



**FACULTAD
DE INGENIERIA**
Universidad de Buenos Aires

[75.06] ORGANIZACIÓN DE DATOS

1º CUATRIMESTRE 2020

CURSO ÚNICO

TP1 - Real Or Not ?

AUTORES - Grupo COVIgData - 19

Carbón Posse, Ana Sofía - #101 187
<scarbon@fi.uba.ar>

Giampieri, Leonardo - #102 358
<lgiampieri@fi.uba.ar>

Gojman, Lautaro - #101 185
<lgoijman@fi.uba.ar>

Rodriguez Dala, Tomás - #102 361
<trodriguez@fi.uba.ar>

DOCENTES

Argerich, Luis

Golmar, Natalia

Ramos Mejia, Martín

Martinelli, Damian

Índice

1. Introducción	3
1.1. ¿Qué es Twitter?	3
1.2. Elementos principales de Twitter	3
2. Set de Datos	4
2.1. Filas:	4
2.2. Columnas:	4
2.3. Manipulación de Datos	5
3. Análisis Exploratorio	6
3.1. Longitud de los Tweets	7
3.2. Palabras comunes en los Tweets	8
3.2.1. ¿Por qué nos interesan?	8
3.2.2. Detalles sobre las análisis	9
3.2.3. ¿Cuáles son las palabras más usadas?	9
3.2.4. Palabras únicas	10
3.2.5. ¿Qué pasa cuando separamos los tweets en verdaderos y falsos?	11
3.3. Hashtags en los Tweets	12
3.3.1. ¿Qué es un hashtag? ¿Por qué nos importan?	12
3.3.2. Detalles sobre las análisis	13
3.3.3. Vista general	13
3.3.4. ¿Cuáles son los hashtags mas frecuentes en tweets verdaderos?	14
3.4. Menciones en los Tweets	16
3.4.1. ¿Qué son las menciones? ¿Por qué nos interesan?	16
3.4.2. ¿Cuáles son los usuarios más mencionados?	16
3.4.3. Análisis de @YouTube	17
3.4.4. ¿Qué pasa con el resto de las menciones?	17
3.5. Análisis de las Keywords	19
3.5.1. ¿Qué son las Keywords?	19
3.5.2. Detalles sobre el análisis	20
3.5.3. Top 10 Keywords más frecuentes en el set	21
3.5.4. Frecuencia de keywords por su veracidad	22
3.5.5. Keywords que aparecen solo en tweets verdaderos o solo en tweets falsos	23
3.5.6. Keywords en proporción similar en tweets verdaderos y falsos	25
3.5.7. Cantidad de palabras	26
3.5.8. Relación con la longitud de los tweets	28
3.6. Análisis de las Locations	29
3.6.1. Detalles del análisis	30
3.6.2. Análisis por país	32
3.6.3. Análisis por ciudades	34
3.6.4. Análisis por estados de Estados Unidos	36

4. Conclusiones	38
4.1. Insights	38
5. Código	39

1. Introducción

El objetivo de este primer trabajo es realizar un análisis exploratorio del set de datos otorgado de la competencia: <https://www.kaggle.com/c/nlp-getting-started>. Esta misma contiene tweets realizados por distintas personas alrededor de todo el mundo anunciando una posible emergencia. Pero, no siempre está claro si las palabras de las personas realmente anuncian un desastre.

A medida que se exploren los datos, queremos ver que cosas podemos descubrir sobre ellos que puedan resultar interesantes. Se propone inicialmente analizar:

- El contenido de los tweets.
- El formato de los tweets.
- Las distintas ubicaciones dónde se generan los tweets.

Finalmente, entre el análisis exploratorio y los ítems estudiados, se busca obtener *insights* aprendidos sobre los mismos y con ellos poder predecir si un tweet va a ser real o no.

1.1. ¿Qué es Twitter?

Es un servicio de comunicación bidireccional con el que puedes compartir información de diverso tipo de una forma rápida, sencilla y gratuita. En otras palabras, se trata de una de las redes de microblogging más populares que existen en la actualidad y su éxito reside en el envío de mensajes cortos llamados “tweets”. Fue creada por Jack Dorsey y su equipo en 2006 y la idea se inspira en el envío de fragmentos cortos de texto, donde puedes añadir un enlace, imágenes, vídeo, encuestas o incluso gifs.

1.2. Elementos principales de Twitter

- **Tweet** : Es cada uno de los mensajes que se publica. Recordemos que cada uno de ellos contiene hasta 280 caracteres sin contar el material multimedia que incluyas en tus contenidos.
- **Hashtag (#)** : Se representa con el numeral y permite añadir tras él los términos que queramos. Se utiliza para facilitar búsquedas.
- **Mención (@)** : Cuando escribimos un tweet y queremos nombrar a una persona (o varias) que es usuario de Twitter, añadiremos su nombre de usuario precedido de @. De esta manera, el usuario recibirá notificación en sus menciones y podrá respondernos (si así lo desea).
- **Ubicación** : Permite añadir la ubicación de lo comentado en el tweet.

2. Set de Datos

Lo primero a analizar es la estructura general de los datos. De esta manera podemos comenzar a tener una idea de qué aporta cada columna y que podemos hacer con ellas. En la Figura 2.1 se muestra la estructura inicial de nuestro set de datos:

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1
...
7608	10869	NaN	NaN	Two giant cranes holding a bridge collapse int...	1
7609	10870	NaN	NaN	@aria_ahraray @TheTawniest The out of control w...	1
7610	10871	NaN	NaN	M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt...	1
7611	10872	NaN	NaN	Police investigating after an e-bike collided ...	1
7612	10873	NaN	NaN	The Latest: More Homes Razed by Northern Calif...	1

7613 rows × 5 columns

Figura 2.1

- Este mismo pesa **297.5 KB**.
- No sabemos a que periodo de tiempo pertenecen estos tweets.

2.1. Filas:

- Hay 7613 filas.
- Cada fila identifica un tweet distinto.

2.2. Columnas:

- Hay 5 columnas.
- No todas presentan información, algunos fueron completados con **NaN**.

El contenido de las 5 columnas son :

- **id**: Identificador único para cada tweet.

- **keyword:** Palabra clave del tweet que indica que trata de una emergencia.
- **location:** La ubicación de donde fue el tweet publicado.
- **text:** El texto del tweet.
- **target:** Identifica si el tweet habla de una emergencia o no.

2.3. Manipulación de Datos

Decidimos manipular el set de datos para facilitar su análisis. A pesar de que el set de datos es pequeño, esto nos beneficia no solo a ocupar menos memoria, si no que también a tener un tiempo de ejecución mas corto.

- **Conversión de tipo de datos :** La columna target hace referencia a si un tweet anuncia una emergencia (1) ó, caso contrario, no anuncia una emergencia (0). Nos pareció conveniente convertirlo al tipo booleano ya que se hace mas llevadero trabajar con 'True' y 'False', que con '1' y '0' . Además tiene la ventaja de reducir uso de memoria.
- **Data Mining :** Se generan nuevos sets de datos y se extraen columnas importantes de los proporcionados. Unas de estas modificaciones son: la eliminación de la columna id ya que no revela ningún tipo de información importante para nuestro análisis, también agregamos nuevas columnas tales como 'Country', 'City', 'State' para poder analizar de mejor manera las ubicaciones de los tweets.

Cada uno de estos procesos serán detallados en sus respectivas secciones.

3. Análisis Exploratorio

Como dijimos en la introducción, el objetivo de este trabajo es realizar un análisis exploratorio del set de datos. Vamos a ver qué cosas podemos descubrir sobre los datos que puedan resultar interesantes.

Para empezar, dividimos el set original en dos: Uno que contiene únicamente los tweets verdaderos y el otro los tweets falsos. De esta manera, resultó más fácil poder analizar cada caso por separado.

Empezamos por contar la cantidad de tweets que había para cada target y lo que pudimos observar fue que **hay más tweets Falsos que Verdaderos**, como podemos observarlo en la Figura: 3.1.

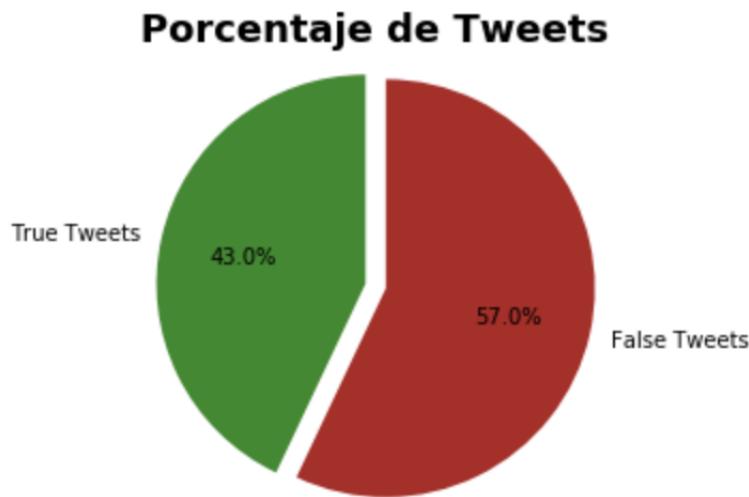


Figura 3.1

De un total de **7613** tweets (100%), **4342** tweets son falsos (57%) y **3271** tweets son verdaderos (43%). Esto nos beneficia ya que, como nos dice la ecuación del desvío estándar de De Moivre, si el tamaño del conjunto de datos es chico entonces es más propenso a que sus elementos estén más alejados de la media. Como la diferencia de tamaños entre ambos es pequeña, van a tener desviaciones similares. Esto significa que no debería haber problemas a la hora de comparar cada uno de los casos. En las siguientes secciones vamos a proceder a analizar el contenido de los tweets, para ver si encontramos información interesante sobre ellos.

3.1. Longitud de los Tweets

En esta sección se busca analizar si la longitud (en caracteres) de los tweets tiene relación alguna con la veracidad de los mismos.

Como bien dijimos mas arriba, el conjunto de datos para cada target es lo suficientemente grande para poder calcular ciertas estadísticas sin obtener un gran desvío. Vamos a calcular la longitud máxima, la longitud mínima y la longitud promedio. Estas las podemos observar en la Figura: 3.2.

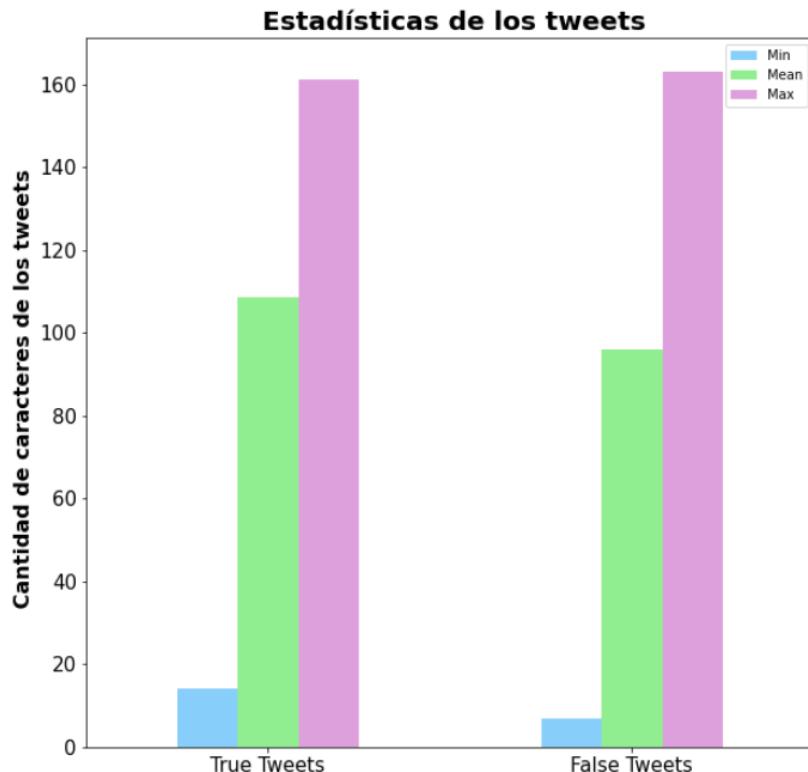


Figura 3.2

Con estos resultados podemos observar que primero los tweets verdaderos tienden a ser levemente más largos y están más concentrados en un rango de caracteres, mientras el rango de concentración de los tweets falsos se encuentra mas distribuido a lo largo de la cantidad de caracteres. Estos resultados también se pueden ver plasmados en la Figura 3.3.

