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 de Situación #1

Juan Carlos Pereira Kohatsu

Enero 2017

Índice de contenidos.

Índice de Figuras.

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[1] 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*[2] *para la Detección de Delitos de Odio.*

Para ello se pretende seleccionar un procedimiento de aprendizaje maquinal supervisado (naïve Bayes, vecinos próximos, árboles de clasificación,…) que detecte aquellos tuits que contengan mensajes denigratorios o amenazadores para los grupos[3] 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[4] o Streaming[5]) y disponemos de un conjunto de tuits de tamaño adecuado[6].

---

[1] El Ministerio del Interior [12] clasifica los delitos de odio en 8 categorías: Antisemitismo, aporofobia, religión, discapacidad, identidad u orientación sexual, racismo/xenofobia, ideología y discriminación por sexo/género.

[2] Usamos *tuit* por ser el término admitido en la RAE para los mensajes de Twitter® [13].

[3] 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.).

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

[5] 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.

[6] 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[7] 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 **¡Error! No se encuentra el origen de la referencia.¡Error! No se encuentra el origen de la referencia.**.

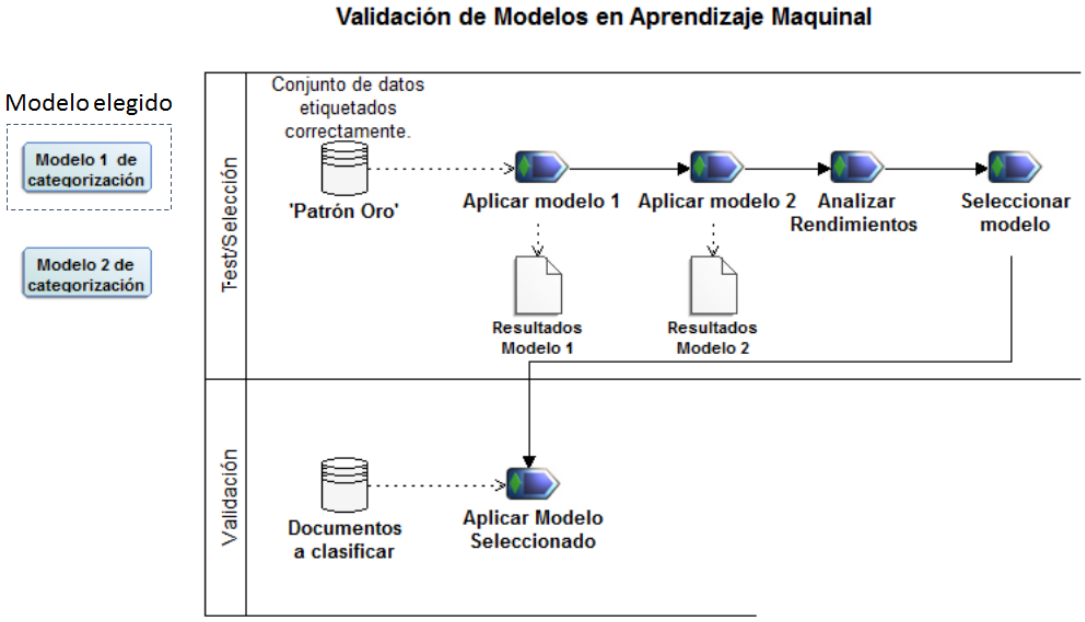

Figura 1 Validación y selección de modelos

Según las primeras observaciones, la proporción de tuits que tratan temas relacionados con el odio es habitualmente muy limitada[8] 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.

[7] 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.
[8] La frecuencia de aparición de estos mensajes es muy variable y suele dispararse, bien tras producirse cierto tipo de eventos como manifestaciones, atentados, etc. [14] o bien previamente en campañas para preparar agresiones planificadas a ciertos grupos [16].

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

- *Wikicorpus* [5] [6]
- Cess_esp contiene 6.030 sentencias anotadas.
- *Corpus* TASS[9]: 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[10] [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[11]. Su utilidad se limita a poder utilizarlos como auxiliares para el etiquetado gramatical[12].
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[13], 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:

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

[9] https://gplsi.dlsi.ua.es/sepln15/es/taller-de-analisis-de-sentimientos-en-la-sepln-tass
[10] Tuits Tass Usuario: `corpus_data_tass`. Contraseña: `tass2016`
[11] Es decir, la muestra de textos que es un corpus no es representativa de todos los tipos de habla, situación o variedad.
[12] Los dos primeros, puesto que el TASS no contiene tal etiquetado.
[13] Como es el caso en que las clases no son excluyentes.

