FACULTAD DE INGENIERIA
Universidad de Buenos Aires

# U.B.A. FACULTAD DE INGENIERÍA

Departamento de Computación
Organización de Datos 75.06/95.58
1er Cuatrimestre de 2020

# Trabajo práctico Nº1


| *Apellido y Nombre:* | *Padrón:* |
|---|---|
| Coronel, Dalma | 92257 |
| Gonzalez, Francisco | 71607 |
| Peña, Alejandro Nicolas | 98259 |
| Minino, Alan Nahuel | 99599 |

*Grupo: Unidos por los datos*


Fecha de Entrega : 21/05/2020

# Índice

# 1. Introducción

En el siguiente informe se detalla el análisis exploratorio realizado sobre el set de datos de la competencia de Kaggle: `https://www.kaggle.com/c/nlp-getting-started/overview/description`, en la cual se propone analizar tweets para poder monitorear emergencias en tiempo real.

Para este trabajo nos enfocamos en el análisis del set de datos "train"; el cual contiene tweets de personas donde ya se determinó si el tweet resultaba ser de una emergencia verdadera o no, como se indica en la columna "target". En el análisis realizado, se buscó encontrar características que compartieran los tweets con el mismo grado de verdad y que pudiera servir de indicio a la hora de determinar si un tweet sobre una emergencia es verdadero o falso.

El análisis presentado se hizo utilizando Python3 con las siguientes librerías:

- Pandas
- Numpy
- Matplotlib
- Seaborn

A lo largo del informe se pueden ver distintas visualizaciones de los datos, las cuales fueron hechas con la intención de facilitar la comprensión de la información disponible.

Se utilizó un repositorio en GitHub: `https://github.com/alepenaa94/TP1_Real_or_Not`, como herramienta de integración, donde se puede encontrar el notebook utilizado para el análisis exploratorio.

# 2. Limpieza y manejo de los datos

El set de datos proporcionado por la competencia estaba compuesto por los siguientes campos:

- id - un identificador único para cada tweet
- keyword - una palabra clave particular del tweet
- location - la ubicación desde donde el tweet fue enviado
- text - el texto del tweet
- target - indica si un tweet es sobre un desastre verdadero (1) o no (0)

Al leer el archivo csv con los tweets, especificamos los tipos de datos que contenían los campos *ID* y *TARGET* para reducir el espacio que ocupaba la información del set. Convertimos el campo *ID* al tipo entero sin signo de 16 bits (ya que el id comprendía valores del 1 al 10873) y *TARGET* al tipo booleano (ya que es el campo que indica si el tweet es verdadero o falso).

Luego, vimos la información del dataframe para tener un detalle de los campos, la cantidad de registros y ver cuáles contenían nulos.

```
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        7613 non-null   uint16
 1   keyword   7552 non-null   object
 2   location  5080 non-null   object
 3   text      7613 non-null   object
 4   target    7613 non-null   bool
```

Pudimos ver que estábamos trabajando con un set de datos de 7613 registros y que las columnas *KEYWORD* y *LOCATION* contenían valores nulos.

Para el caso del campo *KEYWORD*, cómo solo eran 61 registros nulos de todo el set, decidimos descartar dichos registros para los análisis particulares de ese campo. Además, al analizar los distintos valores que tomaba el campo, notamos que aparecían los caracteres " %20" en los lugares en donde iba un espacio como se puede ver a continuación:

| | |
|---|---|
| body%20bags | 41 |
| oil%20spill | 38 |
| burning%20buildings | 37 |
| cliff%20fall | 36 |
| nuclear%20reactor | 36 |
| dust%20storm | 36 |
| buildings%20burning | 35 |
| emergency%20plan | 35 |
| airplane%20accident | 35 |
| structural%20failure | 35 |

Esto se debe a que el " %20" representa al espacio en código URL. Por lo tanto, decidimos realizar el reemplazo de los caracteres por el espacio para una mayor claridad a la hora de visualizar la información.

Para el caso de *LOCATION*, cómo eran más de 2500 registros con valores nulos, completamos dichos registros con el valor "Unknown" a la hora de analizar esta columna. Por otro lado, observamos que la ubicación ingresada en muchos casos no es algo coherente como se puede observar a continuación:

| | |
|---|---|
| 5255 | Clean World |
| 7496 | United States |
| 2713 | NaN |
| 4427 | Glenview to Knoxville |
| 2539 | Istanbul |
| 6333 | Asia |
| 2041 | Boston, MA |
| 815 | Waterfront |
| 28 | NaN |
| 2608 | Illinois, USA |
| 4186 | a van down by the river |
| 5121 | NaN |
| 2789 | en el pais de los arrechos |
| 1046 | Bushkill pa |
| 2796 | Philadelphia, PA USA |
| 1797 | Melbourne, Australia |
| 1512 | Bouvet Island |
| 7344 | Washington State |
| 2424 | DC |
| 990 | #EngleWood CHICAGO |
| 176 | West Wales |
| 1600 | Kolkata, India |
| 766 | NaN |
| 5211 | don't buy the s*n |
| 1140 | WorldWide |
| 4109 | Newton Centre, Massachusetts |
| 4201 | Kenya |
| 341 | NaN |

Viendo esto, se podría esperaba que en esos casos el tweet tenga mayor probabilidad de ser falso.

# 3. Análisis exploratorio

## 3.1. Target

Como primer paso, decidimos ver comó estaba distribuída la cantidad de tweets del set de datos según si son verdaderos o falsos.


Figura 1: Cantidad de tweets según target.

Pudimos ver en la figura 1 que casi el 43 % de los tweets describían realmente un desastre en su mensaje (4342 tweets falsos contra 3271 verdaderos).

## 3.2. Keyword

Analizando el campo keyword vimos que, si bien hay 7552 registros que tienen valor no nulo, este solo puede tomar 221 valores únicos. De entre estos valores, la keyword "FATALITY" era la de mayor frecuencia, con 45 apariciones, y "RADIATION EMERGENCY" la de menor, con 9 apariciones.

Hicimos un análisis de las 20 keyword con mayor frecuencia para ver de que tipo de palabras se trataba.

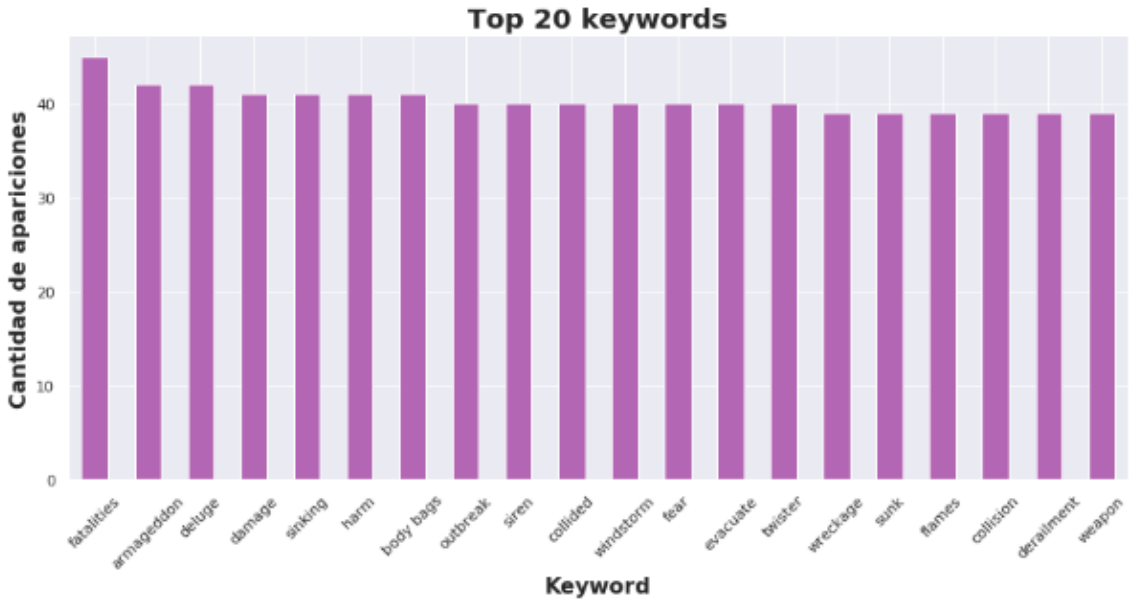


Figura 2: Keywords con mayor frecuencia.

