UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Detección de Tuits de Odio.

Juan Carlos Pereira Kohatsu
Tutor:
Ponente:

JUNIO 2017

Tabla de contenido

Índice de Figuras

Índice de Tablas

Resumen.

Las llamadas *redes sociales* constituidas por plataformas tales como **Facebook™, Twitter™** que operan sobre Internet constituyen el soporte de los *medios (de comunicación) sociales* que facilitan el intercambio y la discusión de información, experiencias y opiniones entre individuos de manera rápida y masiva, nunca antes vista en la historia de la humanidad.

El abanico de medios sociales abiertos al uso público es variadísimo y creciente y sus usos son múltiples:

- artículos en *wikis*,
- opiniones sobre la calidad de hoteles, restaurantes (tripadvisor, yelp),
- contactos sociales y profesionales (facebook, linkedin),
- *blogs* o bitácoras *web* (wordpress).

Ciertamente, como todo lo nuevo, la explosión de los medios sociales ha tenido consecuencias que han sido valoradas tanto positiva como negativamente para el conjunto de la sociedad.

Entre los efectos generalmente considerados como negativos, los medios sociales han hecho persistentemente *'visibles'* algunas actitudes de ciertos grupos sociales que, hasta la fecha, solo se mostraban de una manera velada y/o esporádica. Entre ellas destacan las que se traducen en ataques a personas o colectivos en razón de su pertenencia a un determinados grupos definidos por características de nacionalidad, preferencias sexuales, raza, religión…

Este fenómeno junto con un cambio de actitud frente a ciertas conductas o grupos sociales ha motivado que, en muchos países, surja una nueva categoría delictiva: los llamados *delitos de odio* que, en España han sido regulados en 2015 mediante modificación del Código Penal, (1)

Desde el trabajo de Gary Becker (2) sobre crimen y castigo sabemos que los resultados que se derivan de los modelos indican que un *incremento en la probabilidad de sanción o arresto*, sin importar la disposición al riesgo del infractor, *tiene un efecto negativo sobre la oferta de delito*. Sin embargo, *el efecto de un incremento de la pena es indeterminado* o ambiguo sin más suposiciones y las suposiciones que se haga sobre la posición frente al riego son la clave del efecto de una mayor severidad sobre el crimen. Este es incierto para los amantes al riego, mientras para los adversos al riesgo, un incremento en la severidad de la pena reduce el delito.

De manera que los medios sociales:

1. si somos capaces de analizar masiva y automáticamente mensajes y detectar aquellos que puedan constituir delito de odio;
2. facilitarán enormemente la identificación de los infractores de las leyes y la obtención de pruebas.

Por lo tanto, como corolario (2) aumentará la probabilidad de sanción al infractor y, por consiguiente, disminuirá la frecuencia de este tipo de delitos.

Cabe, incluso, ir más allá y plantearse a futuro la aplicación de la *justicia maquinal* mediante la cual, el infractor es incluso *juzgado* por un sistema informático que utiliza las herramientas desarrolladas en el campo de la inteligencia artificial para determinar su inocencia o culpabilidad, al menos como fase previa a la iniciación de un proceso legal convencional.

Este problema de clasificación de tuits presenta un claro desequilibrio de clases, ya que la constituida por los tuits sospechosos de delito de odio, son muchos menos que los neutros (3). Este tipo de problemas se denominan '*la aguja en el pajar*'.

Fig. 1 La aguja en el pajar

Palabras clave*:* redes sociales, medios sociales, grupos sociales, colectivos, etnia, raza, nacionalidad, religión, orientación sexual, odio, discriminación, violencia, delitos de odio, desequilibrio entre clases.

Abstract.

The so-called "social networks" built-up by platforms such as **Facebook™, Twitter™** which operate on the Internet underpin the *social media* that facilitate quick and mass exchange and discussion of information, experiences and opinions between individuals in a way never before seen in human history.

The range of social media available to the public is varied and growing and can be used for multiple purposes:

- articles in *wikis,*
- opinions about quality in hotels, restaurants (tripadvisor, yelp),
- social and professional contacts (facebook, linkedin),
- web blogs (wordpress).

Like anything new, the explosion of social media has had consequences that have been valued both positively and negatively for society as a whole.

Among the effects generally considered as negative, social media have persistently made 'visible' some attitudes of certain social groups that, to date were present only in a veiled and/or sporadic way. Prominent amongst them are those, which result in attacks on individuals or groups because their affiliation to certain groups defined, by characteristics of nationality, sexual preferences, race, religion...

This phenomenon, coupled with a change of attitude towards certain social behaviors or groups, has led to the emergence of a new criminal category in many countries: the so-called *hate crimes.*

Since Gary Becker's '*Crime and Punishment*' (2) we know that the punishment of criminals is probabilistic. The offender may escape detection or apprehension, or be apprehended but not convicted. His economic theory of crime states that some criminal justice variables are much more effective than others. Increasing *arrest rates*, followed by increasing the likelihood of *being convicted* have the largest impact. On the contrary,

increasing the penalties beyond current levels has an uncertain effect on the crime rate. From this theory, it is clear that public authorities should focus more attention on strategies that increase the risk of arrest and less on strategies that increase the severity of punishment.
So, the social media

1. if we are capable of analyzing automatically mass messages in social media in so that we can detect those that can constitute a hate crime will enormously facilitate the identification of offenders and the collection of evidence.
2. Therefore, as a corollary, both the likelihood of being arrested and that of being convicted will raise for the infringer and, consequentially, the frequency of hate crimes will be reduced.

It is even possible to go further and consider in the future the application of *machine justice* by which the offender is tried by a computer system that uses tools developed in the field of artificial intelligence to determine their innocence or guilt, at least as a phase prior to the start of a conventional legal process.
This classification problem for tweets presents a clear imbalance of classes, since the one formed by messages suspected of hate crime, are much less than its complementary class (3).These types of problems are often called ‘the *needle in a haystack'*.

Keywords*:* Social networks, social media, ethnic groups, race, nationality, religion, sexual orientation, hate crime, discrimination, violence, imbalance of classes.

1 Introducción.

Los *delitos de odio* son un tipo de infracción de la ley cuyo motivo principal es la existencia de *prejuicios* respecto a la víctima del mismo y tienen lugar cuando el perpetrador del delito elige a su víctima en base a su pertenencia a un cierto *grupo*.

