UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE INGENIERÍA

75.06 Organización de Datos

Trabajo Práctico 1
Primer Cuatrimestre de 2020

Grupo 10: The Datalorian

Bobadilla Catalan, German	90123
Briglia, Antonella	90903
Calvani, Sergio Alejandro	98588
Valdivia, Josue Giovanni	93075

Link de GitHub: https://github.com/SergioCalvani/75.06-Datos-TP1-2020

Índice

1. Introducción

Este informe se encarga de analizar los datos de diversos tweets que se obtuvieron de la competencia de kaggle que se encuentra en el siguiente link: https://www.kaggle.com/c/nlp-getting-started, en el se encuentra un CSV que presenta los siguiente campos

- ***id:*** Identificador de cada tweet.
- ***keyword:*** Un identificador de texto específico del Tweet, este valor puede venir en blanco.
- ***location:*** La localización de donde provienen el Tweet, así como el campo keyword, puede venir en blanco.
- ***text:*** El Tweet en sí.
- ***target:*** Denota si el tweet es real (1) o no (0)

El objetivo de todo este informe es analizar los datos descritos anteriormente para observar si se encuentra alguna tendencia en los datos, encontrar si hay alguna irregularidad en alguno de los mismos para luego poder utilizarlos en el Trabajo Práctico Nro 2, referente a Machine Learning. Sin ir más lejos, este informe también ayudará al lector a comprender la relación entre datos.

2. Análisis General

Para comenzar, haremos un breve análisis general, luego haremos un análisis más en detalle y relacionándolo con los demás campos.

2.1. Keywords

2.1.1. General

Figura 1.1: Keywords más Frecuentes

Completando los vacíos vemos que es la cantidad que predomina, luego a simple vista se ve que mucha diferencia entre las principales no parece ver, para ello, a continuación haremos un análisis un poco más detallado del top 10 de keywords.

2.1.2. Top 10

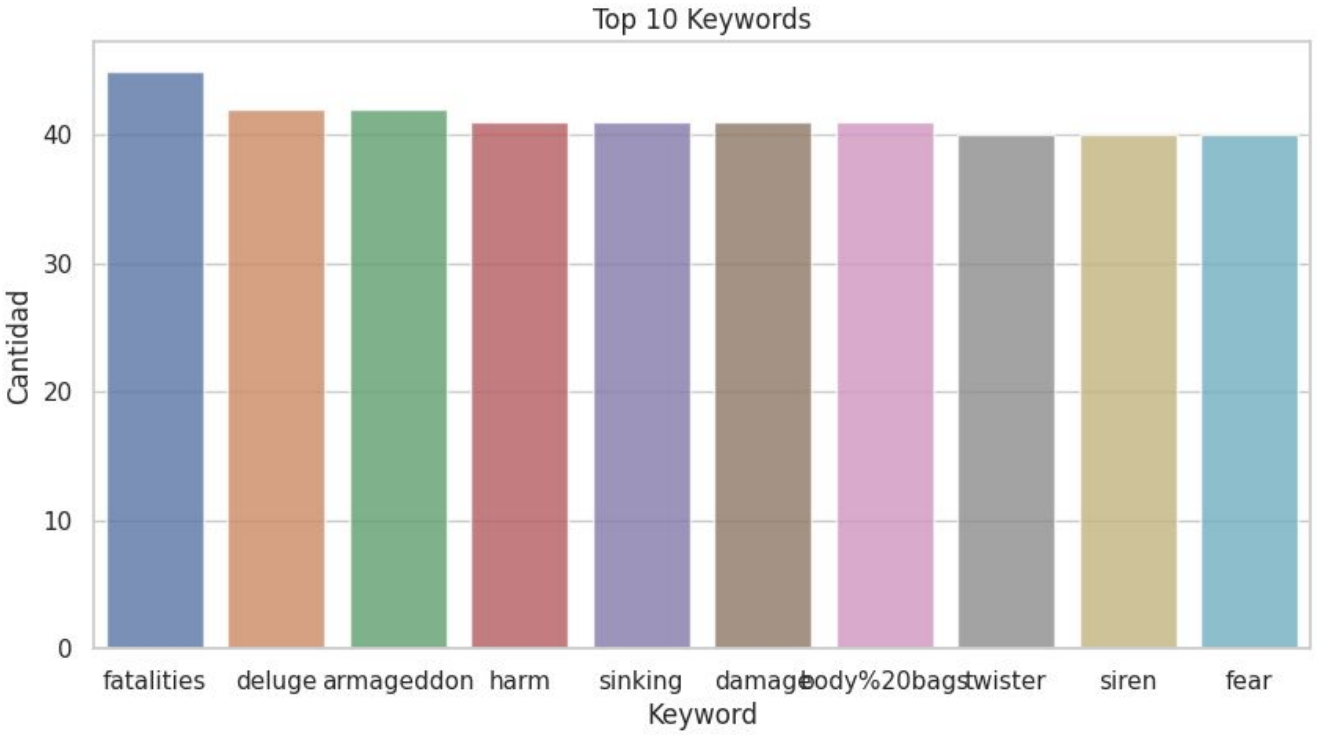

Figura 1.2: Cantidad de Tweets según Keyword

A simple vista se observa que ‘fatalities’ es el keyword con mayor cantidad de ocurrencias, aunque su valor es muy pequeño, de 40 tweets, seguid por ‘armageddon’ y ‘deluge’ con una diferencia mínima. Esto suponemos que es porque hay una gran cantidad de keywords, por lo que tiene sentido que la repetición de los mismos sea realmente poca

2.2. Target

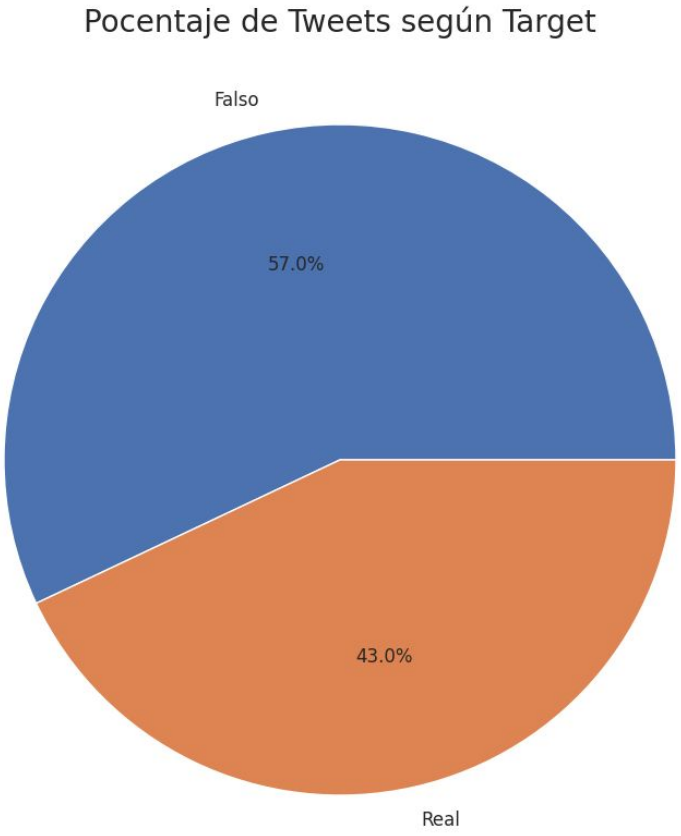

Figura 1.3: Porcentaje de Tweets según Target

En relación con el target, podemos ver que hay una mayor cantidad de tweets Falsos, con un porcentaje aproximado del 57%, siendo el restante 43% el de Reales. Esto puede generar una tendencia a lo largo del análisis, ya que debido a esto, el porcentaje de los tweets falsos en relación a distintos factores puede ser mayor al de tweets reales.

2.3. Longitud de los tweets

Analizamos la relación que pueda existir entre la longitud de los tweets y la veracidad de los mismos, para esto utilizamos la longitud máxima, mínima y promedio para ver si existe alguna relación con el target que nos indica la veracidad

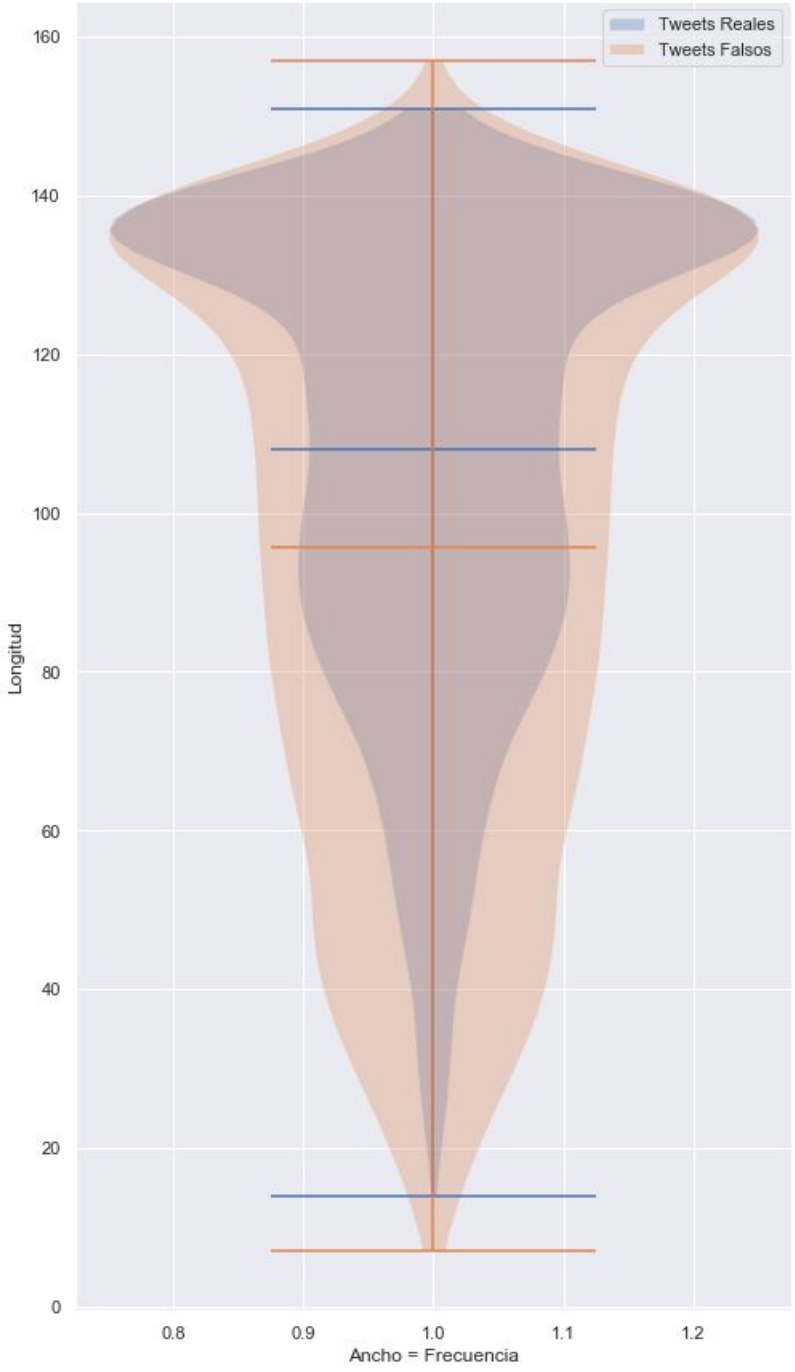

Figura 1.4: Relación del Target con la Longitud

2.4. Conclusión

Se vió que fatalities es el Keyword con mayor cantidad de apariciones aunque su valor es realmente pequeño y la diferencia con los anteriores es realmente pequeña. En cuanto al Target, la distribución es pareja, habiendo una predominancia de tweets falsos, lo cual puede generar una tendencia a lo largo del informe. En relación a la longitud de los tweets, vimos que cuanto más corto es el mismo, más posibilidades de ser falso es.

3. Locación

Veremos que sucede con la cantidad de tweets por zona, pudimos observar que la gran mayoría de los tweets están en inglés. por lo que se puede suponer que la gran mayoría de los mismos van a provenir de países cuya lengua principal es la misma, tal como los Estados Unidos, Inglaterra o Canadá.

3.1. Top 20

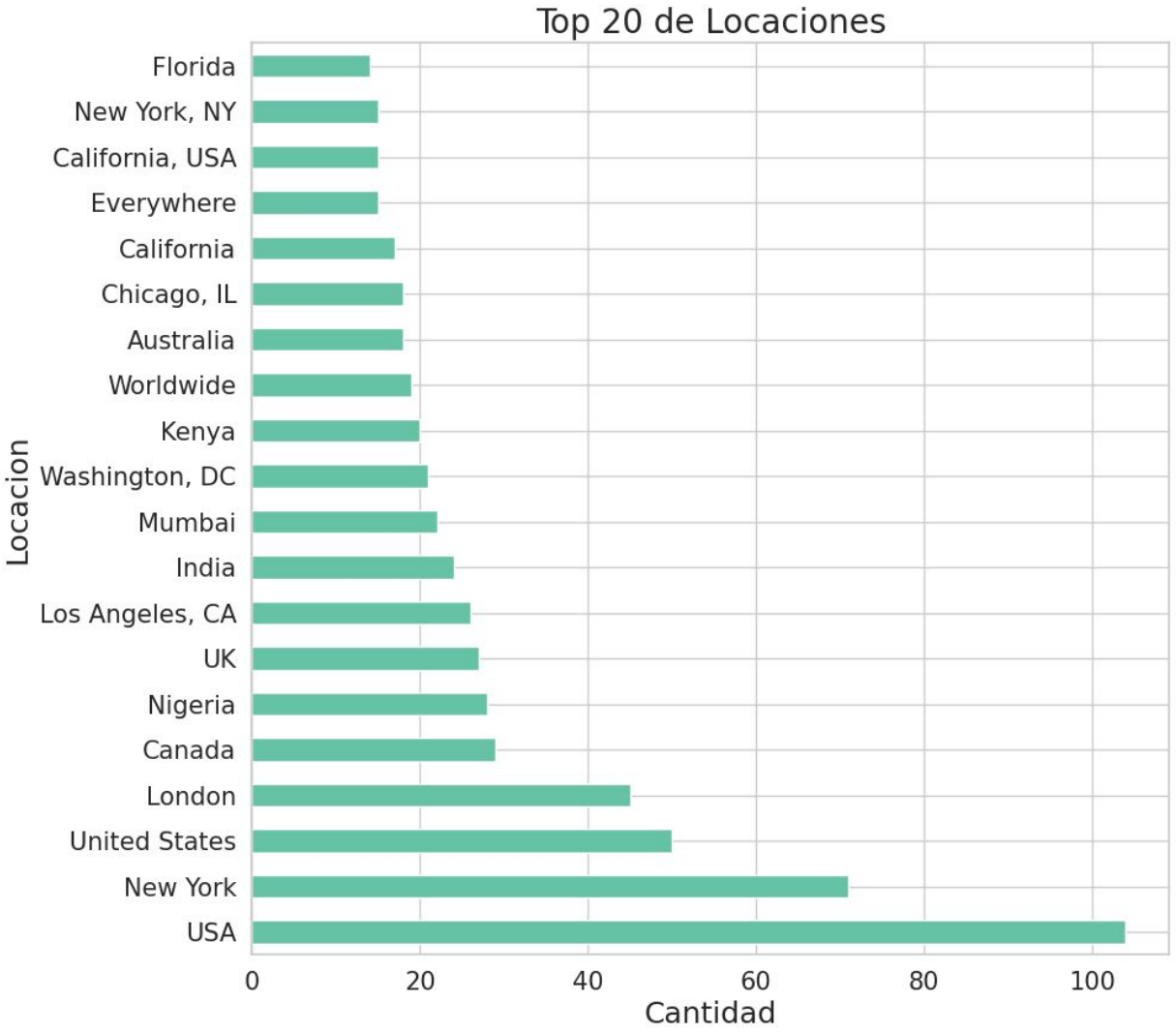

Figura 3.1: Las 20 de Locaciones

Podemos ver así en crudo que USA es el que se encuentra en mayor cantidad de en campo Location, seguido por New York y United States, los cuales, si analizamos a nivel país,

lo mismo. Luego le sigue London y Canadá, esto también creemos que es así por la amplia predominancia de los tweets escritos en inglés.

3.2. Países

3.2.1. Cantidad de Tweets

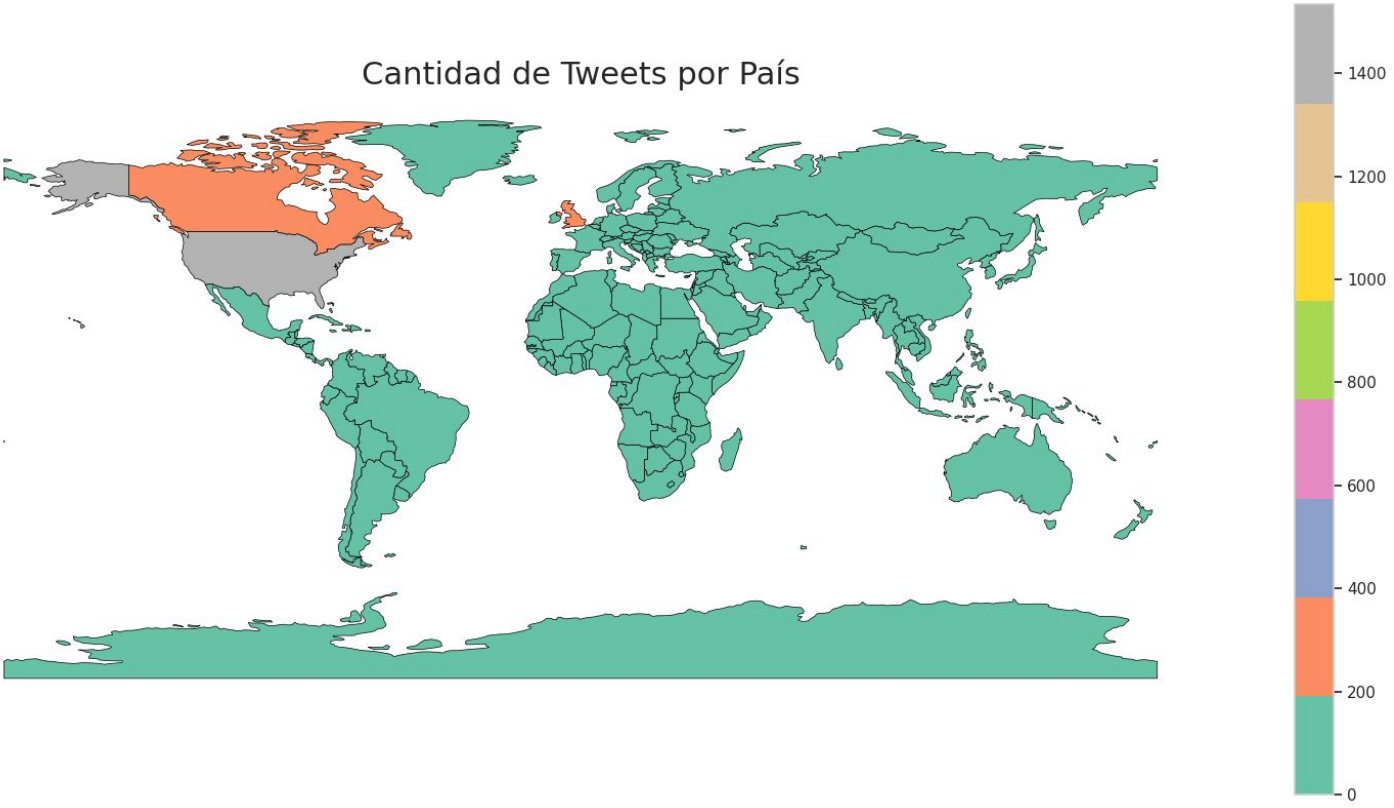

Figura 3.2: Cantidad de Tweets por País

Tal como habíamos previsto, la mayor cantidad de tweets provienen de los Estados Unidos, llegando casi a los 1400, seguido por muy lejos Canada y Inglaterra, luego podemos ver que los demás países presentan una menor cantidad de tweets. Si bien acabamos de ver los países que tienen la mayor cantidad de tweets, sería más interesante y así mismo más importante para lo que va a ser el trabajo de Machine Learning, el porcentaje de tweets reales y falsos, ya que, si se cuentan la cantidad por target, no estamos analizando los datos correctamente, ya que no es lo mismo tener 5 tweets falsos sobre 100 que 6 sobre 1000.

