



UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE INGENIERÍA

**75.06 Organización de Datos**

**Trabajo Práctico 1**

**Primer Cuatrimestre de 2020**

Grupo 27

Joaquin Lopez Saubidet	99252
Santiago Tadini	104439
Tomas Sabao	99437
Zugna, Federico	95758

**Link de GitHub:** [https://github.com/tomasSabao/Organizacion de datos tp1](https://github.com/tomasSabao/Organizacion_de_datos_tp1)

## Índice

Índice .....	2
1. Introducción .....	3
2. Informe.....	4
3. Conclusión Final .....	20

## 1. Introduccion

Este informe se encarga de analizar los datos sobre un conjunto de los tweets del set de datos de la competencia: <https://www.kaggle.com/c/nlp-getting-started>.

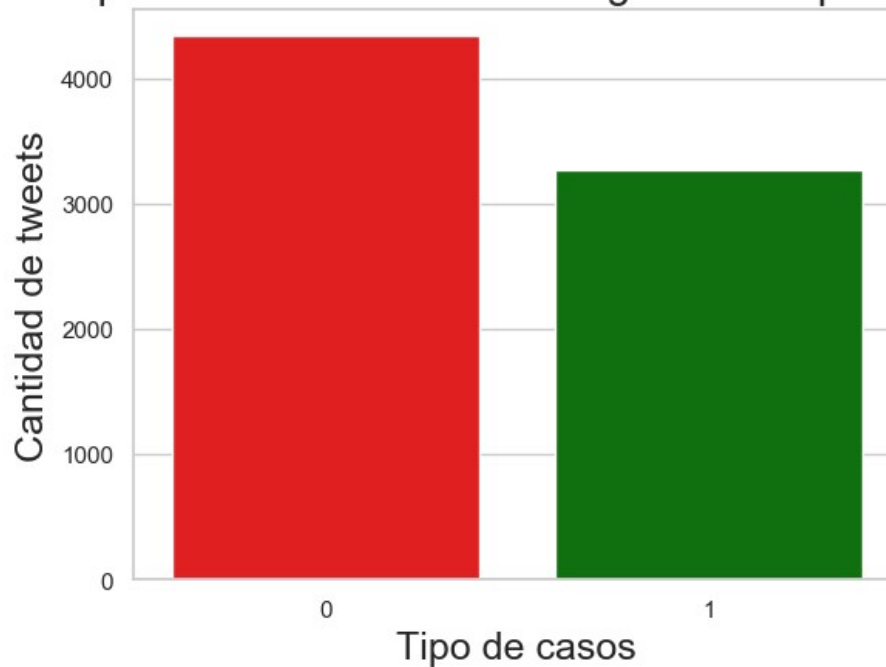
- `id` - identificador unico para cada tweet
- `text` - el texto del tweet
- `location` - ubicación desde donde fue enviado (podría no estar)
- `keyword` - un keyword para el tweet (podría faltar)
- `target` - en train.csv, indica si se trata de un desastre real (1) o no (0)

El objetivo del primer TP es realizar un análisis exploratorio del set de datos. Queremos ver qué cosas podemos descubrir sobre los datos que puedan resultar interesantes. Estas cosas pueden estar relacionadas al objetivo del TP2 (predecir si un cierto tweet es real o no) o no, ambas son de interés.

## 2. Informe

Dentro del archivo podemos ver como cada uno de los tweets presentes tienen un campo target que nos indica si es un caso real o no lo es. De todo el set de datos, comparamos la cantidad de tweets que hay para cada caso.

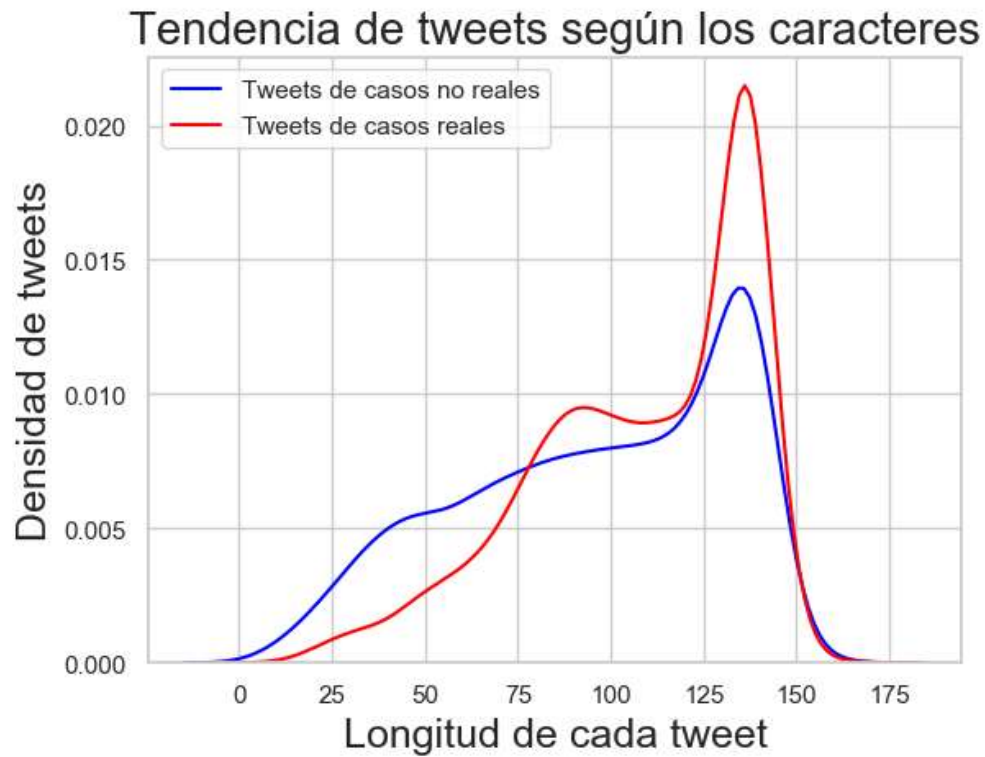
Comparación total de tweets según cada tipo de caso



Se puede observar que hay una mayor cantidad de tweets acerca de casos no reales. En un 57% del total de los tweets son de casos no reales, y por consecuencia, un 43% son casos reales.

