**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.

[Tareas emprendidas. 3](#_Toc472671990)

[1. Descripción y nombre del proyecto. 3](#_Toc472671991)

[2. Características del proyecto. 3](#_Toc472671992)

[Primeros resultados. 3](#_Toc472671993)

[Referencias 4](#_Toc472671994)

[Glosario 5](#_Toc472671995)

Índice de Figuras.

## Tareas emprendidas.

### Descripción y nombre del proyecto.

Esencialmente, el proyecto consistirá en desarrollar un sistema de clasificación de tuits en las siguientes clases[[1]](#footnote-1) que difieren ligeramente de las utilizadas por el ministerio del Interior:

* 1. Clase 1: mensaje de odio.
     1. Clase 1A: Racismo/xenofobia.
     2. Clase 1B: Sexo/género.
     3. Clase 1D: Identidad/orientación sexual.
     4. Clase 1F: Discapacidad.
     5. Clase 1G: Clase social.
  2. 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.

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]](#footnote-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]](#footnote-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]](#footnote-4) o Streaming[[5]](#footnote-5)) y disponemos de un conjunto de tuits de tamaño adecuado.

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.

Al requerirse el etiquetado de los mensajes y, según las primeras observaciones, siendo habitualmente la proporción de tuits que tratan temas relacionados con el odio muy limitada[[6]](#footnote-6) 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 automático o semiautomático. 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 (1).

Tal enfoque requiere del uso de herramientas de procesamiento del lenguaje natural (NLP) tales como el paquete Natural Language Toolkit (2) (NLTK) descrito en el libro de Steven Bird, Ewan Klein, and Edward Loper (3).

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:

* Wikicorpus[[7]](#footnote-7) (4) (5)
* Cess\_esp contiene 6030 sentencias anotadas.
* Corpus TASS[[8]](#footnote-8): 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 incluía 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 que se ha pedido[[9]](#footnote-9).

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, salvo el Corpus TASS

Pero se trata de *corpora* más bien restringidos, y puesto que, por un lado, cada aplicación requiere que el corpus utilizado sea representativo para el tipo de hablante, situación o variedad que es el objetivo del proyecto y, por otro, los *corpora* no son equilibrados[[10]](#footnote-10). Esto es especialmente importante cuando el grupo de textos que nos interesan son de pequeño tamaño.

Como alternativa, contemplamos la utilización de unigramas o bigramas como criterio principal de clasificación.

### 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. Tabla de confusión
2. Precisión
3. Exhaustividad
4. ROC

## Primeros resultados.

# Referencias

1. **Youngjoong Ko y Jungyun Seo.** Asociation for Computational Linguistics. [En línea] http://www.aclweb.org/anthology/C00-1066.

2. **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.

3. **R.A.E.** RAE. [En línea] 2017. http://dle.rae.es/.

4. *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.

.

# 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

E

etiquetado

Añadir un camo - etiqueta - que identifique alguna característica. En este caso la clase de pertenencia del tuit. 3

R

RAE

Real Academia de la Lengua Española. 3

X

*xenofobia*

Miedo, rechazo u odio al extranjero 3

1. El Ministerio del Interior (3) clasifica los delitos de odio en 8 categorías: Antisemitismo, aporofobia, religión, discapacidad, identidad u orientación sexual, racismo/xenofobia, ideología y discriminación por sexo/género. [↑](#footnote-ref-1)
2. Usamos *tuit* por ser el término admitido en la RAE para los mensajes de Twitter® (4). [↑](#footnote-ref-2)
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.). [↑](#footnote-ref-3)
4. La *REST API* consiste básicamentye en una consulta (QUERY) a su servidor que devuelve una respuesta en JSON, XML, etc. [↑](#footnote-ref-4)
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. [↑](#footnote-ref-5)
6. La frecuencia de aparición de estos mensajes es muy variable y se dispara tras producirse cierto tipo de eventos como manifestaciones, atentados, etc. (4) [↑](#footnote-ref-6)
7. <http://www.clips.ua.ac.be/pages/using-wikicorpus-nltk-to-build-a-spanish-part-of-speech-tagger> [↑](#footnote-ref-7)
8. <https://gplsi.dlsi.ua.es/sepln15/es/taller-de-analisis-de-sentimientos-en-la-sepln-tass> [↑](#footnote-ref-8)
9. [Tuits Tass](http://www.sepln.org/workshops/tass/2016/) Usuario: corpus\_data\_tass. Contraseña: tass2016 [↑](#footnote-ref-9)
10. Es decir, la muestra de textos que es un corpus no es representativa de todos los tipos de habla, situación o variedad. [↑](#footnote-ref-10)