Trabajo Práctico I 'Real or Not? NLP with Disaster Tweets' Análisis Exploratorio

75.06/95.58 Organización de Datos
Primer cuatrimestre de 2020

Equipo NaN:
Ignacio Ibarra
Alejandra Toscano

Repositorio:
https://github.com/Ignacio-Ibarra/NLP-Disasters

Índice

1. Introducción

Este es un trabajo práctico realizado para la materia 7506 - Organización de Datos de la Facultad de Ingeniería de la Universidad de Buenos Aires. Se trata de una competencia lanzada en la plataforma Kaggle, la cual consiste en predecir si un tweet está hablando de una desastre o no, implementando técnicas de Natural Language Processing (NLP).
El objetivo del Trabajo Práctico I es realizar un análisis exploratorio de los datos mediante procesamientos y visualizaciones de las distintas relaciones entre las variables proporcionadas. Se cuenta con un dataset de entrenamiento (`train.csv`). Las variables presentes en el mismo son:

- `id`: una identificación única de cada una de las observaciones.
- `keyword`: palabra clave para clasificar al tweet
- `location`: ubicación del usuario. Dicho dato es arbitrario ya que el usuario define este input, no necesariamente es la ubicación real
- `text`: texto del tweet
- `target`: variable binaria que toma valor 1 cuando el tweet corresponde con un desastre y 0 cuando no.

2. Primer análisis

Se comienza analizando la estructura del DataFrame: cuenta con 7613 filas y 5 columnas correspondientes a las variables mencionadas en el punto anterior. Tanto `id` como `target` son del tipo `int`, mientras que el resto son del tipo `object`. El uso de memoria es bajo: 2297kb.
Por otro lado, `keyword` y `location` presentan elementos nulos. A estos objetos se les asignarán los valores 'no_keyword' y 'no_location' respectivamente.
Se analiza a continuación las cantidades de tweets con Target 0 y 1:

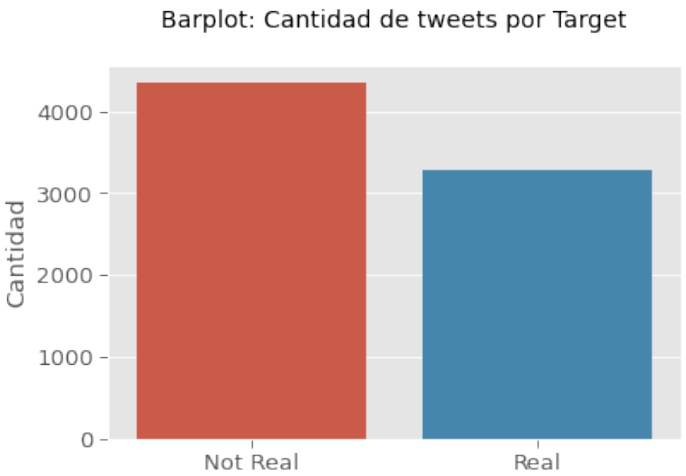

Figura 1: Cantidades de tweets por Target

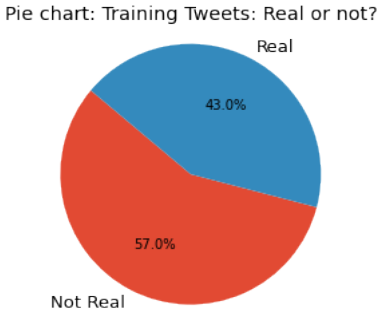

Figura 2: Proporciones de tweets con Target 0 y 1 respectivamente

Se observa mayor proporción de Target 0 sobre Target 1.
Se busca ver si la variable `location` presenta información útil para el análisis. Se muestra a continuación las 20 ubicaciones más usadas, incluyendo 'no_location'.

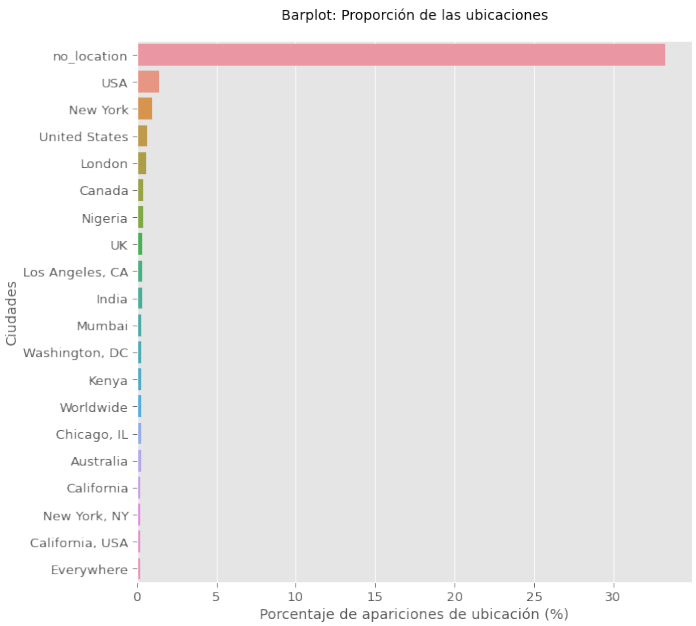

Figura 3: Proporciones de tweets con Target 0 y 1 respectivamente

Se observa que:

- El 33 % de las muestras no tienen ubicación
- El 67 % restante se reparte entre una variedad grande de ubicaciones (3342), con bajos porcentajes de repetición.
- Las ubicaciones encontradas son de países con idioma inglés, siendo 'USA' el más popular. Se observa en el barplot varias ciudades de este país.

De acuerdo a lo observado, la variable location no representa valor relacionándola con el resto de las variables. Esto concuerda con el hecho de que la ubicación no es un input automático de una geolocalización, si no que es elegido por el usuario sin ser necesario un formato específico: puede elegir el país, ciudad, o cualquier otra cosa (ejemplo: 'Everywhere').

3. Cleaning

Previo a continuar con el análisis exploratorio, es necesario realizar una limpieza del DataFrame.

3.1. Registros duplicados - Primera parte

En primer lugar se comprueba que dado el id no se repite en ninguna fila.
En segundo lugar, ignorando la columna 'id', se analiza si los tweets de la columna 'text' tienen más de una ocurrencia. También se controla de que en caso de estar repetido, no esté clasificado con target 1 y target 0 simultáneamente. Se obtiene la siguiente tabla:

	text	target	
		count	nunique
646	11-Year-Old Boy Charged With Manslaughter of T...	10	1
45	#Bestnaijamade: 16yr old PKK suicide bomber wh...	6	1
6131	The Prophet (peace be upon him) said 'Save you...	6	2
3589	He came to a land which was engulfed in tribal...	6	2
4589	Madhya Pradesh Train Derailment: Village Youth...	5	1
...	...	...	...
2507	Bamenda Floods Kill Animals Birds - http://t.c...	1	1
2506	Baltimore City : I-95 NORTH AT MP 54.8 (FORT M...	1	1
2505	Bairstow dropped his buffet ticket there. Deva...	1	1
2504	Bad day	1	1
7502	âÈMGN-AFRICAâ¨ pin:263789F4 âÈ Correction: Ten...	1	1

7503 rows × 3 columns

Figura 4: Tabla de tweets repetidos

Por un lado, se observa que existen tweets que se repiten hasta 10 veces en el dataset. Por otro lado, efectivamente hay tweets repetidos que fueron clasificados de dos formas distintas, lo cual representa una contradicción. Estos elementos son eliminados del DataFrame.
Los tweets repetidos sin contradicción de targets, serán eliminados salvo su

primera ocurrencia (el número de repeticiones es guardado en el DataFrame).

3.2. Limpieza de texto

El siguiente paso consistió en identificar y eliminar urls, menciones y emojis. Previo al filtro, urls y menciones fueron contadas y guardadas dentro del DataFrame. Los hashtags no fueron eliminados ya que suelen formar parte del texto. También se modificaron las contracciones (Ejemplo: can't –>can not) y se suprimieron los signos de puntuación.

3.3. Registros duplicados - Segunda parte

Se realizó el mismo análisis de registros duplicados con el texto limpio.

	text_clean	target	
		count	nunique
6271	Watch This Airport Get Swallowed Up By A Sands...	24	1
6497	Wreckage Conclusively Confirmed as From MH37...	20	1
3181	Families to sue over Legionnaires More than 4...	19	1
1607	hot Funtenna hijacking computers to send da...	17	2
2758	Christian Attacked by Muslims at the Temple Mo...	15	1
...	...	...	...
2333	Ancient Mayan Tablet Found in Jungle Temple	1	1
2332	An outbreak of Legionnaires disease in New Yo...	1	1
2331	An outbreak of Legionnaires disease in New Yo...	1	1
2330	An optical illusion clouds rolling in over t...	1	1
6939	åÈMGN AFRICAå¨ pin 263789F4 åÈ Correction Ten...	1	1

6940 rows × 3 columns

Figura 5: Tabla de tweets repetidos - Paso 2

Se observa que existen tweets que tenían el mismo texto pero distintas urls. Se vuelve a encontrar tweets repetidos con contradicción de targets. Estas ocurrencias son eliminadas. Los tweets repetidos con mismo target son eliminados salvo su primera ocurrencia (el número de repeticiones es guardado en el DataFrame).

3.4. Eliminación de stopwords

Al texto limpio, se le puede aplicar un nuevo filtro que consta en eliminar palabras genéricas que en términos de NLP se consideran ruido": stopwords Esta nueva limpieza del texto es guardada en una columna nueva del DataFrame, de modo que en la estructura se tienen el texto original, el texto post primera limpieza y el texto post limpieza sin stopwords.

