FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1

[75.06] Organización de datos
21 de mayo de 2020
Primer cuatrimestre de 2020

Repositorio: github.com/d-bizari/Organizacion-de-datos

Alumno:	Bizari, Daniel
Número de padrón:	100445
Email:	dbizari@fi.uba.ar
Alumna:	Cruz, Milagros
Número de padrón:	101228
Email:	mvcruz@fi.uba.ar
Alumno:	Torresetti, Lisandro
Número de padrón:	99846
Email:	ltorresetti@fi.uba.ar
Alumno:	Ignacio Carol Lugones
Número de padrón:	100073
Email:	icarol@fi.uba.ar

Índice

1. Introducción

En el presente informe se propone analizar tweets que han sido publicados y, para ello, se utilizará el set de datos de la competencia: https://www.kaggle.com/c/nlp-getting-started

El objetivo principal de este trabajo es llevar a cabo un análisis exploratorio. Primero, se preparará la información a ser examinada; luego, se formularán algunas preguntas interesantes, las cuales guiarán el comienzo de la investigación. Finalmente, a través de comparar y establecer relaciones entre las distintas variables, se llegará a resultados que permitan responder las interrogantes mencionadas y formular nuevas a medida que transcurre el proceso. Además, durante el desarrollo se confeccionarán visualizaciones que permitan comprender mejor lo investigado. Por último, se cerrará el trabajo extrayendo conclusiones sobre lo estudiado

2. Propiedades del set de datos

Para empezar, se debe conocer el set de datos con el que se trabajará. Se observaron las siguientes columnas:

- id - identificador único para cada tweet
- text - el texto del tweet
- location - ubicación asociada al tweet
- keyword - un keyword para el tweet
- target - en train.csv, indica si se trata de un desastre real (1) o no (0)

Con un panorama más claro, se podrán formular preguntas interesantes que guiarán el comienzo del análisis exploratorio. Sin embargo, antes de poder llevar esto a cabo, es necesario trabajar sobre el data set para asegurar que la calidad de los datos a trabajar sea óptima. Para ello, se cargó la información en un dataframe y se procedió a operar sobre el mismo.

3. Limpieza de la información

1. **Detección y eliminación de anomalías**

   Se define como dato anómalo todo aquel que tiene un valor imposible o conocidamente erróneo, el cual puede impactar negativamente en el análisis llevando a conclusiones equivocadas. Por ello, se estudió cada columna para determinar qué valores serían corregidos, o en caso de no ser posible, eliminados. Se definieron los siguientes criterios:

   - **Columna text**: únicamente se considerarán anómalos aquellos que tengan este campo vacío.
   - **Columna location**: en cuanto a los caracteres que forman la palabra, si hay espacios al principio, al final o caracteres "extraños" (cualquiera que no sea alfabético), se corregirá el dato. En este caso, no se descarta ninguna anomalía. Respecto al significado de la palabra en sí, durante el proceso de exploración de los datos se llegaron a observar algunos valores que no tenían sentido (por ejemplo, la palabra 'news' aparecía como 'location', la cual no representa ninguna ubicación real), por lo que fueron filtrados.
   - **Columna keyword**: mismo criterio que el aplicado a los caracteres de las palabras correspondientes a la columna location.
   - **Columna target**: como solamente puede tomar dos valores (0 si el tweet es falso, 1 si es real), cualquier otro valor será anómalo y será descartado dado que no se puede corregir ya que no sabemos el valor real del tweet y esto afectará las estadísticas de manera directa.

2. **Tratamiento de la información faltante**

   a) **Detección de información faltante**

   Al explorar el dataframe, se observó que en las columnas keyword y location faltaban una cantidad considerable de datos; es decir, ambas columnas poseían valores NaN. Se procedió a analizar estos casos:

   - **Columna keyword**: se observó que, si bien el valor de la keyword aparecía como NaN, en varias ocasiones el texto del tweet correspondiente contenía una keyword o alguna palabra que pudiera ser usada como tal.
   - **Columna location**: de manera análoga, en distintos tweets se encontraron ubicaciones que podrían ser utilizadas como valor en la columna *location*.

   Teniendo en cuenta esto, se decidió intentar obtener la información faltante a partir del campo text.

   b) **Extracción de información**

   Para realizar esto, se armaron varios conjuntos de datos, los cuales contienen posibles candidatos que podrían llenar el valor de los campos faltantes. Para el caso de la columna *Keyword*, se armó utilizando los valores no *Nan* de esta columna, y en el caso de *Location*, se armaron 3 conjuntos: uno corresponde a los datos no *NaN* extraídos de esa columna; los otros dos fueron extraídos de datos externos de nombres de países y ciudades oficiales.
   En ambos casos, se le aplicó un método al data set para detectar aquellas filas que tuvieran valor *NaN* en una de estas columnas para luego buscar dentro del campo *text* de la misma algún valor que cumpliera con los criterios definidos para poder ser utilizado en su lugar.

   Los criterios definidos para la columna *Keyword* son:

   1) Si el tweet incluye una palabra que coincide con algún valor perteneciente al conjunto de valores de *Keyword*, se la utilizará para reemplazar el dato faltante.
   2) Si el tweet tiene un numeral ('#') seguido de una palabra (también denominado hashtag), entonces se utilizará como keyword la palabra sin el numeral.
   3) En caso de no poder usar ninguna de las opciones presentadas, se pondrá 'unknown' como keyword.

   Los criterios definidos para la columna *Location* son:

   1) Si el tweet incluye una palabra que coincide con algún valor obtenido previamente del conjunto de valores de *Location*, se la utilizará para reemplazar el dato faltante.
   2) Si el tweet incluye una palabra que coincide con algún valor perteneciente al conjunto de valores de nombres de ciudades oficiales, se la utilizará para reemplazar el dato faltante.
   3) Si el tweet incluye una palabra que coincide con algún valor perteneciente al conjunto de valores de nombres de países oficiales, se la utilizará para reemplazar el dato faltante.
   4) En caso de que no se haya podido asignar una palabra con los criterios anteriores, se le asignara un 'unknown' como location.

3. **Filtrado de datos no representativos**

   Finalmente, es importante conservar solo aquellos datos que sean relevantes y contribuyan a brindarnos conclusiones reales y representativas.

Por otro lado, si se tienen pocas muestras de un mismo dato, cada muestra tendrá un peso mayor en el conjunto total y el valor no será representativo, como por ejemplo si se está analizando la relación entre verdad y los hashtags, y hay una sola muestra del hashtag 'hola' con valor de tweet verdadero, esto impacta directamente en el análisis ya que dará como resultado que los tweets que tuvieron este hashtag son siempre verdaderos. Por lo que incluir este tipo de datos podría provocar una conclusión errónea. Por ello, se eliminaron aquellos datos que tuvieran muy pocos valores (en relación al resto).

Finalmente, se consideró que la columna id no aportaba nada al análisis propuesto, por lo que esta columna fue eliminada del dataframe a trabajar.

4. Análisis exploratorio de los datos

Se comenzó el análisis partiendo de ciertas preguntas que surgieron al examinar las propiedades del set de datos: ¿Existe alguna relación entre la ubicación desde la que se envió el tweet y la keyword que contenía el texto? ¿Está esto relacionado de alguna forma con la veracidad del tweet? ¿Cómo se relaciona longitud del texto con esta última variable?

No obstante, a medida que se fue desarrollando el proceso, nuevas interrogantes fueron surgiendo, provocando que los aspectos a analizar aumentaran. Por ello, para un mejor entendimiento, se dividió esta sección del informe en distintas partes para poder explicar con detalle lo explorado.

4.1. Análisis de veracidad

4.1.1. Veracidad según la longitud del tweet

Dado que se tiene un set de datos con tweets de distintas longitudes, una pregunta que surgió fue: *¿La longitud de los tweets influye en la veracidad?* Realizando las operaciones correspondientes para obtener la longitud de los tweets, agrupándolos por longitud y filtrando aquellos grupos con una cantidad de tweets menor o igual a 10, se obtuvo el siguiente resultado.

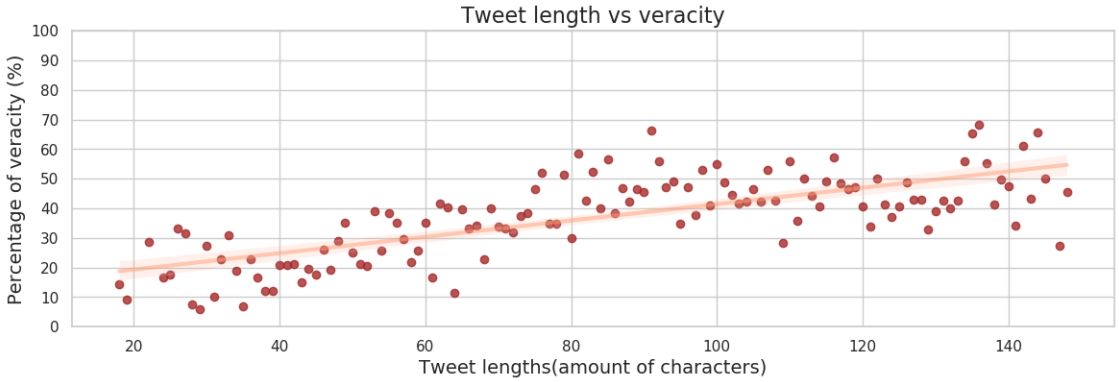

Figura 1: Longitud vs veracidad

En el eje 'y' se encuentra el porcentaje de veracidad, el cual indica la proporción de tweets reales y falos en relación a determinada longitud del tweet. Se aprecia en la figura que al aumentar la longitud de los tweets, aumenta la veracidad de los mismos; en otras palabras, la cantidad de tweets verdaderos aumenta más rápido que la de los que no lo son.

4.1.2. Tipos de keywords utilizadas en tweets con un nivel de veracidad muy alto o muy bajo

De manera análoga, otra inquietud que surgió fue si existe una relación entre las keywords y el nivel de veracidad del tweet. Para investigar este tema, se agrupó por keyword y se confeccionó el siguiente gráfico:

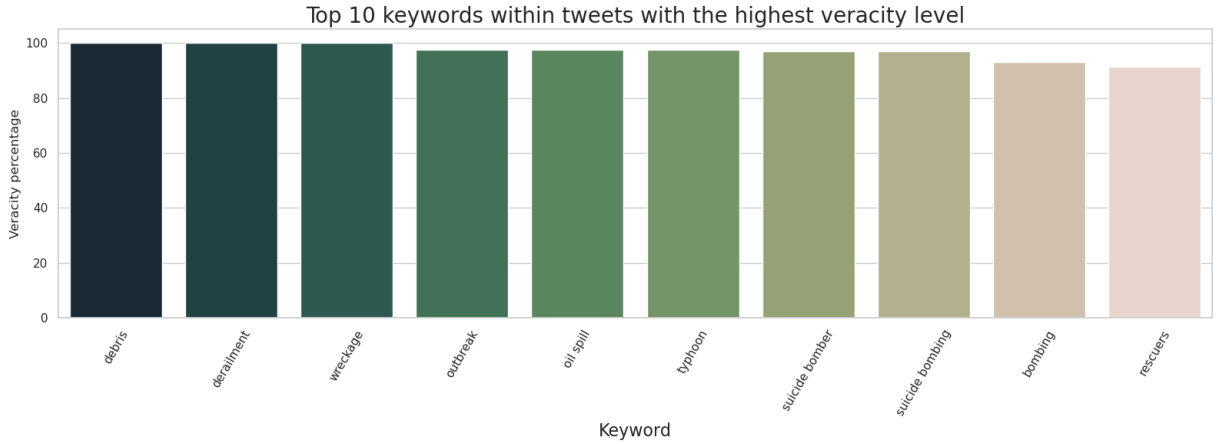

Figura 2: Keywords en tweets con alto nivel de veracidad

Se observa que los tweets que contienen las keywords presentadas son, en general, reales ya que el porcentaje oscila entre un 90 % y un 100 %. A su vez, es interesante la temática a la que aluden estas keywords: debris (ruinas), deraliment (descarrilamiento), wreckage (destrucción), etc. Todos ellas son palabras asociadas a accidentes, amenazas o situaciones del índole peligroso para la sociedad.