Se sabe que la cantidad máxima de caracteres por cada tweet es de 280 caracteres. El promedio de longitud de todos los tweets es de 101 caracteres aproximadamente.

Se puede ver que los tweets que tienen entre aproximadamente 80 caracteres y 150 son en los que predominan los casos reales, en cambio cuando es menor a 75 caracteres y mayor a 150 caracteres la mayoría de tweets son de casos no reales



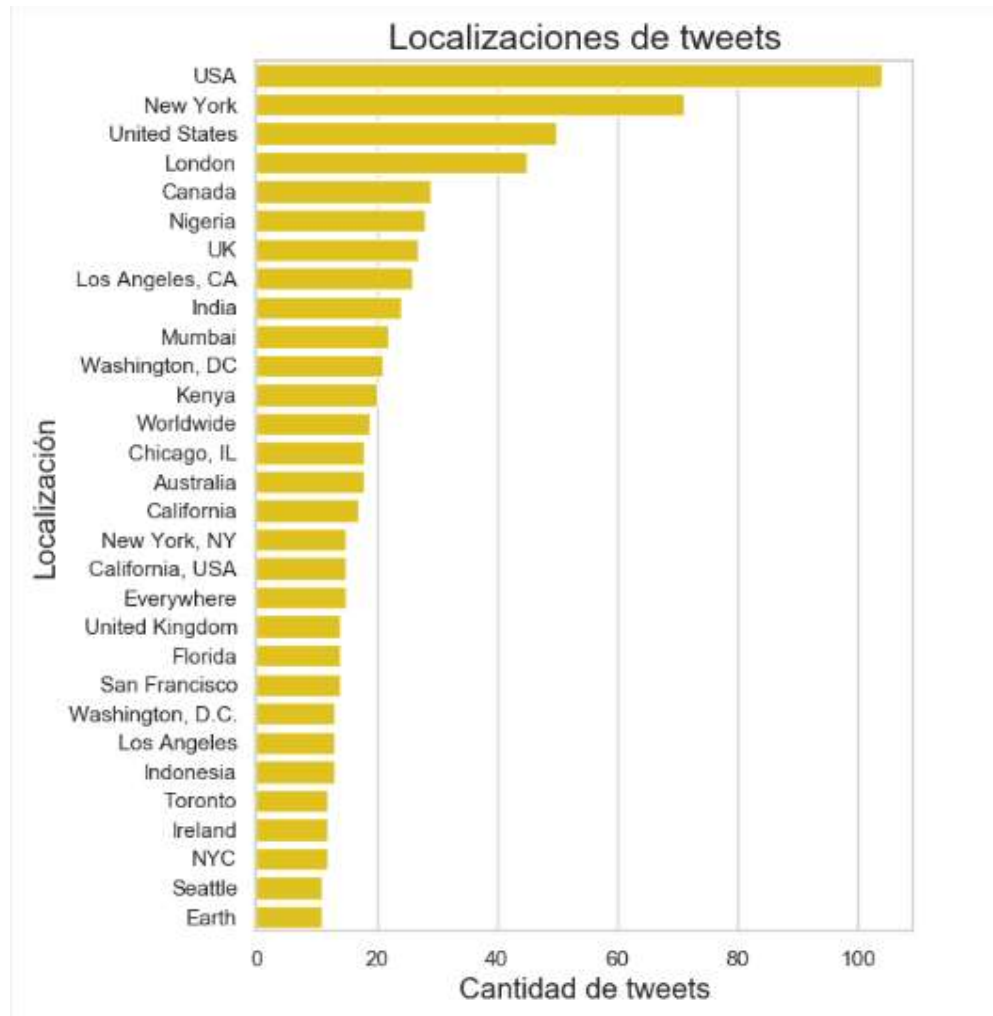
Pues tanto los tweets con menor (7 caract.) longitud como el mayor (163 caract.) se corresponden a tweets de casos no reales.

También se puede ver la cantidad de palabras por tweets, predominan los tweets que están entre 10 y 20 palabras.