Los atributos principales que definen el grupo de pertenencia de la víctima suelen ser el sexo, la etnicidad o raza, la nacionalidad, el idioma, la orientación sexual, la religión, la discapacidad, la apariencia física o la identidad de género, entre otros.

Existen evidencias de que tales delitos de odio están influidos por eventos *singulares* de *amplia difusión* (4) (atentados terroristas, migración incontrolada, manifestaciones, revueltas,…). Este tipo de sucesos suelen actuar como detonadores de manera que la frecuencia de este tipo de ilícitos aumenta espectacularmente tras ellos. Por ello, parece razonable dotar a los responsables de la seguridad pública de herramientas que permitan evaluar la probabilidad de tales delitos y, si es posible, su localización geográfica y temporal.

Los medios sociales de comunicación juegan un importante papel en la comisión de estos delitos en tanto en cuanto las redes se llenan de mensajes de individuos afines a los perpetradores que incitan a castigar al grupo elegido como diana (5) que, recogidos a lo largo de un periodo temporal posterior al incidente detonante, pueden servir para analizar la evolución de la amenaza: escalada, estabilización, duración y descenso.

Tal es la importancia de estos medios que, en muchos países, se han tipificado recientemente[1] también como pertenecientes a la categoría de delito de odio aquellas *manifestaciones públicas* que puedan considerarse una incitación al odio hacia ciertos colectivos.

Uno de los servicios más utilizados para realizar manifestaciones abiertas mediante la publicación de *microblogs* es Twitter™, motivo por el cual este servicio se ha seleccionado como fuente básica de datos para el desarrollo de un *Sistema para la Detección de Indicios de Delitos de Odio*.

Como en todo proyecto relacionado con la *Ciencia de los Datos*, es evidente que, antes de trabajar con datos es preciso capturarlos, lo que se hará mediante la utilización de la API de Twitter.

A continuación, se realiza un análisis exploratorio de datos que servirá de base para la depuración de los mismos, su formateo y modelización.

A partir de los datos depurados, se procederá a su análisis mediante técnicas de *Procesado de Lenguaje Natural* (NLP) para extraer patrones y atributos de los textos para, finalmente, clasificar los mensajes mediante técnicas de *Inteligencia Artificial* (AI) como positivos (que son indicio de una mayor *oferta* de delitos[2]) o negativos/neutros que no aportan pistas al respecto.

El elemento clave para tal clasificación es el *contenido del mensaje* en el que el redactor del mismo[3] –en este caso del *tuit* - manifiesta su sentimiento u opinión respecto a una *entidad* o aspecto de la misma.

El caso que nos ocupa, se enmarca dentro de un grupo de problemas de clasificación binaria - dos clases: contenido de odio o neutro - caracterizado por un ***desequilibrio*** muy

[1] En España en 2015 (1)
[2] Utilizamos la terminología económica de G. Becker (2)
[3] O *fuente de opinión.*

pronunciado entre el número de instancias en cada clase[4], cuya proporción puede alcanzar valores superiores a 1:1000.
Este tipo de situaciones es de importancia en el mundo real en situaciones en que el coste de una clasificación errónea de las instancias de la clase *minoritaria* es muy elevado. Como ejemplos citaremos el diagnóstico de enfermedades o la detección de fraudes en tarjetas de crédito. En el primer caso, unos pocos píxeles del conjunto que constituye una imagen son la base del diagnóstico y en el segundo la proporción de fraudes sobre el total puede ser inferior al 1%. En el primer caso, un falso negativo puede llevar a la muerte del paciente.
Esta situación presenta tres problemas importantes para la clasificación de instancias:

1.1. Los mensajes de odio sobre un colectivo concreto *varían* a lo largo del tiempo *ligados a ciertos eventos* (atentado terrorista→mensajes antiislámicos, casos de corrupción→mensajes antipartidistas, premios Goya→cine español,…) lo que hace que los atributos relevantes para la clasificación sean variables y deban revisarse continuamente.
1.2. Se dificulta la aplicación de *métodos de clasificación supervisada* ya que para el etiquetado manual de unos cientos de casos, se requiere el examen de cientos de miles de tuits lo que alarga y encarece el etiquetado.
1.3. Por otro lado, el desequilibrio entre las clases provoca que el algoritmo que entrenamos sobre un conjunto de con muy pocas instancias de la clase minoritaria sea, con frecuencia, incapaz de generalizar el comportamiento de esta clase y, por tanto, puede tener una escasa capacidad predictiva.

Estos problemas se tratan en este trabajo.

2 Estructura.

La memoria explicativa del proyecto se estructura de la siguiente manera:
Comenzamos revisando algunos conceptos básicos referidos a la clasificación de en conjuntos con desequilibrio y revisaremos el *estado del arte* al respecto así como las herramientas de programación existentes para manejar el proyecto.
Seguidamente, en el apartado *descripción del proyecto* se detallan aspectos fundamentales del mismo tales como

- Especificación del producto,
- Herramientas utilizadas
  Metodología de gestión de proyectos utilizada
  Fuentes de datos,
  Software y hardware,
  Control de versiones

A continuación se expone el diseño del proyecto

3 Problemas de clasificación y desequilibrio de clases.

3.1 Conjuntos de datos.

Lo ideal es recoger *varios conjuntos de datos independientes*, si ello no es posible, debemos conformarnos con un solo conjunto de datos que habremos de dividir en dos o tres conjuntos.

[4] Es decir, diferencias significativas entre las probabilidades *a priori*.

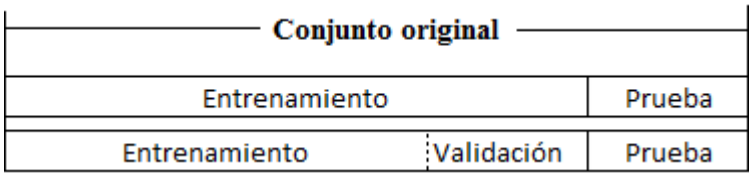

Fig. 2: División del Conjunto de Datos

La estrategia consistente en usar

1. un *conjunto de entrenamiento* para aprender y estimar los parámetros del modelo;
2. un *conjunto de validación* para evaluar modelos y seleccionar uno de ellos y
3. un *conjunto de prueba* o test para valorar la capacidad de predicción de los modelos.