A su vez, interesa ver si se puede encontrar un patrón en las keywords utilizadas en los tweets que tienden a ser falsos

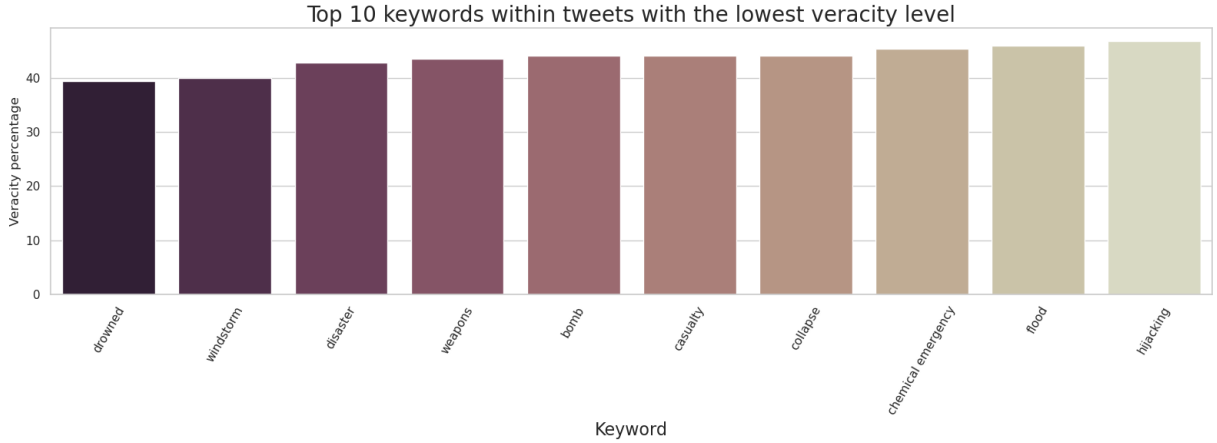

Figura 3: Keywords en tweets con bajo nivel de veracidad

Sorprende ver que estas keywords siguen con la misma temática que las anteriores, a pesar de que la veracidad de los tweets que la contienen tienen una alta probabilidad de ser falsos. Esto

nos indica que no existe una relación directa entre la temática de las keywords y la veracidad del tweet que las contiene y, por lo tanto, no se puede usar como parámetro para predecir si un tweet es real o no.

4.1.3. Tipos de hashtags más utilizados por veracidad

De lo analizado en cuanto a tópicos recurrentes en los tweets, interesa saber qué hashtag predomina en los tweets categorizados como reales según el target.

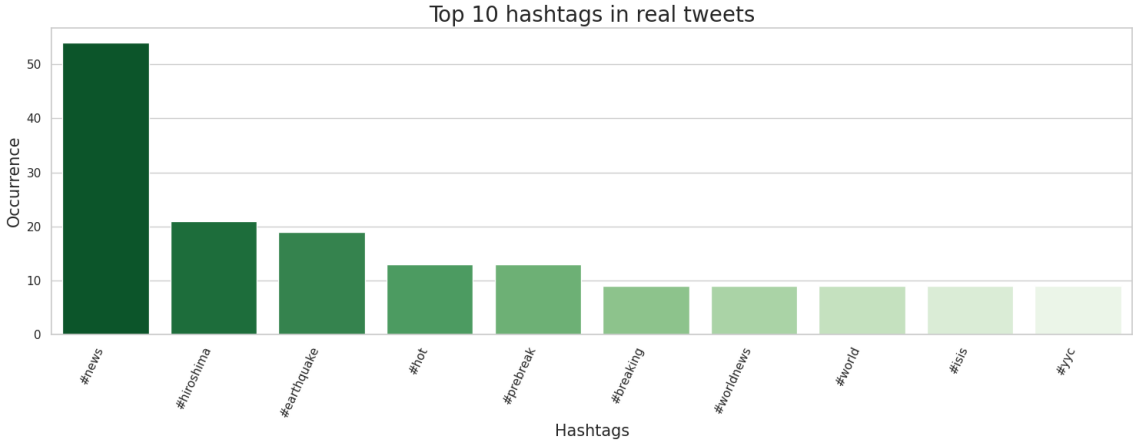

Figura 4: Hashtags más usados en tweets reales

Se puede apreciar una amplia diferencia entre el más utilizado (#news) y el resto de los hashtags, lo cual induce a pensar que hay una tendencia a twittear noticias. Además, se observan ciertos tags que, si bien no son tan frecuentes, también se relacionan al mundo del periodismo: #prebreak, #breaking (ya que 'breaking news' es una frase muy utilizada en este ámbito) y #worldnews.

Por otra parte, llaman la atención algunos hashtags que podrían dar indicios de lo que tratan las noticias: #hiroshima y #earthquake. Se podría suponer en una primera instancia que las noticias publicadas son, en general, sobre catástrofes.

Otro tag a mencionar es #isis, el cual es un gupo terrorista que sostiene ellos son los verdaderos creyentes y los no creyentes quieren destruir su religión, justificando de esa forma sus ataques contra otros musulmanes y no musulmanes. Esto podría estar conectado con #news, ya que no es raro encontrar noticias sobre ataques de este estilo

En contraposición a esto, también es importante investigar sobre los hashtags más utilizados en tweets falsos:

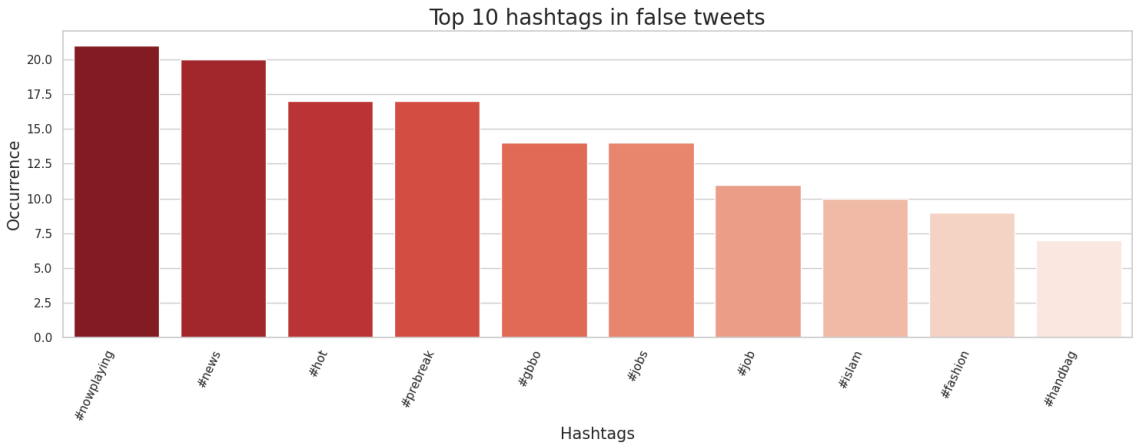

Figura 5: Hashtags más usados en tweets falsos

En este caso, en cambio, se ve una mayor variación de temas que se tratan en los tweets, por lo que analizaremos algunos de ellos por separado. El hashtag más mencionado es #nowplaying el cual, al buscarlo en la sección 'Top' de twitter, muestra contenido de este estilo:

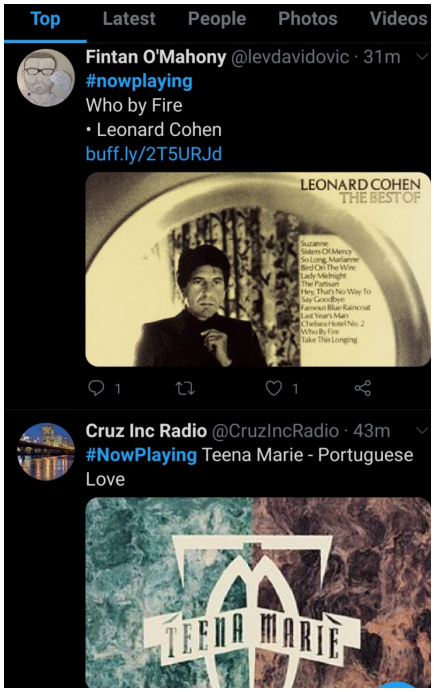

Figura 6: Uso frecuente del hashtag nowplaying

En los tweets encontrados, lo que se pudo observar es que este hashtag se utiliza más que nada para compartir canciones. Investigando un poco más sobre el tema, se encontró un análisis realizado por Michele Zappavigna sobre HERMES Twitter Corpus, un conjunto de textos destinado a la investigación científica que contiene 100 millones de palabras de Twitter.[1] La autora afirma que este hashtag es uno de los más utilizados en HERMES y que los patrones asociados al tag #nowplaying

[1]El mismo puede conseguirse en http://www.michelezappavigna.com/p/corpora.html

muestran a los usuarios compartiendo las pistas musicales que más les gustan, creándose así una selección de canciones comunitaria.

En otras palabras, este hashtag puede utilizarse para viralizar álbumes o canciones. Teniendo en cuenta que es el más usado en tweets falsos, se puede plantear la hipótesis de que la mayoría de los tweets recolectados que contienen ese hashtag son bots, utilizados para viralizar algún nuevo álbum o canción que haya sacado un (o una) cantante determinado (o determinada).

El hashtag que le sigue al analizado previamente, por una mínima diferencia, es news, el cual también es el más usado en los tweets reales. Lo mismo ocurre con prebreak, el cuarto con más apariciones. Esto podría deberse a que, si bien hay una tendencia a publicar noticias, parte de ellas son falsas. Esto, a su vez, puede relacionarse con el tag #islam, ya que, al igual que en el caso de #isis en los tweets reales, hay antecedentes de muchos atentados ocurridos y notificados por el sector de periodismo pero, al parecer, no toda la información divulgada es verdadera.

Los últimos hashtags de los tweets falsos a explorar son #hot y #jobs/#job. Al buscarlos en Twitter, aparecieron tweets de mujeres publicando fotos del índole sexual y ofertas de trabajo, respectivamente. No obstante, al estar muy presentes en tweets falsos, se podría pensar que son distintos tipos de estafas.

En el primer caso, los perfiles falsos de mujeres "provocativas"son frecuentemente utilizados para intentar hacer entrar a la gente a un link determinado y así infectar el equipo de la víctima; otro uso común es crear un vínculo de confianza lo suficientemente fuerte como para pedir dinero. No obstante, es importante destacar el tag "hot"también aparece entre los más usados en los tweets reales; aunque su valor no sea tan alto en comparación a otros, podría tratarse de usuarios compartiendo fotos.

En el segundo, las falsas propuestas de trabajo se utilizan para conseguir datos personales, financieros o incluso pedir dinero con la excusa de tener que pagar una cuantía para poder conseguir el puesto.

4.1.4. Concentración de tweets reales por ubicación

Surge una pregunta similar con la columna *Location*: ¿Cuáles son las ubicaciones desde las cuales se escribieron más tweets verdaderos? Para responder esto de una manera correcta, lo que se hizo fue buscar un porcentaje de verdad, es decir, la cantidad de tweets verdaderos en comparación con los tweets totales por ubicación. Esto se debe a que si se muestran sólo las ubicaciones con más tweets reales, estos podrían ser insignificantes en comparación con la cantidad de tweets totales de esa ubicación, lo que podría provocar conclusiones equivocadas.