3.5. Resultado del Cleaning

De esta forma, nuestra cantidad de muestras se reduce: de las 7613 filas originales, ahora se cuenta con 6889. Se analiza el impacto que esto tiene en las proporciones de tweets por target:

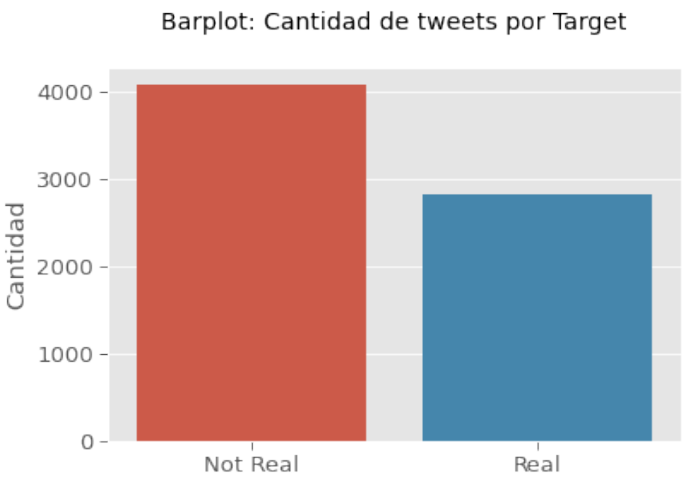

Figura 6: Cantidades de tweets por Target (Post Cleaning)

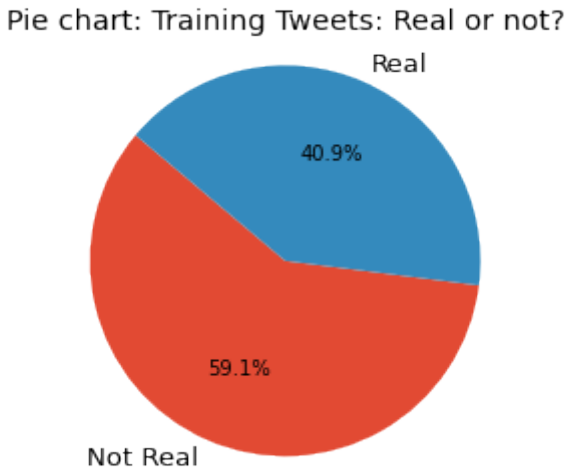

Figura 7: Proporciones de tweets con Target 0 y 1 respectivamente (Post Cleaning)

Sobre un total de 6,889 registros 4,072 pertenecen a Target Not Real (target 0) y 2,817 a Target Real (target 1), lo que equivale a 59.1 % y 40.9 % respectivamente. Se observa que las proporciones se mantienen pese al cleaning.

4. Análisis de Keywords

Se plantea la siguiente pregunta: ¿existen keywords que se usen más para tweets relacionados con desastres que para tweets 'no desastrosos'? En caso de ser así ¿en qué se diferencian estos keywords con los demás? Se presenta el siguiente gráfico:

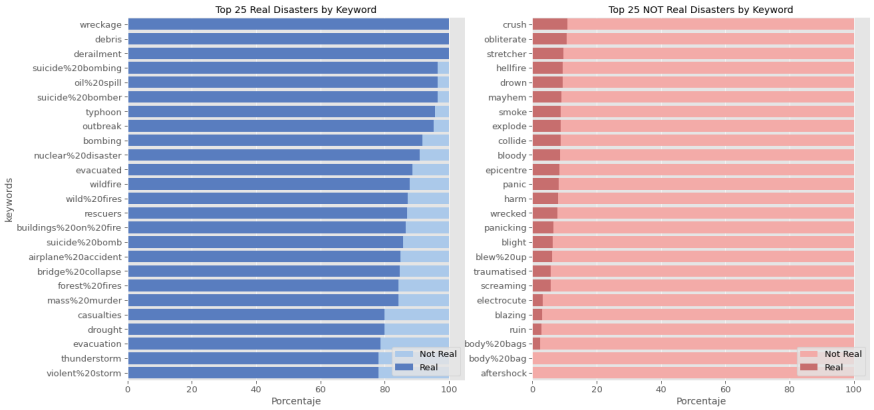

Figura 8: Proporciones de target 0 y target 1 por Keyword)

En la figura de la izquiera se ilustran 25 keywords con porcentajes de Target Real"mayores al 75 %. Se observan keywords referentes a desastres naturales (drought, forest fire, violent storm) o edilicios (bridge collapse, buildings on fire). También se encuentran tragedias mortales (mass murder, airplane accident, nuclear disaster, suicide bombing). Otros keywords incluyen lenguaje de noticias (casualties, evacuated, rescuers).
Por otro lado, en la figura de la derecha, se observan los 25 keywords con menores porcentajes de veracidad. En su mayoría se trata de palabras que servirían para describir tragedias pero que a su vez pueden usarse en muchos otros contextos. Un ejemplo de esto es el keyword "body bag", con un porcentaje de targets Reales menor al 5 %. El término por si solo sonaría alarmante, pero veamos las acepciones de la palabra:

1. Bolsa para guardar cadáveres
2. Una clase de cartera/mochila

3. Según Urban Dictionary, "To boddybag an oponent.[es] asegurar la victoria sobre el oponente (algo así como "kick ass"), slang muy utilizado en las batallas de rap.

4. Otra acepción de Urban Dictionary, "Bodybagging.[es] robarle un chiste a otro comediante y hacerlo pasar como propio.

Se observa también mayor cantidad de verbos que en target 1, donde predominan sustantivos.

5. Análisis del texto

Quien realice un análisis exploratorio a partir de una "tabula raza"para econtrar *insights* acerca del objeto estudiado conduce su análisis sin una brújula. El análisis exploratorio es más bien como el trabajo de un detective, parte desde un saber, un pre-concepto y aborda la realidad desde los anteojos de su teoría. La exploración de los datos en el presente trabajo es guiada por la intuición de que los tweets que son reales se caracterizan, en cierto modo, por tener textos con un estilo informativo. En cambio los tweets que no son reales tienen textos que carecen de dicho estilo. Siempre habrá oportunidades para grises y contraejemplos, pero si es posible encontrar tendencias generales entonces estaremos en un buen comienzo. Para ello esta sección tiene tres tipos de análisis:

1. Características generales que hacen al formato del texo, es decir, la cantidad de caracteres, la cantidad de palabras, la cantidad promedio de caracteres por palabra por tweet, cantidad de URLs, hashtags, mentions y si posee números.

2. Cuáles son las clases de palabras (sustantivos, verbos, adjetivos, etc.) que se encuentran con mayor frecuencia, qué relaciones encontramos entre distintas clases de palabras.

3. Revisión de los *stopwords* más comunes.

4. Análisis sobre qué palabras o combinaciones de palabras son más frecuentes.

5.1. Formato del Texto

5.1.1. Cantidad de caracteres

Cantidad de caracteres pre-cleaning
Hay una distribución en ambos casos con una asimetría negativa, pero la del target 1 es más marcada. El coeficiente de asimetría de la distribución del largo del tweet para target 1 es -0.74 y la de target 0 es -0.39. Esto nos marca que hay una mayor prevalencia por los tweets largos en el caso del target 1.

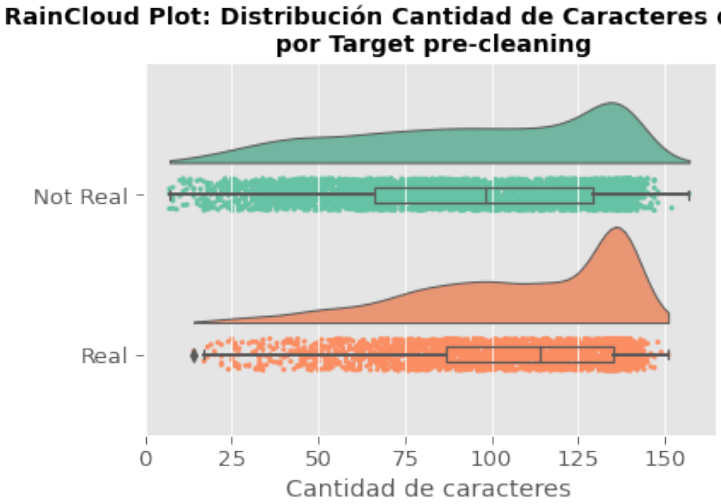

Figura 9: Distribuciones por target de la cantidad de caracteres pre-cleaning

Como se observa en el gráfico la punta de la distribución para target 1 es más empinada y por lo tanto acumula más registros alrededor de los valores altos. Por otro lado la media también es mayor para target 1 que para target 0.La dispersión lograda con el método *jitter* muestra que para el target 0 hay una concentración a lo largo de todo el rango de la variable con una leve dispersión en los valores más bajos. Mientras que para el target 1 la dispersión en los valores bajos es mayor todavía y la mayor concentración de puntos es notoria para valores medios y altos.

Cantidad de caracteres post-cleaning.
Si consideramos los textos luego de sacarles hashtags, mentions, puntuaciones, dígitos y stopwords, las distribuciones cambian radicalmente.

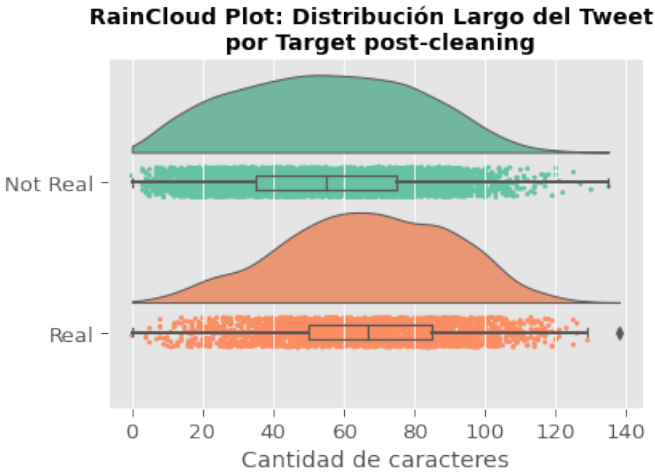

Figura 10: Distribuciones por target de la cantidad de caracteres post-cleaning

El coeficiente de asimetría de target 1 pasa a -0.14, con lo cual sigue conservando su asimetría hacia la derecha. Sin embargo, para target 0 se vuelve positivo y vira su asimetría levemente a la izquierda. Luego del cleaning se nota aún más el sesgo de los tweets que hablan de desastres reales a ser más largos que los tweets que hablan de desastres no reales.

5.1.2. Cantidad de palabras

En este apartado se realiza el mismo análisis que en el anterior pero sobre la cantidad de palabras de cada tweet.

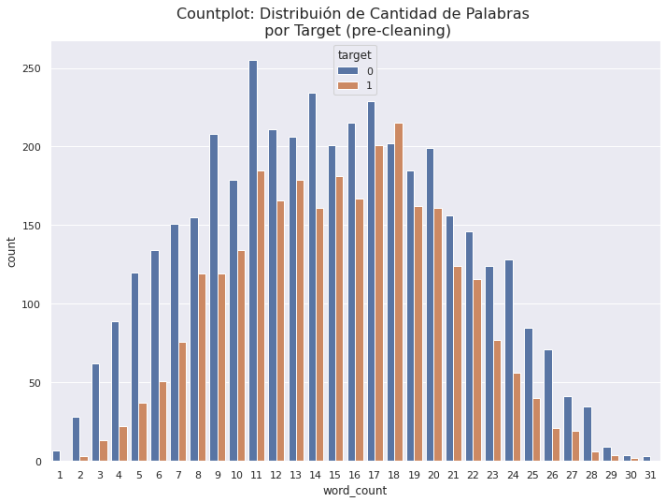

Figura 11: Distribuciones por target de la cantidad de palabras pre-cleaning

La cantidad de caracteres promedio para target 1 ronda los 15, mientras que para target 0 ronda los 14.

