FACULTAD DE INGENIERIA
Universidad de Buenos Aires

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ingeniería
Departamento de Computación

ORGANIZACIÓN DE DATOS (75.06)

Cuatrimestre y año: 1C 2020
Grupo: Superficial intelligence

Integrantes:

Padrón	***Nombre***	***Email***
104288	Maximiliano Levi	mlevif@fi.uba.ar
101168	Damian Ganopolsky	dganopolsky@fi.uba.ar
104342	Andrés Jalife	ajalife@fi.uba.ar
92190	Jorge Sandez	jsandez@fi.uba.ar

Link al repositorio de GitHub:
https://github.com/maxilevi/tp1-datos

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Índice

1. Introducción

2. Análisis General
   - *2.1. Keywords*
     - 2.1.1. Keywords nulas
     - 2.1.2. Análisis de la distribución
     - 2.1.3. Relación con la veracidad
   - *2.2. Ubicación*
     - 2.2.1. Tweets con ubicación
       - 2.2.1.1. Longitud de la ubicación y la veracidad
       - 2.2.1.2. Ubicaciones no válidas
       - 2.2.1.3. Ubicaciones más frecuentes
     - 2.2.2. Tweets sin ubicación
   - *2.3. Texto*
     - 2.3.1. Longitud del texto
     - 2.3.2. Apariciones de
       - 2.3.2.1. Palabras
       - 2.3.2.2. Hashtags
       - 2.3.2.3. Mayúsculas
       - 2.3.2.4. Caracteres Numéricos
       - 2.3.2.5. Arrobas
       - 2.3.2.6. Signos de pregunta
     - 2.3.3. Veracidad en el uso de links

3. Conclusiones

1. Introducción

La idea de este análisis es descubrir aspectos interesantes dentro de un set de datos en el que se encuentran una variedad de tweets que alertan sobre posibles desastres que están presenciando.

Este set de datos consta de 4 tipos de información:

- Texto (el tweet en sí)
- Ubicación del supuesto desastre
- Keyword
- Target (veracidad del tweet, representado con 1 (uno) si es verdadero y 0 (cero) en otro caso)

Nosotros vamos a examinar hasta el más mínimo detalle para tratar de deducir cuando un verdadero es verdadero, y cuando un falso efectivamente lo es.

Esta investigación se divide en distintas etapas, una por cada tipo de información y también relacionando los tipos de información entre sí.

Cabe aclarar que el set de datos trae consigo aproximadamente 4300 tweets falsos y 3300 tweets verdaderos, por lo que hay una mayor proporción de tweets falsos, teniendo esto en cuenta, pasamos a analizar los datos.

Dicho esto, esperamos que les resulte interesante y sencillo de entender dado que ese fue uno de nuestros objetivos durante su realización.

2. Análisis General

2.1 Análisis sobre la columna keyword

Keywords nulas

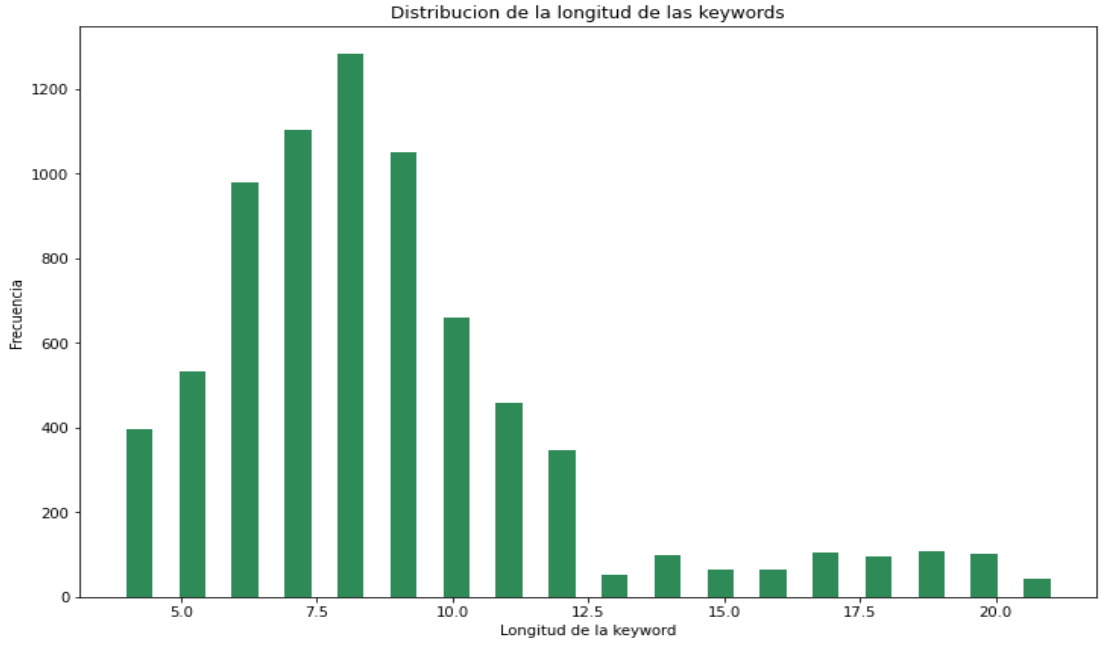
Para este análisis, lo primero que se realizó fue la cuenta de los elementos nulos totales en la columna, y se llegó al resultado de que se trataba solamente de 61.

Se puede pensar en buscar una relación entre los elementos nulos y la veracidad de los tweets, pero al tener una cantidad nula tan pequeña no tenemos una muestra lo suficientemente significativa como para abstraer una conclusión realmente verdadera. Esto es así debido a la "ecuacion mas peligrosa de la historia", igualmente si todos los tweets o la gran mayoría tuviera target 1, esto me daría un posible indicio de que hay una relación. Es por esto, que igualmente se calculó la media del target, siendo esta 0.68 por lo que se ve que no hay una relación directa entre únicamente estos dos factores.

Análisis de la distribución de variables

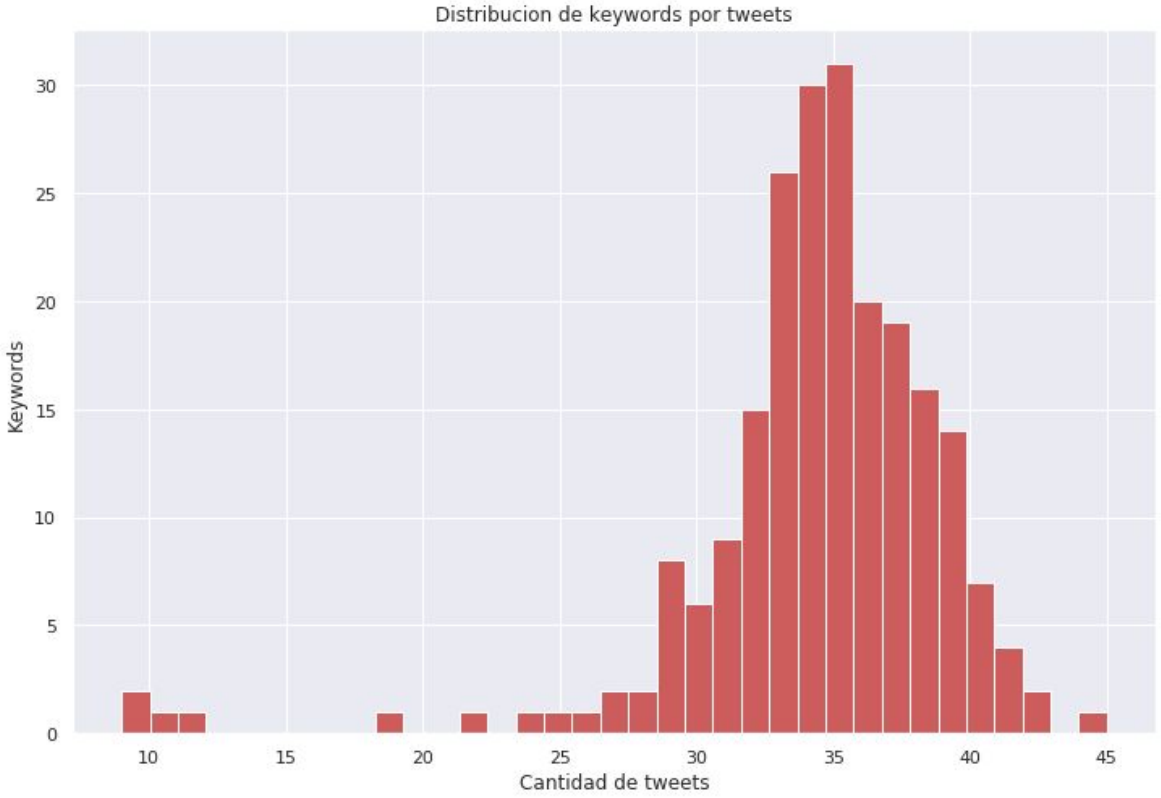
Una incógnita a resolver para tener un mayor entendimiento de estos datos fue saber cómo estaba distribuida la longitud de las keywords.

Para analizar esta distribución se creó una columna con la longitud de la keyword de la fila, para luego analizar la distribución usando un histograma.



Se puede ver que esta longitud está acotada entre 4 y 21 aproximadamente por ser únicamente una palabra. La longitud más frecuente es de 8 caracteres y se puede ver como mientras más próximo se esté a este valor, más frecuencia tiene. Además, se calculó la media de la longitud siendo esta 8,73.

Luego, para lograr un mayor entendimiento sobre la columna se vio que la cantidad de keywords era 221 por lo que se pensó en un histograma para visualizar los datos, ya que esa cantidad era demasiado grande para ver en un Bar Plot.



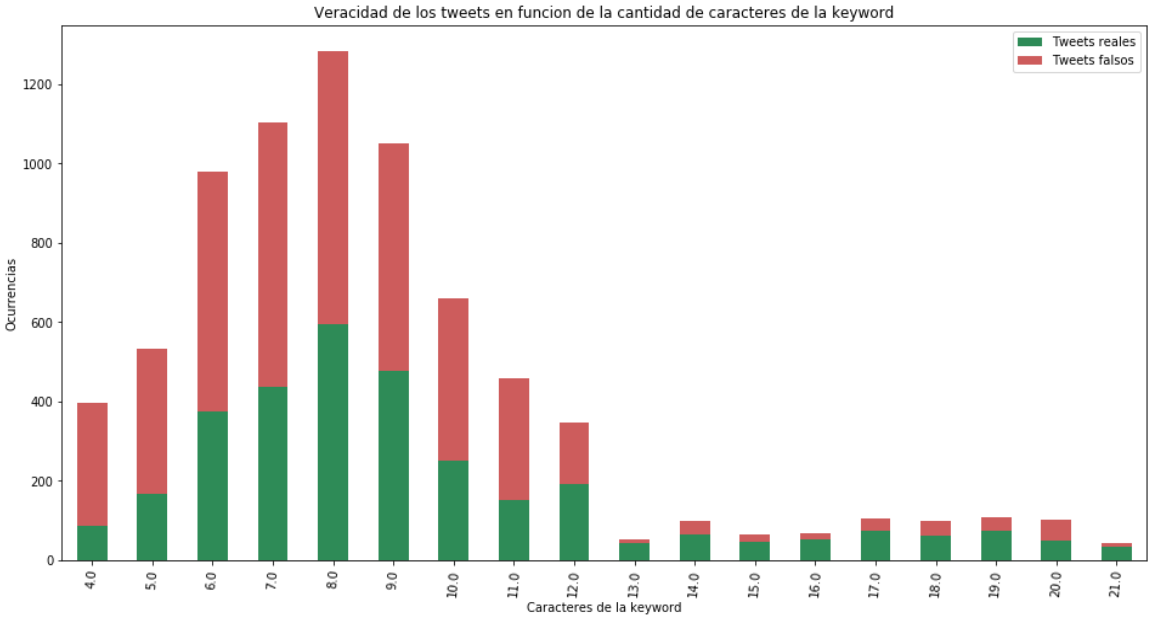
Al visualizar la distribución, se observó que la cantidad de tweets por cada keyword varía entre 10 y 45 aproximadamente. Además, se puede ver que la mayoría de las palabras clave cuenta en promedio con 30 a 40 tweets, siendo el promedio exacto 34,17.

Relación con la veracidad

Una cuestión muy importante que se trató para el análisis de la columna y que generó muchas preguntas, fue saber que muchas palabras clave se usan de manera cotidiana metafóricamente. Y, al ser las palabras claves relacionadas con catástrofes las que desatan que la computadora “piense” que son verdades, se trató de buscar dependencias entre estas keywords y su grado de veracidad.

La primera pregunta que nos planteamos fue, ¿Hay alguna relación entre el largo de la keyword y la veracidad?. Nos realizamos esta pregunta al pensar que no debe ser tan común usar palabras muy largas para darles una connotación no real, y puede que también se aplique la inversa.

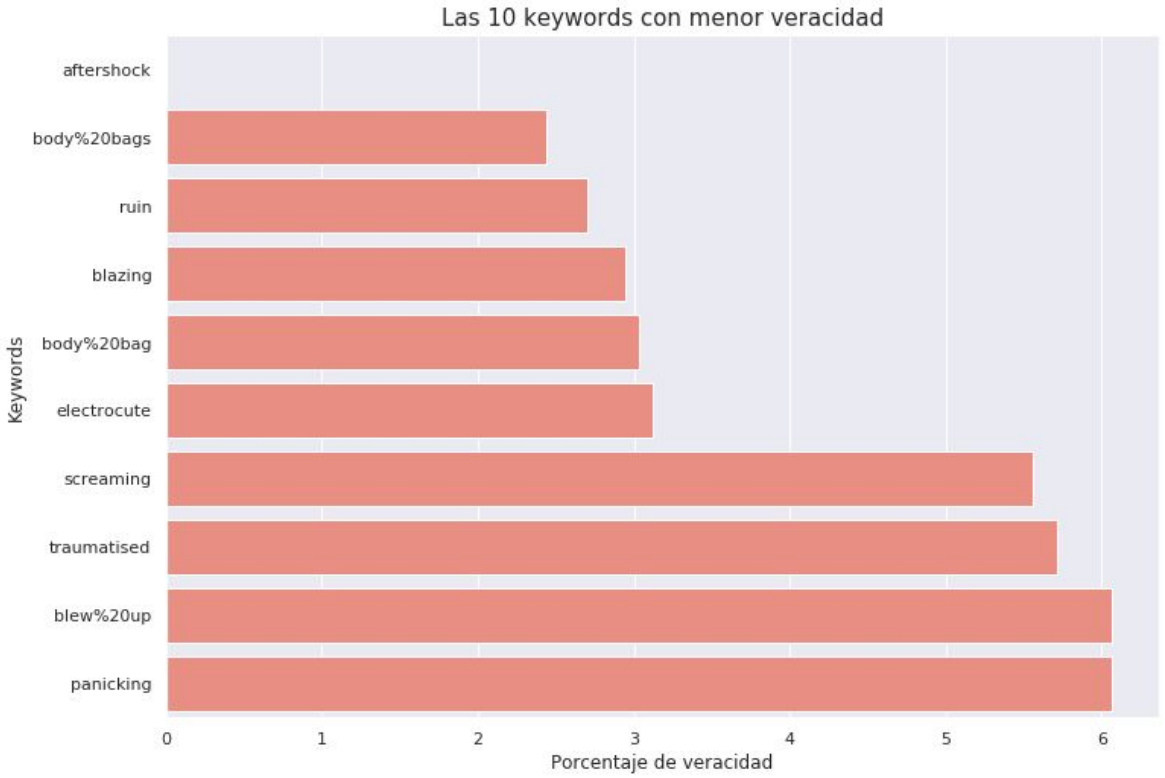
Para resolver esta pregunta, se separó en grupos por longitud de la keyword, contando la cantidad de ocurrencias de sucesos(si es real o no) para cada cantidad de caracteres. Luego, se graficó en un 'stacked bar plot' la cantidad de tweets falsos sobre la cantidad de tweets reales.



Se puede ver que en donde la muestra es mayor(caracteres de la keyword=8.0) la proporción de veracidad se encuentra bastante balanceada, se calculó la media para dicho caso y es 0.463, ligeramente distinta de 0,5.