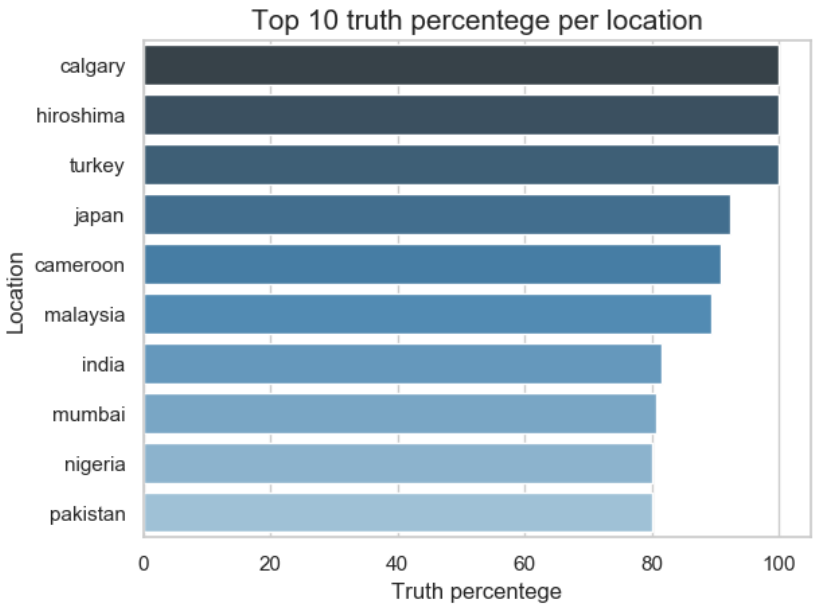

Figura 7: Top 10 de locations por porcentaje de verdad

En un primer acercamiento, se observa que la proporción tweets reales de los totales varía entre un 80 % y 100 %, los cuales son valores bastante altos. No obstante, es importante recordar que la cantidad de tweets analizados por ubicación juega un papel importante. Es por esto que es lógico que la ubicación con más tweets reales sea Calgary (ciudad de Alberta, la cual es una provincia de Candadá) o Hiroshima (ciudad de Japón) ya que la misma no tendrá tantos tweets como un país entero como Pakistán o Nigeria, quienes ocupan los puestos más bajos de este top 10.

Para entender un poco mejor lo que está ocurriendo, se decidió realizar otra visualización: un ranking que muestre las 10 ubicaciones con más tweets y el porcentaje de ellos que son reales.

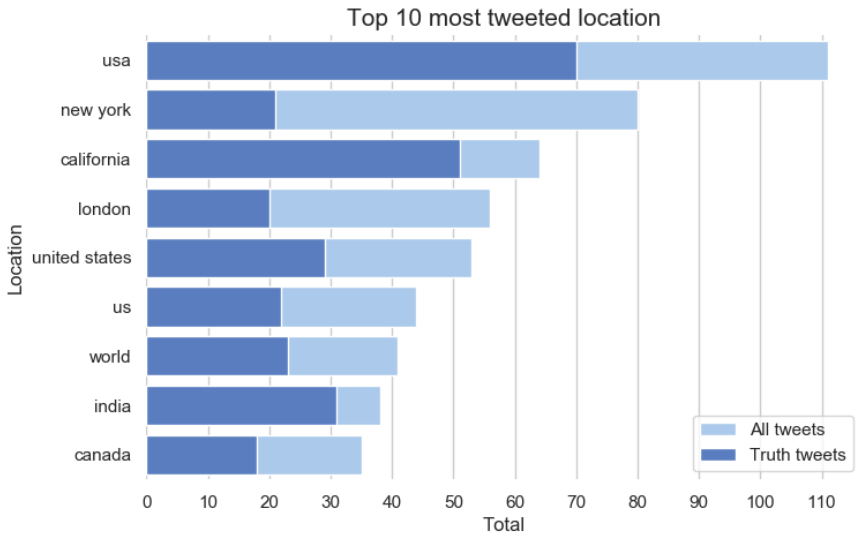

Figura 8: Top 10 de locations por cantidad de tweets

Se puede ver que Estados Unidos es la ubicación con mayor cantidad de tweets, ya que aparecen

tanto el país como estados dentro de él: usa, new york, california y united states. No obstante, a pesar de ser los mas frecuentes en el data set, también son los que más tweets falsos tienen.

En contraposición a esto se encuentra el caso de India, el cual sí se encuentra en el top 10 de ubicaciones de mayor porcentaje de verdad y se puede ver a simple vista que la propoción de tweets verdaderos es mayor que la de tweets falsos.

4.1.5. Concentración de tweets falsos por ubicación

Ahora se quiere ver cuáles son las *Location* más "mentirosas", es decir, las que posean un menor porcentaje de verdad.

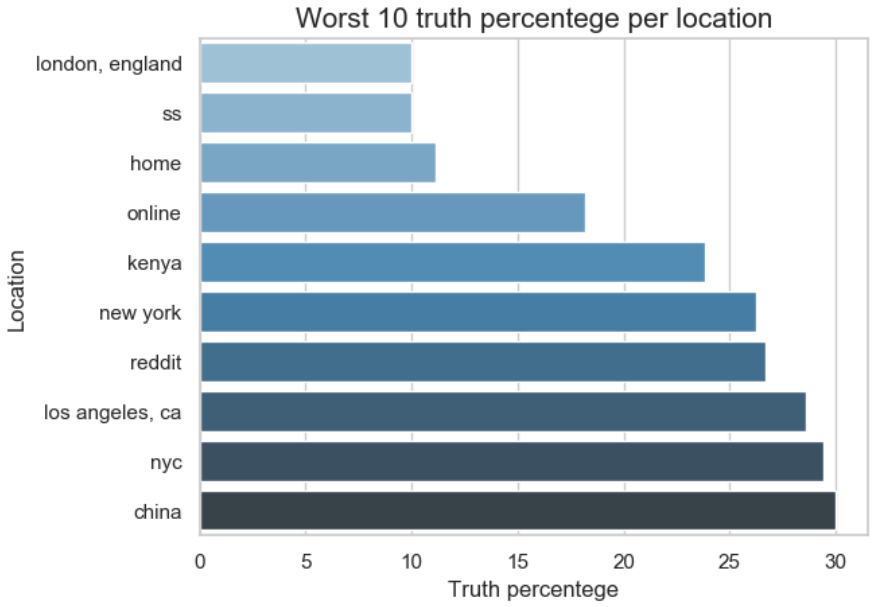

Figura 9: Peores 10 locations por porcentaje de verdad

Se puede observar que algunos de estos valores no representan ubicaciones físicas o que son inexactas. Ejemplos de esto son *reddit*, la cual es una página de humor, memes, disputas, etc; *online*, que simplemente muestra que se está navegando en internet y *Home*, que podría ser cualquier parte del mundo.

Una ubicación interesante a destacar es *ss*, que es la abreviatura de Schutzstaffel o Escuadras de protección. Esta fue una organización fundada en 1925 por Adolph Hitler, la cual se encargaba de la seguridad, la identificación del origen étnico, la política de establecimiento demográfico, y la recopilación y el análisis de información de inteligencia con el objetivo de resolver el supuesto problema de raza que había en Alemania e instaurar un nuevo orden. Esto incita a pensar el por qué de la aparición de tweets con estos posibles tintes políticos y cómo se relacionaría esto con su veracidad. Una posible teoría es que estos tweets fueron generados por un programa informático con el objetivo de divulgar una ideología en particular como es el neonazismo.

También resulta interesante la aparición de China en este ranking dado que en varias ocasiones ha recibido fuertes acusaciones de estar mintiendo y/o ocultando información, como por ejemplo el caso de *Huawei*, o más recientemente las acusaciones respecto al origen del virus *covid-19* y la cantidad de infectados y muertos totales en esa región.

4.1.6. Cantidad de tweets por ubicación y su veracidad

Habiendo analizado las ubicaciones con que concentran mayor cantidad de tweets falsos y verdaderos, interesa ver si existe alguna relación entre la cantidad de tweets por ubicaciones y la verdad del mismo.

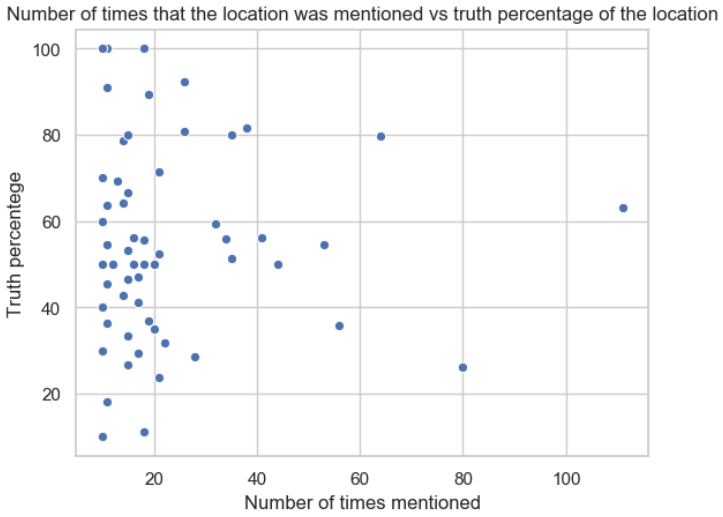

Figura 10: Cantidad de tweets por *Location* relacionado con su porcentaje de verdad

Se observa que la mayoría de los puntos se concentran entre los 10 a 30 tweets por *Location* y se encuentra todo tipo de porcentaje de verdad para ese rango. Luego, hay ubicaciones con mayor cantidad de menciones pero se encuentran muy dispersos y no siguen ninguna tendencia, por lo que no pareciera haber alguna relación entre estas dos variables.

4.2. Análisis geográfico

4.2.1. Ubicaciones y sus niveles de veracidad de tweets

La primera interrogante a plantear fue si existe alguna relación entre la ubicación desde donde se mandó el tweet y la veracidad del mismo. Esto llevó a realizar una visualización a escala mundial, en donde se muestra el porcentaje de verdad por país tal y como se observa a continuación:

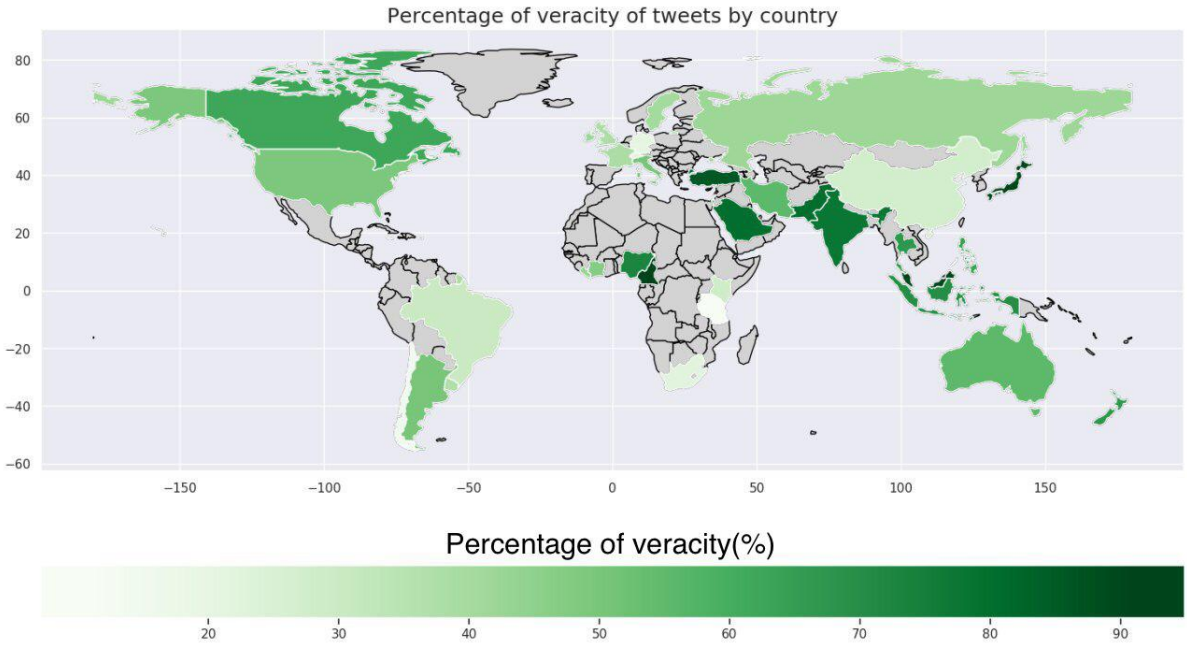

Figura 11: Porcentaje de veracidad por país

En un primer acercamiento, se percibe la mayoría de los tweets reales de este set de datos provienen de India, Pakistán, Arabia Saudita y Turquía. En cambio, los tweets falsos se concentran en China, Brasil, Sudráfrica, Kenia, Tanzania y Alemania.