3.2.2. Porcentaje de Tweets Reales

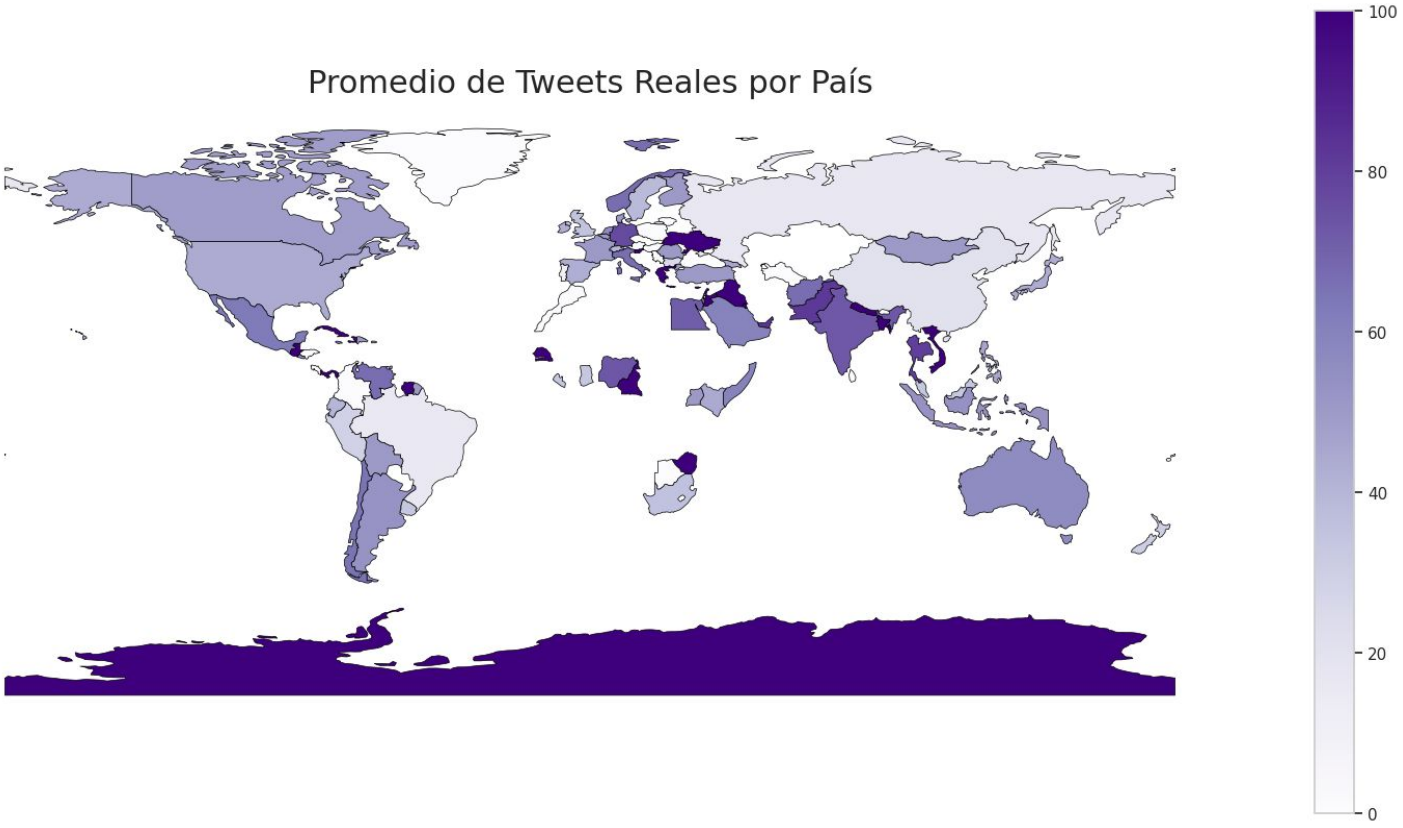

Figura 3.3: Porcentaje de Tweets Reales por País

Se puede observar que los países donde hay una mayor cantidad de tweets totales, el porcentaje parece estar cerca de un 50%, por lo cual la distribución entre los tweets reales y falsos parece ser uniforme, mientras que los países con menor cantidad de tweets totales suelen tener un porcentaje bastante alto. En la India podemos ver que tiene porcentaje bastante alto. En relación a los continentes, América en líneas generales suele tener cantidad de países con un porcentaje bastante similar, cercano al 50%, parecido a Europa aunque en ese continente, el porcentaje de algunos países parece ser superior. En gran parte, África no parece tener muchos países con datos, ya que, la gran mayoría tiene 0%, por lo que puede ser o porque no tienen datos, o realmente el porcentaje de tweets reales es 0, para ello hay que ver que sucede con el porcentaje de tweets falsos. En el medio oriente, la mayoria de paises suele tener un porcentaje bastante elevado, mientras que en Asia, se encuentran porcentajes por debajo de la media.

3.2.3. Porcentaje de Tweets Falsos

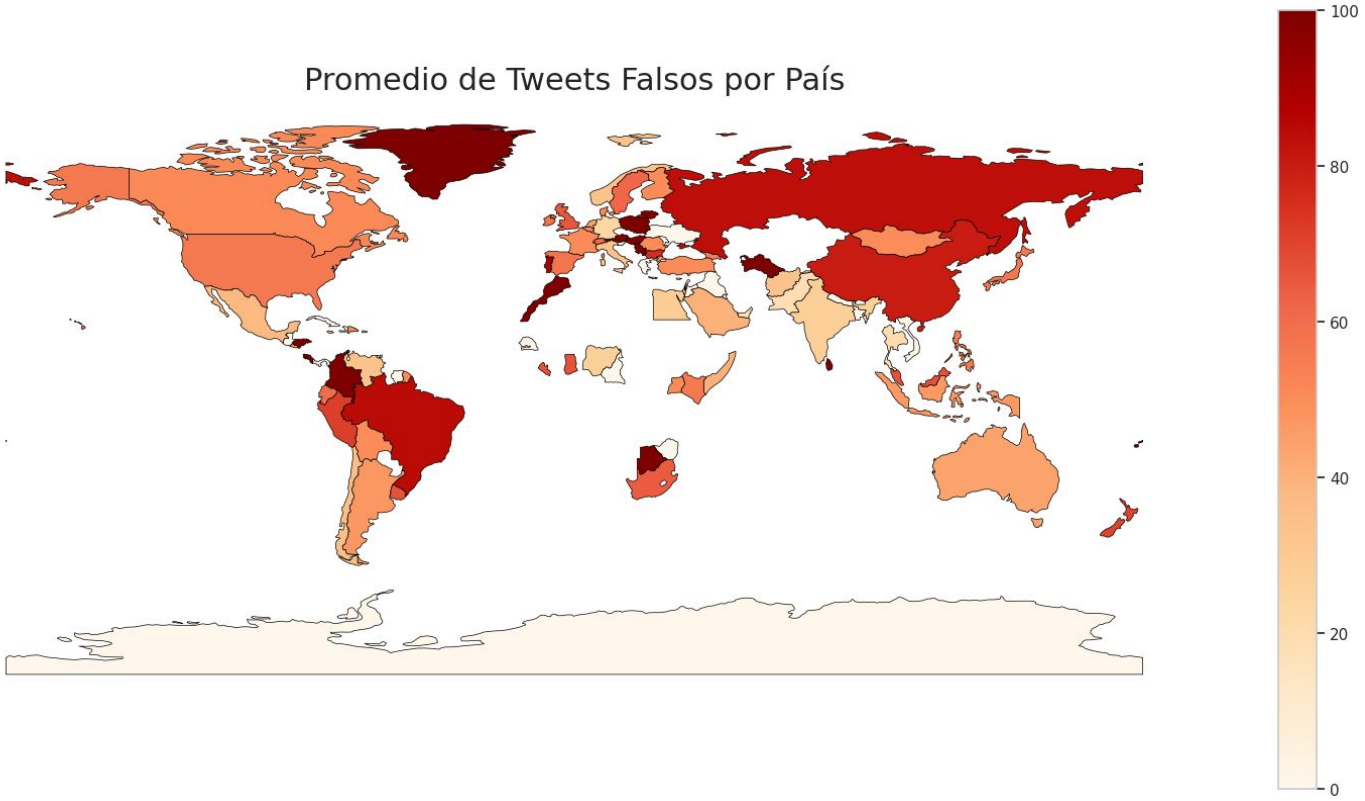

Figura 3.4: Porcentaje de Tweets Falsos por País

En este caso, los países con mayor cantidad de tweets tienen un porcentaje similar al de los tweets reales. Así mismo, se denota a simple vista que Brasil, Venezuela, Rusia y China son los que presentan un mayor porcentaje de tweets falsos, mientras que también se puede notar que, igual a lo que sucedía en el apartado anterior, la gran mayoria de paises africanos no tienen ningún porcentaje, por lo que se puede deducir que no tienen ningún tweet registrado. Si quisiéramos ver que sucede por continente, en América del Sur se encuentra la mayor cantidad de países con porcentaje alto, mientras que en América del Norte es más equilibrado, en Europa sucede algo similar, con algún que otro país con un porcentaje elevado.En el Medio Oriente, la gran mayoría de los países tiene un porcentaje muy chico, lo cual concuerda con los visto anteriormente, en el continente asiático, si bien hay países con un porcentaje bastante alto, la gran mayoría no tiene uno tan elevado.

Ahora bien, ya vimos que la mayor cantidad de tweets se encuentran en los Estados Unidos y, al analizar bien los datos locación, pudimos notar que la gran mayoría de las locaciones con algún indicativo de que es de Estados Unidos estaba ligado al Estado del cual pertenecía, por lo que nos pareció interesante analizar esto.

3.3. Estados de Estados Unidos

3.3.1. Cantidad de Tweets

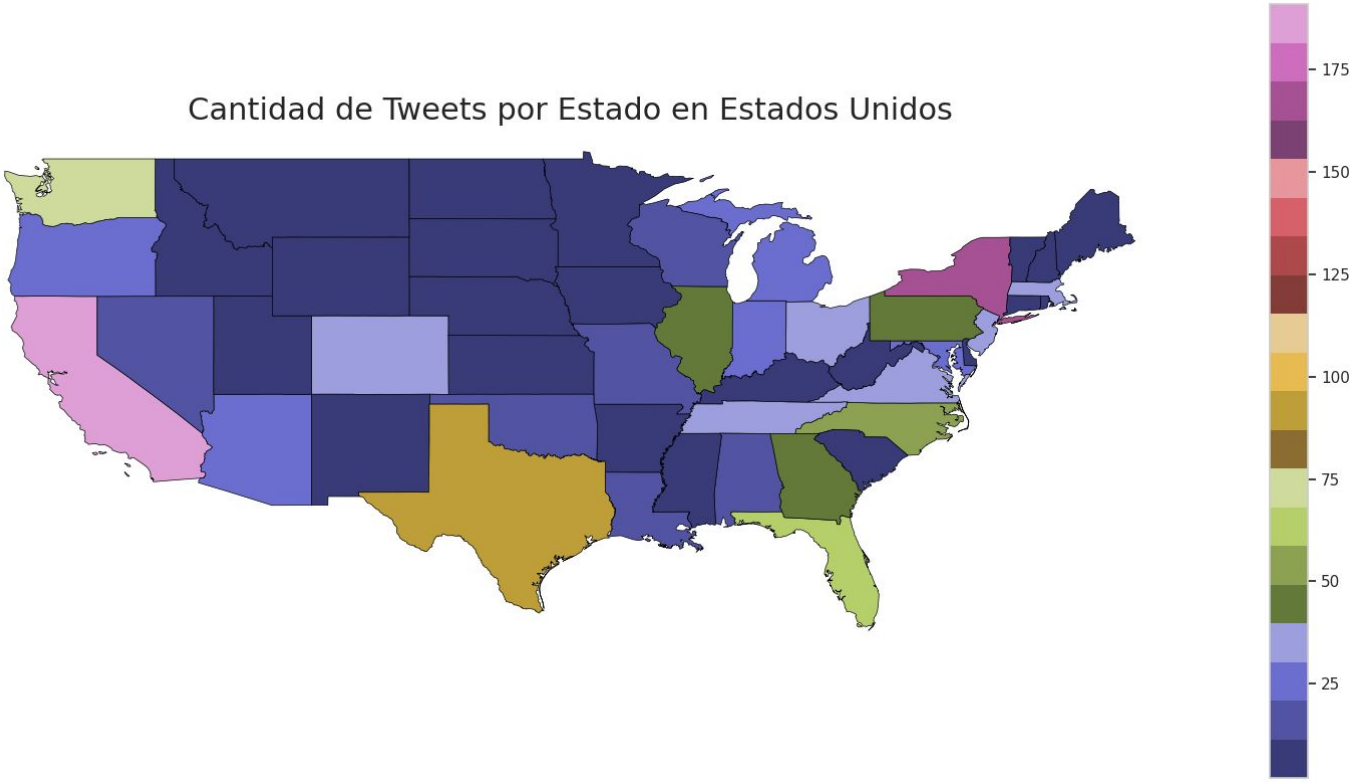

Figura 3.5: Cantidad de Tweets por Estado en Estados Unidos

Se ve a simple vista que los estados de California y de Nueva York son los que tienen la mayor cantidad de tweets, esto no debería sorprendernos en los absoluto ya que son de los estados con mayor cantidad de población del país y por ende, más populares y así mismo, con mayor presencia de tecnología. Luego le sigue el estado de Texas, aunque con una diferencia bastante grande, casi la mitad que en los California y Nueva York, luego le sigue el estado de Washington y Florida. Los demás estados ya podemos ver que tienen una cantidad muy pequeña en comparación a los otros estados.

Ahora, tal y como se hizo con los países, vamos el ver el porcentaje de los tweets reales y falsos por estado.

3.3.2. Porcentaje de Tweets Reales

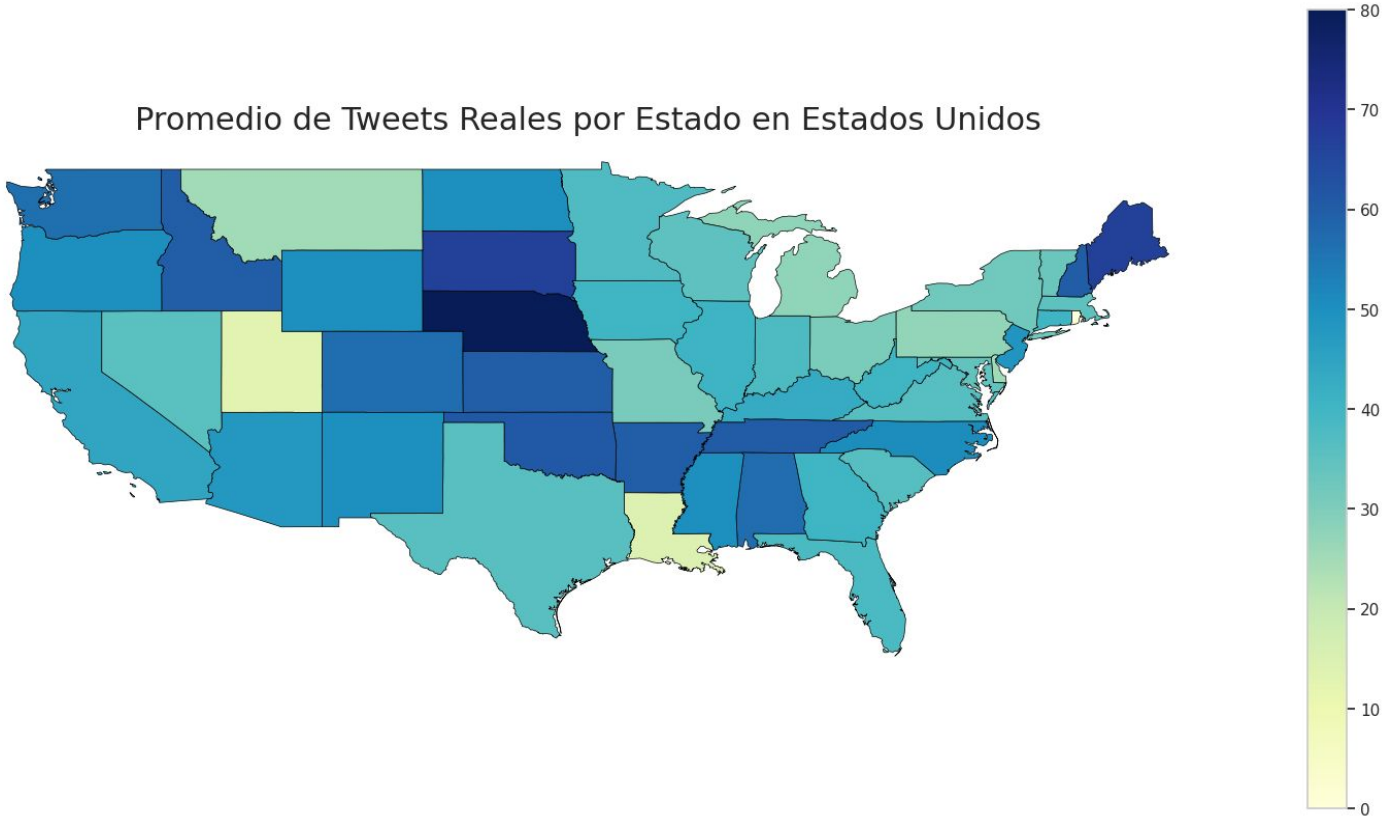

Figura 3.6: Porcentaje de Tweets Reales por Estado en Estados Unidos

Podemos ver que el estado de Nebraska es el que tiene el porcentaje más alto, cerca del 80%, mientras que en los demás estados, el porcentaje suele rondar entre el 60% y 40%, por lo que podría suponerse que la misma tendencia debería verse en el porcentaje de tweets falsos. En este caso, el estado de Utah y Louisiana son los que presentan el porcentaje más bajo, cercanos al 10%.

