

Trabajo Practico Número 1: Grupo EconoDatos

Martín Yanquilevich, Federico Paci, Leonardo Esteban Delgado, Martín Gerschenfeld

May 21, 2020

Contents

1	Introducción	2
2	Análisis Exploratorio de Datos	2
3	Palabras más Utilizadas	3
4	Análisis por Ubicación	5
5	Análisis por Longitud del Tweet	7
6	Análisis de los Keywords	9
7	Más Mencionados Según Veracidad	12
8	Análisis Heatmap	14
9	Conclusión	14

1 Introducción

Link del Repositorio: <https://github.com/Martingerschenfeld/Tp-Datos/>

Link al Collab: <https://colab.research.google.com/drive/1uZfD0SyaQiBpMIWrtTNCIgQrNMA1D4HW>

El trabajo práctico 1 de la materia se basa en el análisis de los tweets del set de datos de Kaggle. El objetivo es hacer un análisis exploratorio sobre los datos que surgen de twitter, relacionado con eventos del tipo desastres naturales.

El set de datos contiene 10,000 tweets seleccionados a mano por la pagina Kaggle. Los tweets se encuentran relacionados con palabras claves de desastres naturales. En muchos casos, usuarios utilizaron palabras metafóricas para referirse a otras situaciones, pero ajeno a desastres naturales. Tal como podemos ver en el siguiente ejemplo, donde la usuaria "Anna K" se refiere al cielo naranja como si se estuviera incendiando. Este tipo de tweet puede ser captado por el data-set. El objetivo del dato es extraer estadísticas y llegar a conclusiones sobre la posibilidad de determinar si Twitter es un lugar confiable para informarse acerca de catástrofes naturales.



Figure 1: Tweet extraido de Kaggle.com

2 Análisis Exploratorio de Datos

Arrancamos el ejercicio cargando los datos como un Pandas Dataframe desde el csv de la pagina de Kaggle.

El dataframe consiste en 5 columnas:

- id - identificador único para cada tweet
- text - el texto del tweet

- location - ubicación desde donde fue enviado (podría no estar)
- keyword - un keyword para el tweet (podría no estar)
- target - en train.csv, indica si se trata de un desastre real (1) o no (0)

Haciendo un info sobre el DataFrame podemos observar que contiene 5 columnas y 7613 filas no nulas. Las columnas id es un numero (integer con longitud máxima de 64 caracteres) , así como el target (que además es binario, vale o bien 0 o 1), por lo que tiene sentido. Las otras tres columnas, corresponden a textos, en Python esto se representa como un string, que es un tipo "object" (que es el formato en Python de todo lo que no es un integer, boolean, ect.). Vemos que las Pandas logró identificar correctamente que tipo de datos son en cada columna.

De las filas, algunas columnas tienen componentes nulos. El id (que funciona como key) y el texto del tweet son los únicos completos. En algunos tweets, no está establecido la ubicación, en otros casos no fue identificado la keyword que mejor identifica el contenido del tweet, ni tampoco si el tweet es veraz o no (target puede ser NaN).

```
# Information about the dataset
df_twitter.info()

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

Figure 2: Info de los Tweets extraído de Kaggle.com

3 Palabras más Utilizadas

Resulta interesante ver cuales son las palabras más utilizadas para identificar la utilidad de twitter en situaciones de catástrofes naturales.

El primer paso consistió en transformar los textos de los tweets en una lista (de la misma dimensión que id). Luego buscamos separar palabra por palabra para obtener una lista con el mismo tamaño a la cantidad de palabras de todos los tweets (113 461). Como paso intermedio a contar palabras, primero convertimos todas las palabras a minúscula. Python es case-sensitive a la hora de realizar operaciones, por lo que dos palabras iguales, donde la primera letra es mayúscula es considerada como una palabra distinta. Finalmente, realizamos un for que toma un diccionario vacío, luego se fija cada palabra en la lista, en caso de no encontrarla en el diccionario, entonces la agrega con un conteo de 1, en caso de que ya exista un registro, entonces le suma el conteo por 1. Finalmente, se ordena el diccionario de mayor value a menor.

Observando el resultado de los 15 resultados más frecuentes vemos que no resulta informativo. Las palabras más utilizadas son artículos, pronombres, ect. Resulta más útil si estas palabras fueran sustantivos o tal vez adjetivos, por lo que debemos volver a filtrar.