Existen múltiples métodos de división de datos, de los cuales el más simple es el *método de retención* (holdout) que consiste en dividir aleatoriamente el conjunto original en dos subconjuntos ($^2/_3$ o $^1/_2$ para entrenamiento y el resto para prueba). Si el conjunto original no es lo bastante grande, el método es ineficiente.
En *aprendizaje estadístico* se utiliza como supuesto básico que *tanto el conjunto de entrenamiento como el de prueba se extraen de una misma distribución subyacente* constituida por la combinación de las distribuciones de la clase mayoritaria y minoritaria. Si se llevan a cabo modificaciones en el conjunto de entrenamiento para tratar de reequilibrar las clases, el *conjunto de entrenamiento y el de prueba tendrán distribuciones diferentes* violando este supuesto.

3.2 Medida del rendimiento de un clasificador.

La *matriz de confusión*, elemento básico de evaluación de clasificadores, se expresa en el caso de que solo existan dos clases como:

		Predicción	
		Positivo	***Negativo***
Real	***Positivo***	Positivo Verdadero (**TP**)	Falso Negativo (**FN**)
	Negativo	Falso Positivo (**FP**)	Negativo Verdadero (**TN**)

Tabla 1: Matriz de Confusión

Las medidas más inmediatas de evaluación del modelo que se nos ocurren son:

- *Exactitud*: $\boldsymbol{Acc} = \frac{TP+TN}{TP+TN+FP+FN}$
- *Tasa de error*: $\boldsymbol{Err} = \frac{FP+FN}{TP+TN+FP+FN} = 1 - \boldsymbol{Acc}$

Sin embargo, cuando *existe un claro desequilibrio entre clases*, como es el caso de los tuits de odio que pueden ser un 2 ‰ del total, es posible obtener una exactitud enorme aun clasificando todos los tuits (erróneamente) como negativos (no de odio), y equivocándonos, por tanto, en todos los positivos:

		Predicción		
		odio	***Neutro***	
Real	***odio***	0	200	200
	Neutro	200	99.800	100.000
		200	100.000	**100.000**

Tabla 2: Matriz de Confusión con Desequilibrio de Clases

Lo que nos daría unos valores de exactitud del 99,8%:

Acc	99,8%
Err	0,2%

Por ello, son más útiles los indicadores:

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

- *Precisión* (p) es el porcentaje de los *tuits* realmente pertenecientes a una clase que se asignan a la misma (aciertos) sobre el total de los asignados a dicha clase por el clasificador:

$$\boldsymbol{p} = \frac{TP}{TP + FP}$$

  Es decir, el porcentaje de predicciones que se acierta.
- *Exhaustividad* (r) es el porcentaje de los *tuits* que han sido clasificados como pertenecientes una clase sobre el total de miembros de dicha clase (porcentaje de instancias de la clase bien clasificadas):

$$\boldsymbol{r} = \frac{TP}{TP + FN}$$

Se utilizan también combinaciones de p y r, tales como la media geométrica:

$$\boldsymbol{G} = \sqrt{p \cdot r}$$

Y la armónica:

$$\boldsymbol{F} = \frac{2}{\frac{1}{p} + \frac{1}{r}} = \frac{2pr}{p + r}$$

Un valor p=1 nos dice que todos los elementos recuperados como relevantes, lo son, pero no nos dice nada acerca de si hemos recuperado todos los documentos relevantes (r).

Por último, mencionaremos otro indicador que se está utilizando cada vez más: el ***ROC*** (6) habitual en Medicina y Biología para hablar de la detección de falsos positivos y negativos.

Ahora a la exhaustividad - $\frac{TP}{TP+FN}$ - se la denomina ***sensibilidad***. Como se ve, es la

$$\Pr(predicción\_\boldsymbol{TRUE}|\boldsymbol{TRUE})$$

y se introduce la ***especificidad***:

$$\frac{TN}{TN + FP} = \frac{d}{(c + d)}$$

De manera que

$$1 - especificidad = \frac{FP}{TN + FP} = \frac{c}{(c + d)}$$

Y es

$$\Pr(predicción\_\boldsymbol{TRUE}|\boldsymbol{FALSE})$$

Por el teorema de la probabilidad total sabemos que:

$$\Pr(predicción\_T|T)\,Pr(T) + \Pr(predicción\_T|F)Pr(F) = \mathbf{Pr(predicción\_T)} = \\ = \boldsymbol{sensibilidad} \cdot Pr(T) + (\mathbf{1} - \boldsymbol{especificidad}) \cdot Pr(F)$$

Si dibujamos el gráfico que relaciona ambas magnitudes, obtenemos la Fig. 3**¡Error! No se encuentra el origen de la referencia.** en que el *ROC* es el *área bajo la curva* que puede tomar valores entre 0 (no acierta nunca) y 1 (la predicción acierta siempre).

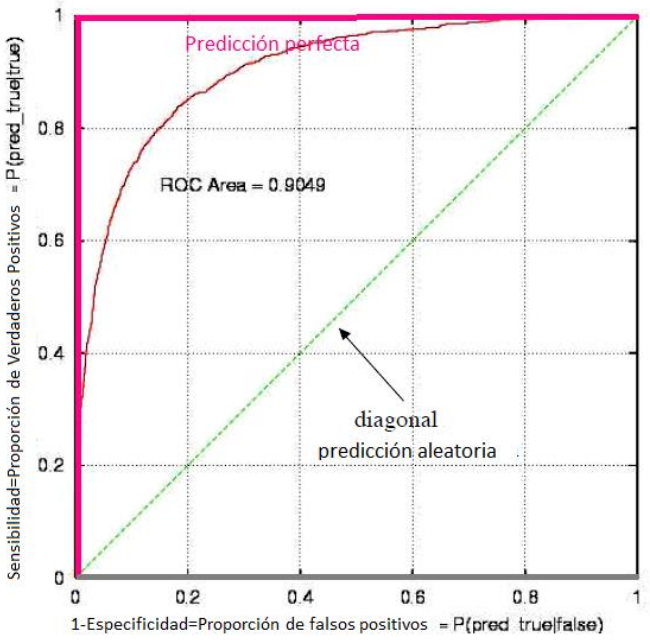

Fig. 3: ROC

3.3 El ciclo de Atención a un Tema y el vocabulario.

Como se ha mencionado anteriormente, los delitos de odio tienden a ser más frecuentes y a crecer en periodos de tiempo posteriores a un suceso antecedente (‘detonador’) (4) y el interés del público sobre un asunto determinado sigue el llamado ciclo de Atención a un Tema que fue descrito inicialmente por Downs (7) y que se muestra en la Fig. 4.

