

**UNIVERSIDAD AUTONOMA DE MADRID**

**ESCUELA POLITECNICA SUPERIOR**



**Grado en Ingeniería Informática**

**TRABAJO FIN DE GRADO**

**Proyecto: *Análisis de Sentimiento en Tuits para la  
Detección de Delitos de Odio***

**Informe Propuesta**

**Juan Carlos Pereira Kohatsu**

**Enero 2017**

## Índice de contenidos.

Objetivo. ....	3
Descripción y nombre del proyecto. ....	3
Características del proyecto. ....	3
Métricas de rendimiento. ....	5
Propuesta de actuación. ....	7
Apéndice: TASS 2016 (8) ....	9
Referencias. ....	10
Glosario ....	11

## Índice de Figuras.

FIGURA 1 VALIDACIÓN Y SELECCIÓN DE MODELOS ....	4
FIGURA 2 VALIDACIÓN Y SELECCIÓN DE MODELOS ....	6
FIGURA 3 RELACIÓN PRECISIÓN-EXHAUSTIVIDAD ....	6
FIGURA 4 INDICADOR ROC ....	7

## Objetivo.

Este primer informe tiene como objetivo presentar el avance realizado en el proyecto y presentar una definición clara del proyecto y una metodología de trabajo coherentes ambas tanto con los objetivos del proyecto como con los recursos – datos, humanos e informáticos – disponibles así como la identificación de fuentes bibliográficas y de otro tipo que puedan ser de utilidad para el proyecto.

## Descripción y nombre del proyecto.

El trabajo se enfoca a la identificación de *tuits* que, por su contenido, pueden contener mensajes que denigren o muestren hostilidad hacia determinados grupos sociales o inciten a la violencia contra ellos.

Esencialmente, el proyecto consistirá en desarrollar un sistema de clasificación de *tuits* en las siguientes clases<sup>1</sup> que difieren ligeramente de las utilizadas por el ministerio del Interior:

- a. Clase 1: *mensaje de odio*.
  - i. Clase 1A: Racismo/xenofobia.
  - ii. Clase 1B: Sexo/género.
  - iii. Clase 1D: Identidad/orientación sexual.
  - iv. Clase 1F: Discapacidad.
  - v. Clase 1G: Clase social.
- b. Clase 2: *neutro*.

donde hemos eliminado antisemitismo - incluido en *religión y/o xenofobia* - y *aporofobia* - que se incluye en *clase social* -, por su escasa entidad.

Por lo tanto, un mismo mensaje puede pertenecer a dos clases distintas (p.e. *moro* se refiere tanto a la clase *racismo/xenofobia* como a *religión*).

El nombre elegido para el proyecto es: *Análisis de Sentimiento en Tuits<sup>2</sup> para la Detección de Delitos de Odio*.

Para ello se pretende seleccionar un procedimiento de aprendizaje maquina supervisado (naïve Bayes, vecinos próximos, árboles de clasificación,...) que detecte aquellos *tuits* que contengan mensajes denigratorios o amenazadores para los grupos<sup>3</sup> definidos en las clases.

## Características del proyecto.

Partimos del supuesto de que el problema de capturar *tuits* ha sido solventado por cualquier procedimiento que use cualquiera de las APIs de Twitter (REST<sup>4</sup> o Streaming<sup>5</sup>) y disponemos de un conjunto de *tuits* de tamaño adecuado<sup>6</sup>.

---

<sup>1</sup> El Ministerio del Interior (9) clasifica los delitos de odio en 8 categorías: Antisemitismo, aporofobia, religión, discapacidad, identidad u orientación sexual, racismo/xenofobia, ideología y discriminación por sexo/género.

<sup>2</sup> Usamos *tuit* por ser el término admitido en la RAE para los mensajes de Twitter® (10).

<sup>3</sup> Excluimos los casos de personas que pueden ser consideradas símbolo de algún grupo (p.e. Irene Villa de víctimas del terrorismo, Carrero Blanco de la ideología franquista, etc.).

<sup>4</sup> La *REST API* consiste básicamente en una consulta (QUERY) a su servidor que devuelve una respuesta en JSON, XML, etc.

<sup>5</sup> La *Streaming API*, al contrario que la REST, es una *Query* que pervive por largo tiempo sobre una conexión HTTP que se mantiene abierta y va entregando los datos cuando estos están disponibles.