Como era de esperarse, vimos que todas las keywords que aparecían en la figura 2 eran palabras relacionadas a desastres.

Continuando con el análisis, decidimos ver cuales eran las keyword que más se utilizaban en tweets de un target pero que no tuvieran mucho uso en los tweets del target opuesto. En base a esto, decidimos considerar solo los keywords que tuvieran muchas ocurrencias. En principio obtuvimos algo de la siguiente forma:

| | keyword_true | keyword_false |
|---|---|---|
| outbreak | 39 | 1 |
| typhoon | 37 | 1 |
| oil spill | 37 | 1 |
| rescuers | 32 | 3 |
| suicide bomb | 32 | 3 |
| ... | ... | ... |
| body bags | 1 | 40 |
| electrocute | 1 | 31 |
| blazing | 1 | 33 |
| epicentre | 1 | 11 |
| body bag | 1 | 32 |

Usando estos valores, creamos distintos GroupedBarPlot en donde se pudiera visualizar estos keywords.

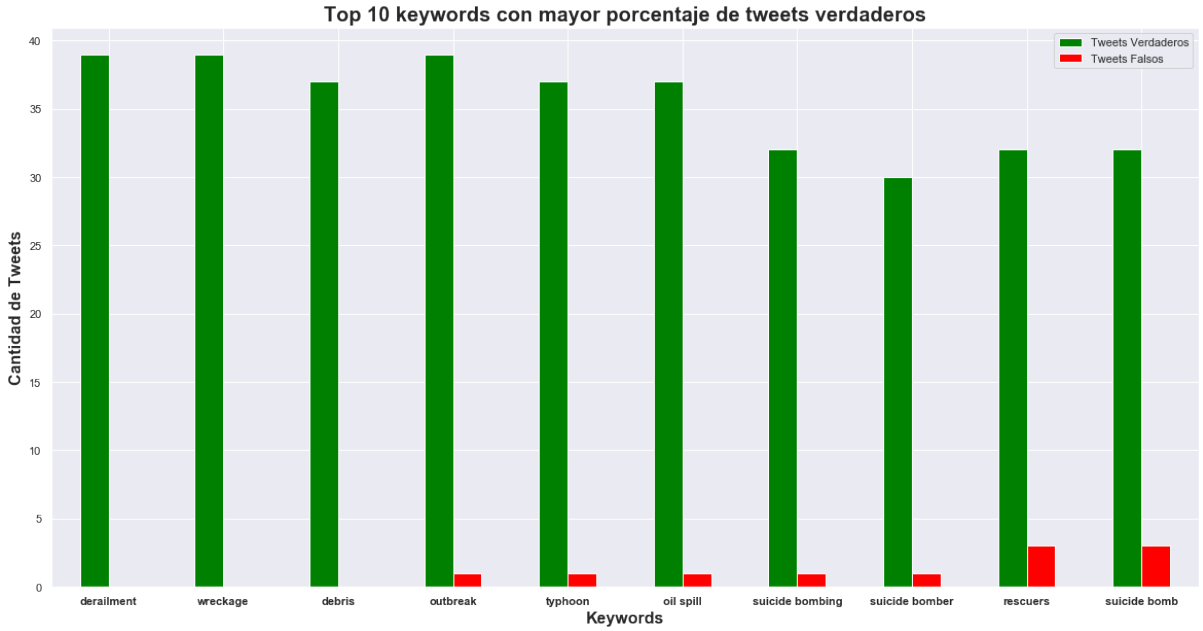


Figura 3: Keywords con mayor porcentaje de tweets verdaderos.

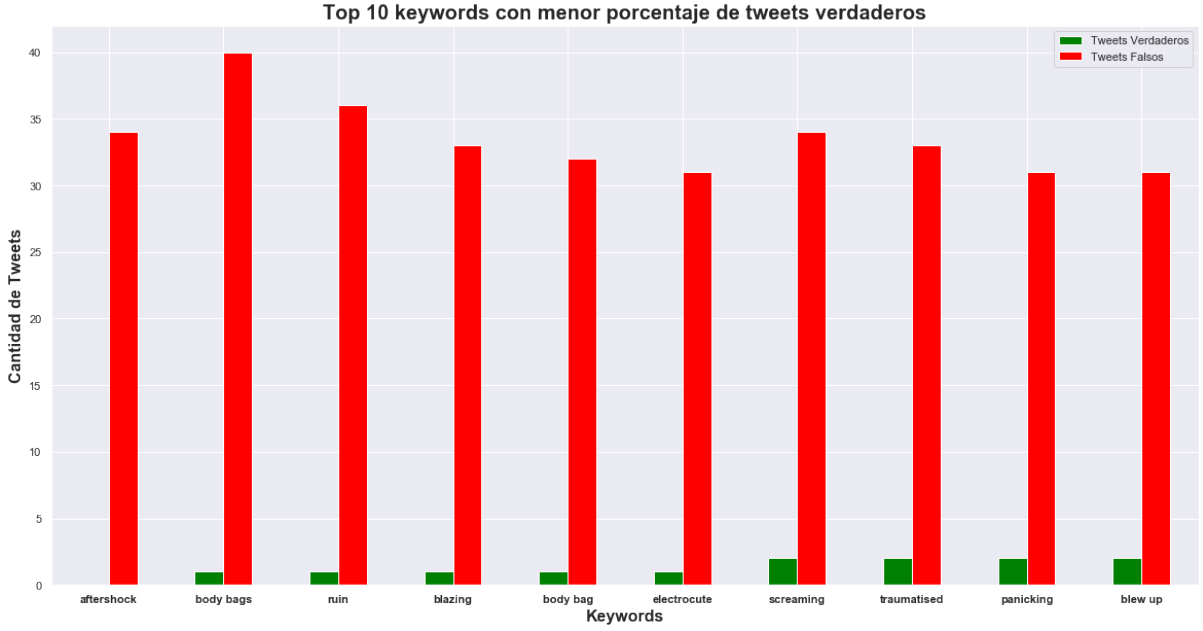


Figura 4: Keywords con menor porcentaje de tweets verdaderos.

En las figuras 3 y 4, pudimos observar que hay temas particulares sobre los cuales suele mentir la gente en los tweets, como terremotos de pequeña escala, cadáveres, incendios, entre otros; si se quiere, desastres que llamen la atención y genere que mucha gente vea el tweet. En contraste, hay temas con muy pocos tweets falsos como descarrilamientos, brotes de enfermedades o tifones, desastres de gran escala y, tal vez, que pueda identificarse si el desastre expuesto realmente sucedió sin mucha dificultad.

Algo a destacar es que, entre las keywords observadas, notamos variaciones de un mismo tema que podría agruparse en un único keyword.

Siguiendo la misma línea de pensamiento, decidimos ver las keywords con una cantidad similar de uso tanto en tweets verdaderos como falsos. Para esto, consideramos como similar cantidad de uso a las keywords con un porcentaje de tweets verdaderos de entre 48 % y 52 %.