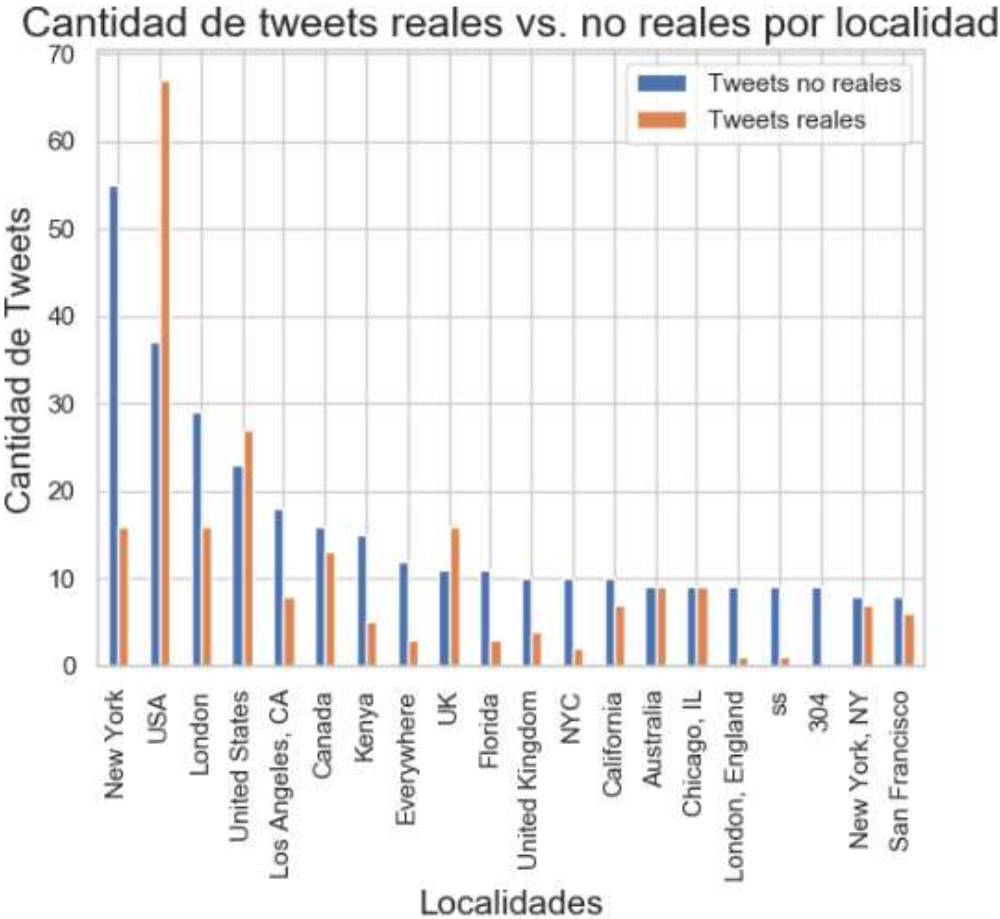
Cantidad de palabras por tweets



Del archivo de tweets tenemos la localización de algunos de estos, por lo que nos fijamos de donde vienen cada uno de estos y notamos que la mayoría fueron realizados en Estados Unidos, pues los tres primeros valores son Usa, United States y New York. También se observa que de los primeros 30 lugares de donde hay mayor cantidad de tweets realizados, el 50% son ciudades de EEUU.



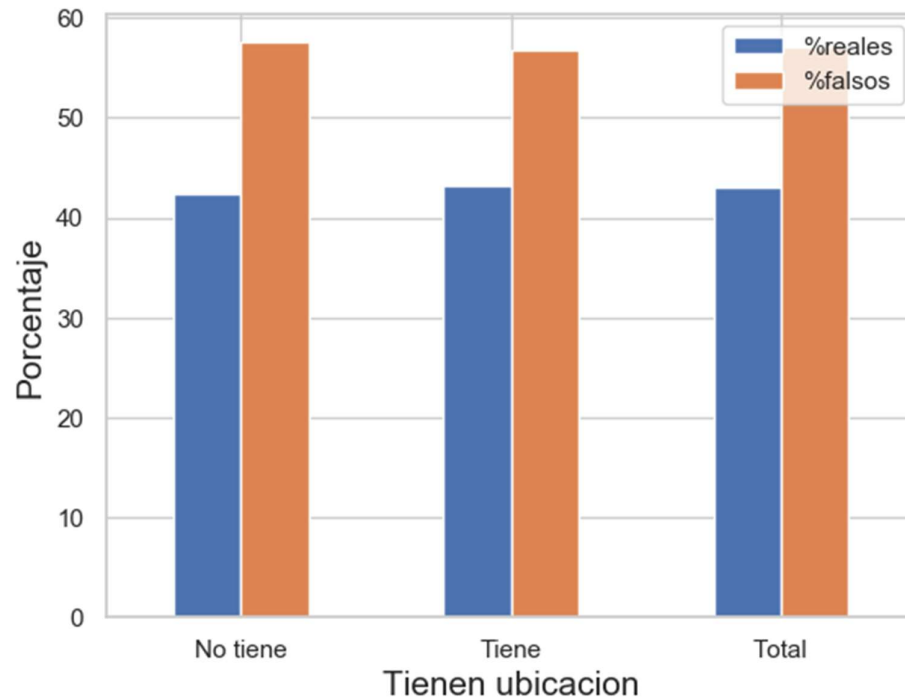
Pero a la vez, separando los casos reales de los no reales, solo en algunos países se logra tener una mayor cantidad de tweets acerca de casos reales.





Haciendo un análisis de la veracidad de los tweets en relación a si tienen ubicación se observa el siguiente gráfico.

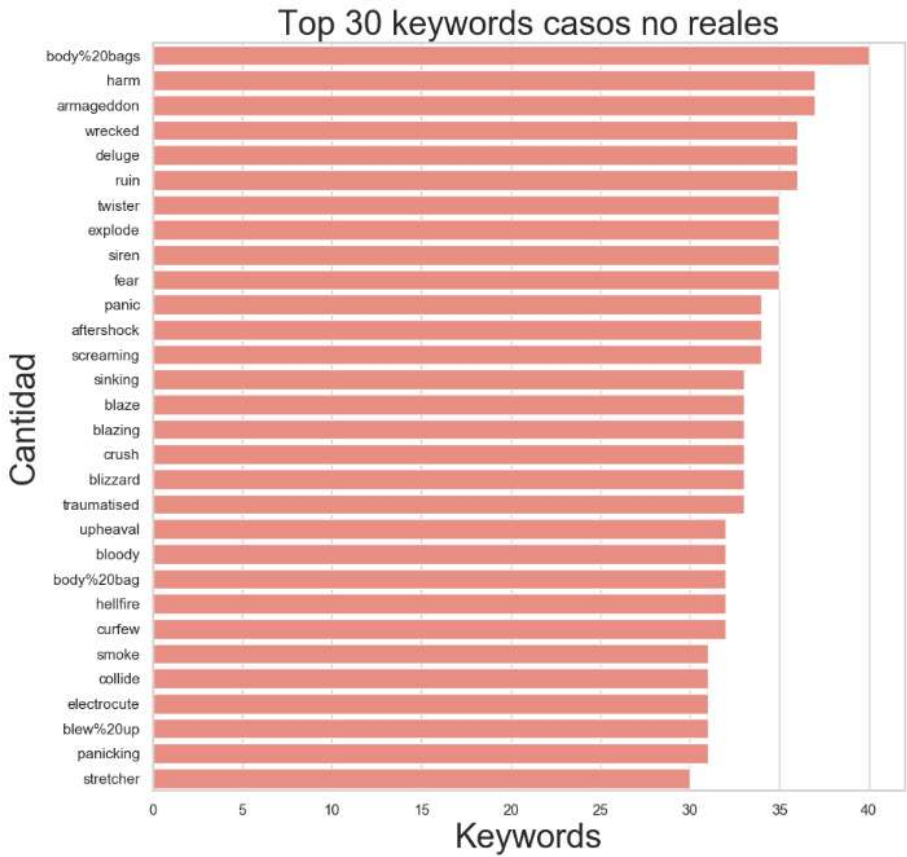
### Porcentaje de veracidad en relacion a si tienen ubicacion

