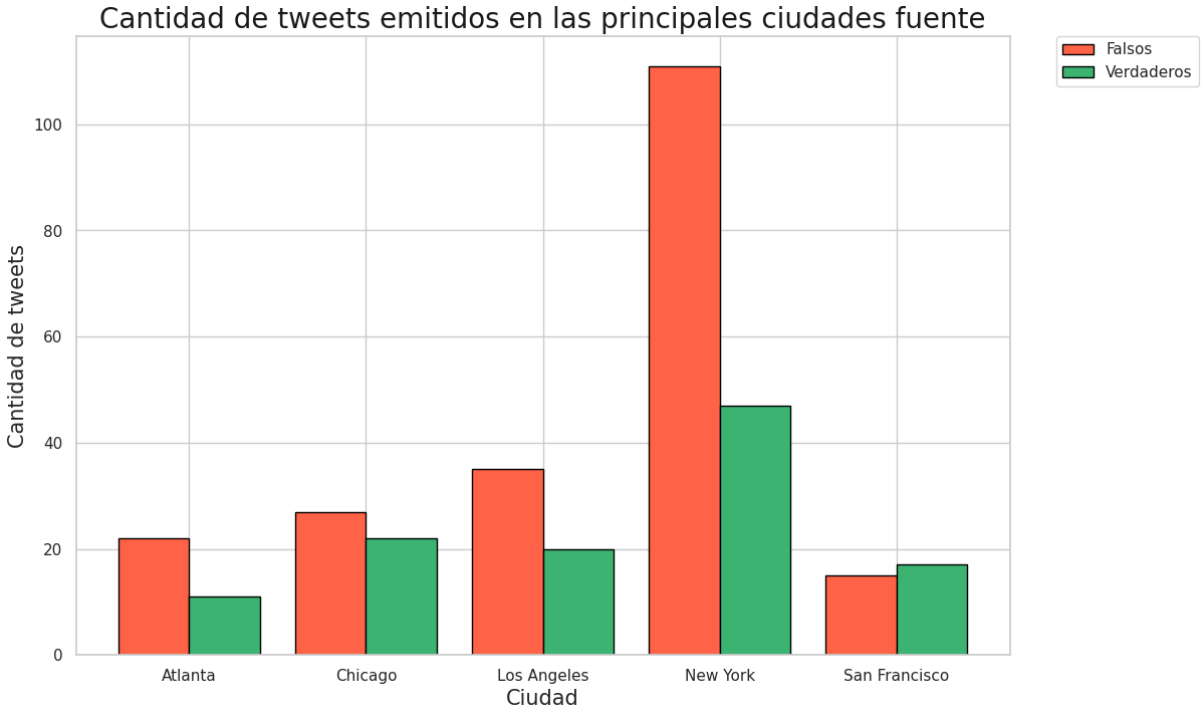

Figura 23: Cantidad de tweets emitidos por target en las principales ciudades emisoras

A pesar de ser la única ciudad con una cantidad significativa de tweets en donde la cantidad de tweets verdaderos supera a los falsos, San Francisco es la única ciudad de las 5 que cumplen con los requisitos en donde la longitud de tweets verdaderos se encuentra por debajo de la longitud de tweets verdaderos a nivel país.

Por otra parte resulta curioso que en las restantes 4 ciudades en donde se da el caso opuesto y la cantidad de falsos supera a los verdaderos, en todas, la longitud promedio de tweets verdaderos supera a la longitud promedio de tweets verdaderos a nivel país.

En cuanto a la emisión de tweets falsos la situación es más variada. En tres de las cinco ciudades la longitud promedio de los tweets falsos no supera al promedio en esa categoría a nivel país, mientras que en las dos restantes sí lo hace.

3.5.6. Hashtags y menciones

Volviendo a un análisis más global considerando la emisión de tweets a nivel país veamos la cantidad de tweets para cada target por cantidad de hashtags.