<sup>6</sup> Una referencia útil a la minería de redes sociales – Twitter, facebook, LinkedIn ...- es (17) y puede encontrarse como IPython notebooks en github (18)

Por tanto, el proyecto se enfoca esencialmente a desarrollar un método que permita clasificar un tuit en una de dos categorías: mensaje de odio o neutral, con subcategorías en el primer grupo.

Como es sabido, los procesos de validación y selección de modelos<sup>7</sup> en aprendizaje estadístico (1) se facilitan enormemente si se dispone de un conjunto de datos que están etiquetados con la salida correcta (*'patrón oro'*) lo que, en muchos casos, requiere la intervención humana. Este proceso se indica en la Figura 1.

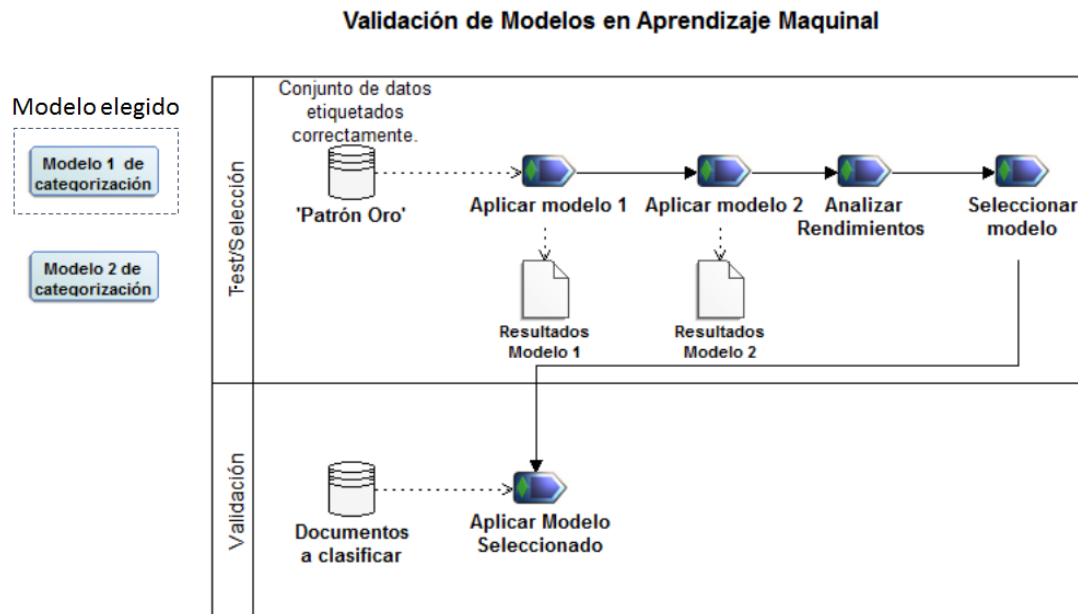


Figura 1 Validación y selección de modelos

Según las primeras observaciones, la proporción de tuits que tratan temas relacionados con el odio es habitualmente muy limitada<sup>8</sup> por lo que un etiquetado manual con selección aleatoria de tuits requeriría revisar una enorme cantidad de mensajes para obtener a cambio un pequeño conjunto de entrenamiento.

Por esta razón, se han explorado varios caminos que permitan, un etiquetado de *tuits no supervisado* o semisupervisado. Estos métodos suelen basarse en un cribado de los tuits basado en reglas que no requiere la intervención humana, siendo un ejemplo el caso expuesto en el artículo (2).

Tal enfoque requiere del uso de herramientas de procesamiento del lenguaje natural (NLP) tales como el paquete Natural Language Toolkit (3) (NLTK) descrito en el libro de Steven Bird, Ewan Klein, and Edward Loper (4).

No obstante, la utilización de las herramientas de procesamiento del lenguaje natural requiere la utilización de *corpora* anotados que permitan realizar análisis, morfológicos, sintácticos, semánticos o de otro tipo para finalidades tan distintas como puedan ser conversión de voz a texto, análisis de opinión, valoración de productos por los usuarios, traducción automática etc.

Desgraciadamente en castellano este tipo de corpora no abundan, de momento hemos identificado los siguientes:

<sup>7</sup> Son dos procesos separados: en la *selección* de modelos se estima el rendimiento de los diferentes modelos para elegir el mejor. En *validación* se trata de estimar la bondad de un modelo ya elegido.

<sup>8</sup> La frecuencia de aparición de estos mensajes es muy variable y suele dispararse, bien tras producirse cierto tipo de eventos como manifestaciones, atentados, etc. (11) o bien previamente en campañas para preparar agresiones planificadas a ciertos grupos (13).