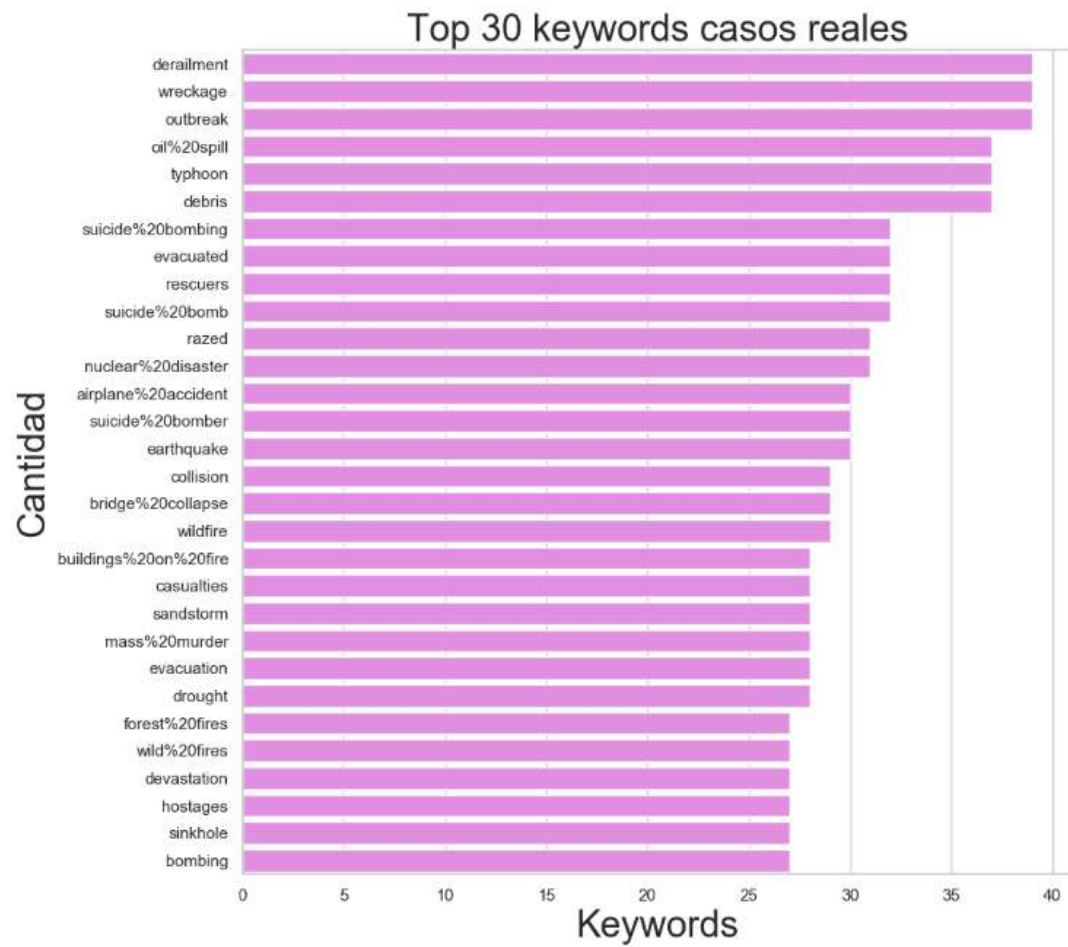


Se puede ver que no hay relación entre si tienen ubicación y su veracidad. El porcentaje de tweets reales se mantiene casi constante en ambos casos.

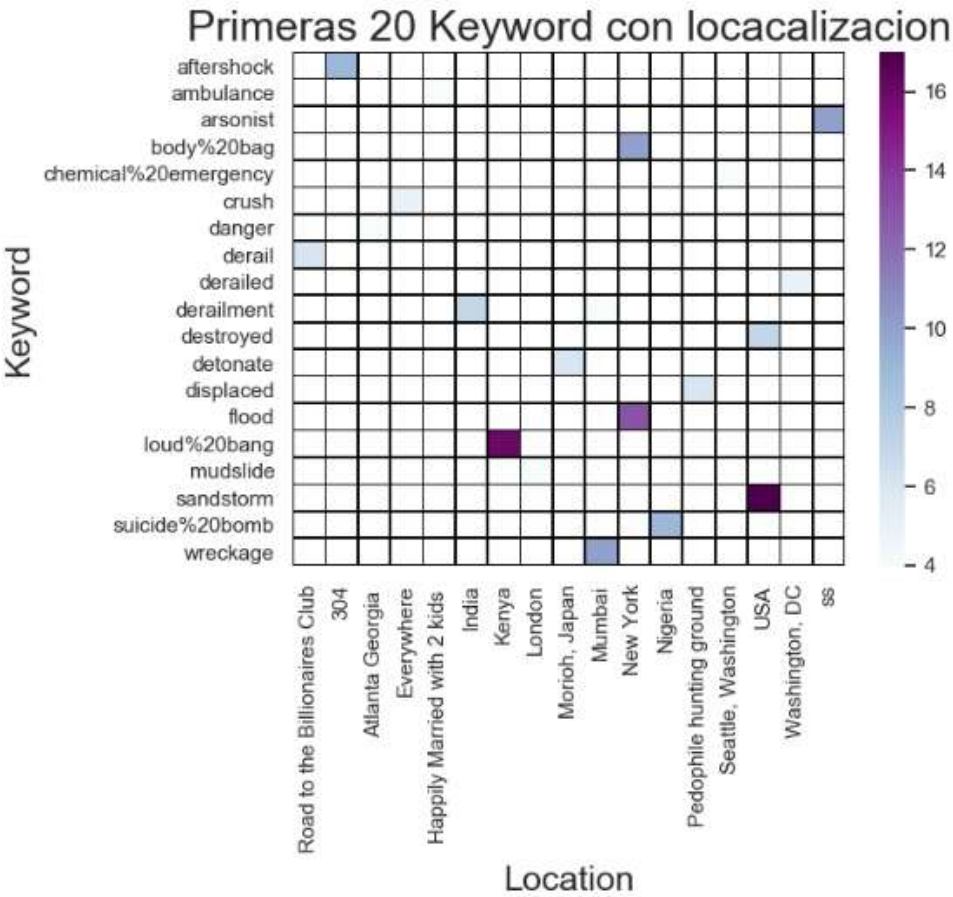
Separando las keywords de los casos reales y no reales, ninguna keyword está dentro del top 30 en ambos casos, siempre pertenece a algún grupo más fuerte que al otro.