Si nos centramos en las keywords con una cantidad de caracteres mayor a 13 podemos observar que proporcionalmente son mayores los casos en los que los tweets son reales. Pero, también notamos que la cantidad de ocurrencias es significativamente menor por lo que al ser una cantidad de keywords pequeña el error no es tan despreciable. Debido a esto, no se puede tener una conclusión certera de que al tener mayor longitud en la keyword hay mayor veracidad, pero para la pequeña cantidad analizada se podría decir que si.

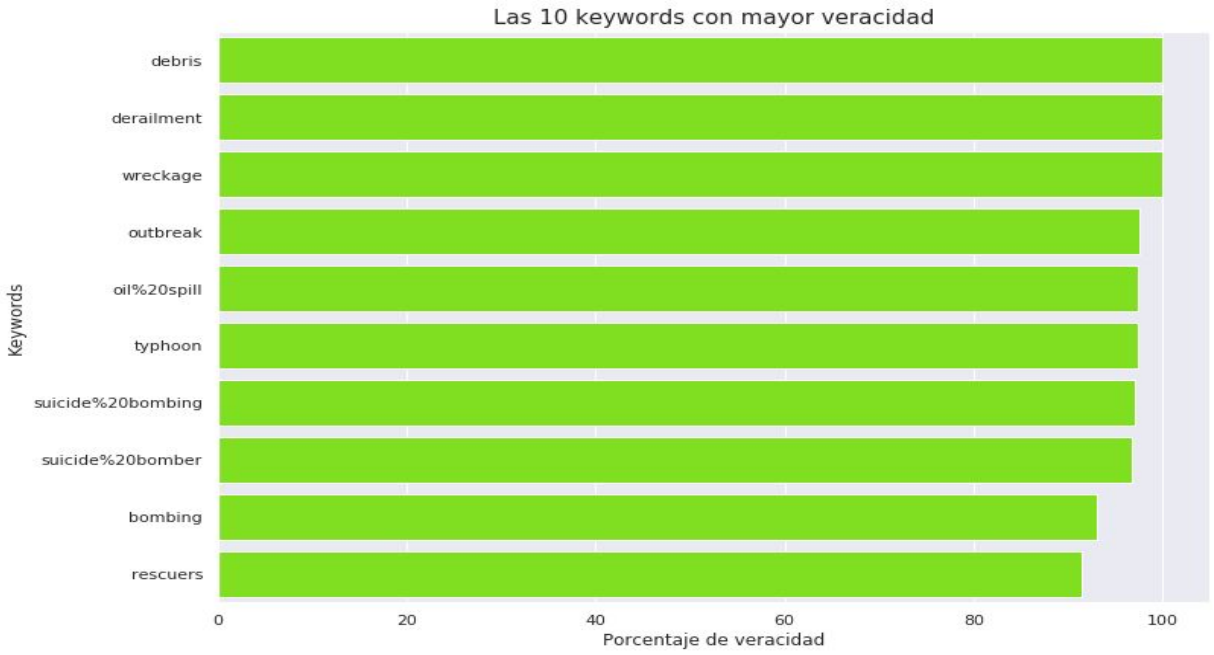
También, se analizaron los tweets cuyas keywords tienen mayor y menor promedio de veracidad. Primero, se visualizaron las keywords que tenían menor porcentaje de veracidad.



A primera vista, se pudo notar que las 10 keywords que contienen el menor porcentaje de veracidad no superan el 6%. Además, se encontró que 'aftershock' tiene un 0% de veracidad en todos los tweets del dataset, por lo que posteriormente se buscó si había una explicación, ya que, la palabra puede significar el temblor posterior a un terremoto.

Al analizar los tweets con esa keyword, se vio que todos la usaban de forma metafórica. Se puede desprender de estos resultados, que 'aftershock' al ser una keyword muy específica, no tiene un uso habitual para hechos catastróficos debido a que el hecho al que se refiere sucede en casos muy excepcionales .

Luego, se analizaron las 10 keywords con mayor porcentaje de veracidad



Al visualizar estos datos, se encontró la keyword “outbreak”, la cual también es usada frecuentemente para denotar hechos no catastróficos. Debido a esto se vieron los tweets a los que hace referencia, y se observó que casi la totalidad de los tweets se refieren a la “Enfermedad de Legionario” la cual al buscar las fuentes se pudo ver que la enfermedad surgió de un brote repentino.

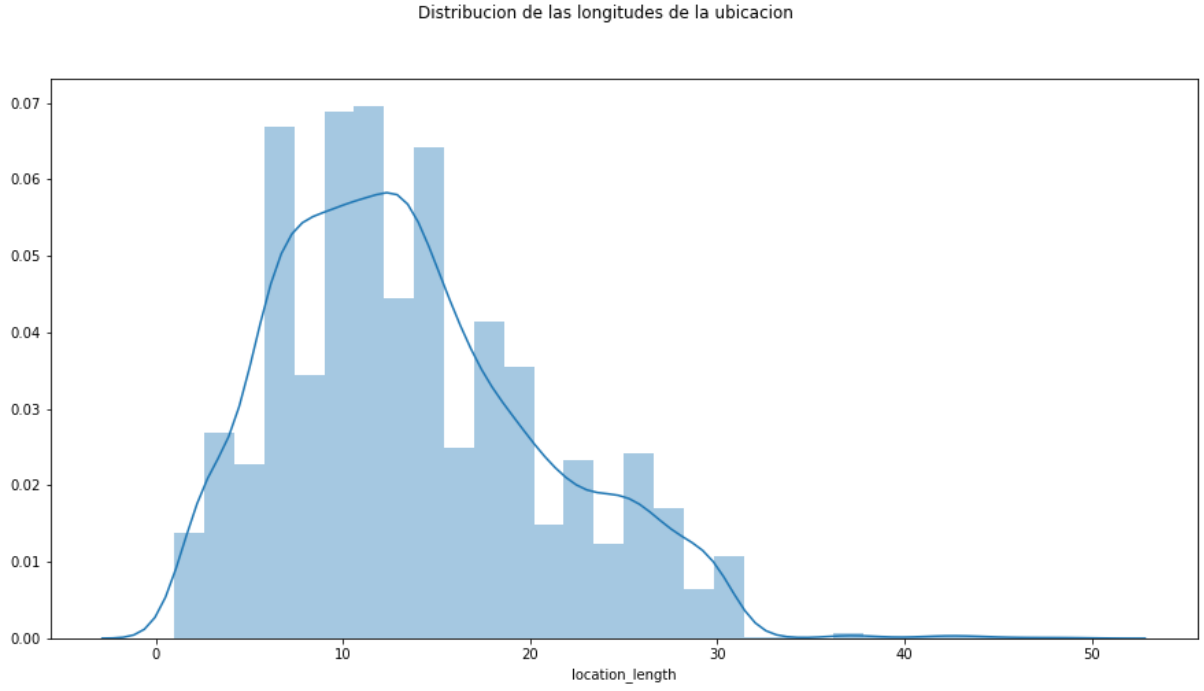
Además, se analizó el promedio de veracidad de la palabra clave ‘quarantine’ para poder saber en qué magnitud la situación del momento puede condicionar la veracidad de los tweets. Obtuvimos como resultado una veracidad media del 27%, por lo que abstraemos la conclusión de que los tweets no son actuales y que el momento del análisis del set de datos está fuertemente relacionado con la veracidad de las distintas keywords.

2.2 Análisis sobre la ubicación de los tweets

Para el análisis de la ubicación de los tweets primero de todo se analizó el estado de los datos. De los 7613 tweets que se encuentran en el dataset solamente 2533 no tenian ubicacion y 5080 si contenían. De acá el análisis de dividió en 2 etapas.

Tweets con ubicación

De los tweets que contenían ubicación primero se quiso ver cómo era la distribución de longitudes de la ubicación, es decir en qué rangos de longitud de caracteres se encontraban los strings que el tweet tiene como ubicación. Para esto se grafico un histograma donde el y-axis es la cantidad de tweets que se encuentran en ese bucket y los buckets del x-axis representan la longitud de las ubicaciones.

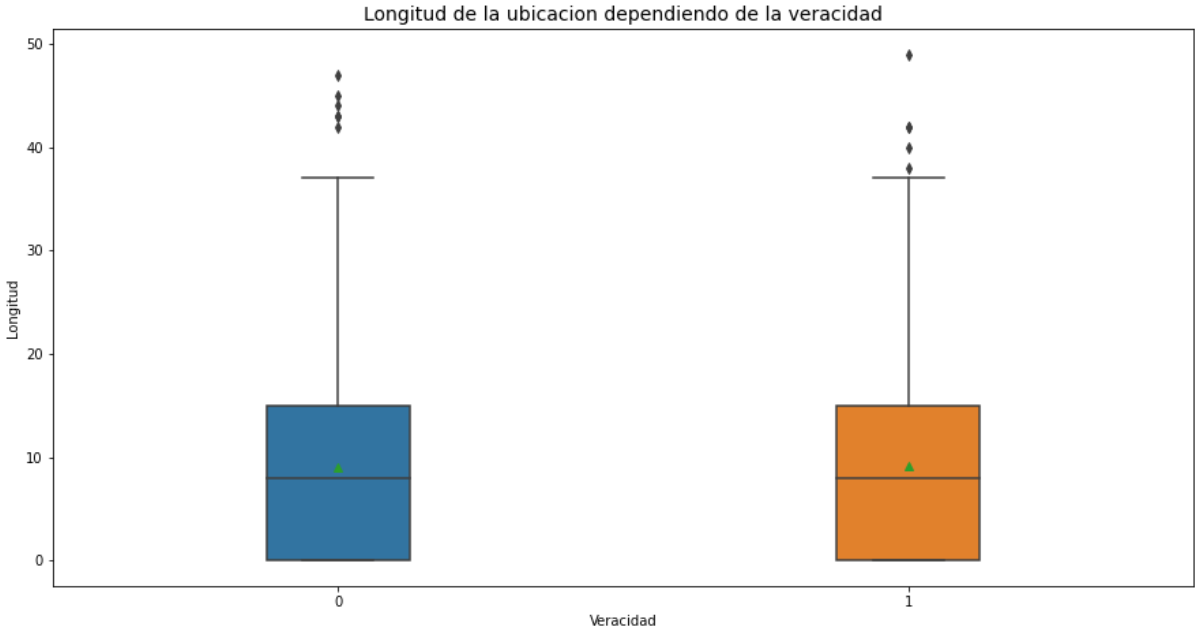


Como podemos observar las mayoría de las longitudes de las ubicaciones se encuentran en el en el rango [5, 20] mientras que muy pocas superan los 35 caracteres. La mayoría se encuentran cercano a los 10 caracteres y la cantidad se reduce bastante luego de los 20 caracteres.

Longitud de la ubicación y la veracidad

Luego del grupo de tweets con información de ubicación se separó en 2 subgrupos más dependiendo de la veracidad del tweet y de estos se analizó información sobre la longitud del keyword de ubicación.

También se grafico un boxplot para poder visualizar mejor los promedios, y los límites mínimos y máximos, también como los 25% y 75%



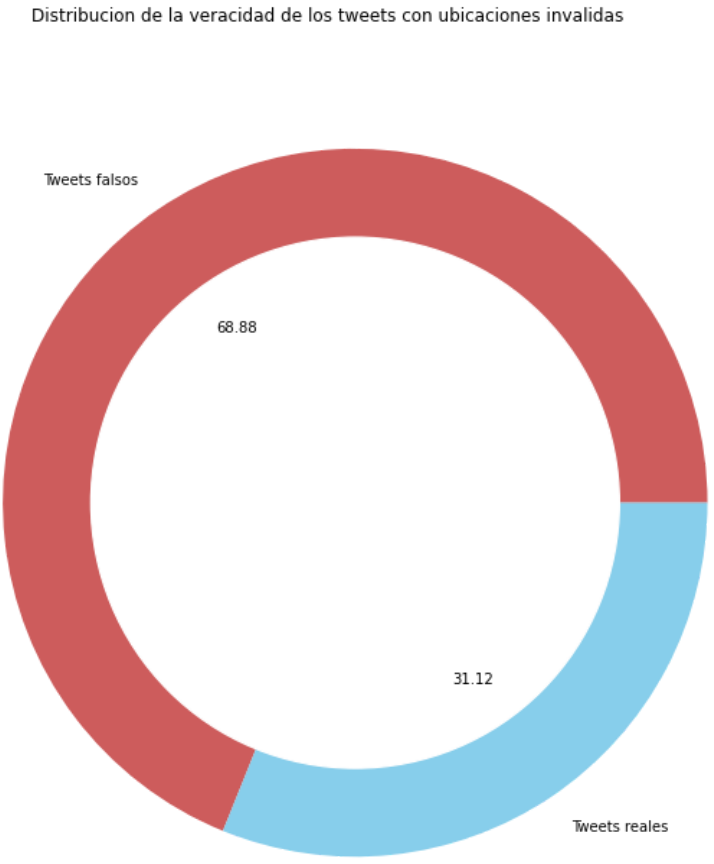
Para los tweets falsos el promedio de longitud fue de 13.63 caracteres, mientras que el mínimo fue de 1 y el máximo de 47. Por el otro lado los tweets verdaderos tuvieron un promedio de longitud de 13.65 caracteres, mientras que el mínimo fue de 1 carácter y el máximo de 49, algo muy similar a los tweets falsos.

Ubicaciones no válidas

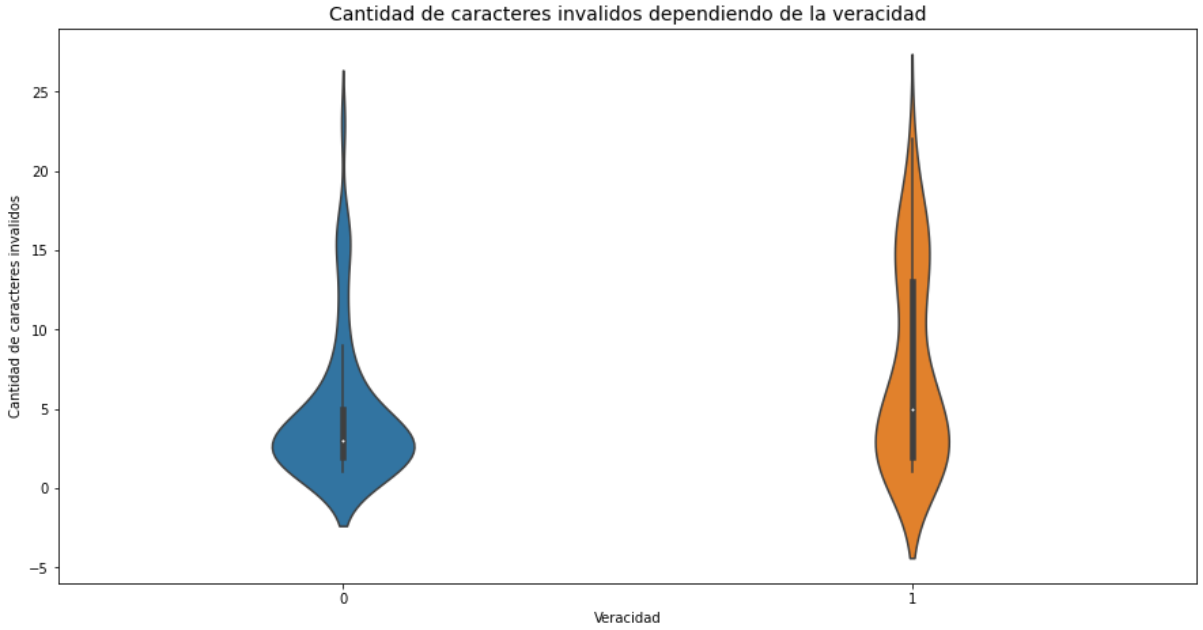
En esta sección nos concentramos en analizar tweets con ubicaciones no estándar, es decir ubicaciones que no son frecuente y que contienen caracteres normales para una ubicaciones. Estos caracteres inusuales puede ser números, signos de exclamación o pregunta entre otros. En este caso filtramos por ubicaciones que contengan uno de los siguientes (números, ?, ;, %, #, |, /, @, +, *, \, $, #) se podría ampliar para añadir más símbolos como (&, =, ¿, “) entre otros.

Se encontró que en el dataset hay 196 tweets con esta caracteristicas de los cuales la mayoría (el 68.88%) son falsos. Esto es una relación bastante significativa ya que 7 de cada 10 tweets que contienen una ubicación no estándar tienden a ser falsos.

