Comparación de clasificadores con y sin filtrado

Sea que tras el filtrado del conjunto inicial, hemos construido un clasificador que ahora queremos probar con un conjunto de prueba de tamaño 100.000.

Podemos proceder de dos maneras:

1. Aplicar el clasificador directamente a las 100.000 instancias del conjunto de prueba.

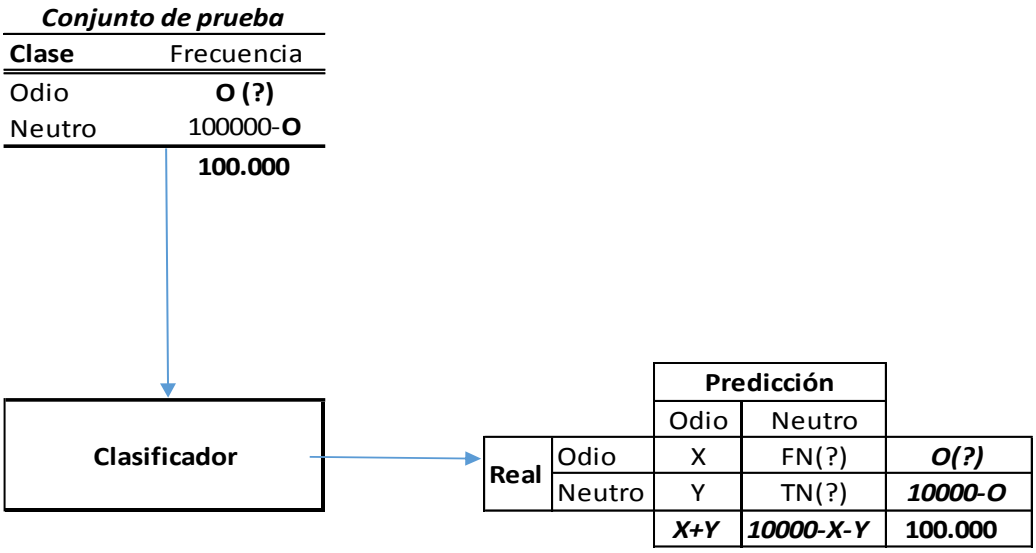

Fig. 1: Clasificación en Conjunto de Prueba sin Filtrado

Como se observa, debido a su tamaño y no estar etiquetado ignoramos el número de tuits de odio - **O**- contenidos en el fichero de prueba pero, tras la clasificación, mediante simple inspección, de los tuits clasificados como de odio (que serán pocos) podemos averiguar los valores X - TP - e Y - FP – y, por lo tanto el valor total de los neutros predichos por el modelo (1000-X-Y).

Es claro que FN y TN no pueden calcularse directamente ya que ello exigiría la lectura de casi 100.000 tuits lo que es impensable.

No obstante, es posible llevar a cabo una estimación de FN y TN ya que sabemos que la proporción de tuits de odio en el conjunto de prueba similar al de la población) se encuentra en torno al 2 ‰, podemos construir la tabla para un error ε y un nivel de confianza de 1-α tal que:

$$P(|\hat{p} - p| < \epsilon) \leq 1 - \alpha$$

que proporciona los siguientes resultados:

P	0,10%	0,20%	0,30%
ε	**n**	**n**	**n**
0,10%	3.696	7.122	10.306
0,15%	1.678	3.296	4.859
0,20%	951	1.881	2.793

Tabla 1: Tamaño de Muestra Binomial

Y dado que $v = n \cdot \hat{p}$ sigue una $N\left(n \cdot p; \sqrt{np(1-p)}\right)$ tendremos que

$$P\left(\frac{|\frac{v}{n} - p|}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}}\right) = P\left(-z_{\frac{\alpha}{2}} < \frac{\epsilon}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{\alpha}{2}}\right) \leq 1 - \alpha$$

Y como en la N (0;1) para α=0,025 $z_{\frac{\alpha}{2}} = 1{,}96$ tendremos que el tamaño de muestra se puede calcular también por:

$$\frac{\epsilon}{\sqrt{\frac{p(1-p)}{n}}} = 1{,}96 \rightarrow n = p(1-p)\left(\frac{1{,}96}{\epsilon}\right)^2$$

Con similares resultados:

P	0,10%	0,20%	0,30%
ε	**n**	**n**	**n**
0,10%	3.838	7.668	11.490
0,15%	1.706	3.408	5.107
0,20%	959	1.917	2.873

Tabla 2: Tamaño de Muestra con la Normal

Así pues, podemos estimar FN y TN con una muestra del conjunto de prueba de tamaño 3.000, lo que es viable.