- Wikicorpus (5) (6)
- Cess\_esp contiene 6.030 sentencias anotadas.
- Corpus TASS<sup>9</sup>: corpus de unos 70.000 tuits, escritos en español por más de 150 personajes de la política, economía, medios de comunicación y el mundo de la cultura en España, entre noviembre de 2011 y marzo de 2012. Cada mensaje incluye su identificador de tuit, la fecha de creación, el usuario y el propio contenido. Cada mensaje ha sido etiquetado con una polaridad global, indicando si el texto expresa un sentimiento positivo, negativo o neutral en 5 niveles. Se precisa autorización para acceder<sup>10</sup> (7).

Pero, en primer lugar, estos *corpora* no incluyen mensajes hablados mientras que en las redes sociales tipo Twitter, el lenguaje es más próximo al hablado que al escrito y, aun cuando el *corpus* TASS solo contiene tuits, estos están escritos por personas de cierta relevancia pública y no contienen mensajes de odio, sino, en todo caso, críticas a estos. Se trata, por tanto, de *corpora* más bien restringidos no equilibrados<sup>11</sup>. Su utilidad se limita a poder utilizarlos como auxiliares para el etiquetado gramatical<sup>12</sup>. Mediante estas técnicas sin supervisión extraeremos los tuits con una presunta inclinación hacia el odio que, esta vez sí, se etiquetarán a mano.

### Métricas de rendimiento.

Se utilizarán los indicadores clásicos en clasificación binaria ya que un texto dentro de la clase 1 puede pertenecer también la clase 1A y 1B:

1. Precisión
2. Exhaustividad
3. ROC

Si los documentos se asignan solo a una clase<sup>13</sup>, se usan las conocidas magnitudes

- Precisión (p): porcentaje de los documentos recuperados (clasificados como relevantes) que realmente lo son (aciertos).
- Exhaustividad (r): porcentaje de los documentos relevantes existentes en la población que han sido recuperados correctamente.

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).

La **¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia. ¡Error! No se encuentra el origen de la referencia.** muestra gráficamente estos valores y la **¡Error! No se encuentra el origen de la referencia.** representa cómo habitualmente están negativamente relacionados.

Ambos indicadores se combinan equilibradamente mediante su *media armónica* en F:

$$F(p, r) = \frac{2pr}{p + r}$$

<sup>9</sup> <https://gplsi.dlsi.ua.es/sepln15/es/taller-de-analisis-de-sentimientos-en-la-sepln-tass>

<sup>10</sup> [Tuits Tass](#) Usuario: corpus\_data\_tass. Contraseña: tass2016

<sup>11</sup> Es decir, la muestra de textos que es un corpus no es representativa de todos los tipos de habla, situación o variedad.

<sup>12</sup> Los dos primeros, puesto que el TASS no contiene tal etiquetado.

<sup>13</sup> Como es el caso en que las clases no son excluyentes.