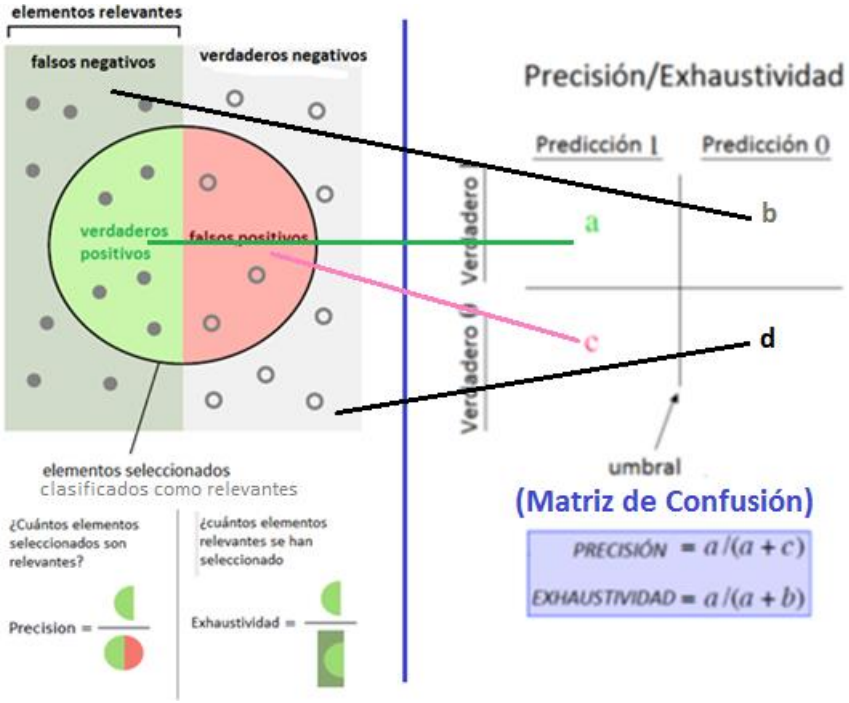

Figura 2 Validación y selección de modelos

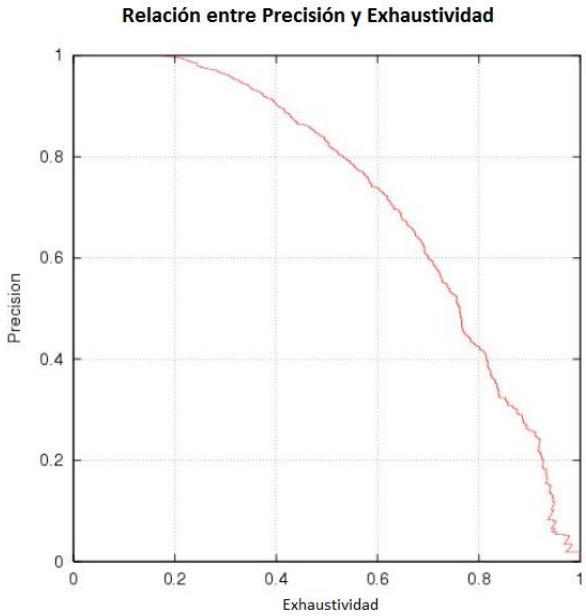

Figura 3 Relación Precisión-Exhaustividad

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

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

y se introduce la *especificidad:*

[14] Receiver Operator Characteristic.

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

De manera que

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

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 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).