4.2.2. Ubicaciones con mayor cantidad tweets

A su vez, es importante tener en cuenta la cantidad de tweets que fueron emitidos desde cada ubicación. Debido a esto, se confeccionó un ranking con los 10 países con mayor cantidad de tweets. El resultado obtenido se muestra en la siguiente figura.

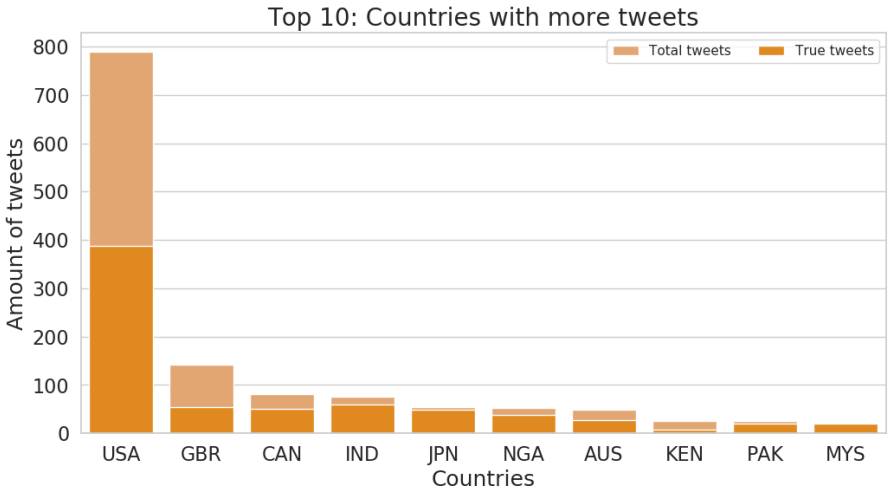

Figura 12: Top 10: países con mayor cantidad de tweets

Se puede observar que el país con mayor cantidad de tweets es Estados Unidos, por lo que se

decidió realizar un análisis sobre este caso en particular. Lo primero que se hizo fue ver cómo se distribuyeron los tweets en los distintos estados, teniendo en cuenta a su vez las longitudes de los mismos.

Caso de estudio particular: Estados Unidos

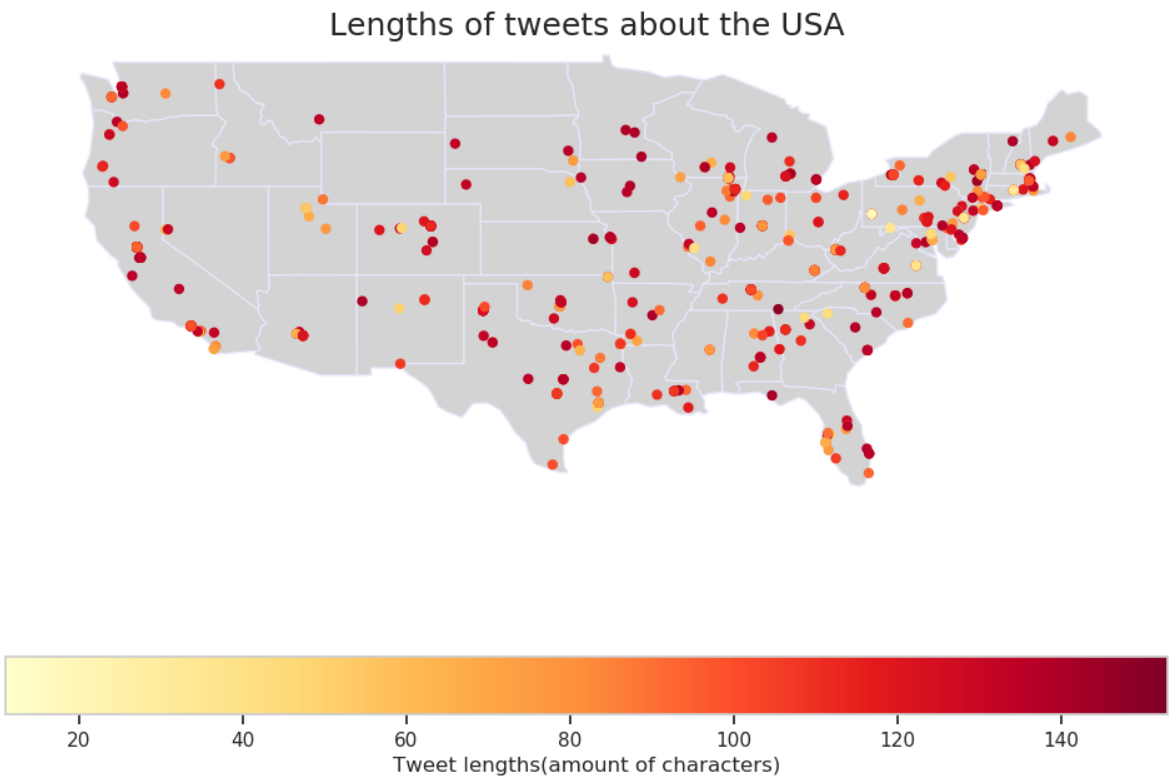

Figura 13: Tweets sobre Estados Unidos

Se puede observar que, en general, los tweets publicados tienen entre 80 y 140 caracteres, siendo los más largos los predominantes. Luego, analizando la figura 13 y en referencia a los resultados mostrados de longitud de los tweets versus veracidad (figura 1), se espera, por ejemplo, que los estados de California y Texas posean un porcentaje de veracidad mayor, dado que tienen tweets con longitudes relativamente largas.

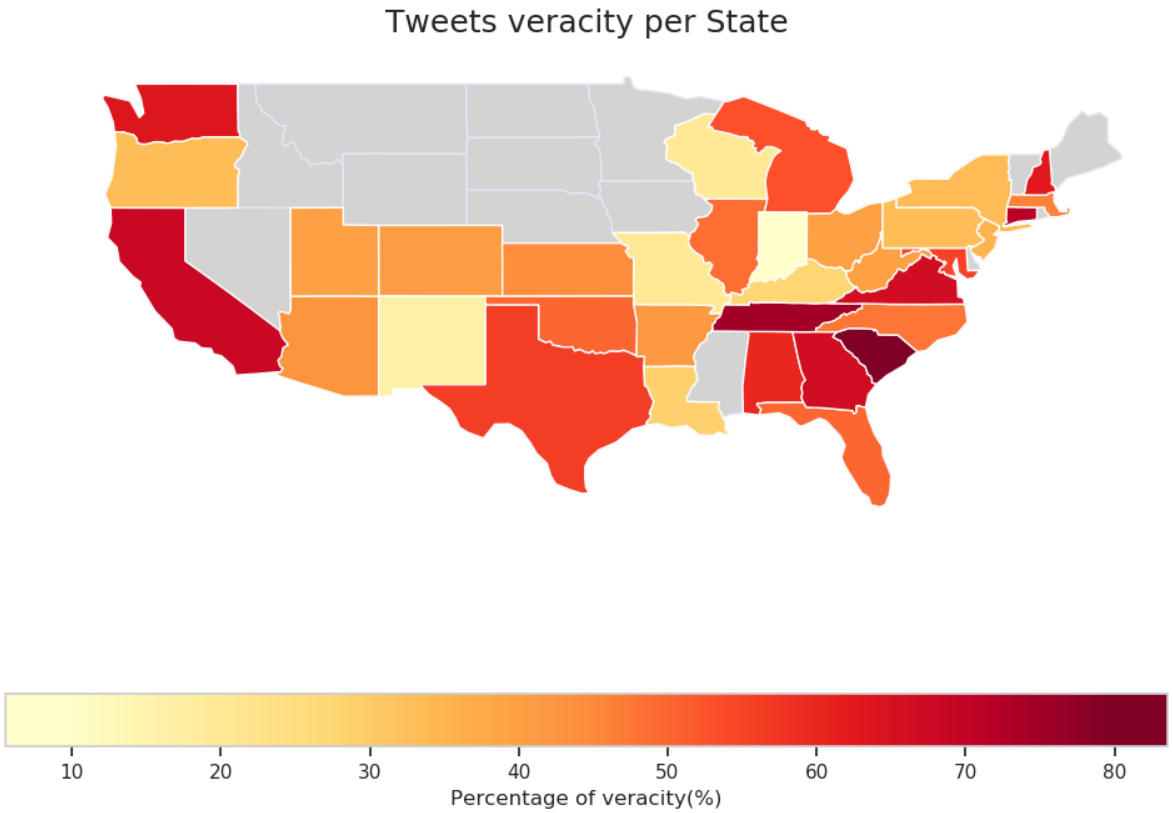

Figura 14: Porcentaje de veracidad por estado

Para el porcentaje de veracidad por estado no se tuvieron en cuenta los estados con una cantidad de tweets menor a cinco.
Se puede concluir que lo supuesto para los estados de California y Texas es correcto, aunque el valor de Texas es un poco más bajo de lo esperado. Por otra parte, no solo estos estados presentaron un nivel de veracidad importante: Carolina del Sur, Tenesse y Connecticut también muestran un nivel alto.

4.2.3. Ubicaciones con mayor cantidad de keywords

Otra pregunta interesante que surgió es si existe una relación entre las distintas ubicaciones mencionadas en el tweet en base a la cantidad de keywords que contiene. Para ello, se confeccionó la siguiente visualización:

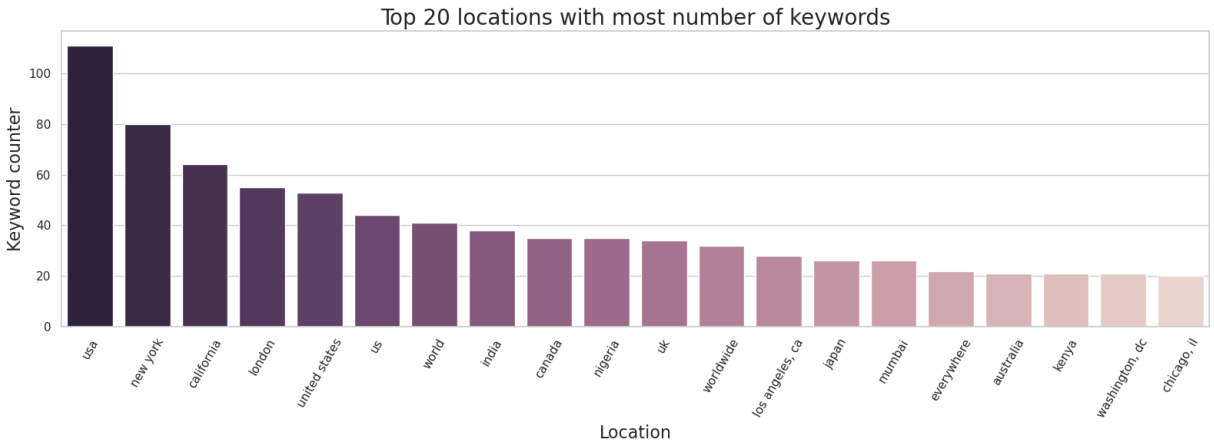

Figura 15: Ubicaciones con mayor cantidad de keywords

En primer lugar, se observa que la cantidad de keywords por ubicación va disminuyendo de forma exponencial. Además, otra cosa a destacar es el tipo de ubicaciones predominantes en el gráfico: de las 20 ubicaciones que se muestran, 7 son de Estados Unidos (usa, new york, united states, washington dc) Esto indica que los tweets relacionados a este país contienen muchas palabras clave, por lo que también se podría analizar qué palabras clave son las más frecuentes y si se puede encontrar alguna conexión entre ellas.