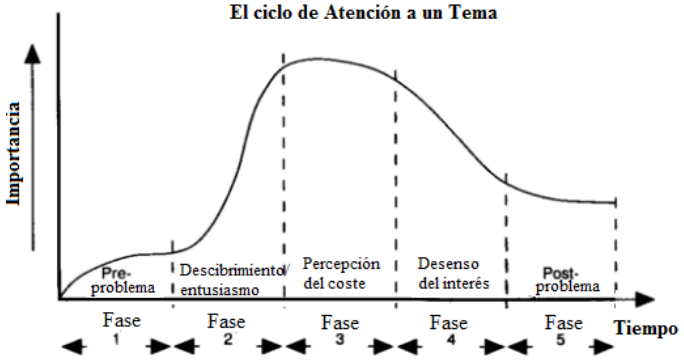

Fig. 4: Ciclo de Atención a un Tema

Pueden verse las diferentes etapas por las que pasa la relevancia del tema para el público a lo largo de un periodo de tiempo que, además no suele ser muy largo.
Este punto es relevante por cuanto afecta tanto a la recolección de tuits como a los términos - atributos – a utilizar que se encuentran relacionados.
Sobre el primer punto, si se desea el seguimiento de un tema concreto una vez ocurrido un suceso (p.e. aparición de un nuevo caso de corrupción), hemos de recoger los tuits en la cresta de la ola y, además, hemos de tener en cuenta que, dependiendo de los asuntos que sean más *trendy* el vocabulario que se usa para su comentario es diferente.

3.4 Desequilibrio de clases, etiquetado y clasificación.

En aprendizaje supervisado, tenemos un conjunto - muestra - de datos $\mathcal{S}$ sobre el cual queremos construir un modelo de clasificación binario. La primera suposición que haremos es que las instancias del conjunto positivas y negativas observadas ($\mathcal{S}^+$ y $\mathcal{S}^-$) se extraen de dos distribuciones diferentes $\mathcal{P}$ y $\mathcal{G}$. Las instancias positivas son las minoritarias. En estas circunstancias es fácil ver que, por una parte,

1. el etiquetado manual de las instancias resulta engorroso puesto que para encontrar un tuit – la *aguja* - de la clase minoritaria se precisa examinar una cantidad ingente de tuits de la otra clase – la *paja*. y, por otra, se explica
2. el motivo por el cual un modelo clasificador sobre el conjunto $S = S^+ \cup S^-$ produce una baja exhaustividad: la distribución positiva está subrepresentada y los valores *atípicos de* $\mathcal{G}$- de mucha mayor cardinalidad - aunque sean una pequeña fracción, influirán en el clasificador ya que se considerarán por este como pertenecientes a $\mathcal{S}^-$ y el clasificador inducido estará sesgado hacia la clase minoritaria, es decir más cercano a los puntos de esta de lo que debiera, produciendo un rendimiento bajo del clasificador (Fig. 5 (A)).

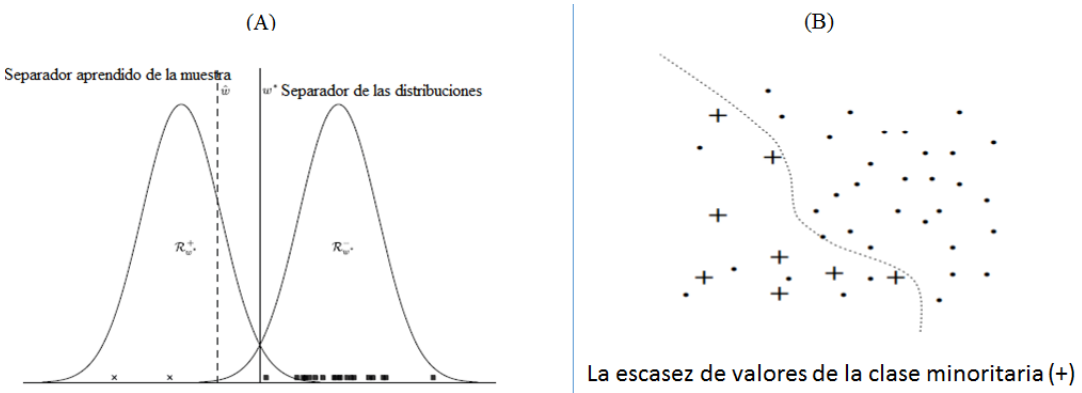

Fig. 5: Errores de Clasificación en Muestras Desequilibradas

En la **¡Error! No se encuentra el origen de la referencia.** (B) se muestra otro ejemplo cuando se usa 1-NN como clasificador en un conjunto desequilibrado. Lo mismo puede decirse de la clasificación bayesiana.
Debido a la dificultad de etiquetado manual, hemos utilizado un procedimiento que *filtra* el conjunto inicial extraído de Twitter mediante el uso de un vocabulario con términos de odio obtenidos de diferentes fuentes y el conjunto filtrado, de una cardinalidad mucho menor, se *etiqueta* y usa como conjunto de *entrenamiento*.

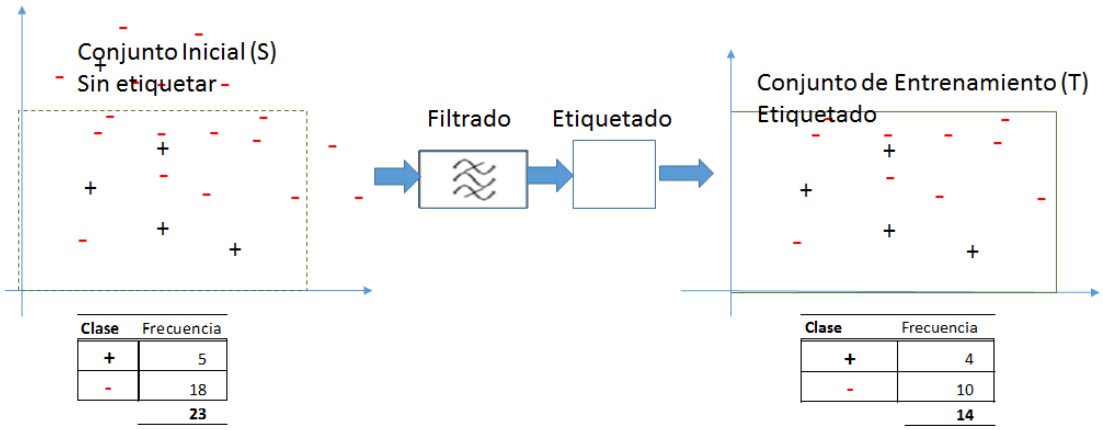

Fig. 6: Filtrado del Conjunto Inicial

Seguidamente, el conjunto de entrenamiento se usa para estimar los parámetros del clasificador y, son este, se clasifican los tuits del conjunto de entrenamiento.
Para la utilización del clasificador a otros conjuntos de tuits, caben dos opciones:

1. Aplicarlo tras filtrar el nuevo conjunto, asignando a todos los tuits que no pasan el filtro a *C-*.
2. Aplicarlo al conjunto completo con lo cual cabe la posibilidad de que algunos de los tuits que no pasan el filtro, se clasifiquen en *C+* (correcta o incorrectamente).