También se puede notar que siempre las primeras 6 keywords de cada grupo están por encima de las 35 repeticiones:



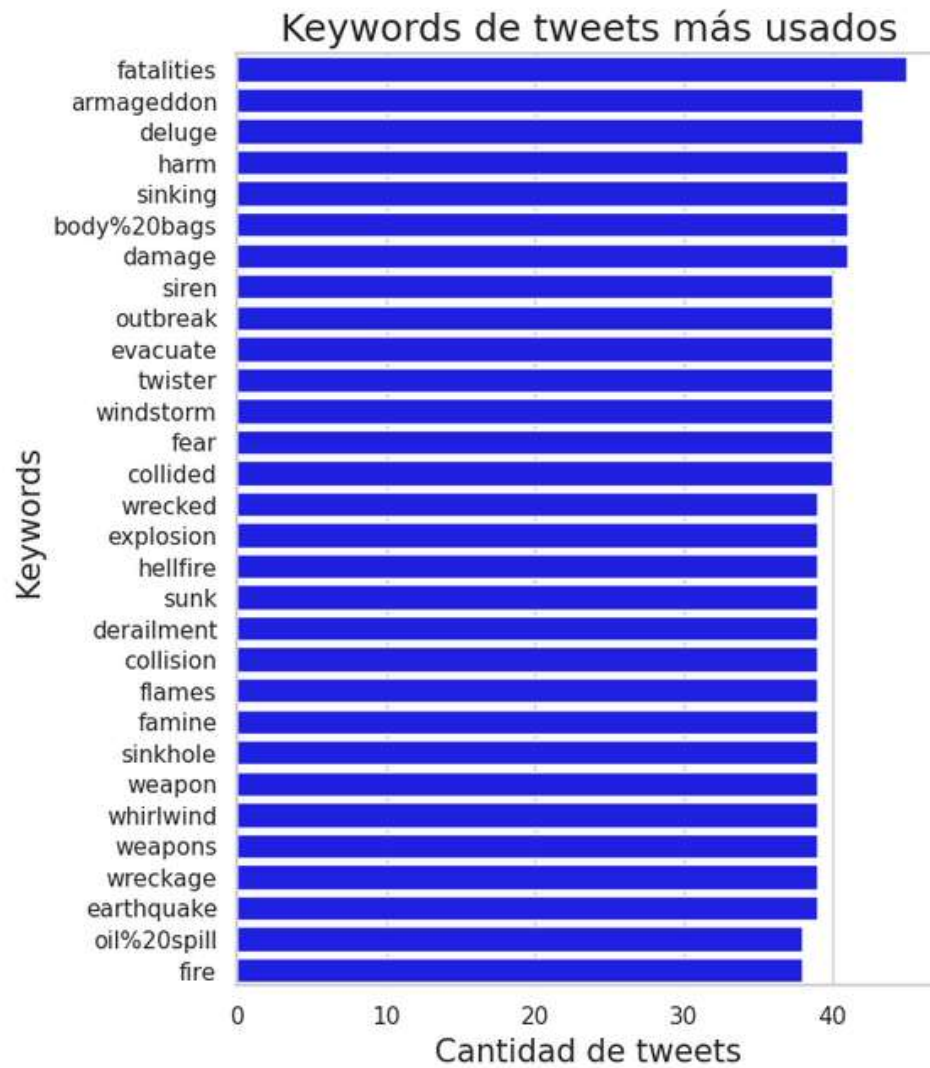


Para las keyword en cada localidad, solo en la localización USA hay dos palabras con dentro del top 20 de keywords. También se puede ver que hay localizaciones que no son ciudades o países, y tienen un numero bajo de repeticiones de la misma keyword, por lo que se puede pensar que es la misma persona o gente cercana a esta.