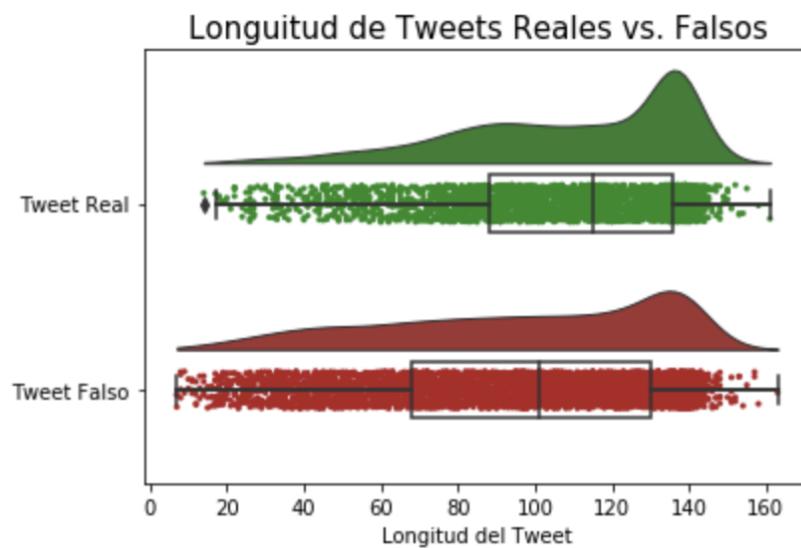


Figura 3.3

En base a la Figura 3.3 se puede observar a simple vista que los tweets verdaderos tienden a tener una longitud de caracteres mayor que los falsos. Los tweets verdaderos tienen una densidad ampliamente mayor en el rango de los 120 a 150 caracteres, mientras que los tweets falsos están distribuidos de manera un poco más uniforme pero también tienen un pico menor al de los tweets verdaderos, en el rango de 130-140 caracteres.

Frente a esto podemos concluir que un tweet que se encuentre en el rango de 120-150 caracteres tiene altas probabilidades de ser verdadero. De todas maneras las diferencias entre las distribuciones y densidades no son tan grandes y tienen formas muy parecidas. En base a esto no creo que la longitud de los tweets juegue un rol importante en determinar la veracidad del mismo.

3.2. Palabras comunes en los Tweets

3.2.1. ¿Por qué nos interesan?

Por más que se puedan agregar fotos o videos, Twitter sigue siendo una plataforma donde la principal forma de expresarse es mediante texto. No todas las personas escriben de la misma manera y no todos los tweets hablan sobre lo mismo, pero todos tienen algo en común, utilizan palabras. Podemos extraer mucha información de las palabras que cada persona decide utilizar. ¿Cuáles son las palabras más comunes? ¿Hay palabras que se repiten más en los tweets verdaderos? ¿De qué temas hablan las palabras más usadas?

Esperamos que las próximas visualizaciones nos ayuden a sacar algunas conclusiones sobre las propiedades del set de datos.

3.2.2. Detalles sobre las análisis

A la hora de decidir cuales son la palabras a tener en cuenta y cuales no decidimos utilizar una lista de stopwords que nos proporciona la biblioteca *nltk* (Natural Language Toolkit). Ignoramos mayúsculas y minúsculas así juntamos a todas las palabras iguales. Además agregamos algunas palabras que no nos aportaban nada de información (como *i'm* y ???).

3.2.3. ¿Cuáles son las palabras más usadas?

Empecemos con un panorama general (Figura 3.4) de las palabras más usadas.

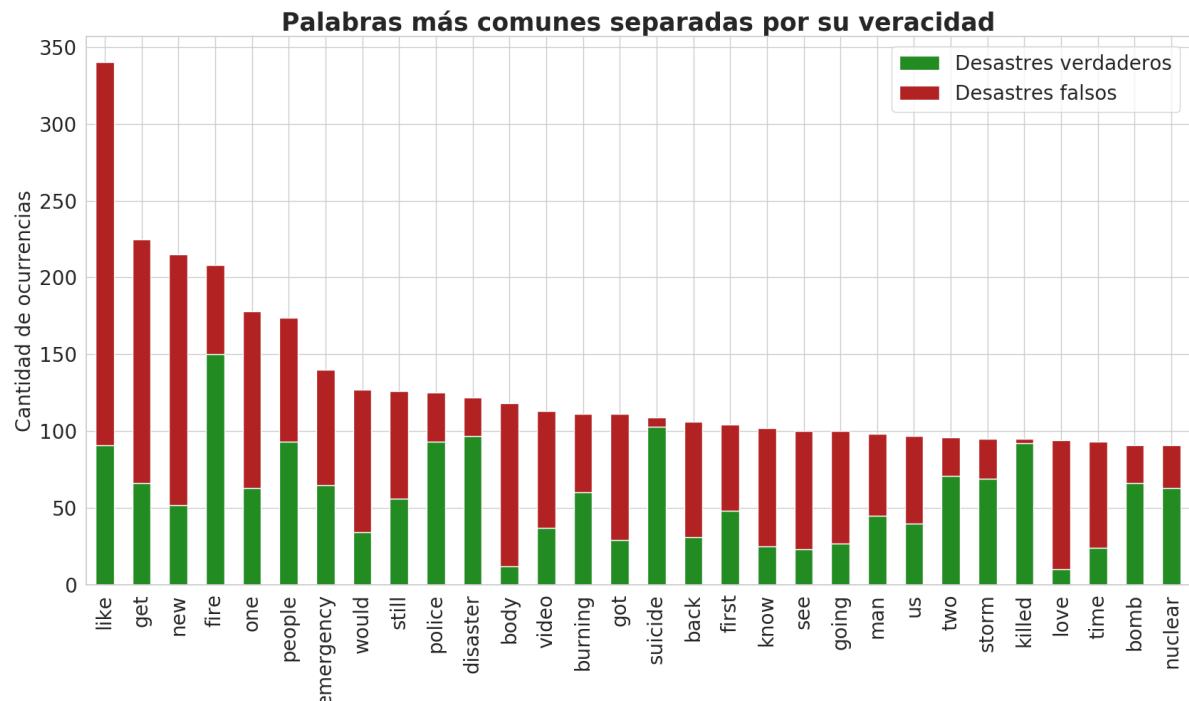


Figura 3.4

Encontramos palabras bastante comunes e inocentes al principio (*like*, *get*, *new*) pero también hay algunas que tienen más sentido en el contexto de un desastre (*fire*, *emergency*, *police*). Algo curioso que podemos ver es que hay algunas palabras que tienen connotaciones más serias como *emergency* o *burning* tienen cantidades muy parecidas de tweets verdaderos y falsos. Esto se debe a que son palabras que dependen mucho del contexto. Aquí hay algunos usos de estas palabras en tweets falsos:

- *It's raining outside I'm burning my favorite candle and there's hot cocoa to sip. Nap time = awesomesauce.* - id: 1877
- *Ah yes the gays are totally destroying America. I can see buildings burning and meteors crashing into schools wow* - id: 1956
- *God forbid anyone in my family knows how to answer a phone. I need new emergency contacts.* - id: 4511

Otro caso especial es el de la palabra *body* que puede ser usada para referirse a un cadáver, pero se ve que en estos tweets la gran mayoría no lo utiliza de esa manera por lo que terminan en tweets falsos como por ejemplo:

- *Why tf did I decide to workout today? My **body** feels like it's been engulfed by a mass of fiery disdain.* - id: 4679
- *Forsure back in the gym tomorrow. **Body** isn't even at 50%. Don't wanna risk injuries.*
- id: 6515

3.2.4. Palabras únicas

Además de analizar cuáles son las palabras que más aparecen, también puede ser interesante la cantidad de palabras únicas utilizadas en cada caso.

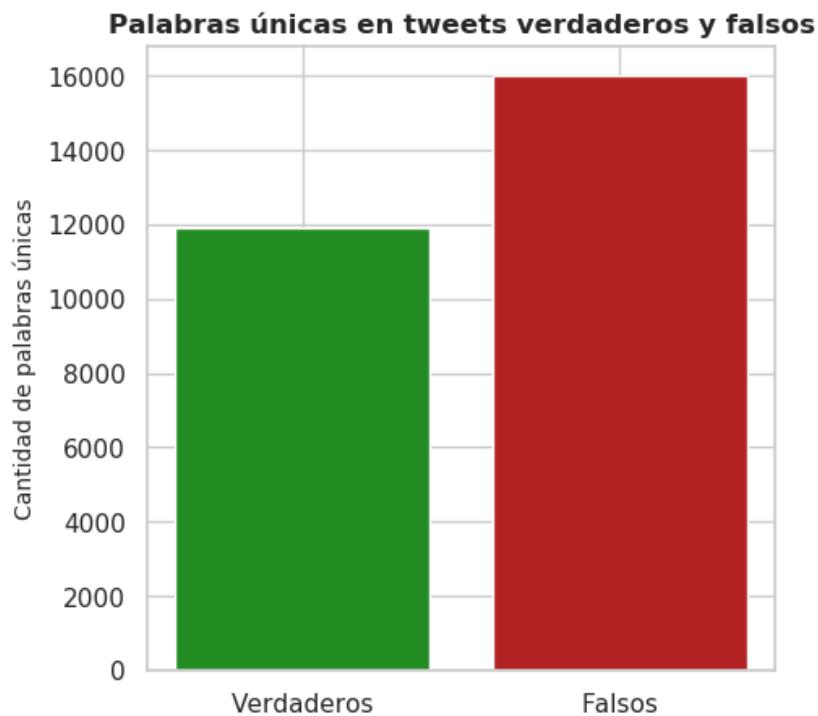


Figura 3.5

De la Figura 3.5 podemos sacar la conclusión de que el vocabulario de los tweets falsos es más extenso. Tiene sentido ya que si el tweet está hablando de un desastre real probablemente utilice un vocabulario más acotado con términos relacionados al tema. Ahora faltaría analizar bien cual es ese vocabulario utilizado para desastres con la esperanza de poder familiarizarnos con el set de datos.



Figura 3.7

Estas palabras en la Figura 3.7, contrastan mucho con las de la visualización anterior. No tratan de ningún tema en particular y son de uso más cotidiano.

Creo que se ve claramente que el vocabulario de los tweets verdaderos y los falsos es muy diferente. La información que nos proporcionan estas visualizaciones pueden ser de gran utilidad a la hora de predecir tweets en el futuro. Sabemos cuales palabras son una señal de alerta y cuales podemos pasar por alto.

3.3. Hashtags en los Tweets

3.3.1. ¿Qué es un hashtag? ¿Por qué nos importan?

Un hashtag se usa para categorizar tweets y facilitar su búsqueda. En general se añaden con palabras clave que resumen el tema de conversación. Por ejemplo, si yo quiero tweets que hablen de perros bastaría con buscar `#perros`.