En el primer caso (Fig. 7), la *matriz de confusión* muestra como FN a las instancias que no pasan el filtro que debieran estar etiquetadas con (+), una en nuestro caso.

		Predicción				
		+	-			
Real	+	3	2	***5***	**Precisión (p)**	75,0%
	-	1	17	***18***	**Exhaustividad (r)**	60,0%
		4	***19***	**23**	**F**	66,7%

Tabla 3: Matriz de Confusión del Conjunto Inicial

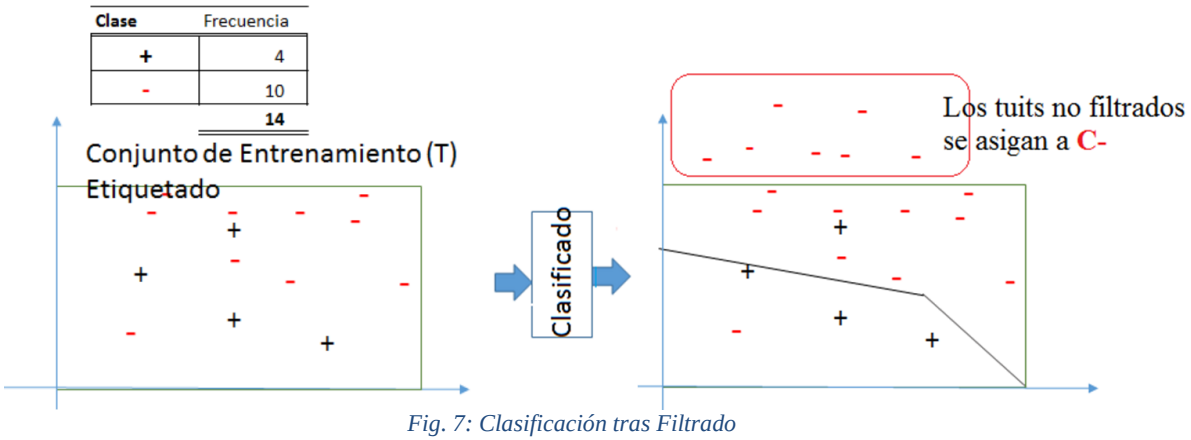

Fig. 7: Clasificación tras Filtrado

El segundo caso, como veremos más adelante (4.2), requiere un recalibrado de las probabilidades del modelo entrenado.

4 Estado del arte

Se trata de un trabajo que persigue la *clasificación* de una colección de tuits en las categorías:

- De odio y

- Neutra

Mediante procedimientos de clasificación *supervisada* de Aprendizaje Máquina, para lo cual debe procederse a un etiquetado previo de un conjunto de entrenamiento.

El proyecto se mueve, por una parte en el terreno problemático de los conjuntos no equilibrados, tema sobre el que existe una abundante literatura que se refiere fundamentalmente a:

- cómo simplificar el etiquetado de las instancias y
- cómo paliar los efectos de la asimetría entre clases.

Por otro lado, el proyecto de clasificación de tuits utiliza herramientas del campo del *Procesamiento del Lenguaje Natural* (PNL), rama de la Inteligencia Artificial que tiene sus orígenes en los años 50 del pasado siglo cuando aparecieron las primeras computadoras. Su desarrollo ha venido acompasado a los avances en la capacidad de cómputo (*Ley de Moore*) y en los algoritmos de *aprendizaje estadístico*.

Por último, conviene pasar revista a las *herramientas informáticas* que facilitan el tratamiento de este tipo de problemas.

4.1 Selección de instancias.

Como hemos dicho, el *corpus* de tuits es un conjunto en el que la clase de tuits de odio es muy minoritaria respecto a la de tuits neutros. Este problema - junto con el de selección de instancias y desequilibrio de clases – ha sido estudiado desde muchos puntos de vista y Selecting *Representative Data Sets* (7) proporciona un resumen de los procedimientos desarrollados para seleccionar conjuntos de entrenamiento *equilibrado* en los casos en que existen grandes desequilibrios entre las clases de clasificación que básicamente son de tres tipos:

1. A nivel de datos
2. A nivel de algoritmo y
3. Conjuntos

Los primeros se basan en el sobremuestreo o submuestreo de la clase minoritaria o mayoritaria. Los segundos en ponderar de modo diferente la importancia de la instancia en función de la clase a que pertenezca y el tercero utiliza una combinación de métodos.

La fase inicial de *selección de instancias* (8) es un proceso de reducir el conjunto de datos original.

La salida ideal de la selección de instancias es una muestra mínima independiente del modelo que pueda cumplir su objetivo con el menor deterioro posible, es decir, que el rendimiento P de un modelo M sea aproximadamente el mismo sobre la muestra S que sobre la población W.

$$P(M_s) \approx P(M_w) \quad \text{(Ecuación 1)}$$

El método clásico de obtención de muestras se basa en las técnicas de muestreo (aleatorio simple, estratificado, adaptativo,…)

Con frecuencia, la muestra puede reducirse para generar un *conjunto de entrenamiento* más manejable (9).

