Trabajo Práctico 1 - Análisis Exploratorio de Datos

75.06 95.58 - Organización de Datos

Primer Cuatrimestre de 2020

Nombre Apellido	Padrón	mail
Hamma AALI CHTOUKI	106607 - Regular, Intercambio	haali.ext@fi.uba.ar
Aymeric COUSAERT	105464 - Regular, Intercambio	acousaert.ext@fi.uba.ar
Andrés VERA	Oyente	andru.vra@gmail.com
Mariana VINYOLAS	Oyente	marianavinyolas@gmail.com

Código : `https://colab.research.google.com/drive/14dN9iBhw2uE44mrPFiQcBQoCOk6-Phfs?usp=sharing`

Documentos : `https://github.com/aymericcousaert/analisis-tweets`

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Índice

1. Introducción

En el presente informe analizamos un set de datos sobre un conjunto de tweets que han sido publicados para informar en tiempo real situaciones de emergencia o desastres, y si éstos resultaron verdaderos o falsos. El mismo se encuentra en la dirección `https://www.kaggle.com/c/nlp-getting-started/data`. Analizaremos el impacto de la ubicación del autor del tweet sobre la realidad de éste y veremos en detalle el contenido de los tweets. Podemos imaginar diferentes aplicaciones a este trabajo : estimar el resultado de una elección e imaginar la tendencia para cada provincia u otra unidad de corte de la localización o estimar la opinión pública sobre un producto. En nuestro caso, los tweets están clasificados como verdaderos o falsos.

2. Información general sobre los datos

El set de datos que tenemos contiene 7613 registros distribuidos en 5 columnas. Se observan aproximadamente 60 valores nulos en la columna **keyword** y mas de 2500 en la columna **location**.

En relación a la veracidad o falsedad del tweet, verificamos que el dataset se encuentra balanceado, un 57 % de los tweets fueron clasificados como falsos y un 43 % de los tweets se clasificaron como reales. Nos parece valido aclarar que en la mayoría de los análisis denominaremos a los tweets reales como **target_1** y a los falsos como **target_0**.

Cada registro posee además un numero que lo identifica como tweet, y que no esta relacionado con el autor del mismo. Por lo tanto el dataset contiene 7613 IDs diferentes, uno por cada registro.

```
RangeIndex: 7613 entries, 0 to 7612
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        7613 non-null   int64
 1   keyword   7552 non-null   object
 2   location  5080 non-null   object
 3   text      7613 non-null   object
 4   target    7613 non-null   int64
dtypes: int64(2), object(3)
memory usage: 297.5+ KB
```

Figura 1: Informaciones sobre el dataset

Los atributos son :

- **id** : Un identificador único para cada tweet.
- **text** : El texto del tweet.
- **location** : La ubicación desde donde fue enviado (que podría no estar).
- **keyword** : Un keyword que identifica el tweet (podría faltar).

- **target** : Indica si se trata de un desastre real (1) o no (0).

3. Limpieza de datos

El pre-procesamiento de los datos se realizó sobre las columnas **location** y **text** y tuvo tres objetivos principales:

- Simplificar el texto del tweet para poder aplicar un análisis de NLP sobre el contenido del mismo.
- Crear columnas derivadas **link**, **contacto** y **hashtag** con la información contenida en el texto, que indican si el tweet contiene un contacto y/o hace referencia a un link y/o contiene un hashtag (value 1) o no (value 0).
- Unificar los registros en **location** para un posterior análisis geográfico.

Utilizamos la biblioteca nltk para crear una nueva columna **text_clean** que contiene las palabras del tweet correspondiente, pero además hacemos una lematización de las palabras, es decir, unificamos las palabras con la misma raíz. En este proceso también se eliminaron los plurales.

Al observar en detalle los datos contenidos en **location**, nos dimos cuenta que para realizar un análisis geográfico de los datos necesitábamos encontrar una manera de armonizar estos datos sin tener que hacerlo de forma manual. Este problema de falta de formato que genera esta variedad infinita de localidades ('Nowhere', 'Everywhere', 'World', 'Married with two kids', por citar algunos ejemplos), creemos que se debe a que se toma la información que carga un usuario cuando se registra en Twitter y puede completar ese campo con lo que se le ocurra en ese momento.

Por este motivo decidimos trabajar sobre otro dataset que contiene información de los países, sus ciudades y la población de las mismas, que se puede encontrar en el siguiente link : `https://www.kaggle.com/juanmah/world-cities`. De esta manera, logramos fusionar los dos dataset para estandarizar el nombre de la ciudad y además conservamos el dato del país como otro atributo derivado de **location**.

Un dato importante que debemos resaltar, es que si bien el dataset es de 7600 registros, cuando trabajamos con **location** partimos de solamente 5000 registros porque aproximadamente 2500 son valores nulos. Incorporar la información del país en nuestro dataset, nos permitió realizar un análisis mas intuitivo al reagrupar los tweets por países y trabajar directamente sobre la distribución por país en vez de ciudad.

Si bien en todo este proceso se perdieron algunos registros (en general debido a que existen ciudades con el mismo nombre en diferentes países), consideramos que la ganancia de los insights obtenidos fue mayor que si no hubiéramos preprocesado los datos de esta manera.

4. Análisis geográfico

4.1. Origen geográfico de los tweets

En esta parte nos hacemos las siguientes preguntas :

- De dónde provienen los tweets ?
- De qué países ?
- De qué ciudades ?