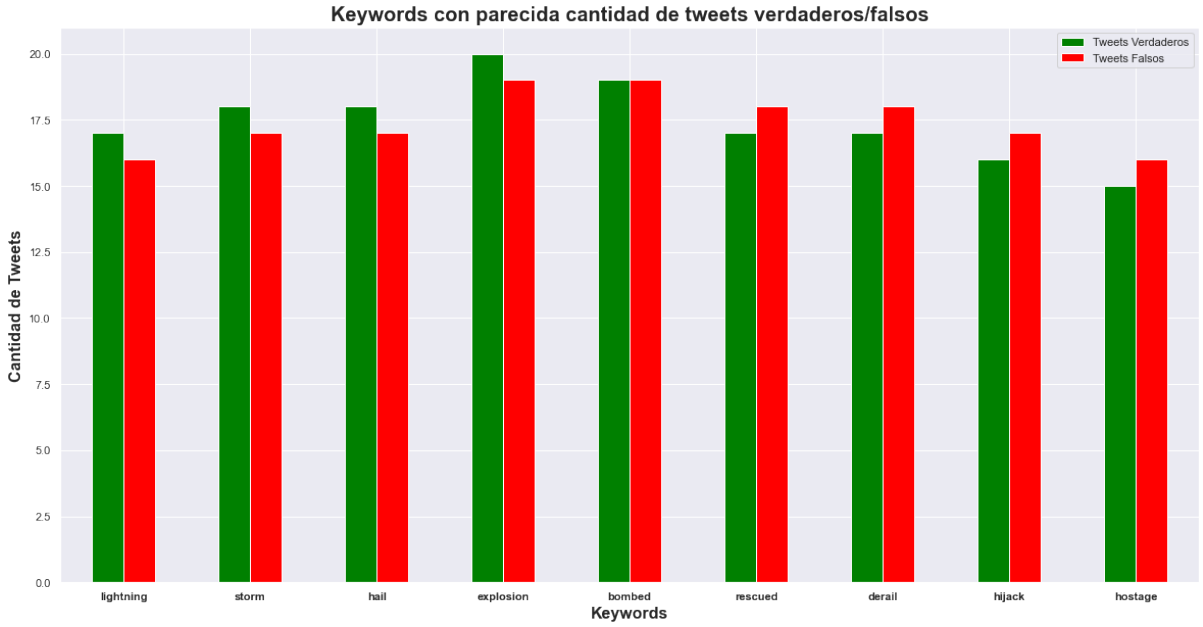


Figura 5: Keywords con parecido porcentaje de tweets verdaderos y falsos.

En la figura 5, vimos que los temas que tienen uso constante en ambos target son más variados yendo desde tormentas, rayos y explosiones; hasta secuestros, rehenes y rescates.

Como no es completamente claro el comportamiento de las keywords analizando simplemente los extremos, deicimos ver la relación de ocurrencias entre keywords con tweets verdaderos y falsos.

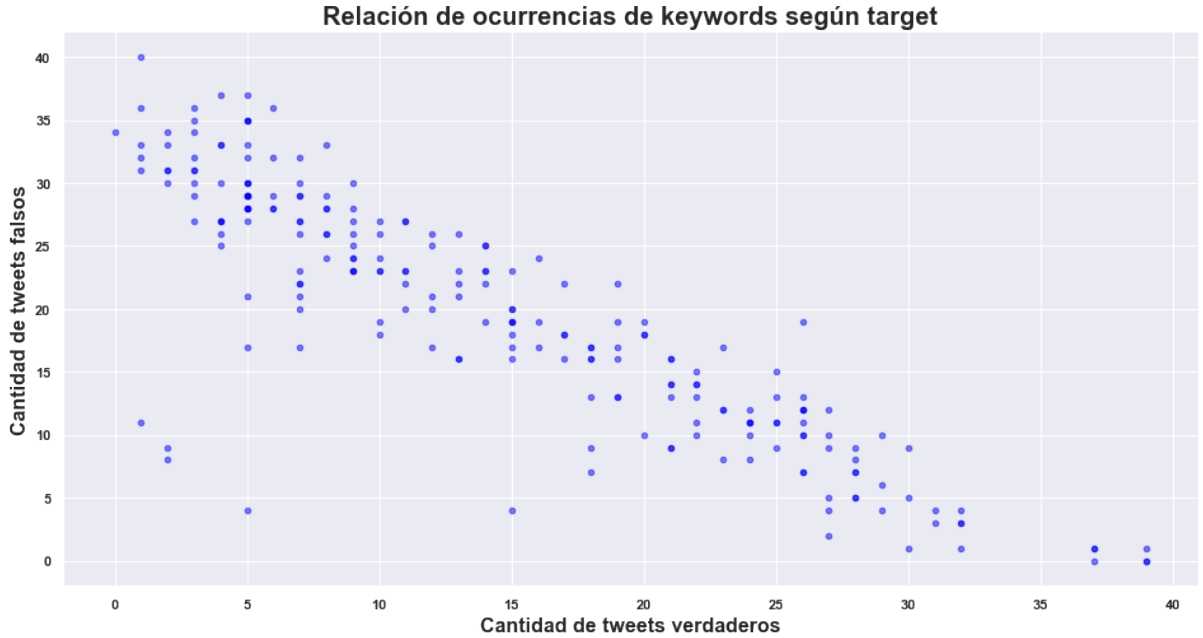


Figura 6: Relación de ocurrencias entre keywords con tweets verdaderos y falsos.

Pudimos observar que a medida que aumenta la cantidad de tweets verdaderos en una keyword, disminuye la cantidad de tweets falsos y viceversa.

Por otra parte, también pudimos corroborar la relación entre la *keyword* y el *text* del tweet, en donde obtuvimos exactamente 6700 registros con coincidencia de la keyword en el texto y el resto vimos que algunas keyword resultan ser una interpretación de lo que dice el tweet y otras no.

## 3.3. Location

De la misma forma que comenzamos con el campo keyword, empezamos el análisis viendo las ubicaciones no nulas que tienen mayor cantidad de tweets.

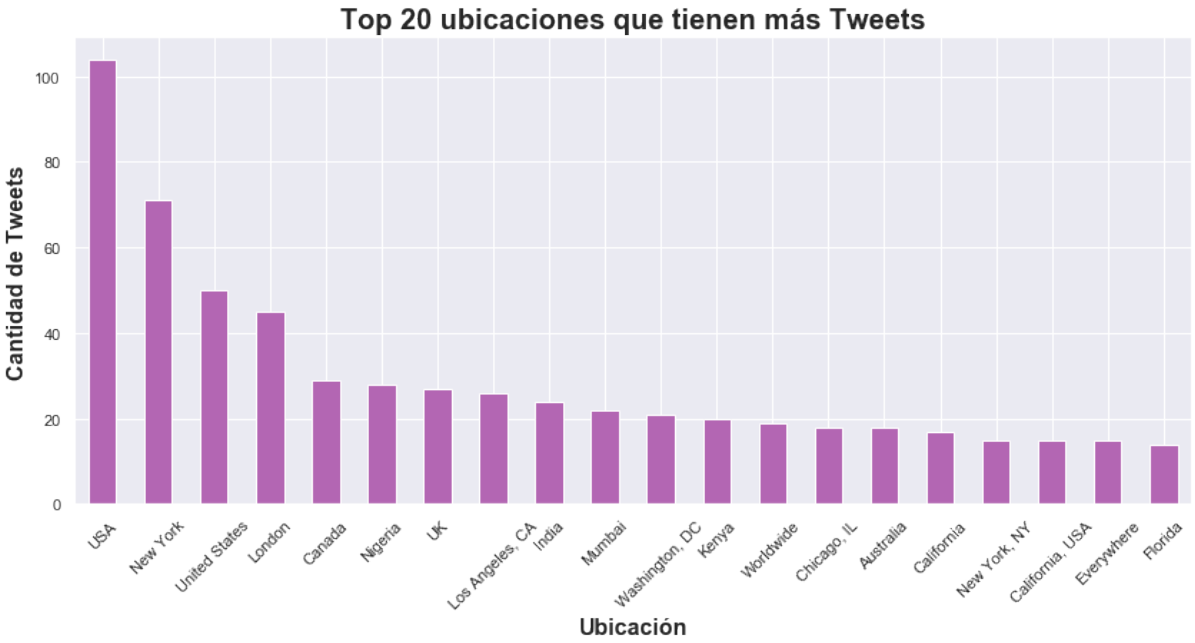


Figura 7: Top 20 ubicaciones que tienen más tweets.