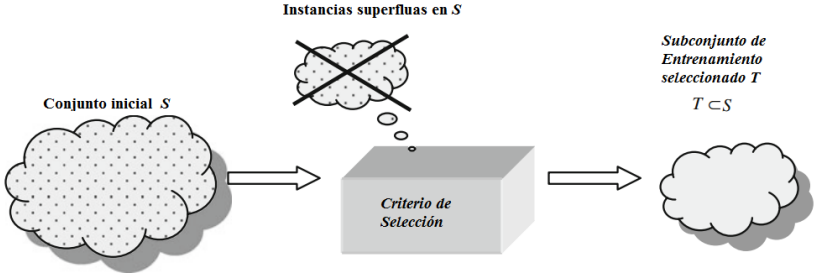

Fig. 8: Selección de Instancias

Esto puede hacerse mediante procedimientos de *selección* utilizando algún tipo de algoritmo de selección de instancias bien relacionado con

1. el rendimiento de algún algoritmo de clasificación (*wrapper methods*) o con
2. el vector de atributos de la instancia con independencia del algoritmo utilizado (*filter methods*)

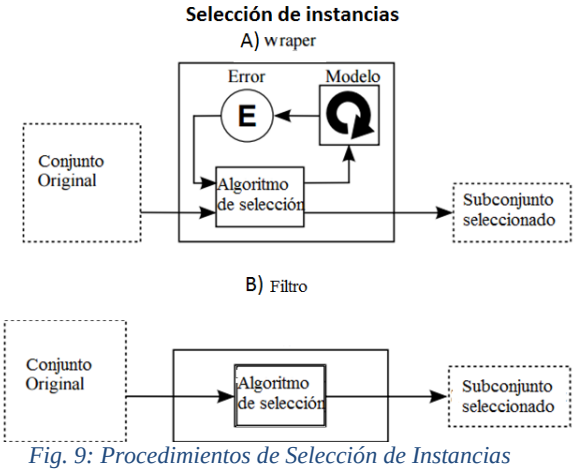

Fig. 9: Procedimientos de Selección de Instancias

Los enfoques del primer tipo (*wrapper*) subrayan el aspecto de minería de datos del problema ejecutando un algoritmo específico de tal campo para disparar la selección de instancias, p.e. seleccionando un subconjunto inicial, ejecutando un algoritmo sobre este subconjunto inicial, evaluando sus resultados y ampliando incrementalmente el subconjunto inicial hasta que los resultados del algoritmo sean lo bastante buenos.

Los enfoques del tipo filtro, son más simples e independientes del algoritmo de clasificación.

4.2 Clases con probabilidades *a priori* no equilibradas.

Hemos mencionado en 3.3 tanto la dificultad de etiquetar conjuntos con gran desequilibrio de clases como la de entrenar algoritmos en conjuntos de entrenamiento ya que ello conlleva (ver 3.3) un elevado *nivel de ruido* y un pobre rendimiento del clasificador.

P.e. si estamos realizando un experimento para el diagnóstico de alguna enfermedad cuya tasa de prevalencia del 1 ‰ no elegiremos un conjunto de entrenamiento de 1 enfermo y 999 individuos sanos, sino un conjunto 50%-50%, donde los sanos constituyen el grupo de control y los enfermos el experimental, lo que modifica las probabilidades a priori de cada clase en el conjunto de entrenamiento ye introduce sesgo en el clasificador.

En (7) se hace referencia a los siguientes grupos de métodos para vencer este problema:

1. Métodos de *nivel de datos*. Se utilizan en el preprocesado y se basan en varios tipos de *remuestreo*. Buscan aumentar el número de instancias de la clase minoritaria (sobremuestreo) y/o reducir los de la clase mayoritaria (submuestreo).
2. Métodos de *nivel de algoritmo*. Se basan sobre todo en dar una sobreponderación a la clase mayoritaria.
3. Métodos *conjuntos* que usan una combinación de métodos.

Todos ellos generan conjuntos de entrenamiento equilibrados sobre los que se entrenan algoritmos.

Resulta evidente que al reequilibrar artificialmente el conjunto de entrenamiento, las distribuciones en el conjunto de entrenamiento y en el de prueba son diferentes de manera que se viola una de las hipótesis básicas de aprendizaje estadístico: que *tanto el conjunto de entrenamiento como el de prueba siguen la misma distribución*. Esto motiva que la aplicación de un modelo entrenado sobre un conjunto de entrenamiento reequilibrado por el analista a un conjunto donde esto no se ha llevado a cabo (prueba) exige calibrar las probabilidades obtenidas del conjunto de entrenamiento retocado para poder aplicar el modelo recalibrado directamente al conjunto de test (no modificado).
En el artículo (10) se analiza este problema y se desarrolla un método para corregir el sesgo$P(M_s) \approx P(M_w)$ (Ecuación 1 introducido.

4.3 Etiquetado de instancias.

Un serio problema que afecta a los conjuntos desequilibrados, es el del *etiquetado del conjunto* de entrenamiento (3.3). Si bien resulta relativamente fácil recopilar cientos de miles de *tuits no etiquetados*, no es tan fácil proceder a su clasificación manual, especialmente cuando, como en este caso, la clase de interés es muy minoritaria.
Se han ensayado algunos métodos para vencer esta dificultad como el descrito en (11) que, a pesar de ser un método *no supervisado* está relacionado con la selección de instancias y ha servido de guía a nuestro proyecto y cuyo esquema es el siguiente (Fig. 10):