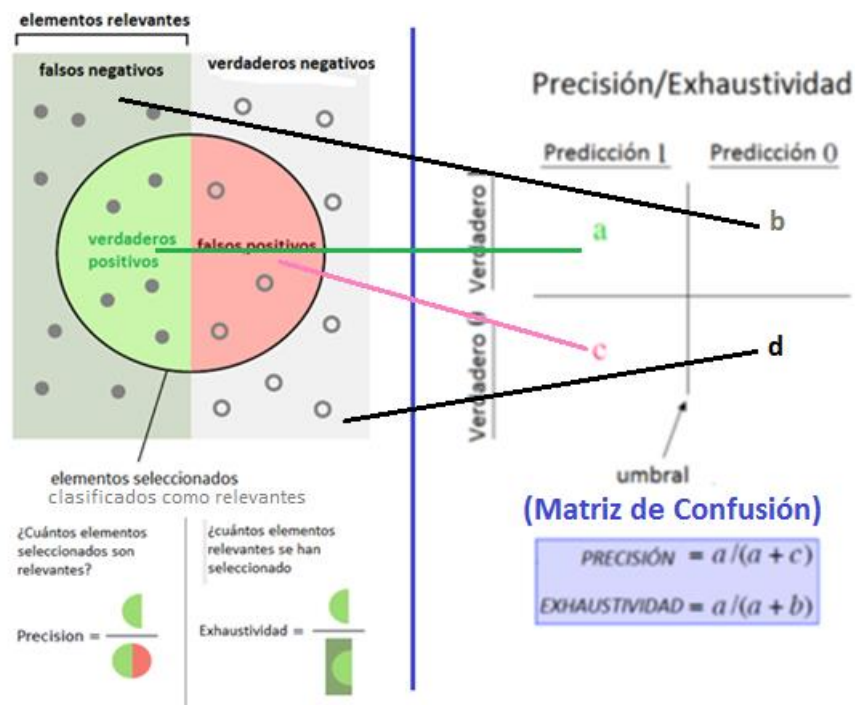


Figura 2 Validación y selección de modelos

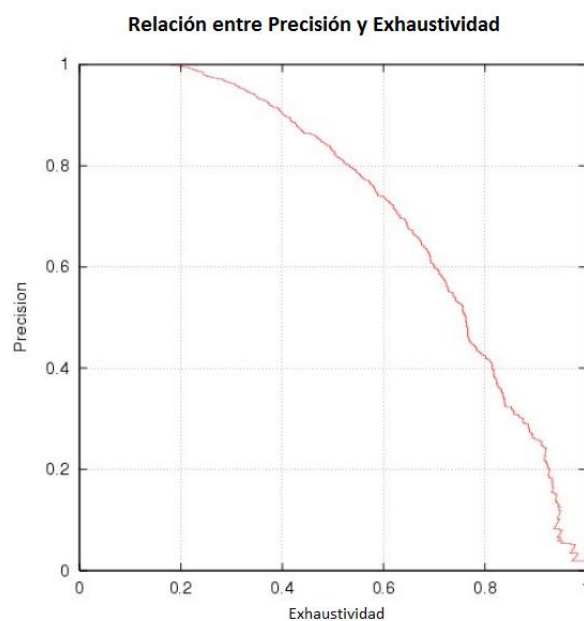


Figura 3 Relación Precisión-Exhaustividad

Otro indicador que se está utilizando cada vez más es el  $ROC^{14}$  habitual en Medicina y Biología para hablar de la detección de falsos positivos y negativos.

Ahora a la exhaustividad -  $\frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$  - se la denomina *sensibilidad*. Como se ve, es la

$$\Pr(\text{predicción\_TRUE} | \text{TRUE})$$

y se introduce la *especificidad*:

<sup>14</sup> Receiver Operator Characteristic.

$$\frac{\text{Verdaderos Negativos}}{\text{Verdaderos Negativos} + \text{Falsos Positivos}} = \frac{d}{(c + d)}$$

De manera que

$$1 - \text{especificidad} = \frac{\text{Falsos Positivos}}{\text{Verdaderos Negativos} + \text{Falsos Positivos}} = \frac{c}{(c+d)}$$

Y es

$$\Pr(\text{predicción\_TRUE}|\text{FALSE})$$

Por el teorema de la probabilidad total sabemos que:

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

Si dibujamos el gráfico que relaciona ambas magnitudes, obtenemos la Figura 4 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).

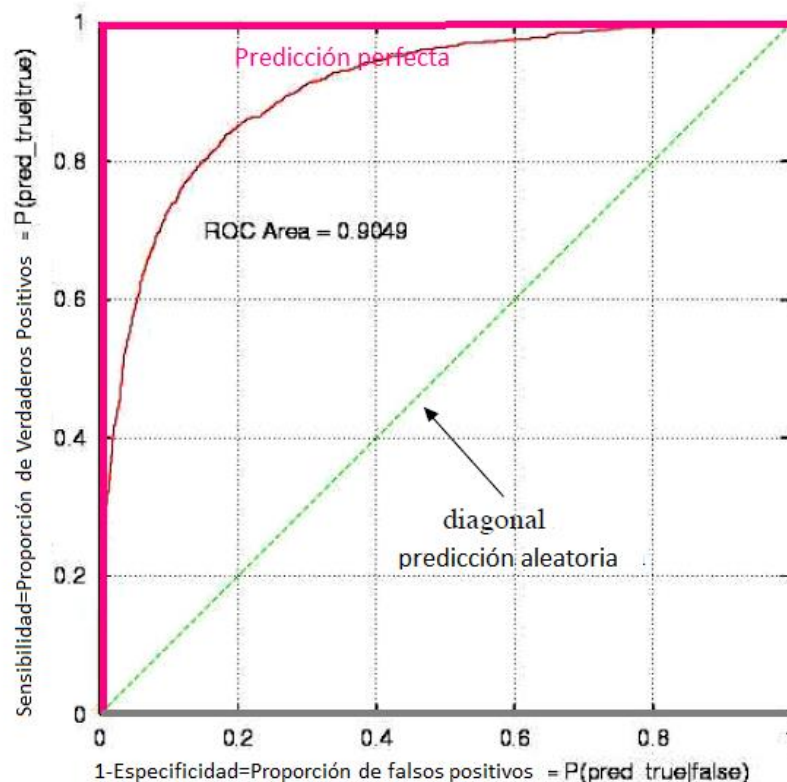


Figura 4 Indicador ROC

### Propuesta de actuación.

Por lo antes dicho, el proyecto contempla dos etapas diferenciadas:

1. La primera, tal como se expone gráficamente en la Figura 5 se refiere a la formación del 'patrón oro' para validación de modelos. Esta parte está basada en el método expuesto en (2) modificado en el sentido de realizar un cribado automático previo de los datos del cual se extraen los que aparentemente presentan alguna relación con mensajes de odio. Mensajes que son posteriormente revisados y etiquetados manualmente, dando lugar al conjunto de sentencias de entrenamiento.
2. La segunda consistirá en seleccionar en base a las métricas de rendimiento el método de clasificación más adecuado entre tres:
  - a. Bayes naïve,
  - b. K-vecinos más próximos y

c. Árboles de decisión.

Utilizando para la validación y selección el conjunto de entrenamiento previamente construido.

### Categorización de tuits semisupervisada

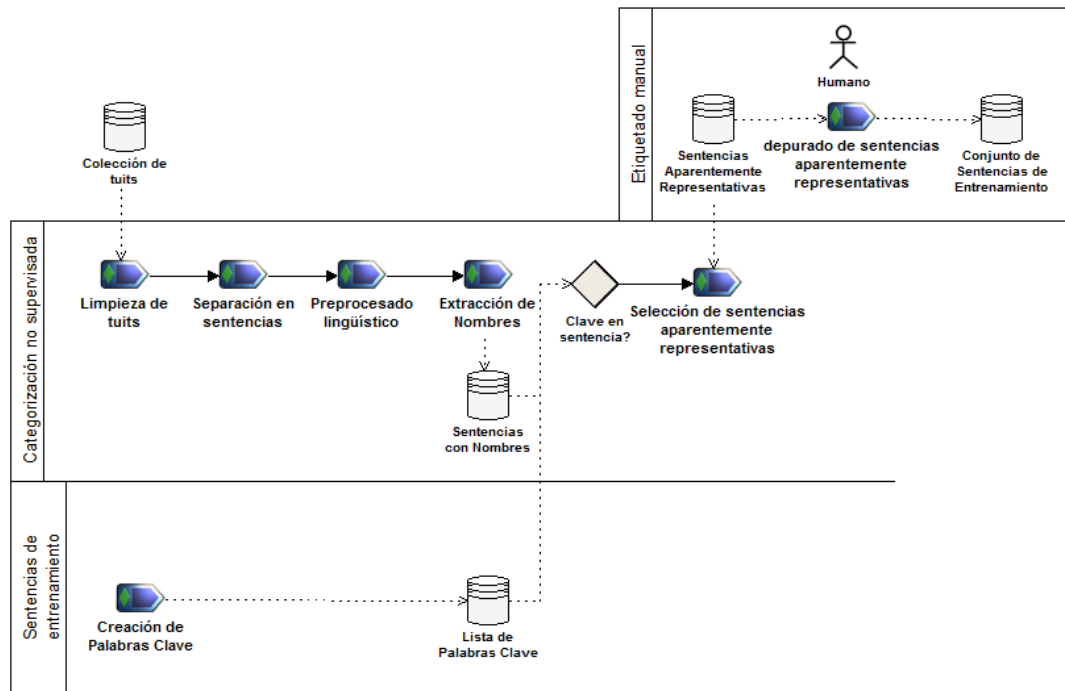


Figura 5 Método de Etiquetado de Tuits Semisupervisado



## Apéndice: TASS 2016 (8)

El Taller de Análisis de Sentimientos en la SEPLN 2016 proporciona dos conjuntos de tuits:

1. Un *corpus* general compuesto por dos conjuntos de datos en formato XML con tuits uno de entrenamiento y otro de prueba, con las siguientes características:

Atributo	Valor
<b>Tuits totales</b>	<b>68.017</b>
Tuits (entrenamiento)	60.798
Tuits (test)	7.219
<b>Tópicos</b>	<b>10</b>
<b>Usuarios</b>	<b>154</b>
Fecha de comienzo	02/12/2011
Fecha de finalización	10/04/2012

Los cuales se refieren a los siguientes tópicos:

- Política,
  - Entretenimiento,
  - Economía,
  - Música,
  - Fútbol,
  - Películas,
  - Tecnología,
  - Deportes,
  - Literatura,
  - Otros.
2. Un *corpus* denominado *STOMPOL* (Spanish Tweets for Opinion Mining at aspect level about POLitics) para el minado de opiniones políticas de los partidos más importantes del país.