De estas ubicaciones, hay 2818 que sólo aparecen una vez, representando el 84.32 % de las ubicaciones como se puede apreciar en la siguiente visualización.

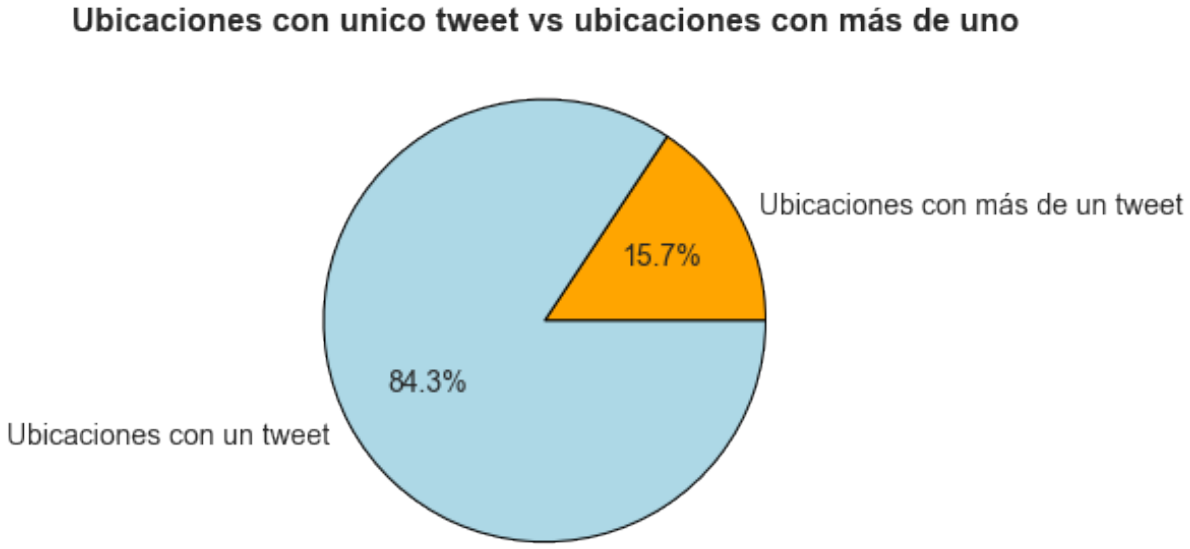


Figura 8: Porcentaje de ubicaciones con un sólo tweet vs ubicaciones con más de uno.

Estos tweets de ubicaciones que aparecen una única vez representan el 39.5 % de los tweets falsos del dataset.


Figura 9: Veracidad de las ubicaciones que sólo aparecen una vez.

Como se mencionó previamente, para estos análisis no se consideraron los tweets provenientes de una ubicación desconocida.

Al analizar considerando dichos tweets, pudimos ver que no se pueden descartar esos tweets ya que representan un volumen importante de los tweets.

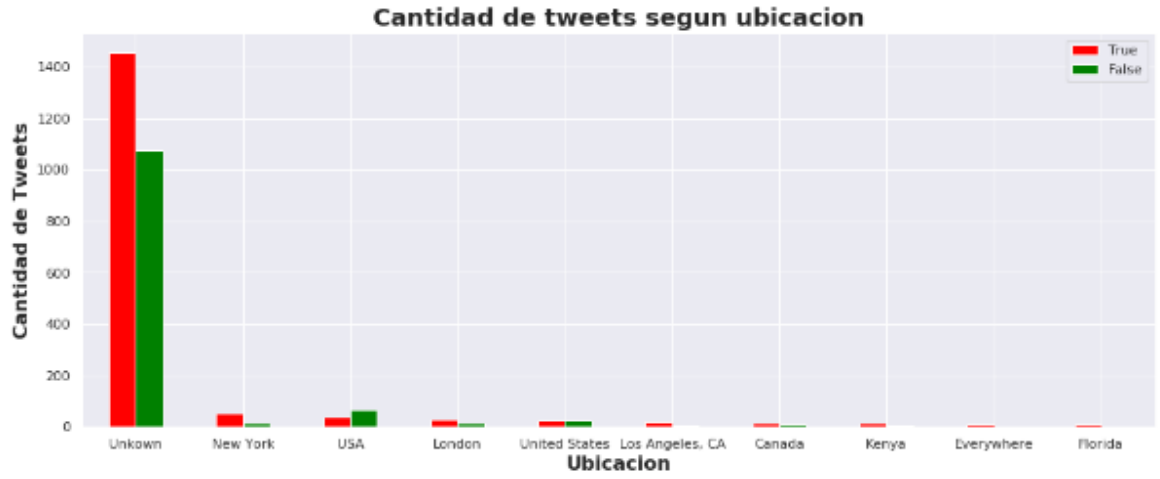


Figura 10: Cantidad de tweets por ubicación.

Observamos que hay 2533 tweets que tienen ubicación desconocida, de los cuales, el 57.56 % corresponde a tweets falsos.

## 3.4. Text

El mayor foco de nuestro análisis lo pusimos en el texto del tweet ya que es en el contenido del mismo en donde supusimos que ibamos a encontrar la mayor cantidad de indicadores sobre si es verdadero o no.

### 3.4.1. Longitud del tweet

Analizamos si había alguna relación entre la longitud del tweet y la veracidad del mismo. Comenzamos echando un vistazo a las estadísticas de la longitud para cada target a través de un boxplot.