Algunos usos de los hashtags pueden ser:

- Hablar sobre lugares: *24 killed in two simultaneous rail crash as acute floods derail the two trains #India #mumbai...* <http://t.co/4KBWPCmMbM> - id: 3415
- Reemplazar palabras importantes: *13,000 people receive #wildfires evacuation orders in California* - id: 6

- Referirse a eventos actuales: *Court back in session. Testimony continues with med. examiner discussing gunshot wounds #KerrickTrial* - id: 10675

Pero sin importar el uso esta clara la razón por la cual nos interesa analizar sus ocurrencias. Denotan palabras clave en el texto que podemos usar para categorizar el tweet. Queremos saber, ¿Cuáles son los que se usan más frecuentemente? ¿Cuáles aparecen más en los tweets verdaderos? ¿Que nos dicen los hashtags encontrados sobre el tweet? Entre otras cosas. Buscamos crear visualizaciones que reflejen la información que nos proporciona el set de datos y que nos ayuden a responder estas mismas preguntas. Esperamos encontrar hashtags con temas más serios en los tweets verdaderos con palabras que se usen para describir desastres o lugares donde ocurrieron, mientras que en los falsos se hable sobre temas más variados.

3.3.2. Detalles sobre las análisis

Algunas cosas para tener en cuenta:

- No estamos diferenciando entre mayúscula y minúscula. Esto es para evitar separar hashtags que hablan de exactamente lo mismo (por ejemplo, #News y #news).
- Estamos ignorando hashtags que creemos que no nos proporcionan nada de información (por ejemplo, #Icesx89Û_ y #????????).

3.3.3. Vista general

Para la primera visualización (Figura 3.8) queríamos ver cuales eran los hashtags más frecuentes y su proporción de verdaderos y falsos.

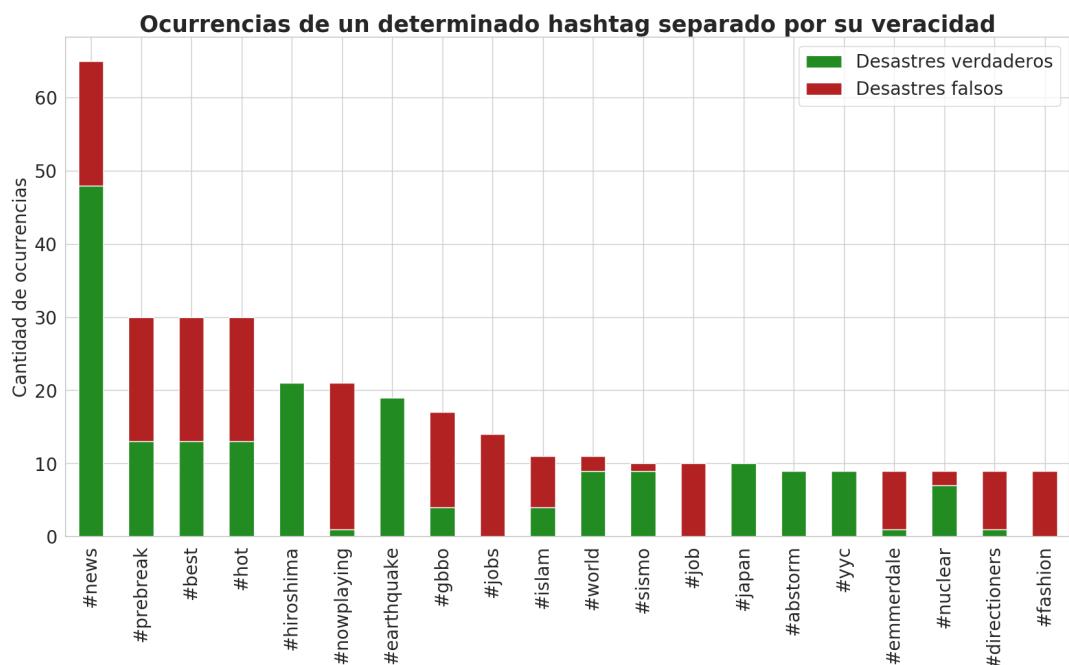


Figura 3.8

- **#beyhive y #directioners:** Son nombres de los fandom de Beyoncé y One Direction respectivamente.

3.4. Menciones en los Tweets

3.4.1. ¿Qué son las menciones? ¿Por qué nos interesan?

En Twitter además de los hashtags se pueden agregar menciones. Estas se utilizan cuando nos referimos a otra persona que también se encuentra en Twitter y queremos que le llegue una notificación. Cualquier persona puede mencionar a cualquier otro usuario en la plataforma y estos usuarios pueden ser de personas particulares u organizaciones. Algunos ejemplos de su uso son:

- *Italy: Three dead after landslide in the Italian Alps: http://t.co/42MawZb8T9 via @YouTube* - id: 6686
- *@ArianaGrande @justinbieber OMGGGG IM SCREAMING* - id: 8541

Tiene sentido pensar que la probabilidad de que un usuario sea mencionado en un tweet real sea siempre la misma. Después de lo que vimos en la sección de los hashtags, probablemente veamos que menciones a medios de comunicación u organizaciones gubernamentales. Esperamos que en análisis de las menciones nos digan que usuarios de Twitter son más propensos a ser mencionados en tweets sobre desastres reales.

3.4.2. ¿Cuáles son los usuarios más mencionados?

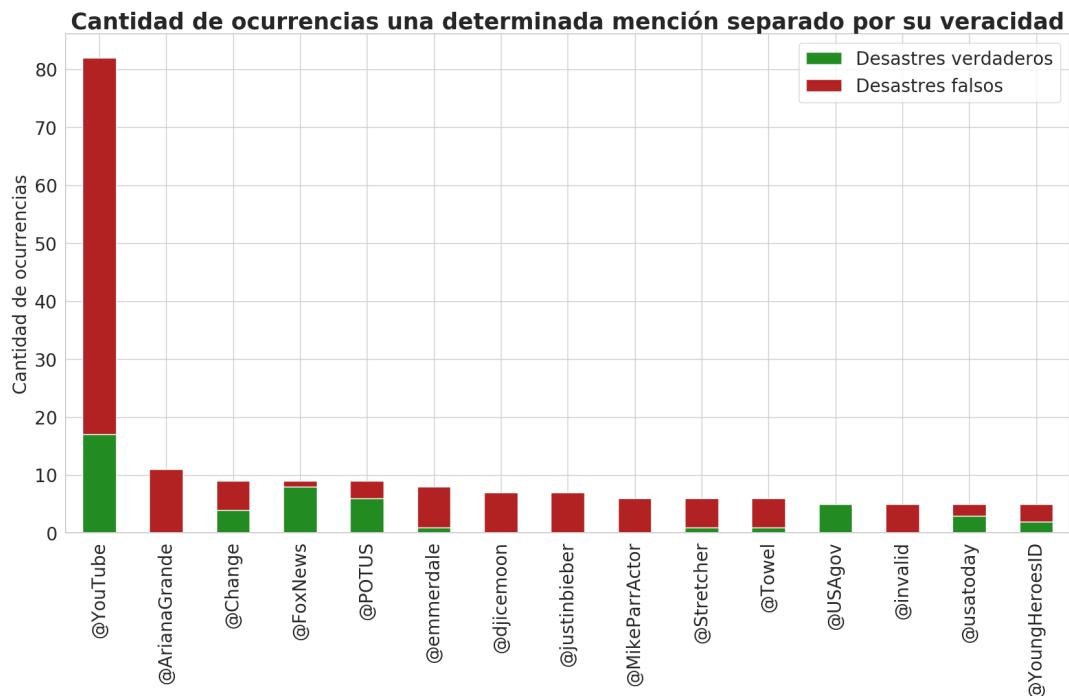


Figura 3.11

Vemos en la Figura 3.11 que @YouTube está primero y tiene casi 8 veces más que el segundo. La ley de los grandes números nos dificulta las cosas para realizar el análisis porque no podemos comparar las menciones de forma justa. Por esta razón vamos a analizar a YouTube en una sección separada. Algo que si podemos observar es que de los que aparecen en la Figura 3.11 hay muy pocos con cantidades parecidas de falsos y verdaderos (como tiene @Change). Casi todos tienen una clara mayoría para un lado o para otro. Del lado de los verdaderos tenemos algunos usuarios como @FoxNews, @POTUS, @USAgov y del otro tenemos a @ArianaGrande, @emmerdale, @justinbieber.

3.4.3. Análisis de @YouTube

YouTube es una plataforma muy popular para compartir videos. A veces cuando alguien comparte o habla sobre un video por Twitter mencionan a @YouTube. Esto significa que, como vimos en la visualización anterior, @YouTube tiene muchas más instancias que el resto de las menciones así que vamos ver su proporción de tweets verdaderos y falsos.

Tweets con la mención @YouTube separados por veracidad

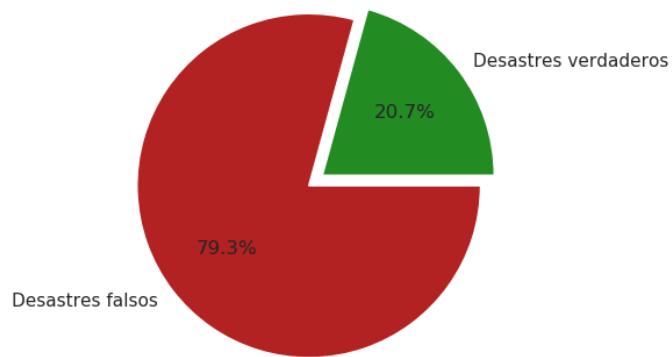


Figura 3.12

Vemos que cerca de 1 cada 5 tweets que mencionan a @YouTube son verdaderos, no es una gran proporción así que no lo podemos usar para predecir desastres verdaderos. El problema es que por más que se utilice YouTube para compartir videos de desastres, siempre va a haber más gente usándolo para otras cosas así que no es un indicador claro.

3.4.4. ¿Qué pasa con el resto de las menciones?

Recordemos que el resto de las menciones no aparecían en cantidades tan grandes por lo que no van a ser tan significativas a la hora de predecir los desastres. Sin embargo, analizarlas puede ser útil para familiarizarnos más con el set de datos y con las propiedades de los tweets verdaderos y falsos.