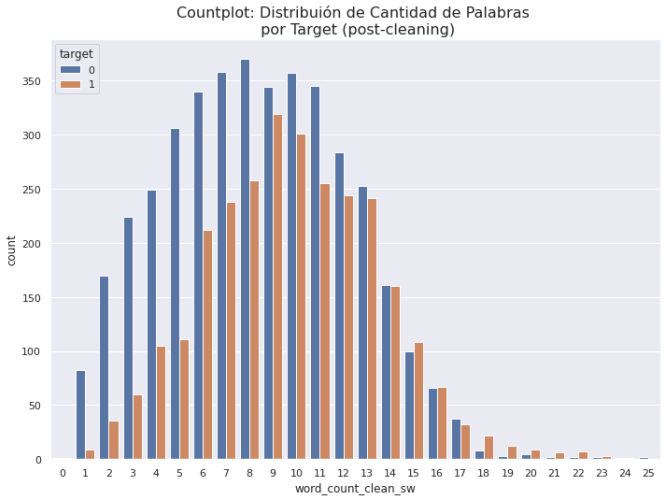

Figura 12: Distribuciones por target de la cantidad de palabras post-cleaning

En cambio, cuando limpiamos cada uno de los textos, las distribuciones cambian bastante. Ambas se vuelven más “normales”. Sobre todo la distribución de target 1 ya que sus colas no son tan pesadas.

Por otro lado, la diferencia entre las medias es mayor, pasa de 1 en el caso before cleaning a 2 en after cleaning

5.1.3. Cantidad de caracteres promedio por palabra

En el siguiente gráfico, las distribuciones de la cantidad de caracteres promedio por palabra para target 0 y target 1, muestran un sesgo mayor hacia la izquierda por parte de target 0.

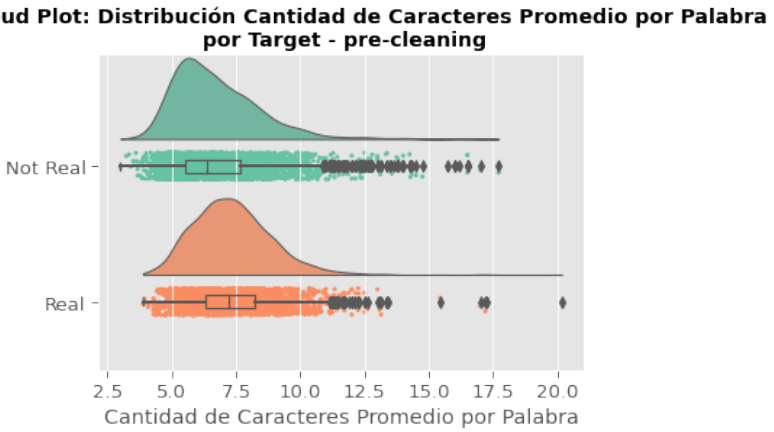

Figura 13: Distribución de Cantidad de Caracteres Promedio por Palabra (pre-cleaning)

La asimetría marcadamente hacia la izquierda de target 0 indica que prevalecen palabras çortas". Mientras que la asimetría ligeramente hacia la izquierda indica que prevalecen palabras çortas o intermedias". Por otro lado, las medias se aproximan a 6 y 7.5 para target 0 y target 1 respectivamente.
En el siguiente gráfico observamos la distribución de la misma variable pero luego de hacer el cleaning (que incluyó eliminar *stopwords*).

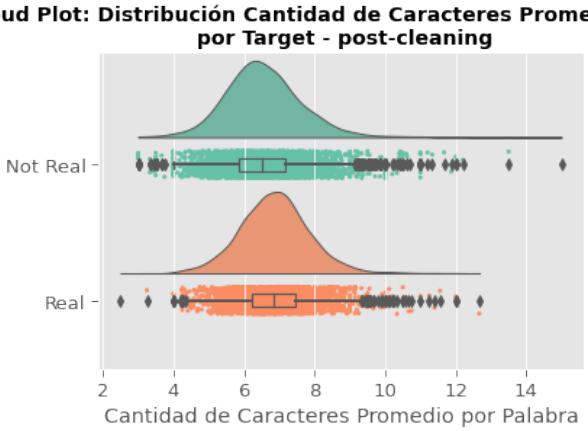

Figura 14: Distribución de Cantidad de Caracteres Promedio por Palabra (post-cleaning)

Las *stopwords* son palabras cortas que carecen de significado por si mismas. Buena parte del uso de los stopwords en el *natural language* son los pronombres personales o posesivos: i, you, he, she, we, them entre los primeros, my, your, their, our, entre los segundos. Hay muchas palabras más dentro de la lista de *stopwords* y esta varía según la librería de NLP que se esté utilizando. En el segundo gráfico se observan distribuciones más similares y mayor simetría. Esto indica que en target 0 es probable que haya una fuerte presencia de *stopwords* y que al quitarlos la cantidad promedio de caracteres de las palabras tiendan a igualarse en todo su dominio.
Este hallazgo es un indicio de la hipóstesis esbozada en un comienzo. El uso de pronombres es poco común en los textos con estilo informativo.

5.1.4. Cantidad de URLs, Hashtags, Mentions y Dígitos

A continuación se observa la correlación de las distintas variables con la variable target.

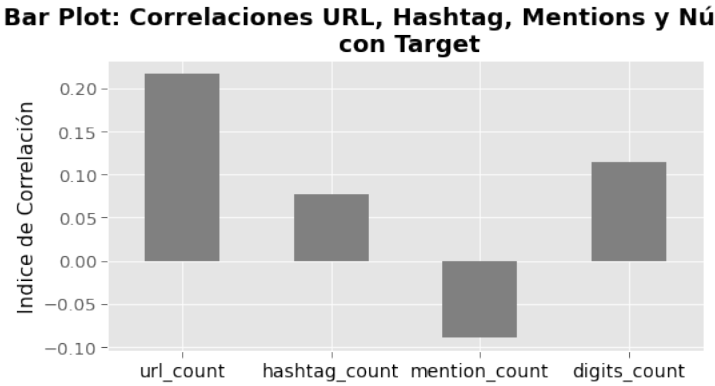

Figura 15: Correlaciones de URL, Hashtag, Mention y Dígitos con la variable Target

Observando algunos tweets es posible intuir que la presencia de URLs responde a que contienen noticias o redacciones similares en su formato y que el usuario pega el url acompañando el mismo. Pareciera que los url funcionan como un re-aseguro de la información, de ello se desprende la correlación positiva entre la frecuencia de URLs por tweet y la variable target.

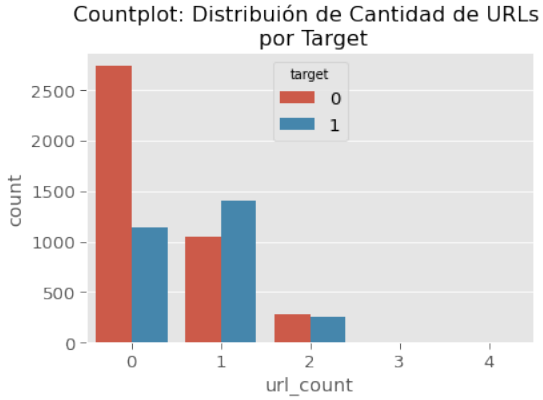

Figura 16: Distribucion de Cantidad de URLs, por Target

Eso también se puede observar comparando las distribuciones. Para el

caso de target 1 la frecuencia relativa es mayor para los tweets que tienen un url que para los que tienen cero. Para target 0 es al revés. Es mayor la cantidad de casos en que tienen cero url que los que tienen una. Podría concluirse que las url funcionan como un 'reaseguro de la información", también pueden ser tweets que refieren a noticias publicadas en la web.

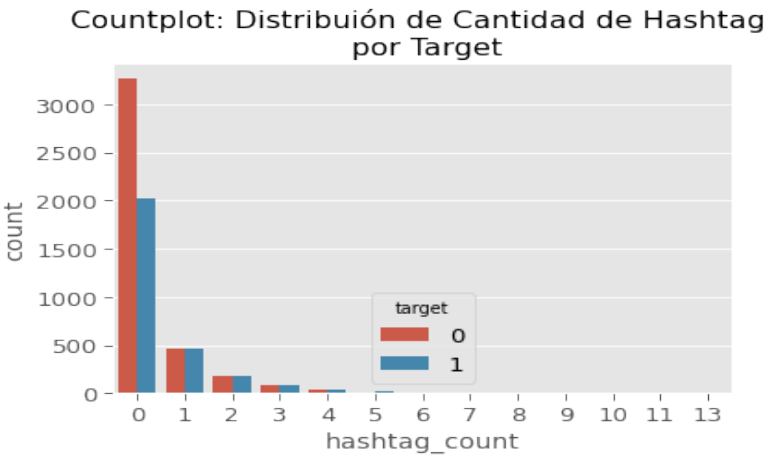

Figura 17: Distribucion de Cantidad de Hashtags, por Target

Los hashtags, por su parte, acompañan en menor medida la relación positiva que tienen las URL con el target. Es por eso que el coeficiente de correlación es de 0.19 entre hashtags y url. De todas maneras es una relación debil con un coeficiente de correlación entre target y hashtag por debajo de 0.10. No queda claro que los hashtags sean un tipo usual de las noticias reproducidas en los tweets.

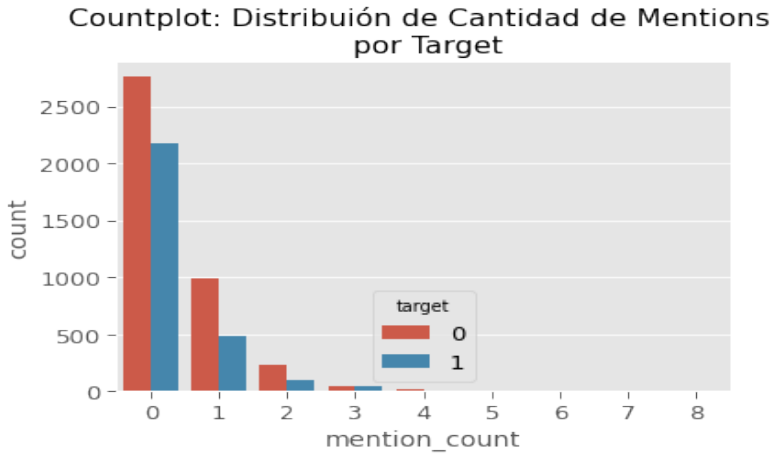

Figura 18: Distribucion de Cantidad de mentions, por Target

Las mentions suelen utilizarse para comunicarle algo a alguien, tienen un costado informal que aparece más en los tweets que no son reales que en los reales.