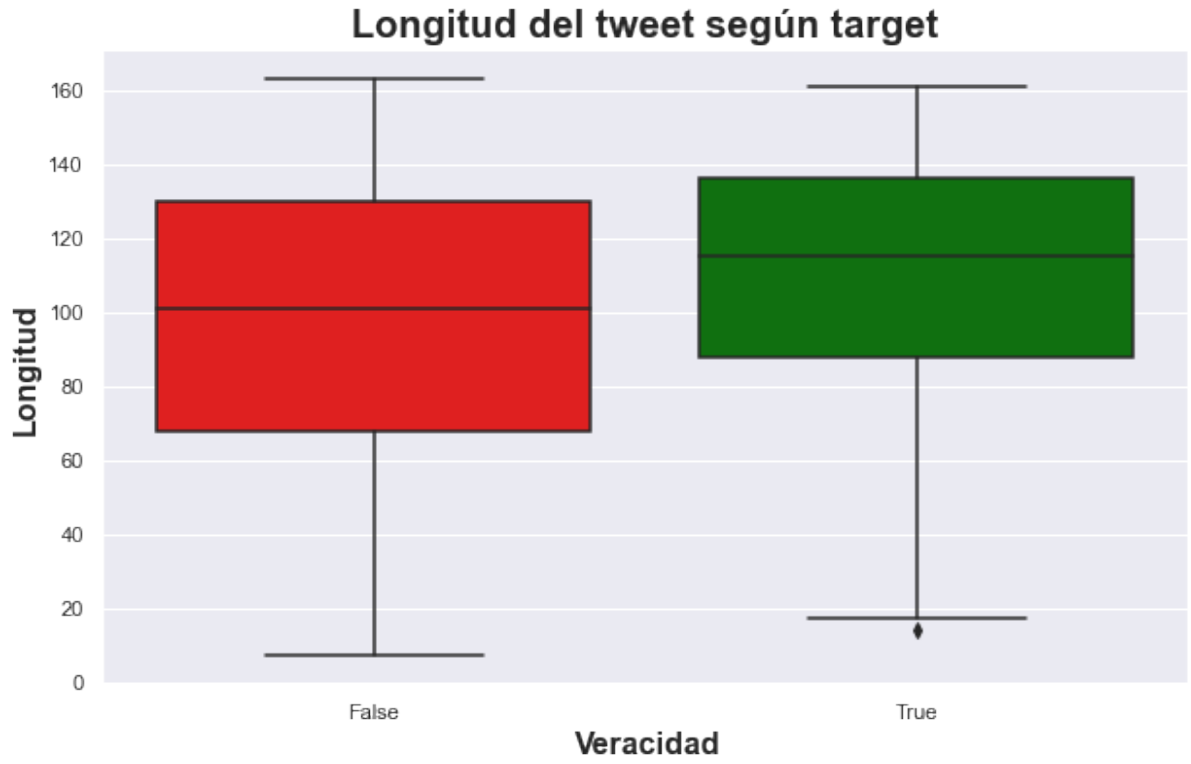


Figura 11: Estadísticas de la longitud del tweet según target.

Viendo el gráfico pudimos observar que tanto el mínimo como los cuartiles y la media de los tweets verdaderos son superiores a los de los falsos, por lo que en principio consideramos que entre mayor fuera la longitud de los tweets más probable sería que el tweet fuera verdadero.

Luego, realizamos una visualización de la distribución de los tweets según su target en función de la longitud de los mismo para evidenciar mejor esta suposición.

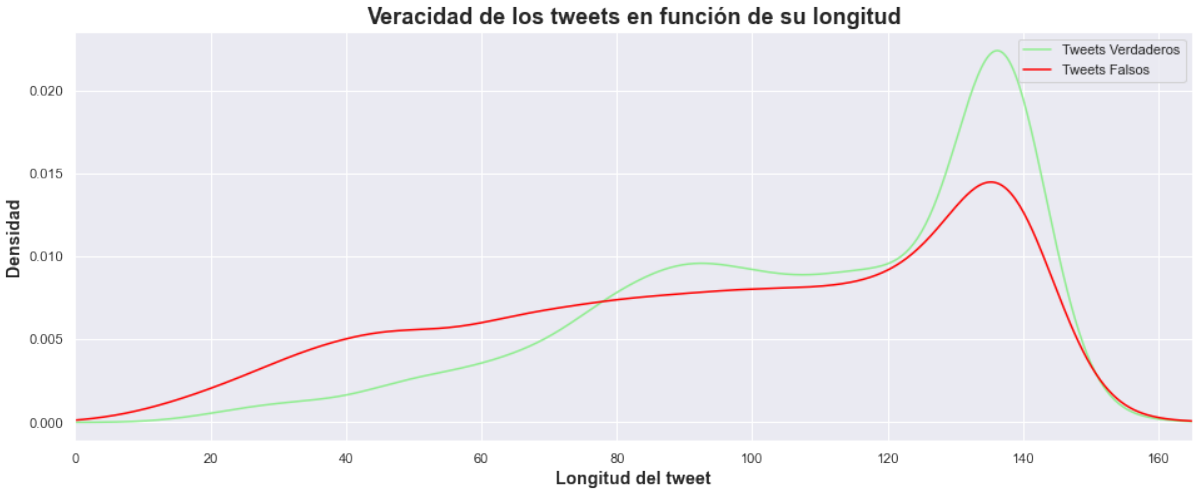


Figura 12: Densidad de los tweets en función de la longitud.

En la figura 12, vimos que los tweets que tienen una longitud menor a 80 caracteres son en su mayoría falsos, mientras que los que tienen una longitud superior tienden a ser verdaderos. También pudimos ver que la longitud de los tweets tiene un pico llegando a los 140 caracteres, lo cuál tiene sentido considerando las limitaciones que la plataforma solía tener respecto a la longitud de los mismos.

Por último, decidimos analizar más en detalle que va sucediendo a medida que la longitud crece; nos preguntamos que sucedía con el target, era "más real.º no el tweet? Para esto, creamos categorías de

longitud en base a los cuartiles de la longitud de los tweets.

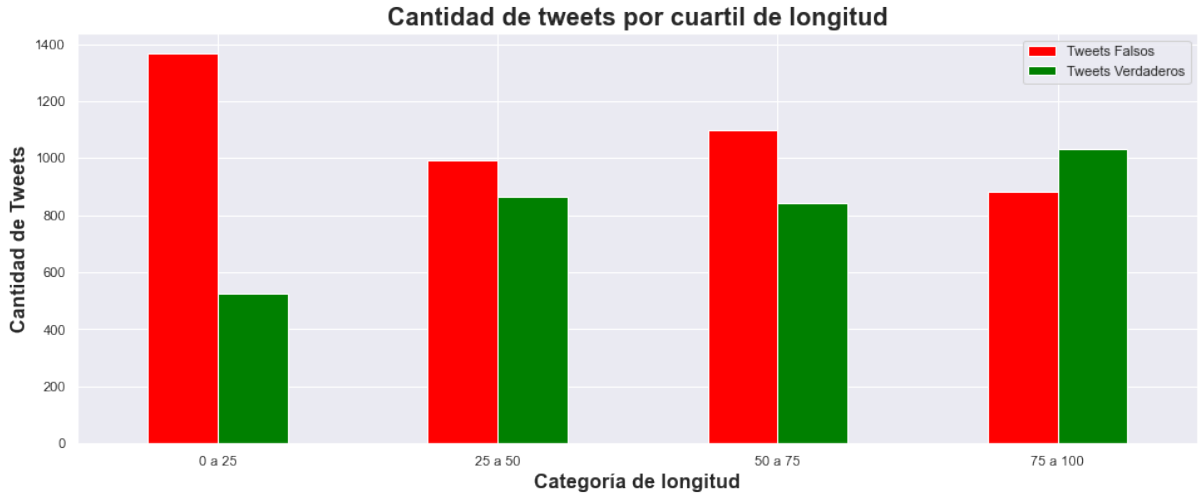


Figura 13: Cantidad de tweets por cuartil de longitud.

En la figura 13, nuevamente pudimos notar que a una longitud menor es más probable que sea un tweet falso y que cuanto más largo es el tweet mayor su probabilidad de ser real. Pero este comportamiento no se mantuvo para la tercer categoría de tweets (tweets con longitud entre 107 y 134 caracteres); en donde hubo un aumento en la cantidad de falsos y disminución en la cantidad de verdaderos, contrario al crecimiento/decrecimiento que venía ocurriendo en el resto de las categorías.

### 3.4.2. Cantidad de palabras del tweet

Otro análisis realizado en relación a la composición del texto, fue ver si había alguna relación entre la cantidad de palabras del tweet y el target.