Esta conclusión tiene sentido ya que ubicaciones como “304”, “1/10 Taron squad” o “?Gangsta OC / MV RP; 18+.?” tienen sentido en muy pocos contextos sintácticos como ubicaciones y parecen ocurrir en tweets con baja cohesión sintáctica, lo cual no hace que sean falsos pero aumenta la probabilidad. Este fenómeno se puede apreciar mejor con el gráfico a continuación.



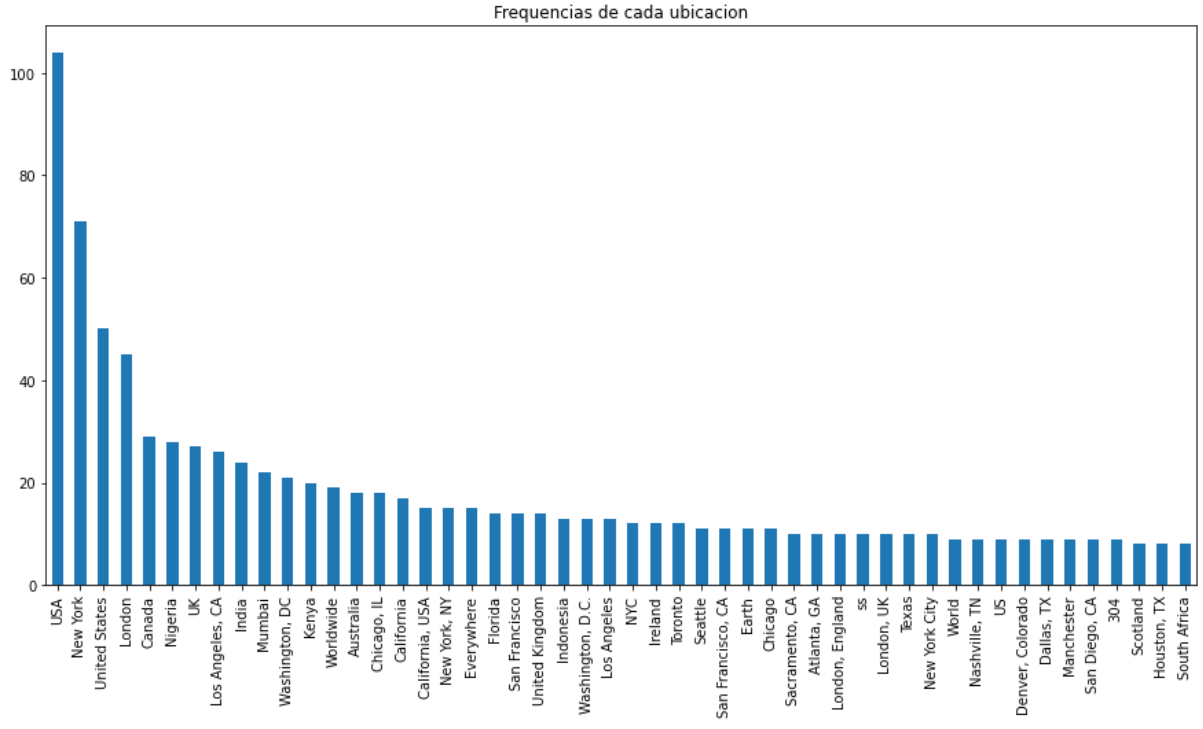
Por el otro lado, también se analizó la distribución de caracteres inválidos sobre los tweets, dependiendo de la veracidad, lo que nos lleva al siguiente violinplot



Como podemos ver en el gráfico, parecería que los tweets verdaderos tienen más de estos caracteres inválidos, mientras que los tweets falsos suelen tener una cantidad acotada [2,5] de caracteres inválidos

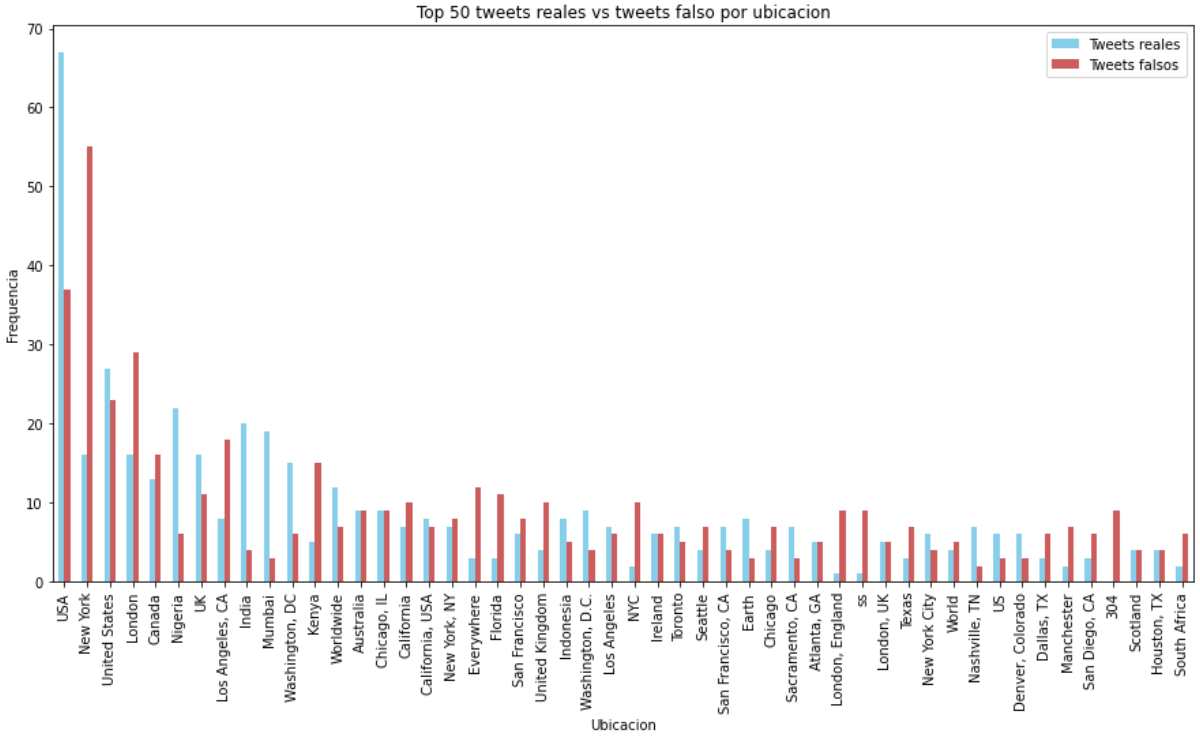
Ubicaciones más frecuentes

Para analizar las ubicaciones más frecuentes se armo un barplot con la cantidad de ocurrencias de cada ubicación.



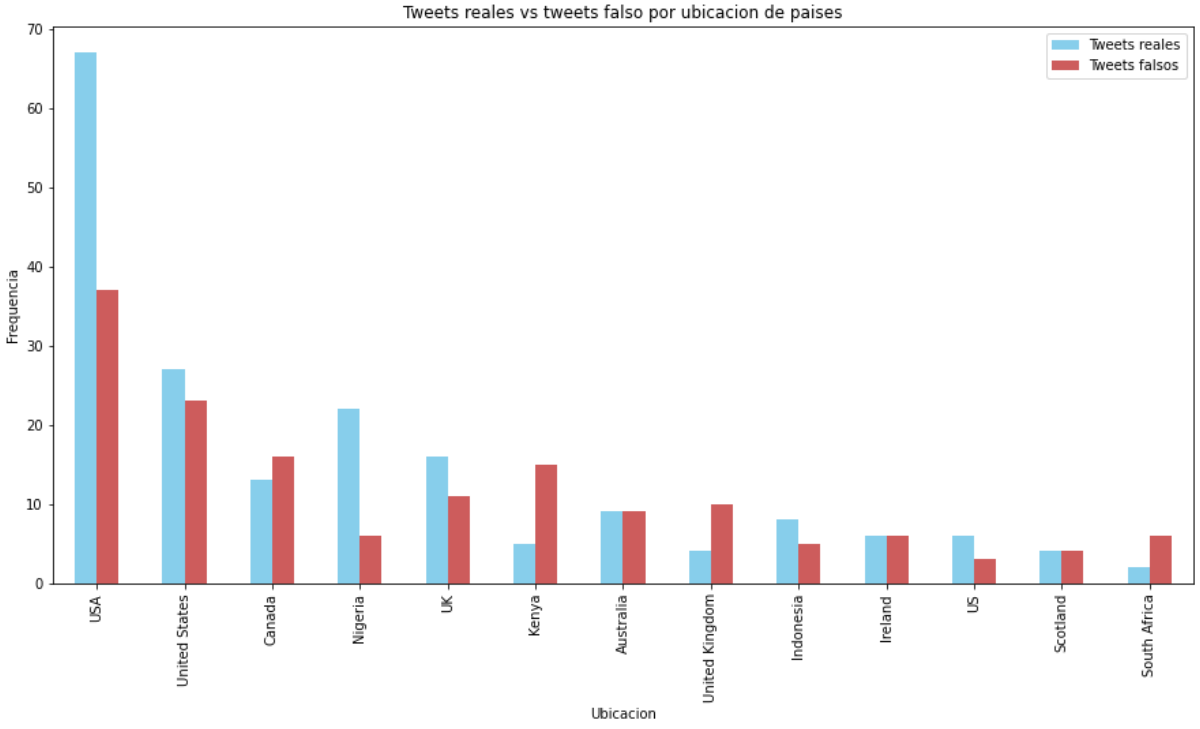
Como puede apreciarse, la mayoría de las ubicaciones se concentran en lugares como ciudades o estados de EEUU aunque también hay de otras partes del mundo como India o Canadá, pero se concentra mayormente en zonas donde la lengua inglesa predomina.

Por el otro lado, se buscó si existía algún tipo de relación entre la veracidad de un tweet y si su ubicación era alguna de las más comunes. Una ubicación es de las "más comunes" si se encuentra dentro de las top 50 más frecuentes. El siguiente barplot demuestra la relación entre ambos.



Si bien el gráfico muestra algunas anomalías como en el caso de India o Nigeria al analizarse en conjunto no puede determinarse ninguna relación.

Vamos a analizar el gráfico más profundamente viendo un subconjunto de las ubicaciones, los paises mas comunes como Nigeria o India. Esto nos debería dar una idea de si existe alguna relación en estas anomalías.

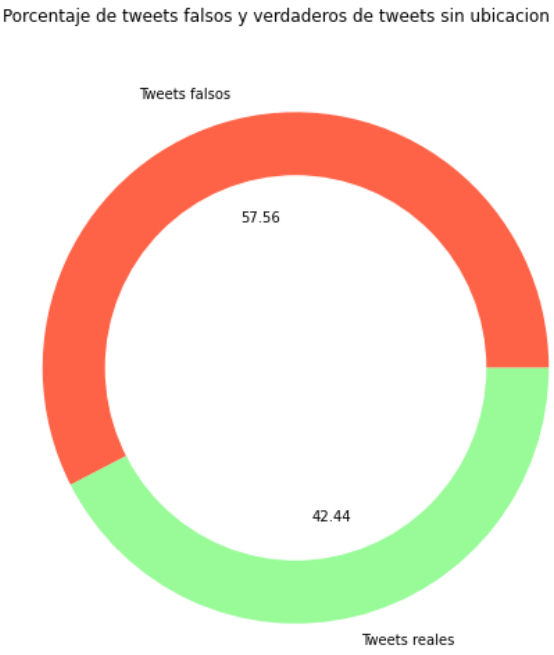


En este subgrupo no encontramos ninguna relación muy notable pero algo que es remarcable es que la veracidad cuando la ubicación se trata de países está bastante balanceada, no como en otros casos que se analizaron antes.

Tweets sin ubicación

De los 2533 tweets sin información sobre la ubicación que hay se tienen 1458 falsos y 1075 verdaderos, esto representa que el 57% de los tweets sin información de ubicación no son reales.

Se preparó un piechart que demuestra los porcentajes. Se preparó un pie chart que nos muestra la comparación de porcentajes.



Si bien no es un incremento significamente se podría teorizar que la falta de información sobre un tweet o incrementa las posibilidades de que el tweet no sea verdadero. En este caso muchos de los tweets que no tenían ubicación tampoco tenían una keyword en el dataset.

2.3 Análisis sobre el texto de los tweets

El análisis del texto se va a dividir en 4 secciones analizando cómo varían dependiendo de su veracidad.

Las cuatro secciones que elegimos son las siguientes:

- Longitud (Cantidad de caracteres por tweet)
- Apariciones de palabras
- Apariciones de hashtags
- Frecuencia de tipo de caracteres (mayusculas, numeros, etc)

Longitud

Primero, lo que se hizo fue analizar el promedio de las longitudes de los tweets falsos versus los tweets verdaderos.
Se puede observar que el promedio de longitud de los tweets verdaderos es un poco mayor (unos 12 caracteres).

	len
target	
0	95.706817
1	108.113421

Después vimos que la longitud máxima de caracteres en un tweet es 157, por lo que se decidió separar los tweets en 3 subgrupos dependiendo de su longitud.

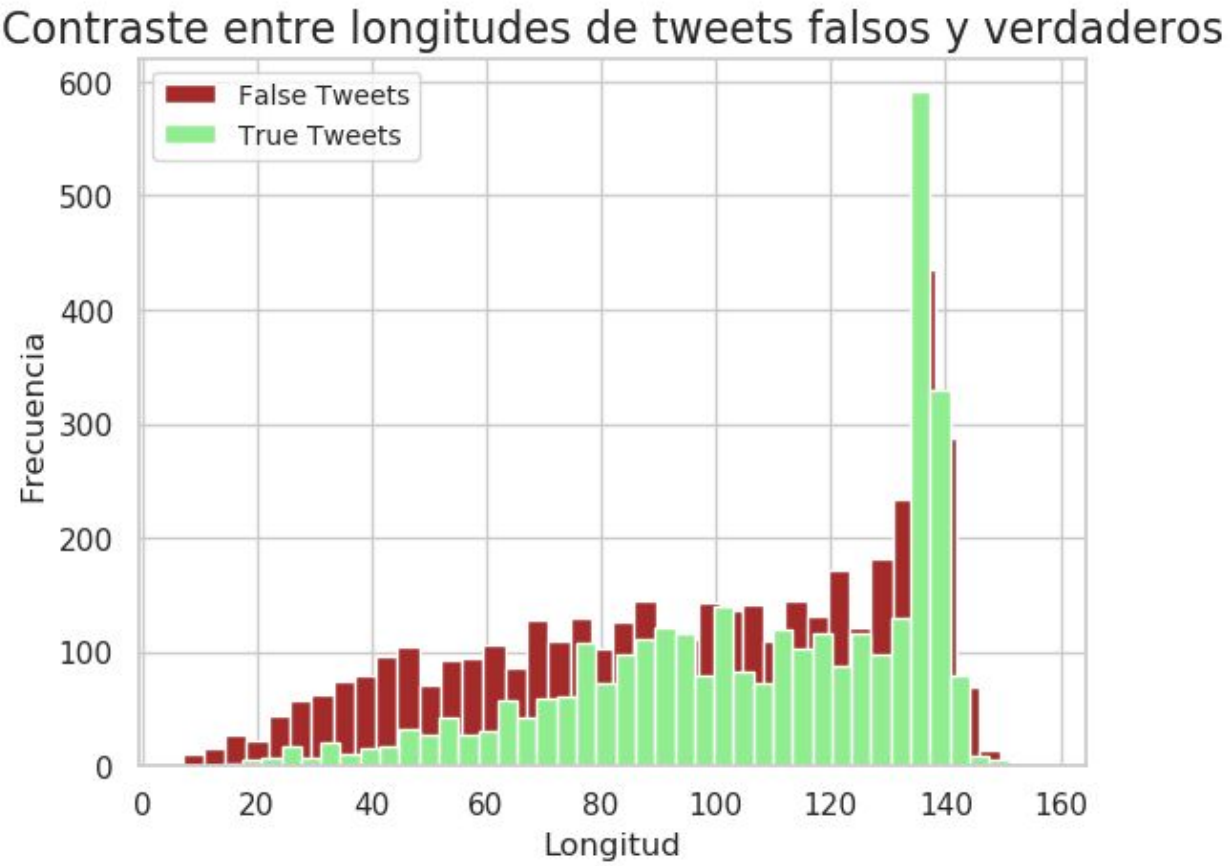
1. Longitud menor a 50 caracteres.
2. Longitud entre 50 y 100 caracteres.
3. Longitud mayor a 100 caracteres.

