**UNIVERSIDAD AUTONOMA DE MADRID**

**ESCUELA POLITECNICA SUPERIOR**

** **

**Thesis submitted in fulfillment of the requirements for the grade in *Computer Science***

**Sentiment Analysis for Hate Crime Detection**

**Author: Juan Carlos Pereira Kohatsu**

**JUNE 2017**

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](http://www.tripadvisor.es), [yelp](https://www.yelp.es)),
* contactos sociales y profesionales ([facebook](https://es-es.facebook.com/), [linkedin](https://es.linkedin.com/)),
* *blogs* o bitácoras *web* ([wordpress](https://es.wordpress.com)).

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, que, en su artículo 510 dice:

(Ley Orgánica 1/2015, 2015)

1. Serán castigados con una pena de prisión de uno a cuatro años y multa de seis a doce meses:

a) Quienes *públicamente* *fomenten*, *promuevan* o *inciten* directa o indirectamente al *odio*, hostilidad, *discriminación* o *violencia* contra un grupo, una parte del mismo o contra una persona determinada por razón de su pertenencia a aquél, por motivos racistas, antisemitas u otros referentes a la ideología, religión o creencias, situación familiar, la pertenencia de sus miembros a una etnia, raza o nación, su origen nacional, su sexo, orientación o identidad sexual, por razones de género, enfermedad o discapacidad.

Desde el trabajo de Gary Becker (Becker, 1968) 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 riego, 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 (Becker, 1968), 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 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 (Goodman, 2016).

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

Abstract.

The so-called social networks made up of platforms such as Facebook ™ and Twitter ™ that operate on the Internet are the support of the social media that facilitate the exchange and discussion of information, experiences and opinions between individuals in a fast and massive way, this has never been seen before in the History of Humanity.

The range of social media open to public use is varied and growing and its uses are multiple:

• articles in wikis,

• opinions on the quality of hotels, restaurants (tripadvisor, yelp),

• social and professional contacts (facebook, linkedin),

• blogs or weblogs (wordpress).

Of course, like everything new, the explosion of social media has had consequences that have been valued both positively and negatively for society as awhole.

Among the effects generally considered negative, social media have persistently made 'visible' some attitudes of certain social groups that, up to now, had only been shown in a veiled occasional way. Among them are those which would lead to attacks on individuals or clusters because of their membership in a group defined by characteristics such as nationality, sexual preferences, race, religion ...

This phenomenon, together with a change of attitude towards certain social behaviors or groups, has led to the emergence of a new category of crime in many countries: the so-called hate crimes that have been regulated in many countries, Spain among them since 2015.

The OSCE defines *hate crime* as , more specifically (OSCE, s.f.):

*Hate crimes are criminal acts motivated by bias or prejudice towards particular groups of people. To be considered a hate crime, the offence must meet two criteria: First, the act must constitute an offence under criminal law; second, the act must have been motivated by bias.*

*Bias motivations can be broadly defined as preconceived negative opinions, stereotypical assumptions, intolerance or hatred directed to a particular group that shares a common characteristic, such as race, ethnicity, language, religion, nationality, sexual orientation, gender or any other fundamental characteristic. People with disabilities may also be victims of hate crimes.*

*Hate crimes can include threats, property damage, assault, murder or any other criminal offence committed with a bias motivation. Hate crimes don't only affect individuals from specific groups. People or property merely associated with – or even perceived to be a member of – a group that shares a protected characteristic, such as human rights defenders, community centres or places of worship, can also be targets of hate crimes*.

From Gary Becker`s work (Becker, 1968) on *crime and punishment*, we know that the results derived from their models show that an increase in the probability of punishment or arrest, regardless of the offender's risk attitude, has a negative effect on the crime *supply*. However, the effect of an increase in the severity of penalties is ambiguous without additional assumptions about offender’s position vis-à-vis risk. This effect is uncertain for risk-loving agents, but not for risk-averse, in which case an increase in the severity of the penalty reduces their crime supply.

*Keywords:* Social networks, social media,

1. Introducción.

Los *delitos de odio* son un tipo de delito 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 basados de odio están influidos por eventos *singulares* de *amplia difusión* (Pete Burnap , Matthew L. Williams, 2015) (atentados terroristas, migración incontrolada, manifestaciones, revueltas, …). Este tipo de sucesos suelen actuar como detonadores de manera que la frecuencia de este tipo de delitos aumenta espectacularmente tras ellos. Por ello, parece razonable dotar a los responsables de mantener la seguridad de herramientas que permitan evaluar la probabilidad del alza 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 sociales se llenan de mensajes de individuos afines a los perpetradores que incitan a castigar al grupo elegido como diana (King, R.D., G.M. Sutton, 2013) 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]](#footnote-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 de Análisis de Sentimiento[[2]](#footnote-2) para la Detección 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 indican una mayor *oferta* de delitos[[3]](#footnote-3)) o negativos/neutros que no aportan pistas al respecto.