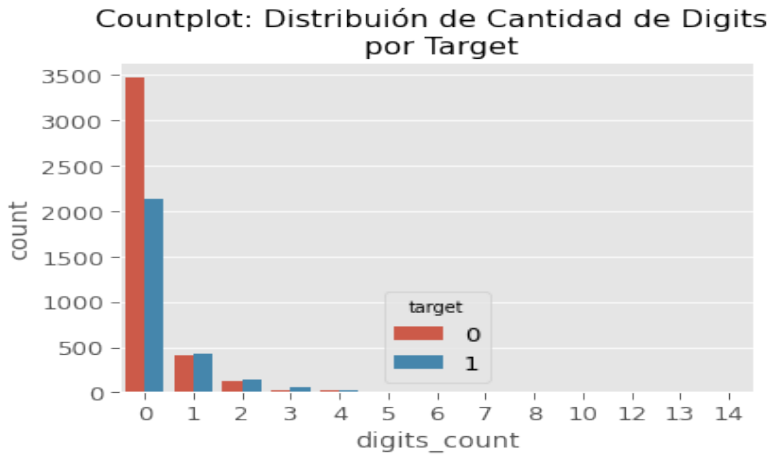

Figura 19: Distribucion de Cantidad de Digits, por Target

Por último los dígitos están presentes mayoritariamente en los tweets reales. Del total de los registros en los cuales hay tweets que tienen números, el 53.22 % corresponde a target 1. De los que no tienen dígitos el 70.45 % de los que no tienen dígitos en su texto son target 0. Es por ello que esta variable correlaciona positivamente con target.

5.2. Clases de Palabras

Una parte importante del Natural Language Processing (NLP) es la que refiere al análisis de las partes de un texto (*parts of speech tagging*). Se trata de identificar la clase de palabra según el lugar que ocupa en el texto. La pregunta que guía esta sección es qué relación tienen con el target.
En el siguiente plot observamos el promedio de apariciones de cada clase de palabra.

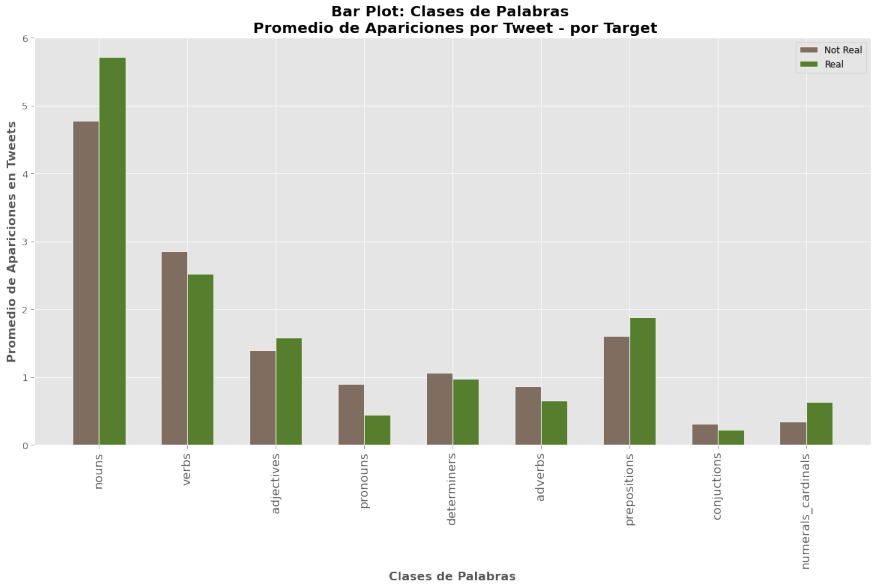

Figura 20: Frecuencia media de clases de palabra por Target

Sustantivos y adjetivos aparecen más veces en cada tweet en los de target 1 que en los de target 0. Mientras que en verbos y adverbios son más frecuentes en target 0.

Un dato llamativo es que los tweets con target 1 pasan del 40.9 % de la muestra a 35.2 % cuando sólo tomamos en cuenta aquellos registros en los cuáles el texto posee al menos 1 adverbio y 1 verbo conjuntamente. Inversamente, la probabilidad de que un tweet sea target 0 cuando en un mismo tweet se combinan verbo y adverbio pasa de 60 al 65 % aproximadamente.

Esto último, se relaciona con la hipótesis de que los textos de target 1 pueden tener una escritura periodística, es decir, el verbo se usa en su significado más prototípico.

Otro clásico de la escritura periodística es que hay sustantivos y no hay pronombres. Es decir, en vez de utilizar "They are making a vaccine against the Covid-19", es más común que el pronombre "they" sea reemplazado por un sustantivo, al estilo "Scientists are making a vaccine against the Covid-19". Cuando se verifica esta situación el porcentaje de registros con target 1 aumenta de 40.9 % a 50.57 %.

En la relación sustantivo-verbo podemos ver que la importancia relativa es mayor en target 1 que en target 0. Se puede suponer que se debe a que los sustantivos proveen más información que los verbos, hacen mención a cosas.

Estas relaciones también se pueden observar a través de las correlaciones de cada clase de palabra con la variable target. Correlaciones positivas muestran que a mayor (o menor) uso de determinada clase mayor (o menor) target. Correlaciones negativas muestran una relación inversa, a mayor (menor) frecuencia de una clase de palabra, menor (mayor) target.

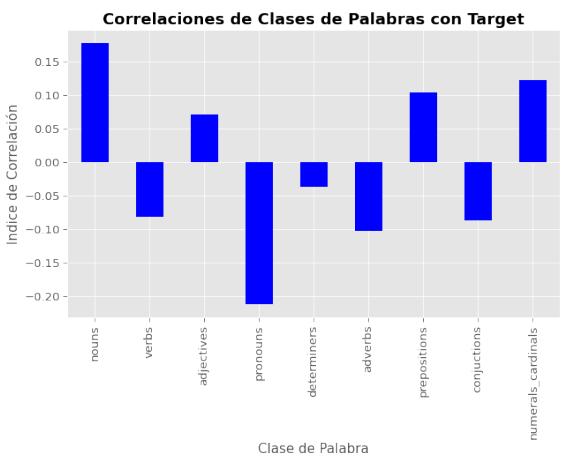

Figura 21: Correlaciones de Clases de Palabras con Target

Jugando a la probabilidad condicional se llega a 67.46 % de probabilidad teniendo textos con al menos un sustantivo, una preposición, un adjetivo, un cardinal y cero adverbios y cero pronombres, con el resto de las variables caeteris paribus. Aunque realmente, uno no pueda decir qué tipo de texto se enmarca en este juego, lo que sí es posible afirmar es que la intersección de estos conjuntos representa el 7.89 % del total de los registros solamente.

Uno más. La intersección de los siguientes conjuntos tiene una probabilidad condicional de 80 % de que el tweet sea target==0: ninguna preposición, ningún cardinal, ningún adjetivo y un adverbio, con el resto de las variables caeteris paribus. Pero dicha intersección representa un 0.5 % del dataset. Se incurre en un costo muy alto de representatividad para alcanzar esa probabilidad alta.

Un buen juego sería encontrar un equilibrio entre combinaciones de conjuntos que tengan una probabilidad alta y que, al mismo tiempo, no genere un trade off tan importante en la representatividad de dicha intersección.

Por otro lado, observando las subclases de palabras podemos incorporar más alementos.

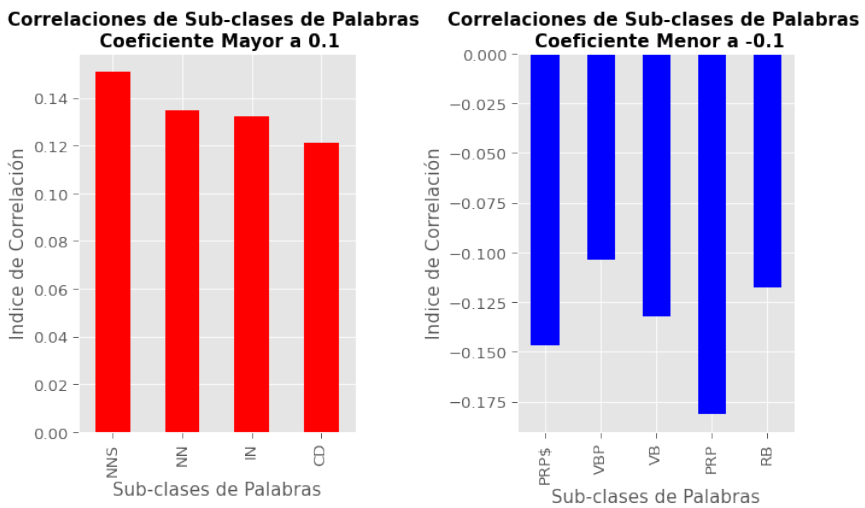

Figura 22: Correlaciones de subclases

Dentro de las sub-clases que tienen un coeficiente mayor a 0.1 se observan sustantivos comunes plurales (‘NNS’), sustantivos comunes singulares

('NN'), preposiciones ('IN') y cardinales ('CD'). En las sub-clases que correlacionan negativamente las cinco más importantes son pronombres personales ('PRP'), pronombres posesivos ('PRP$'), verbos ('VB'), verbos en presente singular sin tomar tercera persona y adverbios ('RB').

En el siguiente apartado se trata el caso especial de los pronombres dado que es muy llamativo el uso que se le da en target 0 y en target 1 comparativamente.

5.2.1. Análisis de Pronombres

Se propone un análisis de cantidad de apariciones de pronombres en tweets según su target: ¿Los tweets que refieren a desastres, hablan en primera o segunda persona, o esto es más común en los tweets de target 0?