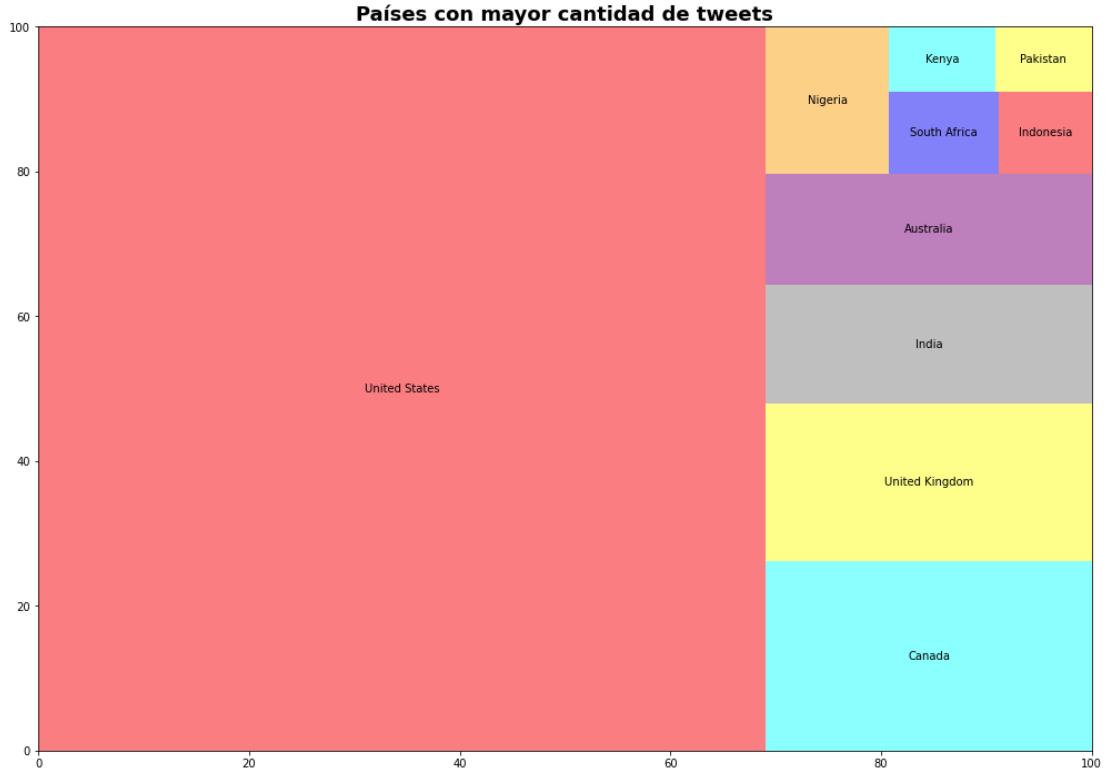

Figura 2: Países del mundo donde están enviado los tweets

La mayoría de los tweets que tienen una location en este dataset provienen de los Estados Unidos. Aproximadamente el 70 % de los tweets provienen de los Estados Unidos. Después siguen con un porcentaje similar (0,3*0,2=6 % del total) Canadá, Reino Unido, India y Australia. Teniendo en cuenta la superioridad numérica de los tweets estadounidenses, queremos lograr una visualización similar excluyendo los Estados Unidos.

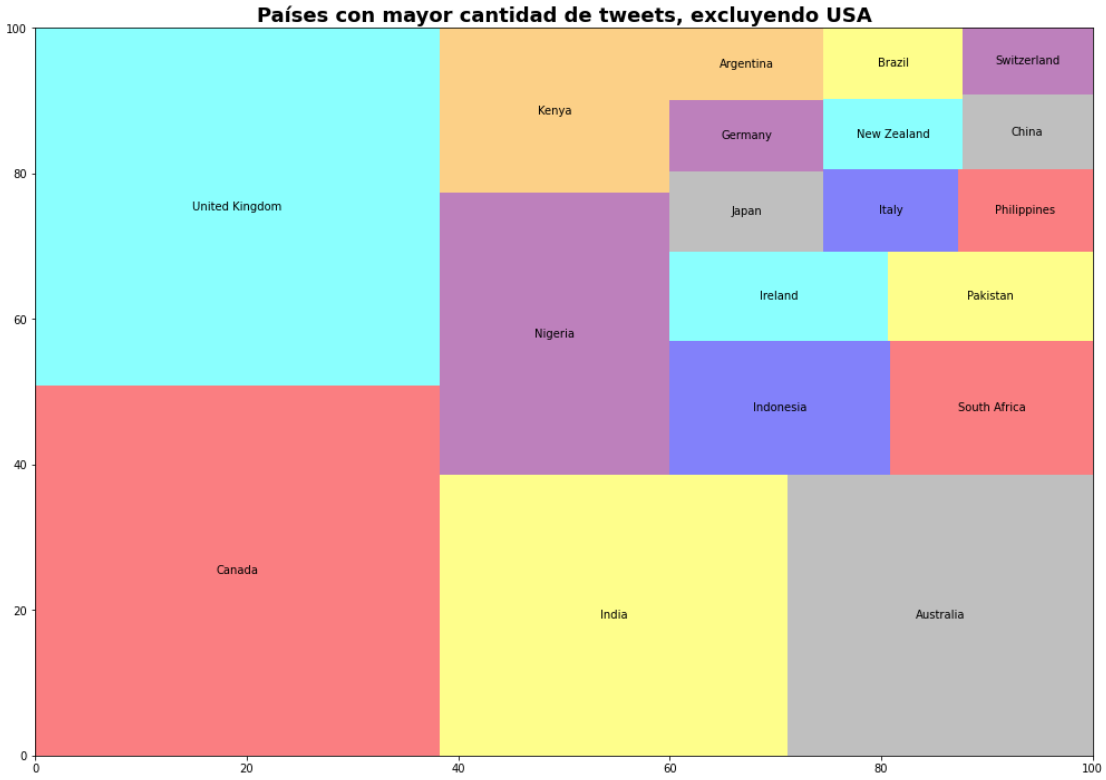

Figura 3: Países del mundo donde se generaron los tweets, excluyendo EE.UU.

Al eliminar EEUU en esta visualización, logramos una mejor representación de cuales son los otros países de donde vienen los tweets. Después de los ya mencionados, también están en orden Nigeria, Kenya, Indonesia, Sudafrica, Irlanda y Pakistan.

Canadá, Reino Unido, India, Australia, Sudafrica e Irlanda son países angloparlantes. Para esta razón, podemos imaginar que el conjunto de tweets es solo en idioma ingles. Para los países que no tienen ingles como idioma, podemos suponer que los tweets fueron enviados por un habitante del país que habla ingles o por un extranjero de viaje.

Por ultimo, aprovechando la gran cantidad de tweets provenientes de EEUU, realizamos un análisis para visualizar cuales son las ciudades estadounidenses con mayor cantidad de tweets.

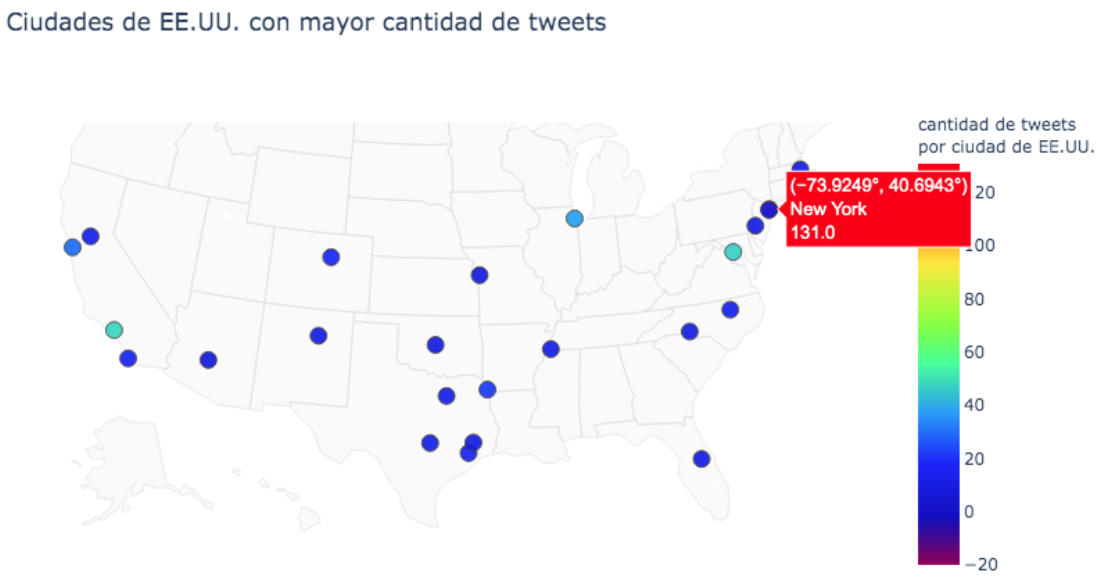

Figura 4: Ciudades de los Estados Unidos donde se han enviado los tweets

Pensamos que la mejor manera de visualizar como se distribuyen los tweets generados dentro de los EEUU es mostrándolos en un mapa y por tal motivo preparamos un gráfico interactivo donde uno puede posicionarse en un punto y el gráfico muestra la cantidad de tweets del punto elegido, como muestra la imagen sobre la ciudad de New York.