3.3.3. Porcentaje de Tweets Falsos

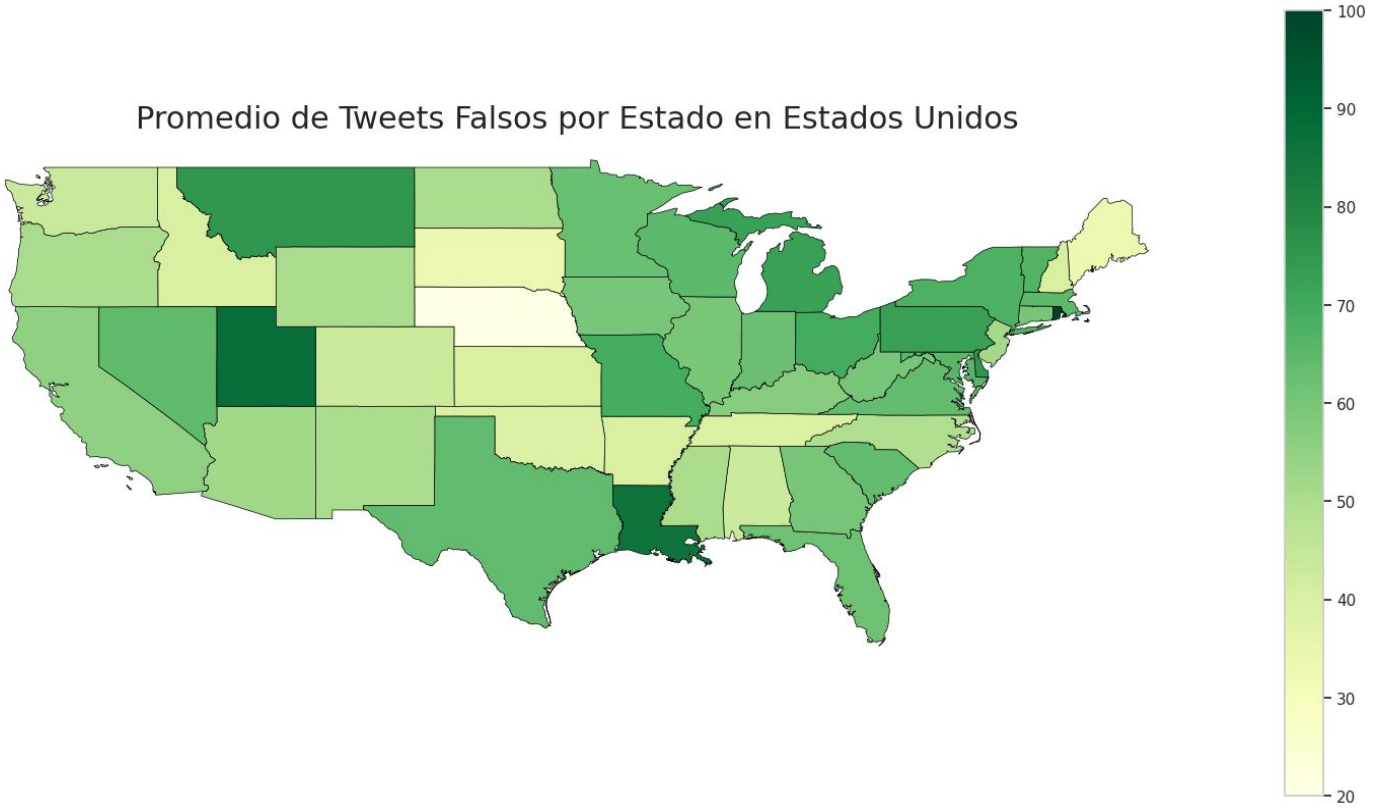

Figura 3.7: Porcentaje de Tweets Falsos por Estado en Estados Unidos

Tal y como se había observado con los países, los estados en el que el porcentaje de tweets verdaderos era alto, aquí el porcentaje es bajo, siendo los estados de Utah y Louisiana los de mayor porcentaje. En este caso, el estado de Nebraska es el que tiene el menor porcentaje y los demás estados salvo los tres ya mencionados, el porcentaje suele ser bastante equilibrado como sucedía con el porcentaje de los reales.

3.4. Conclusión General

En resumen, los países cuyo idioma predominante es el inglés son los que presentan la mayor cantidad de tweets, tales como Estados Unidos, Inglaterra y Canadá. Luego, se pudo ver que en estos países con mayor cantidad, el porcentaje tanto de tweets falsos como verdaderos fue muy similar, mientras que India es uno de los países con mayor porcentaje de tweets verdaderos y la zona de Medio Oriente, mientras que en el resto del mundo el porcentaje es cercano al 50%, habiendo excepciones. En cuanto al porcentaje de tweets falsos, Brasil, Venezuela, China y Rusia son los países que presentan el mayor porcentaje,

mientras que en América del Sur es la zona que tiene la mayor cantidad de países con porcentaje alto. En relación a los estados de Estados Unidos, California y Nueva York son los que presentan la mayor cantidad de tweets pero sus porcentajes, tanto de tweets falsos como reales es muy similar, en el caso del porcentaje de los reales, el estado que predomina es Nebraska, mientras que Utah y Louisiana los de mayor porcentaje de tweets falsos.

4. Análisis Sintáctico

Para el siguiente análisis utilizamos Spacy que es una librería para el procesamiento de texto, con ella podemos extraer los sustantivos, adjetivos, verbos, símbolos, etc de un texto dado, para este análisis se asume que el lenguaje de los tweets es inglés.

4.1. Palabras

4.1.1. Populares Generales

Para un primer análisis general (sin considerar sustantivos, adjetivos, etc.) utilizamos el método llamado stemming que consiste en considerar palabras parecidas como por ejemplo waiting o waited, como una sola, llevándolas a una única raíz: wait.

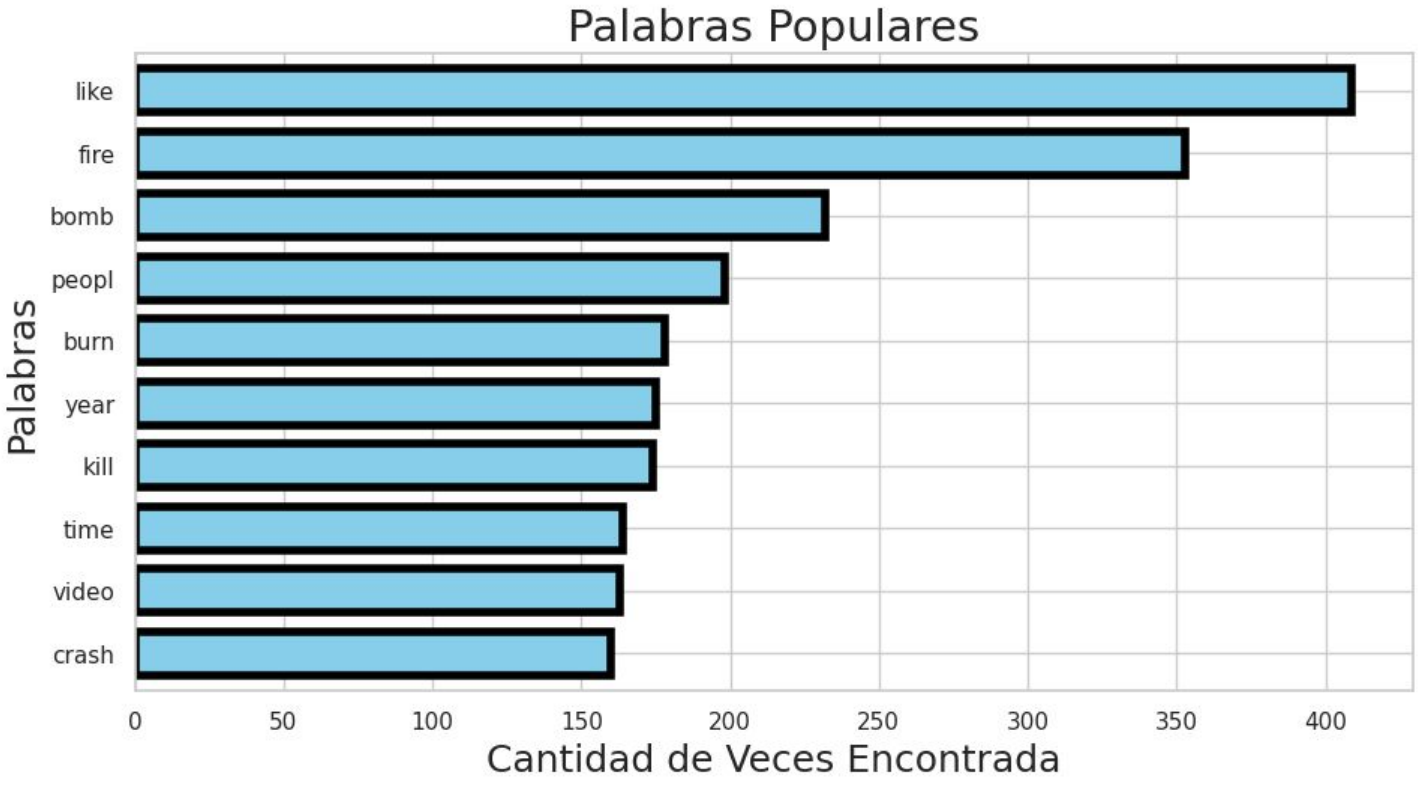

Figura 4.1: Palabras Populares

4.1.2. Populares en Tweets Reales

Dado que existen palabras (stems) populares tanto para tweets reales como para falsos, nos pareció interesante armar un top ten solo con palabras que sean populares para tweets reales, y no para tweets falsos.

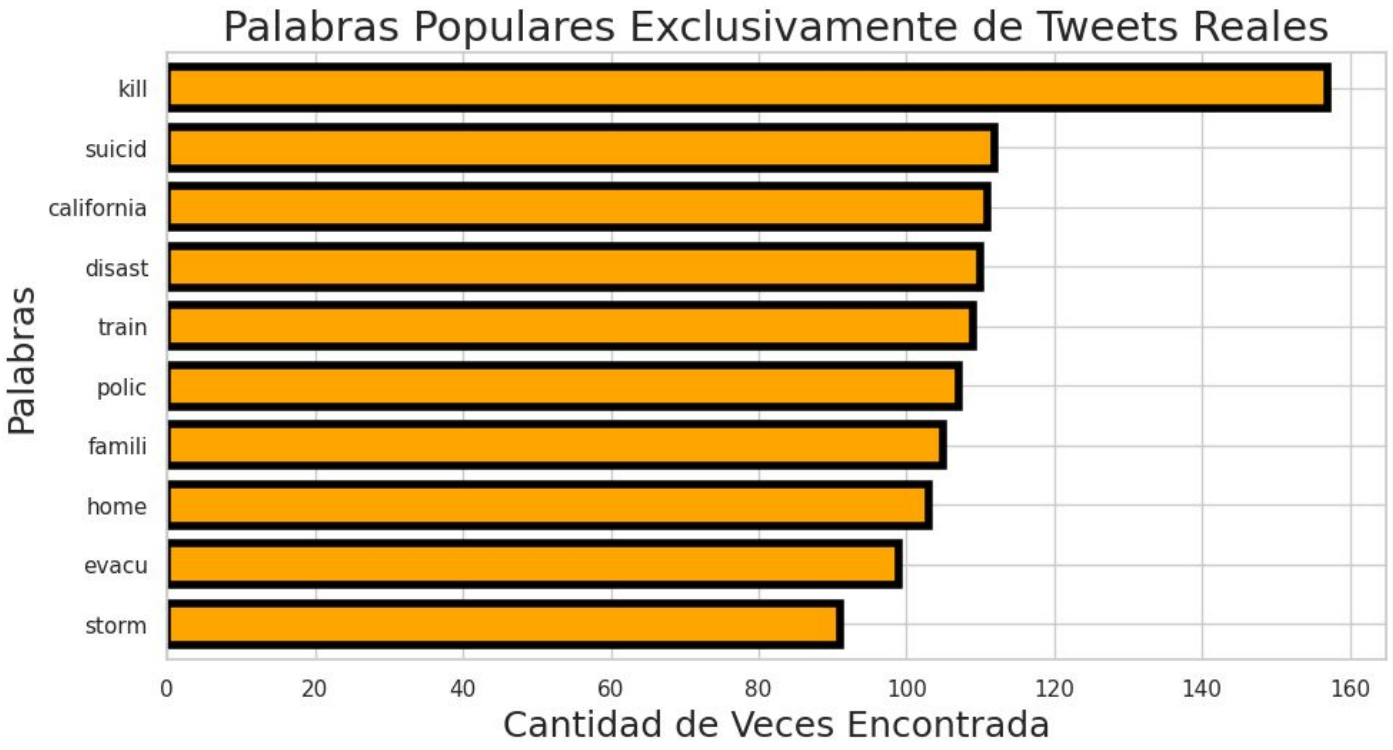

Figura 4.2: Palabras Populares Exclusivamente de Tweets Reales

4.1.3. Populares en Tweets Falsos

De forma similar a como hicimos en el gráfico anterior, ahora armamos un top ten con palabras populares solo de los tweets falsos, y no de los reales.