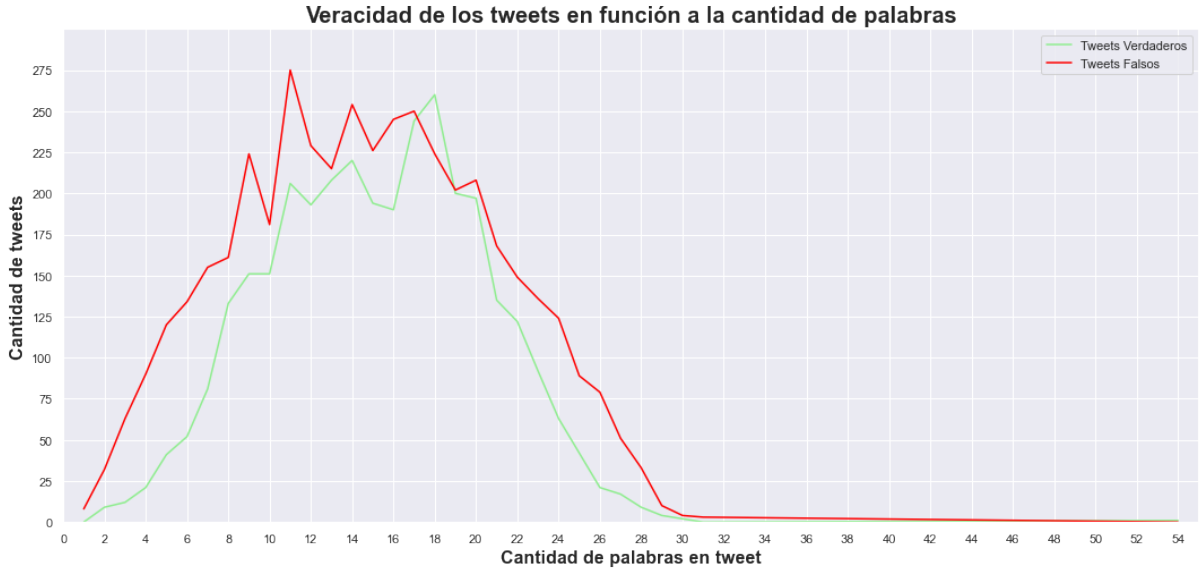


Figura 14: Tweets por cantidad de palabras.

En la visualización observamos que, salvo en contados casos (cuando el tweet tiene alrededor de 18 palabras), hay mayor cantidad de tweets falsos que verdaderos independientemente de la cantidad de palabras. Por lo tanto, no pudimos identificar que haya alguna relación entre la cantidad de palabras del tweet y su target como sí vimos en el caso de la longitud del mismo.

### 3.4.3. Tweets con menciones

En los tweets pueden realizarse menciones a otros usuarios, decidimos analizar si esto era un factor determinante en la veracidad de los mismos. En poco más de un cuarto de los registros con los que trabajamos se realizaban menciones; mediante un Pie Chart visualizamos como se distribuía el target.

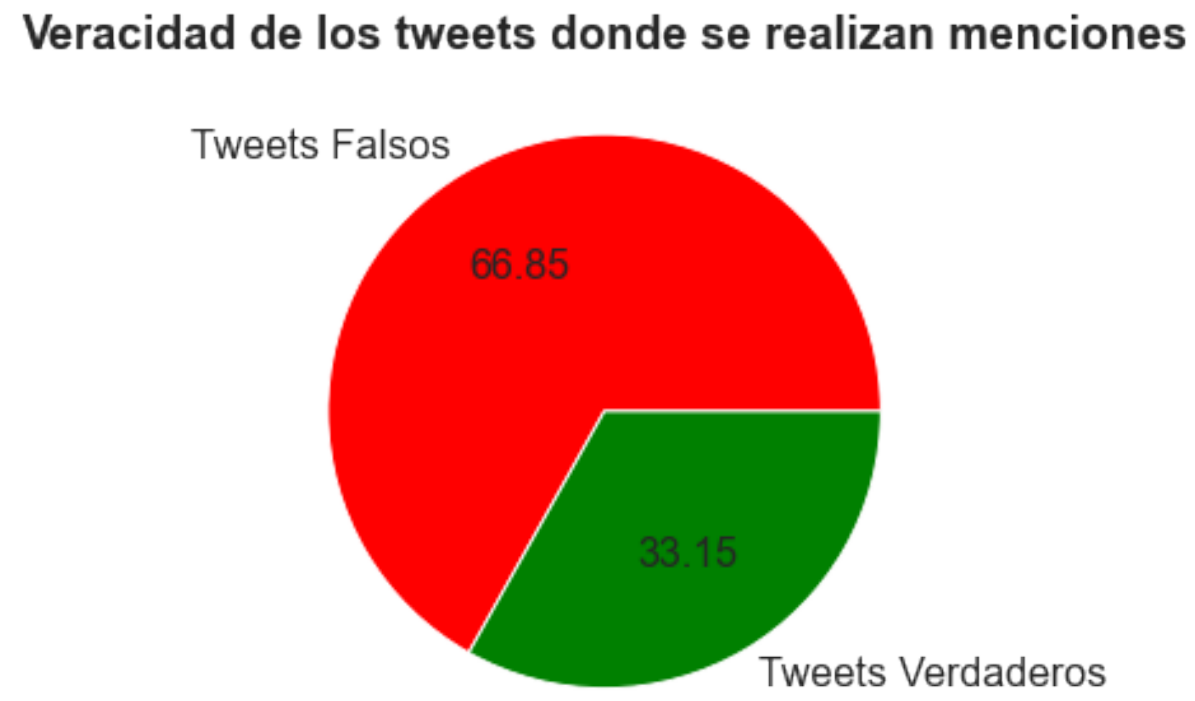


Figura 15: Distribución del target de los tweets con menciones.

En la figura 15, vemos que aproximadamente 2/3 partes de los registros con menciones resultaban ser falsos. En principio, esperabamos que fuera mayor la cantidad de tweets falsos debido a que sean tweets entre usuarios particulares y no noticias de desastres reales; pero la cantidad alta de tweets verdaderos se puede deber a que en los tweets sobre desastres se realizan menciones a cadenas de noticias o a otras cuentas relacionadas al desastre en cuestión.

### 3.4.4. Tweets "expresivos"

Consideramos como "tweet expresivo" a los tweets en donde se utiliza múltiples signos de exclamación o interrogación consecutivos. Debido a que suele ser una forma más "informal" de expresarse, supusimos que podía ser un indicador para tweets sobre desastres falsos. Cabe destacar que tan solo 641 registros (menos del 10 % de los tweets) se trataban de "tweets expresivos".


Figura 16: Distribución del target de los tweets expresivos.

En la visualización, pudimos observar que efectivamente la gran mayoría de los tweets analizados resultaban ser falsos. Esto puede deberse a que los tweets en donde se utilizan signos de puntuación de manera exagerada, en general para resaltar una emoción del usuario, no suelen ser sobre noticias reales.

### 3.4.5. Palabras destacadas del texto

En esta parte del informe vamos analizar cuales son las palabras más destacadas del texto del tweet y ver si estas presentan alguna relación contra los hashtags que posea el tweet o keywords del mismo.

Antes es necesario comentar un pre-procesamiento que se le efectuó al texto de cada tweet. Hicimos uso de la librería "NLTK" la cual fue necesaria para encontrar lo que se llaman *stopwords*, las cuales vendrían a ser artículos, preposiciones, etc. En el pre-rpocesamiento primero lo que se hizo fue remover los links por espacios en blanco, remover comillas simples y paginas web que no comiencen con http o https. Con esto nos quedan palabras que vamos a procesar con la librería de python "TextBlob" la cual nos genera una lista de las palabras de común uso, eliminando así caracteres no deseados como corchetes o símbolos que no nos vendrían a interesar. Tambien se filtraron algunas palabras que se fueron viendo en el analisis de procesamiento del texto y que se interpretaron como posible ruido y se optaron por descartar. Por último la idea es realizar una normalización de cada palabra usando "WordNetLemmatizer" de la librería "NLTK", que con la cual vamos a "modificar el tiempo" en el que esta escrita una palabra por ejemplo "played" pasaría a ser "play" esto es para poder tener un mejor análisis y poder generalizar estas palabras. Ahora si, con el pre-procesamiento que se le efectuó al texto de cada tweet obtuvimos un grupo de palabras normalizadas para cada tweet y nos interesa analizar que sucede con estas palabras como cuales son las más frecuentes.