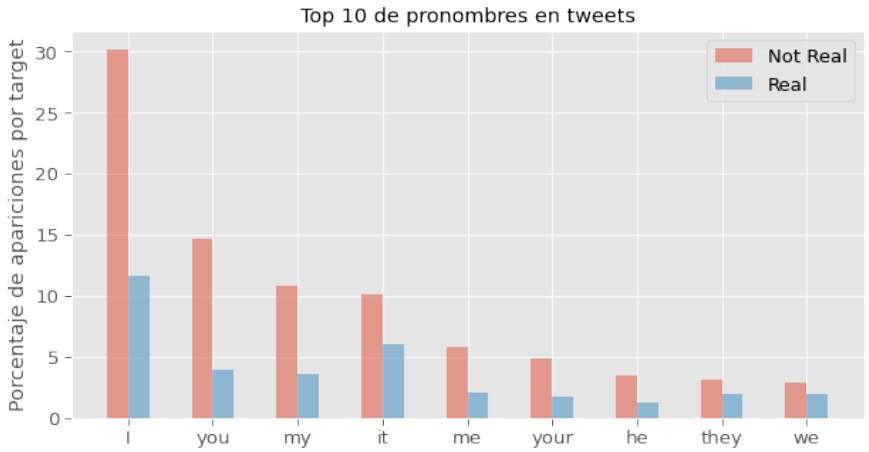

Figura 23: Top 10 pronombres en tweets

En la figura se observan los pronombres con más apariciones para ambos targets. Se observa que los mismos 10 pronombres aparecen en ambos rankings. Sin embargo, todos los porcentajes de aparición de target Not Real superan a los de target Real. En partircular, los valores de 'I', 'you' y 'my' para los targets Not Real triplican en porcentaje a los de target Real. La naturaleza de un tweet que expresa un desastre responde a un estilo de redacción con poca cantidad de pronombres respecto a un tweet 'normal'. Esto se relaciona con que al informar un desastre, se deben dar datos concretos,

por ende debe haber mayor cantidad de sustantivos propios o comunes. En los siguientes apartados se utilizará la técnica de n-grams para abordar el uso de las palabras. Dado que para ello es necesario no utilizar los stopwords, es que se destinó una sección especial para pronombres.

5.2.2. Análisis de Stopwords

Los stopwords para el análisis de N-gramas (próxima sección) se consideran ruido. Sin embargo, vale la pena preguntar: ¿Ese ruido, es el mismo para los dos targets? O hay palabras que suelen usarse más en una emergencia que en otro? Se realiza un análisis a través de un heatmap:

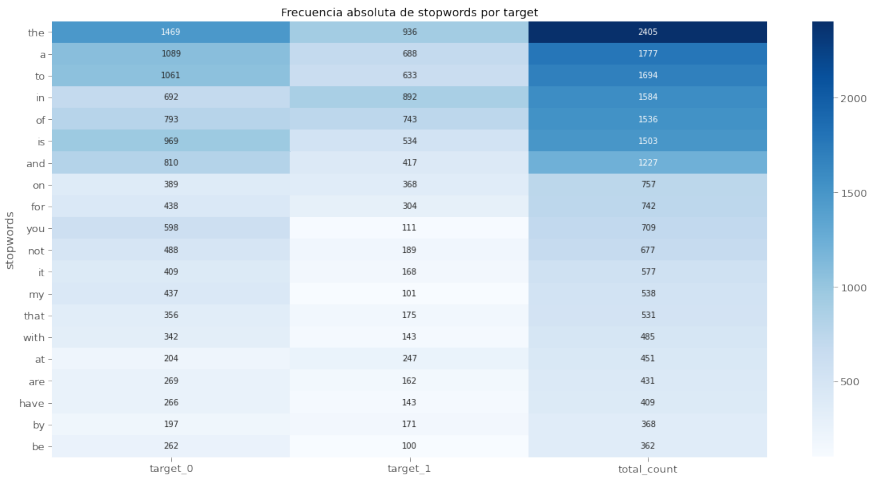

Figura 24: Top 20 de stopwords en ambos targets

Se observa una gran cantidad de stopwords en ambos, siendo mayor su cantidad en el target 0. Palabras como 'the', 'a', 'to', 'and' en target 0 superan ampliamente en cantidad de ocurrencias. Preposiciones referentes a ubicaciones ('in','at') predominan en target 1. Este resultado motiva a realizar el análisis mostrado en la siguiente sección.

6. Análisis de N-gramas

Al realizar un cleaning de los textos, queda como resultado la estructura básica de las oraciones. Vale preguntarse entonces: ¿Los tweets que hablan de desastres usan las mismas expresiones que los que están hablando de otra cosa? ¿Hay palabras que sean comunes a ambos? Una primera aproximación a responder estas preguntas se ilustra en los siguientes wordclouds:

Figura 25: Wordclouds Real / Not Real

Sin demasiado análisis, ya se observa que las palabras del Wordcloud 'Real' son bastante negativas respecto al Wordcloud 'Not Real'. Efectivamente hay palabras que se repiten en ambos. Para un mejor análisis se propone una descomposición de los textos en unigramas, bigramas y trigramas.

6.1. Unigramas

6.1.1. Unigramas únicos de cada target

Se muestran a continuación los Unigramas más repetidos que únicamente aparecen en el target 1.

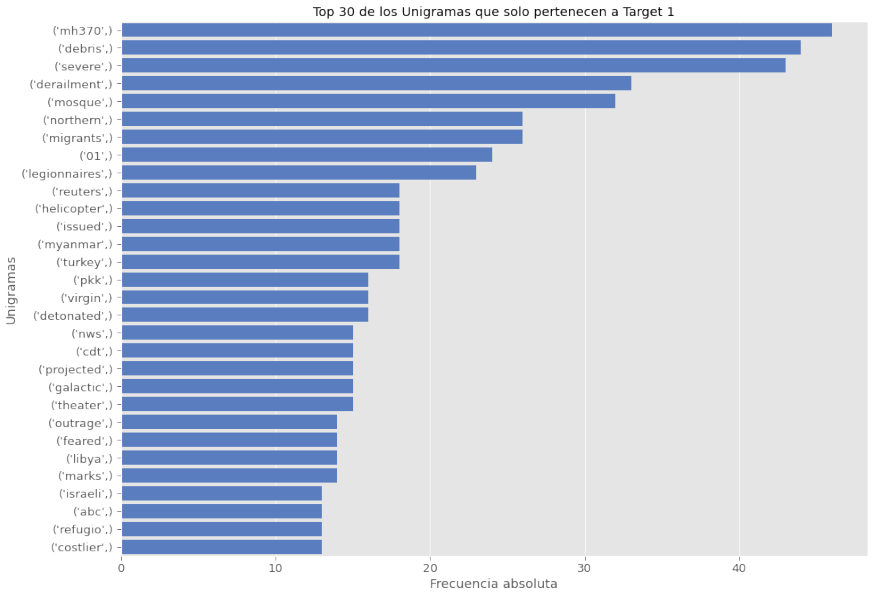

Figura 26: Top 30 unigramas que solo aparecen en target 1

Los unigramas carecen de contexto, sin embargo las palabras por si solas llaman la atención y es posible relacionarlas con un desastre. 'MH370' fue el vuelo de Malaysia Airlines que terminó en tragedia.'Reuters' y 'abc' son canales de noticias. También aparecen gerundios y países como Myanmar, Turkey, israeli y Lybia. El resto de los unigramas son en su mayoría sustantivos y es vocabulario que podría usarse en el contexto de un desastre. Se analizan a continuación los unigramas para target 0:

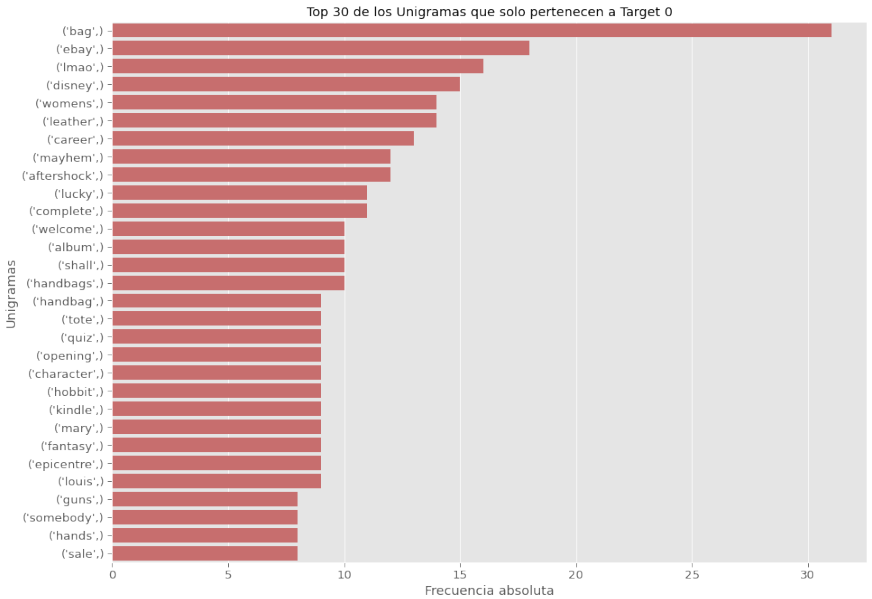

Figura 27: Top 30 unigramas que solo aparecen en target 0

Los unigramas se diferencian en gran medida de los anteriores. Por un lado, se encuentran marcas como Ebay, Kindle y Disney. Otras palabras que aparecen son 'handbag' o 'sale'. Todo esto muestra un lado comercial-publicitario de los tweets de target 0. En general los unigramas encontrados tienen caracter positivo y no es usual verlos en el contexto de un desastre: es muy improbable encontrar palabras como 'fantasía' o 'hobbit' para informar sobre un bombardeo en Libia.

6.1.2. Unigramas comunes en ambos targets

En el siguiente Heatmap se observan los unigramas con más ocurrencias y que pertenecen a ambos targets.