Como sabemos que estos tweets se refieren a información en tiempo real de posibles desastres, suponemos que la distribución en el mapa no es aleatoria. Por un lado se observan tweets generados sobre la costa oeste, conocida por sus frecuentes incendios y ocasionales terremotos. Sobre la costa este, se encuentran las ciudades con frecuentes alertas de huracanes y tornados. También observamos que la ciudad de Nueva York posee la mayor cantidad de tweets de este dataset. Suponemos que están relacionados tanto a temas relacionados con accidentes de transito, como a alertas por sospechas de atentados.

Otro ejemplo de visualizar la distribución de los tweets, es graficando la relación del numero de registros respecto a la cantidad de habitantes que posee la ciudad. Como podemos ver en el siguiente gráfico, al realizar esta operación NYC pierde el primer puesto.

Figura 5: Ciudades de los Estados Unidos donde se han enviado los tweets, normalizado por la población

Si bien en esta etapa no estamos evaluando la veracidad de los tweets generados en estas ciudades, podemos concluir que la distribución dentro de los EEUU es como lo esperábamos.

4.2. Influencia de la location sobre el target de un tweet

Como mencionamos anteriormente, los tweets están clasificados según si predijeron o no algún desastre/accidente/catástrofe y queremos relacionar esta clasificación al origen geográfico del mismo.

Por lo tanto, nos hicimos las siguientes preguntas :

- Hay una parte del mundo donde el target es mas frecuente 0, es decir, existen lugares donde en general emiten falsamente alertas de desastres ?

Una observación importante en este punto, es que tuvimos que tener el cuidado de mostrar en la visualización solamente países con porcentaje de tweets falsos mayor al 50 %. En caso contrario, estaríamos representado a los países con mas tweets verdaderos que falsos y no se correspondería con el titulo de esta visualización.

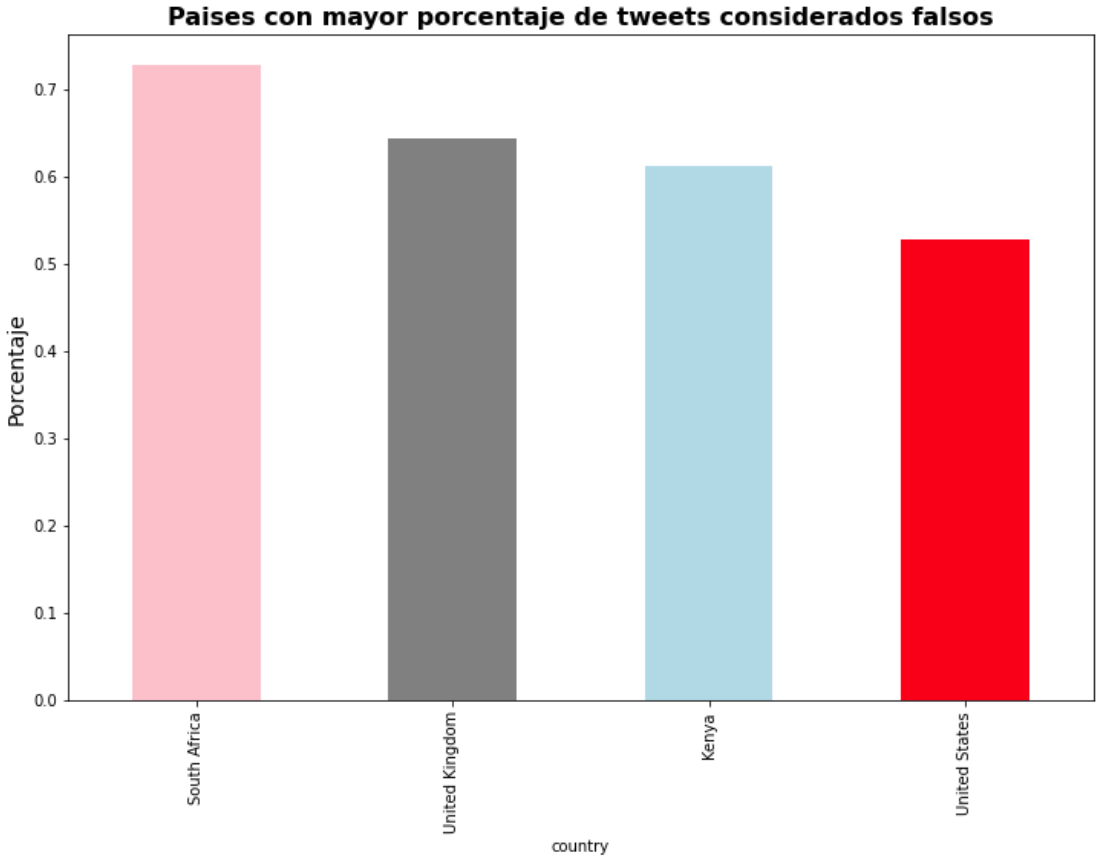

Figura 6: Países del mundo teniendo una mayoría de tweets falsos

En la visualización anterior, se ve que África del Sur es el país que tiene el mas de tweets falsos proporcionalmente (mas de 70 % de los tweets son falsos). Después lo siguen Reino Unido, Kenya y Estados Unidos poseen también mayor proporción de tweets falsos.

Ahora nos hacemos la misma pregunta, pero mirando el caso opuesto. También en este caso, solo consideramos los países con un porcentaje de tweets verdaderos mayor al 50 % para respetar el titulo de la visualización.

- Existen regiones del mundo donde se generen alertas de desastres verdaderos ?