2. Aplicarlo tras filtrar.

También podemos clasificar tras el filtrado, asignando *todas las instancias que no pasan el filtro como neutras* como se muestra en la Fig. 2:

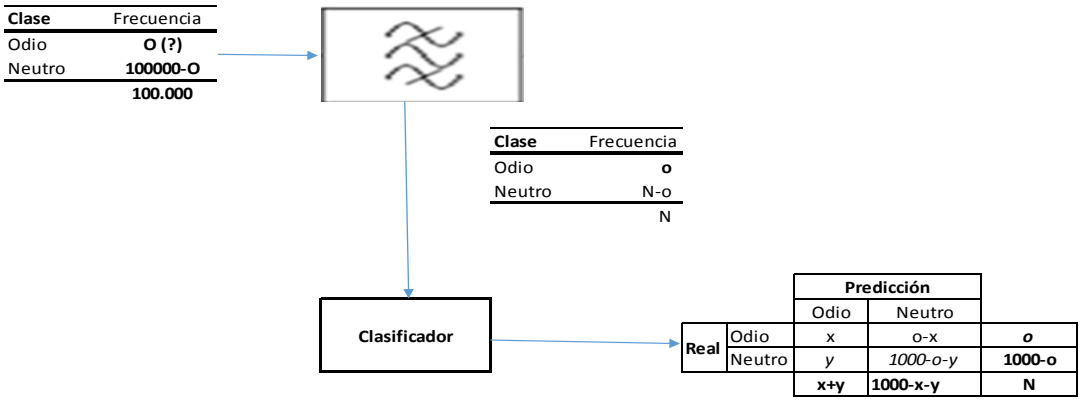

Fig. 2: Clasificación tras Filtrado

En este caso, también será necesario completar la matriz de contingencia con los valores estimados de FN y TN:

		Predicción		
		Odio	Neutro	
Real	Odio	x	FN(?)	***O***
	Neutro	y	TN(?)	**10000-O**
		x+y	**100000-N+o-x**	**100.000**

Tabla 3: Matriz de Contingencia en Clasificación tras Filtrado

Que, como antes, se hará muestreando el conjunto de prueba.

Solo lo haríamos con el clasificador que nos haya dado mejores resultados en conjunto de prueba.

Previo al muestreo, la comparación entre los valores **X –x** e **Y- y** nos da una primera impresión acerca del funcionamiento de cada uno de los procdimientos.

Clasificador bayesiano.

$$\mathrm{P(C|f_1,..,f_m)}P(f_1,..,f_m) = P(f_1,..,f_m|C)\mathrm{P(C)}$$

$$\mathrm{P(C|f_1,..,f_m)} = \frac{P(f_1,..,f_m|C)\mathrm{P(C)}}{P(f_1,..,f_m)}$$

$$classify(\mathrm{f_1,..,f_m}) = \underset{c}{\mathrm{argmax}}\frac{P(f_1,..,f_m|C)\mathrm{P(C)}}{P(f_1,..,f_m)}$$

Es claro que $P(f_1,..,f_m|C)$ y *P(C)* son diferentes en el conjunto de prueba y en el filtrado.

Si p.e. con un solo atributo *f* tuviésemos las siguientes probabilidades:

	P(c)	
Clase	**Conjunto**	**Filtrado**
Odio	1%	35%
No odio	99%	65%

	P(f|C)	
Clase	**Conjunto**	**Filtrado**
Odio	80%	90%
No odio	7%	3%

Fig. 3: Probabilidades bayesianas en conjuntos inicial y filtrado

Es claro que un tuit del *conjunto inicial* con el atributo *f* se clasificaría como **No odio** ya que

$$\mathrm{P(Odio|f)} = \frac{P(f|Odio)\mathrm{P(Odio)}}{P(f)} \propto P(f|Odio)\mathrm{P(Odio)} = 0{,}8 \cdot 0{,}01 = 0{,}008$$

$$\mathrm{P(No\ odio|f)} \propto P(f|No\ odio)\mathrm{P(No\ odio)} = 0{,}07 \cdot 0{,}99 = 0{,}069 > \mathrm{P(Odio|f)}$$

Sin embargo, en *el conjunto filtrado* sucede al revés:

$$\mathrm{P'(Odio|f)} \propto P'(f|Odio)\mathrm{P(Odio)} = 0{,}9 \cdot 0{,}35 = 0{,}315$$

$$\mathrm{P'(No\ odio|f)} \propto P'(f|No\ odio)\mathrm{P(No\ odio)} = 0{,}03 \cdot 0{,}65 = 0{,}02 < \mathrm{P(Odio|f)}$$

Y se clasificaría como de **Odio.**