En el primer grupo, el cual está compuesto por 765 tweets (una baja cantidad en relación al total) se observó que predominaron los falsos notoriamente, con un resultado de 618 falsos y 147 verdaderos.

En el segundo grupo, ahora con una cantidad de tweets un poco más significativa (2522), nuevamente hubieron más tweets falsos que verdaderos. [1486 falsos y 1036 verdaderos]

Por último, en el tercer grupo es donde cayeron la mayoría de los tweets, con un total de 4219 tweets, en donde siguieron habiendo más tweets falsos (2180) que verdaderos (2039), pero esta vez la brecha fue notablemente menor.

Para poder entender estos resultados, y ver que no se nos haya escapado ninguna información interesante, vamos a visualizar los dos sets de datos (verdaderos y falsos) en un histograma superpuesto.

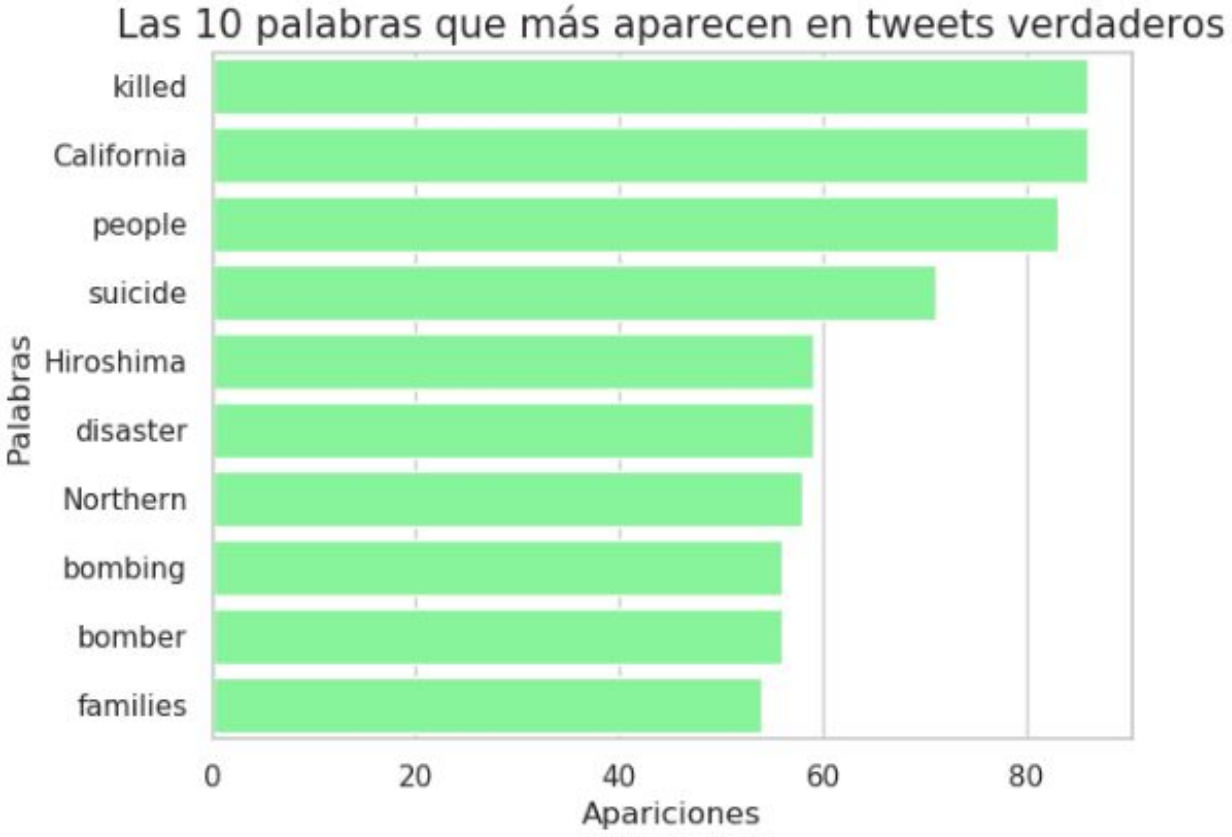


Viendo este gráfico podemos darnos cuenta como efectivamente en la mayoría de las longitudes menores se concentran más tweets falsos, pero vemos que en el rango de longitud [135-145] pareciera predominar el tweet verdadero, por lo tanto se analizó este rango y se observó que en efecto había más tweets verdaderos que falsos (811 tweets verdaderos y 716 tweets falsos), siendo este entonces, el rango de longitudes más "verídico".

Apariciones de palabras

En esta sección se va a comentar un poco lo que se notó al examinar las palabras que más aparecen, tanto falsas como verdaderas.

Separando entre tweets falsos y verdaderos, obtenemos los siguientes gráficos de las 10 palabras que más aparecen. Decidimos filtrar el resultado para las palabras de longitud mayor a 5 para quedarnos con palabras interesantes, y no únicamente proposiciones, conjunciones, pronombres, etc…



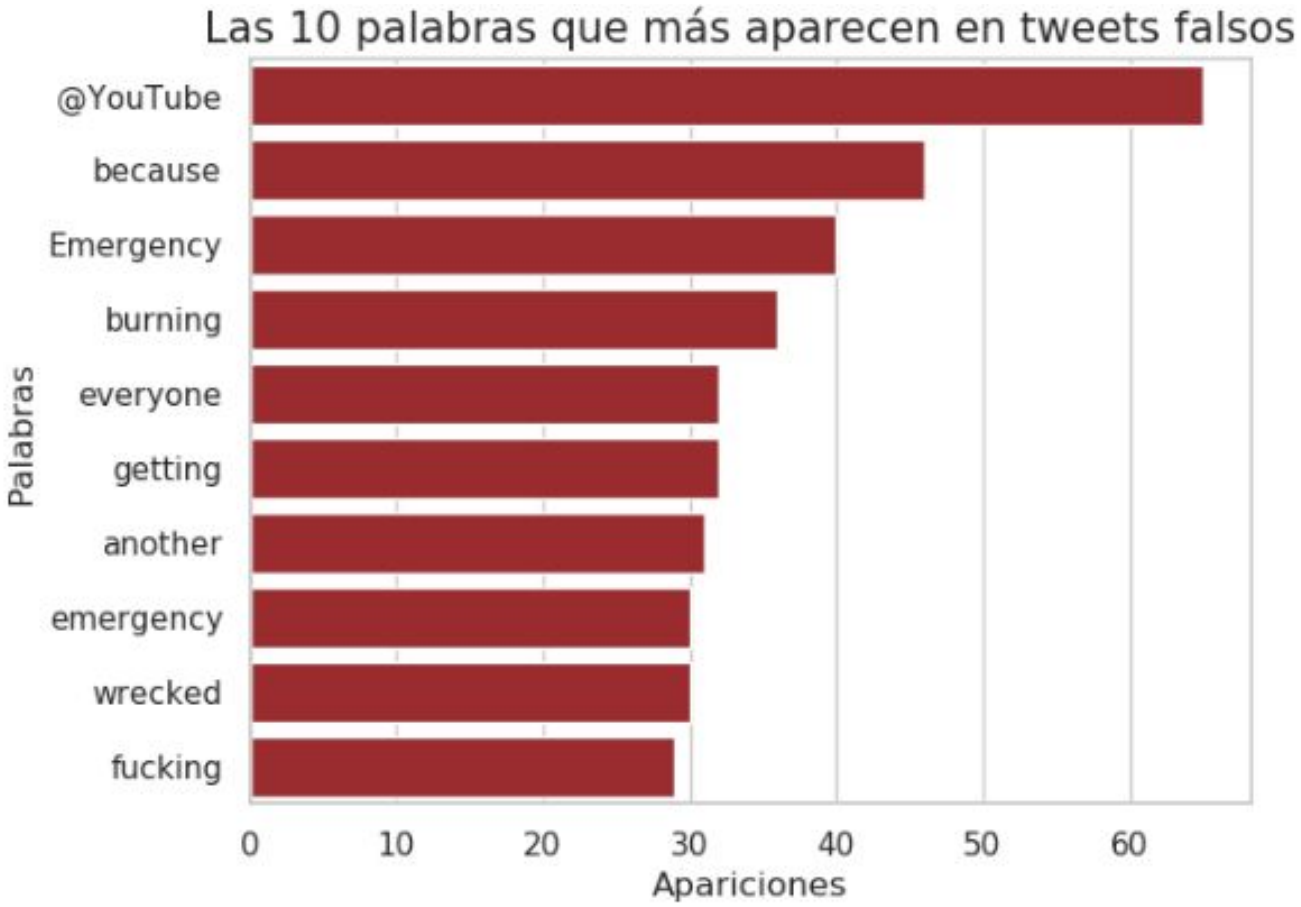
Viendo este gráfico de barras basado en tweets verdaderos, podemos ver que hay distintos tipos de palabras, lugares en los que pueden haber sucedido los desastres, tipos de desastres, entre otros. Pero todas estas tranquilamente pueden estar relacionadas a desastres.

Como vemos, las tres palabras que más se repiten son “killed”, “California”, “people” y “suicide”. Esto nos puede decir que dentro del set de datos hubo muchos desastres que

verdaderamente ocurrieron en California, y que también muchos pueden haber sido asesinatos.

En cuanto al resto de las palabras, siguen siendo una gran cantidad de apariciones (rondan las 50/60), y estas nos hablan principalmente desastres como suicidios y bombas.

Por otro lado observamos el gráfico de las apariciones de palabras falsas.



De aquí vemos que la palabra que más aparece es “Youtube”, esto ya nos da un parámetro de su veracidad dado que no es una fuente de información confiable.

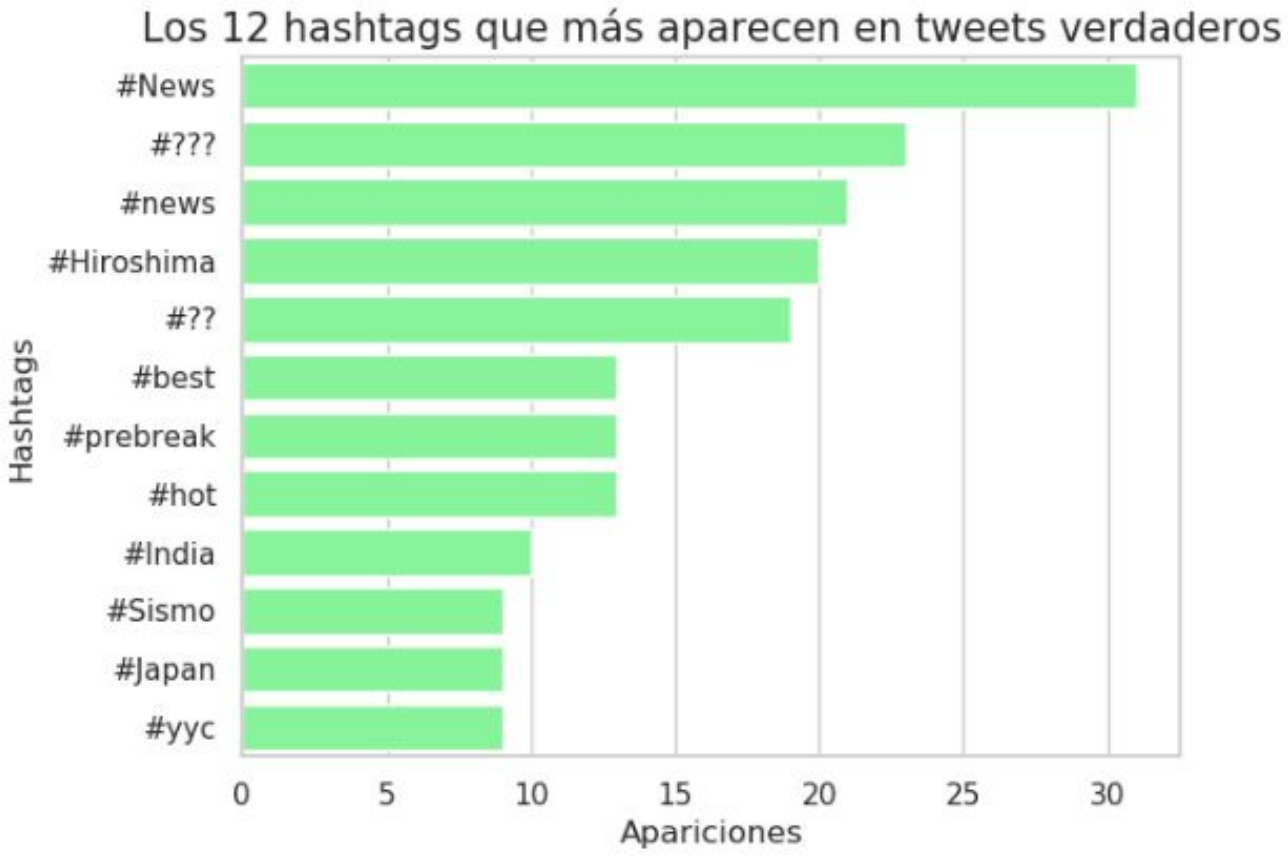
Por debajo de ella se encuentran otro tipo de palabras que no aportan mucho al análisis, se menciona bastante la palabra “emergencia”, y dado que se las menciona en tweets falsos, deben ser falsas emergencias. (¿O acaso nos encontramos bajo un caso de falsos negativos?)

Apariciones de hashtags

Ahora vamos a pasar a ver si encontramos algo que resulte atractivo en los hashtags que más aparecen en nuestro set.

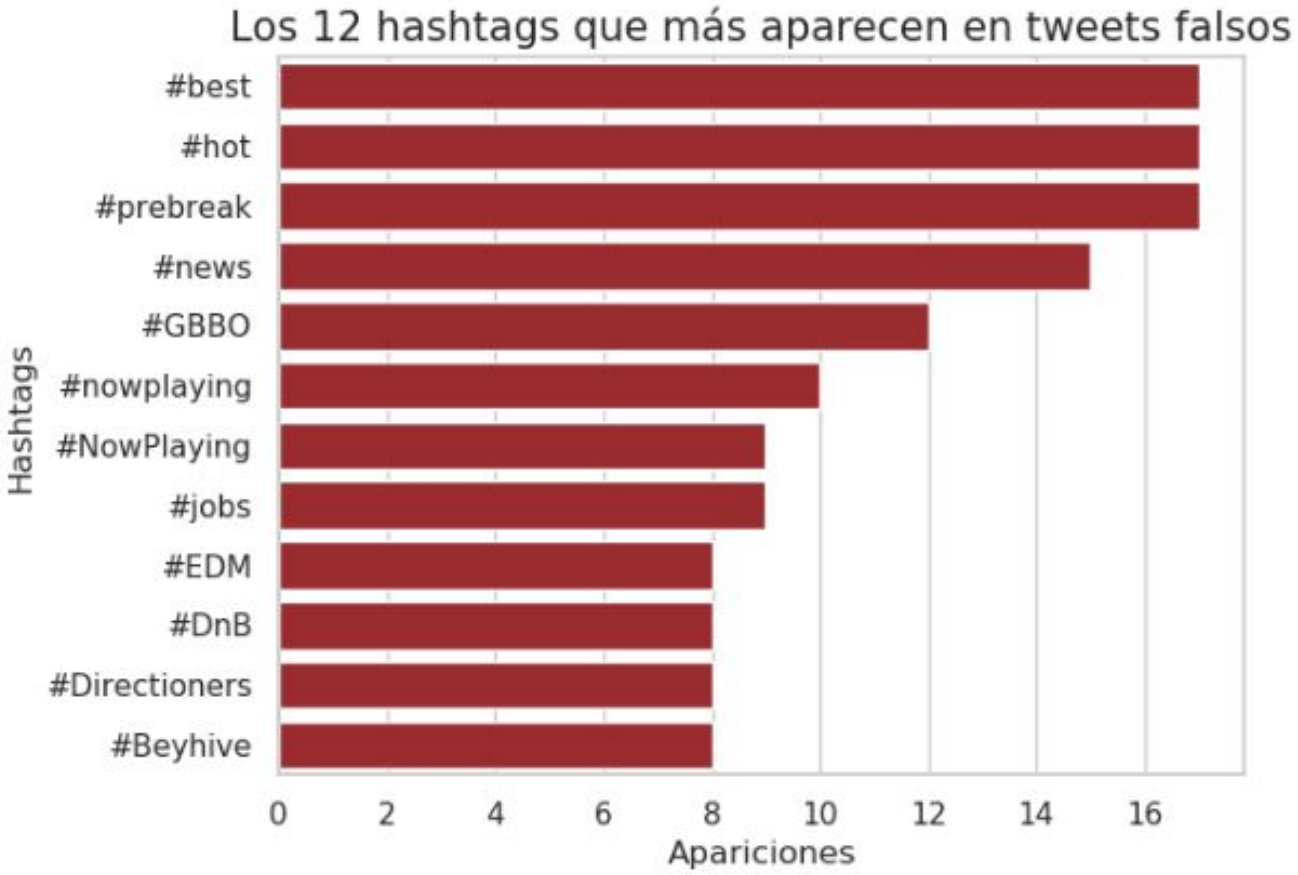
Como en los recientes análisis, vamos a separar los datos en verdaderos y falsos e examinarlos por separado.