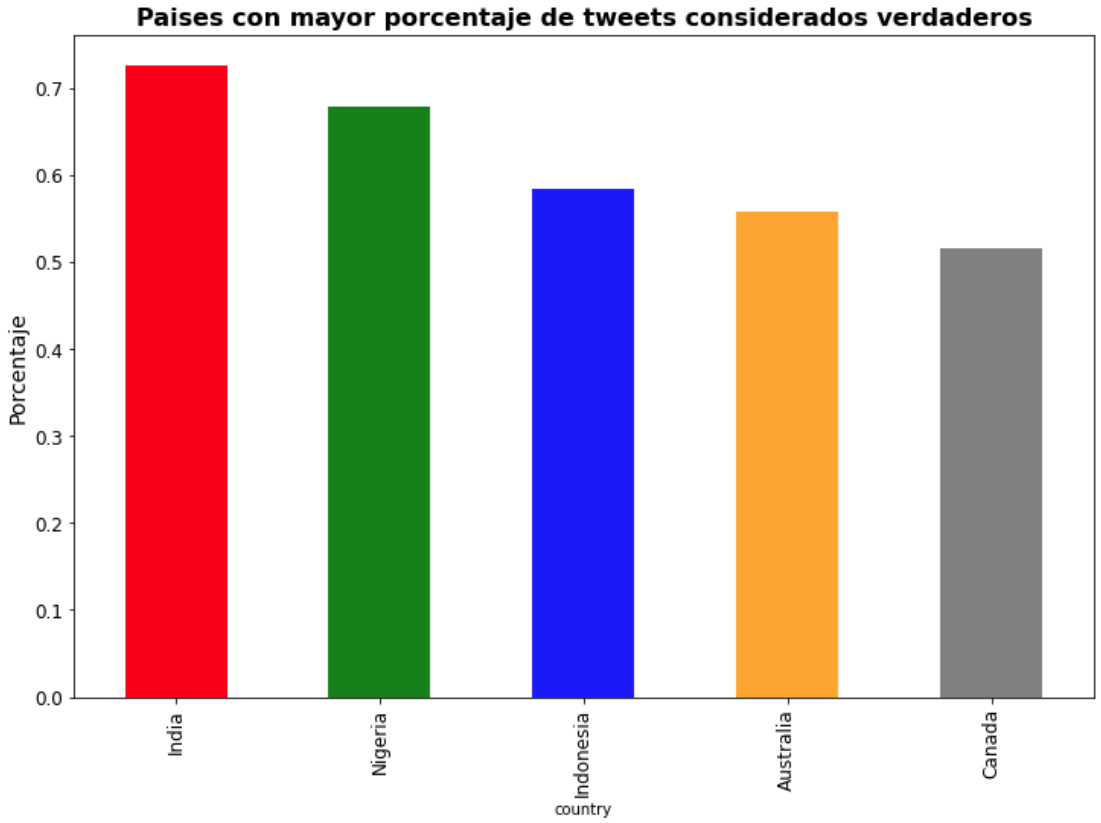

Figura 7: Países del mundo teniendo una mayoría de tweets verdaderos

Con esta visualización, vemos que los países con mayor porcentaje de tweets que resultaron en un desastre real, son India, Nigeria, Indonesia Australia y Canadá.

El análisis de esas dos visualizaciones dice que sorprendentemente, se generan más noticias falsas (Fake news) en los países del primer mundo (United Kingdom, United States) que en los países del tercer mundo (India, Nigeria, Indonesia).

En este análisis, fue necesario filtrar los países del dataset considerando únicamente a los países que tuvieran como mínimo 20 tweets y de esta manera obtener un porcentaje que sea representativo. Si un país tiene solamente 1 tweet, el porcentaje según nuestro análisis seria o bien 0 % o bien 100 % y no reflejaría la tendencia de un país a producir tweets falsos o tweets verdaderos.

4.3. Longitud de los tweets en el mundo

En esta parte nos preguntamos si hay una parte del mundo donde la longitud de los tweets es mas grande en promedio. Consideramos la longitud del tweet calculando la suma de todos sus caracteres. El promedio global de la longitud del tweet resulto ser de 102 caracteres.

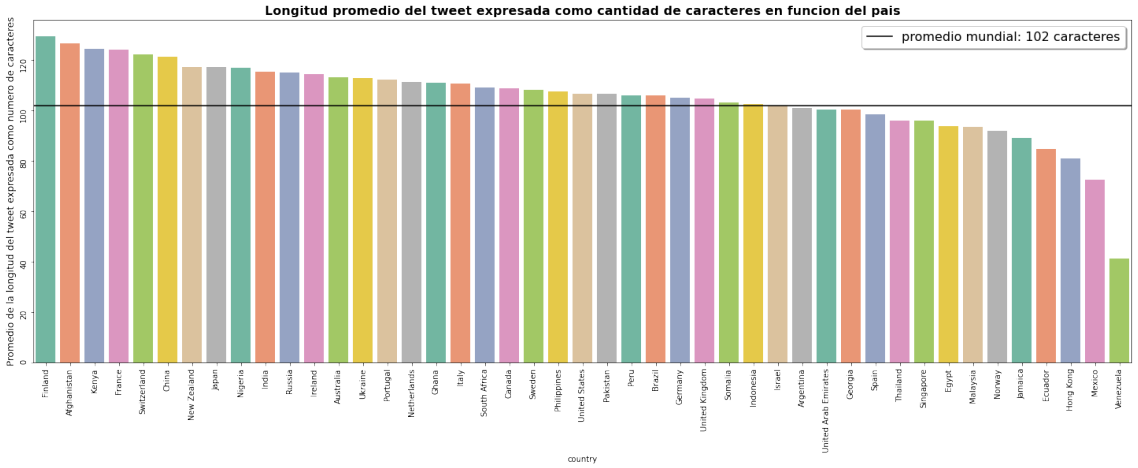

Figura 8: Longitud promedio en los países del mundo

Para esta visualización elegimos tomar en cuenta solamente países que registraron más de 3 tweets, porque de lo contrario no podríamos decir que tenemos un promedio de la longitud de los tweets.

Los países con mayor promedio son Finlandia, Afghanistan, Kenya y Francia, con mas de 120 caracteres en un tweet en promedio. Nos resulta llamativo el caso de Venezuela, que tiene un promedio muy bajo (40 caracteres de promedio en un tweet).

5. Análisis de los tweets

En esta sección de nuestro análisis, trabajamos sobre el contenido del texto del tweet con el objetivo de encontrar algún patrón que los relacione con la veracidad de los mismos. Primero intentamos tener una impresión general del contenido de ciertos indicadores para luego analizarlos en cuanto a su relación con la clasificación del tweet. Decidimos entonces identificar el contenido de hashtags (#), la presencia de notificación a otros contactos (@), si el tweet incluye enlaces externos a través de un link, y por ultimo el contenido de caracteres numéricos.

Luego, con los datos de estos análisis intentamos encontrar una relación con la clasificación del tweet como real o falso.

Por ultimo realizamos un pequeño análisis de las palabras mas frecuentes, utilizando el texto que obtuvimos en el pre-procesamiento de los datos.

5.1. Frecuencia de aparición de los keywords

En esta parte clasificamos cuales son los keywords que aparecen con mayor frecuencia al analizar los tweets verdaderos y luego analizando los falsos.