## Estructura.

La memoria explicativa del proyecto se estructura de la siguiente manera:

* 1. Comenzamos revisando algunos conceptos básicos del Análisis de Sentimiento y el *estado del arte*.
  2. Seguidamente, en el apartado *descripción del proyecto* se procede a detallar aspectos tales como
     1. Especificación del producto,
     2. Herramientas utilizadas
        1. Metodología de gestión de proyectos utilizada
        2. Fuentes de datos,
        3. Software y hardware,
        4. Control de versiones
  3. A continuación se expone el diseño del proyecto

## Estado del arte.

El elemento clave es el *sentimiento* u *opinión* que el redactor del mensaje[[4]](#footnote-4) –en este caso del t*weet* - expresa en el mismo respecto a una *entidad* o aspecto de la misma.

La *orientación[[5]](#footnote-5)* de tal opinión puede ser positiva, negativa o neutral.

Una entidad puede representarse como una jerarquía de componentes tal como la mostrada en la Figura 1.

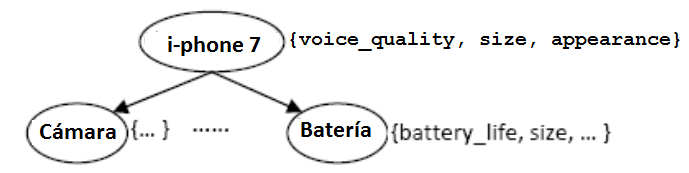


Figura Estructura jerárquica de una entidad

Cada nodo representa un *componente* y está asociado a un conjunto de *atributos* de dicho componente (p.e. para la batería, su vida útil y tamaño). Una opinión puede aplicarse sobre cualquier nodo de la entidad o atributo del nodo (p.e. la cámara - *nodo* - es buena, pero el tamaño (size)de la batería - *atributo* - es excesivo. Para simplificar, suele utilizarse el mismo término: *aspecto* o *característica* indistintamente para nodos y atributos.

Una opinión es básicamente una 5-tupla**:**  donde sus componentes representan lo siguiente:

* es una entidad.
* es un aspecto/característica de la entidad .
* es el valor (positivo, negativo o neutro) del sentimiento de la fuente de opinión sobre la característica en el tiempo .
* es el redactor del mensaje (titular de la opinión).
* es el instante de tiempo en que se expresa la opinión.

De todos estos valores, la dupla que relacionaentidad y atributo se denomina *objeto de la opinión*.

Naturalmente, pueden añadirse campos adicionales a la tupla de opinión tales como *género*, *edad*, *geolocalización*, etc.

Si el texto se refiere a *una sola entidad,* el objetivo del análisis de sentimiento se simplifica y consiste en descubrir en cada texto con independencia de a, h y t**.**

Este tipo de análisis es p.e. el efectuado en Amazon cuando los comentarios de los clientes sobre *un producto* - entidad en este caso - asignan a este un número determinado de  como puede verse en la Figura 2.



Figura Evaluación de productos en Amazon

## Descripción del proyecto.

## Diseño.

## Desarrollo.

## Pruebas.

## Resultados.

## Referencias.

Contenido

[*Palabras clave:* 3](#_Toc471990161)

[*Keywords:* 3](#_Toc471990162)

[2. Estructura. 4](#_Toc471990163)

[3. Estado del arte. 4](#_Toc471990164)

[4. Diseño. 4](#_Toc471990165)

[5. Desarrollo. 4](#_Toc471990166)

[6. Pruebas. 4](#_Toc471990167)

[7. Resultados. 4](#_Toc471990168)

[8. Referencias. 4](#_Toc471990169)

[Bibliografía 6](#_Toc471990170)

# Bibliografía

*BOE*. (30 de Marzo de 2015). Obtenido de https://www.boe.es/boe/dias/2015/03/31/pdfs/BOE-A-2015-3439.pdf

1. En España en 2015 (Ley Orgánica 1/2015, 2015) [↑](#footnote-ref-1)
2. O *Minería de Opiniones* [↑](#footnote-ref-2)
3. Utilizamos la terminología económica de G. Becker (Becker, 1968) [↑](#footnote-ref-3)
4. O *fuente de opinión.* [↑](#footnote-ref-4)
5. También llamada *orientación semántica* o *polaridad*. [↑](#footnote-ref-5)