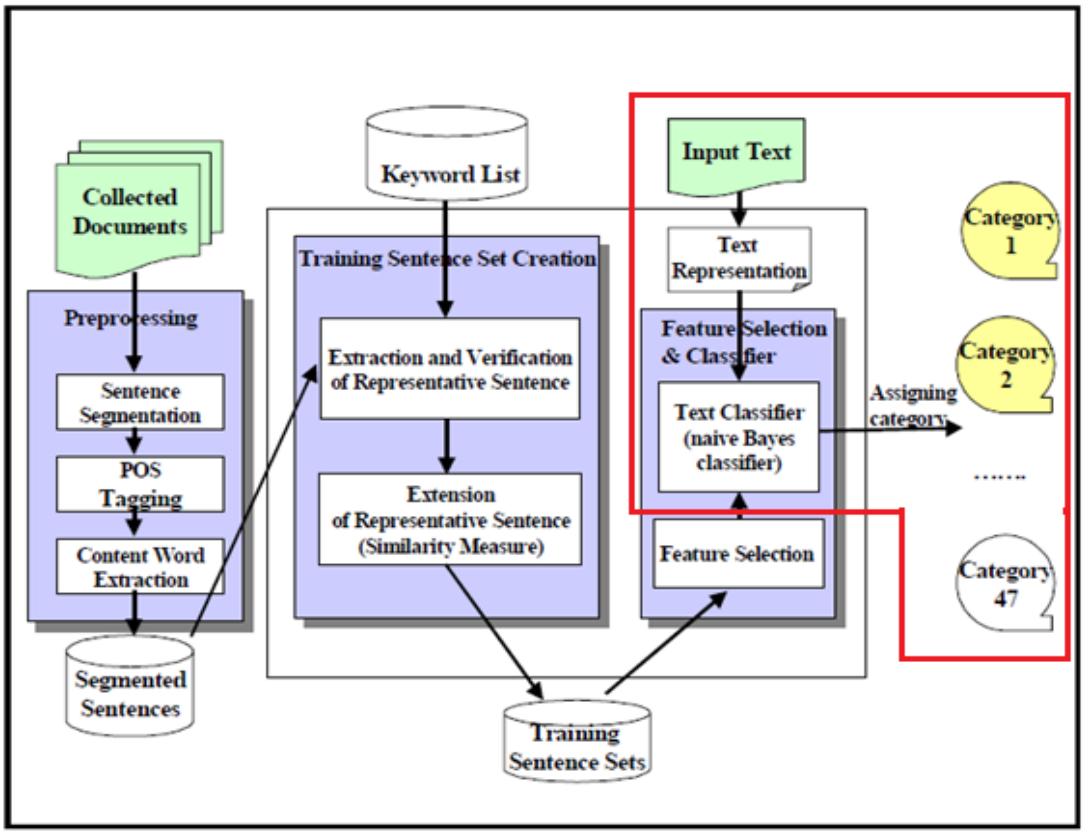

Fig. 10: Método de Etiquetado no Supervisado

4.4 Selección de atributos.

Una de las etapas más importantes en clasificación es la *selección de atributos*.

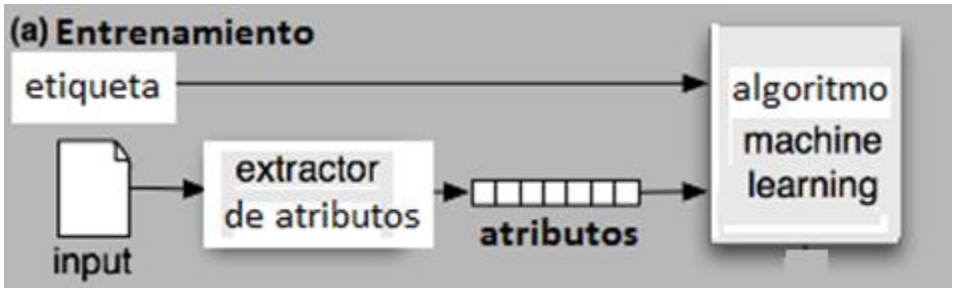

Fig. 11: Selección de Atributos

Yan y Pedersen (12) hacen un estudio comparativo de diferentes procedimientos para la selección de atributos en clasificación de textos comenzando por el más simple: umbral de frecuencia documental en el que se eliminan aquellos términos (*atributos*) que aparecen muy raras veces en los documentos de una clase. Además, destaca por su simplicidad el método de la *Chi cuadrado* ($\chi 2$) que usa tablas de contingencia que muestran la frecuencia con que aparece cada atributo en cada clase etiquetada y contrasta la hipótesis de independencia entre *término* y *clase* utilizando la f.de D. $\chi 2$.

4.5 Herramientas informáticas disponibles.

Para la ejecución del proyecto, nos interesan las herramientas para minería de datos disponibles para el analista. Tales herramientas se refieren tanto al campo del *PNL* propiamente dicho como al de los *algoritmos* de clasificación y recuperación de información que son necesarios para clasificar los tuits.

A continuación mencionamos algunas de las plataformas existentes que se encuentran en permanente evolución.

4.5.1 Plataformas de código abierto.

- Basadas en Python:
  - Natural Language Tool Kit (NLTK) (13)
  - gensim (14)
  - Scikit-learn (15)
  - TensorFlow (16)
- Basadas en Java:
  - CoreNLP (17)
  - MALLET (18) programas para modelizado temático y clasificación de textos.
  - LingPipe (19)
  - Weka (20)
  - yTextMiner (21) : plataforma desarrollada en la universidad de Yonsei, Corea que integra parte de los modelos anteriores con otras librerías tanto para modelización temática como clasificación de textos y análisis de sentimiento.
  - Apache Lucene (22)

4.5.2 Plataformas comerciales.

Probablemente la más interesante es la que forman los diferentes paquetes del Sistema *SAS™*[5], más concretamente:

- *SAS™ Enterprise Miner*: paquete de aprendizaje automático y minería de datos enfocado a *business analytics*

[5] Existe una versión - *SAS university edition* - que puede utilizarse temporalmente de manera libre (https://www.sas.com/en_us/software/university-edition.html) pero no incluye las herramientas de minería de datos.

- *SAS™ Text Miner*: utiliza la técnica SVD (la misma que gensim) para simplificar y acelerar la clasificación de textos y extracción de información.

5 Descripción del proyecto.

5.1 Especificaciones.

El objetivo del proyecto es seleccionar y desarrollar un método para clasificar un conjunto de tuits ya recogidos de Twitter en dos categorías:

- Próximos a mensaje de odio,
- Neutros.

5.2 Etapas.

Comenzaremos por seleccionar los campos que nos interesan de los tuits que son:

5.2.1 Identidad del tuit,

5.2.2 Texto del *tuit* ('documento')

Los documentos han de ser depurados

6 Diseño.

7 Desarrollo.

8 Pruebas.

9 Resultados.

10 Utilización.

Los programas y documentación del proyecto se encuentran contenidos en [repositorio GitHub]. Tal repositorio ha sido dividido en dos partes:

1. La primera contiene los programas utilizados en la fase de ensayo y selección de modelos y
2. Una segunda en la que se contienen los programas operativos, es decir aquellos que son de aplicación directa a los tuits y que permiten distinguir aquellos microblogs que pueden considerarse de odio hacia algún grupo o persona.

En este segundo directorios existe un Wiki que contiene las instrucciones necesarias para descargar e instalar los programas y correrlos para clasificar tuits en las categorías mencionadas.