Veamos que obtuvimos como resultado de esto:


Figura 17: Wordcloud - Palabras destacadas.

La idea del "wordcloud" es poder a simple vista identificar que palabras aparecen con mayor frecuencia, en nuestro caso luego que el texto de tweet fue procesado. Esto a su vez nos interesa para poder cruzar informacion con los mismos textos limpios por cada tweet o hashtags que contengan o keywords, para poder ver si existe o no alguna relacion y que informacion podemos obtener de esto.

En un principio vamos a ver estas palabras más destacadas(un top15) para ver cuales estan contenidas en el texto del tweet(pre-procesado) ya que no todos los tweets puede contenerlas y queremos analizar las que las contienen que sucede con la veracidad del tweet.

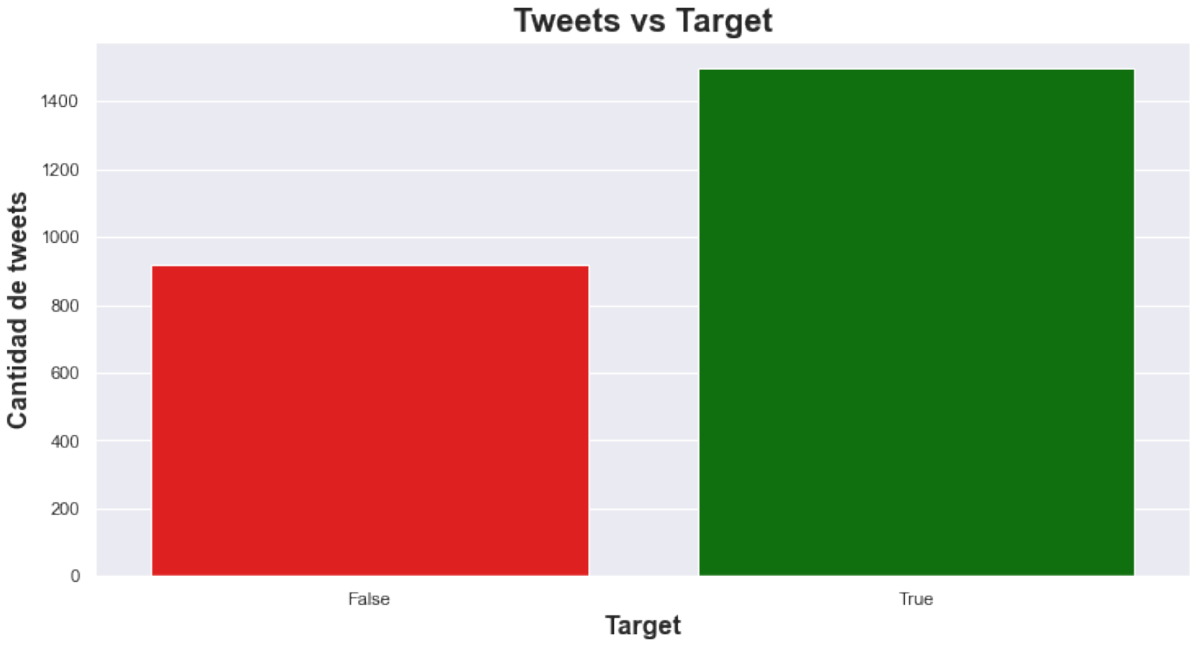


Figura 18: Top15 palabras importantes - tweets.

Si bien se puede observar que se "acota" el set de datos, vemos que hay mayor probabilidad a ser un tweet verdadero que uno falso con estas palabras obtenidas. Veamos que sucede al cruzar estas mismas contra las keywords de cada tweet.

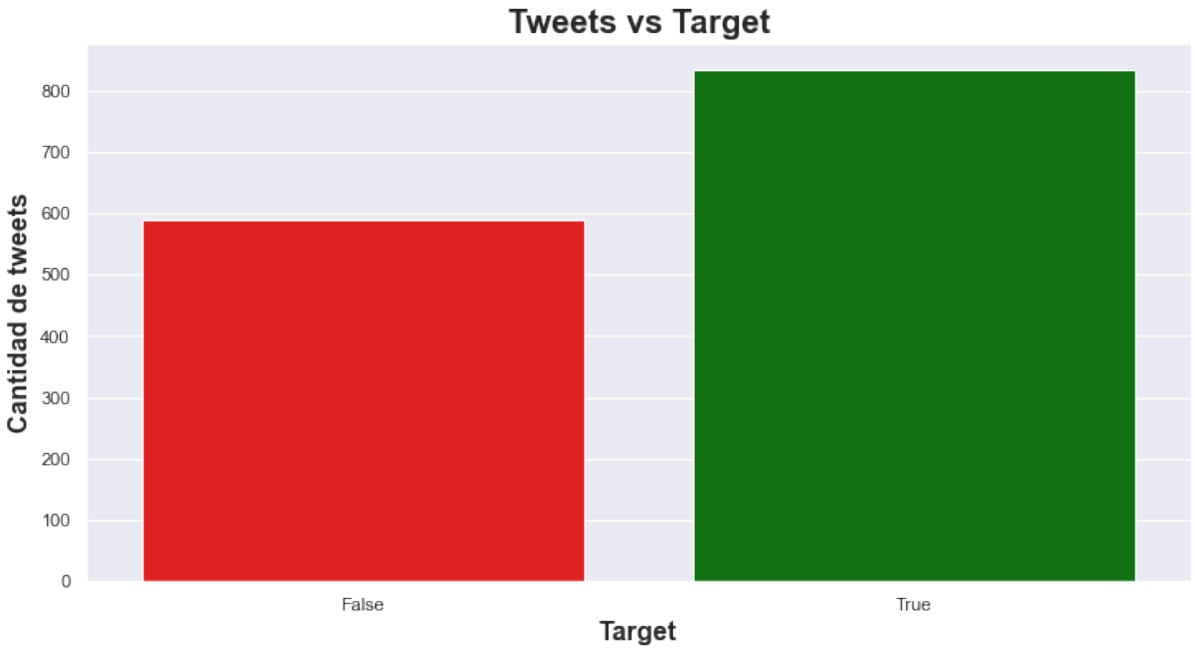


Figura 19: Palabras destacadas en keyword

El subset es aun más chico pero vemos aun una leve tendencia de tweets verdaderos similar a la que se observa en la figura 18. Algo para mencionar es que se relacionó lo obtenido del wordcloud vs los hashtags encontrados en el texto y no se pude concluir algo, debido a la gran "pérdida" de información en el subset.

Por último lo que quisimos era ver que sucedía al utilizar los datos de la figura 18 y de la figura 19 y lo que obtuvimos fue muy similar a la figura 19.

Como parte de la conclusión vemos que ciertas palabras destacadas de forma general para todos los tweets, con el procesamiento que explicamos anteriormente, podemos ver que los tweets que tienen una coincidencia con las más destacadas tienen mayor probabilidad de ser tweet verdaderos como se muestra en la figura 18. El resultado obtenido cuando buscamos coincidencias con los hashtags perdimos muchos registros del set de datos por lo que no tiene validez asegurar algo pero si vimos que para este subset la tendencia fue tweets verdaderos.

### 3.4.6. Tweets con links

Otro análisis realizado fue determinar si el hecho de que un tweet tuviera un link afectaba en su target. Notamos que el 52.2 % de los tweets tenían links en su texto.