```
[('the', 3207),  
 ('a', 2135),  
 ('in', 1949),  
 ('to', 1934),  
 ('of', 1814),  
 ('and', 1405),  
 ('i', 1336),  
 ('is', 930),  
 ('for', 880),  
 ('on', 834),  
 ('-', 763),  
 ('you', 746),  
 ('my', 671),  
 ('with', 562),  
 ('that', 538)]
```

Figure 3: Las 15 palabras más frecuentes en los tweets del data set

Para obtener mejor información sobre las palabras más frecuentes, observamos que las 15 palabras más frecuentes tenían 4 letras o menos. Por ende el filtro que aplicamos fue tomar las palabras que tenían por lo menos 5 letras. Este filtro no es el más preciso y existe el riesgo de que vayamos a eliminar palabras de 3 o 4 letras relevantes, sin embargo consideramos que es efectivo para poder tener información rápida.

En el siguiente histograma presentamos las 15 palabras más frecuentes de por lo menos 5 letras y su frecuencia.

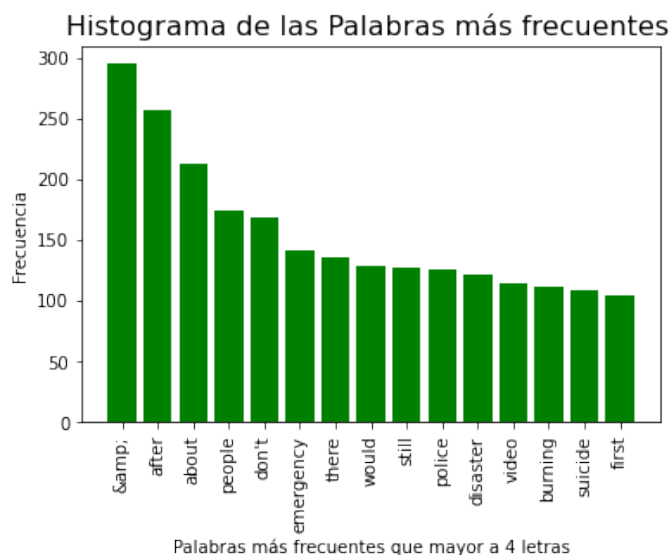


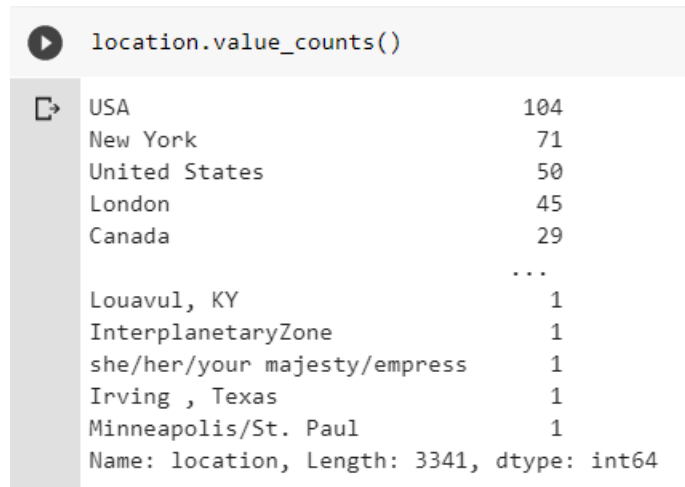
Figure 4: Las 15 palabras de más de 4 letras más frecuentes en los tweets del data set

Si bien no podemos identificar si los tweets son relacionados con catástrofes naturales, y habría que verlas en su contexto, las palabras que figuran en el histograma son, por lo menos pertinentes. Para ejemplificar, observamos las palabras desastre, emergencia y quemando, que son descripciones del hecho. Encontramos las palabras personas y policía, que son sujetos intervinientes, y encontramos palabras como suicidio, y después que hablan de las consecuencias de la catástrofe. Como hecho negativo del filtro, encontramos en el listado también la palabra &, que luego de buscar el significa en internet, resulta que es parte de un código de html, por ende fue un error en el cargado de los tweets al csv.

4 Análisis por Ubicación

La columna 'location' del DataFrame contiene información sobre el lugar donde se envió el tweet. Resulta interesante analizar, puesto que si tuviéramos información acerca de catástrofes naturales que ocurrieron, entonces se podría comparar, y ver con mayor detalle la importancia de twitter. La información de ubicación es auto-reportada y no está sujeto a verificación.

Un análisis preliminar muestra la siguiente información.