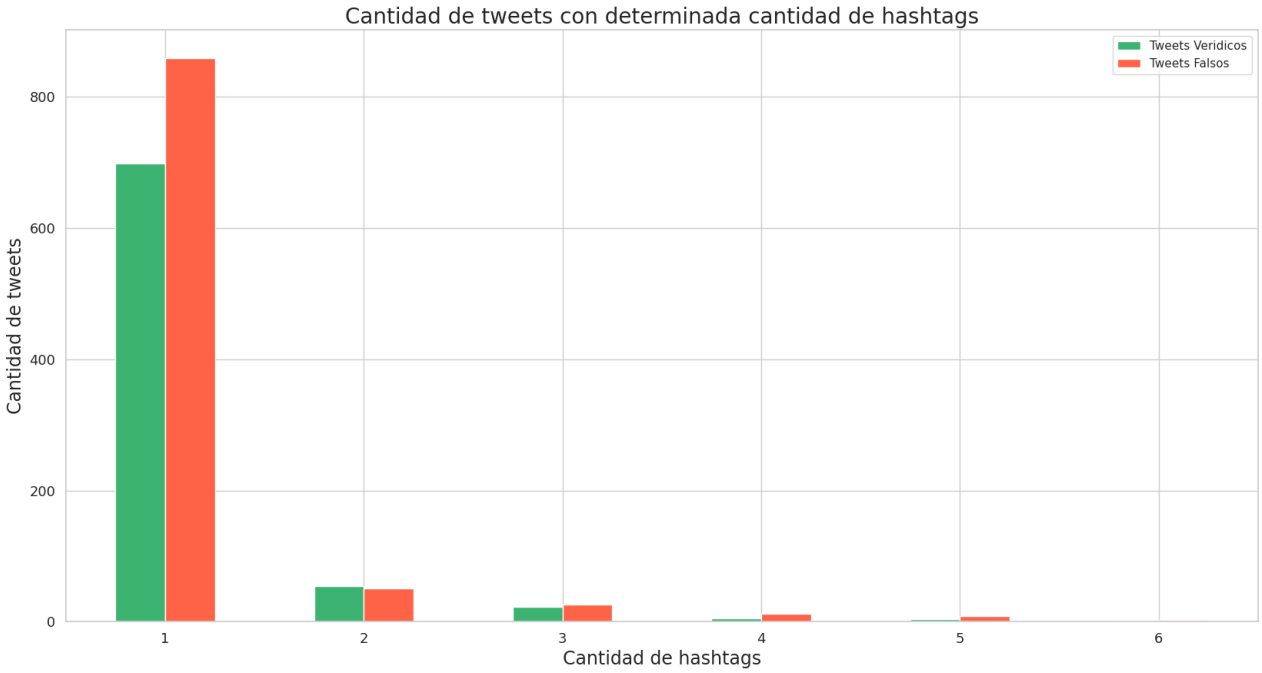

Figura 24: Cantidad de tweets por target que contiene cada cantidad de hashtags

Claramente la mayoría de los tweets tanto verdaderos como falsos contienen un solo hashtags observándose una enorme diferencia con los que tienen dos o más. Por otra parte entre la cantidad de tweets que contienen un solo hashtag se observa que una mayor cantidad de tweets falsos que verdaderos. Un análisis similar hacemos para las menciones [Figura 25].