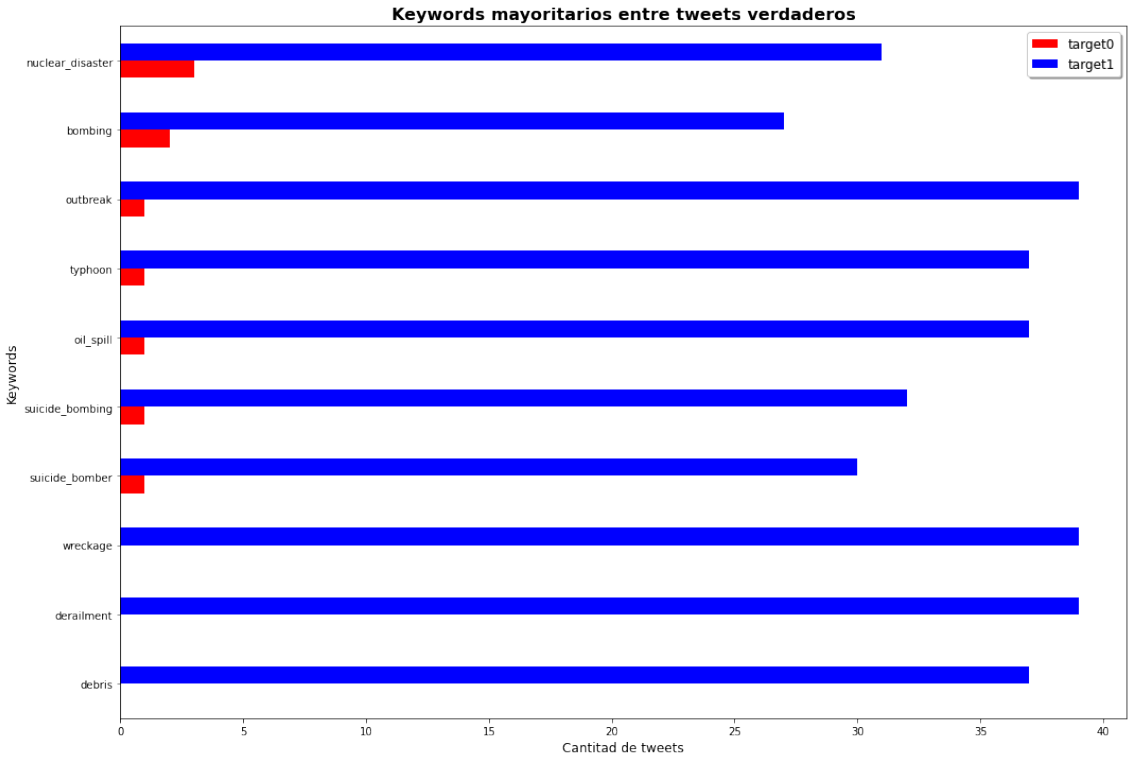

Figura 9: Keywords mas frecuentes entre los tweets verdaderos

Vemos que wreckage, derailment y debris son exclusivamente keywords de tweets con target 1. Hacen referencia a un estado de una situación después de un accidente.

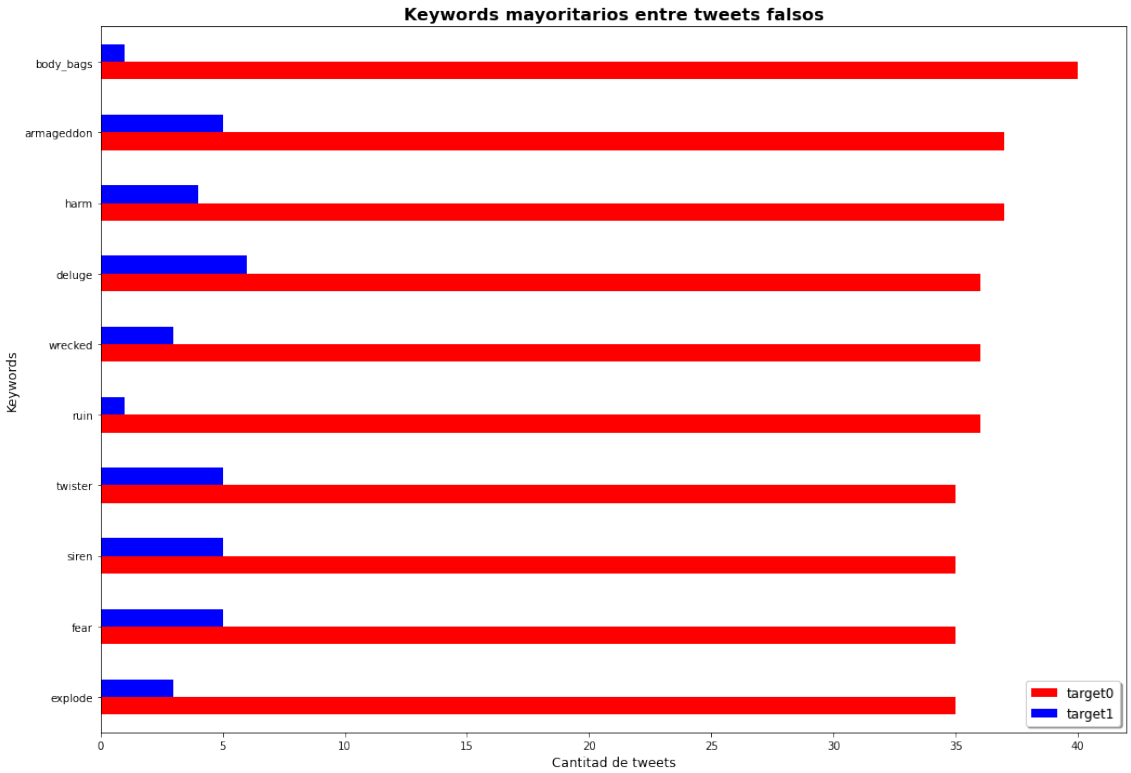

Figura 10: Keywords mas frecuentes entre los tweets falsos

Con los tweets falsos, vemos que las palabras armageddon, body-bags, harm y deluge son keywords que aparecen con mayor frecuencia en los tweets de target 0. Sabiendo que armageddon es una película, es esperable tener una mayor frecuencia de aparición en los tweets de target 0. Body-bags es un keyword que se relaciona con un numero de muertos, cantidad de bajas por decirlo de otra manera, pero no es frecuente encontrarla en tweets reales.

También tenemos keywords que se pueden utilizar en diferentes contextos. Ruin puede hacer referencia a una demolición de un edificio publico para el cual se puede decir que el tweet es verdadero o algo mas personal para el cual no se puede chequear la veracidad del tweet. Fear también puede hacer referencia a algo publico o a algo mas personal.

5.2. Frecuencia de los hashtag

En esta parte nos hacemos las siguientes preguntas :

- Cuales son los hashtags mas frecuentes en los tweets verdaderos ?
- Cuales son los hashtags mas frecuentes en los tweets falsos ?

