

Procesamiento de lenguaje natural tóxico en redes sociales.

**Nombre Estudiante:** Esperanza Rodríguez-Correa Fernández

Inteligencia de Negocio y Big Data

B2.593 TFM Big data

**Nombre Consultor/a** José Luís Gómez

**Nombre Profesor/a responsable de la asignatura** Josep Curto Díaz

**Fecha Entrega** 06/2019

  
Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada [3.0 España de Creative Commons](http://creativecommons.org/licenses/by-nc-nd/3.0/es/)

© (Esperanza Rodríguez-Correa Fernández)

Reservados todos los derechos. Está prohibido la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la impresión, la reprografía, el microfilme, el tratamiento informático o cualquier otro sistema, así como la distribución de ejemplares mediante alquiler y préstamo, sin la autorización escrita del autor o de los límites que autorice la Ley de Propiedad Intelectual.

**FICHA DEL TRABAJO FINAL**

|  |  |
| --- | --- |
| **Título del trabajo:** | *Procesamiento de lenguaje natural tóxico en redes sociales.* |
| **Nombre del autor:** | Esperanza Rodríguez-Correa Fernández |
| **Nombre del consultor/a:** | *José Luis Gómez García* |
| **Nombre del PRA:** |  |
| **Fecha de entrega (mm/aaaa):** | 06/2019 |
| **Titulación:** | *Plan de estudios del estudiante* |
| **Área del Trabajo Final:** | *El nombre y código de la asignatura de TF* |
| **Idioma del trabajo:** | *Español* |
| **Palabras clave** | *NLP, Machine Learning*  *Toxic messages* |
| **Resumen del Trabajo (máximo 250 palabras):** *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.* | |
| La finalidad de este proyecto es el reconocimiento de lenguaje tóxico en redes sociales o cualquier comunicación entre humanos en Internet como puede ser un foro universitario.  Ésta propuesta está inspirada en ‘perspective’ tecnología creada por Jigsaw y el equipo de tecnología de lucha contra el abuso de Google a partir de un proyecto de investigación colaborativo llamado ‘Conversation-AI’, finalmente ha sido integrada en la biblioteca Tensorflow. Este estudio se diferencia de múltiples encontrados en que se hace con una comparativa entre un dataset en español recopilado de twitter y un dataset en ingles haciendo una comparativa de comportamiento con las actuales herramientas del mercado. | |
| **Abstract (in English, 250 words or less):** | |
| The main purpose of this project is to recognize toxic language in the social network or any conversation between humans, such as interactions in a university forum.  This approach has been inspired by 'perspective' technology created by Jigsaw and Google’s Counter Abuse Technology team in a collaborative research project called Conversation-AI and finally integrated into the Tensorflow library.  What this study contributes is not only a project which compares the performance of the gold standard algorithms on two datasets with different languages (Spanish and English) but also probes the necessity of a library similar to Vader in Spanish with the most common toxic comments. | |

**Índice**

[1. Introducción 1](#_Toc12255943)

[1.1 Contexto y justificación del Trabajo 1](#_Toc12255944)

[1.2 Objetivos del Trabajo 1](#_Toc12255945)

[1.3 Enfoque y método seguido 2](#_Toc12255946)

[1.4 Planificación del Trabajo 2](#_Toc12255947)

[1.5 Breve sumario de productos obtenidos 1](#_Toc12255948)

[1.6 Breve descripción de los otros capítulos de la memoria 1](#_Toc12255949)

[2. Estado de los modelos de análisis de sentimientos. 2](#_Toc12255950)

[2.1. Librerias de análisis de sentimientos, modelos no supervisados. 2](#_Toc12255951)

[2.1.1 LIWC 2](#_Toc12255952)

[2.1.2 VADER 3](#_Toc12255953)

[2.1.3 IBM Watson Sentiment Analyzer. 4](#_Toc12255954)

[2.1.4 TextBlob 4](#_Toc12255955)

[3. Descripción del dataset del challenge Jigsaw Unintended Bias in Toxicity Classification 5](#_Toc12255956)

[3.1 Descripción de los datos del dataset 6](#_Toc12255957)

[3.2 EDA 7](#_Toc12255958)

[4. Descripción del dataset de twitter 11](#_Toc12255959)

[4.1 Creación del corpus de twitter 11](#_Toc12255960)

[4.2 Preprocesamiento de los datos de twitter 11](#_Toc12255961)

[4.2 Preprocesamiento de los datos de twitter 12](#_Toc12255962)

[5. Estudio de modelos de Machine Learning 16](#_Toc12255963)

[5.1 Pasos iniciales 16](#_Toc12255964)

[3. Conclusiones 17](#_Toc12255965)