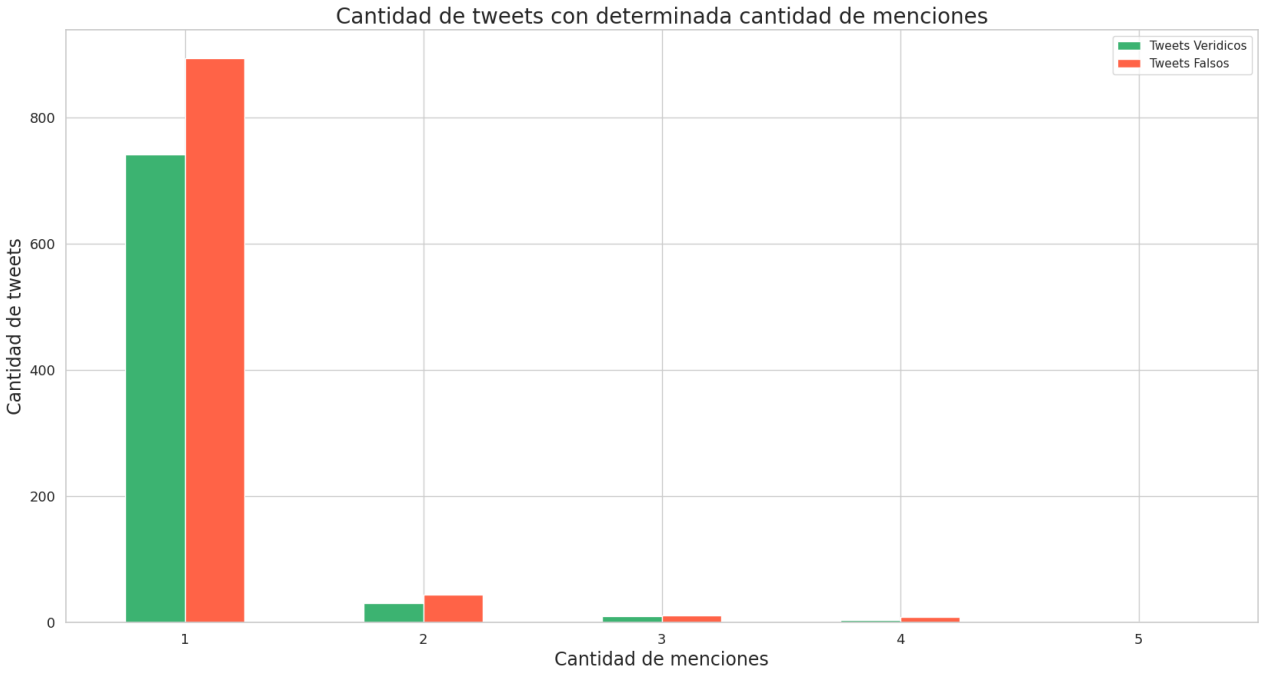

Figura 25: Cantidad de tweets por target que contiene cada cantidad de menciones

Si comparamos las gráficas correspondientes a hashtags y menciones vemos que se comportan de manera muy similar, de hecho los gráficos son casi idénticos. La mayoría de los tweets contiene una sola mención habiendo una enorme diferencia con la cantidad de tweets que contienen dos o más menciones.
Por otra parte también es superior la cantidad de tweets falsos que contiene una mención con respecto a la cantidad de tweets verdaderos que contienen la misma cantidad.
Entre ambas gráficas hay solo una pequeña diferencia y es que para los hashtags es levemente superior la cantidad tweets verdaderos que contiene dos hashtags con respecto a los falsos mientras que en el caso de las menciones nuevamente como a lo largo de todo el análisis la cantidad de tweets falsos vuelve a ser superior.

3.6. URLs

Aquí encontramos una diferencia notoria ya que si miramos los tweets que no contienen ninguna URL, se observa una superioridad grande por parte de la cantidad de tweets falsos sobre los verdaderos.

Si miramos la cantidad de tweets que tiene un URL la tendencia se invierte y la cantidad de tweets verdaderos es superior a la cantidad de falsos, aunque la diferencia no es tanta.

Con dos URLs la situación vuelve a invertirse y nuevamente la cantidad de falsos pasa a superar a la cantidad de verdaderos, pero en ambos casos la cantidad de teeets disminuye significativamente con respecto a los dos casos anteriores.
La cantidad de tweets con 3 URLs para ambos target es muy pequeña.