4.3. Análisis del contenido de los tweets

4.3.1. Keywords más utilizadas

Un aspecto interesante para explorar es qué tópicos se tratan en los tweets. Para ello, se empezó analizando las keywords y cuáles fueron las más utilizadas. Luego, se tomaron las 20 más usadas y se construyó el siguiente gráfico:

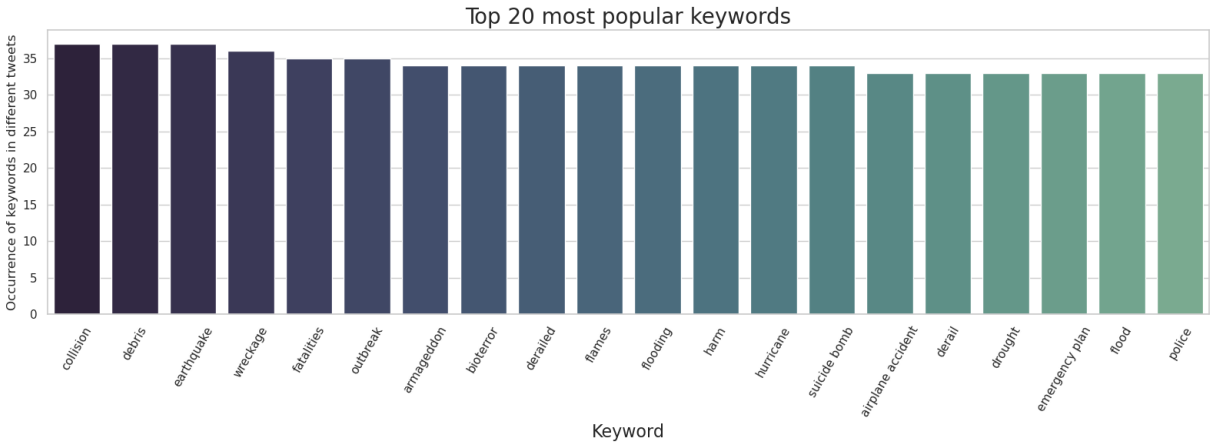

Figura 16: Keywords más recurrentes

Lo primero a observar es la clase de keywords más usadas: collision (choque), debris (escom-

bros), earthquake (terremoto), wreckage (destrucción) entre otras. Todas ellas están relacionadas en la medida que tratan temas del índole catastrófico: desastres naturales, atentados, accidentes, etc.

Además, se puede ver que las apariciones de dichas keywords oscilan en una cantidad de entre 33 y 37 valores, lo que indica cierta uniformidad entre las veces que se repitieron en diferentes tweets tal y como se puede apreciar en la figura 6. En otras palabras, no existe una concentración de keywords de una tragedia en particular, sino que el tópico recurrente, en base a lo analizado, es el de distintos eventos que podrían ser peligrosos para la sociedad.

Teniendo en cuenta lo analizado en la figura 4 sobre los tweets que contienen noticias, se podría pensar que hay una tendencia a publicar noticias sobre tragedias que tienen un impacto directo sobre la sociedad. Además, esto puede reafirmarse observando otros de los hashtags más usados que guarden relación con carástrofes tales como #hiroshima y #earthquake.

Caso de estudio particular: Desastres naturales

Se puede observar que, de todos los eventos mencionados en las keywords, uno de los tópicos presentes en el set de datos está relacionado con los desastres naturales. La información de los mismos se obtuvo de la página de EM-DAT[2], en dicha página los clasifican en un tipo principal (natural o tecnológico), en un subgrupo y luego las categorías principales de ese subgrupo. Como el objetivo fue analizar únicamente los desastres naturales, se realizó una clasificación diferente, el subgrupo pasó a ser el grupo y las categorías principales pasaron a ser los subgrupos. A partir de estos datos extraídos surgieron preguntas tales como:

- ¿Cuál es el subgrupo de los desastres naturales con mayor veracidad?

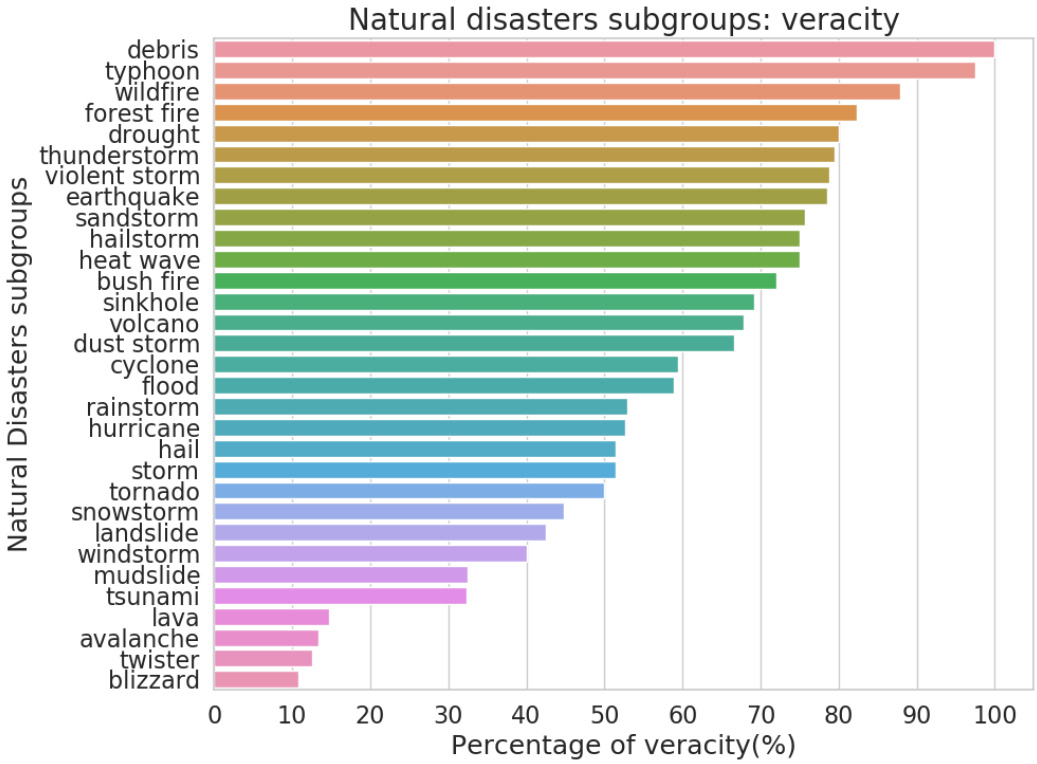

Figura 17: Veracidad de los subgrupos de desastres naturales

Se observa en la figura 17 que el subgrupo con mayor veracidad es debris (escombros) con un 100 % de veracidad, mientras que el de menor porcentaje de veracidad es blizzard (tormenta de

[2]Emergency Events Database: https://www.emdat.be/classification

nieve).

¿Cuáles son las características de los 5 subgrupos con mayor veracidad?

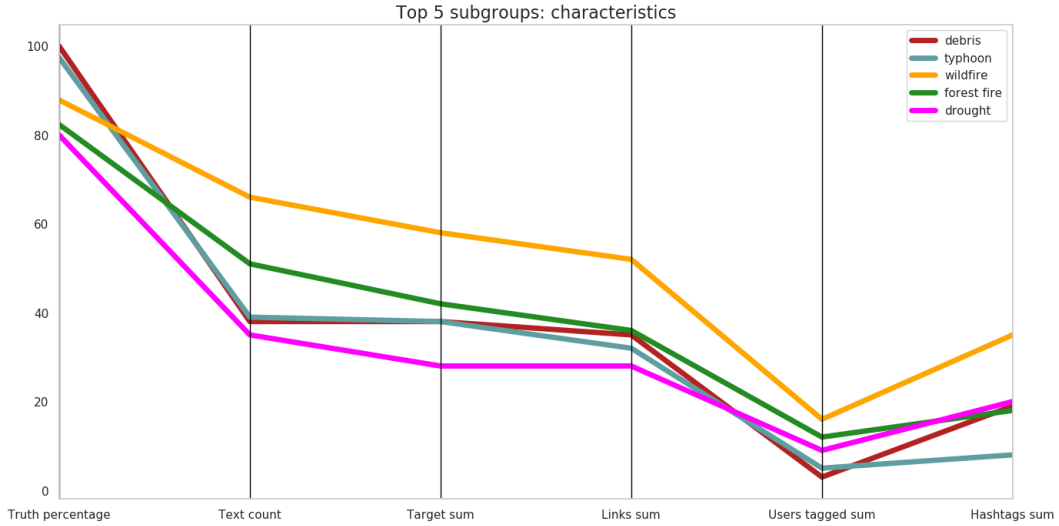

Figura 18: Características de los 5 subgrupos con mayor veracidad

Como se dijo anteriormente, debris es el que posee el mayor porcentaje de veracidad. Analizando sus características se puede observar que la cantidad de links es semejante a la cantidad de tweets, se podría decir que poseen un link por tweet. Del top 5, es el que menos usuarios etiquetados posee en sus tweets, y con respecto a los hashtags se encuentra en un punto medio.

Las características del segundo subgrupo con mayor veracidad, typhoon (tifón), tiene un porcentaje de veracidad muy levemente por debajo de debris, ya no se cumple la relación anterior de un link por tweet,la cantidad de usuarios etiquetados también es baja pero se encuentra por encima de debris y la mayor diferencia entre ellos se da con la cantidad de hashtags, typhoon del top 5 es el que menos hashtags posee.

Por último, si se observan las características de wildfire (fuego salvaje) se encuentra tercero en porcentaje de veracidad, sin embargo es el que posee mayor cantidad de links, usuarios etiquetados y hashtags.

Con esto se puede concluir que: muchos elementos en los campos de links, hashtags y usuarios etiquetados, no garantiza que el subgrupo vaya a poseer mayor porcentaje de veracidad, sino que se debe tener aproximadamente un link por tweet, etiquetar a pocos usuarios, como por ejemplo canales de noticias ya que etiquetar a influencers o famosos provocaría que se dude mucho acerca de la veracidad del tweet, y utilizar una cantidad moderada de hashtags para poder lograr que se vuelva trending topic, como es el caso de *Earthquake* que su porcentaje de veracidad se encuentra muy cerca del 80 % y en la figura 20 se aprecia que se encuentra dentro de los trending topics.

- ¿Dónde ocurrieron en el mundo los tweets sobre desastres naturales?