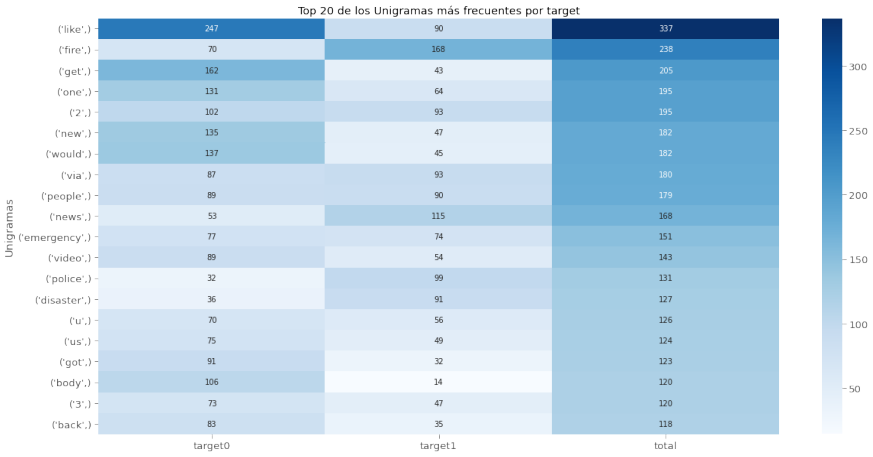

Figura 28: Heatmap: Top 20 unigramas que aparecen en ambos targets

Este ranking muestra que hay palabras que pueden usarse tanto en desastres como en otros contextos, lo cual implica que por si solas, no nos dan pauta de que se trate o no de un desastre. Ejemplo de esto es la palabra 'emergency'. A continuación se muestra un ejemplo de ocurrencia de esta palabra en target 0: I'm glad when I call someone it's not an emergency since they never answer their phones or call back??
A continuación se muestra un barplot con la probabilidad de que al encontrar un unigrama, pertenezca a target 0 o 1.

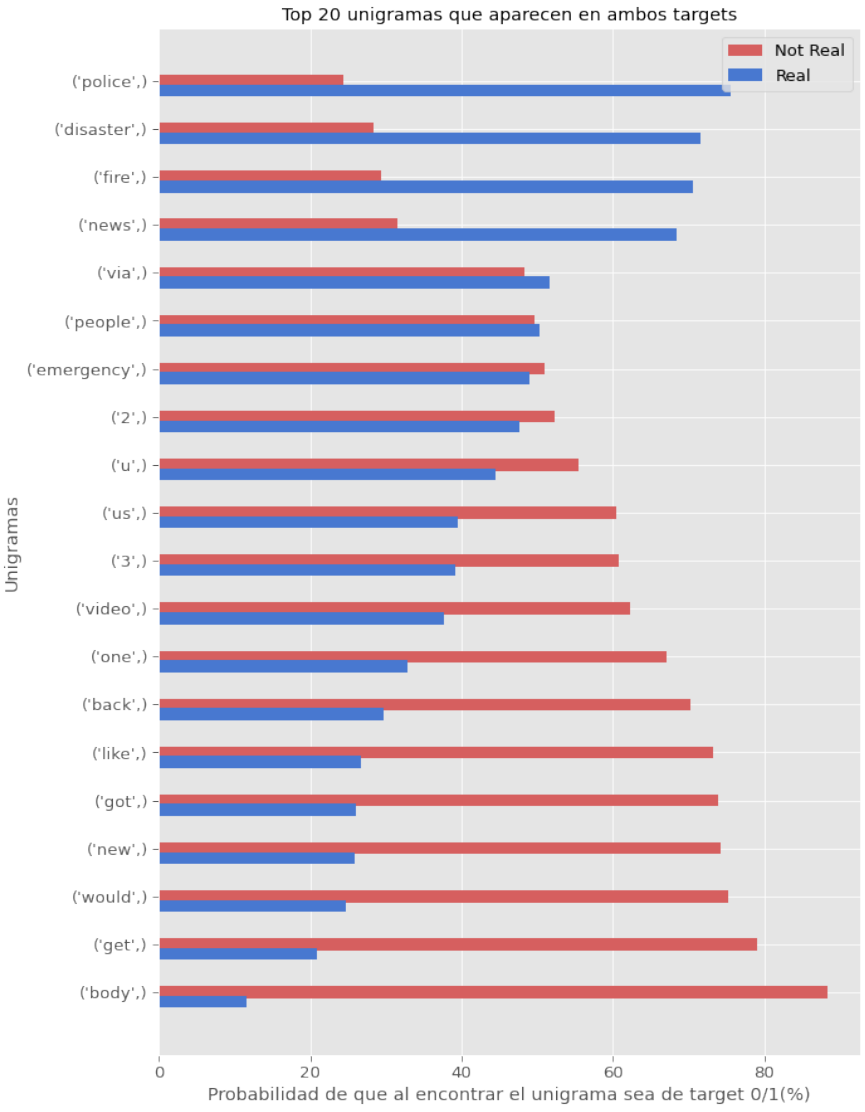

Figura 29: Top 20 unigramas que aparecen en ambos targets con sus probabilidades de ocurrencia

Las palabras 'police', 'disaster', 'fire' y 'news', tienen una clara tendencia a pertenecer a tweets desastrosos. 'Via', 'people' y 'emergency' tienen la misma probabilidad de pertenecer a las dos clases. El resto de los unigramas son más probables de encontrar en target 0. En su mayoría se tratan de verbos en sus formas conjugadas, lo cual es de esperar ya que provienen de individuos.

6.2. Bigramas

6.2.1. Bigramas únicos de cada target

Se muestran a continuación los Bigramas más repetidos que únicamente aparecen en target 1.

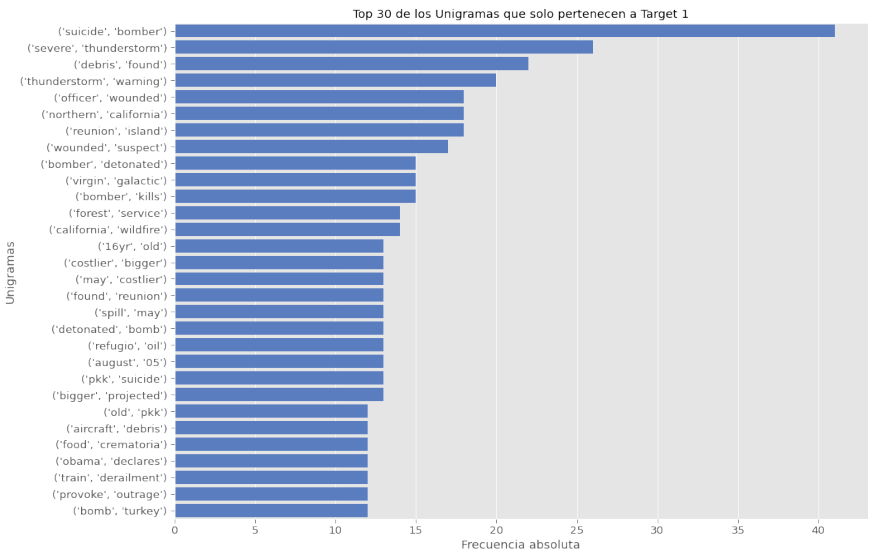

Figura 30: Top 30 bigramas que solo aparecen en target 1

Los bigramas por si solos ya describen situaciones que despiertan una alarma. Varias de las palabras también aparecen en la columna keywords: suicide bomber, debris, wildfire son ejemplos de esto.
Se analizan a continuación los bigramas para target 0:

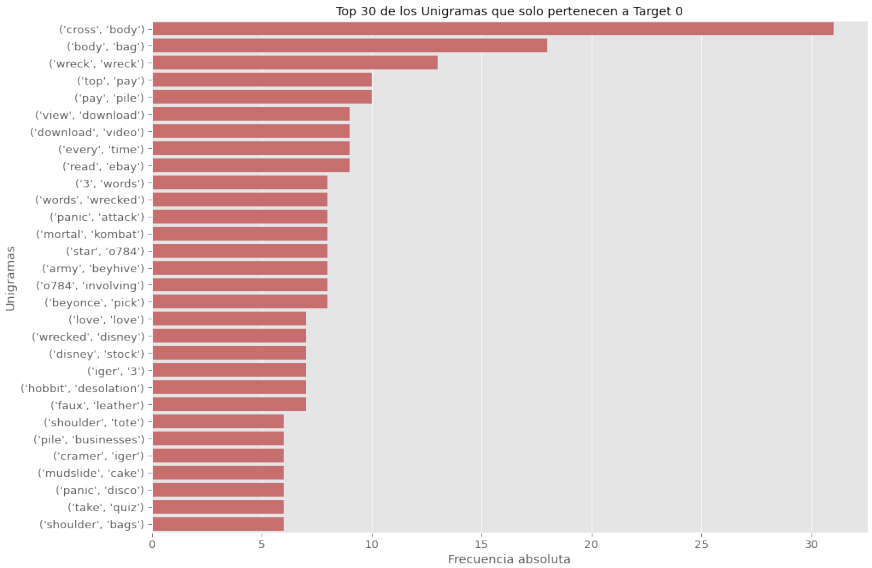

Figura 31: Top 30 bigramas que solo aparecen en target 0

Para target 0 se obtienen bigramas con contenidos que semánticamente no se relacionan con desastres.
Otro punto a resaltar es que comparando ambos rankings, las frecuencias absolutas de target 1 son mayores que las de target 0, lo que demuestra que los bigramas de target 1 son menos variables que los de target 0.

6.2.2. Bigramas comunes en ambos targets

Se analizan los bigramas que se encuentran en ambos targets. Se muestra a continuación un heatmap con los bigramas y cómo se reparten entre clases.