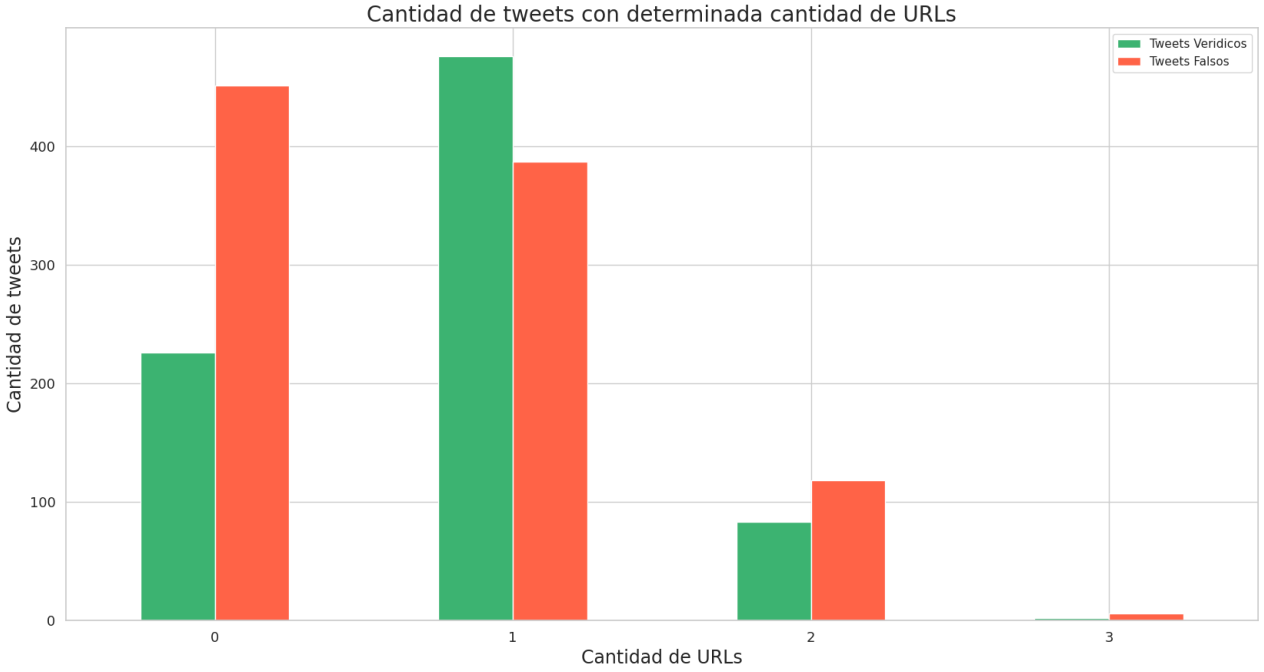

Figura 26: Cantidad de tweets por target que contiene determinada cantidad de URL

4. Conclusión

Dentro de este análisis podemos obtener varias conclusiones:

En primer lugar , tal como lo describimos a lo largo del informe podemos destacar como la mayor parte de los tweets se originaron en Estados Unidos. Aquí las tendencias de veracidad dieron muy parejas, pero hubo una leve inclinación hacia los tweets falsos.

En segundo lugar, podemos notar como los tweets verdaderos tienden a ser más descriptivos brindando la mayor cantidad de datos posibles y es por eso que su promedio de caracteres es superior al de los tweets falsos. En cuanto al contenido de los mismos, observamos que los verdaderos son más concisos con la información que brindan utilizando menos menciones y más URLs las cuales en su mayoría son imágenes que podrían referir a las distintas catástrofes. Por otra parte, los Hastags tienen una distribución pareja y podemos notar una diferencia en aquellos cuya temática es de noticias dado que tienden a estar asociados a los tweets verdaderos.

Por último , encontramos las palabras claves y podemos notar que todas las que se relacionan con problemas climáticos, muertes y bombas (entre otras) forman parte del contenido de los tweets verdaderos con mayor frecuencia.