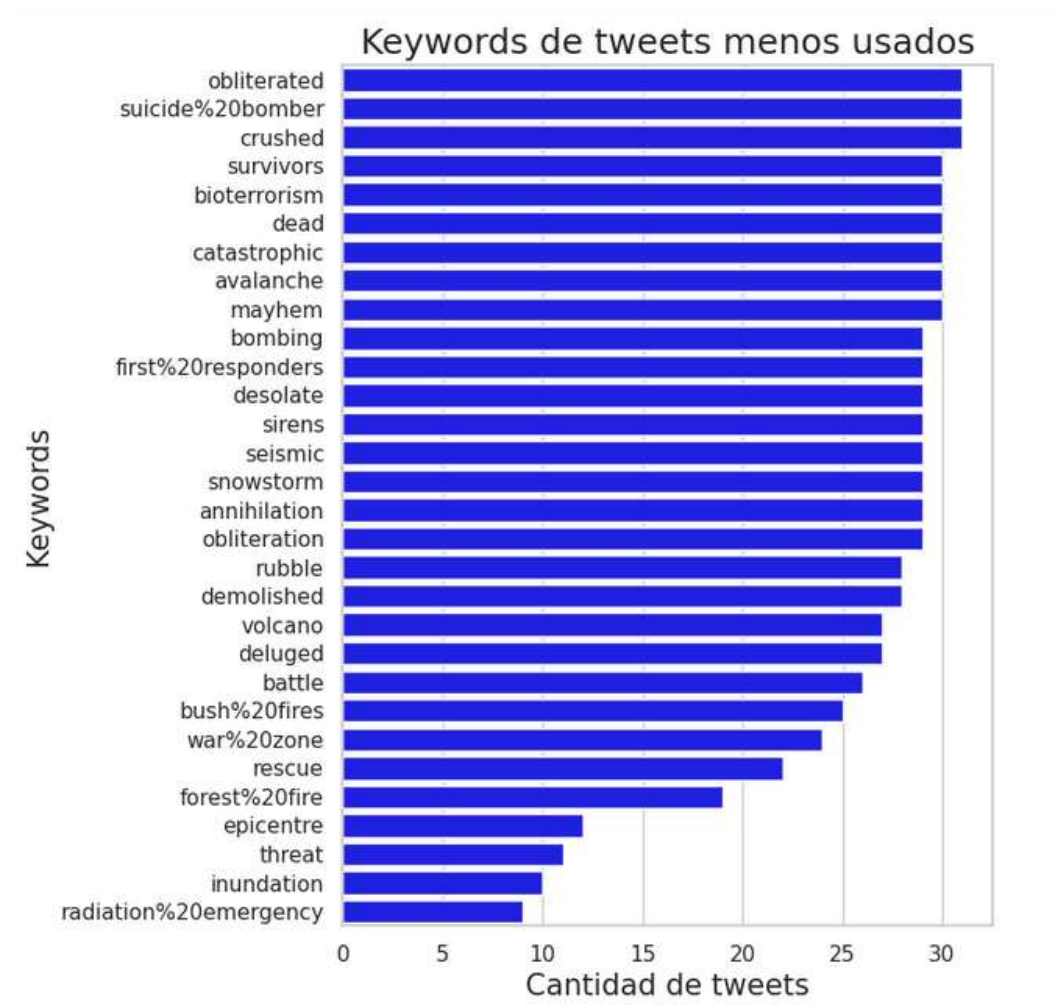


El set de datos analizado posee un campo llamado keyword. Analizamos este campo:

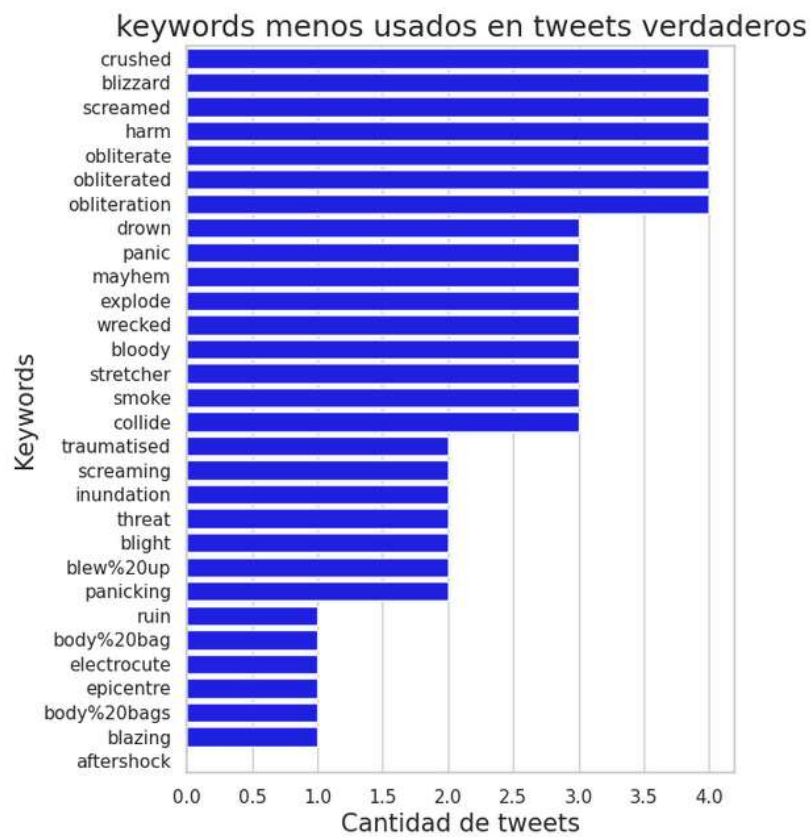
Hay 221 keywords únicos en nuestro set de dato. Nos interesa saber la distribución de los mismos, para ello contamos la cantidad de apariciones de cada uno. Mostramos los 30 keywords mas y menos usados.



La distribución de los keywords es bastante uniforme entre aquellos que más apariciones tuvieron, como puede verse, rondando las 40 apariciones.



Anteriormente habíamos calculado cuáles eran los 30 keywords que más aparecían para casos reales. Vamos a analizar cuáles son los 30 keywords que menos aparecen en casos reales:



Se quiere ver si hay una correlación entre las palabras usadas en un tweet y si este es real o no. Para esto se determino un coeficiente de veracidad establecido como :

$$\text{coeficientede veracidad} = \frac{\text{aparicionesentweetsverdaderos}}{\text{aparicionestotalesentweets}}$$

Dado que no todos los usuarios de twitter escriben una misma palabra de la misma manera, se pueden encontrar duplicados de una misma palabra (ejemplo: tweet y Tweet ), tomando esto en cuenta, se pasa cada palabra a minúscula y se procede a calcular su cantidad de apariciones de acuerdo al target del tweet. Una vez realizado esto, nos quedan 27983 palabras únicas, cada una de las cuales posee un coeficiente asociado.

Se tiene que tomar en cuenta la posibilidad de que haya palabras cuya cantidad de apariciones sea muy baja, lo que puede llevar a valores de coeficientes muy elevados (casos en los que una palabra aparece una única vez en un tweet verdadero sería un buen ejemplo). Es por ello que se realiza un estudio de las apariciones totales de cada una de estas palabras únicas. El

resultado obtenido es el siguiente:

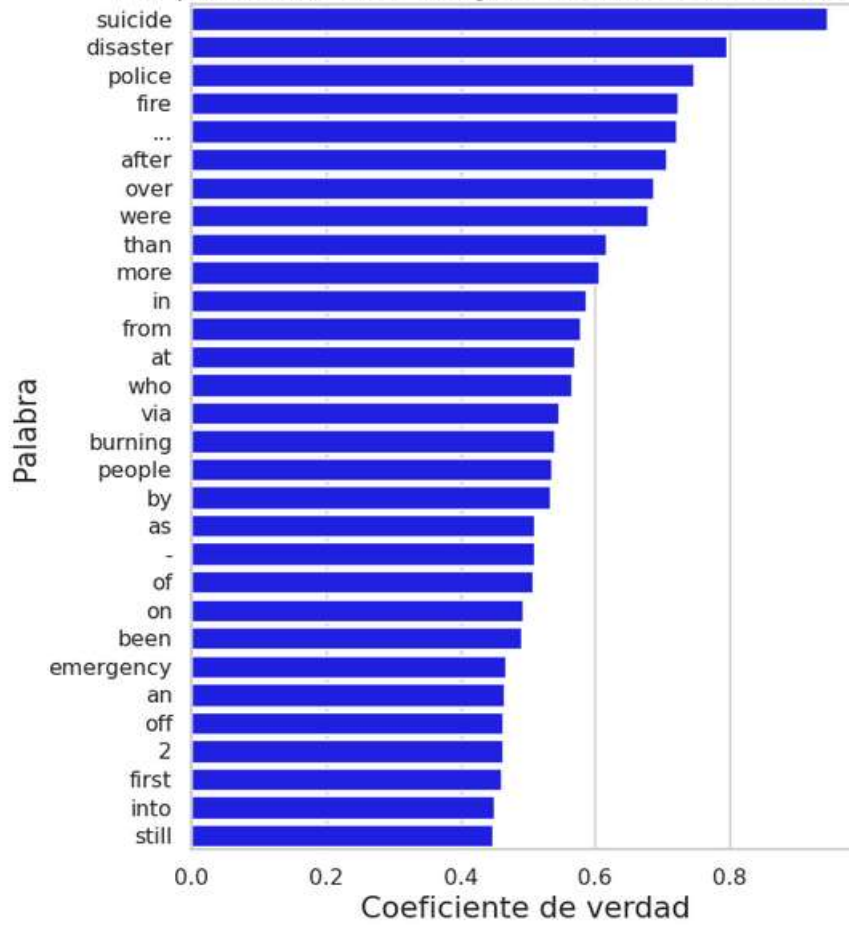
count	27,983.00
mean	4.05
std	37.58
min	1.00
25%	1.00
50%	1.00
75%	2.00
max	3,207.00

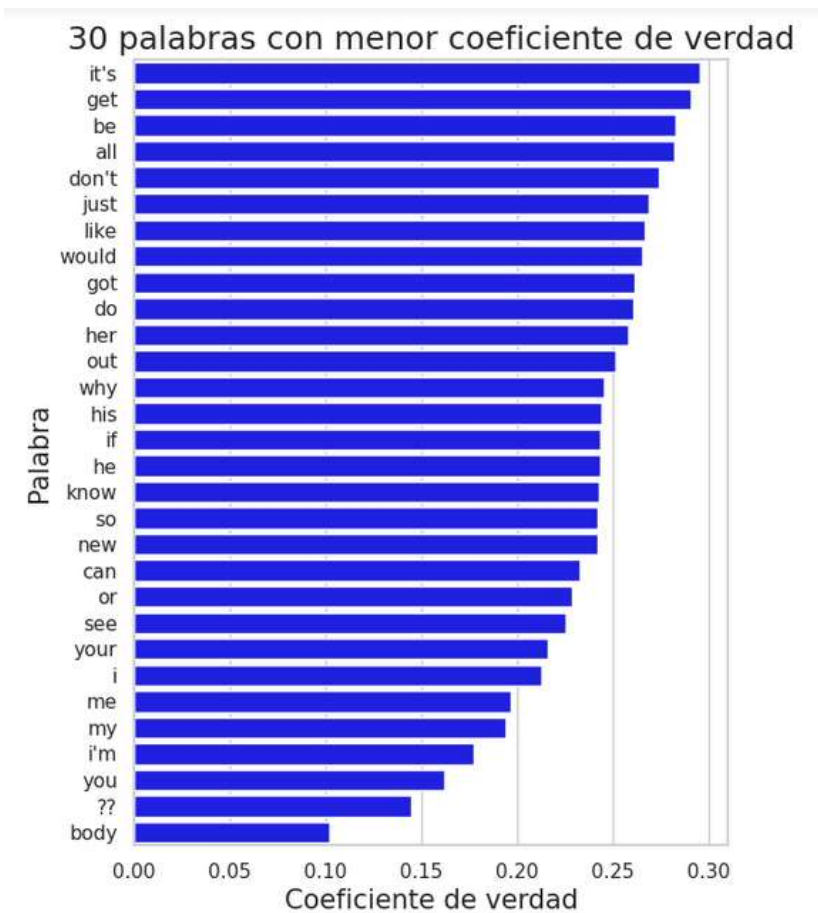
Puede verse que la distribución de la cantidad de apariciones favorece en gran medida valores bajos, lo que convierte a esos coeficientes en valores que verdaderamente no aportan información. Se decide imponer una restricción a los datos que se van a analizar: para considerarse valido el valor de veracidad, la palabra analizada debe de haber aparecido en al menos 100 tweets.

Son 94 las palabras que cumplen con esta restricción. De las mismas nos interesa saber cuáles son las 30 palabras que mayor y menor coeficiente tienen.