Frecuencia de las menciones en los tweets reales

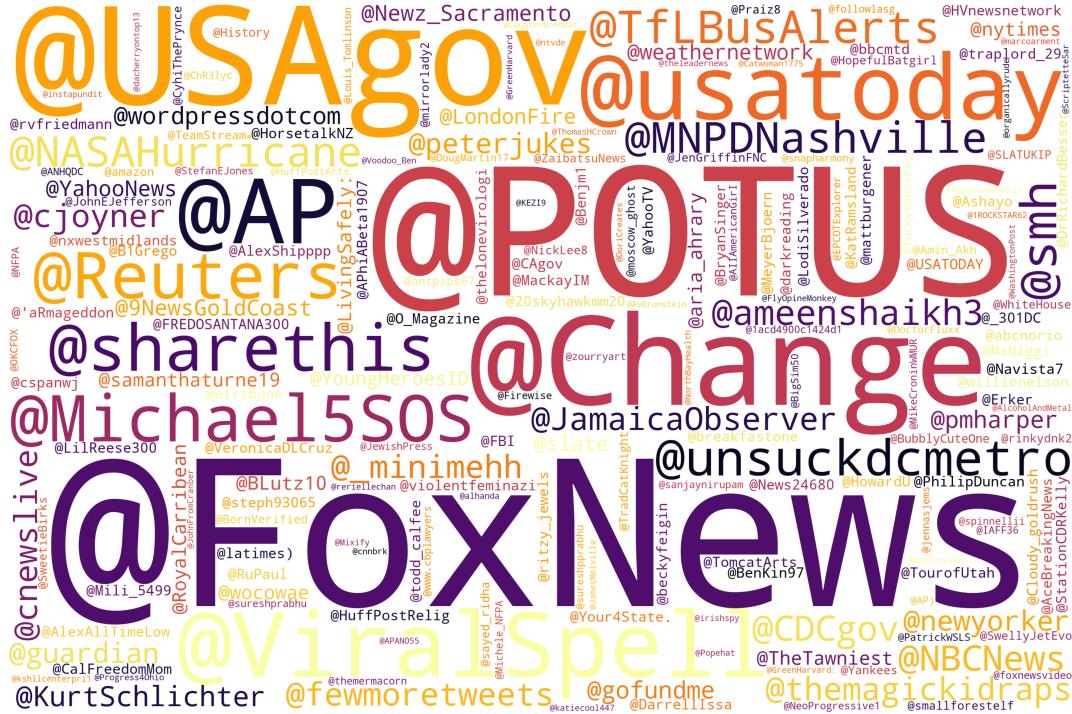


Figura 3.13

Frecuencia de las menciones en los tweets falsos

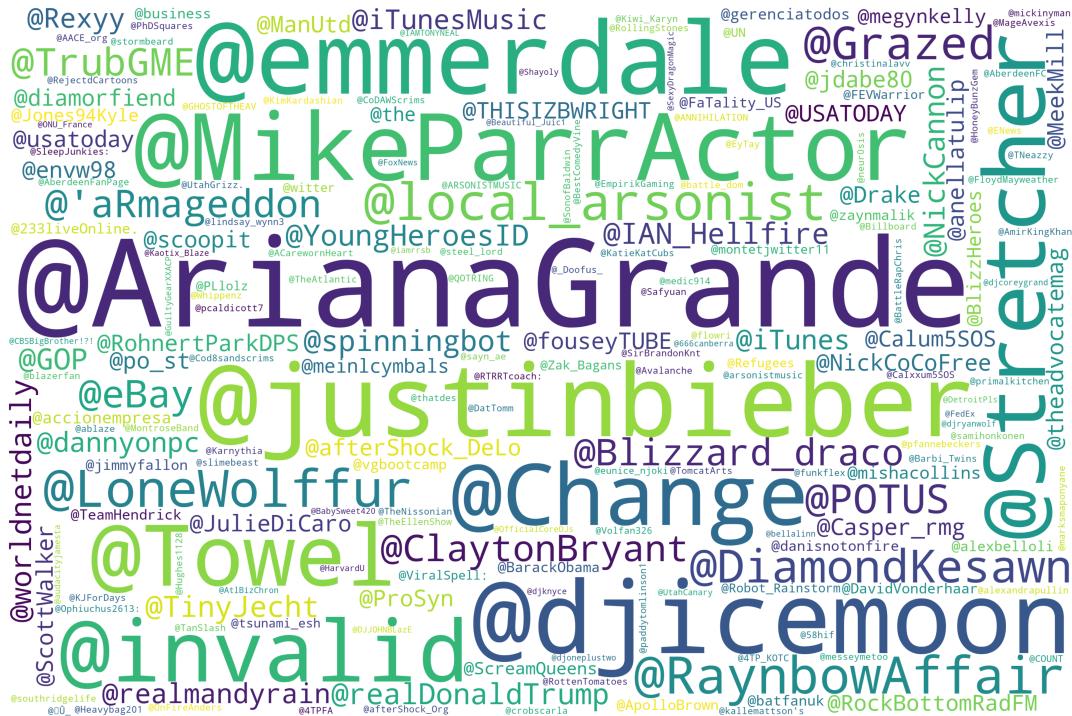


Figura 3.14

Volvemos a tener la misma separación que en las Figuras 3.9 y 3.10. Por el lado de los tweets verdaderos tenemos usuarios que pertenecen a organizaciones gubernamentales o de medios de comunicación. Algunos de estos usuarios son:

- **@FoxNews:** Fox News es un canal de noticias televisado en los Estados Unidos.
- **@POTUS:** Cuenta del presidente de los EEUU (siglas para “President Of The United States”).
- **@USAgov:** Utilizada por el gobierno de los EEUU para difundir información en las redes sociales.
- **@Change:** change.org es una plataforma que ayuda crear peticiones para la junta de firmas para incentivar cambios en el mundo.

Tal y como pasaba con los hashtags, para los tweets falsos tenemos cuentas de celebridades relacionadas con la música o televisión:

- **@ArianaGrande:** Cantante y compositora musical estadounidense.
- **@emmerdale:** Cuenta utilizada para publicitar la telenovela inglesa Emmerdale. Al parecer la muerte de uno de los personajes generó varios tweets mencionandola.
- **@MikeParrActor:** Actor inglés que trabajó en la telenovela Emmerdale.
- **@djicemoon:** Cuenta del compositor de música electrónica Ice Moon. Todos los tweets que lo mencionan tienen el mismo texto y están promocionando una de sus canciones llamada “Aftershock”.

3.5. Análisis de las Keywords

3.5.1. ¿Qué son las Keywords?

Las Keywords que se utilizan en el set de datos son palabras clave que se encuentran en el texto del Tweet. Puede que sean parte de una mención, de un hashtag o simplemente que pertenezcan al cuerpo del Tweet. Estas reflejan potencialmente el contenido del texto del Tweet, ya que fuera de contexto no indican si este se refiere a una catástrofe o no, pero nos pueden dar un buen indicio. Damos un ejemplo del set, en donde un mismo keyword aparece en dos tweets, uno que refiere a una catástrofe y otro que no.

On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE - id: 53

*How the West was burned: Thousands of wildfires **ablaze** in California alone - id: 66*

En el primer Tweet, vemos que ablaze se utiliza para describir al cielo nocturno, mientras que en el segundo refiere a las llamas de un incendio forestal, que claramente es una catástrofe. A esto nos referimos cuando decimos que los keywords están fuera de contexto sin su Tweet. Sobre los keywords queremos ver cuales son más frecuentes, cuales aparecen más en tweets verdaderos que en tweets falsos y viceversa, cuales aparecen en la misma proporción entre

tweets verdaderos y falsos, cuales aparecen únicamente en tweets verdaderos o en tweets falsos y qué es lo que les da unicidad. También queremos ver si la cantidad de palabras en un keyword impactan sobre la veracidad del tweet (algunas keywords tienen más de una palabra), y cual es la relación de los keywords con la longitud del texto del tweet.

3.5.2. Detalles sobre el análisis

- Para el análisis de las keywords no vamos a tener en cuenta aquellos tweets que no tengan una asignada, ya que si tratamos de filtrar nos encontramos con que algunos tweets pueden contener en su texto más de un keyword.

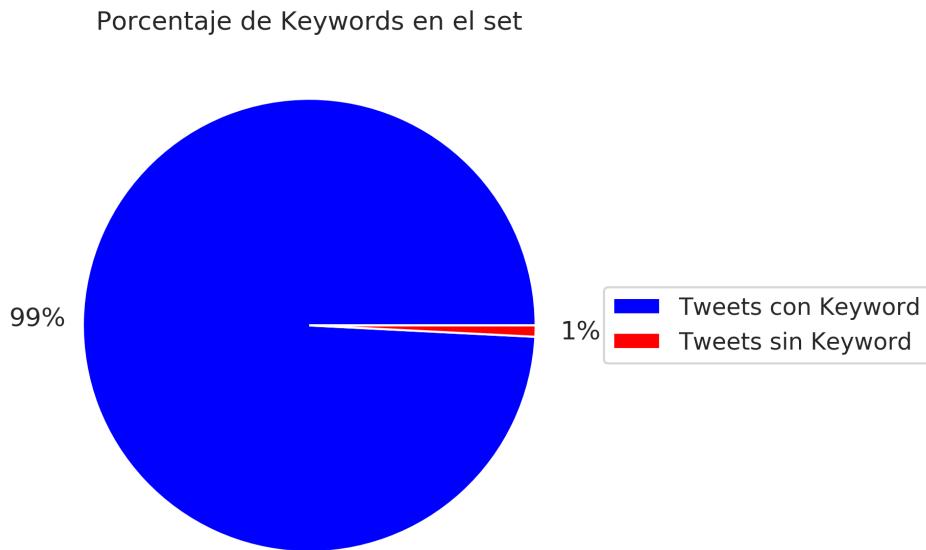


Figura 3.15

Como solo representan el 1% de todos los tweets, simplemente los eliminamos.

- Muchos tweets tienen texto muy similar, salvo por algún enlace, de modo que decidimos sacar esos tweets para que no tengamos agrupamiento de keywords que pueda meter basura en los datos.

3.5.3. Top 10 Keywords más frecuentes en el set

Queremos ver cuales son las keywords más frecuentes.

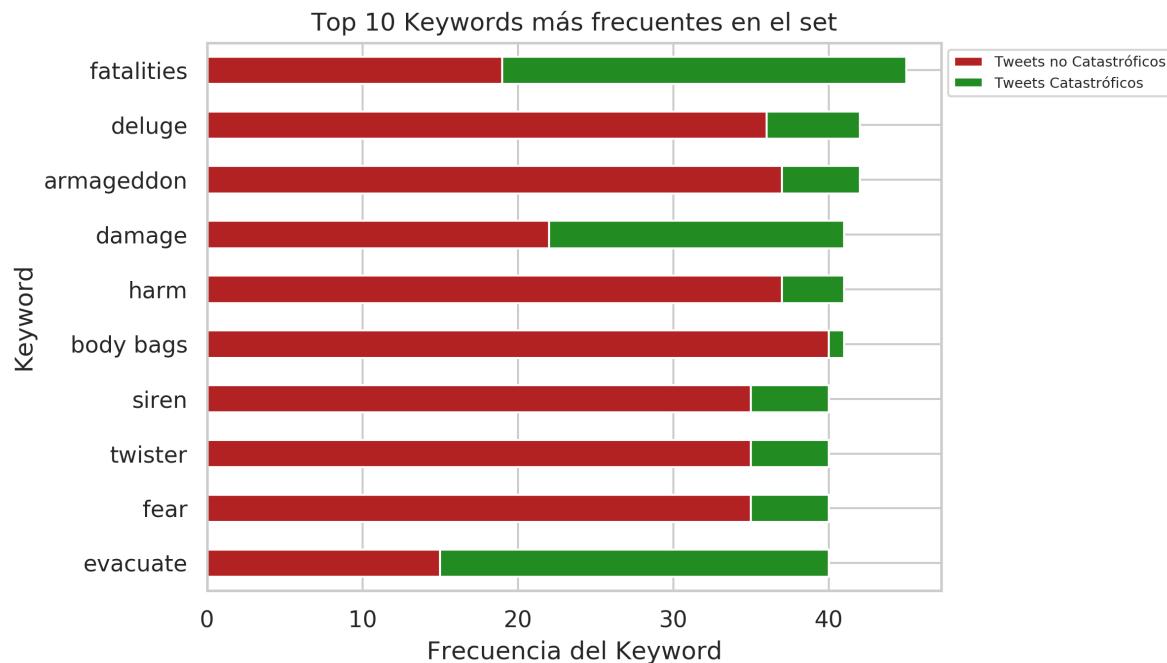


Figura 3.16

Vemos en principio que la distribución de la veracidad no es uniforme sobre los keywords. Algunos son mas frecuentes en tweets verdaderos que en falsos y viceversa, y algunos tienen una gran tendencia a alguno de los dos extremos, como es el caso de **body bags** o **evacuate**. Vemos que tiene sentido que una palabra como **armageddon** este en una cantidad alta de tweets que no refieren a catástrofes ya que por lo general se utiliza para engrandecer o exagerar alguna situación, mientras que **evacuate**, que está en su mayoría en tweets que refieren a catástrofes, se podría utilizar para describir la evacuación de gente de algún sitio peligroso, es un verbo más técnico. Es algo que se puede esperar del set, que keywords un tanto más específicos sobre alguna enfermedad o algún desastre aparezcan en su mayoría en tweets catastróficos. Algo que puede sorprender es la cantidad de tweets falsos en los que aparece el keyword **body bags**. Otro caso especial que aunque no está en el top nos parece interesante mostrar es el de la palabra **bloody** que significa 'sangriento' pero es utilizado, en su gran mayoria, en tweets falsos. Esto es porque también puede ser usado para expresar enojo, como lo vemos en los siguientes ejemplos:

- *I'm awful at painting.. why did I agree to do an A3 landscape in bloody oils of all paints ?? - id: 1307*
- *Bloody hell what a day. I haven't even really done anything. Just. Tired. Of everything. Thought vaca would help but it only did so much. =/ - id: 1301*

3.5.4. Frecuencia de keywords por su veracidad

Ahora queremos ver, para todas las keywords, cual es la dependencia entre las frecuencias de aparición en el set por verdadero/falso. Es decir, queremos ver si existe relación cuantitativa entre las veces que aparece un keyword en un tweet verdadero y en un tweet falso. Esto lo podemos ver en un scatter.

