## **Primeros Resultados según Matriz de Confusión**

## Medidas de desempeño de clasificadores.

Como es sabido, la *matriz de confusión* mide el rendimiento de un clasificador.

y se expresa en el caso de que solo existan dos clases como:



Las medidas más inmediatas de evaluación del modelo son:

* Exactitud:
* Tasa de error:

Sin embargo, cuando existe un claro desequilibrio entre clases, como es el caso de los tuits de odio, se utilizarán los indicadores clásicos en clasificación binaria:

1. Precisión (*p*recision)
2. Exhaustividad (*r*ecall)

* **Precisión** (p) es el porcentaje de los *tuits* clasificados *correctamente* como de odio - *TP* -del total de los asignados a dicha clase por el clasificador - *TP+FP* - (% de aciertos).
* **Exhaustividad** (r) es el porcentaje de los *tuits* de la clase odio existentes en el fichero que han sido clasificados correctamente.

## Proceso seguido.

Según lo que se deduce del conjunto inicial de muestra, el contenido de los tuits utilizados para etiquetado es el siguiente



Donde se aprecia claramente que estamos en un clarísimo desequilibrio de clases (0,2% frente a 99,8%).

El proceso que se ha seguido para la obtención del conjunto de entrenamiento parte del conjunto S (muestra) que se filtra mediante una lista de vocabulario obteniendo un nuevo conjunto R de mucha menor cardinalidad que se etiqueta manualmente, produciendo el conjunto de entrenamiento T, a partir del cual se procederá a la estimación de los modelos de clasificación como muestra la Figura 1.

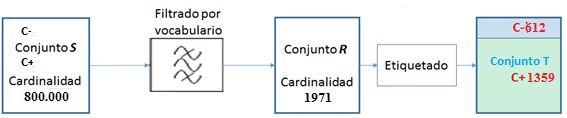


Figura 1

Seguidamente se procede a la prueba de los modelos (Bayes, K-NN, Redes Neuronales,…) a fin de seleccionar el más adecuado sobre un conjunto de cardinalidad≈100.000.

## Resultados obtenidos.

En nuestro caso, los resultados son los siguientes:

1. **Matriz de confusión**:





1. **Medidas de desempeño**:



Puede verse que, como era de esperar, el error es muy pequeño, sin embargo, el rendimiento del clasificador es bastante pobre sobre todo debido a su baja *exhaustividad*, lo que implica que ***el 80% de los tuits de odio originales no son detectados.***

La Figura 2 muestra gráficamente los diferentes conceptos.

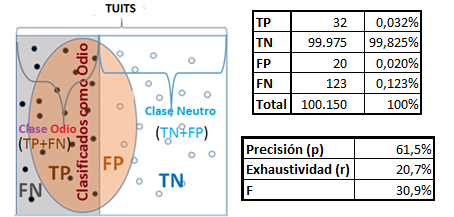


Figura 2

## Conclusión.

Si los resultados mostrados anteriormente son correctos, es claro que, como consecuencia del gigantesco desequilibrio existente entre las clases - 2: 1000 - , la *exhaustividad* de la clasificación es muy pequeña: solamente el 16% de los tuits de odio son clasificados correctamente y la precisión tampoco puede decirse que sea muy elevada.

Es preciso analizar más en detalle los resultados y ver dónde está el fallo, si en el filtro o en el clasificador.

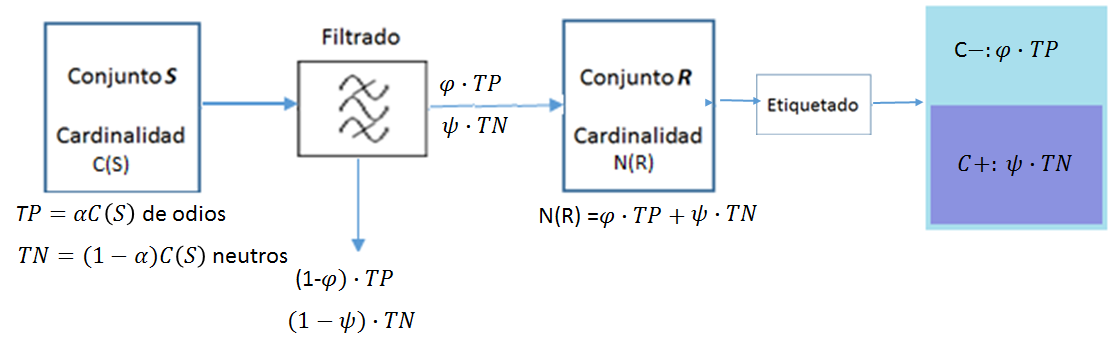


Figura 3: Filtrado de Tuits

Para ello es preciso ver

1º Cuántas instancias de odio (φ∙*TP*) pasan el filtro y

2º Cuántos de los que pasan se clasifican erróneamente.

Todo ello en los dos posibles métodos de clasificación:

1. Sin filtrado previo:



Figura 4: Clasificación sin Filtrado Previo

1. Con filtrado previo y posterior clasificación:



Figura 5: Clasificación con Filtrado Previo