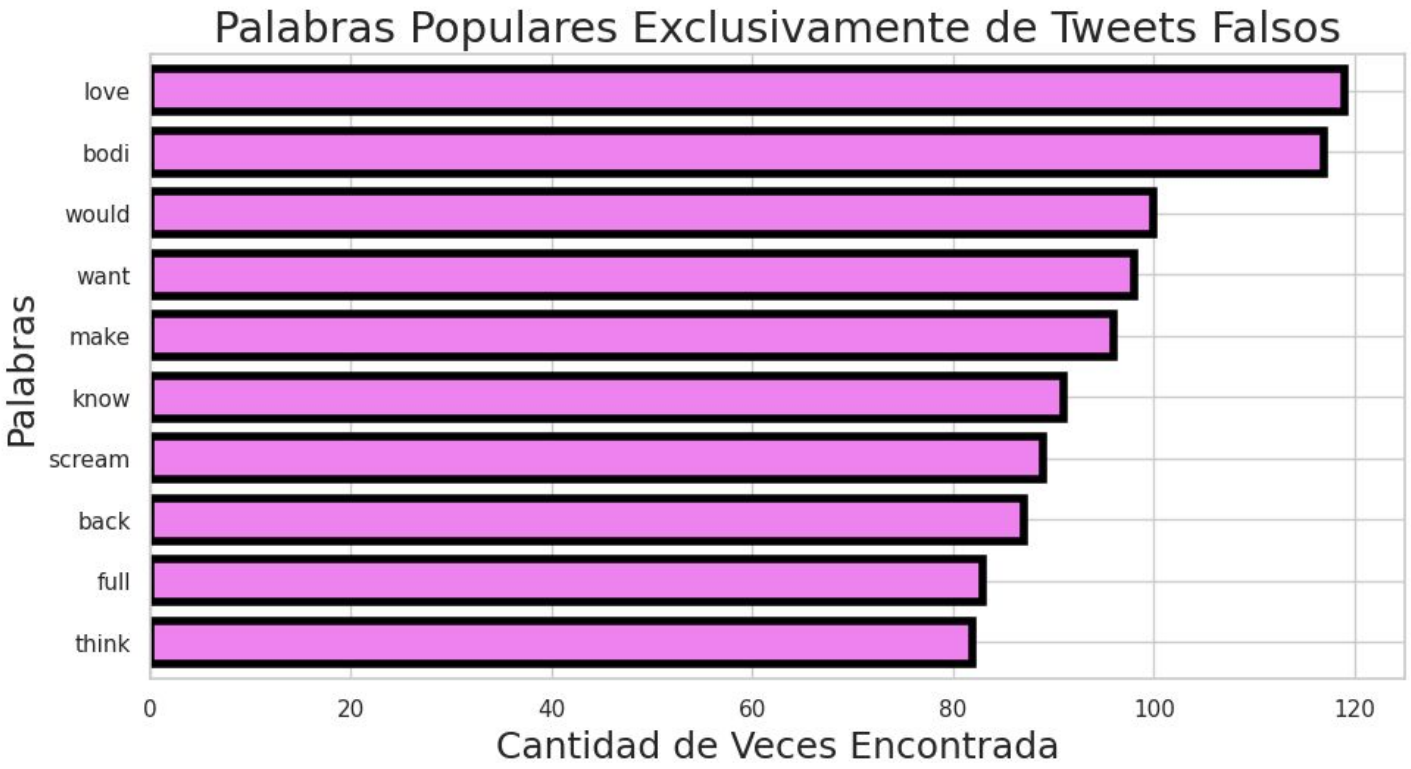

Figura 4.3: Palabras Populares Exclusivamente de Tweets Falsos

4.2. Cantidad Promedio

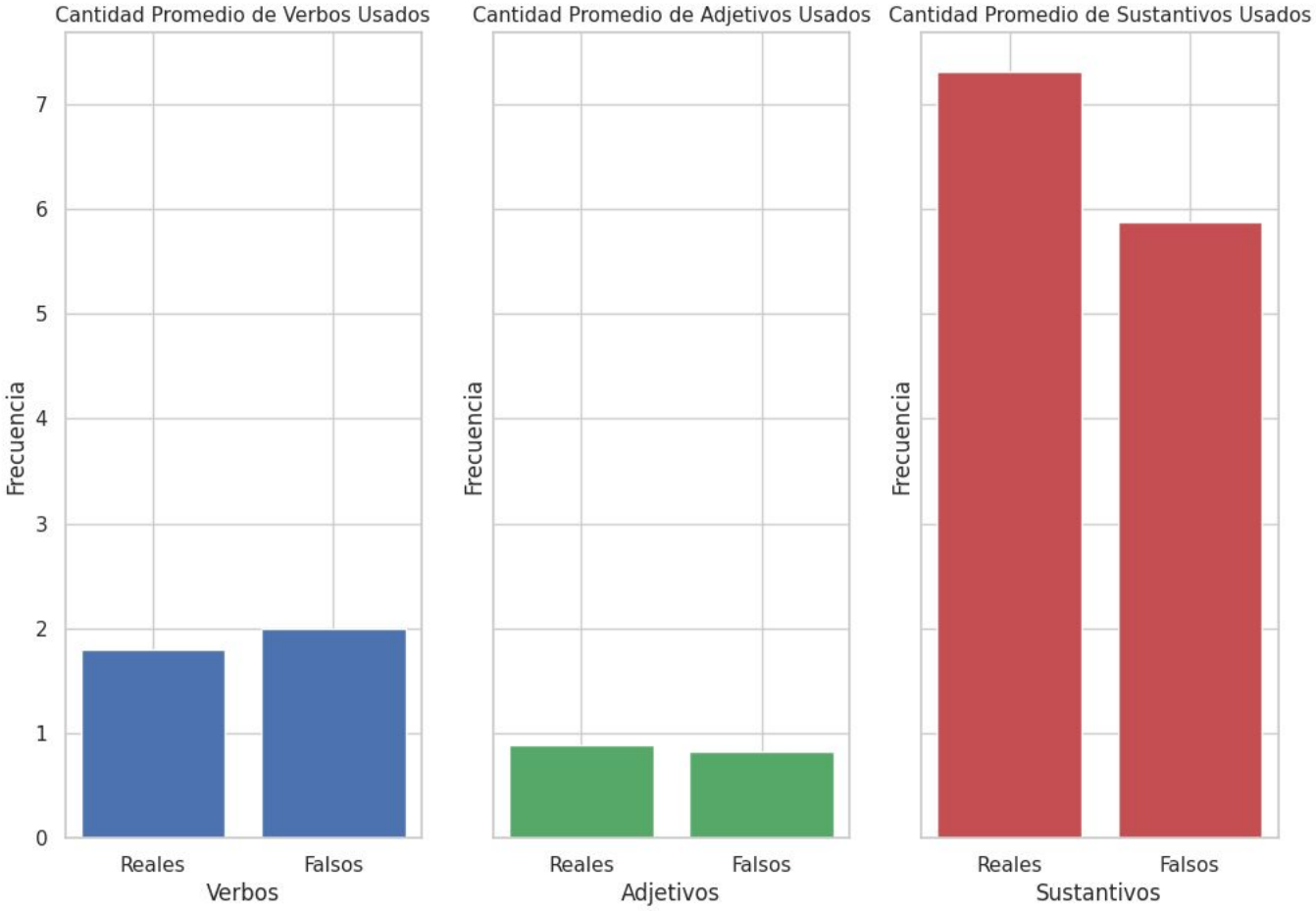

Figura 4.4: Cantidad Promedio de Verbos, Adjetivos y Sustantivos usados

Se ve a simple vista que salvo en los Verbos, el promedio de Sustantivos y Adjetivos es mayor en los tweets reales que en los falsos, habiendo una amplia diferencia en el primero, mientras que en el segundo es prácticamente ínfima. En cuestión con los Verbos es a la viceversa, el promedio es mayor en los tweets falsos antes que en los reales, aunque como sucede con los Adjetivos, la diferencia no es mucha.

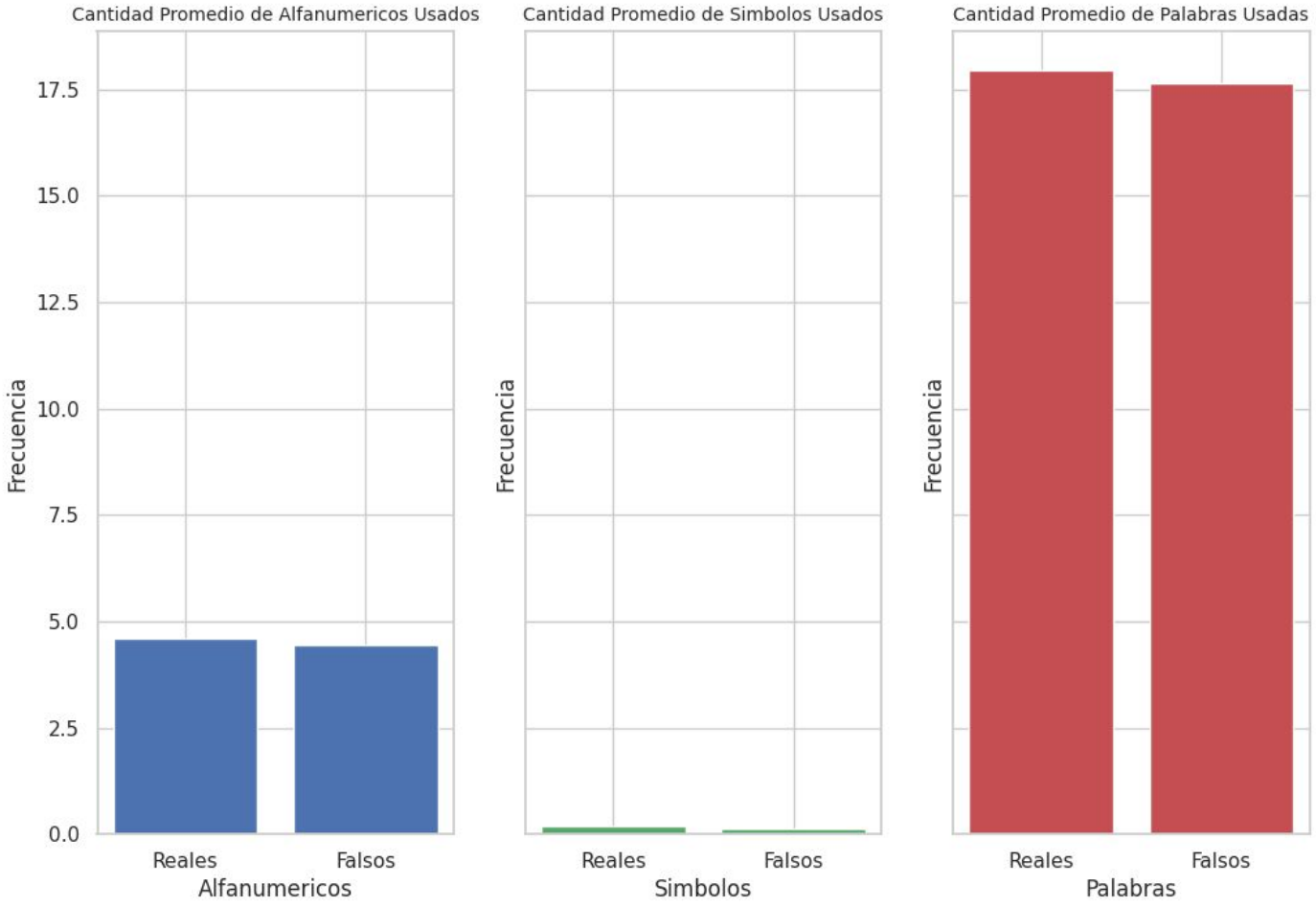

Figura 4.5: Cantidad Promedio de Alfanuméricos, Símbolos y Palabras usados

Con respecto a los Alfanuméricos, Símbolos y Palabra, el promedio es mayor siempre en los reales, lo de las palabras se relaciona directamente con la longitud de los mismos, como se vio en el Análisis General, los tweets más largos suelen ser reales, por lo que tiene sentido que el promedio de palabras sea mayor en los reales antes que en los falsos.

4.3. Verbos

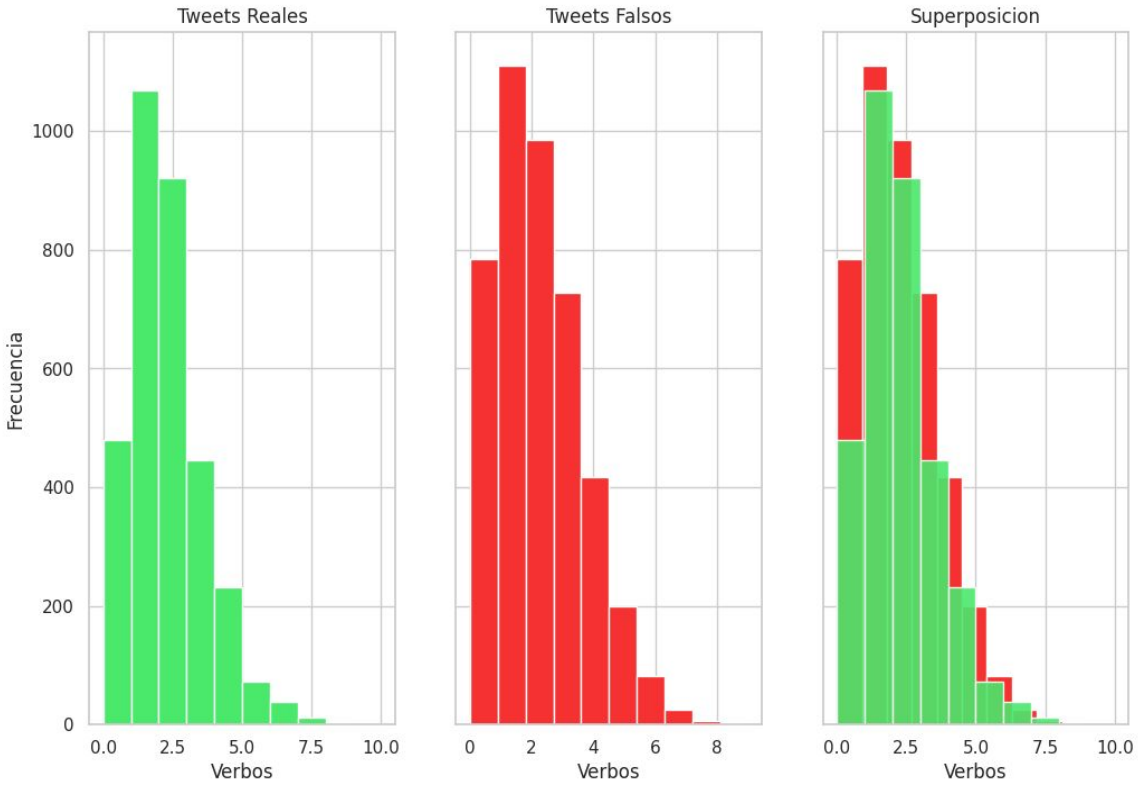

Figura 4.6: Histograma de Verbos Por Target

En el histograma no se aprecia ninguna diferencia notable entre la cantidad de verbos que se utilizan cuando se redacta un tweet sea falso o verdadero, pero se puede ver la cantidad de verbos que suelen utilizarse que está entre uno a tres verbos

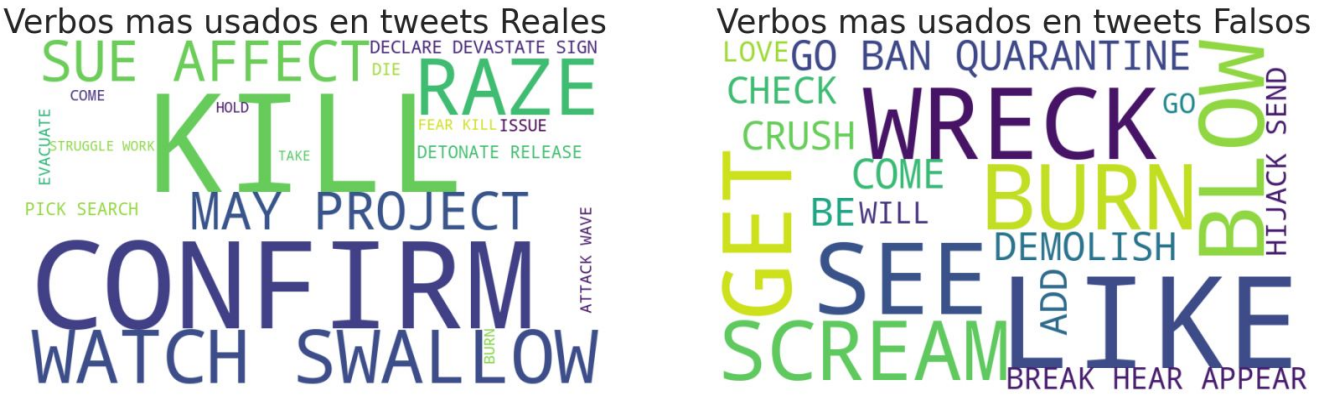

Figura 4.7: Verbos más usados en Tweets Reales y Falsos

Estos son los 20 verbos mas usados para cada caso ya sean tweets verdaderos o falsos, estos verbos han sido preprocesados a su forma infinitiva, ya que se encontraban muchas veces conjugados en sus distintas variantes

4.4. Adjetivos

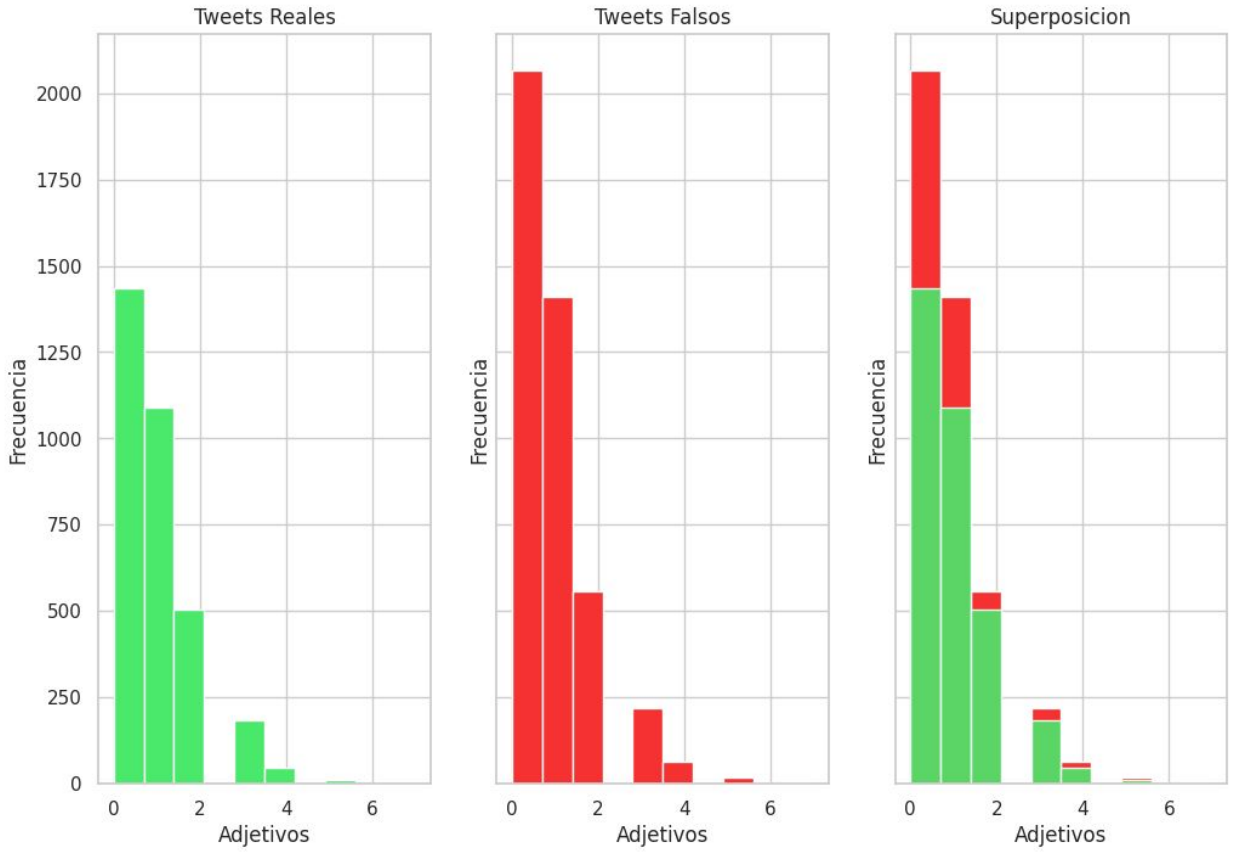

Figura 4.8: Histograma de Adjetivos Por Target

En la cantidad de adjetivos que se utilizan al redactar un tweet, parece haber una gran cantidad de tweets falsos donde no se escriben adjetivos en los mismos, que es un poco menos del doble de veces más grande que la cantidad de tweets verdaderos en donde no se usaron adjetivos para los casos donde se escribieron al menos uno o más, no hay una diferencia notable por lo que no podría sacar alguna conclusión de la veracidad del mismo en esos casos

Figura 4.9: Adjetivos más usados en Tweets Reales y Falsos

Estos son los 20 adjetivos mas usados para cada caso ya sean tweets verdaderos o falsos, estos verbos han sido preprocesados a su forma infinitiva, ya que se encontraban muchas veces conjugados en sus distintas variantes

4.5. Sustantivos

Figura 4.10: Sustantivos más usados en Tweets Reales y Falsos

Estos son los 20 sustantivos más usados para cada caso ya sean tweets verdaderos o falsos, estos verbos han sido preprocesados a su forma infinitiva, ya que se encontraban muchas veces conjugados en sus distintas variantes

La utilización de la palabra FIRE en los tweets falsos es notoria por lo menos en el set de datos con el que contamos

4.6. Alfanuméricos

Figura 4.11: Alfanuméricos más usados en Tweets Reales y Falsos

Estos son los 20 palabras alfanuméricas usadas para cada caso ya sean tweets verdaderos o falsos, estos verbos han sido preprocesados a su forma infinitiva, ya que se encontraban muchas veces conjugados en sus distintas variantes

La cantidad de signos de pregunta que se ponen en los tweets falsos parece ser mucho mayor que en los verdaderos junto con el NOT que esta puesto como alfanumérico pero lleva un símbolo junto por eso figura aca