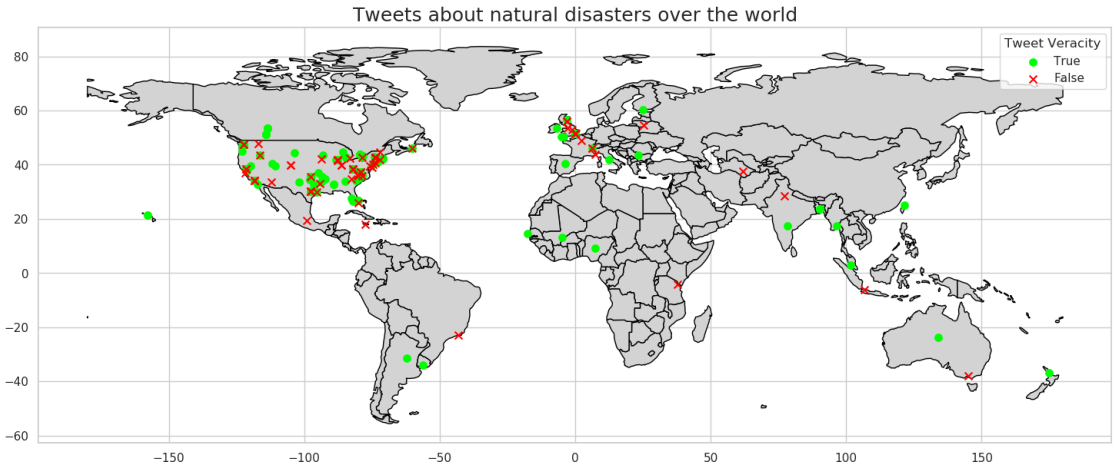

Figura 19: Desastres naturales alrededor del mundo

El hecho de que Estados Unidos posea mayor cantidad de puntos en comparación con el resto de los países del mundo tiene sentido ya que por la figura 12 se sabe que es el que más tweets posee. Otro hecho que hace que haya tan pocos puntos marcados en el mapa es que para este gráfico es necesario que las *Location* sean mas precisas, o sea, que en lo posible aclaren ciudad y país, muchos desastres naturales indicaban como *Location* un país lo que llevó a que sean descartados al momento de realizar el gráfico. Un análisis que no pudo realizarse por este hecho era ver si los tweets relacionados a terremotos o actividad sísmica se daban más a lo largo del cinturón de fuego del Pacífico, dado que esta zona posee una intensa actividad sísmica en comparación con otras partes del mundo.

4.3.2. Hashtags más utilizados

Siguiendo esta idea de ver ver cuáles son los *trending topics*, se tiene particular interés en los hashtags, los cuales permiten etiquetar o clasificar los mensajes de Twitter, agrupando aquellos tweets que refieran a una misma temática. En la figura 7 se muestran los *trending topics* para este set de datos, mostrando la cantidad total de ocurrencias de los mismos y cuantas de estas ocurrencias son verdaderas.

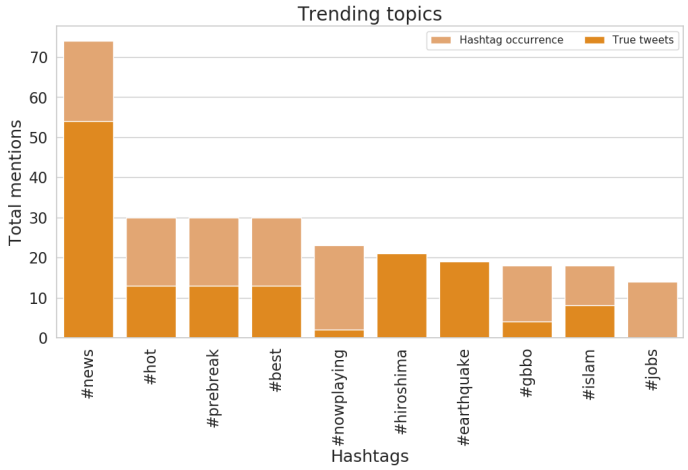

Figura 20: Trending topics: cantidad de apariciones y cantidad de tweets reales en los que fueron usados

Aquí se puede ver que el tag más recurrente es el de news y que la cantidad de tweets reales que lo usa representa un porcentaje importante de la cantidad total de tweets que lo utilizan. El hashtag prebreaks, en cambio, está más balanceado, por lo que la mitad del total es de tweets reales y la otra, de tweets falsos. Teniendo en cuenta que ambos tags están relacionados al mundo del periodismo y sumando lo analizado en la subsección 4.1.3 (Tipos de hashtags más utilizados por veracidad), se podría afirmar que si bien Twitter se usa frecuentemente como herramienta para informar sobre noticias reales y que superan en cantidad a las fake news, la cantidad de información falsa es considerable.

4.3.3. Cantidad de hashtags por keyword

Los anteriores análisis aún dejan una duda sin contestar sobre las keywords: ¿Cómo se relacionan las keywords con los hashtags?

Para empezar a responder esto, primero hay que investigar lo siguiente: ¿Cuántos hashtags tiene un tweet que contiene determinada keyword? Para ello, se decidió analizar sólo las 20 keywords que tienen más hashtags, por lo que se filtró el dataframe obteniendo el siguiente resultado:

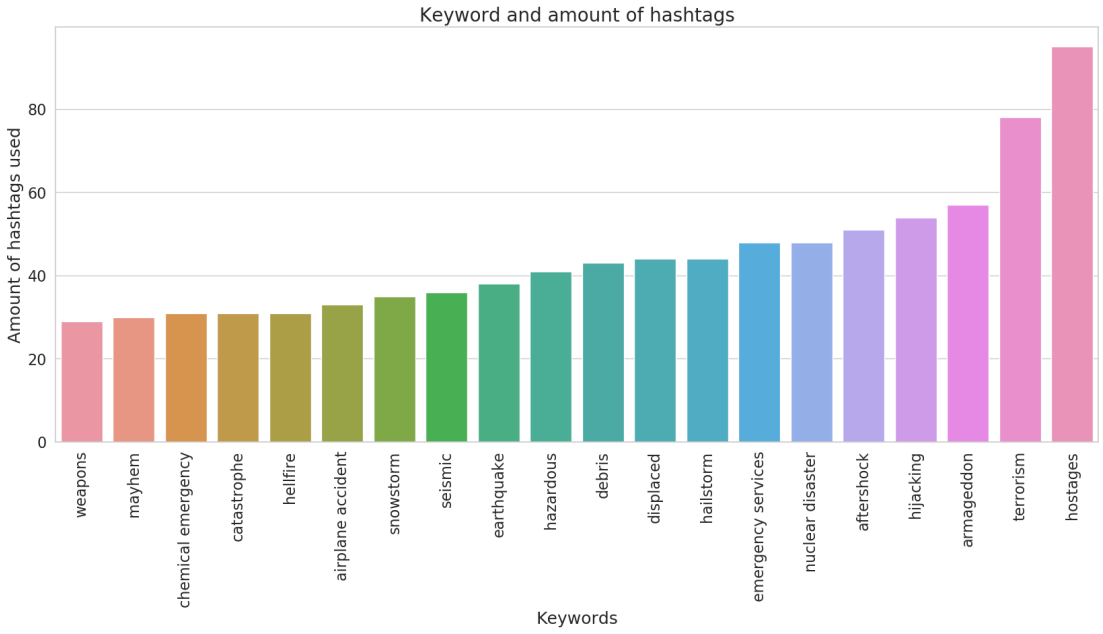

Figura 21: Keywords y la cantidad de hashtags que usan los tweets que la contienen

Estas son las 20 keywords cuyos tweets en los que se encuentran dentro utilizan más cantidad de hashtags; es decir, que tienen más tags por tweet.

En este gráfico se puede ver que 'hostages' (rehenes) y 'terrorism' (terrorismo) son las que tienen más hashtags asociados. Como los tags sirven no sólo para agrupar a los tweets por temática sino también para viralizar ciertos tópicos en Twitter, se podría suponer que se trataba de difundir algún atentado que hubiera ocurrido. Esto se lo podría relacionar con lo analizado en la figura 8, en donde se pudo observar que Estados Unidos es la ubicación con mayor cantidad de tweets. Este país ha tenido diversos ataques de esa índole, como se lo fue el atendado de las torres gemeleas en el 2001 o, un caso más reciente, el de tiroteo masivo en un Walmart de Texas en 2019.

Además, se lo puede conectar con lo investigado en la sección 4.1.3 Tipos de hashtag por veracidad, en la que se vio que algunos de los hashtags más usados en los tweets reales y falsos fueron, respectivamente #isis y #islam. Juntando todo esto, una posible conclusión sería que se utiliza twitter para divulgar noticias de atentados que ocurrieron en Estados Unidos, utilizando

los hashtags para más difusión; no obstante, una cantidad considerable de estas noticias podrían ser falsas.

Esto hace que se plantee otra interrogante: ¿Eso significa que las keywords que más utilizan hashtags son las más importantes o más veraces?

4.3.4. Veracidad de keywords que más hashtags poseen

Para responder a esta pregunta, se buscó solamente la veracidad de las keywords que están siendo usadas en el gráfico anterior, ignorando las que no aparecen en él. Se obtuvo la siguiente visualización:

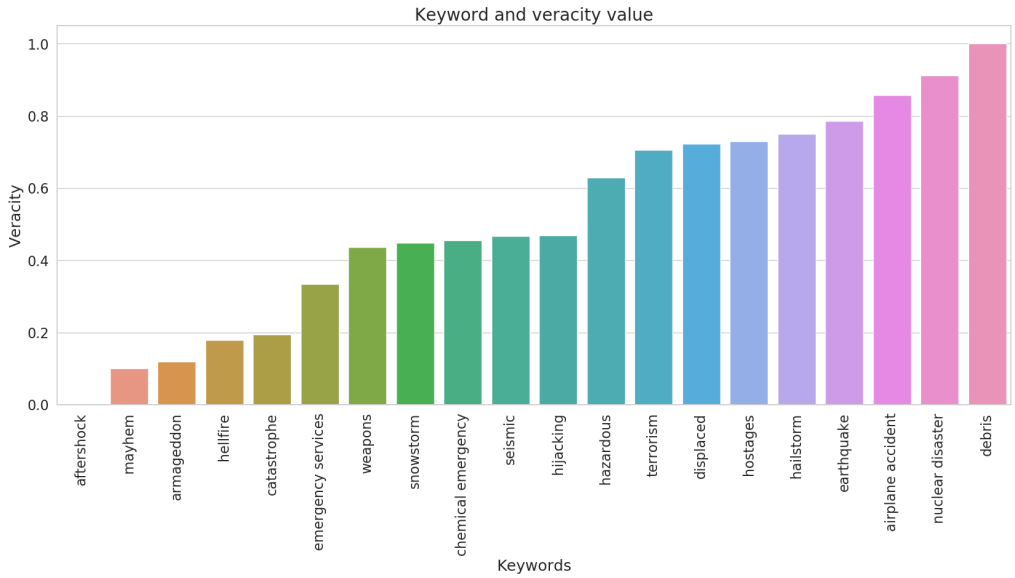

Figura 22: Keywords con mayor cantidad de hashtags en los tweets que las contienen y su nivel de veracidad

Interesa ver los valores de veracidad obtenidos para los tweets que contienen las keywords 'hostages' y 'terrorism', ya que son las que más cantidad de hashtags asociados tienen. Se puede observar que su nivel de veracidad oscila entre 0.6 y 0.8, lo cual, siguiendo con la hipótesis de divulgación de atentados vía Twitter, podría llevar a pensar que las posibles noticias que involucran estos tópicos tienden a ser verdaderas.

Al pasar a un análisis más general de este gráfico, se observa que no existe una relación clara con respecto a la cantidad de hashtags usados por keyword con la veracidad de la keyword. Eso se puede ver de forma más clara con 'aftershock', que es la quinta keyword con más hashtags pero tiene una veracidad de 0 entre todos los tweets en los que aparece.