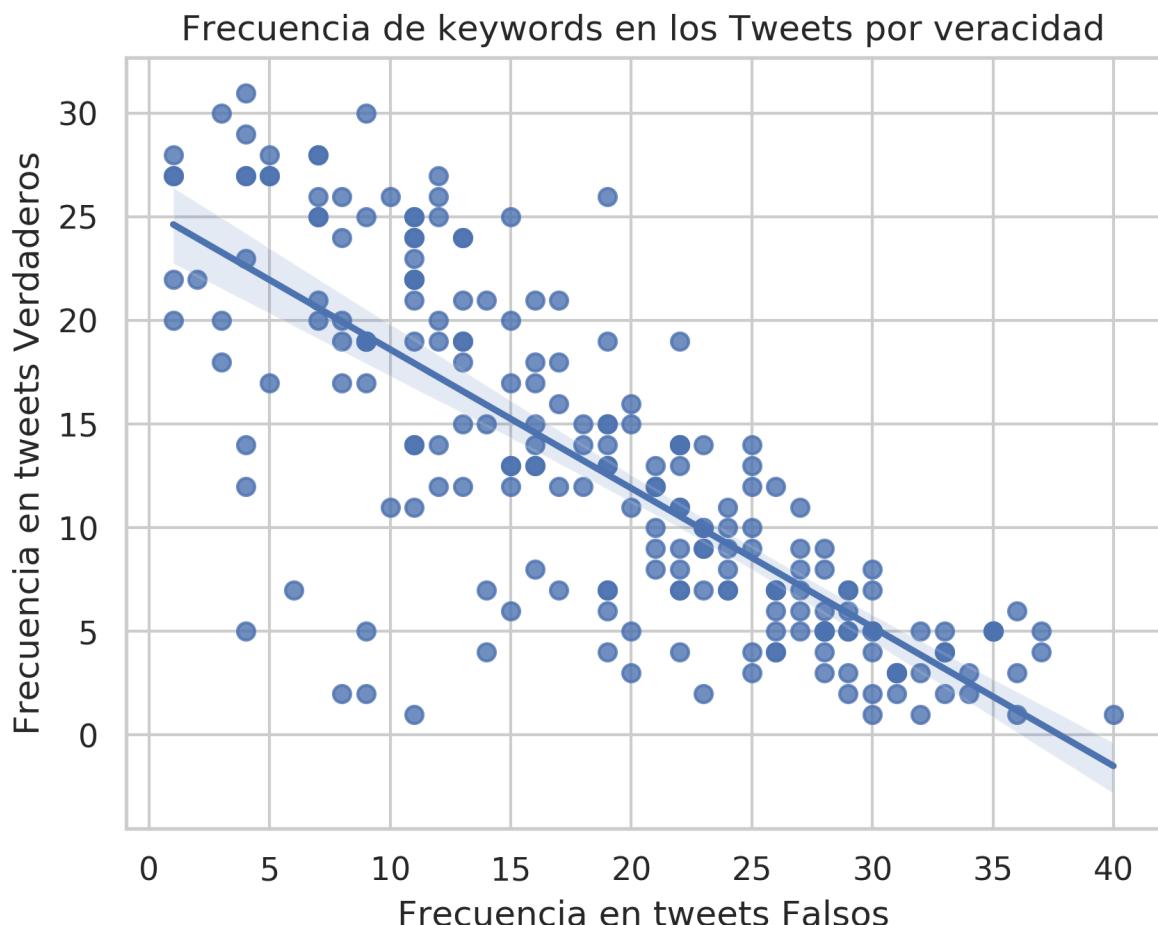


Figura 3.17

Observamos un poco más de dispersión en aquellos keywords que aparecen mas frecuentemente en tweets verdaderos. Vemos que los keywords que son mas frecuentes en tweets falsos están más juntos en el gráfico. En principio nos indica que los keywords más frecuentes en tweets falsos aparecen en una proporción similar. Algunas palabras son comunes en ambos casos, mientras que hay otras que son mas comunes dependiendo la veracidad del tweet. Lo que si sabemos es que la mayoría de los keywords aparecen en proporción distinta, y que a medida de que a cierto keyword lo encontramos en muchos tweets catastróficos, será menos probable encontrarlo en tweets no catastróficos, y viceversa, de ahí la tendencia de proporcionalidad inversa del gráfico.

3.5.5. Keywords que aparecen solo en tweets verdaderos o solo en tweets falsos

Vamos a analizar ahora los keywords que aparecen solo en tweets verdaderos o en tweets falsos.

	keyword	frequency_false_tweets	frequency_real_tweets
2	aftershock	27.00	nan
27	body bag	24.00	nan
218	debris	nan	33.00
219	derailment	nan	31.00
220	wreckage	nan	16.00

Figura 3.18

Empezamos por **aftershock**. Hace referencia a una réplica sísmica, pero también puede utilizarse para describir el impacto de alguna situación particular. En el set, los tweets hacen referencia sin embargo a una bebida alcohólica, a nombres de usuarios, a una montaña rusa en un theme park y a una película, entre otros. Mostramos algunos ejemplos.

- *Aftershock was the most terrifying best roller coaster I've ever been on. *DISCLAIMER* I've been on very few.* - id: 162
- *Tried orange aftershock today. My life will never be the same* - id: 185
- *Aftershock ↗ (2010) Full↗ Streaming - YouTube* - id: 176
- *@KJForDays I'm seeing them and Issues at aftershock ??* - id: 170
- *@afterShock_DeLo scuf ps live and the game... cya* - id: 146

Estos, como el resto de los tweets con este keyword, no refieren a un sismo. Hay muchas referencias a temas no catastróficos ya que si bien es una palabra utilizada técnicamente para describir una catástrofe, es frecuentemente utilizada en otros ámbitos. Es importante notar que los tweets que tengan el keyword en una mención probablemente no se refieran a una catástrofe.

La siguiente es **body bag**. Refiere a una bolsa para cadáveres. En el set, muchos tweets tienen este keyword ya que se confunde con **Crossbody bag**, una cartera. Vemos unos ejemplos.

- *New Ladies Shoulder Tote Handbag Women Cross Body Bag Faux Leather Fashion Purse - Full re* - id: 1379

- *Louis Vuitton Monogram Sophie Limited Edition Clutch Cross body Bag - Full read by eBay* - id: 1384
- *#handbag #fashion #style Vintage Coach Purse Camera Bag Cross Body #9973 16.99 (0 Bids)* - id: 1415

Vemos que poco tienen que ver con una catástrofe. En este caso es porque muchos tweets tienen en su texto **cross body bag**.

Veamos ahora **derailment**. Este keyword refiere a un descarrilamiento. Y vemos que si tomamos un sample de los tweets con este keyword ese es el significado que se le da.

- *After the green line derailment my concern for track that looks like this goes up a bit...*
@cta @CTAFails - id: 3543
- *25 killed 50 injured in Madhya Pradesh twin train derailment* - id: 3531
- *@AlvinNelson07 A train isn't made to withstand collisions! Immediate derailment. It's totally fucked.* - id: 3528

Todos los tweets que tienen este keyword refieren a un descarrilamiento. Se puede concluir que este keyword es más técnico y que en caso de no tener el target de los tweets, se lo podría buscar en el texto de cada tweet para obtener con alta probabilidad potenciales catastróficos. **Wreckage** hace referencia a una destrucción.

- *MH370 victim's family furious the media was told about wreckage confirmation first* - id: 10739
- *Wreckage is MH370: Najib #MH370 #najibrazak #MalaysiaAirlines* - id: 10768
- *Wreckage 'conclusively confirmed' as from MH370: Malaysia PM* - id: 10774
- *The first piece of wreckage from the first-ever lost Boeing 777 which vanished back in early March along with the 239 people on board has...* - id: 10763

Si nos fijamos en este sample, vemos que todos los tweets hacen referencia a la destrucción de un avión, y varios al vuelo MH370 de Malaysian Airlines que desapareció en 2014. Analizando un poco más los tweets, vimos que todos los tweets que tienen como keyword 'wreckage' refieren a este vuelo, incluso el último tweet del sample que si bien no tiene el string 'MH370' en su texto lo referencia mediante el avión, la fecha y la cantidad de gente desaparecida, que coinciden con las del vuelo en cuestión. De esta manera podemos decir que cualquier tweet que tenga en su texto el string 'MH370' hace referencia directa al vuelo y por ende es una catástrofe. Para confirmar esto filtramos todos los tweets y encontramos que el 100% de estos tienen marcado el target como catástrofe. Por último, **debris**. Se refiere a escombros o ruinas. Tomamos un sample.

- *Discovered Plane Debris Is From Missing Malaysia Airlines Flight 370 | TIME* - id: 3107

- *MH370: debris found on reunion island. ?? #sad #tragedy #innocent #crash #mh370 - id: 3122*
- *Experts leave lab as Malaysia confirms debris is from #MH370 - id: 3136*
- *Plane debris is from missing MH370 - id: 3145*

De nuevo vemos mucha referencia al MH370. Haciendo el mismo filtro que con 'wreckage', encontramos que todos estos tweets hacen referencia al vuelo.

Como conclusión, sabemos que en este set en particular, las palabras que aparecen únicamente en tweets verdaderos se utilizan en su sentido literal. Sabemos que si en un tweet encontramos el keyword derailment se tratará de un descarrilamiento de tren, o si encontramos el keyword debris se tratará entonces de los restos de un avión.

3.5.6. Keywords en proporción similar en tweets verdaderos y falsos

Ahora vamos a ver que keywords están en una proporción similar entre tweets catastróficos y falsos. Cuando nos referimos a similar, decimos que un keyword tendrá que tener un porcentaje de catástrofe mayor a 45 % y menor a 55 %, buscamos el cuantil 45 y el cuantil 55 (esto es, aparece entre un 45 % y un 55 % en tweets catastróficos). Esto lo hacemos para ver qué keywords no impactan tanto sobre el target de los tweets.

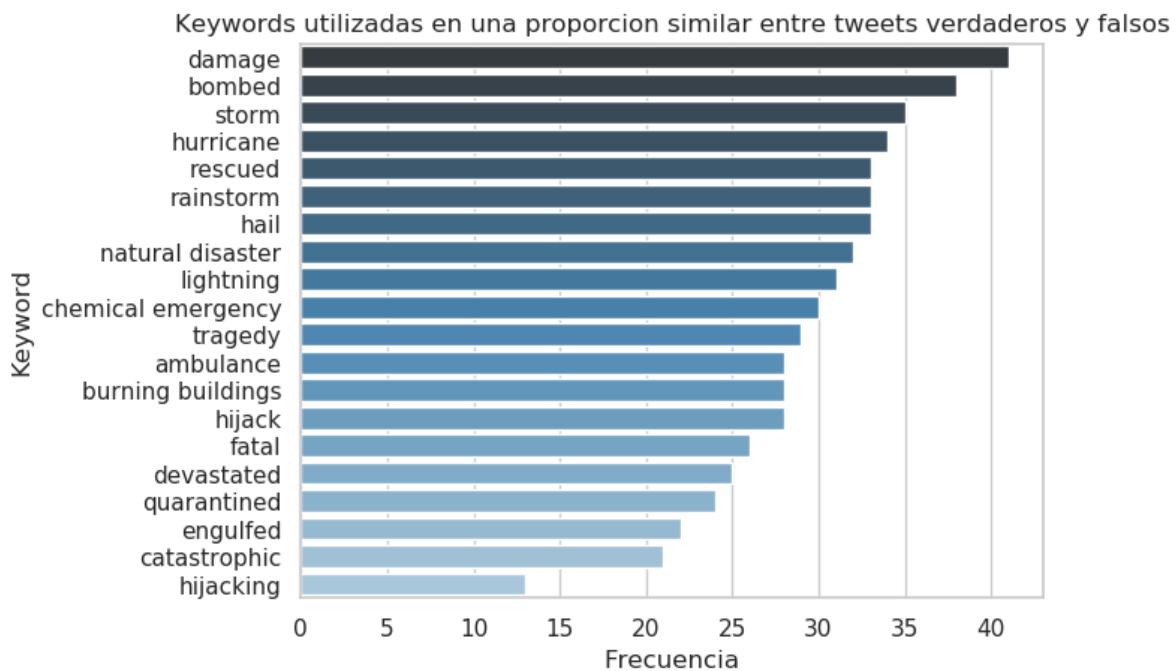


Figura 3.19

Si en algún momento filtramos los tweets por keywords, sin tener el target, se espera que, agrupándolos, 50 % sean catastróficos y el resto sean falsos.