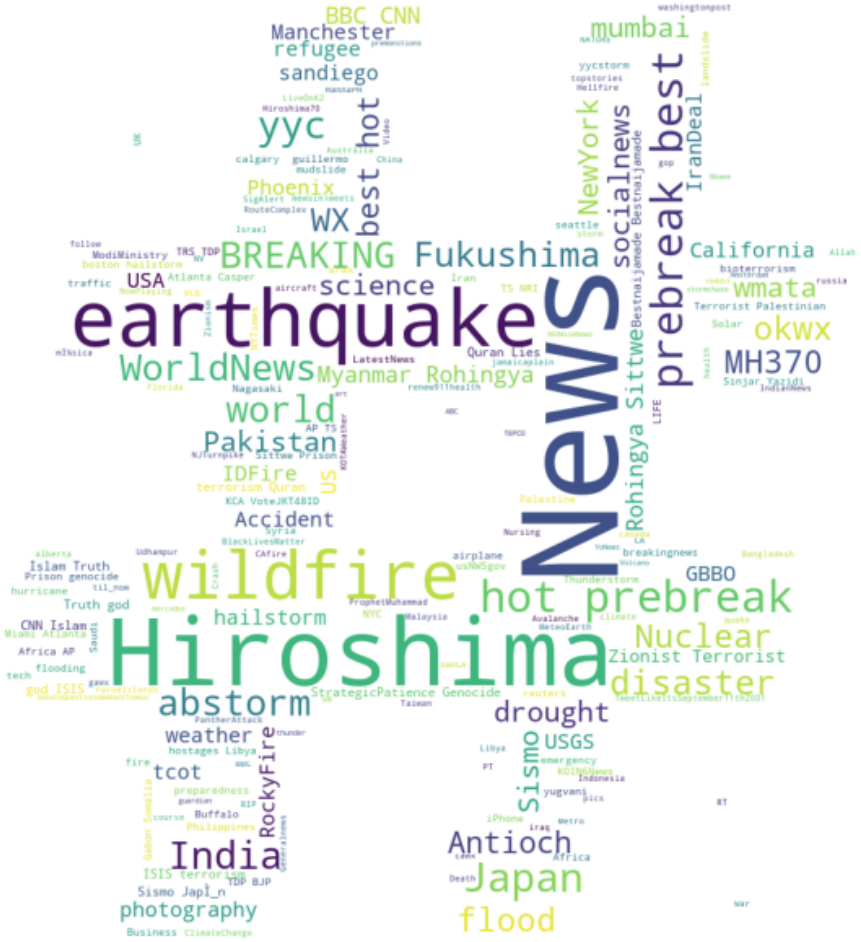

Figura 11: Hashtags mas frecuentes entre los tweets verdaderos

Podemos ver que los hashtags mas frecuentes son News, Hiroshima y Earthquake. Hiroshima era un hahstag esperable porque la cuidad es conocida por la destrucción atómica de la que fue víctima. News y Earthquake no hacen referencia a un cierto evento, pero news describe el evento y earthquake es un tipo de desastre. Entre las otras palabras que describen eventos que realmente tuvieron lugar, podemos notar Fukushima (accidente nuclear), MH370 (desaparición de un avión civil en vuelo de la Malaysia Airlines), Rohingya (Éxodo de los musulmanes rohingya para escapar de la campaña militar a gran escala de limpieza étnica), Irandeal (Iran nuclear deal).

También realizamos el mismo análisis sobre los tweets clasificados como falsos.

Figura 12: Hashtags mas frecuentes entre los tweets falsos

Se ve que los hashtags mas frecuentes son NowPlaying, news y Job. NowPlaying es un hashtag famoso de twitter par compartir que musica o que película se escucha al momento. Se ve también Hiring, Soundclound, military, islam que no tienen relación cierta con eventos. Pero en forma general, como era de esperar, no sobresale ningún hashtag que nos haga relacionar con algún desastre o amenaza.

5.3. Influencia de la presencia de números sobre el target

Los números pueden tener una relación con los detalles dados sobre una situación. Por ejemplo la presencia de números en un tweet puede hacer referencia a una cantidad de víctimas, a la temperatura, a un porcentaje de votos, a una hora... etc. En esta parte nos preguntamos si esos detalles tienen una influencia sobre el target.

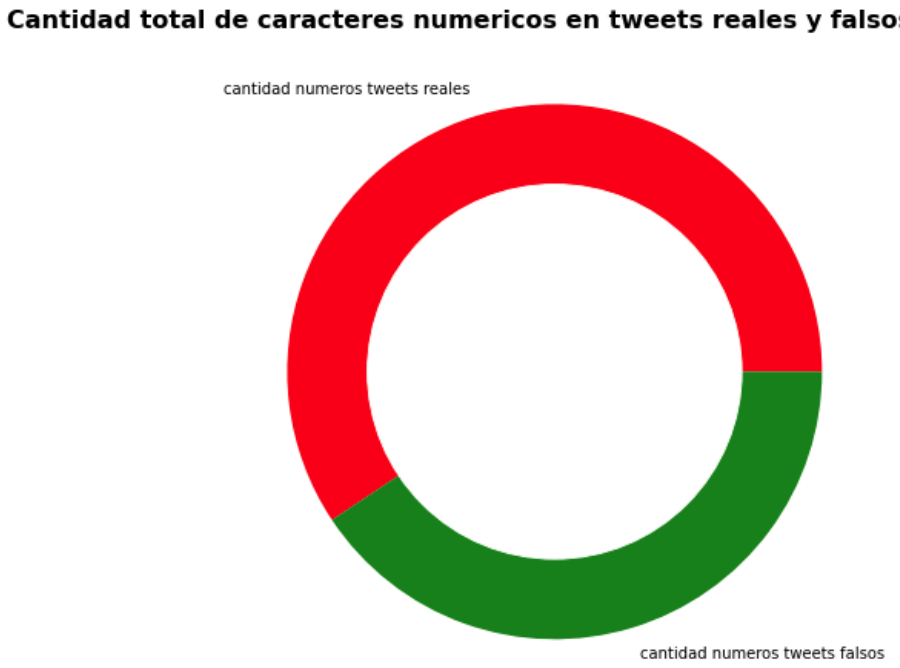

Figura 13: Cantitad total de numeros en los tweets

Como se puede ver en el gráfico, la presencia de números en los tweets reales es ligeramente mayor que en los tweets no relacionados con desastres.

Ahora veamos si la longitud de los números podría influir sobre la veracidad de los tweets.