Luego de procesar un poco la información, llegamos al siguiente gráfico que nos muestra los 12 hashtags que más se repiten.



Aquí observamos que el hashtag que más se usa es "#News" ocupando el primer y tercer puesto. Esto es entendible ya que los tweets hablan de noticias del momento. Un poco más abajo observamos el hashtag "#Hiroshima" que casualmente también es una de las palabras que más aparecen, por lo tanto debe haber sucedido algún desastre en esta ciudad japonesa. Al mismo tiempo "#Japan" puede estar hablando del mismo suceso.

Por otro lado, investigamos ahora los hashtags en tweets falsos.



Lo primero que se observa es que en los tweets falsos hay menos repeticiones de hashtags (observando el valor máximo de 17 repeticiones en contraste con las 32 apariciones verdaderas).

Por otro lado, analizando en sí los hashtags, el que más aparece es “#prebreak” que en la sección de hashtags de tweets verdaderos aparecía pero no le dimos mucha importancia. Para entender un poco más de que se trata este hashtag vamos a ver qué es lo que dicen.

	id	keyword	location	text	target	len
4391	6243	hijacking	perth, australia	#hot Funtenna: hijacking computers to send da...	0	119
4392	6244	hijacking	Mongolia	#hot Funtenna: hijacking computers to send da...	1	119
4393	6245	hijacking	brisbane, australia	#hot Funtenna: hijacking computers to send da...	0	119
4394	6246	hijacking	China	#hot Funtenna: hijacking computers to send da...	0	119
4396	6248	hijacking	Chiyoda Ward, Tokyo	#hot Funtenna: hijacking computers to send da...	0	119
4397	6253	hijacking	rome	#hot Funtenna: hijacking computers to send da...	0	119
4399	6255	hijacking	EastCarolina	#hot Funtenna: hijacking computers to send da...	0	119
4400	6256	hijacking	Brazil	#hot Funtenna: hijacking computers to send da...	0	119
4403	6259	hijacking	NaN	#hot Funtenna: hijacking computers to send da...	1	119
4404	6261	hijacking	France	#hot Funtenna: hijacking computers to send da...	0	119
4405	6262	hijacking	NaN	#hot Funtenna: hijacking computers to send da...	0	119
4407	6265	hijacking	tokyo	#hot Funtenna: hijacking computers to send da...	0	119
4408	6267	hijacking	china	#hot Funtenna: hijacking computers to send da...	0	119
4412	6272	hijacking	Brazil	#hot Funtenna: hijacking computers to send da...	0	119
4414	6274	hijacking	NaN	#hot Funtenna: hijacking computers to send da...	1	119
4415	6276	hijacking	Japan	#hot Funtenna: hijacking computers to send da...	1	119
4420	6283	hijacking	NaN	#hot Funtenna: hijacking computers to send da...	1	119
4453	6336	hostages	Japan	#hot C-130 specially modified to land in a st...	1	126
4454	6337	hostages	Las Vegas, NV	#hot C-130 specially modified to land in a st...	1	126
4461	6344	hostages	Tennessee	#hot C-130 specially modified to land in a st...	1	126
4462	6345	hostages	NaN	#hot C-130 specially modified to land in a st...	1	126
4475	6365	hostages	cuba	#hot C-130 specially modified to land in a st...	1	126

Cuando observan seguramente se preguntan ¿Qué es esto? Lo mismo nos preguntamos nosotros, parecieran ser varios tweets que dicen todos cosas parecidas y usan tanto el hashtag #Prebreak como #hot (nuestro segundo hashtag más usado en los falsos y el octavo en los verdaderos). Al mismo tiempo estos varían respecto a su veracidad, lo que nos da a pensar que deben ser datos erróneos, por lo que los vamos a descartar de nuestro análisis.

Continuando con los hashtags, observamos uno que dice “GBBO” que después de investigarlo descubrimos que es un programa de televisión llamado “The Great British Bake Off” así que algo debe haber sucedido en ese programa que generó este tipo de tweets.

Otro que aparece ahí es “Nowplaying” que muy posiblemente sean radios o reproductores de música que comentan que se está escuchando en el momento. Al mismo tiempo “EDM” es un tipo de música electrónica, así que estos dos hashtags podrían estar asociados.

Por último observamos dos hashtags curiosos [#Beyhive y #Directioners], y al ver sus tweets asociados encontramos lo siguiente.

Para #Beyhive:

	id	keyword	location	text	target	len
346	496	army	NaN	Beyonce Is my pick for http://t.co/nnMQlz91o9 ...	0	89
349	501	army	NaN	22.Beyonce Is my pick for http://t.co/thoYhrHk...	0	89
350	502	army	NaN	17.Beyonce Is my pick for http://t.co/thoYhrHk...	0	89
364	522	army	NaN	Beyonce Is my pick for http://t.co/nnMQlz91o9 ...	0	89
370	530	army	NaN	Beyonce Is my pick for http://t.co/nnMQlz91o9 ...	0	89
371	531	army	NaN	7.Beyonce Is my pick for http://t.co/thoYhrHkf...	0	88
372	533	army	NaN	Beyonce Is my pick for http://t.co/nnMQlz91o9 ...	0	89
373	535	army	NaN	6.Beyonce Is my pick for http://t.co/thoYhrHkf...	0	88

Y para #Directioners:

	id	keyword	location	text	target	len
347	498	army	NaN	One Direction Is my pick for http://t.co/q2eBl...	0	103
351	503	army	NaN	One Direction Is my pick for http://t.co/q2eBl...	0	103
355	512	army	NaN	Vote for #Directioners vs #Queens in the 5th r...	0	107
358	516	army	NaN	One Direction Is my pick for http://t.co/q2eBl...	1	103
360	518	army	NaN	One Direction Is my pick for http://t.co/q2eBl...	0	103
365	523	army	NaN	One Direction Is my pick for http://t.co/q2eBl...	0	103
367	526	army	NaN	One Direction Is my pick for http://t.co/y9Wvq...	0	97
378	543	army	NaN	One Direction Is my pick for http://t.co/q2eBl...	0	103
379	544	army	?	One Direction Is my pick for http://t.co/iMHFd...	0	97

Por lo tanto vemos como son elecciones de gente para lo que creemos que debe ser un concurso o algo por el estilo, y al investigar un poco más dimos con que Beyhive es el grupo de fans de Beyoncé, así como Directioners el de la banda One Direction.

Apariciones de mayúsculas

Dado que en la gramática se considera un correcto uso de mayúsculas para una buena lectura, se consideró interesante estudiar las apariciones de mayúsculas en los textos ya que podemos considerar que un tweet escrito con gran cantidad mayúsculas puede significar algún tipo de exclamación o de mala escritura y si encontramos un error gramático (no ortográfico) podemos pensar que dicha fuente no es fiable.

Como primer paso vamos a dividir el análisis entre tweets con al menos una mayúscula y con ninguna mayúscula. ¿Por qué hacemos esta distinción?

El motivo es que si consideramos que los tweets confiables son escritos por personas que tienen algún tipo de formación periodística, debería estar acostumbrado a escribir de manera correcta y empezar las palabras con mayúscula luego de un punto.

Por ende, la ausencia de mayúsculas también es señal de que algo no anda bien, a menos que la noticia sea de algún tipo de estadística, en dicho caso podría ser que empiece con un número y jamas use una mayúscula en su contenido.

Se encontró que de 7613 tweets que contiene el data frame, solo 278 no contienen ninguna mayúscula y diferenciando los verdaderos de los falsos, nos queda lo siguiente:

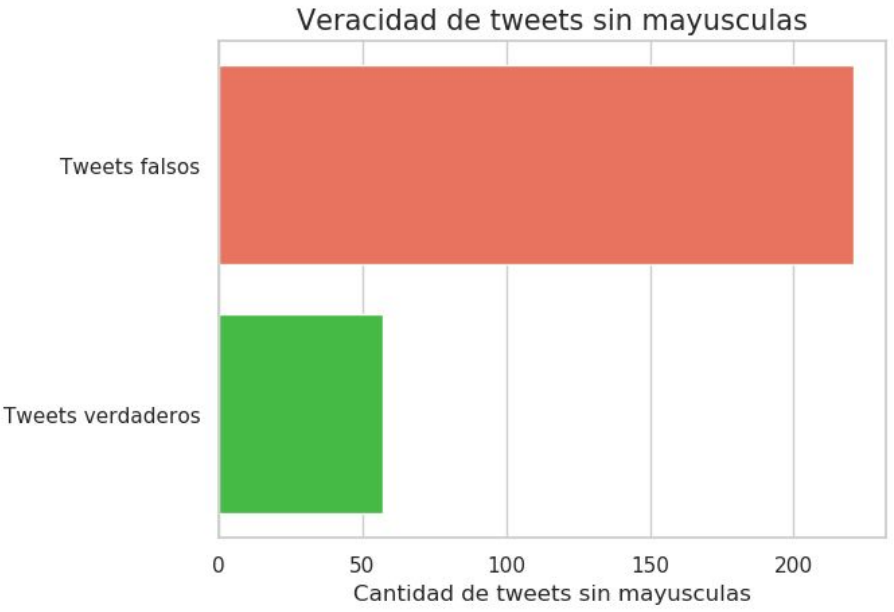
	target-count
target	
0	221
1	57

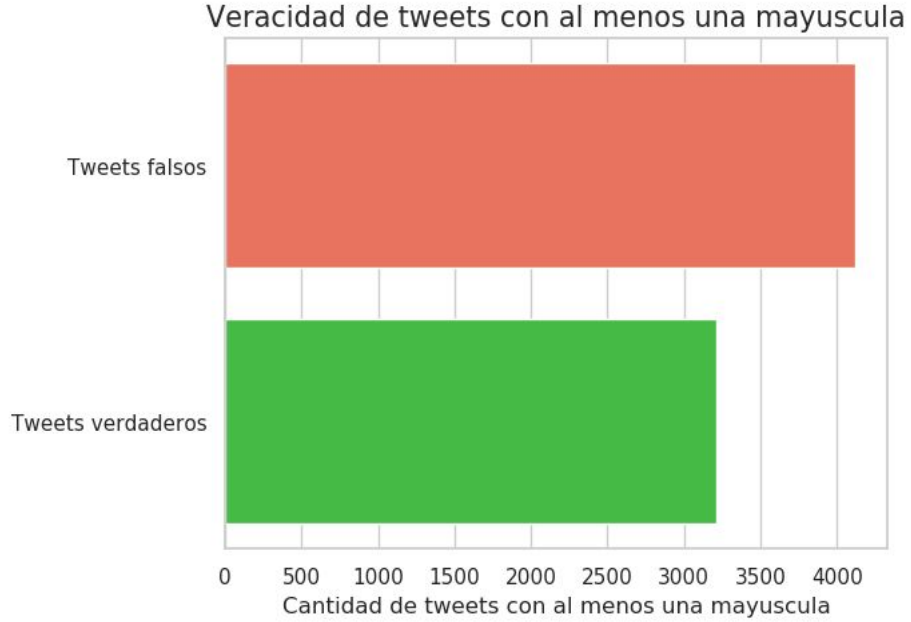
Es decir que tenemos un 79.5% de tweets cuando no hay ninguna mayúscula en su contenido. Esto se corresponde con lo antes mencionado acerca del correcto uso de las mayúsculas en los textos.

Luego, realizaremos el mismo análisis para aquellos tweets que contienen al menos una mayúscula. Pudimos obtener los siguientes valores:

	target-count
target	
0	4121
1	3214

Como se puede ver, ahora tenemos un subset de datos más grandes, es decir que la mayoría de los tweets de nuestro set de datos utiliza alguna letra mayúscula, como era de esperar. Sin embargo vemos que ahora la diferencia no es tan marcada, y no podemos sacar mucho más de lo visto, es decir que el 56.20% de los tweets con al menos una mayúscula son falsos. Para visualizar la diferencia de porcentaje de tweets falsos cuando tenemos una mayúscula y cuando no, utilizaremos dos gráficos que nos representarán cada caso



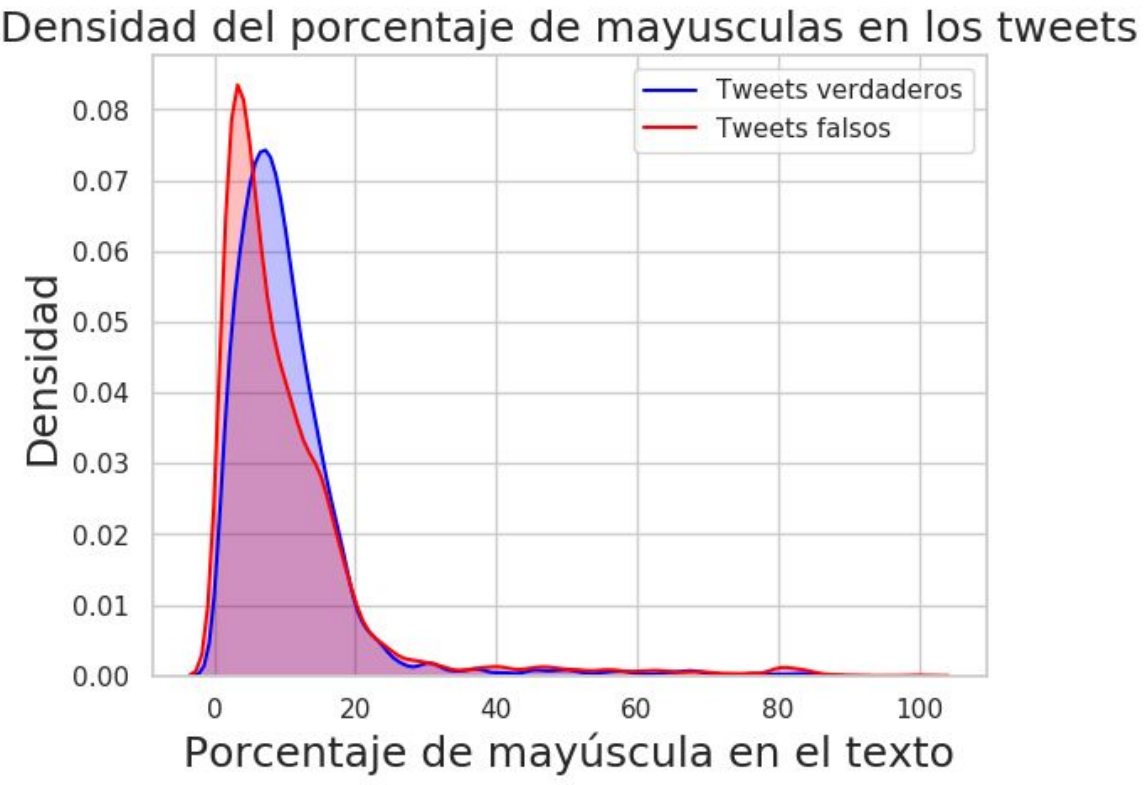


Claramente vemos que para un caso la diferencia es bastante marcada, mientras que para el otro, no dice mucho en cuestiones de probabilidad.