```
location.value_counts()
```

USA	104
New York	71
United States	50
London	45
Canada	29
...	
Louavul, KY	1
InterplanetaryZone	1
she/her/your majesty/empress	1
Irving, Texas	1
Minneapolis/St. Paul	1

Name: location, Length: 3341, dtype: int64

Figure 5: Ubicación auto reportada del data set

Observamos que Twitter no agrupa dicha información en países, sino que copia textualmente el valor. Ocurre que hay ubicaciones que se encuentran abreviadas (Por ejemplo USA, representa el mismo país que United States of America), hay personas que han puesto su estado o ciudad (New York o California son dos estados dentro de Estados Unidos), hay personas que han escrito mal su ubicación, hay personas que escribieron el nombre su país en su idioma natal y cuesta comparar con el nombre en inglés (Brasil en vez de Brazil), espacios en el nombre de los países que lo hace imposible distinguir para Twitter y una gran cantidad de personas pusieron de ubicación lugares inexistentes o frases sin sentido (tal como she/her/your majesty/empress).

Para conseguir mejores resultados en este apartado, hicimos reemplazos para homogeneizar los nombres al idioma inglés. Por un tema de tiempo y eficiencia, fue hecha para todos los valores cuya frecuencia fue mayor a 3. No hemos encontrado una regla sistemática, y si bien se puede seguir haciendo cambios, consideramos que no modifica sustancialmente el resultado.

Para hacer dicho cambio, armamos una diccionario cuyo key es el string del texto que queremos cambiar y el value es el país al cual pertenece. En caso de que no fuera identificable, se le aplicó el valor nulo de NaN. Luego mapeamos el diccionario a location, completando lo no identificado con NaN. Observamos a continuación los resultados.