3.5.7. Cantidad de palabras

Algunos keywords se componen de múltiples palabras. Queremos ver que relación podemos encontrar con la cantidad de palabras del keyword y el target. Para eso agrupamos las keywords que tienen una sola palabra y calculamos que tan frecuentemente aparecen en ambos casos.

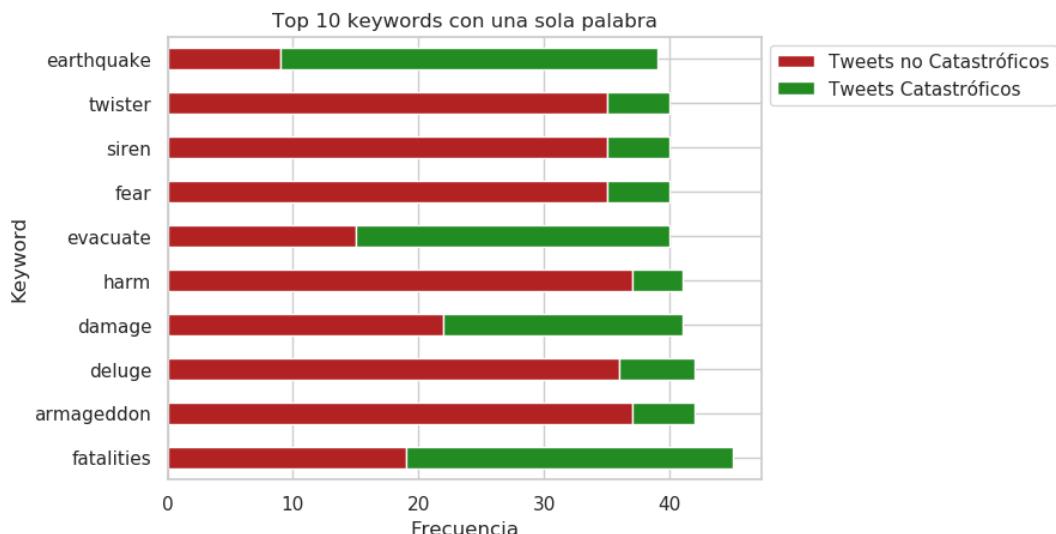


Figura 3.20

Veamos que porcentaje de estos son verdaderos y cuales son falsos.

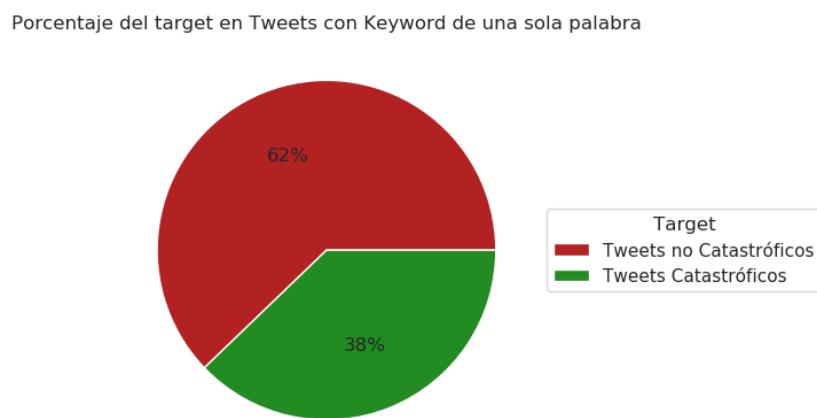


Figura 3.21

Notamos que hay un gran porcentaje de tweets con keyword de una sola palabra que son falsos. En el primer gráfico vemos que hay keywords muy generales, **siren**, **fear**, **harm**. Fuera de contexto, estos keywords no nos dicen mucho sobre el tweet, no como es el caso de **derailment**. Ahora vemos si los keywords con dos palabras son más frecuentes en tweets catastróficos.

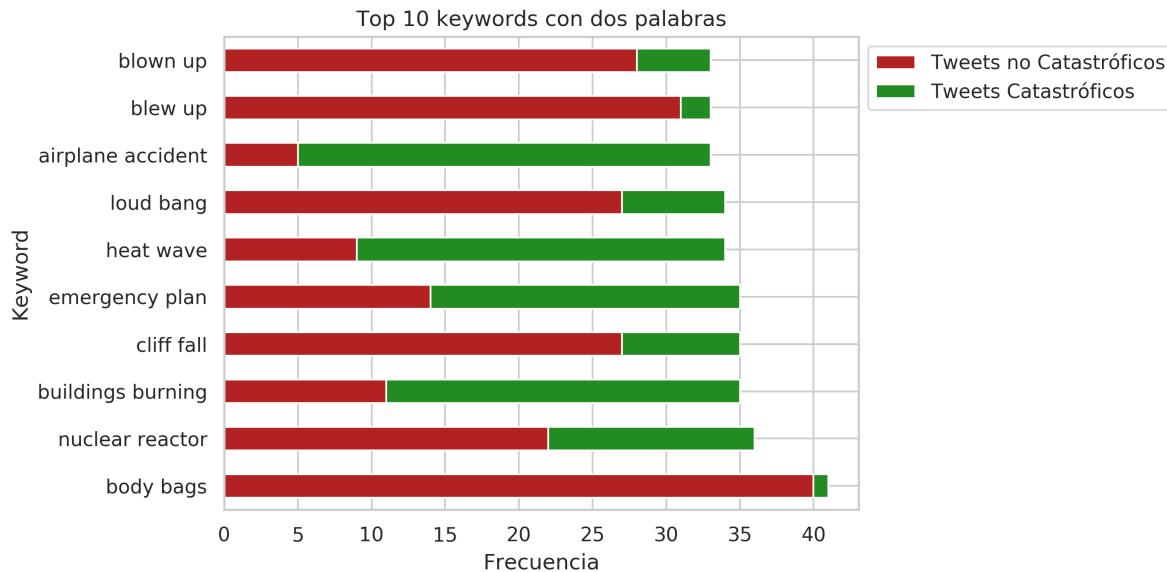


Figura 3.22

Comparando con los keywords de una palabra, vemos que hay mayor frecuencia de tweets verdaderos, y hay keywords más específicos también, como **airplane accident**, **heat wave**, **buildings burning**. Veamos ahora como se distribuye el target.

Porcentaje del target en Tweets con Keyword de dos palabras

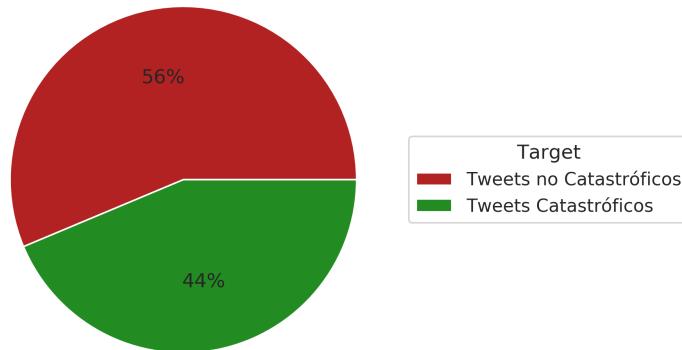


Figura 3.23

Si bien la mayoría de los tweets son falsos, vemos que hay un porcentaje mayor de tweets que son verdaderos si comparamos con los keywords que tienen solo una palabra, esto puede deberse a que los keywords con dos palabras son más descriptivos sobre el texto del tweet que los que solo tienen una. No hicimos el análisis con más de dos palabras ya que hay solo un keyword en esa categoría.

3.5.8. Relación con la longitud de los tweets

Por último, queremos ver si hay relación entre las longitudes de los tweets, los keywords y el target. Para este análisis, agrupamos a todos los keywords, calculamos el promedio de longitud de los textos por keyword, y además el porcentaje de catástrofe de cada uno (esto es, el porcentaje de tweets con ese keyword que son catastróficos).

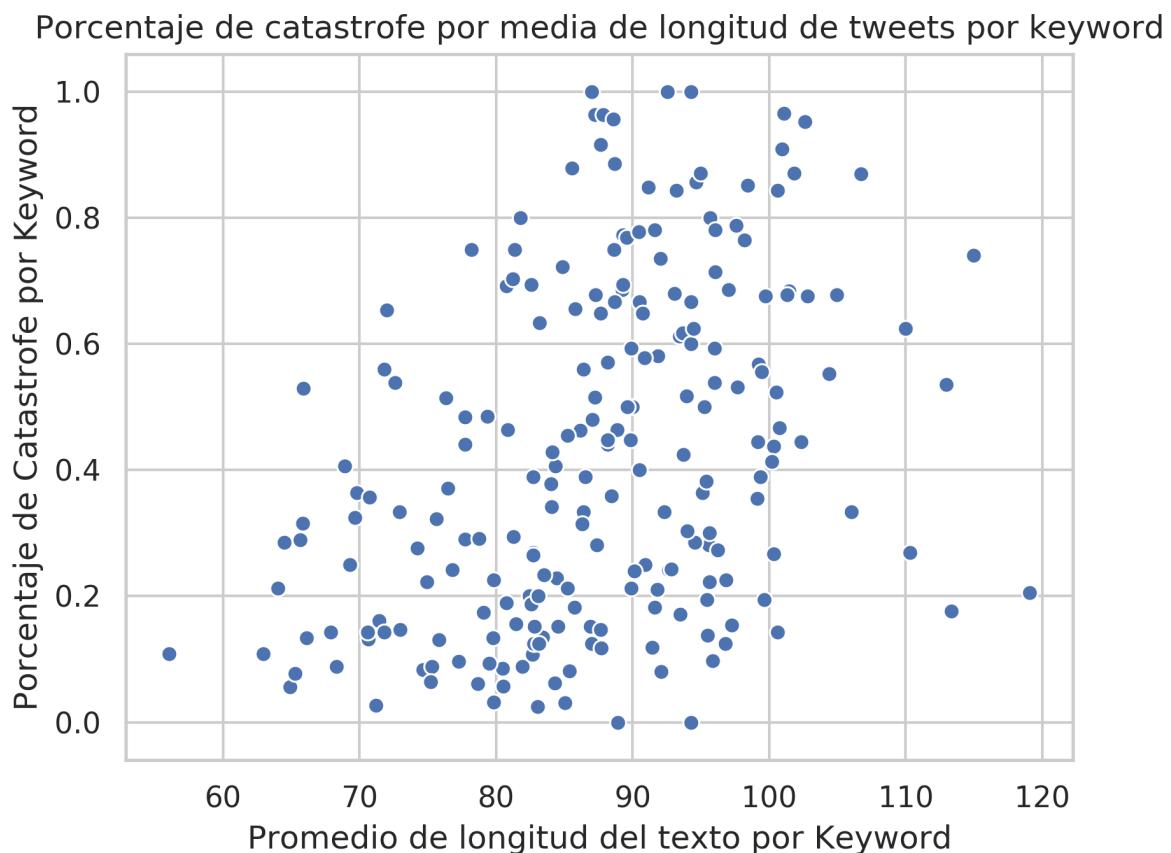


Figura 3.24

Lo que podemos observar es que aquellos keywords para los cuales el promedio de longitud del texto es menor a 90, tienen un porcentaje bajo de catástrofe. Mientras menor es el promedio, menor es el porcentaje de catástrofe. Entre 90 y 100, se puede observar bastante dispersión. Sin embargo, la tendencia del porcentaje de catástrofe aumenta (se puede observar como se desplaza hacia los cuantiles superiores a medida que nos desplazamos por el eje x). Como fue el análisis con la longitud de los tweets y el target, vemos que agregándole la restricción de agruparlos por keyword nos da una idea de cuales son aquellos keywords que tienen menor probabilidad de estar asociados a tweets verdaderos (que son aquellos que tienen un promedio de longitud menor).