4.7. Símbolos

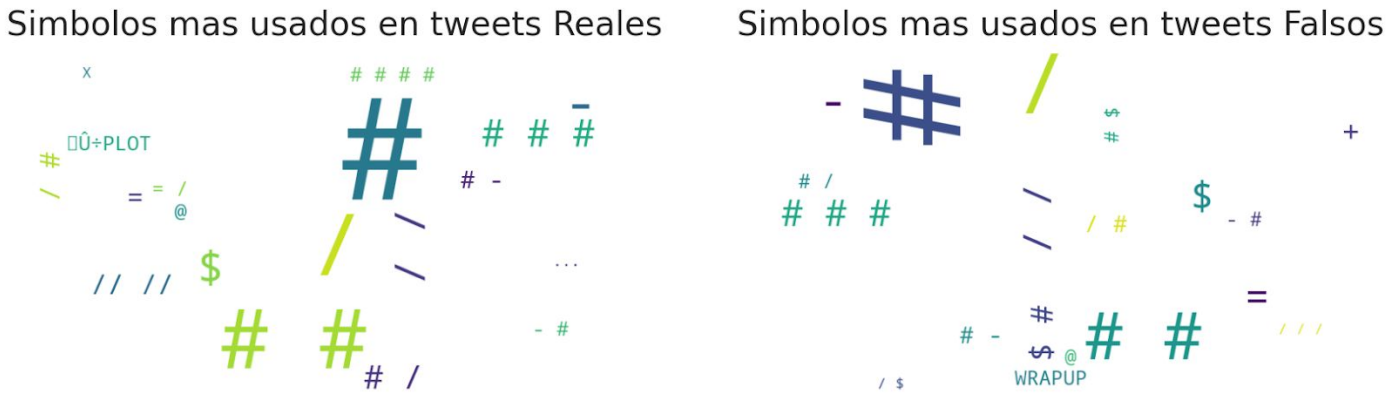

Figura 4.12: Símbolos más usados en Tweets Reales y Falsos

Estos son los 20 símbolos más usados para cada caso ya sean tweets verdaderos o falsos, estos verbos han sido preprocesados a su forma infinitiva, ya que se encontraban muchas veces conjugados en sus distintas variantes

No parece tener alguna diferencia es bastante igual en los dos casos

4.8. Por Locación

Las diez localizaciones de donde provienen la mayor cantidad de tweets son UK ,New York, Nigeria, India, USA, Unite States, Los angeles CA, Mumbai, London y Canada

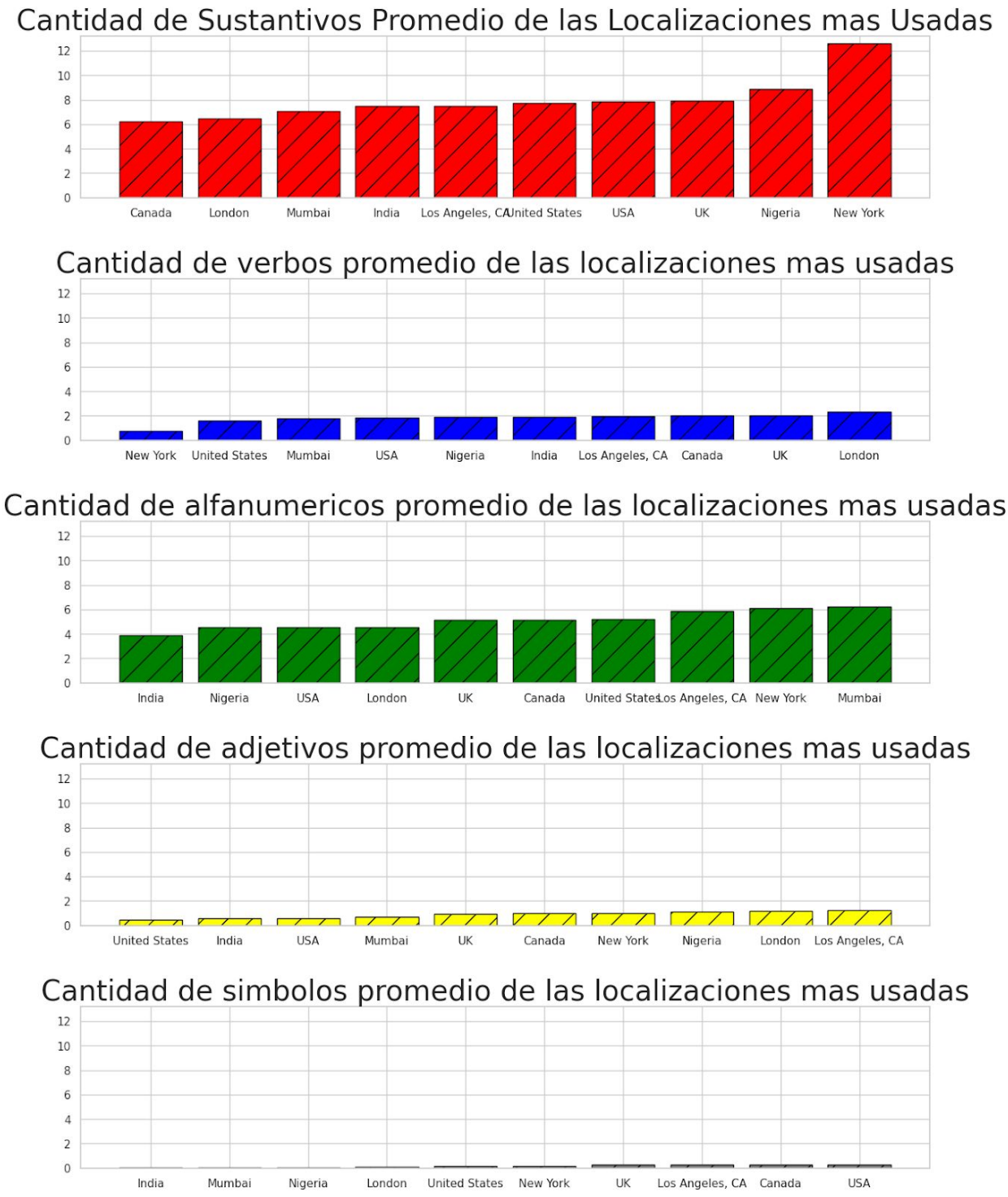

Figura 4.13: Cantidad de Sustantivos, Adjetivos y Verbos Promedio de las Locaciones más Usadas

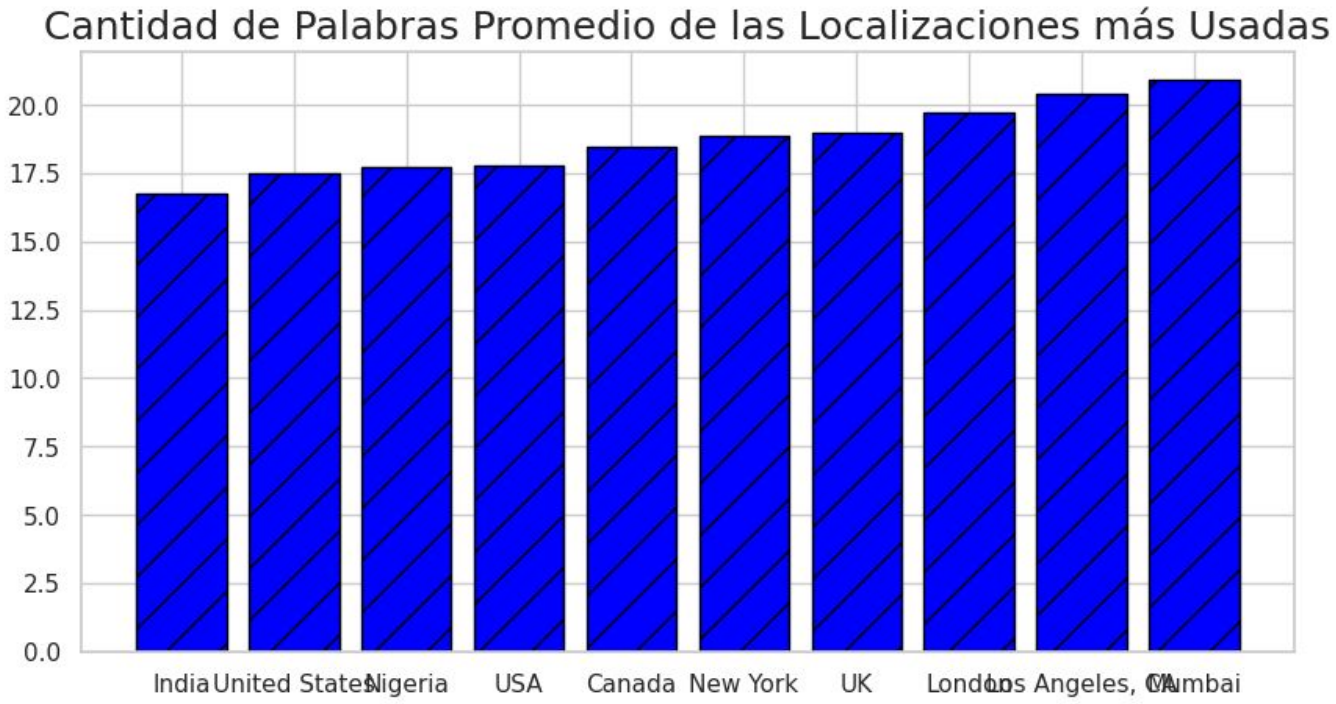

Figura 4.14: Cantidad de Palabras Promedio de las Locaciones más Usadas

4.9. Conclusión General

Viendo el análisis de las palabra en general, like fue la más utilizada, seguido por fire. Si nos enfocamos en las palabras en los tweets reales, kill es por lejos la más popular, seguida muy por detrás por suicid, en cambio, en los falsos, love es la que más veces apareció. Con respecto a la cantidad promedio, salvo en los Verbos, la cantidad promedio se encuentra en los tweets reales y no en los falsos, siendo en los Sustantivos donde la diferencia es mucho más grande que en los demás, así mismo, notamos que, como se podía especular de la longitud de los tweets, la cantidad de palabras es mayor en los tweets reales que en los falsos.En cuanto a los verbos, ni hay ninguna diferencia marcada en la cantidad que se utilizan en los tweets reales o verdaderos, y se pudo ver que kill es el verbo más utilizado en los tweets reales y like en los falsos. En relación a los adjetivos, parece haber una gran cantidad de tweets falsos donde no se escriben adjetivos, siendo Dead el adjetivo más utilizado en los tweets reales y Good en los falsos. Referido a los sustantivos, no hubo ninguno que se destacara en los tweets reales mientras que en los falsos Fire fue el que tomó la delantera con amplia diferencia. Con los alfanuméricos ocurre algo similar, en los tweets reales no hay ninguno que se destaque mientras que en los falsos NOT parece ser el de mayor predominancia, en cuanto a los símbolos, no pudimos sacar ninguna conclusión ya que los resultados son muy similares. Con respecto a las Locaciones, vemos que en New York esta la cantidad de sustantivos promedios más grande, en London la cantidad de verbos, en Mumbai la de alfanuméricos, en Los Ángeles la de adjetivos y en USA la de símbolos. Así mismo, la mayor cantidad de palabras promedio se encuentra en Mumbai.

5. Caracteres

Una de las cuestiones que se pueden analizar son los distintos caracteres que pueden aparecer en los Tweets, tales que denotan un Hashtag(#),si tiene una pregunta (? o ¿), una exclamación (! o ¡), una mención (@) o un link a alguna página y/o video ('https').

5.1. Cantidad de Tweets

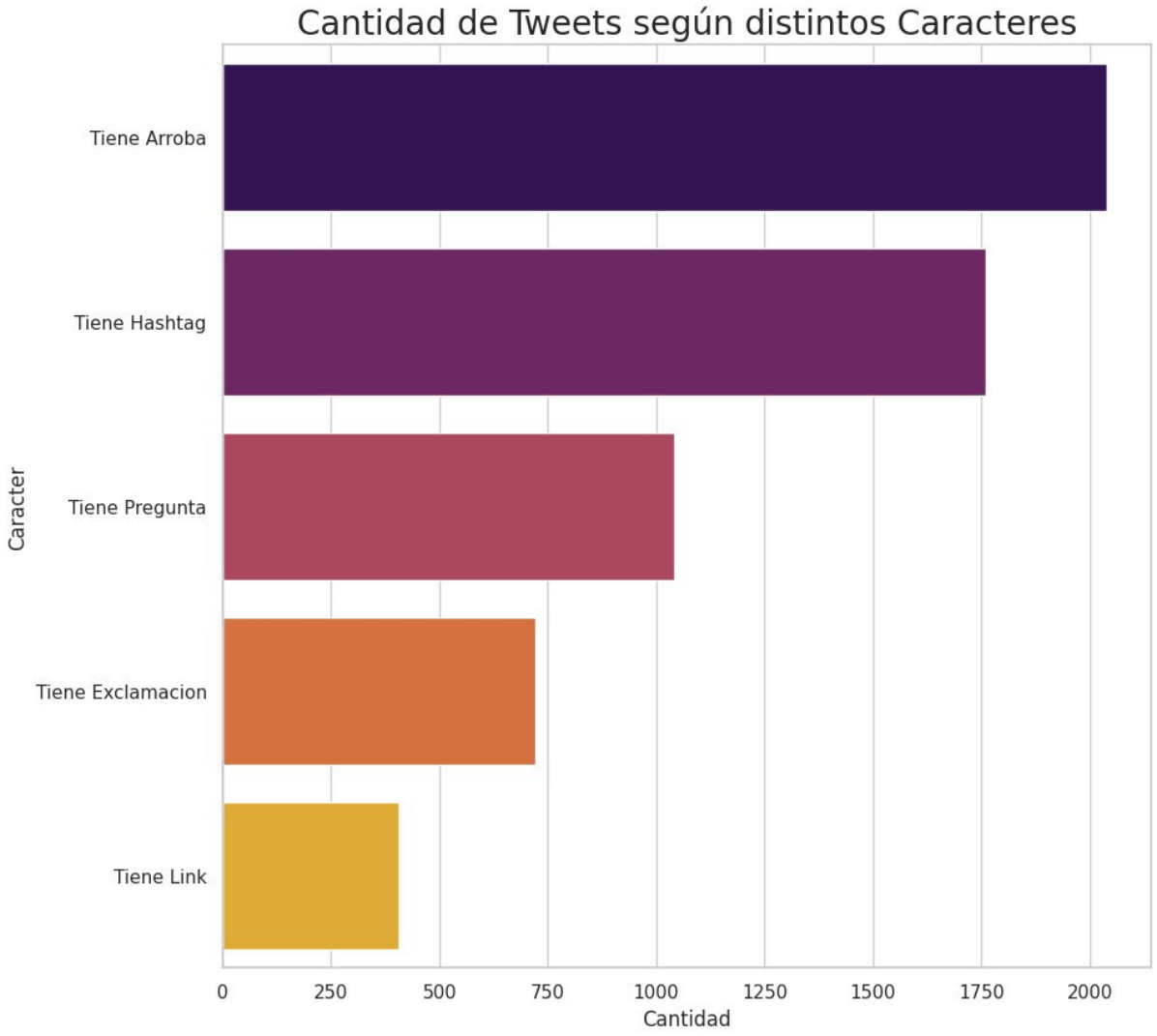

Figura 4.1: Cantidad de Tweets según distintos Caracteres

Vemos que casi 2000 tweets tiene por lo menos un @, lo que simboliza que tiene una mención de algún tipo, luego le sigue el #, relacionado con los hashtags. Esto tiene sentido ya que es una red social en la que estos caracteres son muy utilizados, el @ ya que se

menciona a alguna cuenta en particular y el # para hacer un Hashtag tendencia, lo cual genera que a la gente le aparezca en su Time Line y se vuelva genere algún tipo de revuelo. Los demás símbolos, tales como los de pregunta y exclamación aparecen en menor medida lo cual nuevamente tiene sentido, y por último los tweets que tienen links son realmente muy pocos, nuevamente nos parece normal ya que no es la principal utilización de esta red social la propagación de videos.

5.2. Porcentaje de Tweets Reales y Falsos

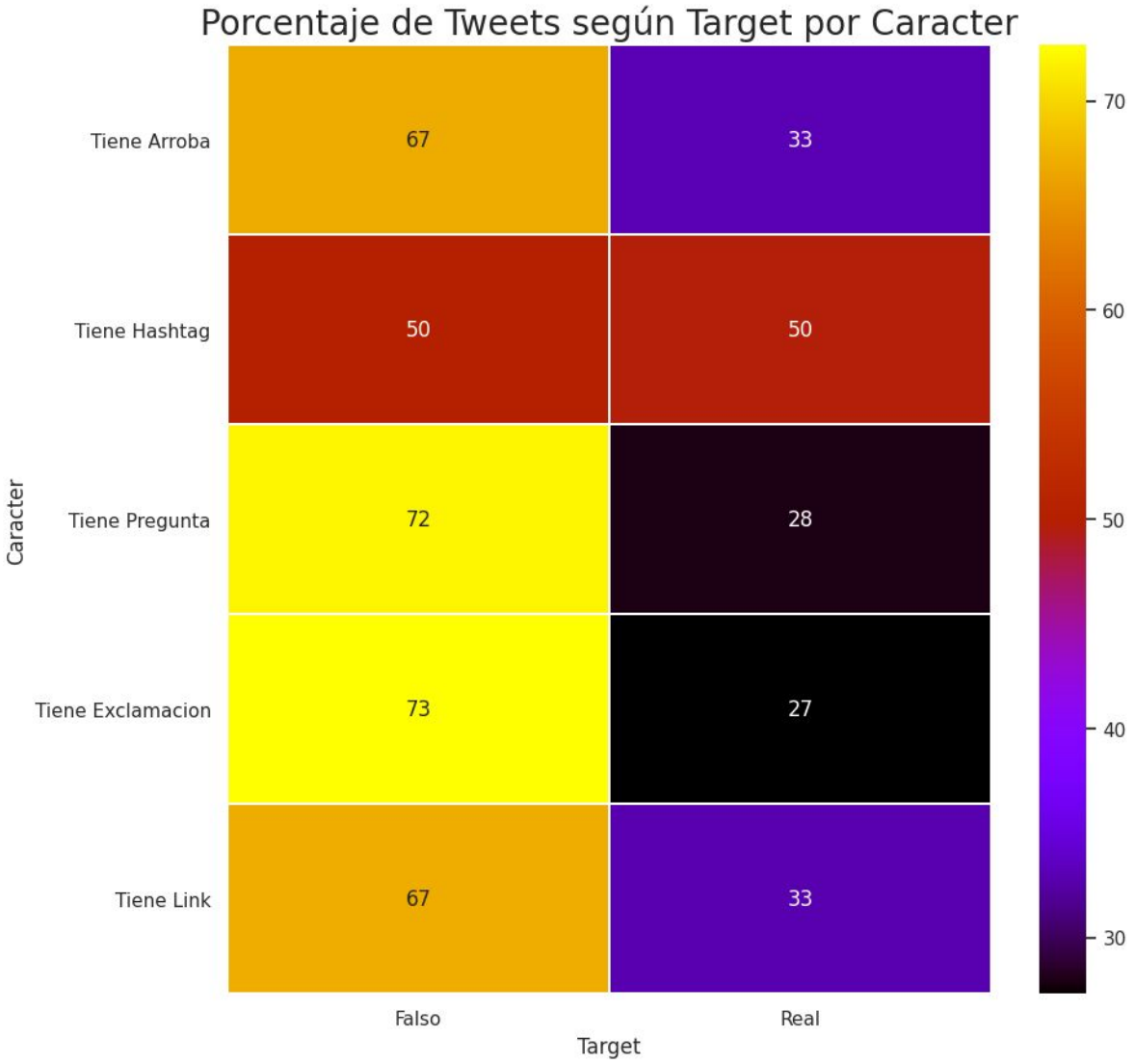

Figura 4.2: Porcentaje de Tweets según Target por Carácter

Aquí se pueden ver unas particularidades, en todos los casos salvo con los Hashtag, el porcentaje de tweets falsos es superior al de tweets reales, y todos con un porcentaje muy similar, siendo alrededor de un 30% el porcentaje de veracidad y un 70% de falsedad. Con respecto a los Hashtag, la distribución es equitativa, por lo que sería interesante analizarlo en profundidad.