```
print(location.value_counts())
```

NaN	6553
United States	747
United Kingdom	164
Canada	41
India	30
Australia	24
Nigeria	14
Japan	10
Kenya	7
United Statess	6
Brazil	5
Switzerland	4
China	4
South Africa	4

Name: location, dtype: int64

Figure 6: Ubicación del auto reportada que realizó el tweet - Corregido

Observamos cambios notorios con respecto a la tabla anterior. Mientras que en antes los Estados Unidos figuraba como las primeras 3 entradas, ahora lo tenemos condensado en una sola entrada y por ende es más representativo. En este punto, decidimos simplemente observar los gráficos, y considerar el resultado cualitativo debido a la poca confiabilidad de los datos NaN. Para concluir de la tabla 2, podemos decir Estados Unidos es el país donde la gran mayoría de los tweets provienen de, luego en una menor cuantía sigue el Reino Unido, luego en una menor medida varios países, que no podemos garantizar significativamente cual es más frecuente.

5 Análisis por Longitud del Tweet

Para el siguiente apartado, analizamos la longitud de los tweets del data set, recordamos que hasta el 2017, los límites de caracteres que se puede tener en un tweet es de 140, luego de dicha fecha el límite fue aumentado a 280. Desconocemos el año del data set, sin embargo es de presumir que es pre 2017.

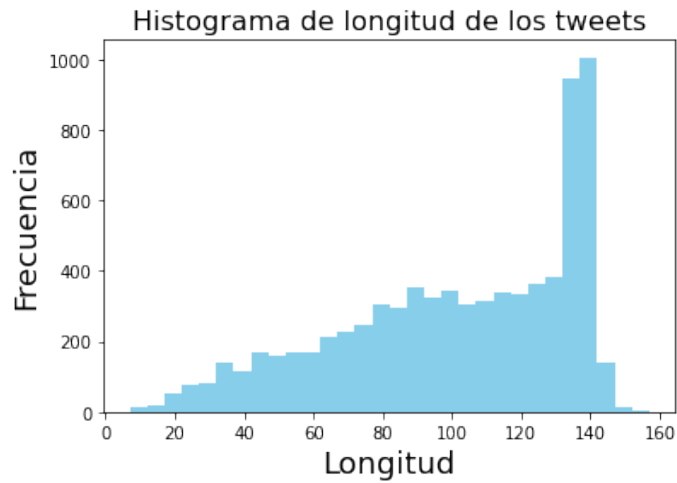


Figure 7: Histograma de longitud de los Tweets

Observamos que la moda son valores cercanos al limite de 140 caracteres. Este resultado tiene sentido, puesto que es esperable que en situaciones extremas, las personas se sientan inclinados por expresar emociones, deseos de recuperación ajena o criticas a gobiernos por no estar suficientemente preparado. Otro punto a tener en cuenta es que existen tweets con longitud mayor a 140, lo más probable es que sean anomalías, o errores en la carga al csv. De todas formas, recomendamos mirar más de cerca los datos para evaluar que hacer con estos datos.