Análogamente, 'debris' no es una keyword que este en el top de cantidad de hashtags, pero tiene el valor de veracidad asociado a los tweets que la contienen es muy alto.

Si no se consideran estos casos extremos, se puede apreciar una leve tendencia que a mayor cantidad de hashtags por keyword, disminuye la veracidad asociada a los tweets que las contienen, lo cual se denota en casos como 'weapons', 'catastrophe' o 'airplane accident'.

Sin embargo, no se tienen pruebas suficientes como para considerar verídica a esta regla de análisis. En otras palabras, las keywords y la cantidad de hashtags que contienen los tweets no muestran una relación directa sobre la veracidad de los mismos.

Como conclusión, se llega a la idea de que, para analizar la legitimidad de un tweet que contiene determinada keyword, no sería totalmente correcto analizar la cantidad de hashtags que contiene. Es decir que el hecho de que haya muchos hashtags en un mismo tweet para determinada keyword no impacta en la veracidad del mismo.

4.3.5. Usuarios más etiquetados

Dado que los tweets además de hashtags pueden contener usuarios etiquetados, se decidió analizar cuáles eran los 10 usuarios más etiquetados y ver a su vez si existía alguna relación de veracidad en torno a su mención en el tweet. La figura 23 muestra los resultados de este análisis.

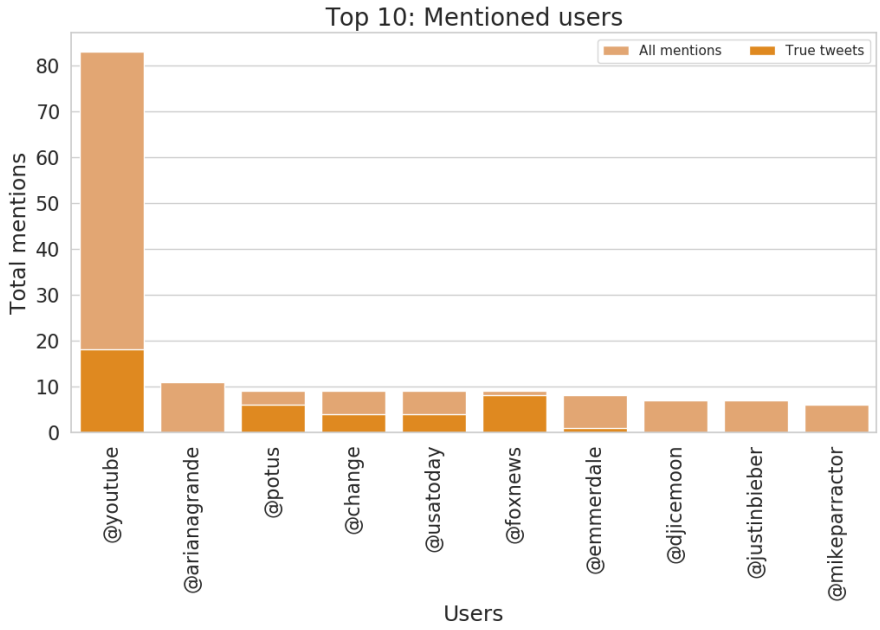

Figura 23: Top 10 usuarios etiquetados

Se observa que, en general, la mayor parte de los tweets que mencionan a los usuarios mostrados son falsos. En particular, se destaca *@youtube* debido a la cantidad de menciones que se encontraron. Al indagar un poco en Twitter, se encontró que la mayoría de los tweets que mencionan a este usuario comparten videos musicales. Teniendo en cuenta lo antes mencionado sobre la veracidad de los tweets en relación con *@youtube*, se podría pensar que la gran cantidad de tweets falsos se debe a que se han utilizado bots para viralizar ciertos videos.

A su vez, también ocurre que los tweets que etiquetan a famosos (*@justinbieber, @arianagrande y otros*) también suelen ser falsos. Esto podría deberse a que, al etiquetar influencers, se puede lograr una mayor visibilidad del tweet y, de esta forma, divulgar cierto contenido, que puede estar relacionado (o no) a la celebridad etiquetada. Un caso conocido de esto es el que se dio hace un par de años con Justin Bieber[3].

En contraposición a esto, los tweets en los que se etiqueta a canales de noticias (*@foxnews. @usatoday*) por lo general suelen ser verdaderos.

4.3.6. Longitud del tweet y su relación con distintas variables

1. Variable: referencias dentro del tweet (hashtags, links y usuarios etiquetados):

   Como es sabido, un tweet puede contener hashtags, links o usuarios etiquetados, en base a esto se analizó como aumenta la cantidad de ellos en función de la longitud media de los tweets

[3]Nos referimos a su conocido caso con el burrito https://www.youtube.com/watch?v=Vs6In7UtyXY.

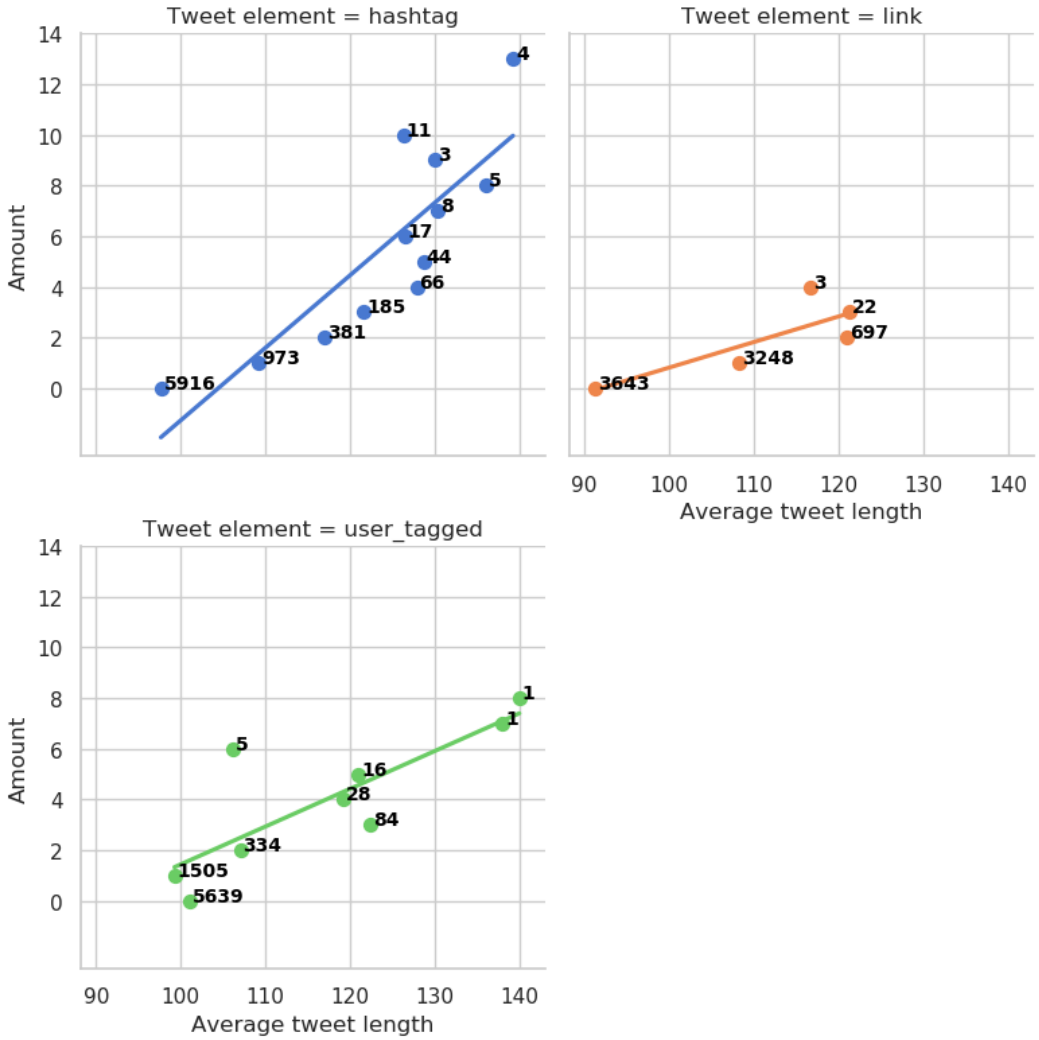

Figura 24: Aumento de hashtags, links y usuarios etiquetados en función de la longitud media de los tweets

Comparando los tres gráficos se puede apreciar que el incremento más rápido se produce para el caso de los hashtags, esto tiene sentido dado que como se vio anteriormente el uso de los mismos puede producir que cierto tema mencionado en los tweets se vuelva un Trending topic y de esta forma llegar a mayor cantidad de usuarios, también tiene sentido que al utilizar más hashtags por tweet aumente la longitud media del mismo, ya que un hashtag es una cadena de caracteres.

En segundo lugar se encuentran los usuarios etiquetados, esto también tiene sentido ya que suelen etiquetarse a influencers para que un cierto tema, sea verdadero o no, se expanda con mayor velocidad a través de las redes sociales.

En el último lugar se encuentran los links, esto se debe a que en *Twitter* los tweets no pueden superar cierta cantidad de caracteres (originalmente la longitud máxima era de 140 caracteres, a partir de 2017 el máximo es de 280) y el uso de muchos links provocaría que el límite permitido se alcance más rápido.

2. Patrones de comportamiento:

Por último, se quiso ver si existían patrones en el comportamiento de los tweets de acuerdo a la longitud de los mismos. Para analizar esto, en primer lugar se obtuvo la longitud máxima y la mínima de los tweets, 157 y 7 respectivamente. Luego se crearon los siguientes intervalos para categorizar a los tweets según su longitud:

- Pequeños - [7, 57) - Tweet Size ID: 1
- Medianos - [57, 107) - Tweet Size ID: 2
- Largos - [107, 157] - Tweet Size ID: 3

Para la búsqueda de patrones no solo se consideró la longitud de los tweets, sino que también se tuvo en cuenta: la cantidad de links, hashtags, usuarios etiquetados y el porcentaje de veracidad. Una vez seleccionado los elementos a analizar de los tweets, se los agrupó por su longitud y se obtuvieron las estadísticas para los distintos campos mencionados. Por último, previo a realizar el gráfico, se realizó un filtrado para eliminar aquellos grupos que contuvieran una cantidad de tweets inferior a 10, dado que de incluirlos obtendríamos resultados poco certeros.