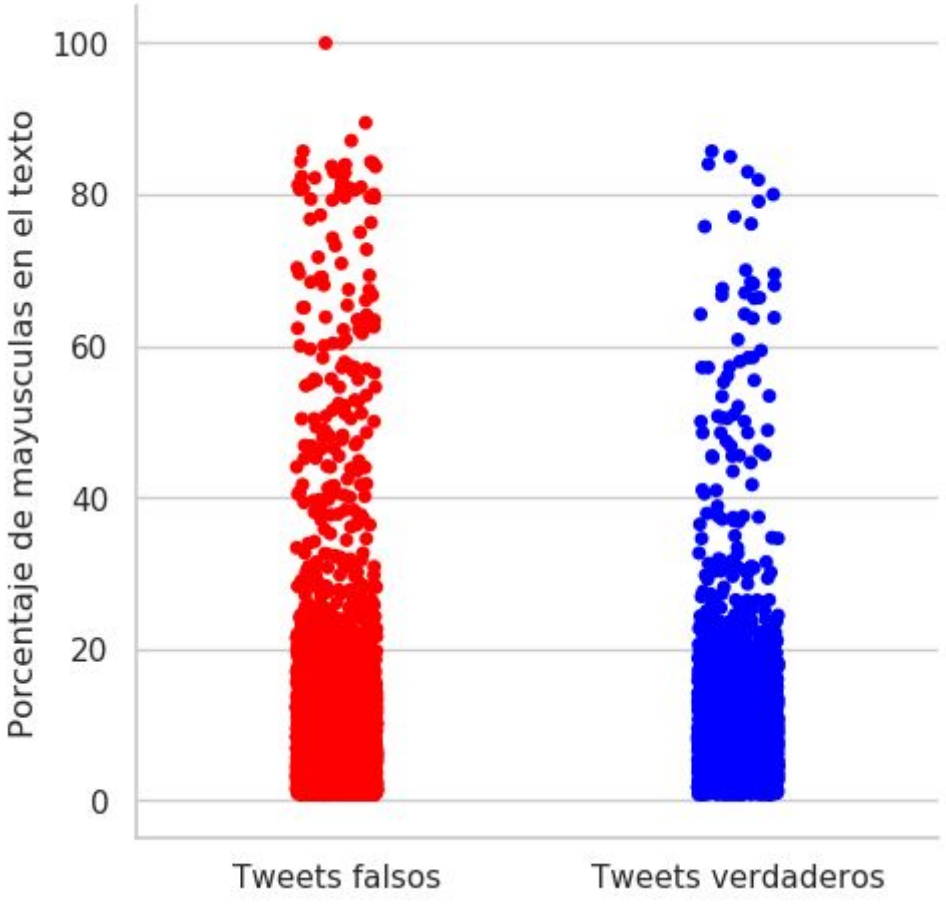
Para el próximo análisis que podemos realizar deberemos tener en cuenta solamente aquellos tweets que tienen una mayúscula como mínimo. De esta forma, se podrá verificar el promedio de cantidad de mayúsculas tanto para los tweets verdaderos como falsos y averiguar si normalmente los tweets con muchas mayúsculas efectivamente son falsos. Si calculamos el promedio de ambos casos obtenemos lo siguiente.

target	uppercase_percentage-mean	uppercase_percentage-count
0	10.322274	4121
1	10.371549	3214

A primera vista podemos ver que los promedios son casi idénticos. Para visualizar cómo se distribuyen podemos usar un histograma solapado y comparar los porcentajes para cada caso.



Se puede ver que la distribución de ambas variables son parecidas, sin embargo los tweets falsos tienen una mayor concentración de promedios de cantidad de mayúsculas entre 0 y 10. Podemos visualizar esto con un plot categórico.



Se puede obtener los textos con mayor cantidad de mayúsculas en porcentaje tanto para los tweets verdaderos como falsos:

- **Verdadero**: UNPREDICTABLE DISCONNECTED AND SOCIAL CASUALTY ARE MY FAVORITES HOW DO PEOPLE NOT LIKE THEM
- **Falso**: LOOOOOOL

Apariciones de caracteres numéricos

Por el lado de los números en los tweets, se puede pensar que son una característica bastante utilizada sobre todo al momento de proporcionar horario, cantidades, distancias, etc. Por lo que se pensó que los caracteres numéricos merecían un correcto análisis de veracidad.

Para empezar tenemos que ver si los caracteres con ningún número son fiables o no, y cual es el porcentaje de los mismos en cuanto a veracidad corresponde.

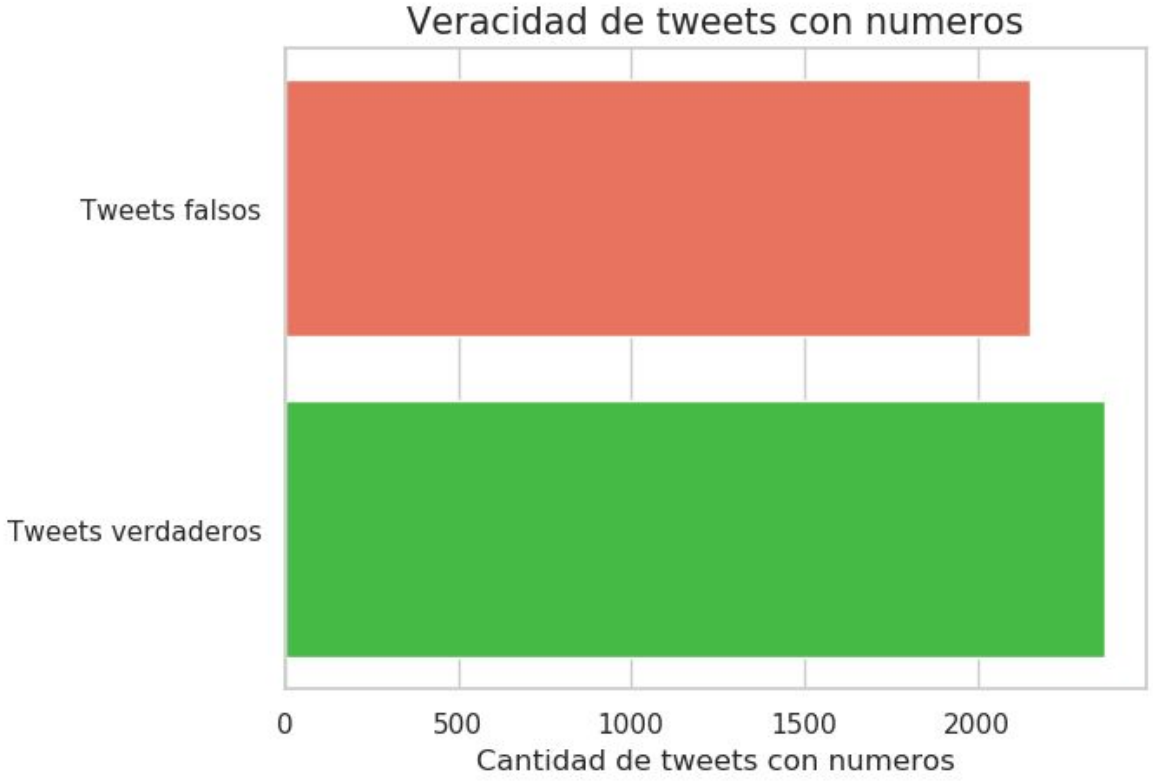
target	target-count
0	2187
1	901

Vemos que casi un 71% de los tweets que no contienen número son falsos. Si se lo toma por el lado de la no aparición de un número, se puede pensar como la falta de precisión en algunos aspectos de la noticia, por lo que no sería fiable, pero esta lógica podría no ser correcta para todos. También podemos ver que el subset encontrado es de 3088 tweets sobre los 7613 de su totalidad. Por ende se supone que ambos subsets estarían equilibrados en cuanto a cantidad de filas con una leve proporción mayor para los tweets con números. Veamos qué pasa con los tweets que tienen al menos un número en su interior:

target	target-count
0	2155
1	2370

Tenemos para este caso que los tweets están mucho más equilibrados en cuanto a la veracidad teniendo un 52,38% de porcentaje. Podemos ver las cantidades de dichos casos en los siguientes gráficos:

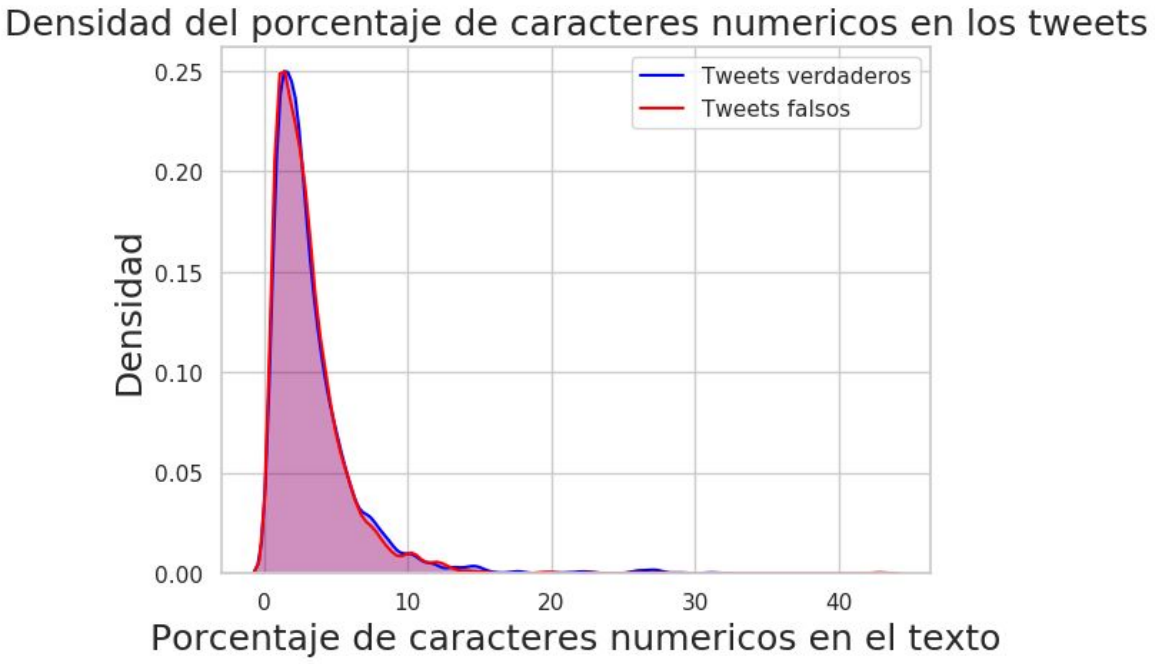

Veracidad de tweets sin numeros
Tweets falsos
Tweets verdaderos
0
500
1000
1500
2000
Cantidad de tweets sin numeros


Veracidad de tweets con numeros
Tweets falsos
Tweets verdaderos
0
500
1000
1500
2000
Cantidad de tweets con numeros

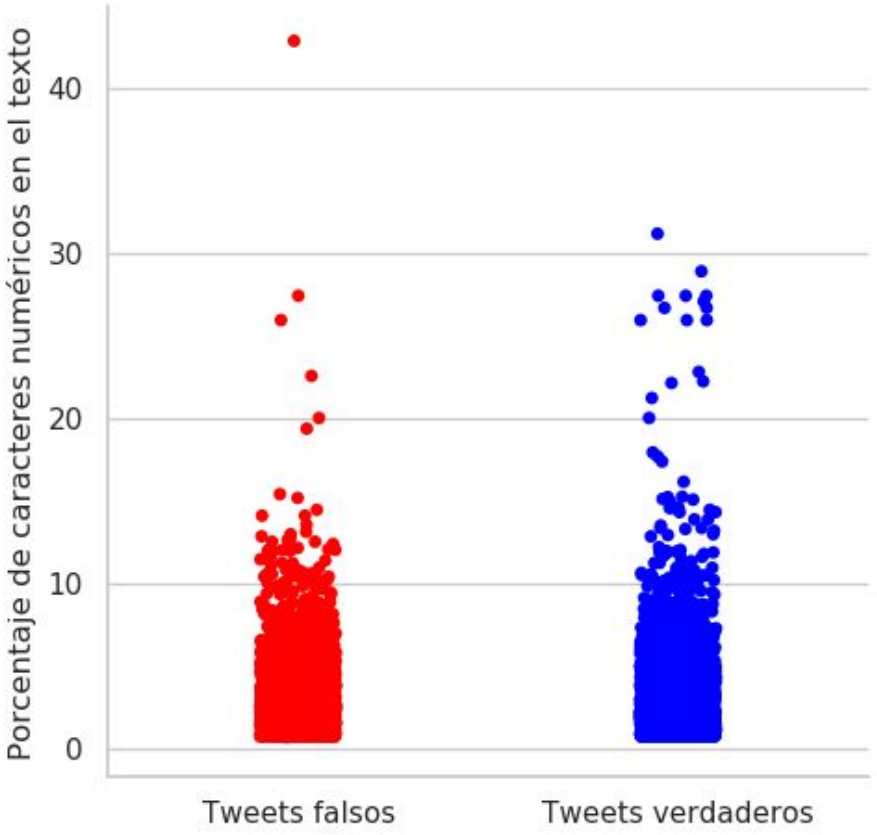
Los gráficos solo sirvieron para comprender mejor la diferencia entre un caso y el otro. Pero, una vez más, se cree que es necesario conocer el porcentaje promedio de los tweets que contienen al menos un caracter numerico. Si calculamos esto obtendremos los siguientes promedios de porcentaje:

target	numeric_percentage-mean	numeric_percentage-count
0	3.140710	2155
1	3.356245	2370

Vemos que están bastantes equilibrados en cuanto a porcentaje se trata, teniendo las siguientes distribuciones solapadas en el siguiente gráfico



Como se advirtió antes, podemos ver que las variables se comportan prácticamente de la misma forma, mucho más que cuando analizamos las mayúsculas. Pero si queremos probar de otra forma como se comportan los porcentajes en cuanto a tweets con al menos un caracter numérico, podemos realizar un catplot para visualizar.



Se puede ver que para valores altos, los tweets verdaderos están separados pero no tanto como los tweets falsos, los cuales prácticamente son 7 los que contienen estos valores separados de la mayoría. A continuación podemos listarlos de mayor a menor porcentaje:

- **Falsos:**
  - Err:509
  - #Sismo M 1.3 - 1km NNE of The Geysers California: Time2015-08-05 23:40:21 UTC2015-08-05 16:40:21 -07:00 a... http://t.co/x6el3ySYcn #CS
  - #USGS M 1.4 - 4km E of Interlaken California: Time2015-08-06 00:52:25 UTC2015-08-05 17:52:25 -07:00 at ep... http://t.co/zqrcptLrUM #SM
  - UNWANTED PERSON at 200 BLOCK OF SE 12TH AVE PORTLAND OR [Portland Police #PP15000266818] 17:10 #pdx911
  - 95-03 BMW 528 530 540 740 Emergency Warning Hazard Switch Button OEM 20177-707D http://t.co/kVNahTHUWZ http://t.co/Y8xkNpqMnJ
  - New Giant Flames (Giant Manly Brown) info/order sms:087809233445 pin:2327564d http://t.co/T1mBw0ia3o http://t.co/CLfa0PY5Lm
  - @DaneMillar1 *screams 666*

Luego, tenemos el tweet verdadero con mayor cantidad de números:

- **Verdadero:** Thu Aug 06 2015 01:20:32 GMT+0000 (UTC)\n#millcityio #20150613\ntheramin sirens

Apariciones de carácter arroba (@)

El carácter arroba '@' es un caracter usado como referencia para otras cuentas de twitter. Dicho carácter normalmente esta presente en los tweets para referenciar algún tipo de cuenta que se relaciona con la noticia mencionada, por ejemplo si se da el clima y se quiere tener una especie de fuente a la que nombrar, se podría arrobar al servicio meteorológico para poder tener referencia y que se esté al tanto de la noticia dada tanto para los usuarios como para los meteorólogos.

Casos como el anterior tenemos muchos pero lo que podemos averiguar con el set de datos proporcionados es cuántos de los que no tienen ninguna referencia son reales y cuántos no.