5.3. Relación entre Caracteres

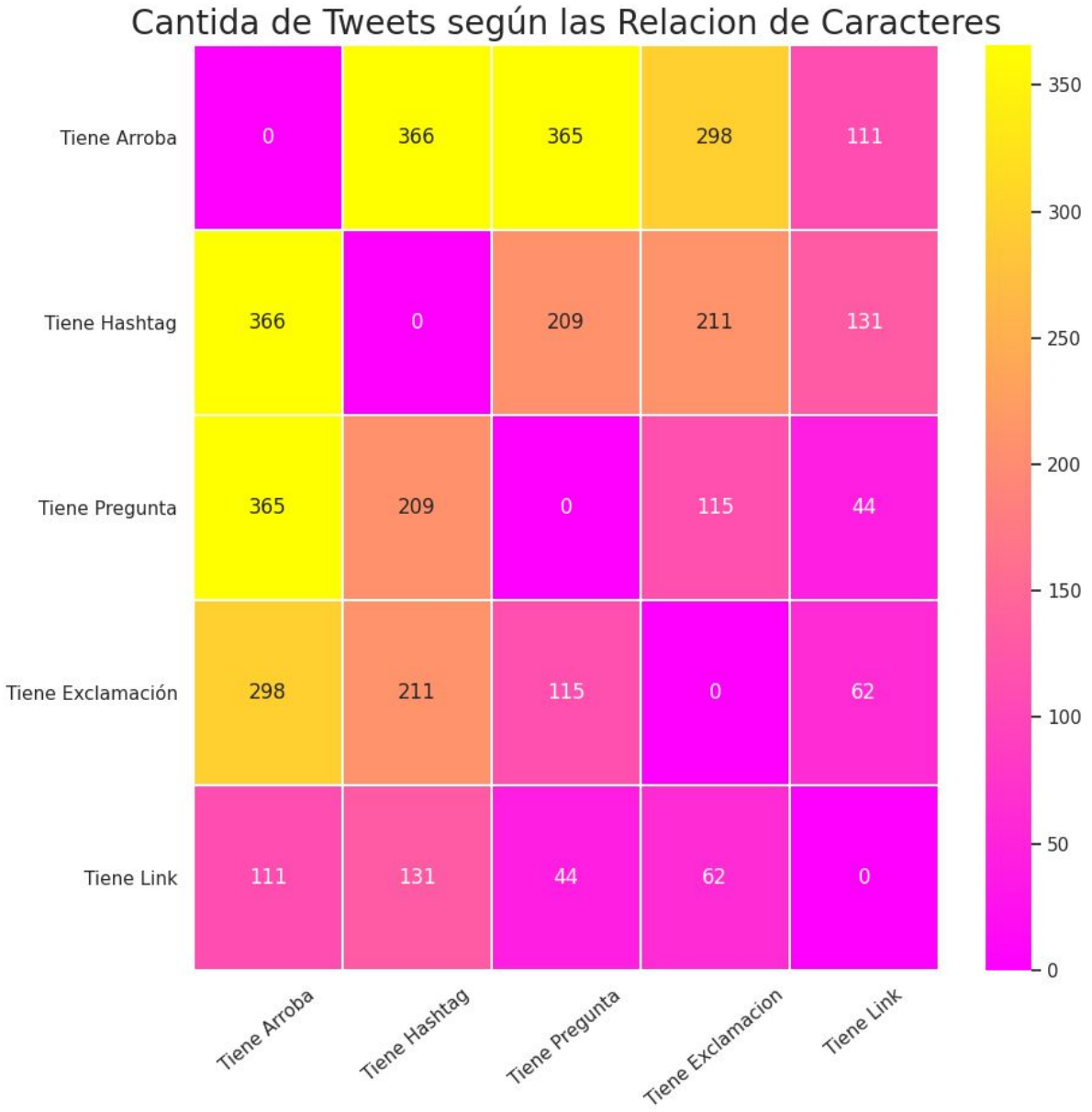

Figura 4.3: Cantidad de Tweets según la Relación entre Caracteres

En este gráfico veremos la relación entre los caracteres, podemos ver que los tweets con menciones son los que más relación tiene con los demás símbolos, relacionándose principalmente con el carácter '#' y '?' o '¿'. Con respecto a los demás, se ve que la relación no es tan grande, siendo las que tienen links la que menor relación tienen con el resto.

Ahora bien, visto estos caracteres especiales, sería interesante analizar el caso particular de los dos que son más utilizados en esta red social, el de mención (@) y el de Hashtag (#).

5.4. Hashtag

Una de las particularidad de las redes sociales es el uso de hashtags, palabras que se encuentran luego del símbolo '#' con el fin de ser utilizado como una 'etiqueta'. Particularmente en Twitter, es donde es más utilizada por lo que cabe analizar que sucede con los tweets en función de la cantidad de hashtags que tengan y cuales son los más populares.

5.4.1. Cantidad de Tweets

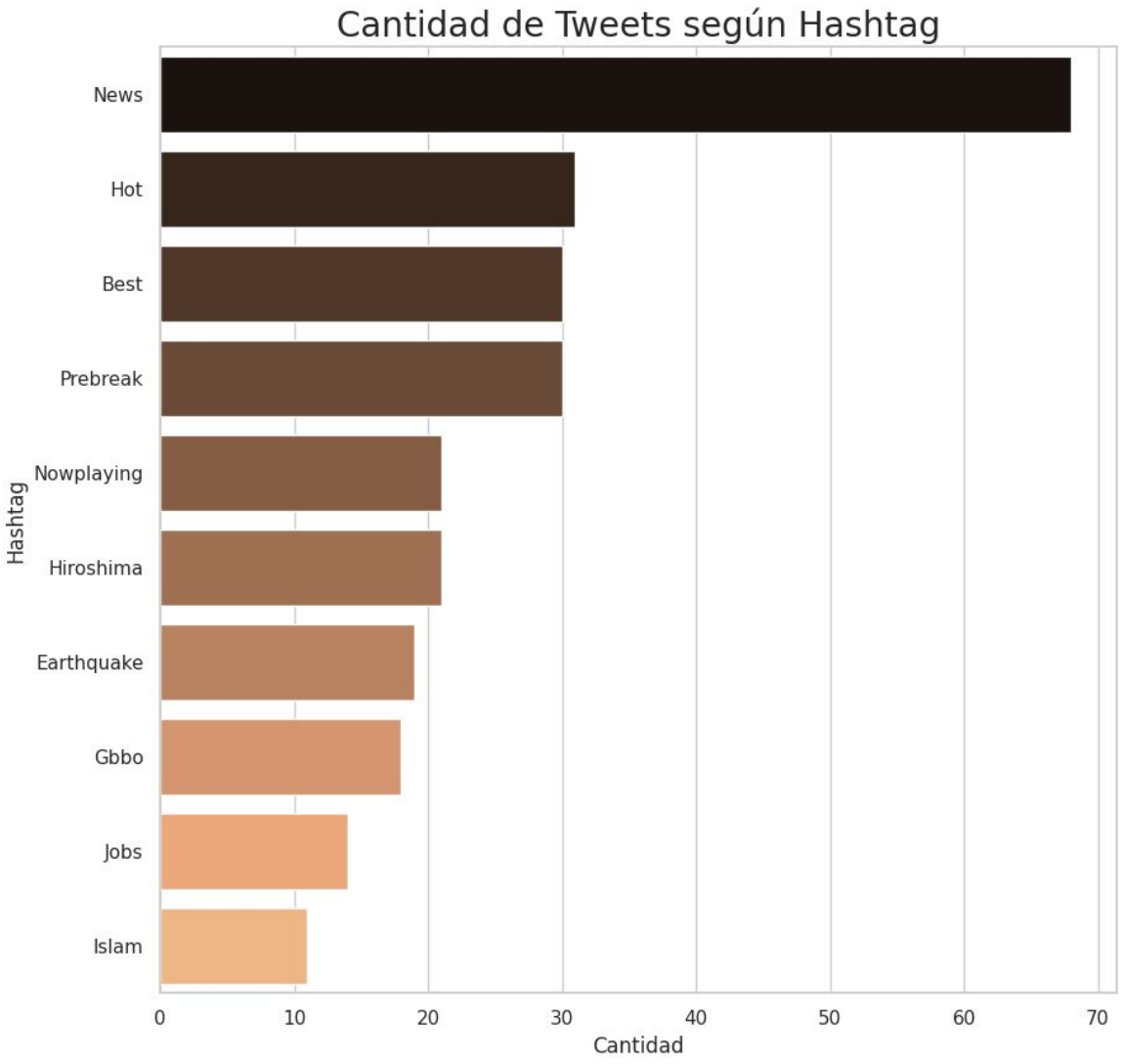

Figura 4.4: Cantidad de Tweets según Hashtag

A simple vista se observa que #News es el hashtag que se encuentra en la mayor cantidad de tweets con un valor de 70, seguido muy por detrás por #Hot, con una cantidad 30. Analizando un poco, se puede deducir que la mayor diferencia es entre el primero y el segundo, mientras que ya con los demás la diferencia es realmente muy poca.

5.4.2. Porcentaje de Tweets Reales y Falsos

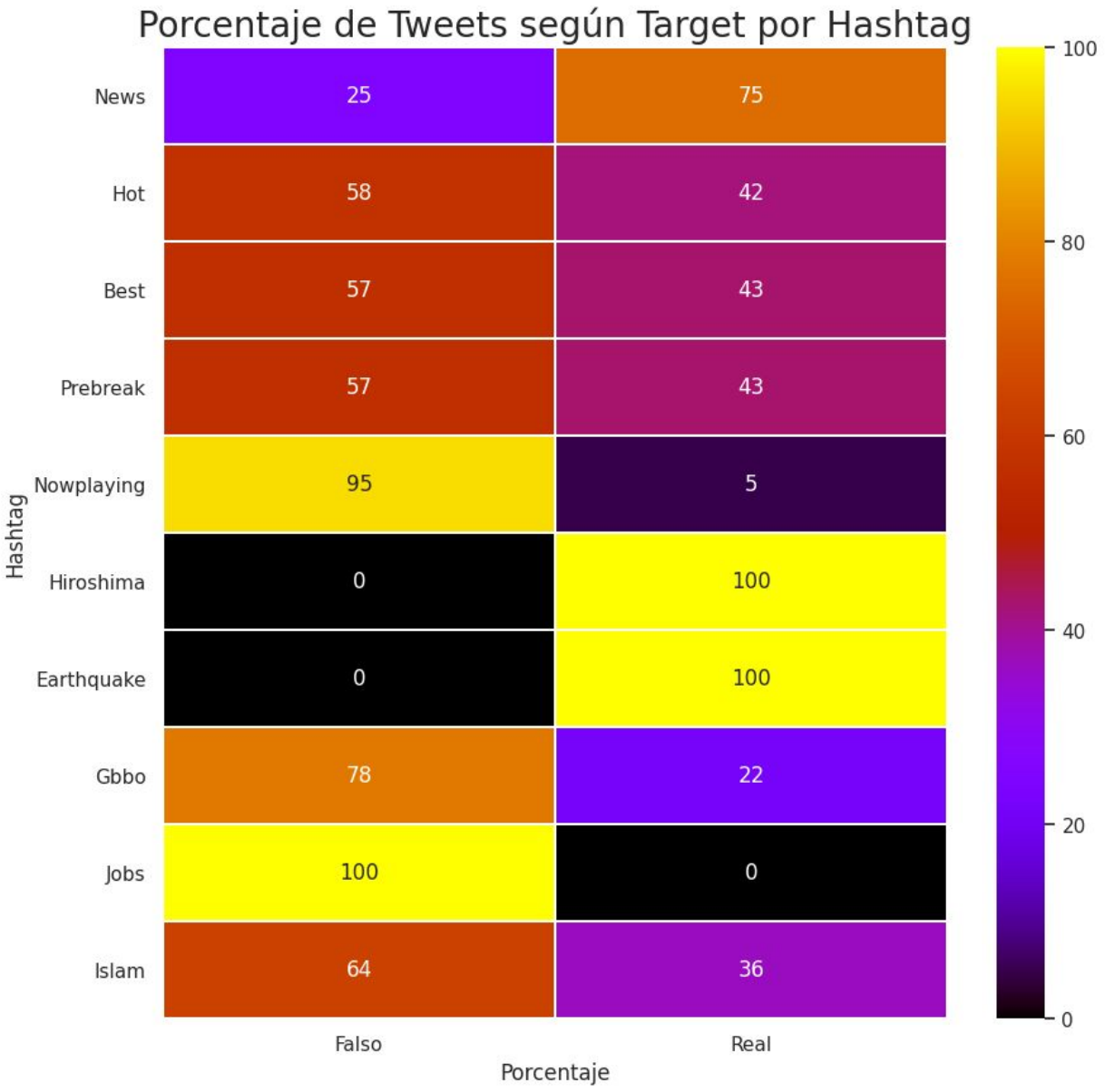

Figura 4.5. Porcentaje de Tweets según Target por Hashtag

Podemos ver que 3 de los 10 hashtag más utilizados tienen un porcentaje de tweets reales más grande que los de los falsos, siendo en #Hiroshima y #Earthquake tienen un 100% de tweets veraces mientras que en #News el porcentaje de veracidad es del 75%, este es el más importa ya que es el hashtag más utilizado. Luego, en los demás, el porcentaje de tweets

falsos es mayor, siendo en #Jobs y #Nowplaying en los que hay mayor diferencia, mientras que los otros, mantienen una paridad entre ambos porcentajes.

5.4.3. Relación entre Cantidad de Hashtags y Veracidad

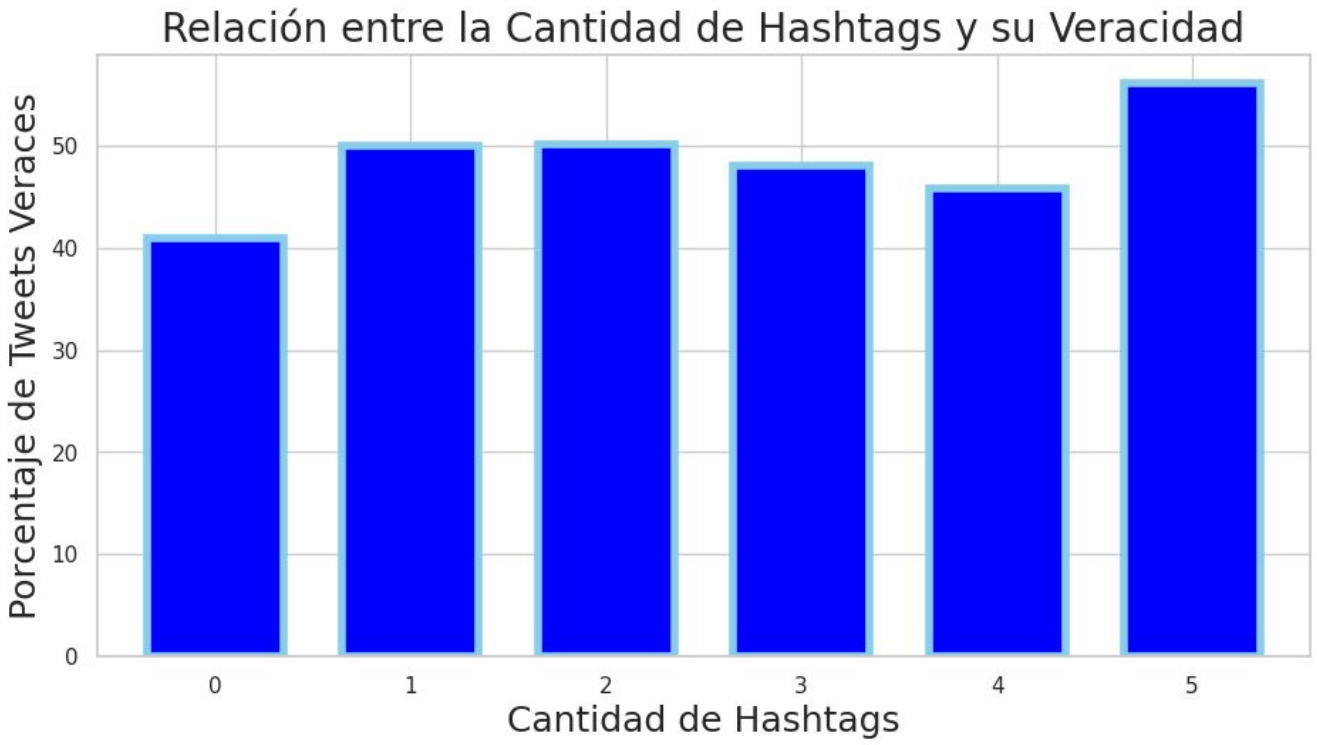

Figura 4.6. Relación entre la cantidad de hashtags y su veracidad

Nos pareció interesante analizar si había una relación entre el número de hashtag en un tweet y el porcentaje de tweets veraces. Sin embargo no parece haber un patrón claro.