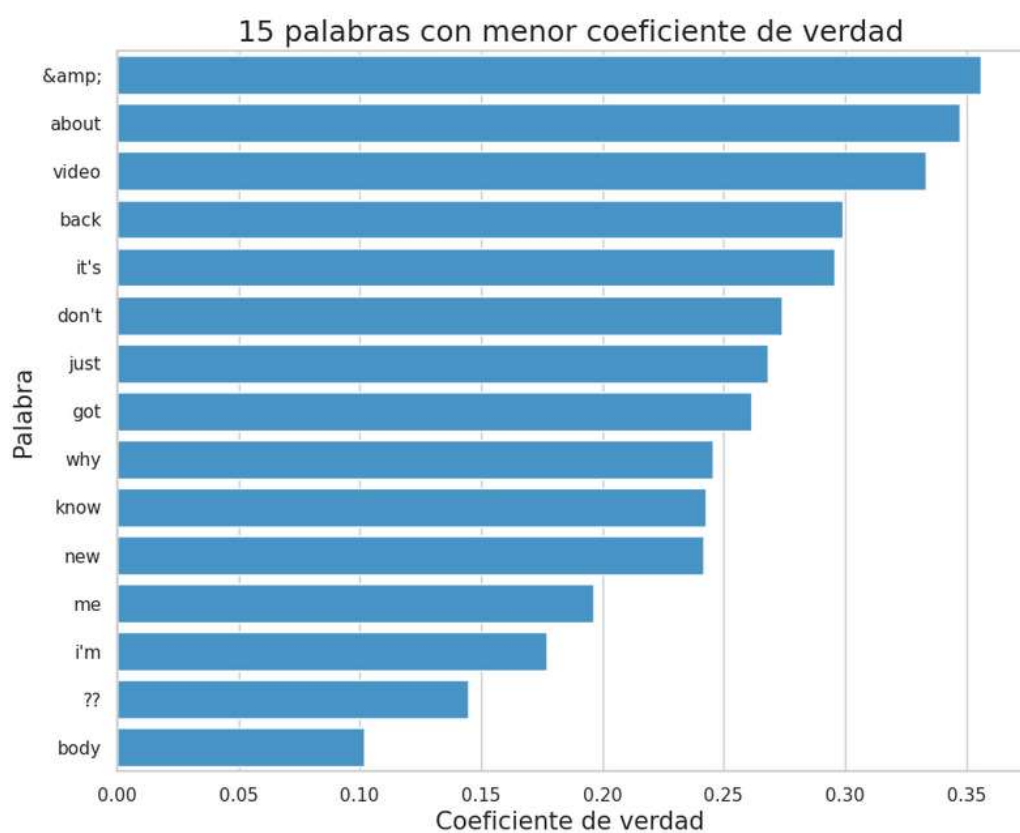
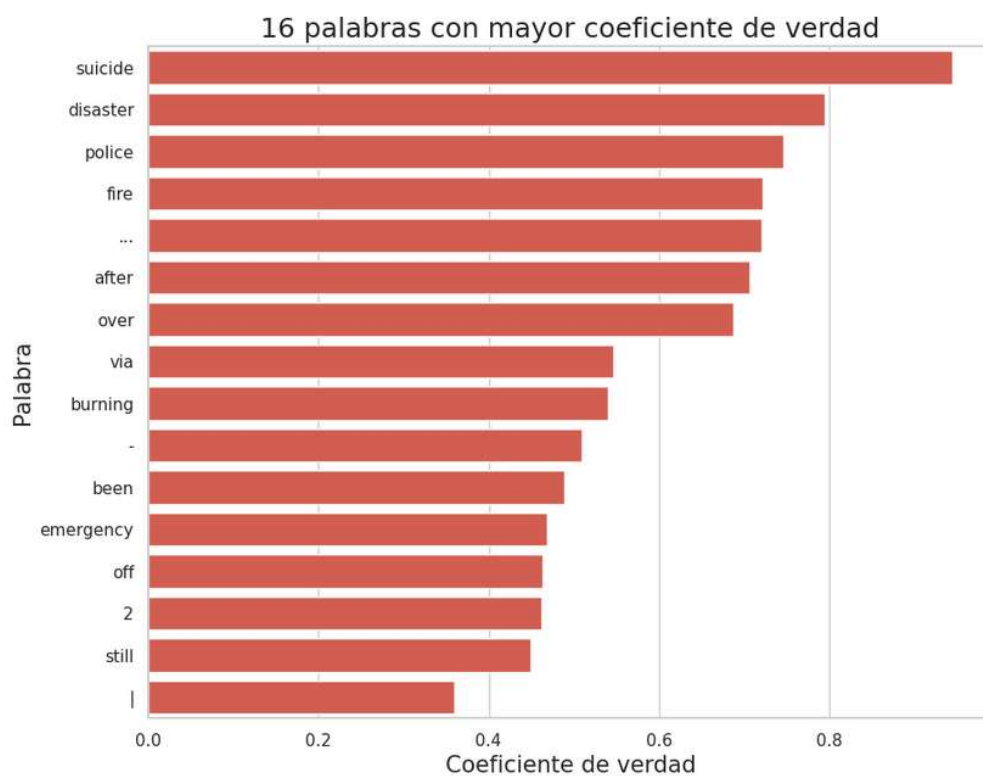


### 30 palabras con mayor coeficiente de verdad





Como puede verse, entre estos valores de coeficientes de verdad, hay artículos, pronombres y preposiciones, lo cuál tiene sentido, pues son una parte fundamental para la comunicación del lenguaje. Estas palabras son muy comunes, por lo que su tasa de aparición naturalmente va a ser bastante alta. Si bien cumplen con el requisito de apariciones, el hecho de ser tan comunes no brinda información útil para un análisis de veracidad basado en palabras de un tweet. Se decide entonces aplicarle un filtro a las palabras que analizamos. Tomamos las 100 palabras más comunes del lenguaje inglés, y filtramos las mismas de las palabras analizadas. Como resultado, nos quedan 31 palabras que no pertenecen a las más comunes y tienen más de 100 apariciones en nuestro set de información.



### 3. Conclusión Final

El filtrado de palabras aparecidas en un tweet en base a su cantidad de apariciones (para eliminar outliers) y en base a qué tan comunes son en el idioma en el que se escribieron los tweets, nos permite asignarle un “peso” a cada palabra, el cuál puede ser usado a futuro para predecir si un tweet es verdadero o no basado únicamente en su contenido. El set de datos analizado posee una fuerte tendencia a palabras que aparecen una o dos veces, razón por la cual se puso como requisito el haber sido usado un número mayor a 100 veces. Suponiendo que se tuviera un set de datos no tan orientado a palabras de poca aparición, este requisito puede ser reducido y en consecuencia expandir el rango de predicción, dado que puede ocurrir que haya tweets que no poseen ninguna de las palabras a las que nosotros asignamos un valor, y sin embargo son verdaderos, generando una predicción errónea.

De los análisis realizados de los tweets, sólo cuando estos se realizan en países, la localización compartida en el set de datos es un país o un conjunto de países como es Gran Bretaña, la cantidad de casos reales es mayor a la de los casos no reales. A su vez, se observa que hay localizaciones que no corresponden a ciudades o países y en estas solo predominan los casos no reales, por ende pueden haber sido una persona o un grupo de personas cercanas creando estos tweets falsos para que predominen.