## Referencias.

1. Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. s.l. : Springer, 2009.
2. Youngjoong Ko y Jungyun Seo. Association for Computational Linguistics. [En línea] <http://www.aclweb.org/anthology/C00-1066>.
3. Natural Language Toolkit. [En línea] 2017. <http://www.nltk.org/>.
4. Steven Bird, Ewan Klein, yd Edward Loper. Natural Language Processing with Python. [En línea] <http://www.nltk.org/book/>.
5. Universidad Politécnica de Cataluña. *WikiCorpus*. [En línea] <http://www.cs.upc.edu/~nlp/wikicorpus/>.
6. *Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus*. Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró. La Valetta : Proceedings of 7th Language Resources and Evaluation Conference (LREC'10), 2010.
7. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). SEPLN. [En línea] 2016. [Citado el: 2017 de Enero de 2017.] <http://www.sepln.org/workshops/tass/2016/>.
8. *Overview of TASS 2016*. Miguel Ángel García Cumberras, Julio Villena Román, Eugenio Martínez Cámara, M. Carlos Díaz. 2016. *TASS 2016 Proceedings*.
9. Ministerio del Interior. Ministerio del Interior. [En línea] 2015. [Citado el: 20 de enero de 2017.] <http://www.interior.gob.es/documents/10180/3066430/Informe+Delitos+de+Odio+2015.pdf>.
10. R.A.E. RAE. [En línea] 2017. <http://dle.rae.es/>.
11. *Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making*. Pete Burnap, Matthew L. Williams. 223-242, 2015, Policy & Internet, Vol. 7.
12. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. Pak, Alexander y Paroubek, Patrick. La Valetta : European Languages Resources Association (ELRA), 2010.
13. *Hatebase: An anti-genocide app*. Brown, Jesse. Toronto : Rogers Media, 2013, MacLean's.
14. Christopher D. Manning, Prabhakar Raghavan y Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge, UK : Cambridge University Press, 2008.
15. Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Englewood Cliffs, New Jersey : Prentice Hall, 2000.
16. *Lexical and algorithmic stemming compared for 9 European Languages with Hummingbird Searchserver at CLEF 2003*. Tomlinson, Stephen. s.l. : Proc. CrossLanguage Evaluation Forum, 2003. págs. 286-300.
17. Russell, Matthew A. *Mining the Social Web*. Sebastopol, CA : O'Reilly, 2014.
18. Mining the Social Web 2nd Edition. [En línea] <https://github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition/tree/master/ipynb>.

## Glosario

### A

#### Análisis de Sentimiento

También llamado minería de opinión consiste en el uso de procesamiento de lenguaje natural para identificar y extraer información subjetiva de unos recursos textuales. El análisis de sentimientos es una tarea de clasificación masiva de textos de manera automática, en función de la connotación positiva o negativa del lenguaje utilizado en el documento..... 1, 4

#### antisemitismo

Enemistad hacia los judíos, su cultura o su influencia.....4

#### aporofobia

Repugnancia u hostilidad ante el pobre, el sin recursos o el desamparado. ....4

#### aprendizaje maquina

Aprendizaje automático, aprendizaje máquina o *machine learning* es la ciencia de conseguir que las computadoras actúen sin haber sido explícitamente programadas.....4

### C

#### corpora

Plural de corpus. Un *corpus* es una gran colección de textos que contienen material escrito o hablado sobre el que se basa el análisis lingüístico. ....6

### E

#### etiquetado

Añadir un campo - etiqueta - que identifique alguna característica. En este caso la clase de pertenencia del tweet..... 5, 6

### N

NLP ..... Véase Procesamiento del Lenguaje Natural.

NLTK..... Véase NaturalLanguage Toolkit

### P

#### procesamiento del lenguaje natural

Campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. ....6

### R

#### RAE

Real Academia de la Lengua Española.....4

### X

#### xenofobia

Miedo, rechazo u odio al extranjero.....4