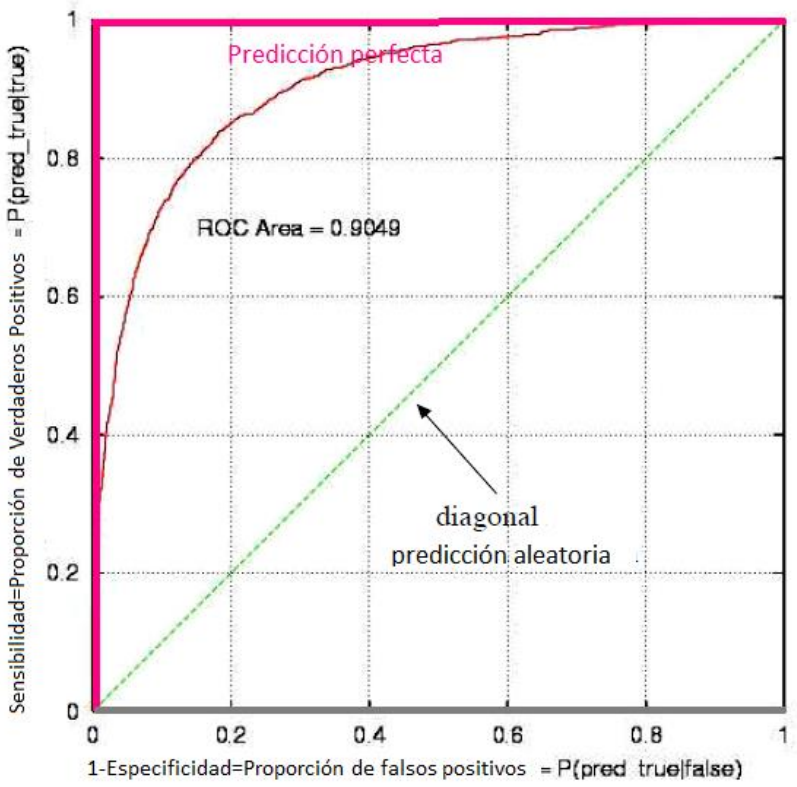

Figura 4 Indicador ROC

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.

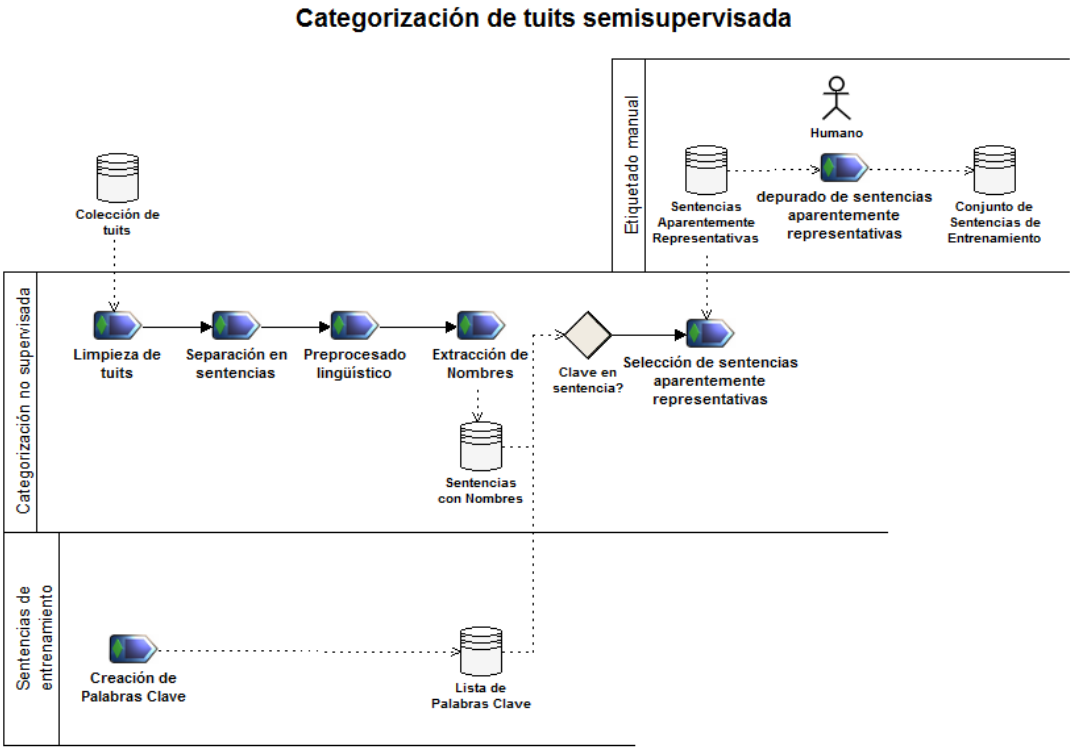

Figura 5 Método de Etiquetado de Tuits Semisupervisado

Apéndice: TASS 2016 [11]

El **T**aller de **A**nálisis de **S**entimientos en la **S**EPLN 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
Tuits totales	**68.017**
Tuits (entrenamiento)	60.798
Tuits (test)	7.219
Tópicos	**10**
Usuarios	**154**
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] R. T. y. J. F. Trevor Hastie, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.