Bibliografía

1. BOE. [En línea] 30 de Marzo de 2015. https://www.boe.es/boe/dias/2015/03/31/pdfs/BOE-A-2015-3439.pdf.
2. *Crime and Punishment: An Economic Approach.* Becker, Gary S. 2, s.l. : The University of Chicago Press, Marzo-Abril de 1968, Journal of Political Economy, Vol. 76, págs. 169-217.
3. *The class imbalance problem in pattern classification and learning.* V. García, J. S. Sánchez, R.A. Mollineda, R. Alejo, J.M. Sotoca. 2007.
4. *Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making.* Pete Burnap , Matthew L. Williams. 2, 2015, Policy & Internet, Vol. 7, págs. 223-242,.
5. *High Times for Hate Crime: Explaining the Temporal Clustering.* King, R.D., G.M. Sutton. 4, 2013, Criminology , Vol. 51, págs. 871–94.
6. *An introduction to ROC analysis.* Fawcett, Tom. s.l. : Elsevier, 2005.
7. *Up and Down with Ecology - The Issue Attention Cycle'.* Downs, A. 1972, Public Interest (28), págs. 28-50.
8. Tomas Borovicka, Marcel Jirina Jr., Pavel Kordik y Marcel Jirina. Selecting Representative Data Sets. [En línea] 2012. http://dx.dio.org/10.5772/50787.
9. *On Issues of Instance Selection.* Motoda, Huan Liu y Hiroshi. 2002, Data Mining and Knowledge Discovery 6(2), págs. 115-130.
10. *A review of instance selection methods.* al., J. Arturo Olvera-López et. 2010, Artif Intell Rev (2010) 34, págs. 133–143.
11. *Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure.* Marco Saerens, Patrice Latinne, Christine Decaestecker. 2002, Neural computation 14(1), págs. 21-41.
12. *Automatic Text Categorization by Unsupervised Learning.* Seo, Youngjoong Ko y Jungyun. 1997.
13. *A Comparative Study on Feature Selection in Text Categorization.* Yiming Yan y Jan O. Pedersen. San Francisco : s.n., 1997. ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning.
14. *An Extensible Toolkit for Computational Semantics.* Dan Garrette, Ewan Klein. Tilburg University, Netherlands : s.n., 2009. Proceedings of the Eighth International Conference on Computational Semantics.
15. *Software Framework for Topic Modelling with Large Corpora.* Řehůřek, Radim and Petr Sojka. Valetta, Malta : s.n., 2010. Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. págs. 46--50.
16. *Scikit-learn: Machine Learning in Python.* al, Gael Varoquaux et. 2011, Journal of Machine Learning Research, págs. 2825-2830.
17. Google Research. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. [En línea] 2015. http://download.tensorflow.org/paper/whitepaper2015.pdf.
18. al., Christopher D. Manning et. The Stanford CoreNLP Natural Language Processing Toolkit. [En línea] http://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf.
19. McCallum, Andrew. MALLET: A Machine Learning for Language Toolkit. [En línea] https://people.cs.umass.edu/~mccallum/mallet/.
20. *Bob Carpenter.* Carpenter, Bob. Valencia,Spain : s.n., 2007. Proceedings of the 2nd BioCreative workshop.
21. *Weka: Practical Machine Learning Tools and Techniques.* al., Ian H. Witten et al. 2007. Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems. págs. 192-196.

22. yTextMiner. [En línea] http://informatics.yonsei.ac.kr/yTextMiner.
23. https://lucene.apache.org/core/. [En línea]
24. [En línea] https://lucene.apache.org/core/.

Glosario

A

AI. *Véase* Inteligencia Artificial.
análisis exploratorio de datos
Proceso al que se someten los datos antes de su modelización. Suele consistir en resumir sus características principales, con frecuencia usando métodos gráficos. Su objetivo es permitir una planificación más adecuada tanto del proceso de recolección como de su tratamiento posterior., 6
aprendizaje estadístico
Rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. Se trata de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos., 13
Aprendizaje Máquina. *Véase* Aprendizaje Estadístico

B

business analytics
Conjunto de habilidades, técnicas y prácticas para la exploración iterativa del desempeño pasado de una empresa a fin de obtener una mejor comprensión de su funcionamiento y su desarrollo futuro. Utiliza tanto la minería de datos como la inteligencia artificial., 17

C

Ciencia de los Datos
Campo interdisciplinario que comprende los procesos y sistemas para extraer conocimiento ,,de grandes volúmenes de datos en sus diferentes formas (estructurados o no estructurados) y formatos (.txt, .dat, .doc, .jpg, etcétera, 6
clasificación de textos
Tarea consistente en asignar un documento a una categoría determinada., 17
clasificación supervisada
Este tipo de clasificación cuenta con un conocimiento *a priori*, es decir para la tarea de clasificar una instancia dentro de una categoría contamos con modelos ya clasificados (instancias agrupadas que tienen características comunes), 7
Se parte de un conjunto de clases conocido a priori. Estas clases deben caracterizarse en función del conjunto de variables mediante la medición de las mismas en individuos cuya pertenencia a una de las clases no presente dudas., 13
conjunto de entrenamiento
Es el histórico de datos que se usa para entrenar al sistema que detecta los patrones. El conjunto de datos se compone de instancias, y las instancias de factores, características o propiedades, 8
conjunto de prueba
Es el usado para evaluar la bondad d las predicciones del modelo., 8
conjunto de validación
Es el usado para evaluar y seleccionar modelos *entrenados* sobre el conjunto de entrenamiento, 8
corpus
En PNL, colección de documentos., 13

D

delitos de odio
Aquellos motivados por prejuicios respecto a la víctima del mismo y tienen lugar cuando el perpetrador del delito elige a la víctima en base a su pertenencia a un cierto grupo., 6

E

entidad
Producto, persona, evento, organización o tópico., 6
estado del arte
Una de las primeras etapas dentro de un proyecto es la construcción de su estado del arte, ya que permite determinar la forma como ha sido tratado el tema, cómo se encuentra el avance de su conocimiento en el momento de realizar una investigación y cuáles son las tendencias existentes, en ese momento, para el desarrollo de proyectos en el mismo campo., 7
etiquetado
Asignar una etiqueta (clase) a cualquier dato del conjuntoada., 15
etnicidad
etnicidad
Una etnia es un conjunto de personas que tienen en común rasgos culturales, idioma, religión, vestimenta, nexos históricos, tipo de alimentación, y, muchas veces, un territorio. Dichas comunidades, a veces, reclaman para sí una estructura política y el dominio de un territorio, 6
Exactitud

F

G

I

L

M

N

O

P

S

T