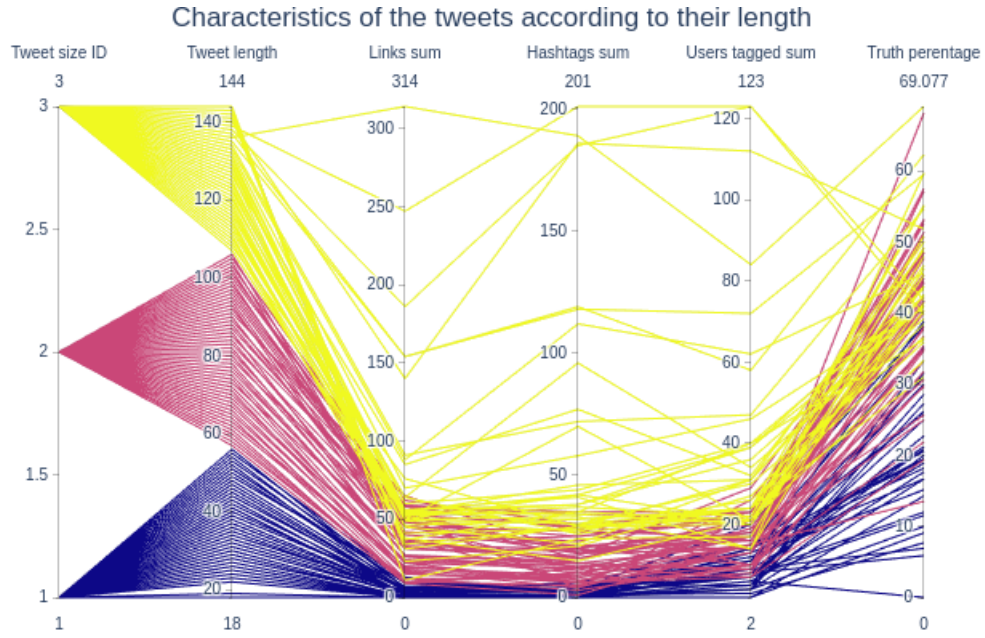

Figura 25: Patrones de los tweets según su longitud

En este tipo de gráficos tendremos una línea por cada instancia (dato) y columnas representando cada dimensión (cada campo mencionado anteriormente seria una dimensión distinta), por cada dato se traza una una línea que pasa por los valores que toma el dato en cada dimensión. Cada dimensión tiene su propia escala de acuerdo a los valores hallados para la misma e indica el máximo valor abajo del nombre de cada una, como en el caso de *Truth percentage* que indica que el valor máximo hallado en los datos es de 69,077 %. Analizando el gráfico se observa que los tweets pequeños y medianos parecen seguir algún patrón, por ejemplo puede verse que sus líneas suelen ir bastante juntas a lo largo de todas las dimensiones, mientras que para los tweets largos ya no se cumple esto, las líneas se dispersan mucho más y en todas las direcciones.

Observación: si se quisiera analizar con mayor rigurosidad, el gráfico mostrado es interactivo y se encuentra a disposición en el repositorio de *Github*, para acceder al mismo se debe abrir el archivo *TP1.ipynb* desde *Jupyter notebook*.

4.3.7. Análisis sintáctico

Habiendo analizado semánticamente algunos elementos particulares de los textos de los tweets publicados tal y como los hashtags, las keywords y los usuarios etiquetados, surgió la necesidad de encarar el aspecto sintáctico. De esta forma, se plantea la siguiente interrogante: ¿Existe algún tipo de palabra que predomine dependiendo de la veracidad del tweet?

Se utilizó la librería NLTK para hacer un análisis genérico de los textos de los tweets. Se tuvo en consideración que podría haber sentencias que no fueran palabras reconocidas por la misma (como cadenas compuestas exclusivamente por caracteres no alfanuméricos, por ejemplo '???!!!!'), por lo que esos casos fueron eliminados y no serán considerados para el siguiente análisis.

Luego, usando la herramienta antes mencionada, se creó un dataframe donde sólo existieran las palabras que fueran reconocidas como formato valido para 'Penn tree banking'. Una vez hecho esto, se separaron las palabras en los siguientes grupos:

- Verbos.
- Pronombres.
- Sustantivos.
- Adverbios.
- Adjetivos.
- Palabras restantes que sean reconocidas como tales.

Finalmente, se procedió a confeccionar el siguiente gráfico:

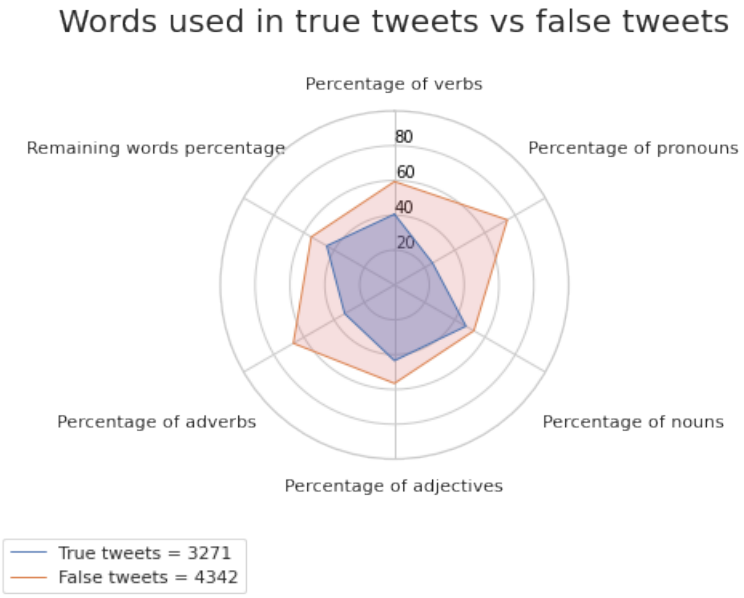

Figura 26: Tweets reales y falsos: distribución del tipo de palabras que los conforman

En primer lugar, se puede ver que hay una cantidad mayor de tweets que son falsos, por lo que la distribución obtenida para este tipo de tweets es más precisa que la de los reales. No obstante, la cantidad de tweets reales es numerosa, por lo que estos datos permiten fundamentar ciertas diferencias drásticas que hay en ambas distribuciones.
Se estimó que si tuviéramos una muestra mayor de tweets verídicos, que no cambiaran las condiciones de análisis, para explorar, entonces habría más sustantivos en sus pero la proporción de pronombres y adverbios seguiría siendo baja

Viendo este gráfico, se puede llegar a la conclusión de que para los tweets verídicos predominan los sustantivos mientras que, para los tweets no verídicos, predominan los adverbios y pronombres.

No se puede afirmar ni negar la relación de los tweets y su veracidad con respecto al resto de tipos de palabras, ya sean adjetivos, verbos o el resto. Esto se debe a que no hay una diferencia tan grande entre ambos como para justificar marcar una tendencia.

5. Conclusiones

Lo primero a mencionar es el contenido de los tweets analizados. Al analizar qué tipo de keywords fueron las más recurrentes se encontró una tendencia a publicar temas relacionados a eventos catastróficos. En particular, los desastres naturales eran los más recurrentes, y se encontró que algunos de ellos eran mencionados más frecuentemente en tweets reales (debris, typhoon, wildfire) y otros, en falsos (blizzard, twister, avalanche).

Luego, se exploraron los hashtags más usados y se observó que varios de ellos estaban relacionados al mundo del periodismo (news, prebreak). Explorando la veracidad de los tweets que contenían estos tags, se notó que casi la mitad de ellos eran falsos. Para profundizar en esto, se analizaron los hashtags más usados en los tweets reales y se encontró que el más usado era 'news' y algunos de ellos eran desastres naturales (earthquake, hiroshima).

Finalmente, se investigó sobre la mención de usuarios y se encontró que los tweets que mencionaban usuarios relacionados a canales de noticias (Fox News, por ejemplo) eran, en general, reales. Sin embargo, Youtube o aquellos que fueran celebridades (Justin Bieber o Ariana Grande) están más relacionados a tweets falsos.

Teniendo en cuenta todo lo mencionado anteriormente, una primera conclusión es que Twitter se utiliza como medio para publicar noticias de eventos catastróficos y, en particular, desastres naturales. Esto se ve reflejado tanto en las keywords utilizadas, las cuales se relacionan con estos eventos, como en los hashtags y en los usuarios mencionados, los cuales se corresponden con el mundo periodístico.

También se puede agregar que gran parte son fake news y, a partir de lo investigado sobre los desastres naturales, se puede deducir que gran parte de las noticias falsas estarán relacionadas a desastres como blizzard, twister, avalanche; mientras que las verdaderas a debris, typhoon y wildfire. A su vez, se podría relacionar esto con el hecho de que en muchos tweets falsos se etiquetan celebridades, ya que al etiquetar influencers los tweets cobran más impacto y podrían volverse virales. De esta forma, se podrían propagar noticias falsas tanto de desastres naturales, tópico recurrente, como de las mismas celebridades, ya que es recurrente que se difundan detalles inventados de la vida privada de los famosos.

Por otra parte, como el usuario más etiquetado es Youtube y el mismo está presente en una cantidad de tweets falsos considerable, se indagó en los hashtags más presentes en los tweets falsos y se encontró que el más utilizado es nowplaying. Este tag se utiliza, en general, para compartir música, por lo que otra posible conclusión es que muchos tweets falsos no son más que bots utilizados por cantantes o grupos musicales para viralizar ciertas canciones entre los usuarios.

Otros tags interesantes que son frecuentemente utilizados en tweets falsos son 'hot' y 'job/jobs'. Esto puede deberse a que son utilizados en tweets diseñados para estafar a la gente, ya se con perfiles falsos de mujeres o con falsas propuestas de trabajo.

En cuanto a las ubicaciones analizadas, se puede concluir que Estados Unidos, al menos en el data set analizado, tiene una gran relevancia en Twitter, ya que una cantidad considerable de tweets proviene o menciona a este país. Sin embargo, estos tweets no tienen un nivel de veracidad tan alto comparado con otros países como India o Japón, ya que pudo observarse que la proporción real-falso es de 65 % y 35 % aproximadamente. También hay que tener en consideración que no hay tantos datos sobre otros países, por lo que se podría inferir que obteniendo más datos sobre tweets relacionados a otros países, se obtendría un análisis más preciso con respecto a la veracidad.

En la búsqueda de patrones de los tweets según su longitud, se vio que los mismos tienden a seguir un mismo patrón cuando sus longitudes se categorizan como pequeños o medianos, por ejemplo para los pequeños hay una menor cantidad de links y hashtags, y mayor cantidad de usuarios etiquetados, para los medianos esta tendencia cambia y comienzan a haber más links y hashtags que usuarios etiquetados. Para los tweets categorizados como largos no ocurre lo mismo, sus valores se dispersan de campo a campo y no se ve una tendencia tan marcada como para los casos anteriores. Con respecto al porcentaje de veracidad, se apreció que los tweets pequeños son los que poseen los porcentajes de veracidad mas bajos y no logran superar el 40 %, los tweets medianos no poseen porcentajes tan bajos como los pequeños y hay una tendencia marcada en torno al 40 %, y por último, la tendencia del porcentaje de veracidad de los tweets largos es levemente superior a los medianos (se aprecia que es mas marcada entre el 40 % y 50 %) y son los que suelen alcanzar los porcentajes más altos de veracidad.

6. Bibliografía

- https://www.marketingdirecto.com/digital-general/social-media-marketing/como-usar-hashtags-en-twitter-una-guia-sencilla-para-los-marketeros
- Discourse of Twitter and Social Media: How we use language to create affiliation on the web - Michele Zappavigna
- Mastering Geospatial Analysis With Python - P. Crickard, E. Van Rees, S. Toms
- Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.
- https://www.infobae.com/america/vice/2018/04/03/que-hay-detras-de-los-perfiles-falsos-de-chicas-sexys/
- https://blogs.imf-formacion.com/blog/recursos-humanos/busqueda-de-empleo/ofertas-falsas-de-trabajo-o-fake-jobs-consejos-para-detectarlas-y-como-actuar/
- https://simplemaps.com/data/world-cities
- https://cs.nyu.edu/grishman/jet/guide/PennPOS.html
- https://www.eleconomista.es/internacional/noticias/10488441/04/20/Europa-se-replantea-su-posicion-sobre-China-por-su-falta-de-transparencia-ante-el-virus-y-su-actitud-con-la-crisis.html
- https://www.infobae.com/america/eeuu/2020/04/10/el-engano-de-china-al-mundo-fue-peor-pruebas-muestran-que-el-coronavirus-ya-existia-desde-antes-de-diciembre-y-el-regimen-lo-oculto/
- https://www.reuters.com/article/us-huawei-tech-usa-pompeo/pompeo-says-huawei-ceo-lying-over-ties-to-china-government-cnbc-idUSKCN1ST1EF
- https://encyclopedia.ushmm.org/content/es/article/ss