Para continuar nuestro análisis de longitud, podemos incorporar el elemento de la veracidad. La variable 'target' es una variable binaria donde si vale 1 entonces el tweet se encuentra relacionado con un hecho real, y si el valor es 0, o bien no se encuentra relacionado, o no se sabe (nuestro supuesto inicial). Repetimos nuestro análisis de longitud de tweets, tomando en cuenta el valor de veracidad.

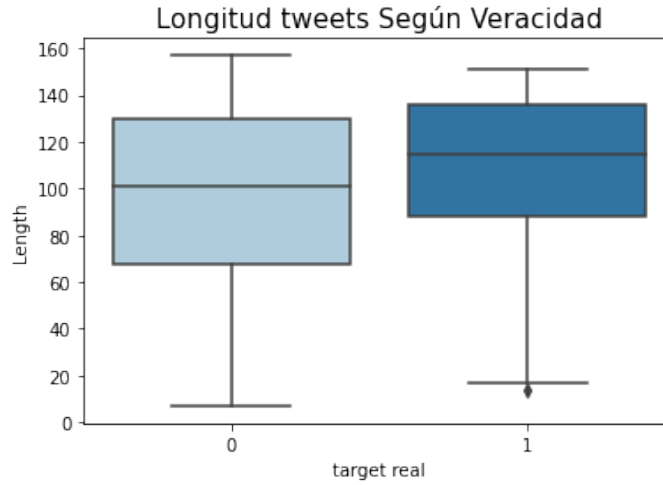


Figure 8: Histograma de longitud de los Tweets en función de la veracidad

Utilizamos un gráfico de caja y bigote. En celeste, son los tweets con veracidad 0, se observa un rango amplio de 10 a 150, con media en 100 y cuartiles en 70 y 130. En la caja azul, se observa un rango entre 20 y 140, con media en 110 y cuartiles en 90 y 135. En primer lugar, pareciera que en el caso de la veracidad igual a 1, los tweets son significativamente y consistentemente más alto. Esto se puede ver por el rango más chico, la media y los cuartiles 1 y 3 más altos. No se puede garantizar si esto es así siempre, pero gráficamente se puede apreciar una intuición: Si el tweet es más largo, entonces es más probable que sea relacionado a un evento real que no.

6 Análisis de los Keywords

Una de las características de los tweets del dataset es el keyword, que es una variable categórica, tweets con el mismo keyword se agrupan entre sí. De los 7613 tweets originales, 7552 tiene categorías, de las cuales se pueden agrupar en 221 categorías distintas. Si bien, muchos de los tweets no están relacionados con catastrofes naturales, decidimos trabajar con un subset de datos para poder filtrar mejor y llegar a conclusiones más significativas.

El primer criterio de filtración fue quedarnos con las keywords que aparecen más veces que el la aparición promedio de keywords. Eso redujo las 221 categorías a 114. Luego elegimos quedarnos con tengan por lo menos 39 frecuencias. A continuación observamos el histograma de las frecuencias.

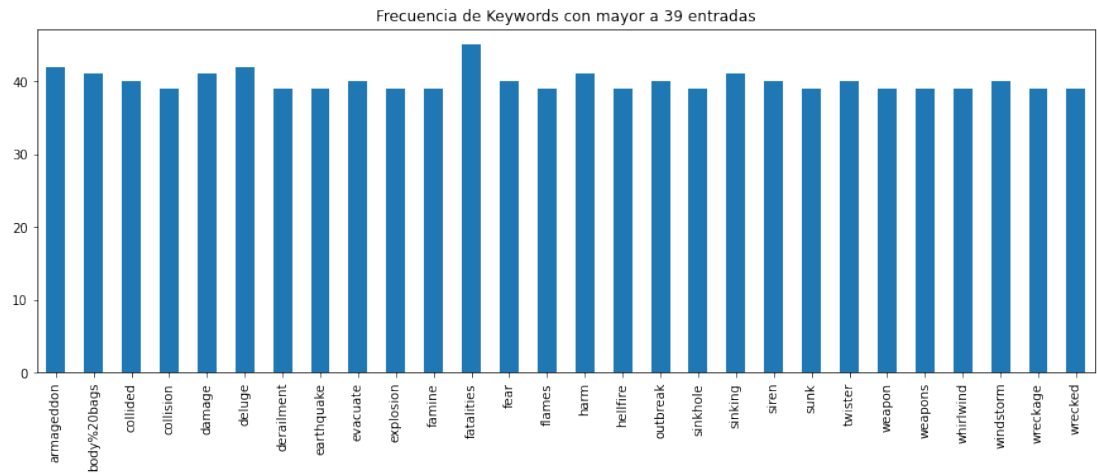


Figure 9: Frecuencia de 38 keywords más frecuentes

Si observamos las keywords de las más frecuentes, observamos que son relevantes con catástrofes naturales, por ejemplo: armageddon, body bags, collision, damage, earthquake, ect.