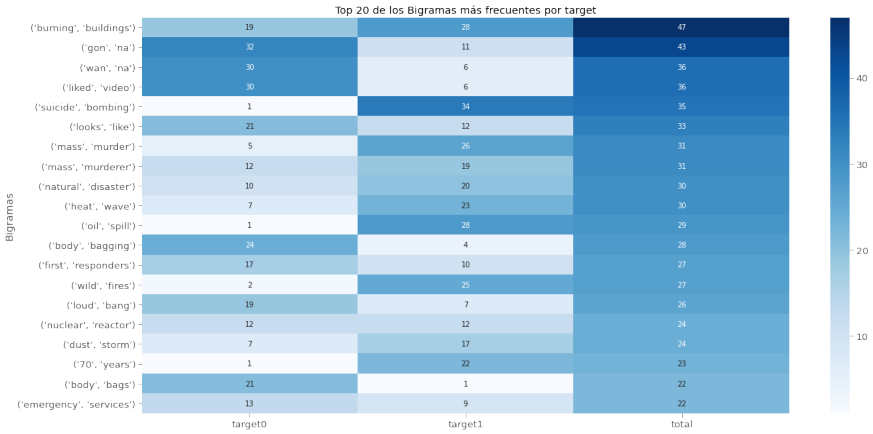

Figura 32: Heatmap: Top 20 bigramas que aparecen en ambos targets

Por un lado, hay bigramas que se inclinan fuertemente a representar desastres pero que tienen alguna ocurrencia de target 0 (suicide bombing, oil spill, wild fires, mass murder). Por otro, ocurre lo contrario con otros (body bags, liked video, wan-na, gon-na). Existen otros bigramas que se encuentran en el medio como nuclear reactor. Esta variabilidad se se ilustra en el siguiente barplot:

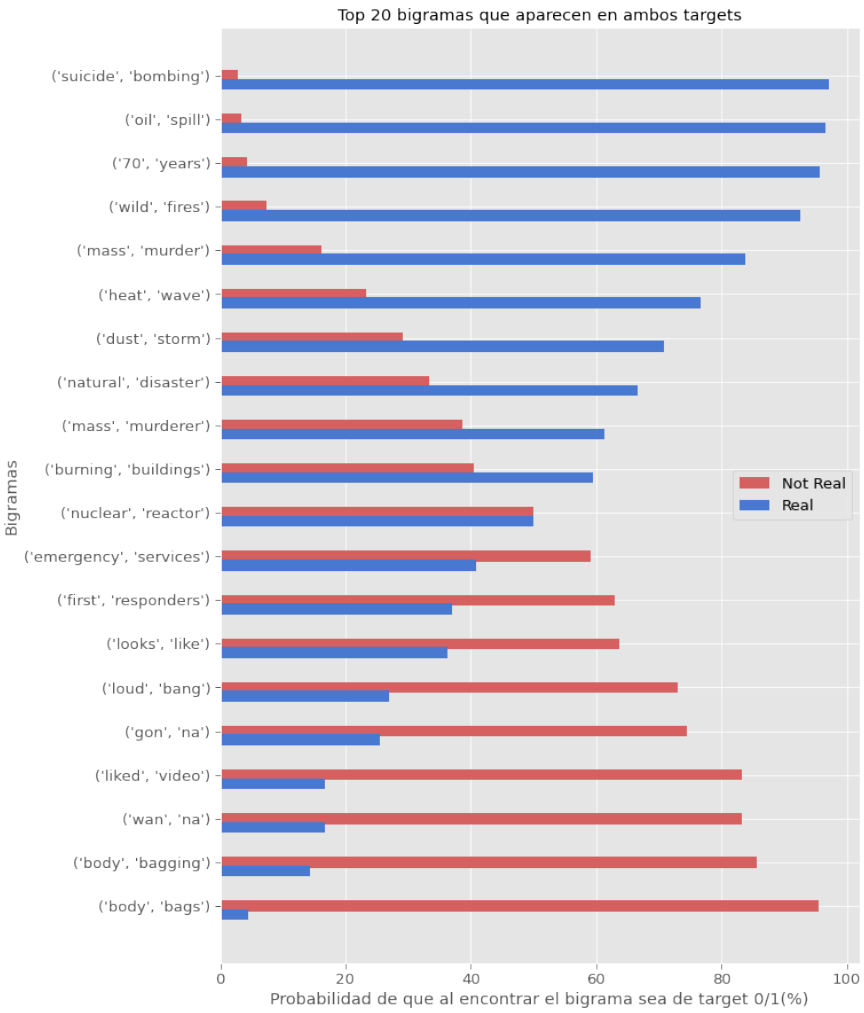

Figura 33: Top 20 bigramas que aparecen en ambos targets con sus probabilidades de ocurrencia

Si observamos los bigramas de arriba para abajo, observamos que se comienza con palabras intuitivamente relacionadas con desastres y con target 1 alto. A medida que se desciende, va bajando la veracidad de los tweets

hasta llegar a palabras que poco tienen que ver con desastres o que suelen usarse en otros contextos.

6.3. Trigramas

6.3.1. Trigramas únicos de cada target

Se muestran a continuación los trigramas más repetidos que únicamente aparecen en target 1.

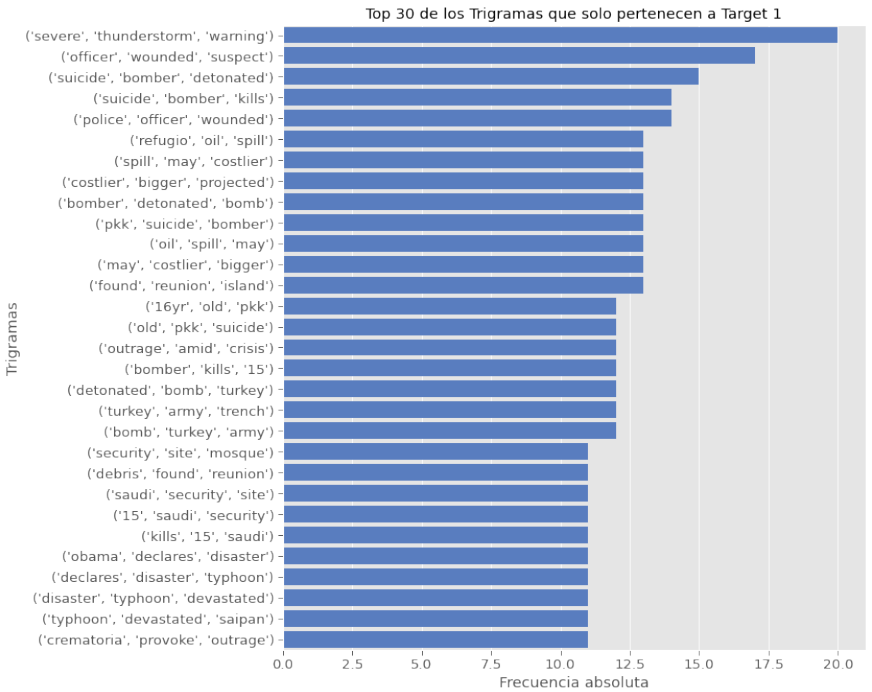

Figura 34: Top 30 trigramas que solo aparecen en target 1

Los trigramas obtenidos tienen un alto nivel de contenido: al leerlos, se puede tener una idea del desastre en cuestión.
Observemos qué ocurre en los trigramas del target 0.

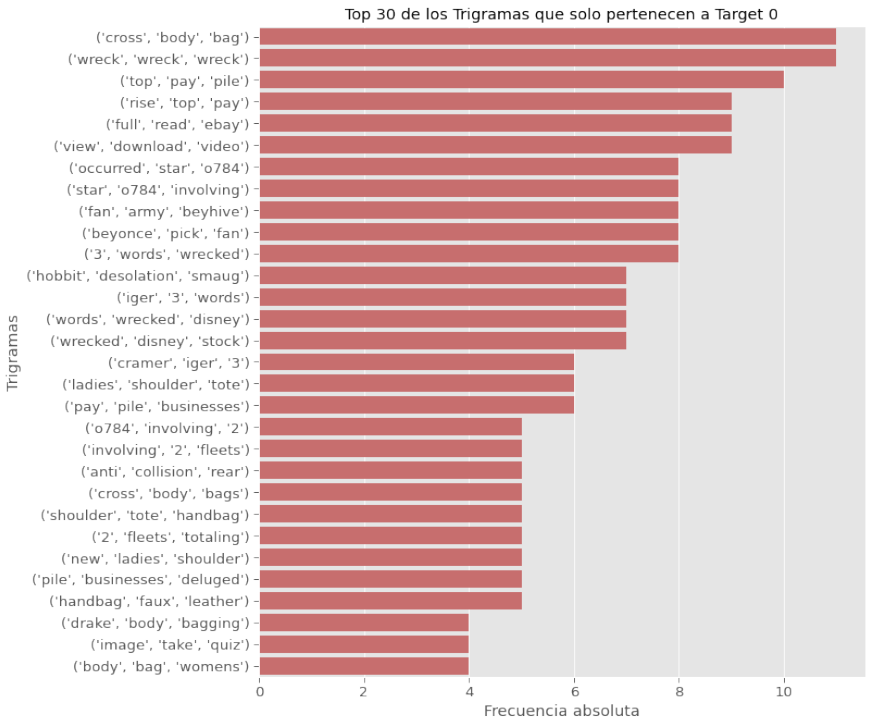

Figura 35: Top 30 trigramas que solo aparecen en target 0

Para target 0 se obtienen trigramas con contenidos que semánticamente no se relacionan con desastres.

6.3.2. Trigramas comunes en ambos targets

Se repite el análisis para descubrir qué trigramas se encuentran en ambos targets. Se muestra a continuación un heatmap con los trigramas y cómo se reparten entre clases.

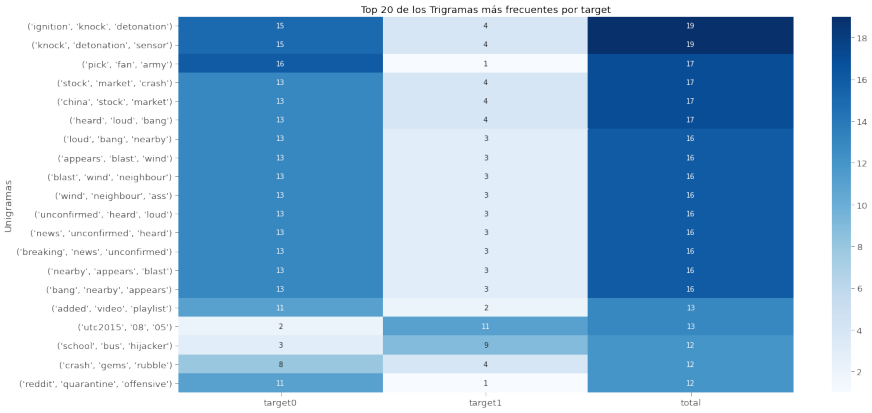

Figura 36: Heatmap: Top 20 trigramas que aparecen en ambos targets

Se observa en la columna total que las cantidades bajan radicalmente respecto al análisis en uni/bigramas. Esto era esperable, ya que sumar palabras se vuelve menor la probabilidad de que dicha secuencia se repita.
El contenido de estos trigramas en general no parece de contenido significativo y varios parecen venir de un mismo tweet, lo cual implica que existen tweets especialmente similares que no fueron filtrados por el cleaning.
A pesar del bajo número de muestras obtenidas, es interesante observar que las cantidades de trigramas en común en target 0 supera en número a target 1. Esto también se ilustra en el siguiente barplot:

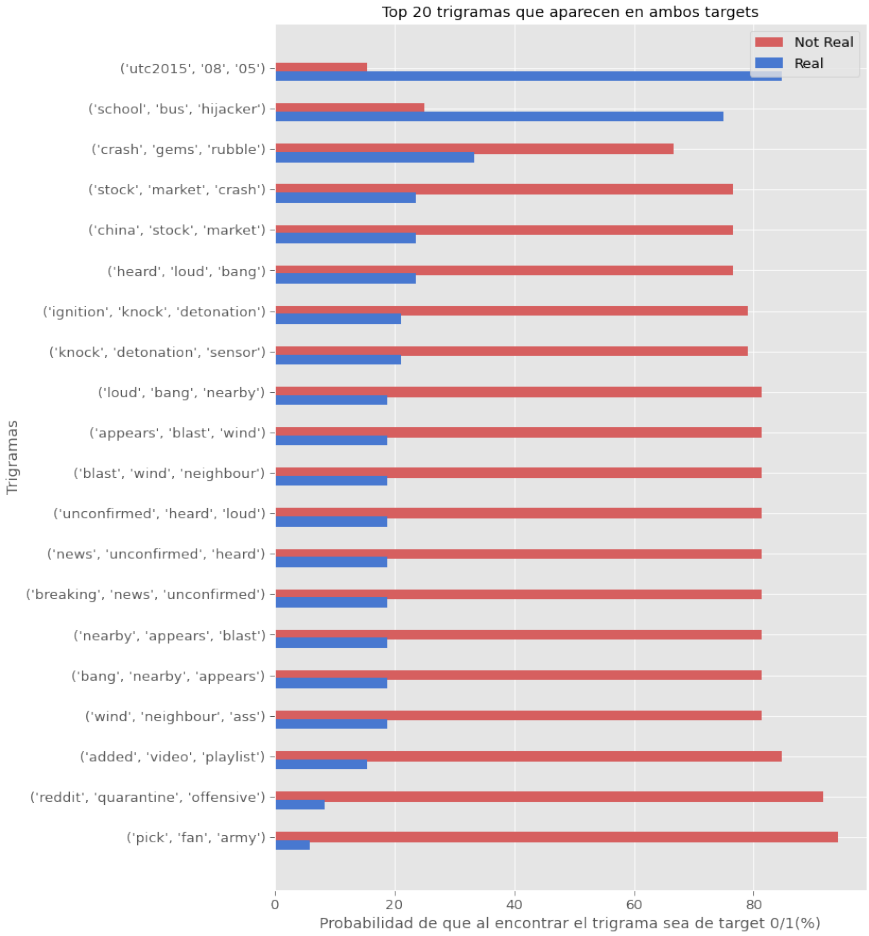

Figura 37: Top 20 trigramas que aparecen en ambos targets con sus probabilidades de ocurrencia

Dado el bajo número de muestras obtenidas, pese a tener una descripción más detallada de la información, los trigramas no suman mayor información respecto a lo obtenido en los bigramas.

7. Análisis de ubicaciones (países/ciudades) en el texto

Dado que en el estudio de N-gramas aparecieron nombres propios de países/ciudades, surge la siguiente pregunta: ¿Los tweets hablan de ubicaciones específicas en su contenido? Intuitivamente, es posible imaginar que al informar de un desastre, es necesario hablar de dónde ocurrió. Se propopone un análisis básico en el que se buscan nombres de países y estados de Estados Unidos (se elige este país ya que fue el que más ocurrencias tiene en la columna 'locations').

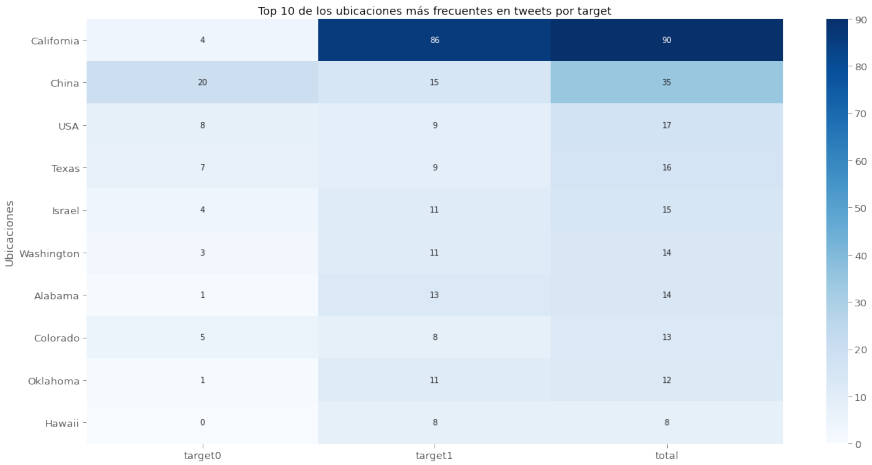

Figura 38: Heatmap: Top 20 de las ubicaciones más frecuentes por target

Sorprende la cantidad de ocurrencias que tiene California en los tweets sobre desastres. En gran parte se explica con que en 2019 hubo casi 7 mil incendios de menor y mayor escala (Fuente: https://disasterphilanthropy.org/disaster/2019-california-wildfires/)
Otras ubicaciones que aparecen pero tienen menor cantidad de ocurrencias. Se analiza a continuación la probabilidad de que al encontrar una ubicación en el texto, sea o no un desastre:

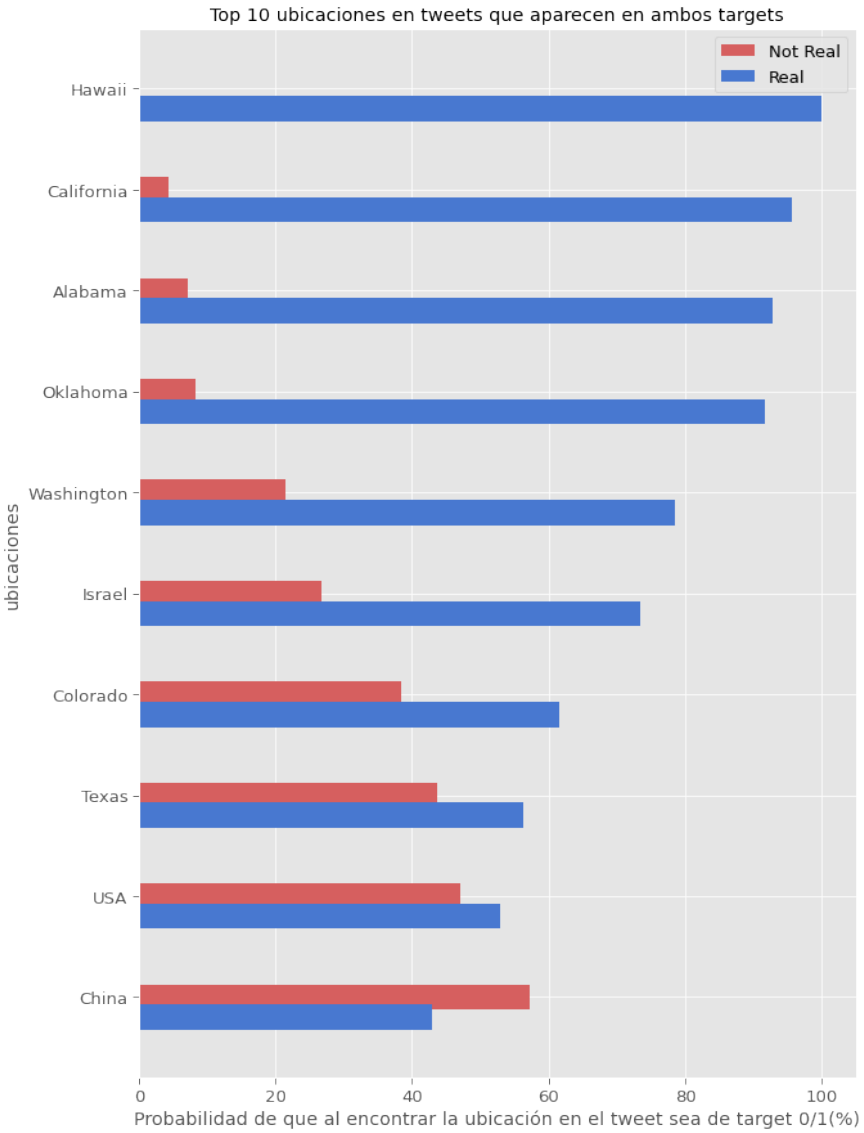

Figura 39: Top 10 ubicaciones que aparecen en ambos targets y sus probabilidades

Se observa que en la mayoría de estas ubicaciones, la probabilidad de que el tweet trate de un desastre supera a la posibilidad que no lo sea. No se puede llegar a esta conclusión con lugares como China, USA, Texas y Colorado.

8. Conclusiones

Se detallan a continuación las conclusiones a partir de los resultados obtenidos:

- No se encontró relación entre la columna 'Locations' y los targets.
- El cleaning de los textos fue un paso clave para lograr un correcto análisis de las características identificadas.
- En los keywords que refieren especificamente a desastres predominan tweets con target 1. Para target 0, los keywords son más dispersos ya que también sirven para describir otra clase de eventos.
- Los tweets reales tienen una tendencia a tener mayor cantidad de caracteres y palabras. A su vez las palabras son levemente más largas.
- URLs y dígitos correlacionan positivamente con la variable target, no así las mentions. Se intuye que las URLs pueden ser utilizadas como reaseguro de la información. Los hashtags tienen una correlación positiva pero baja. Es ciertamente ambigüo su uso.
- En target 1 son más importantes los sustantivos, adjetivos, preposiciones y cardinales, todas clases de palabras que abundan en textos o titulares con formato informativo o periodístico. En cambio en target 0 predominan los verbos, adverbios, los pronombres.
- Los pronombres y adverbios son las clases de palabras que mayor correlación negativa tienen con la variable target
- Los sustantivos y cardinales son las que mayor correlación positiva tienen con la variable target.
- Dentro de las subclases las de mayor correlación positiva son los sustantivos comunes plurales y singulares, las preposiciones que indican lugar y los cardinales.

- En target 0 predominan los pronombres personales y posesivos, verbos y adverbios.
- En particular los pronombres 'I','you' y 'my' para los target 0 triplican en porcentaje a los de target 1.
- El análisis de N gramas demostró que en target 1 las palabras más utilizadas tienen una connotación negativa respecto a las de target 0.
- En target 0 aparecen palabras comerciales
- Hay predominancia de stopwords en target 0, pero en target 1 son mayores las ocurrencias de preposiciones de ubicación.
- Existe una correlación entre nombres propios de países/ciudades con target 1.

Se encontraron diferencias sustanciales entre ambas clases. Los resultados de este análisis serán de gran utilidad a la hora de entrenar un modelo predictivo capaz de clasificar los tweets del set de prueba.