Si dividimos nuestro set de datos y evaluamos aquellos que no tienen ninguna referencia para verificar su veracidad, tenemos lo siguiente:

	target-count
target	
0	2979
1	2595

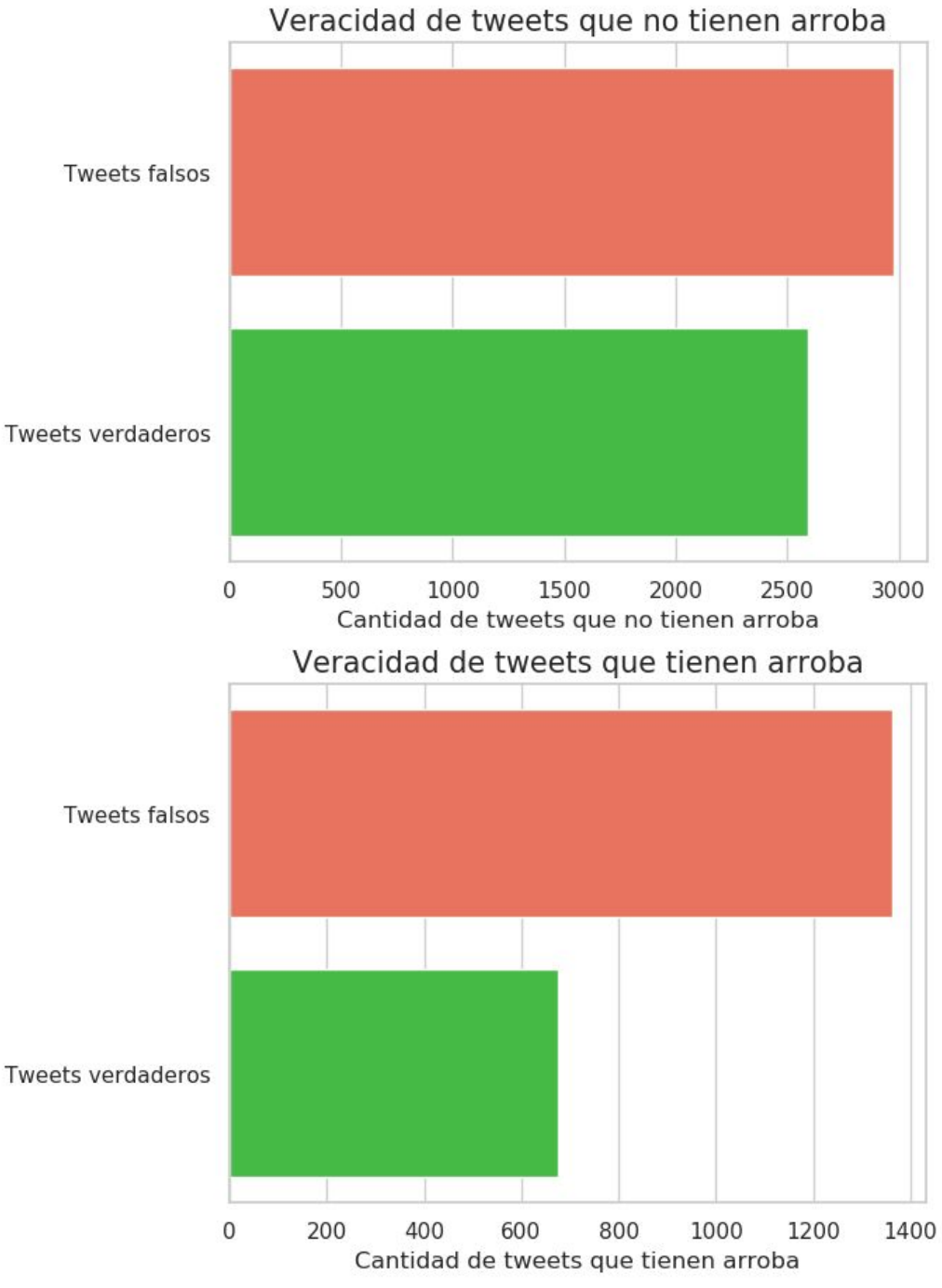
Con lo cual se tiene que el subset de tweets sin ningún arroba es de 5574 sobre 7613. Es decir que la mayoría de nuestros tweets estudiados, no tienen referencias. De este subset para aquellos que no contienen ninguna, no se puede realizar un gran análisis debido a que estos tweets tienen un grado de falsedad de 53.44%.

Pero veamos que sucede para aquellos tweets que si tienen alguna referencia y contemos los que son verdaderos y falsos.

	target-count
target	
0	1363
1	676

Podemos ver que en este caso nuestro subset es mas chico, solo de 2093. Sin embargo, nuestra cuenta arroja que un 66.85% de estos tweets con arroba son falsos. Esto equivale a decir que las referencias muchas veces no aseguran veracidad. Por ende no sirve como parámetro para esto.

A continuación, los gráficos podrían dar una visualización más clara de lo que se quiere expresar y diferenciar.



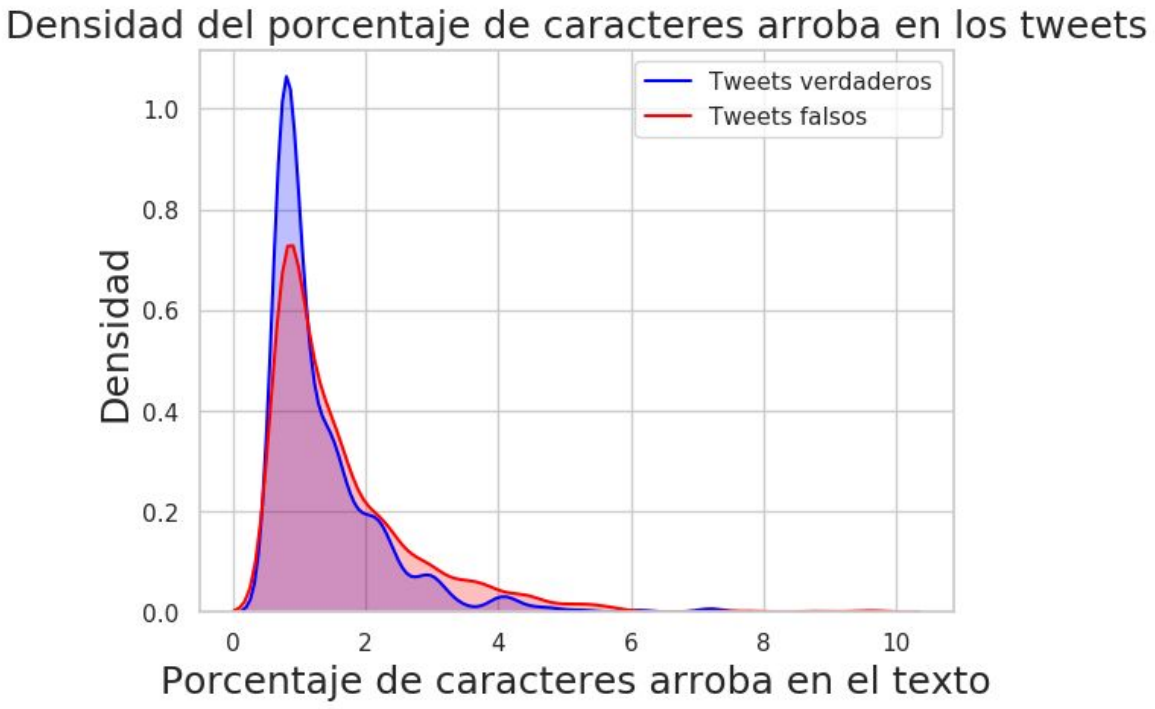
Se puede también hacer un estudio separado para los tweets con referencia y comparar los porcentajes en promedio para ver las diferencias entre los verdaderos y los falsos.

Podemos estudiar el promedio del porcentaje para ambos casos y obtendremos lo siguiente:

target	arroba_percentage-mean	arroba_percentage-count
0	1.621981	1363
1	1.350296	676

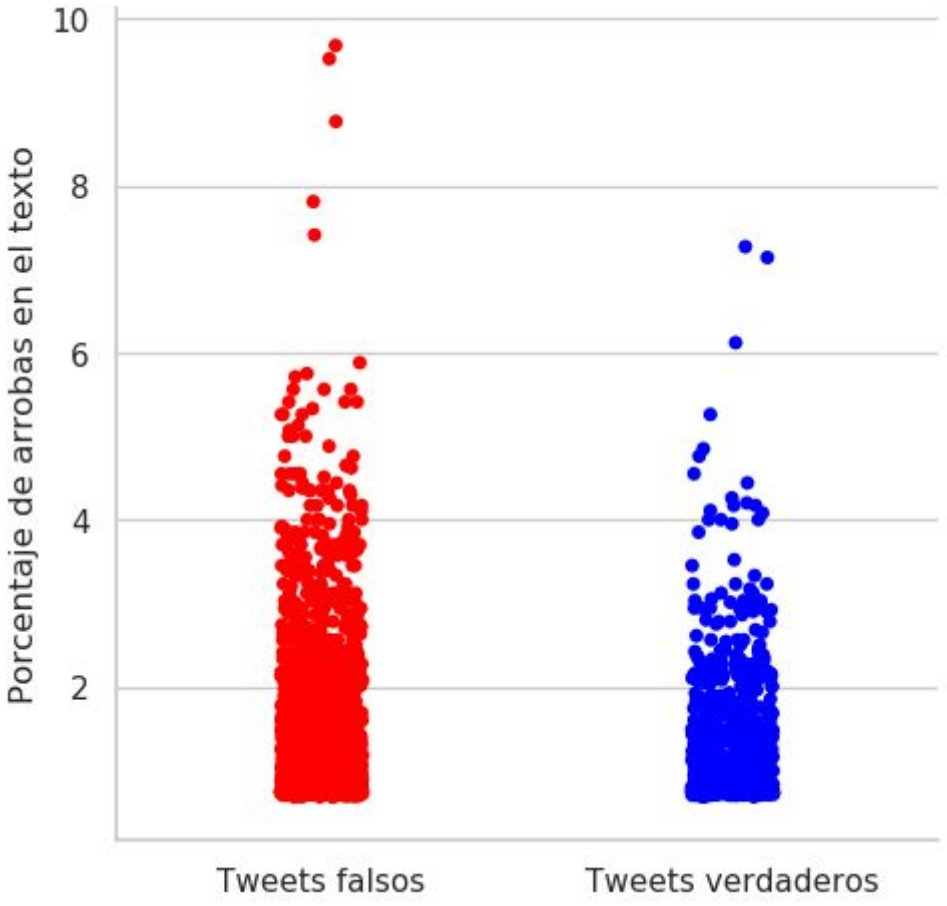
Podemos ver que tenemos un porcentaje promedio un poco mayor para tweets falsos, pero nada muy diferenciado.

Para refinar el estudio, podemos averiguar cómo se distribuyen ambas variables y solaparlas:



Al hacer esto vemos que los porcentajes de los tweets verdaderos estan mucho mas concentrados entre 0 y 1, en cambio los tweets falsos abarcan más valores.

Un plot categórico nos ayudará a poder representar estos puntos y visualizar mejor su distribución:



Vemos que para los tweets falsos se toman valores más altos en el máximo y que los verdaderos se concentran más entre 0 y 1.

Particularmente podemos ver que tenemos 5 tweets falsos con porcentaje mucho más alto que el máximo de los verdaderos. Estos 5 tweets son:

- **Falsos**
    - @PLlolz @Grazed @Stretcher @invalid @witter @Towel still a lot
    - @invalid @Grazed @Towel @Stretcher @PLlolz @witter I can't stop
    - @Grazed @invalid @Stretcher @Rexyy @Towel 'Ben favorited'
    - @Stretcher @invalid @Grazed @Rexyy @Towel I see the programme (:
    - @Stretcher @Rexyy @invalid @Towel let's have babies??!

Y el tweet verdadero con mayor cantidad de arrobas es:

- **Verdadero:** @jasalhad @brianboru67 @Jimskiv92 @hijinks1967 Rioting.

Apariciones de signos de pregunta (¿?)

Como último análisis de caracteres especiales tenemos al signo de pregunta. Dicho caracter se cree que podría reflejar en un tweet un tipo de duda o pregunta dirigida al usuario, o incluso generar interés en cuanto se refiere a su contenido, por ejemplo "¿Cuántas personas resultaron heridas en el huracán katrina?" seguido de un link a otra noticia.

Este linkeo es muy particular cuando se refiere a tener una noticia extensa y que se ha desarrollado en otro lado.

Por ende, en este caso el signo de pregunta puede reflejar también la fiabilidad del texto y las preguntas que se hacen o el porqué de dichas preguntas dirigidas al usuario tanto para generar interés como para también sembrar confusión o generar una sensación de intriga en los lectores o las personas que justo estaban scrolleando en ese momento y se encontraron con dicho tweet.

Una vez más, tendremos que separar los que contienen de los que no, dado que de esta forma tendremos un análisis bastante más rico en cuanto a la veracidad de los tweets dependiendo de este caracter.

Para empezar, estudiaremos aquellos que no tienen un signo de pregunta. En este caso encontramos un subset de 6571. Es decir que tendremos un set de estudio mucho mayor que para aquellos tweets que si tienen.

Si verificamos la cantidad tweets falsos y verdaderos para este subset tenemos los siguientes valores:

	target-count
target	
0	3592
1	2979

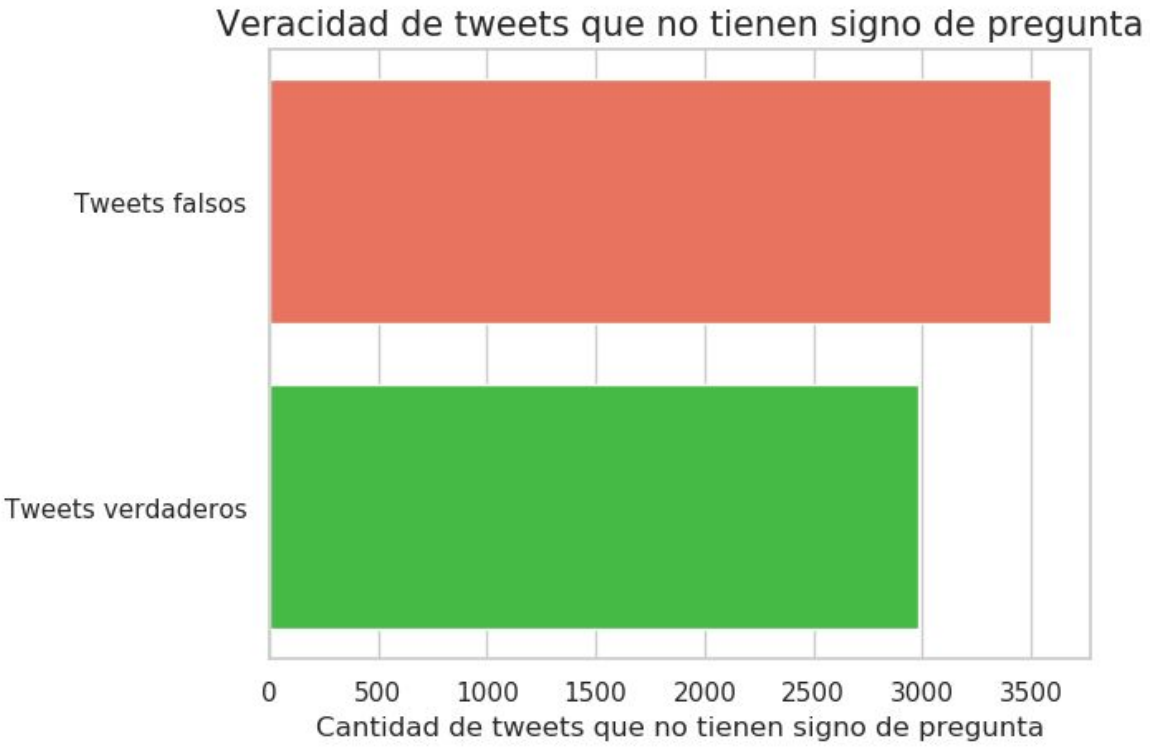
Con estos datos tenemos un 54.66% de tweets falsos para aquellos que no tienen ningún signo de pregunta. Es decir que no se puede decir mucho más de este set y los valores están bastantes equiparados.