Para seguir con nuestro análisis, analizamos como son las keywords en función de la veracidad. Para ello, clasificamos la cantidad de veces que aparece cada keyword en el caso de que el tweet tenga $\text{target} == 0$ y $\text{target} == 1$, para luego encontrar el porcentaje veracidad que tiene cada keyword.

	fake	truth	total	%
armageddon	37.0	5	42.0	0.119048
body%20bags	40.0	1	41.0	0.024390
collided	17.0	23	40.0	0.575000
collision	10.0	29	39.0	0.743590
damage	22.0	19	41.0	0.463415
deluge	36.0	6	42.0	0.142857
earthquake	9.0	30	39.0	0.769231
evacuate	15.0	25	40.0	0.625000
explosion	19.0	20	39.0	0.512821
famine	13.0	26	39.0	0.666667
fatalities	19.0	26	45.0	0.577778
fear	35.0	5	40.0	0.125000
flames	26.0	13	39.0	0.333333
harm	37.0	4	41.0	0.097561
hellfire	32.0	7	39.0	0.179487
outbreak	1.0	39	40.0	0.975000
sinkhole	12.0	27	39.0	0.692308
sinking	33.0	8	41.0	0.195122
siren	35.0	5	40.0	0.125000
sunk	30.0	9	39.0	0.230769
twister	35.0	5	40.0	0.125000
weapon	25.0	14	39.0	0.358974
weapons	22.0	17	39.0	0.435897
whirlwind	25.0	14	39.0	0.358974
windstorm	24.0	16	40.0	0.400000
wrecked	36.0	3	39.0	0.076923
derailment	NaN	39	NaN	NaN
wreckage	NaN	39	NaN	NaN

Figure 10: Frecuencia y % de veracidad por keywords

Yendo en mayor detalle de la veracidad, podemos ver que categorías que eran muy frecuentes en el gráfico anterior, pueden no tener que ver con desastres naturales. Para ejemplificar, el caso de body bags, tiene 41 instancias, de las 40 son falsas y 1 es verdadera. Mientras que categorías como brote (outbreak) tiene 40 instancias, de las cuales 39 fueron relacionados con un hecho real.

Gráficamente podemos observar lo expresado en la tabla.

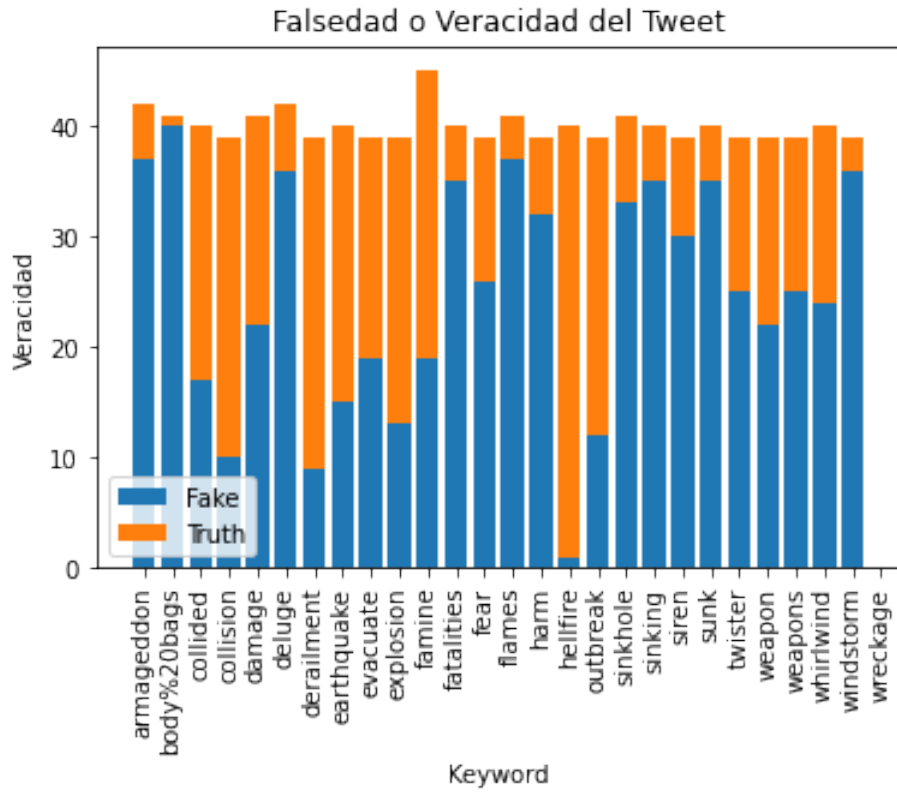


Figure 11: Frecuencia y % de veracidad por keywords

7 Más Mencionados Según Veracidad

Otro análisis que resulta interesante es identificar a que cuentas de twitter arroban al momento de un desastre natural. Para ello extraimos usando el API de Tweepy todas las instancias que se arrobaron cuentas, y luego los graficamos. En los siguientes dos gráficos podemos observar los resultados.