3.6. Análisis de las Locations

En esta ultima sección se busca analizar si la ubicación de los tweets tiene relación alguna con la veracidad de los mismos.

Como bien se ve en la Figura 3.25, hay que tener en cuenta que de **7613** locations de tweets, **2533** (33.3 %) son Nan y **5080** (66.7%) contienen algún tipo de dato, no necesariamente es una ubicación, si no que puede ser basura.

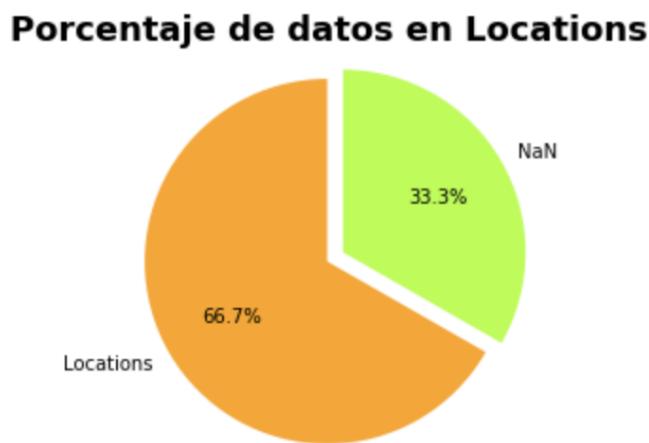


Figura 3.25

Ahora como vemos en la Figura 3.26, de las **5080** (66.7 % de la Figura 3.25) locations de tweets que contienen algún tipo de dato, solo **2954** contienen ubicaciones reales, el resto, que son **2533** (46.2 %) contienen basura, es decir, frases o ubicaciones que no aportan información ya que no existen o no podemos analizar.

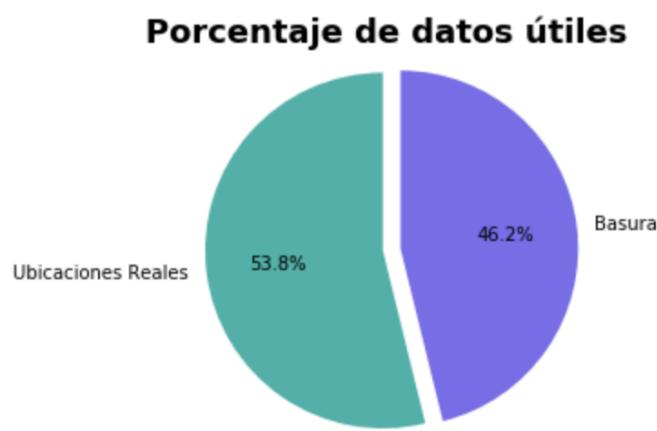


Figura 3.26

Entonces los datos que tenemos disponibles son muy acotados. A demás, dentro de estos pocos datos que tenemos hay un gran porcentaje referido solo a Estados Unidos, y esta es la

razón por la cual decidimos hacerle una sección especial a sus estados. A continuación, en las siguientes secciones vamos a analizar la distribución de los tweets verdaderos y falsos basándonos en distintas localidades. En esta oportunidad, unas de las visualizaciones utilizadas es el 'Tree Map' donde los el tamaño de los rectángulos representa su frecuencia, rectángulos más grandes aparecen más veces.

3.6.1. Detalles del análisis

Debido a que muchos de los datos contenidos en esta columna hacían referencia a la misma ubicación pero estaban expresados de distintas maneras, decidimos limpiar e organizar estos mismos para poder analizar de una manera mas eficiente y rápida las ubicaciones en los tweets. Los procesos involucrados son:

- Reagrupamos las ubicaciones mediante la transformación de la columna 'location' a mayúsculas. Esto une a las cadenas que varían en su forma de escritura pero hacen referencia a la misma ubicación. Un ejemplo es: 'United States' con 'united states' ya que ambas pasan a ser 'UNITED STATES'.
- Importamos dos sets de datos externos para identificar el país, la ciudad o el estado de la columna 'location'.

1) El primer set de datos es 'Countries.csv', lo obtuvimos de <https://gist.github.com/tadast/8827699>. La estructura de este set la podemos observar en la Figura 3.27 :

	Country	code2	code3	Numeric code	Latitude (average)	Longitude (average)
0	AFGHANISTAN	AF	AFG	4	33.00	65.00
1	ALBANIA	AL	ALB	8	41.00	20.00
2	ALGERIA	DZ	DZA	12	28.00	3.00
3	AMERICAN SAMOA	AS	ASM	16	-14.33	-170.00
4	ANDORRA	AD	AND	20	42.50	1.60

Figura 3.27

Estos datos nos permiten reconocer los países que están escritos con alguna abreviatura de dos o tres letras. Un ejemplo podría ser el de 'USA' o 'US' que hacen referencia a 'United States'.

2) El segundo set de datos, contiene a los estados de Estados Unidos, es 'States.csv', lo obtuvimos de <https://github.com/jasonong/List-of-US-States/blob/master/states.csv>. La estructura de este segundo set la podemos observar en la Figura 3.28 :

	State	Abbreviation
0	ALABAMA	AL
1	ALASKA	AK
2	ARIZONA	AZ
3	ARKANSAS	AR
4	CALIFORNIA	CA

Figura 3.28

- Instalamos una librería de Python llamada 'GeoText' <https://github.com/elyase/geotext>. Esta nos permite saber la ubicación del tweet mediante el texto. Un ejemplo de su uso:

```
from geotext import GeoText  
  
lugar = GeoText('London is a great city')  
lugar.cities  
# 'London'
```

Al finalizar esta limpieza de datos, obtuvimos un nuevo set de datos, mucho mas completo, llamado 'TweetsLocations.csv'. Este contiene nuevas columnas: 'country', 'city' y 'state'. Esto nos fue de ayuda a la hora de hacer las visualizaciones.

3.6.2. Análisis por país

Nos interesa tener una vista general de los países más frecuentes. Para empezar en la Figura 3.29 podemos observar los países con mas tweets sin tener en cuenta el target .



Figura 3.29

Podemos destacar que Estados Unidos predomina, es decir, es el país que mas tweets tiene, hay una gran diferencia al resto de los países.

Ahora que ya vimos un panorama general de las ubicaciones más usadas en los tweets, vamos a separar las ubicaciones en los tweets que son verdaderos y los que son falsos y ver cual es la proporción en esos casos.



Figura 3.30



Figura 3.31

A simple vista, en la Figuras 3.30 y 3.31, se puede notar que las proporciones de las dos visualizaciones son muy parecidas. No presenta muchas variaciones en cuanto al área (frecuencia) para cada país. Para las frecuencias mas altas, como en el caso de Estados Unidos, el área en ambos casos se mantienen igual. Hay cambios en los países con frecuencia mas baja, incluso vemos la aparición de nuevos países.

Ahora vamos a ver la veracidad de los tweets para cada uno de estos países mas mencionados.

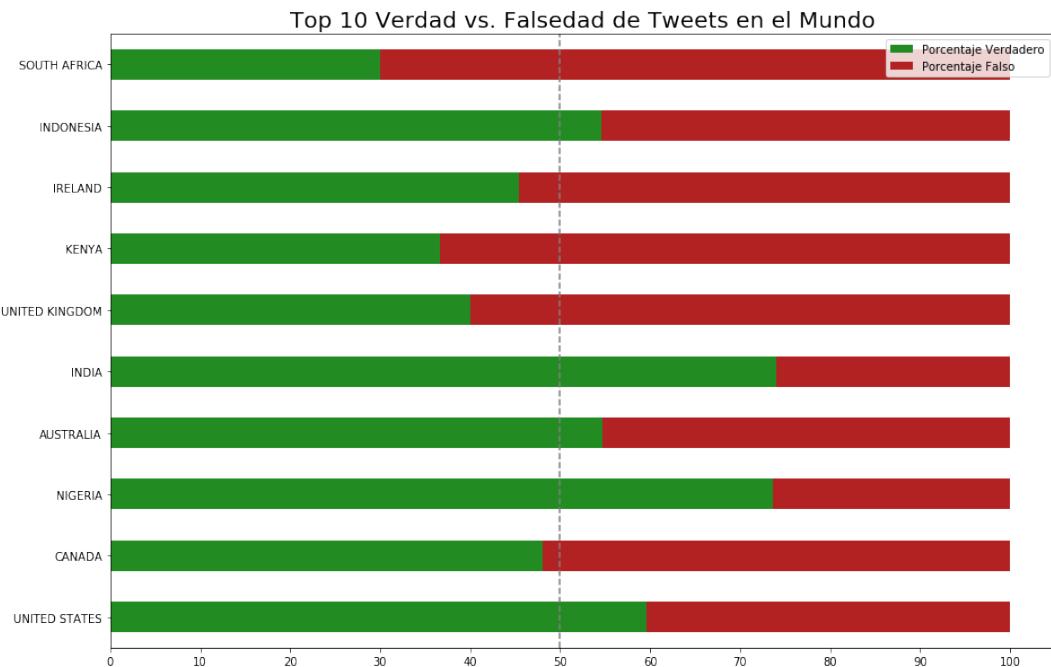


Figura 3.32

En la Figura 3.32 podemos observar los porcentajes de tweets verdaderos y falsos en los diez países con más tweets. La idea de analizar los porcentajes, es útil para poder tener una idea mas detallada de que tan verdadero o no puede ser el tweet en cada país. La linea

punteada sobre el gráfico indica el 50 %. Parádonos sobre ella no hay un patrón específico, pero si podemos ver que en 'Nigeria' e 'India' tienen un porcentaje mayor al resto, con alta probabilidad de que el tweet sea verdadero. Por el otro lado, 'Sud África' y 'Kenya' muestran lo contrario, tienen muy poco porcentaje de tweets verdaderos, por lo tanto, hay pocas probabilidades de que el tweet sea verdadero.

3.6.3. Análisis por ciudades

En esta sección, vamos a poder analizar las ubicaciones en los tweets tanto verdaderos como falsos, pero en esta ocasión por ciudades.



Figura 3.33



Figura 3.34

Acá, en las Figuras 3.33 y 3.34 si podemos notar una diferencia. Vemos que la ciudad de 'Washington' pasa de estar en tercer puesto a sexto, una de las diferencias más notables entre las dos representaciones. También tenemos para destacar que en los tweets falsos, sin contar London, son todos estados de Estados Unidos. En cambio, en los tweets verdaderos, se encuentran ciudades tales como 'Mumbai', 'Toronto', 'Calgary' y 'Melbourne' que son de distintos países.

Nuevamente Estados Unidos se hace destacar, no solo porque aparecen todas sus ciudades, sino que también estas aparecen con gran cantidad de tweets.