5.5. Signos de exclamación

En una suposición previa al análisis, pensamos que los tweets que tienen signos de exclamación no tienden a ser noticias serias. Más aún, pensamos que mientras más signos de exclamación tiene un tweet, más burdo es el intento de llamar la atención, y es más probable que sea falso. El siguiente gráfico parece apoyar esta idea:

Figura 4.7. Relación entre la cantidad de signos de exclamación y su veracidad

5.6. Caracteres en mayúscula

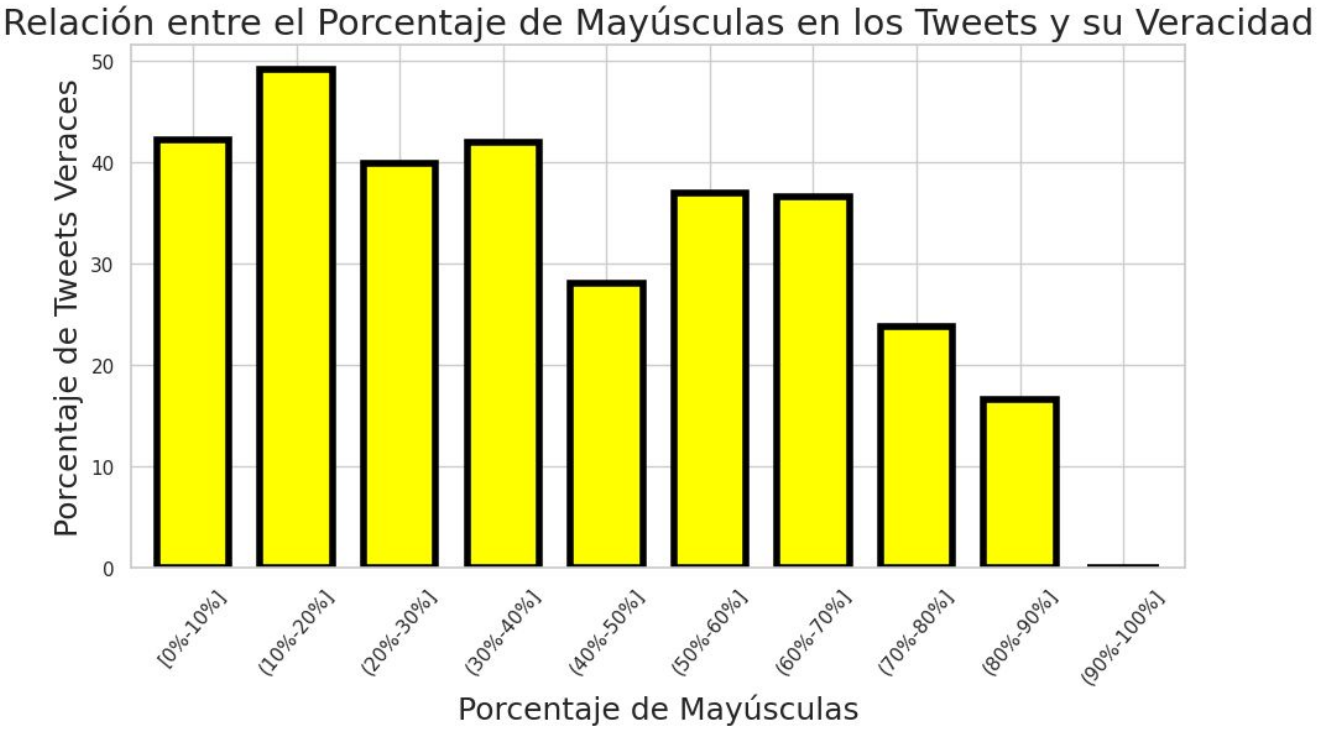

Figura 4.8. Relación entre el porcentaje de mayúsculas y su veracidad

Nos pareció interesante analizar los tweets escritos en mayúscula. Nuevamente, asociamos esta burda manera de hacer llamativos los tweets con la falsedad de estos. Y el análisis parece corroborar nuestra hipótesis: mientras más caracteres en mayúsculas tienen los tweets, su porcentaje de veracidad disminuye.

5.7. Cuentas

Ya hicimos el análisis correspondiente con los Hashtags, ahora lo haremos con las cuentas mencionadas a través del valor @.

5.7.1. Cantidad de Tweets

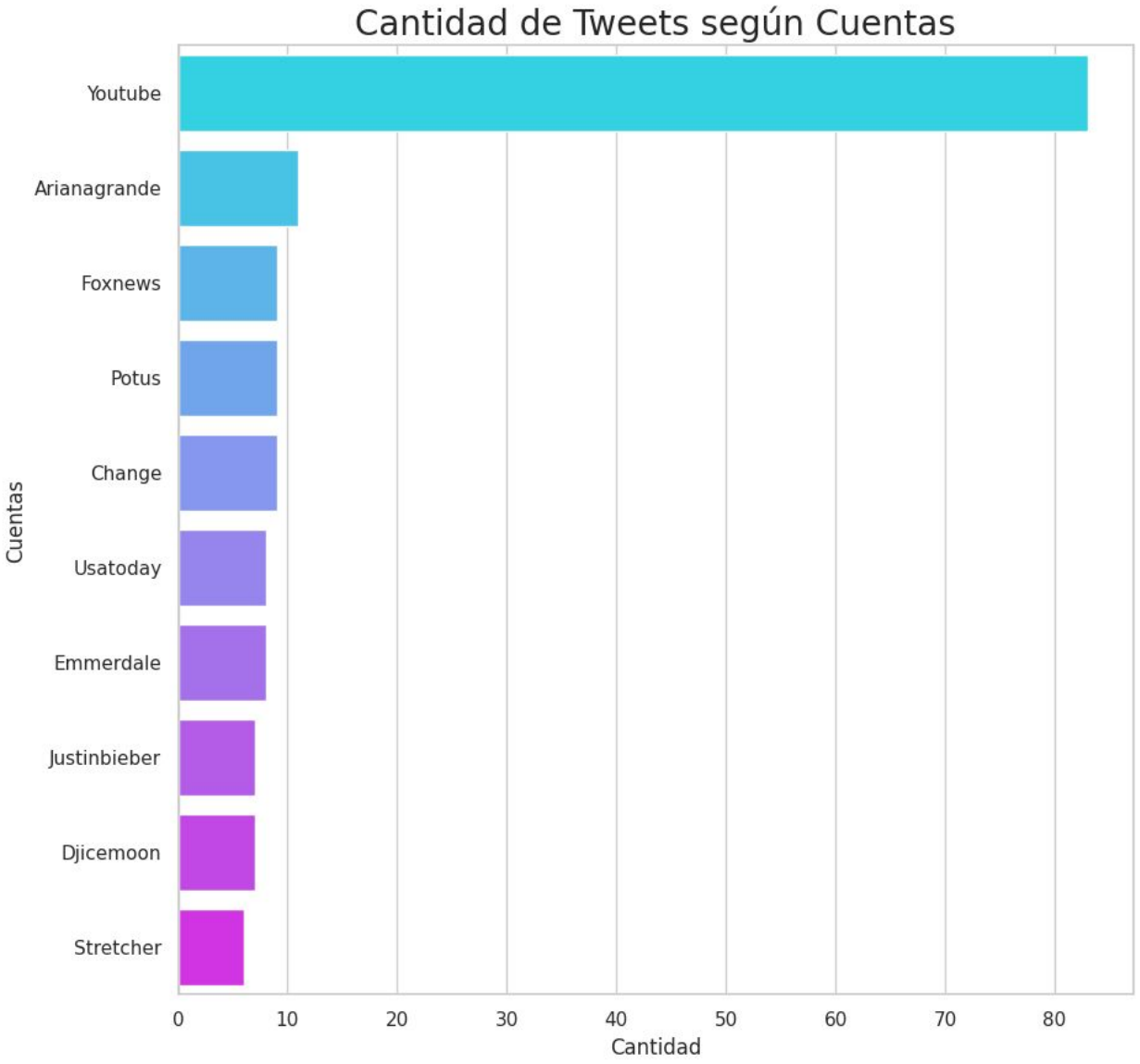

Figura 4.9: Cantidad de Tweets según Cuentas

Con una amplia mayoría @Youtube es la cuenta más mencionada en todos los tweets, con una cantidad cercana a los 85, luego seguido por muy lejos, @Arianagrande y @Potus, este último nos puede parecer extraño ya que como vimos anteriormente, la mayor

cantidad de tweets provienen de los Estados Unidos. Así mismo, se observa que, salvo la diferencia entre el primero y el segundo, la diferencia entre los demás en muy pequeña.

5.7.2. Porcentaje de Tweets Reales y Falsos

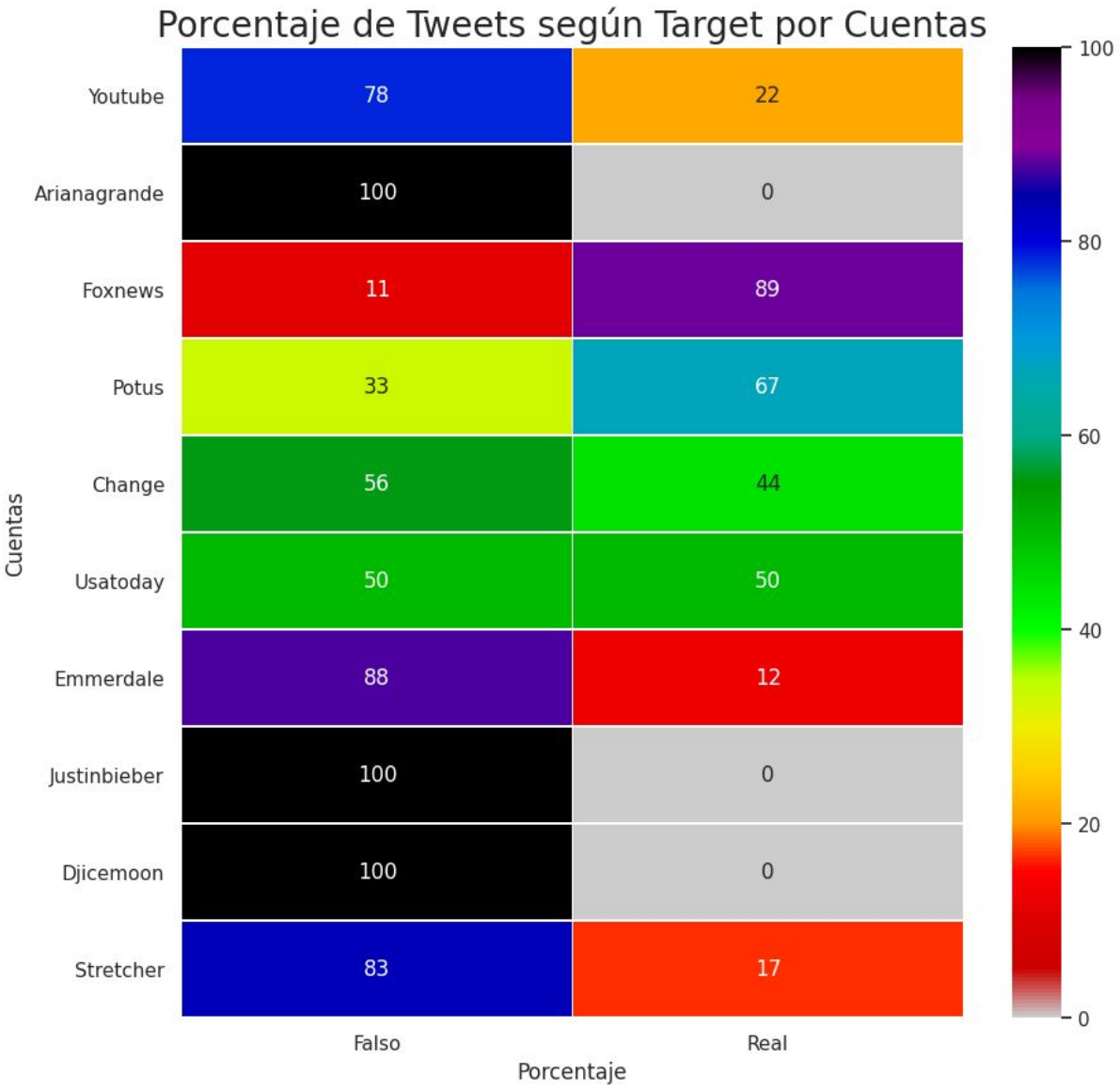

Figura 4.10: Porcentaje de Tweets según Target por Cuentas

Acá se sigue la tendencia vista con los Hashtags, en la gran mayoría de estas cuenta mencionadas se encuentran en tweets falsos, siendo @Arianagrande, @Justinbieber y @Djicemoon las cuentas con un 100% de porcentaje de tweets falsos. En la cuenta más mencionada también predomina un porcentaje de tweets falsos mayor al verdadero, con una diferencia de asi un 55%. Luego vemos casos como @Potus o @Foxnews en el que el porcentaje de tweets verdaderos es mucho mayor al porcentaje de los falsos, si quisiéramos

buscar una relación entre esto y los Hashtag, podríamos deducir que existe una entre #News y @FoxNews, ya que en ambos el porcentaje de tweets verdaderos es mayor y bastante similar.

5.7.3. Relación entre Cantidad de Arroba y Veracidad

Figura 4.11: Relación entre el carácter @ y su veracidad

Del mismo modo que hicimos con los signos de exclamación, analizamos los símbolos de arroba. Con los tweets que tienen de una a tres menciones, no parece haber un patrón claro de comportamiento. Pero luego, mientras más menciones tienen más baja su porcentaje de veracidad. Así como concluimos con los signos de admiración, es posible que el abuso de menciones esté relacionado con el deseo de llamar la atención de manera burda, y esto este relacionado a la falsedad de los mismos.

5.8. Conclusión General

Pudimos ver que #News es el hashtag más utilizado y luego le sigue muy por detrás #Hot. #Japan, #Hiroshima y #Earthquake son los que tienen un porcentaje de tweets veraces del 100% y #News del 75%, mientras que #Jobs y #Nowplaying son las que presentan el mayor porcentaje de tweets falsos. @Youtube es la cuenta con la mayor cantidad de

menciones, seguida por @Arianagrande y @Potus con una diferencia muy grande. En relación a los porcentajes de tweets reales y falsos, en la gran mayoría el porcentaje de tweets falsos fue mayor al de los verdaderos, habiendo en algunos casos un 100% de porcentaje de falsos, aunque pudimos notar que en @Foxnews el porcentaje era favorable a los tweets verdaderos, por lo que se puede establecer una relación con el hashtag #News.

6. Nulos

Ahora veremos los nulos, según lo analizado, solamente los campos Keyword y Location son los únicos campos que presentan valores nulos. En relación a los Keyword nulos, solamente se pudo analizar su relación con el target, ya que la locación para ellos es siempre nula, caso contrario a lo que sucede con la locación nula, en la que si hay valores de Keyword no nulos.

6.1. Relación con Target

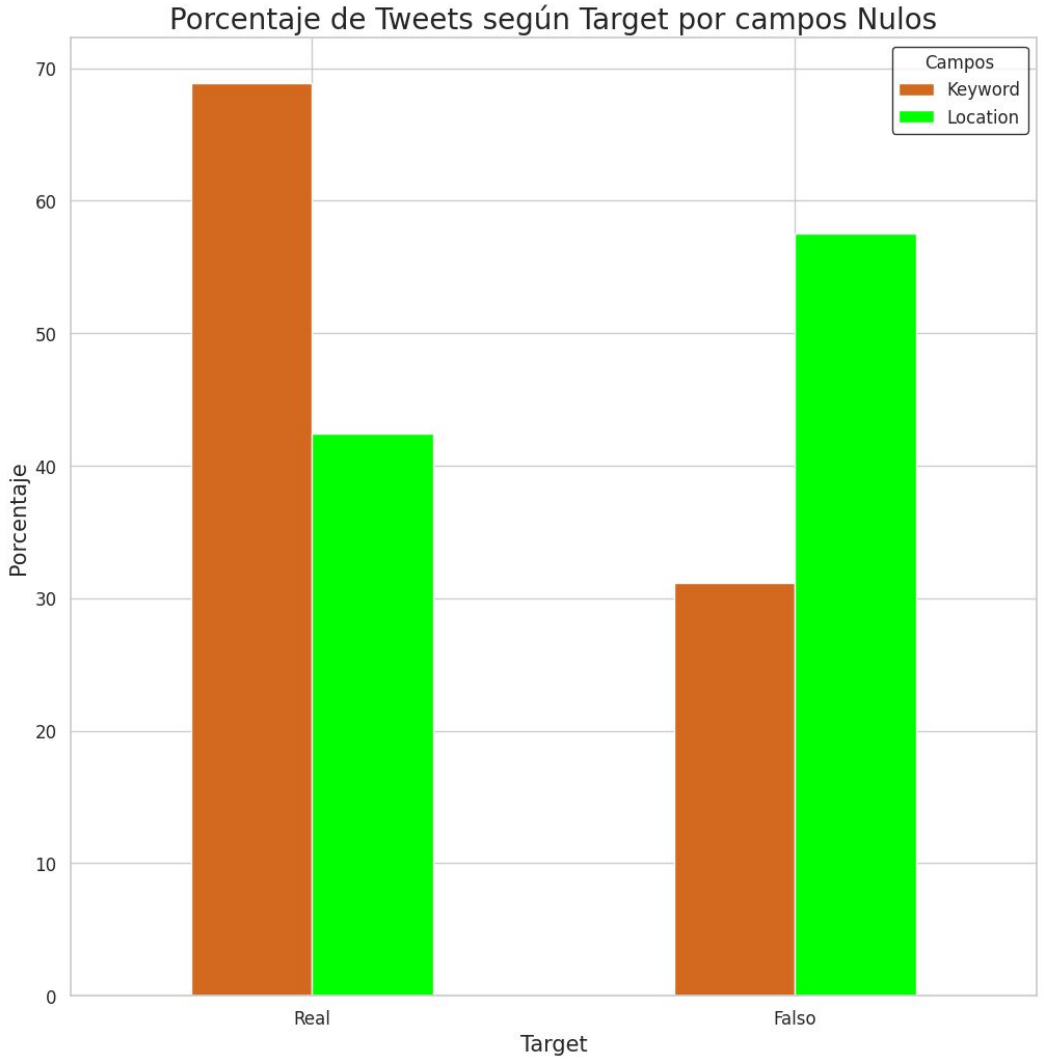

Figura 5.1: Porcentaje de Tweets según Target por campos Nulos

Sorprendentemente vemos que si el Keyword es nulo, el porcentaje de tweets verdaderos es mayor al de los falsos, con una diferencia grande de casi el 40%, caso

contrario sucede con los valores nulos en Location, en el cual, el porcentaje de tweets falsos es mayor al de los verdaderos, con una diferencia del 15%.