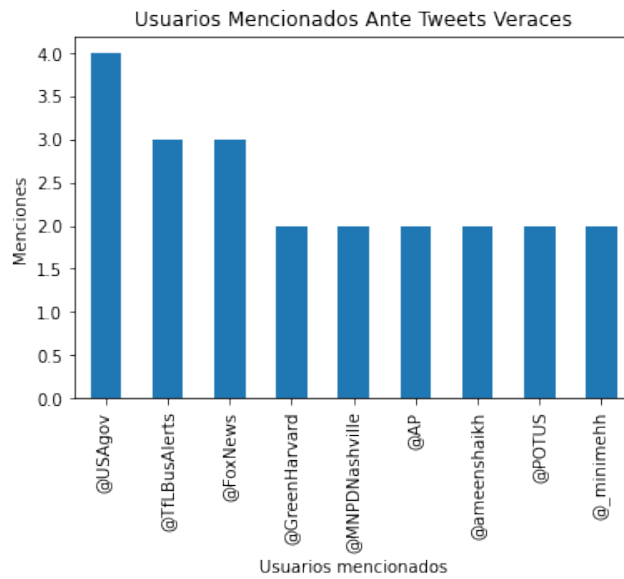


Figure 12: Usuarios Mencionados Ante Tweet Veraces

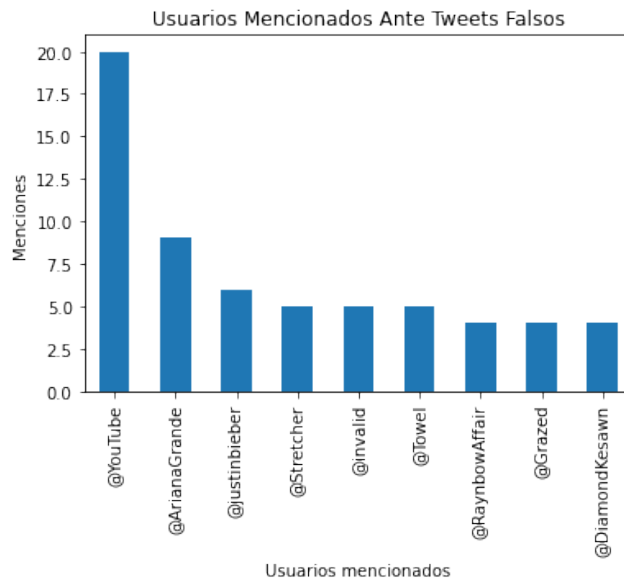


Figure 13: Usuarios Mencionados Ante Tweet Falsos

Podemos observar que durante una catastrofe natural, los usuarios con más menciones incluyen: distintas partes del gobierno, entre ellos el Presidente de los

Estados Unidos, distintos medios periodísticos tal como Fox News y la Prensa Asociada (AP), fuerzas de seguridad, probablemente en lugares donde hayan ocurrido estos hechos ect. Todos estos usuarios son consistentes con el ejercicio.

A su vez, los tweets que no tienen que ver con desastres naturales, etiquetan a distintas personas que no tienen ningún impacto en solucionar el problema, usuarios como la página de Youtube, distintos artistas famosos tal como Justin Bieber y Ariana Grande. Estos tweets no nos permiten sacar conclusiones relevantes sobre los hechos que nos interesa.

8 Análisis Heatmap

Se utilizó un heatmap para ver la intensidad de la correlación entre la variable de longitud del tweet contra el target. Del mismo se obtiene un valor cercano a 0,1. Lo que nos conllevaría a no incurrir en que la cantidad de caracteres utilizados permita ser un indicador sobre la veracidad del mismo.

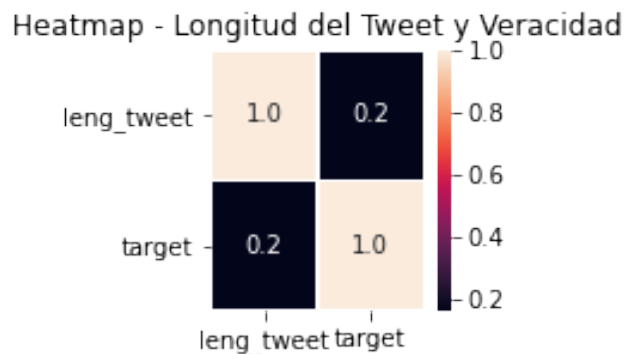


Figure 14: Heatmap entre Longitud del Tweet y Veracidad

9 Conclusión

Luego de realizar análisis exploratorio de los datos, y realizar distintas visualizaciones pudimos encontrar que existen relaciones entre distintas variables. Entre ellas destacamos:

- La mención de cuentas oficiales gubernamentales, o de medios aumentan la probabilidad de que el tweet este relacionado con catástrofes naturales.
- Existen keywords cuyo % de veracidad es muy alto, por lo que habría que hacer un análisis detallado para poder identificarlo.
- La relación entre la longitud del tweet y la veracidad queda demostrado a nivel gráfico sin embargo cuando analizamos la matriz de correlación no

podimos ratificarlo. En este ejercicio ponemos más peso que sacamos de la conclusión visual que el método obtenido por el método estadístico.