[2] Youngjoong Ko y Jungyun Seo, «Association for Computational Linguistics,» [En línea]. Available: http://www.aclweb.org/anthology/C00-1066.

[3] «Natural Language Toolkit,» 2017. [En línea]. Available: http://www.nltk.org/.

[4] E. K. y. E. L. Steven Bird, «Natural Language Processing with Python,» [En línea]. Available: http://www.nltk.org/book/.

[5] «Universidad Politécnica de Cataluña,» [En línea]. Available: http://www.cs.upc.edu/~nlp/wikicorpus/.

[6] G. B. M. C. L. P. Samuel Reese, «Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus,» La Valetta, 2010.

[7] Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), «SEPLN,» 2016. [En línea]. Available: http://www.sepln.org/workshops/tass/2016/. [Último acceso: 2017 Enero 2017].

[8] P. R. y. H. S. Christopher D. Manning, Introduction to Information Retrieval, Cambridge, UK: Cambridge University Press, 2008.

[9] Daniel Jurafsky and James H. Martin, Speech and Language Processing, Englewood Cliffs, New Jersey: Prentice Hall, 2000.

[10] S. Tomlinson, «Lexical and algorithmic stemming compared for 9 European Languages with Hummingbird Searchserver at CLEF 2003,» 2003.

[11] J. V. R. E. M. C. M. C. D. Miguel Ángel García Cumbreras, «Overview of TASS 2016,» de *tTASS 2016 Proceedings*, 2016.

[12] Ministerio del Interior, «Ministerio del Interior,» 2015. [En línea]. Available: http://www.interior.gob.es/documents/10180/3066430/Informe+Delitos+de+Odio+2015.pdf. [Último acceso: 20 enero 2017].

[13] R.A.E., «RAE,» 2017. [En línea]. Available: http://dle.rae.es/.

[14] M. L. W. Pete Burnap, «Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making,» *Policy & Internet*, vol. 7, nº 223-242, 2015.

[15] A. y. P. P. Pak, «Twitter as a Corpus for Sentiment Analysis and Opinion Mining,» La Valetta, 2010.

[16] J. Brown, «Hatebase: An anti-genocide app,» *MacLean's*, 2013.

[17] M. A. Russell, Mining the Social Web, Sebastopol, CA: O'Reilly, 2014.

[18] «Mining the Social Web 2nd Edition,» [En línea]. Available: 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, 3

antisemitismo
: Enemistad hacia los judíos, su cultura o su influencia. 3

aporofobia
: Repugnancia u hostilidad ante el pobre, el sin recursos o el desamparado. 3

aprendizaje maquinal
: Aprendizaje automático, aprendizaje máquina o *machine learning* es la ciencia de conseguir que las computadoras actúen sin haber sido explícitamente programadas 3

C

corpora
: Plural de corpus. Un *corpus* es una gran colección de textosque contienen material escrito o hablado sobre el que se basa el análisis lingüístico. 4

E

etiquetado
: Añadir un camo - etiqueta - que identifique alguna característica. En este caso la clase de pertenencia del tuit. 4

L

lematización
: Proceso lingüístico que consiste en, dada una forma flexionada (es decir, en plural, en femenino, conjugada, etc), hallar el lema correspondiente. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. 5

N

NLP ..... *Véase* Procesamiendo del Lenguaje Natural.

NLTK ..... *Véase* NaturalLanguage Toolkit

Normalización
: Proceso de hacer equivalentes series de caracteres con diferencias superficiales (ONU frente a O.N.U. pe.) o listas de sinónimos (coche, automóvil) 5

P

palabras vacías
: *Nombre que reciben las palabras sin significado como artículos, pronombres, preposiciones, etc. que se elinan en el procesamiento de datos en lenguaje natural por su escaso valor.* 5

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. 4

R

RAE
: Real Academia de la Lengua Española. 3

S

Stemming