6.2. Locación nula con Keyword

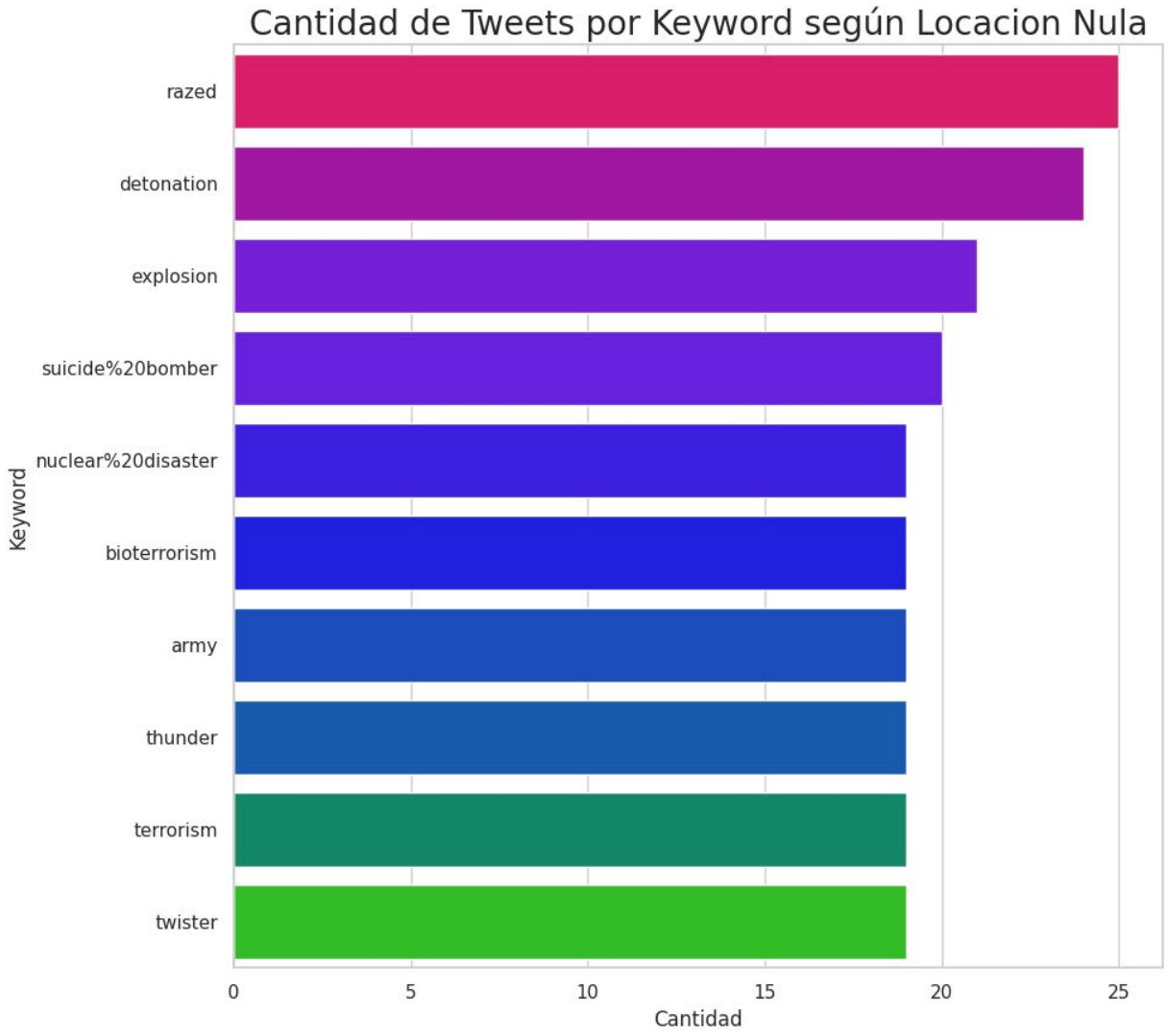

Figura 5.2: Cantidad de Tweets por Keyword según Locacion nula

Se puede ver que el keyword que más aparece en los tweets con locación nula es el de ‘razed’ con una cantidad de casi 25 tweets, seguido muy de cerca por ‘detonation’ y ‘explosion’, con unos 23 y 20 tweets aproximadamente. Se ve que se sigue la tendencia observada en keyword, en la cual la diferencia no es tan grande y la cantidad donde aparecen los principales es corta, y así mismo, ninguno de los keyword que aparecen más veces en los tweets generales se repite aquí.

6.3. Conclusión

Para concluir este capítulo, se vio que en relación al taget, los que tienen el keyword nulo tienen un porcentaje de falsedad más grande que de veracidad, con una diferencia bastante grande mientras que en el caso de que el campo location esté nulo, la relación es al revés, con una diferencia menor. Así mismo, si keyword es nulo, la locación es nula pero no a la inversa, siendo ‘razed’, ‘detonation’ y ‘explosion’ los que más veces aparecen, aunque no está en el top 10 de los keywords más populares de todos los tweets en general.

7. Keywords

Ahora veremos la relación de las Keywords más populares con los otros campos.

7.1. Por Target

7.1.1. Keyword más Frecuentes

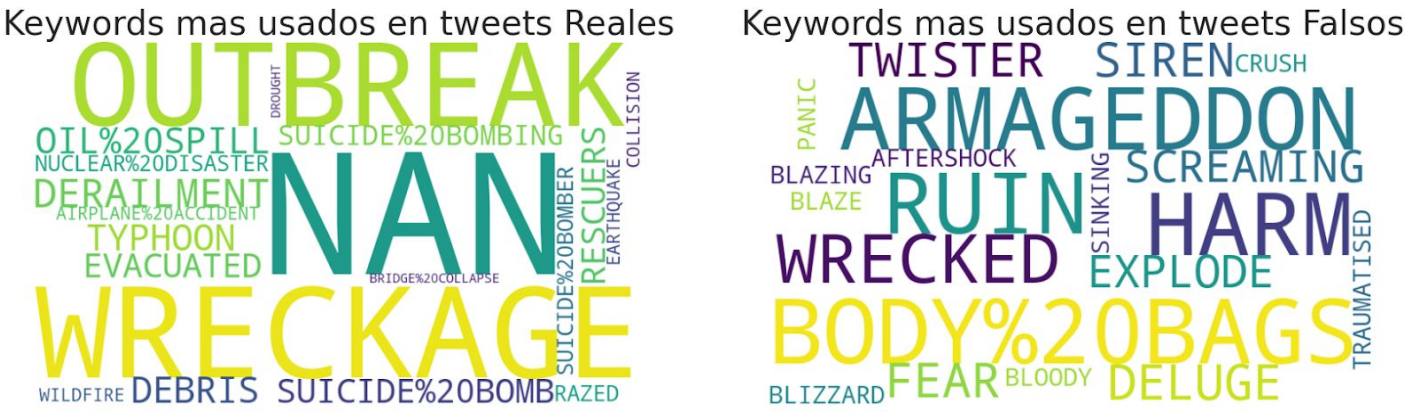

Figura 7.1: Keywords más usados en Tweets Falsos y Reales

Estos son los 20 keywords más usados para cada caso ya sean tweets verdaderos o falsos, estos verbos han sido preprocesados a su forma infinitiva, ya que se encontraban muchas veces conjugados en sus distintas variantes

7.1.2. Porcentaje de Tweets Reales y Falsos del Top 10

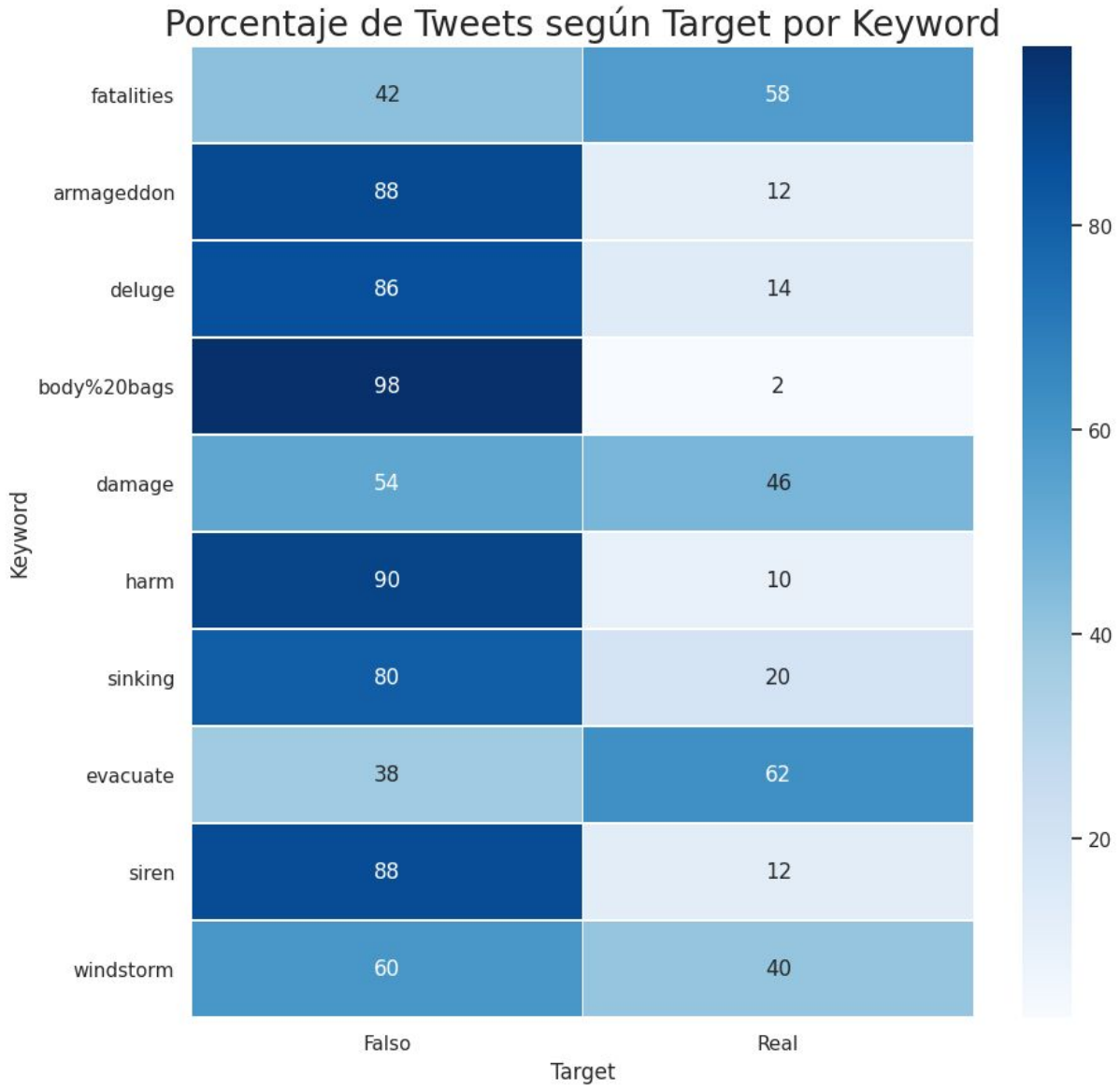

Figura 6.1: Porcentaje de Tweets según Target por Keyword

Salvo con evacuate y fatalities, en los demás keyword predomina el porcentaje de tweets falsos por sobre los verdaderos. En ‘evacuate’ es donde la diferencia es mayor, casi del 25% mientras que en fatalities es más pequeña. En cambio, en las demás, la diferencia es realmente grande, en todas, el porcentaje de tweets falsos es mayor al 80%, siendo ’body%20bags’ el que presenta la mayor diferencia, cerca del 96%.

7.2. Relación con Países principales

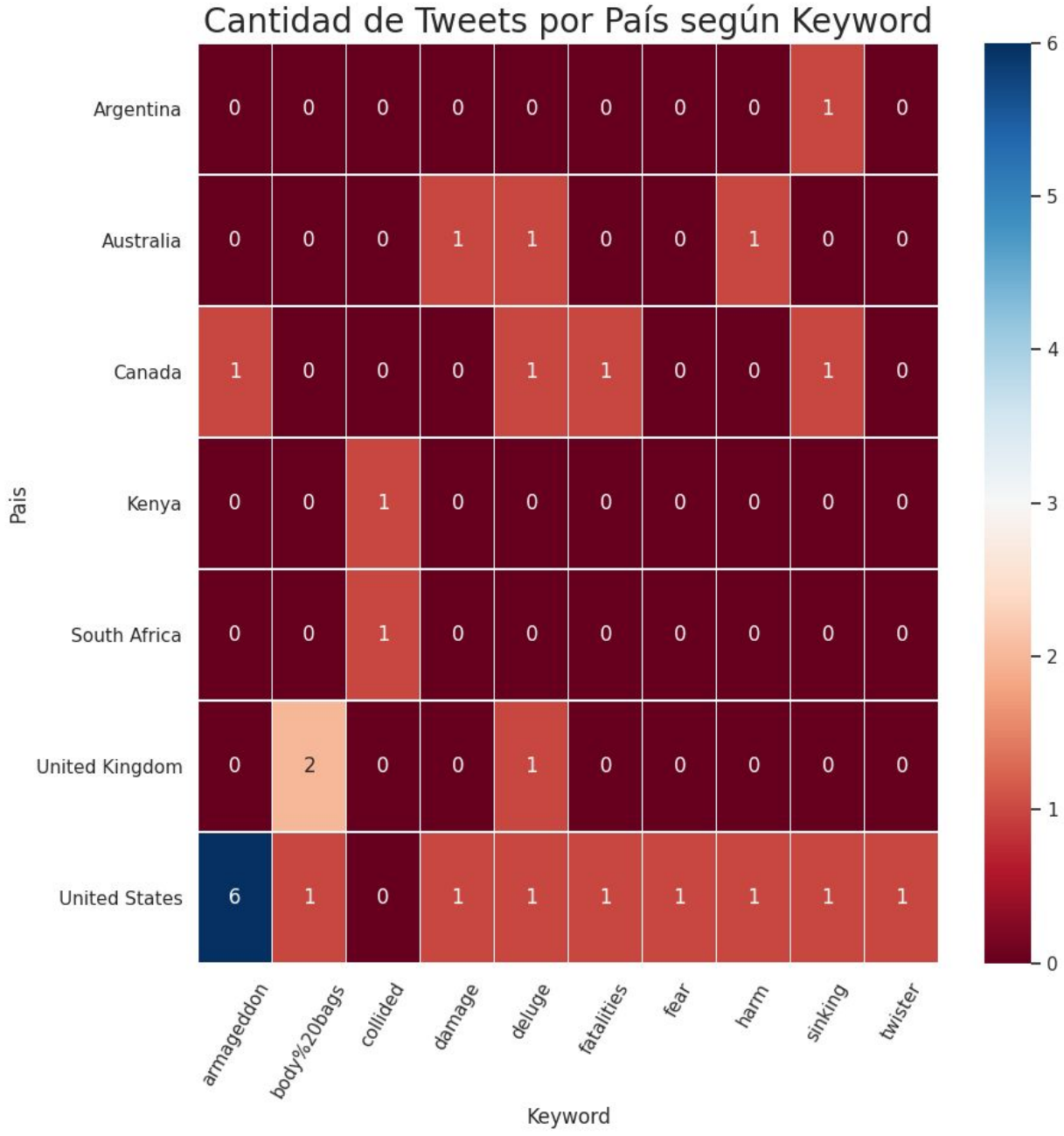

Figura 6.2: Cantidad de Tweets por País según Keywords

Si tomamos los 10 países con mayor cantidad de tweets, vemos que la distribución no nos arroja ningún resultado interesante, ya que en todos los países salvo Estados Unidos, solamente aparece un solo keyword con una solo tweet. Estados Unidos es la excepción ya que, es el que mayor cantidad de tweets tiene en general.

7.3. Conclusión General

Como estuvimos observando a lo largo del informe, la gran mayoría de los keywords principales poseen un porcentaje de tweets falsos mayor al de los verdaderos, salvo en ‘evacuate’ y ‘fatalities’, siendo este último el que predomina en la cantidad de tweets en el que aparece. En relación a los 10 países principales, la relación en casi nula, salvo en los Estados Unidos en la que hay mayor cantidad de ocurrencias, a lo cual pensamos que es debido a que es el país con mayor cantidad de tweets totales.

8. Conclusión Final

Para finalizar, a continuación se enumeran las conclusiones más importantes que logramos obtener:

- Si se completan los NaNs de Keywords, es el que más veces aparece.
- ‘'fatalities' es la keyword en la mayor cantidad de tweets.
- La diferencia entre las keywords es realmente muy pequeña.
- El 57% de los tweets son falsos mientras que el restante 43% verdaderos.
- Existe una relación entre la longitud y la veracidad.
- USA, New York y Unites States son las locaciones en crudo que más aparecen, las cuales, pertenecen al mismo país.
- Estados Unidos, Canadá e Inglaterra son los países con mayor cantidad de Tweets.
- En los países con mayor cantidad de tweets, el porcentaje promedio tanto de rales y falsos es equitativo.
- El medio oriente es la zona donde está la mayor cantidad de países con porcentaje de tweets reales.
- A nivel continente, la distribución de países con tweets reales es equilibrada.
- Brasil, Venezuela, Rusia y China son los países con mayor porcentaje de tweets falsos.
- California y Nueva York son los estados de Estados Unidos con la mayor cantidad de tweets.
- Nebraska es el estado de Estados Unidos con un mayor porcentaje de tweets verdaderos.
- Utah y Lousiana son los estados de Estados Unidos con el mayor porcentaje de tweets falsos.

- Like fue la palabra más popular, Kill la de los tweets reales y Love la de los falsos.

- Salvo en los Verbos, la cantidad promedio de los Adjetivos, Sustantivos, Alfanuméricos, Símbolos y Palabra es mayor en los tweets reales.

- Kill es el verbo más utilizado en los tweets reales y like en los falsos.

- Dead es el adjetivo más utilizado en los tweets reales y Good en los falsos.

- En relación a los sustantivos, en los tweets reales no se destacó ninguno mientras que en los falsos fue Fire.

- Con los alfanuméricos no hubo nada que se destacara, al igual que con los símbolos.

- New York es la locación con mayor cantidad de sustantivos, London de verbos, Mumbai de alfanuméricos, Los Ángeles de adjetivos y USA de símbolos.

- Mumbai es el que tiene la mayor cantidad de palabras promedio.

- En relación a los caracteres especiales, 2000 tienen el símbolo '#'.

- Salvo el caracter '#', los demás tienen el porcentaje de tweets falsos mayor al de los reales.

- El arroba es el que más se relaciona con los demás caracteres.

- #News es el hashtag más utilizado.

- #Hiroshima y #Earthquake tienen un porcentaje de tweets falsos del 100% y #News del 75%, los demás el porcentaje de tweets falsos es mayor al de los reales.

- Los tweets que tienen signos de exclamación no tienden a ser noticias serias.

- Con los tweets que tienen de una a tres menciones, no parece haber un patrón claro de comportamiento pero, mientras más menciones tienen, más baja su porcentaje de veracidad.

- Mientras más caracteres en mayúsculas tienen los tweets, su porcentaje de veracidad disminuye.

- @Youtube es la cuenta más mencionada en todos los tweets.
- En la gran mayoría de las cuentas mencionadas se encuentran en tweets falsos.
- En @Potus o @Foxnews, el porcentaje de tweets verdaderos es mucho mayor al porcentaje de los falsos.
- Si Keyword es nulo, el porcentaje de tweets verdaderos es mayor al de los falsos.
- Sí Location es nulo, el porcentaje de tweets falsos es mayor al de los verdaderos.
- Outbrake es el keyword que más aparece en los tweets reales y Armageddon en los falsos.
- El Keyword que más aparece en los tweets con Locación nula es el de 'razed', seguido muy de cerca por ‘detonation’ y ‘explosion’.
- Salvo con evacuate y fatalities, en los demás keyword predomina el porcentaje de tweets falsos por sobre los verdaderos.