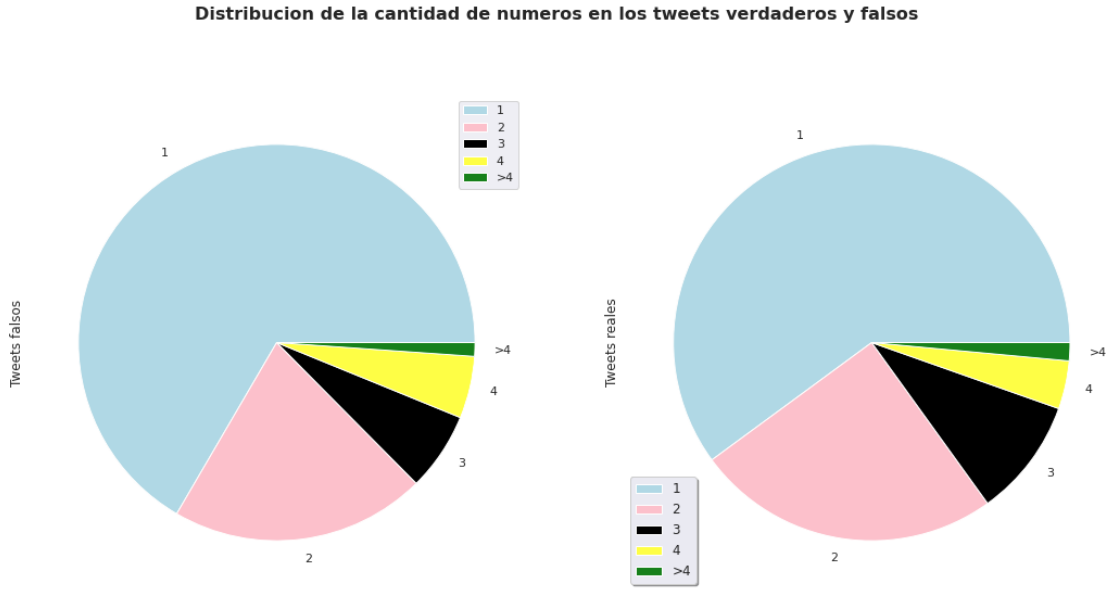

Figura 14: Distribución de la longitud de los números en ambos tipos de tweets

Vemos que los números cortos (tamaño 1) son mas frecuentes en los tweets falsos. En cambio, los números de tamaño 2 y 3 son mas frecuentes entre los tweets reales.

Nos inclinamos a pensar que la presencia o no de caracteres numéricos, no estaría relacionado con la clasificación de un tweet como real o falso.

5.4. Influencia de la longitud de un tweet sobre el target

La longitud del tweet tiene una relación con el target ?

Para evaluar esta relación, pensamos en analizar un histograma porque nos permite observar la distribución de la longitud del tweet considerando la cantidad de caracteres que contiene, según si se trata de tweets reales o falsos.

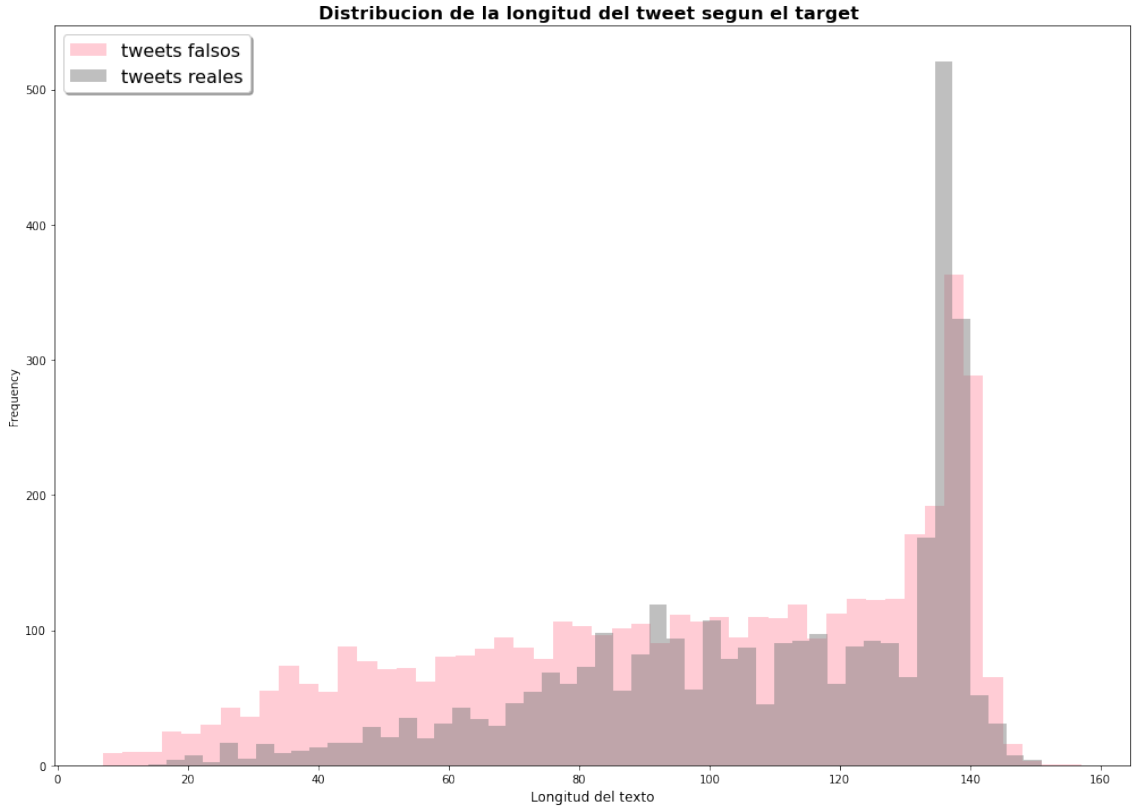

Figura 15: Distribución de la longitud del tweet según su clasificación como falsos o verdaderos

Podemos observar que a valores bajos, es decir, tweets cortos de poca cantidad de caracteres, tiene una mayor proporción de tweets falsos que verdaderos. Recién a partir de los 80 caracteres empieza a aumentar la frecuencia de los verdaderos y entre los 130 y 140 supera la frecuencia de los reales sobre los falsos.

Otra forma de analizar comparativamente las dos clases, es a través de un boxplot, que nos permite visualizar de otra manera la distribución le la cantidad de caracteres según el tipo de tweet.

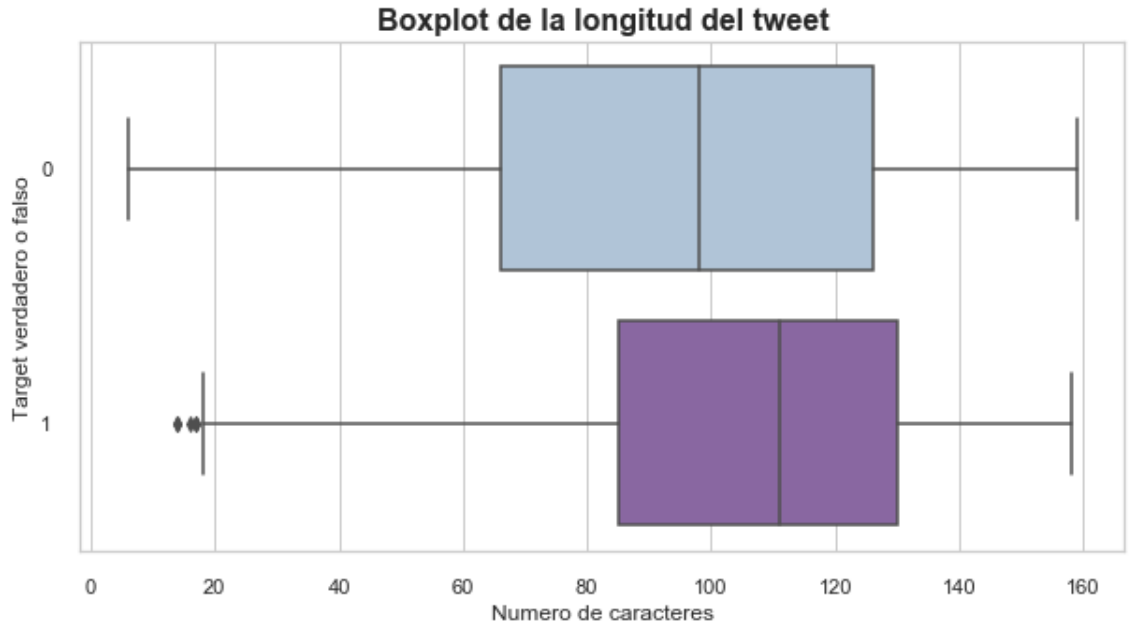

Figura 16: Distribución de la longitud del tweet según su clasificación como falsos o verdaderos