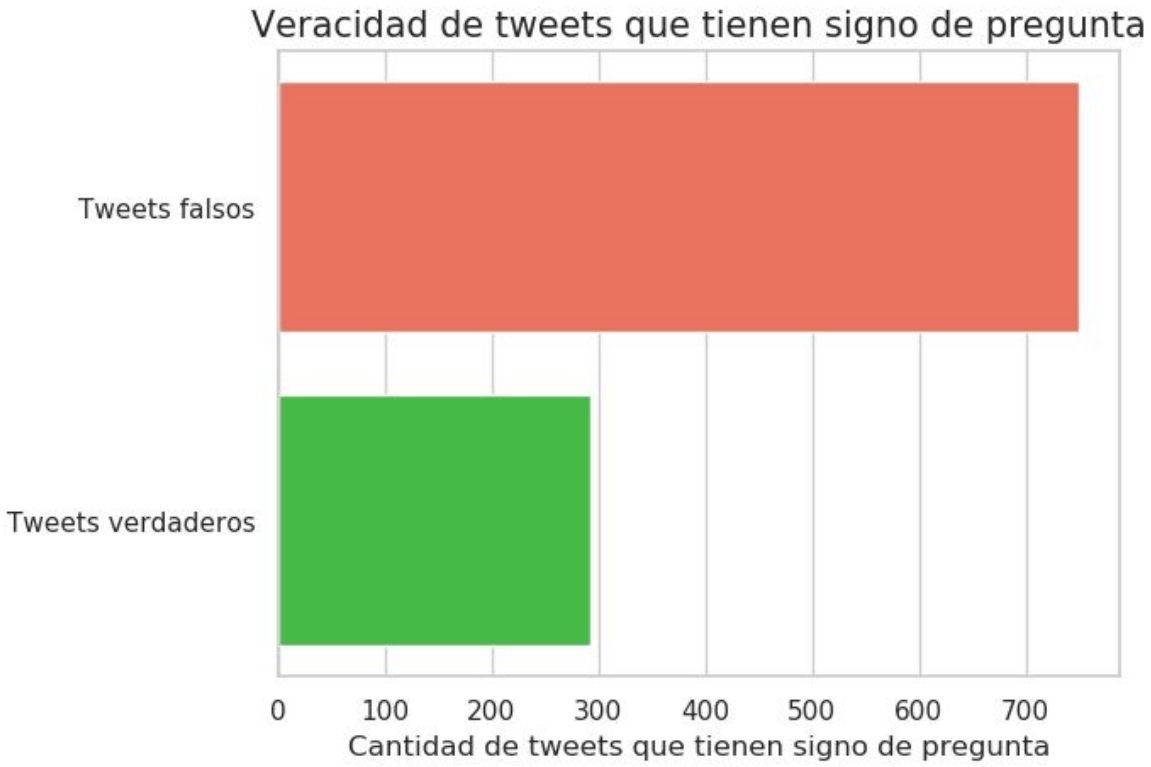
Si ahora estudiamos el subset de los que sí tienen al menos un signo de pregunta, podemos obtener los siguientes valores:

	target-count
target	
0	750
1	292

Es decir, para un subset mas chico, tenemos un nivel de falsedad de casi 72%. Una marcada diferencia entre uno y otro.

Podemos utilizar algunos gráficos de comparación tanto para tweets con y sin signo de pregunta para visualizar mejor estas diferencias:



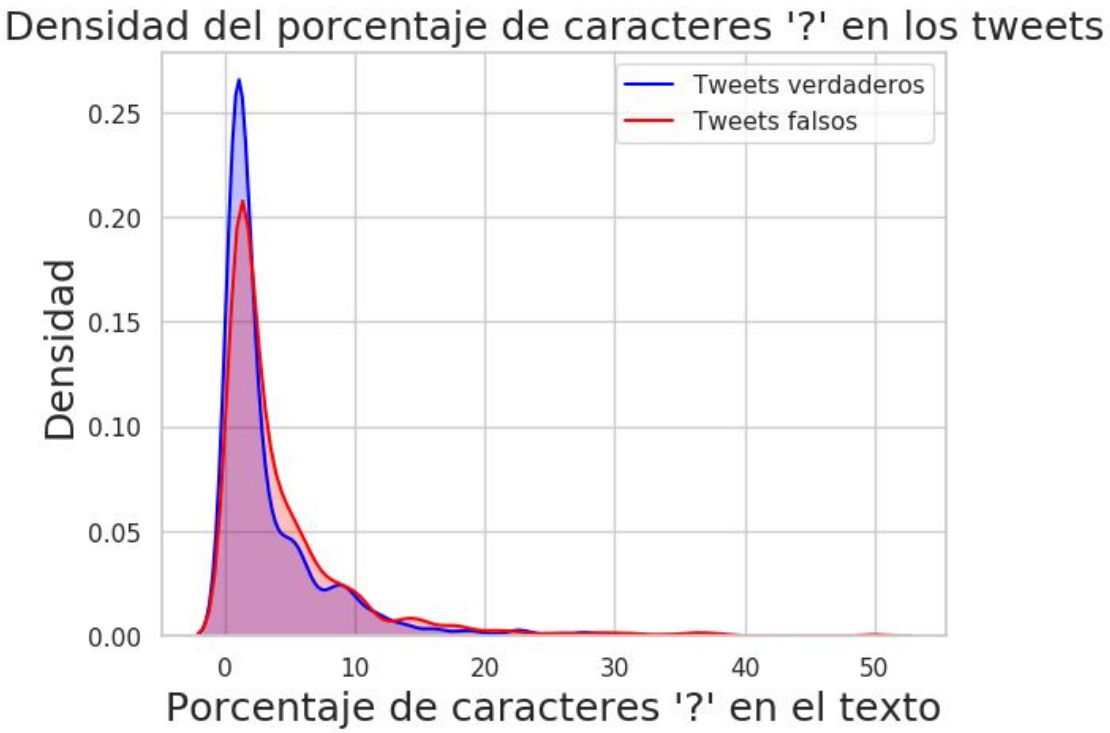


Se puede ver la diferencia en el segundo gráfico acerca de las cantidades.

También se puede realizar el análisis de promedios sobre este último subset. Por ende, si calculamos el promedio de los porcentajes para ambos casos obtendremos lo siguiente:

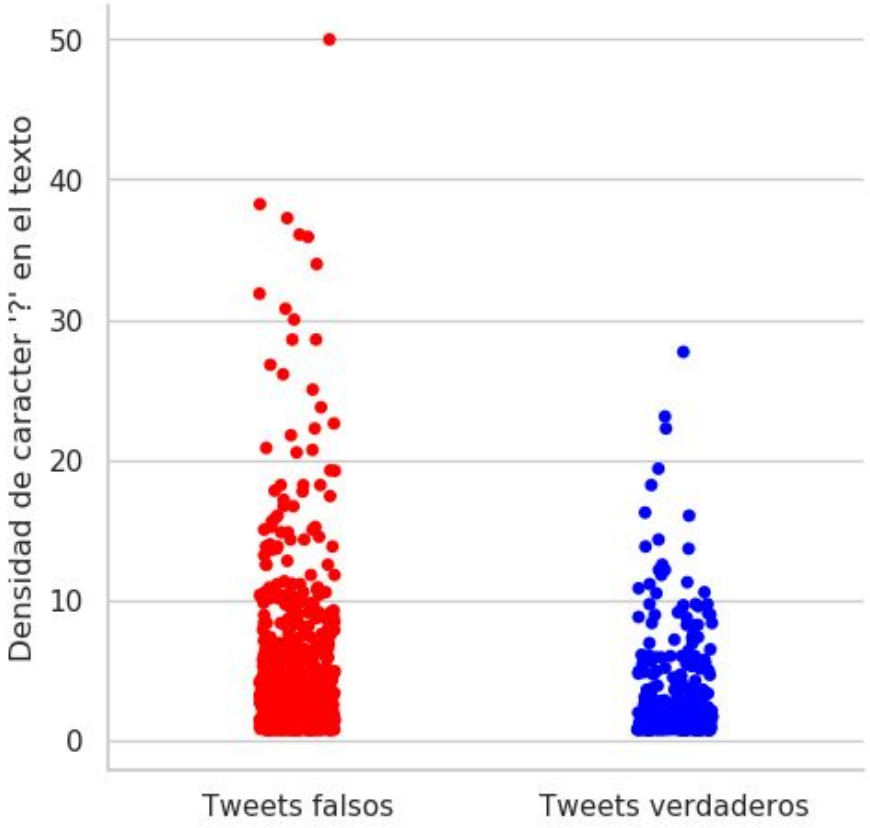
target	question_percentage-mean	question_percentage-count
0	4.438667	750
1	3.281541	292

Se ve un leve crecimiento en los promedios para los valores falsos, pero aun asi no se puede decir mucho de esto, por lo que nos serviría ver cómo se comportan la densidad de dichos promedios.



Se nota en este caso que los valores de los tweets falsos estan un poco más distribuidos mientras que los de los verdaderos se concentran más entre 0 y 5.

Se puede intentar verificar esto con un plot categórico. Obtendremos los siguiente:



Una vez más se puede ver la distribución de los falsos incluso llegando a valores mucho mayores que los verdaderos.

Se pueden obtener los valores del mayor porcentaje de caracter signo de pregunta tanto para los tweets verdaderos como los falsos.

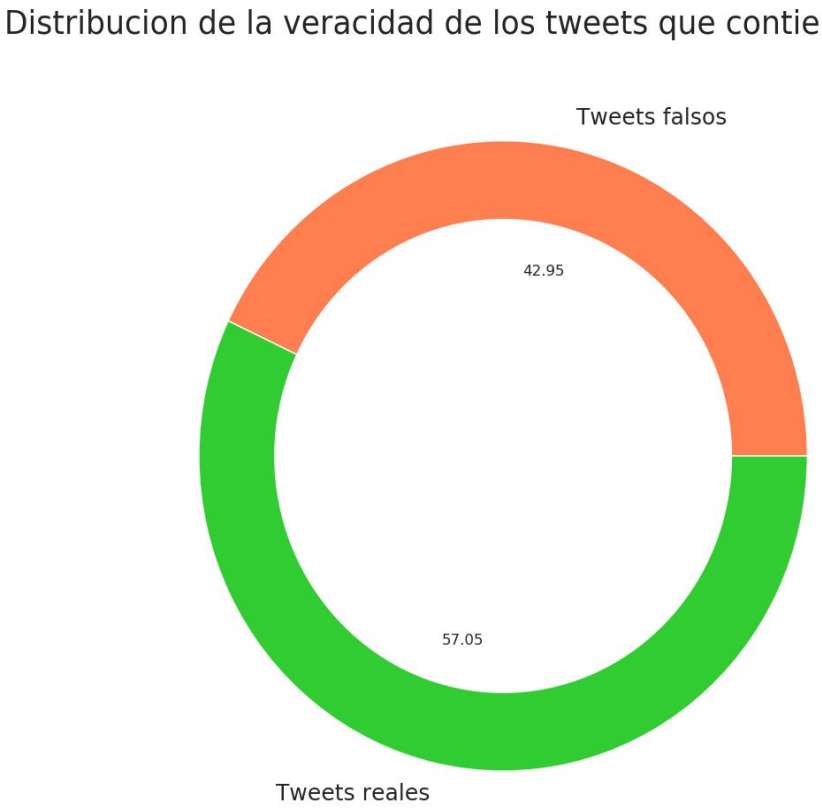
- **Falso:** the best thing at DQ is the cotton candy blizzard ??
- **Verdadero:** #thunder outside my house this afternoon #gawx ??????????????????

Veracidad en el uso de links

Por último, un análisis de veracidad que nos pareció curioso indagar fue el uso de links en los tweets. La pregunta que nos hicimos fue , ¿Estará relacionada la veracidad de los tweets con los links?

Nos basamos en la premisa de que en el dia a dia es muy frecuente observar tweets con links que corresponden a medios de comunicación, y estos medios al centrarse en hechos reales, podrían aumentar el porcentaje de veracidad.

Para este análisis, primero se contó el total de tweets que contenían un link, dando por resultado 3604, lo cual equivale al 47,3% de los tweets del dataset. Luego graficamos la cantidad de tweets reales vs la cantidad de tweets falsos.



En el gráfico pudimos ver claramente que el texto al contener un link es más propenso a ser verdadero. Unos puntos que tuvimos en cuenta fueron que la media total de veracidad del dataset es del 43% y además que la muestra es lo suficientemente grande para no tener un error elevado. Es por esto que concluimos que con los datos aportados por Kaggle, se puede decir que los tweets con links tienen mayor probabilidad de tratarse realmente de catástrofes con respecto a la media de veracidad. Nosotros creemos que esto es causado por la misma premisa que usamos al hacernos la pregunta previamente mencionada.

3. Conclusiones

Para concluir con esta investigación vamos a comentar los datos más interesantantes que logramos recabar.

Se puede destacar en la primera columna analizada, que una keyword aunque no se use mayoritariamente de forma catastrófica, al suceder un hecho que sí lo es las noticias verdaderas se centrarán sobre este y destacarán debido a la importancia del mismo. Esto implica que la veracidad de la keyword a estudiar va a estar fuertemente conectada con el momento en el cual se la analiza.

Tal es el caso, que hoy en día al haber una nueva enfermedad repentina sumamente contagiosa entre personas podría afirmarse que los tweets relacionados a “quarantine” aumentaron su probabilidad de ser reales. Mientras que en el análisis hecho se vio que la palabra clave ‘quarantine’ concentraba únicamente un 27% de veracidad media (cabe aclarar que los datos estudiados no son actuales).

En el ámbito de la ubicación lo que más destaca es la relación encontrada entre caracteres no convencionales para una ubicación de un tweet y la veracidad del mismo tweet. Casi el 70% de los tweets que tienen la característica de tener uno de estos caracteres (números, ?, ;, %, #, |, /, @, +, *, \, $, #) es falso. Aunque el conjunto de tweets que tienen esta característica no tiene una dimensión grande (~200) podemos concluir que es una relación significativa. Esta conclusión tiene sentido ya que ubicaciones como “304”, “1/10 Taron squad” o “?Gangsta OC / MV RP; 18+.?” tienen sentido en muy pocos contextos sintácticos como ubicaciones y parecen ocurrir en tweets con baja cohesión sintáctica, lo cual no hace que sean falsos pero aumenta la probabilidad.

Por el lado del texto resaltan algunas curiosas características.
Primero observamos que tenemos un intervalo de longitudes de tweets que es bastante más verídico que falso, estos son los tweets entre 135 y 145 caracteres.

Con respecto al estudio específico de los caracteres tenemos que los valores observados dieron algunos resultados bastante interesantes en cuanto a este set de disaster tweets. Cabe aclarar que se tuvo que acotar la investigación de los caracteres a los 4 antes mencionados porque se creyó que era mejor elegir menos pero indagar más a profundidad cada uno.
Para la primera investigación tenemos a las mayúsculas y acá encontramos tal vez la mayor diferencia entre falsos y verdaderos: el valor de 80% de falsedad para los tweets que no contienen mayúsculas. Se pensó al comienzo que los tweets que contienen muchas mayúsculas tienden a estar peor escritos por ende, a ser falsos. Pero se pudo encontrar que la realidad es la falta de ellas lo que puede llegar a determinar con más seguridad si un tweet es verdadero o falso.

Por el lado de los caracteres numéricos se lo ha asociado por el valor que se le da a las estadísticas o ciertos horarios que reflejan precisión de algún desastre o accidente. En nuestro set en particular, se encontró que hay más tweets con números que sin números. Para estos tweets con números se encontró que en su mayoría son verdaderos, pero la ventaja es mínima. Ahora para el caso de los tweets que no tienen ningún valor numérico encontramos que hay un 70% que son falsos. Por ende se puede suponer que la falta de valores numéricos también afecta a la veracidad del texto debido a que sin datos estadísticos o de tiempo, la precisión sería menor.
Si lo pensamos desde el lado de la práctica de arrobar, es decir referenciar a cierto perfil, damos a entender que nuestra noticia podría tener una continuación o fuente en otro lado. Por ende, la noticia debería ser un poco más fiable, o eso pensaría uno sin antes haber realizado el análisis de dicho texto. Si separamos entre aquellos tweets que contienen arroba de aquellos que no contienen, tenemos que la diferencia entre verdaderos y falsos es mucho mayor en tweets que contienen arroba que en los que no. Cuando

analizamos y pudimos ver los tweets que tienen en su contenido en proporción mayor cantidad de caracteres @ son del tipo:
"Grazed @invalid @Stretcher @Rexyy @Towel 'Ben favorited'".
Se pudo comprobar que "grazed" es una referencia recurrente en estos tweets con mayor cantidad, así como "invalid".

Por último, pero no por eso sin importancia, se tuvo que realizar estos análisis para aquellos tweets que puedan llegar a generar una pregunta en el lector de la noticia al hacer preguntas o inducir al mismo usuario a indagar aún más sobre el tema. Para este caso, si dividimos el análisis para aquellos que tienen signo de pregunta de aquellos que no, vemos que los que tienen, son en su mayoría falsos. Al analizar los tweets tanto falsos como verdaderos que tienen en su contenido mayor cantidad de caracter con signo de pregunta, encontramos para los falsos "*the best thing at DQ is the cotton candy blizzard ??*" y para los verdaderos "*#thunder outside my house this afternoon #gawx ??????????????????*". Ambos tweets para estos casos no representan una diferencia mayor en su contenido sin embargo su veracidad es diferente.