Figura 20: Distribución del target de los tweets con links en su texto.

Analizando el gráfico, en principio, no parecía que el hecho de que un tweet tenga un link o no sea determinante para la veracidad del mismo.

Decidimos analizar más profundamente los links que aparecen en los tweets para poder determinar si hay links que se repitan en muchos casos y que nos sirvan en la predicción del target de otros tweets en donde aparezcan estos links. En base a esto pudimos ver que hay 72 links que aparecen en más de un tweet, el más recurrente con 9 tweets; por lo que tampoco parecía que hubiera algún caso de un "link que se repita en gran cantidad de tweets falsos o verdaderos"que nos sirviera.

Finalmente, decidimos analizar si los links que se repetían mantenían siempre el mismo valor de verdad o habían tanto tweets verdaderos como falsos que usaran el mismo link.

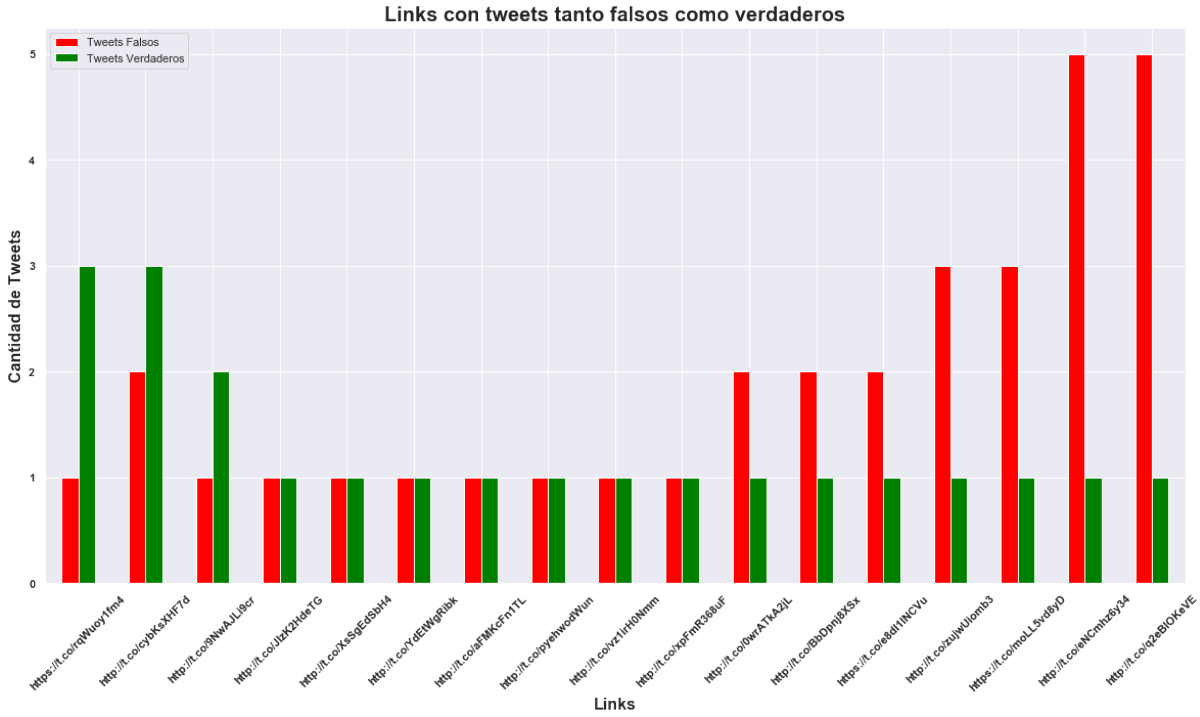


Figura 21: Cantidad de tweets para los links que aparecen tanto en tweets verdaderos como falsos.

En la figura 21, vimos que efectivamente hay 17 links que aparecen en tweets con ambos valores de verdad y que siempre son 1 o 2 tweets en los que varía el target.

Luego de analizar el texto de los tweets en donde aparecían estos links, vimos que en la mayoría de los casos el tweet es casi idéntico más allá de que cambie el target. Por lo tanto, supusimos que el target de estos registros en donde varía la veracidad puede deberse a un error a la hora de clasificar estos tweets. Vamos a tener en cuenta estos tweets para el Trabajo Práctico 2 ya que podrían estar generando ruido en el set de datos.

# 4. Conclusión

Finalizando el análisis, en el cual realizamos algunas observaciones, vamos a mencionar las que consideramos más importantes.

En el análisis de las keywords, por un lado pudimos ver que hay keywords con las que es fácil decir si es el tweet verdadero o falso, mientras que en otras esto resulta difícil predecirlo únicamente con la keyword. A su vez se observa que para las keyowrd *"derailment, wreckage debris"* no tienen ninguna noticia falsa y que para *"outbreak,typhoom y oilspill"* solo poseen una. Podemos destacar algunas keywords que tienen más noticias falsas como *aftershock, body bag, ruin, blazing y electrocute*. Para el caso de *bombed,explosion, storm, derail y rescued( entre otras)*, será necesario hacer un análisis más profundo,con el texto del tweet, su longitud, su contenido.

Otra observación importante que se puede ver en las figuras 11 y 12 es que si la longitud del tweet es menor a 80 caracteres hay mayores probabilidades de que sea falso, ya que los verdaderos suelen ocupar los 140 caracteres que la plataforma permite escribir. Esto que mencionamos también es posible visualizarlo en la figura 13 en donde vemos claramente que a una longitud chica, la cantidad de tweets falsos es mayor a la cantidad de tweets que resultan ser verdaderos. Pero no es algo muy definitivo, ya que como se observó, los tweets falsos suelen usar caracteres expresivos (cosa que no suele verse en tweets de canales de noticias por ejemplo) o hacer menciones.

En la sección de las palabras destacadas vimos lo importante y la ayuda que brinda visualmente el *wordcloud* al momento de ver las palabras más destacadas de todos los tweets fácilmente. Esto nos permitió obtener un listado con el cual buscamos coincidencias, por ejemplo en el mismo texto de cada tweet, y pudimos concluir que se obtiene una mayor probabilidad de validez del tweet cuando alguna de las palabras destacadas obtenidas se encuentra como parte del texto (procesado). En este procesamiento se filtraron algunas links o paginas web, por ejemplo del texto ya que no hacen más que introducir ruido

a la hora de analizar el texto, al igual que otros casos encontrados. También ignoramos las "palabras comunes" con el uso de la librería NLTK y algunas que fuimos viendo durante el análisis para llegar a un texto más "limpio" con palabras de mayor importancia. Cuando analizamos estas mismas, no contra el texto del tweet sino con las keywords, obtuvimos un comportamiento muy similar pero "perdimos" un poco más de información pudiendo así perder un poco de validez para generalizar esto en todo el set. Por otro lado al momento de observar los hashtags , vemos también el mismo comportamiento pero ahora "perdimos" casi todos los datos dejando no valida cualquier conclusión que podamos obtener de ese caso. Resumiendo lo más importante a destacar es que podemos ver que los tweets que tienen una coincidencia con las más destacadas, tienen mayor probabilidad de ser tweet verdaderos como se muestra en la figura 18

A su vez, vemos que la mayoría de los datos que se tienen, muchos no registran ubicación y los que si, obtuvimos que un 83.3 % de las ubicaciones poseen un único tweet y el 15.7 % ubicaciones poseen más de uno. Esto tiene sentido ya que se mostró que las ubicaciones tienen datos no coherentes y debido a esto no son únicas las ubicaciones de los tweets. Como se puede ver en la figura 10 se poseen muchos registros(2533) que no poseen ubicación y el 57.5 % de estos corresponden a tweets falsos.