De esta forma, observamos que los tweets falsos tienen una mayos dispersión respecto de los verdaderos. El 50 % de los tweet posee una longitud entre 65 y 125 caracteres aproximadamente. Los tweets verdaderos tienen menor dispersión, ubicándose la mitad entre 85 y 130 caracteres. También podemos observar los valores máximos y mínimos para cada clase, que en este set de datos son muy parecidos, una leve diferencia en los valores mínimos y casi el mismo valor en los máximos.

Anteriormente calculamos que el promedio global de la longitud de los tweets independientemente de su clase es de 102 caracteres. Pareciera que los tweets verdaderos tienen una leve tendencia a longitudes mayores que el promedio global, pero no creemos que el análisis de solo esta variable tenga un gran impacto sobre la clasificación del tweet, es por eso que seguimos el análisis en busca de otros atributos que puedan tener algún aporte extra.

5.5. Influencia de la presencia de un link, contacto, o hashtag sobre el target

En esta parte nos hacemos las siguientes preguntas :

- La presencia en el tweet de http o https (o sea hay un link) tiene una relación con el target ?
- La presencia de '@' en el tweet tiene una relación con el target ?
- La presencia de hashtag en un tweet tiene una relación con el target ?

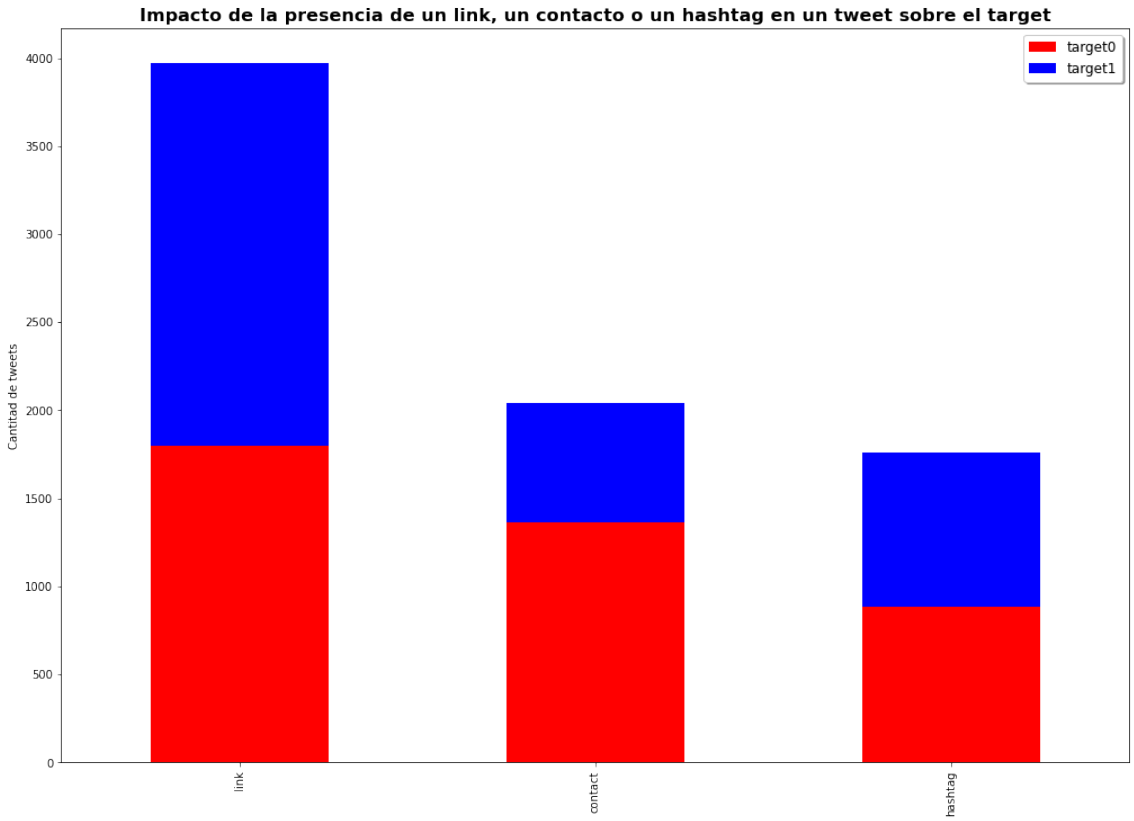

Figura 17: Cantidad de tweets falsos o verdaderos para tweets que contienen links, contactos o hashtags

En la visualización anterior observamos que agregar un link a un tweet puede agregar veracidad al mismo (hay mayor cantidad de tweets verdaderos que tienen un link). En cambio, la presencia de un contacto en un tweet agrega inexactitud (Hay mayor cantidad de tweets falsos con un contacto que verdaderos). La presencia de hashtag no parece condicionar que el contenido del tweet sea real o falso. En general, las diferencias no son tan marcadas, por lo tanto, la presencia de uno de esos tres elementos no nos permite concluir de forma fehaciente sobre la veracidad de un tweet.

5.6. Palabras mas frecuentes en los tweets

Figura 18: Palabras mas frecuentes en los tweets

En la visualización anterior se ve que fire, new, people, time, amp son palabras frecuentes.

6. Conclusiones

Consideramos que el análisis que realizamos abarcó la mayoría de los parámetros que se pueden estudiar de este set de datos que está compuesto por solo 5 atributos.

En cuanto a los resultados obtenidos, pensamos que podrían ser mucho más precisos si el csv contara con más de 7600 registros, teniendo en cuenta que en forma general se estima que 500 millones de tweets son publicados por día.

Otro dato importante, es que hasta fines de 2017 Twitter solo permitía un máximo de 140 caracteres por tweet y luego de esa fecha el valor permitido se duplicó a 280. Revisando la información provista en la competencia de Kaggle, no pudimos encontrar

ninguna referencia en cuanto a la fecha de los datos seleccionados, pero suponemos que deben estar con el viejo limite de 140 caracteres.

Este dato es clave porque limita el aporte de la longitud del tweet al criterio de clasificación como real o falso, y da lugar a que otros atributos nos ayuden a poder clasificar nuevos registros.

En este sentido, concluimos que los insights obtenidos en este análisis, no van a tener un gran aporte de manera individual, pero seguramente la combinación de varios de ellos nos van a ayudar a crear un modelo que pueda clasificar nuevos tweets con una buena performance.