Veamos que sucede con la veracidad en los tweets de estas ciudades

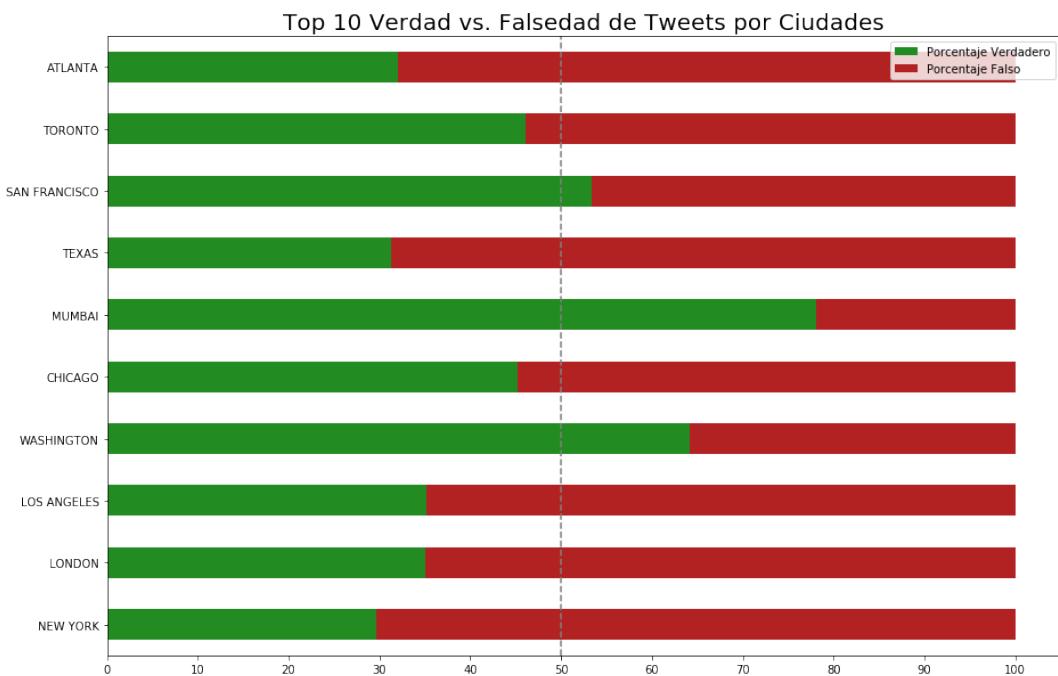


Figura 3.35

Si nos paramos sobre la linea del 50% en la Figura 3.35 vemos que la mayoría de las ciudades tienen mas tweets falsos que verdaderos. Podríamos pensar que por ser ciudades con alta población y concurrida por una mayor cantidad de turistas esto sea normal, ya que, capaz los usuarios de Twitter solo quieran mencionar que se encuentran en esta ciudad o hagan una referencia a algo turístico. Pero esta hipótesis la podemos negar viendo que en 'Washington', una de la ciudades mas conocidas de Estados Unidos, hay un alto porcentaje de tweets verdaderos. Lo mismo podemos decir de 'Mumbai' y 'San Francisco'. Esto nos deja, nuevamente, sin salida porque no podemos determinar algún tipo de patrón para predecir la veracidad de los tweets.

3.6.4. Análisis por estados de Estados Unidos

Por ultimo en esta sección, vamos a analizar el caso particular de los estados de Estados Unidos como bien mencionamos previamente.

Top 10 - Estados de 'USA' con mas tweets verdaderos

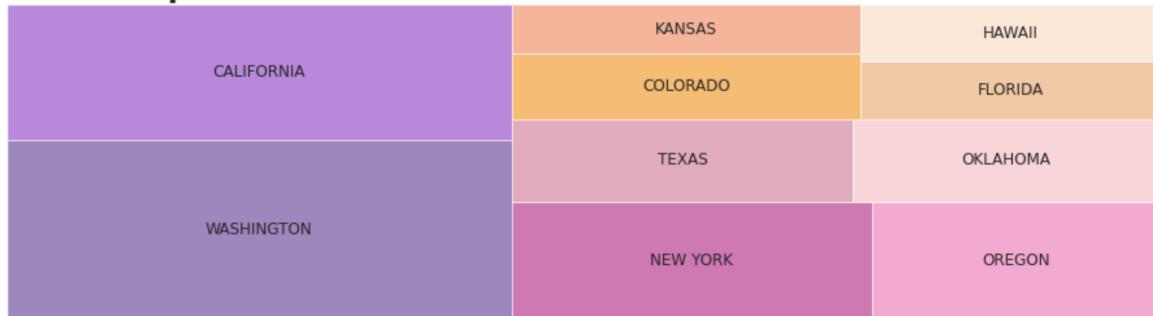


Figura 3.36

Top 10 - Estados de 'USA' con mas tweets falsos

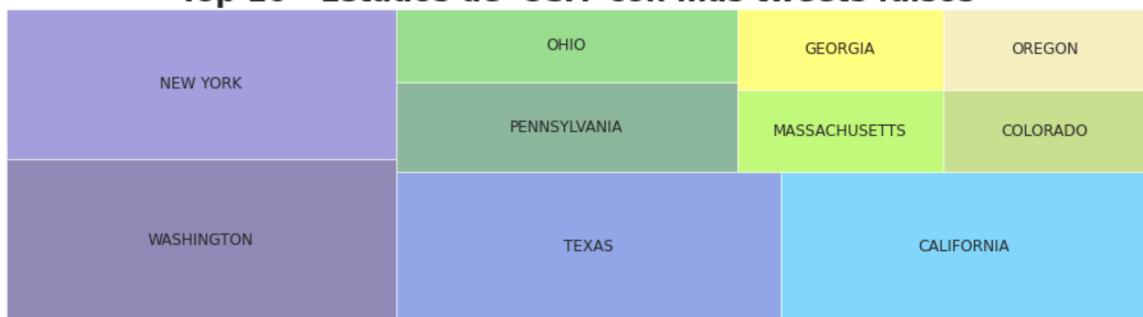


Figura 3.37

Nuevamente podemos ver en las Figuras 3.36 y 3.37 las proporciones de las dos visualizaciones son muy parecidas. Para las frecuencias mas grandes encontramos los estados mas poblados y conocidos tal como 'New York', 'Washington' y 'California' se mantienen con gran frecuencia en ambos targets.

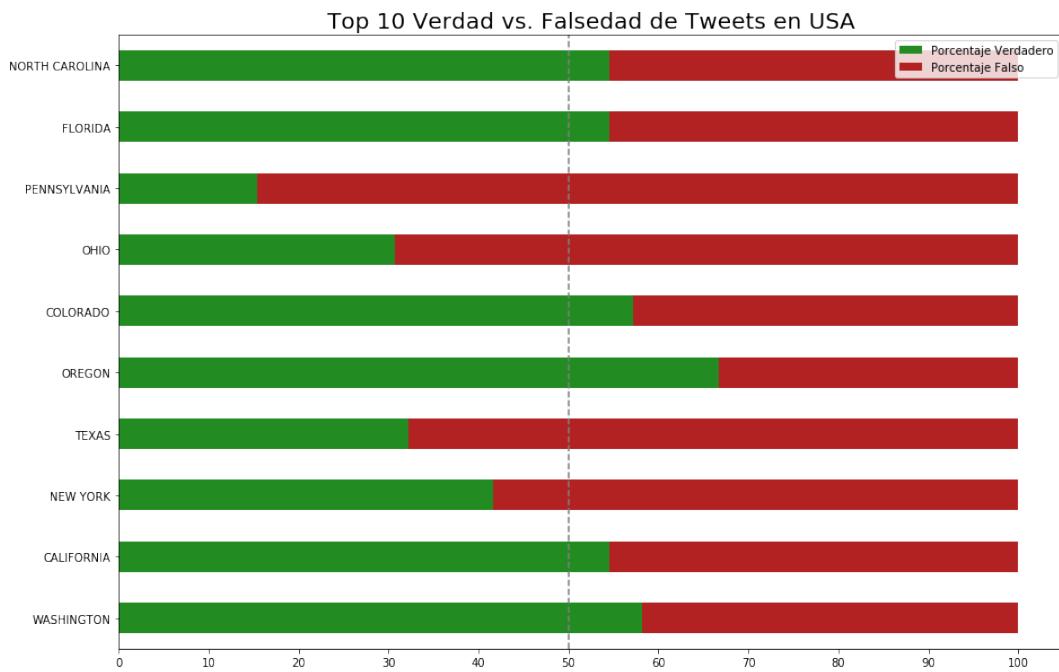


Figura 3.38

En la Figura 3.38 se puede observar los porcentajes de Tweets verdaderos y falsos en los diez estados de EE.UU con mas Tweets. Nuevamente si nos paramos sobre la linea del 50%, la distribucion se mantiene estable y vemos que no hay ninguna tendencia. Los casos mas llamativos son los de 'Pennsylvania', 'Texas' y 'Ohio' porque tienen bajo porcentaje de tweets verdaderos.

Frente a esto podemos concluir que un tweet que se menciona dentro de los Estados Unidos, tenemos la posibilidad de poder llegar a una conclusión de su veracidad. De todas maneras las diferencias entre las distribuciones no son tan grandes y no siguen un patrón estable. En base a esto no creemos que la location de los tweets juegue un rol importante en determinar la veracidad del mismo.

4. Conclusiones

4.1. Insights

El objetivo de este informe era familiarizarnos con el set de datos y al final poder identificar propiedades que nos ayuden a predecir si un tweet está hablando de un desastre real o uno falso. El set de datos parecía pequeño con solo 4 columnas y suponíamos que no íbamos a poder extraer mucha información pero luego de analizar todos los distintos parámetros podemos decir que estábamos equivocados. Logramos encontrar muchos casos donde había distinciones claras entre los tweets que eran auténticos y los que no. Por otro lado, también había casos donde la diferencia no era significativa y podíamos concluir que no debíamos tenerlos tanto en cuenta a la hora de predecir.

- A la hora de analizar las longitudes de los textos encontramos que los tweets verdaderos estaban más concentrados en el rango de los 130-150 caracteres mientras que en el caso de los tweets falsos estaban distribuidos de manera más uniforme con un pequeño pico en el mismo rango. Solo con esta predicción no podríamos decir si un tweet anuncia un desastre o no. Sin embargo, si tenemos en cuenta los otros factores, podría ser de ayuda para terminar de determinar sin un tweet anuncia un desastre o no.
- La separación entre las palabras más utilizadas para los tweets falsos y verdaderos está bien definida. Encontramos palabras relacionadas a desastres y muertes de un lado mientras que del otro no hay un tema común que las une.
- Observando el uso de hashtags y menciones dentro de los tweets encontramos que en los desastres verídicos se referían a organizaciones gubernamentales, medios de comunicación y desastres específicos.
- En cuanto a las keywords, lo que se pudo ver es que a medida que encontramos un keyword en más tweets verdaderos, menos probable será encontrarlo en tweets falsos y viceversa, lo que nos dice que hay una segmentación. Aquellas keywords que aparecen más frecuentemente en tweets verdaderos son más descriptivas sobre estos, por el hecho de tener más de una palabra y/o por usar palabras técnicas. Además encontramos keywords referentes a un sucesos particulares como el vuelo MH370, lo que nos permitió tener un filtro más sobre el resto de los tweets para poder encontrar aquellos que sean catastróficos sin necesidad del target.
- A la hora de hablar de las locations de los tweets, si bien la mayoría de los tweets se encuentran dentro de los distintos estados de Estados Unidos, no hay un patrón definido para determinar si un desastre ocurrió o no. Creemos que este parámetro no es útil a la hora de determinar la veracidad de los mismos.

Finalmente teniendo en cuenta todos estos ítems podremos determinar, con cierto margen de error, si un tweet desconocido habla de un desastre que verdaderamente ocurrió o no. Esto nos será de gran ayuda ya que en el próximo trabajo práctico debemos crear un programa que pueda hacer exactamente eso.

5. Código

En esta sección dejamos un acceso directo a nuestro Google Colab donde van a poder ver el código que utilizamos para poder obtener todas las visualizaciones en el presente informe.
<https://drive.google.com/drive/folders/1NwSCcLG5qkRq8CxUnI4dPhdd55ss77MT?usp=sharing>