[4. Glosario 18](#_Toc12255966)

[5. Bibliografía 19](#_Toc12255967)

[6. Anexos 20](#_Toc12255968)

**Lista de figuras**

[Ilustración 1 Sesgo en la herramienta perspective. Fuente: civic.mit.edu 1](#_Toc12255932)

[Ilustración 2 Acoso/Toxicidad en redes sociales (Fuente: AMERICAblog) 2](#_Toc12255933)

[Ilustración 3. Pantalla usada po la plataforma Civil Comments 6](#_Toc12255934)

[Ilustración 4 Cantidad de comentarios tóxicos 8](#_Toc12255935)

[Ilustración 5 Número de palabras por comentario 9](#_Toc12255936)

[Ilustración 6 Distribución de los comentarios tóxicos 9](#_Toc12255937)

[Ilustración 7 Matriz de confusión 10](#_Toc12255938)

[Ilustración 8 Distribución temporal por grupo de identidades 11](#_Toc12255939)

[Ilustración 9 100 palabras consideradas como tóxicas 14](#_Toc12255940)

[Ilustración 10 100 palabras consideradas como no tóxicas 15](#_Toc12255941)

[Ilustración 11 Coeficiente de correlación de Pearson 16](#_Toc12255942)

**Lista de tablas**

[Tabla 1 Ejemplos de funcionamiento Vader 3](#_Toc12255922)

[Tabla 2 Resultados sobre 241 palabras tóxicas en castellano 3](#_Toc12255923)

[Tabla 3 Rendimiento de IBM Watson 4](#_Toc12255924)

[Tabla 4 Análisis de BIAS en IBM Watson 4](#_Toc12255925)

[Tabla 5 Análisis de reconocimiento de toxicidad 4](#_Toc12255926)

[Tabla 6 Análisis de Bias en TextBlow 5](#_Toc12255927)

[Tabla 7 Análisis de reconocimiento de toxicidad 5](#_Toc12255928)

[Tabla 8 Tamaño de los dataset 7](#_Toc12255929)

[Tabla 9 Nuevas columnas del dataframe de Twitter 13](#_Toc12255930)

[Tabla 10 Resultados Bigram 15](#_Toc12255931)

# 1. Introducción

## 1.1 Contexto y justificación del Trabajo

En la actualidad nos encontramos con un problema de lenguaje tóxico y acoso en las redes sociales que herramientas como perspective intentan solventar. Uno de los problemas a resolver es la existencia del sesgo en determinadas palabras que según el contexto tienen significados diferentes, ejemplo de ello es que las palabras interesante y enriquecedor son positivas pero la frase ‘La película no fue ni enriquecedora ni interesante’ es negativa que la mayoría de los desarrollos hayan sido realizados sobre la lengua inglesa o alemana hace que los resultados sobre el análisis en la lengua española dependan de lo bueno o malo que sea un traductor o de si expresiones populares han sido incorporadas al dataset que se utilizó para el entrenamiento.

La plataforma Civil comments fue la comenzó con este estudio y planteaba como habían demostrado que cuando a un usuario se le indicaba que el lenguaje que estaba escribiendo era tóxico, modificaba su mensaje de forma más respetuosa.

Queremos comprobar sobre un dataset proporcionado por perspective para Kaggle y sobre una muestra de tweets extraídos desde su API en Español si podríamos utilizar las mismas herramientas o sería necesaria la realización de un dataset más amplio en un futuro.

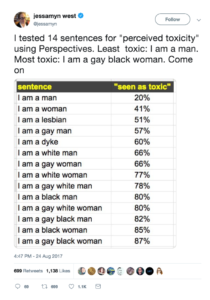


Ilustración 1 Sesgo en la herramienta perspective. Fuente: civic.mit.edu

## 1.2 Objetivos del Trabajo

El objetivo del proyecto es poner en práctica los conocimientos adquiridos sobre análisis de datos, así como la ampliación de ellos con un análisis en profundidad del estado de las técnicas de NLP en el lenguaje español. Se quiere obtener un método de análisis de lenguaje tóxico inclusivo sin sesgos asociados debido al incremento de acoso en redes sociales:



Ilustración 2 Acoso/Toxicidad en redes sociales (Fuente: AMERICAblog)

Se concretan en los siguientes puntos:

* Análisis actual de las librerías de Python que permiten el análisis de sentimientos tanto en ingles como en español.
* Análisis actual de los métodos NLP de análisis de sentimientos
* Análisis de los datos con bias de Kaggle
* Generación de resultados y modelo sobre el anterior dataset
* Descarga de muestras en Español de Twitter
* Análisis de los resultados obtenidos con el modelo anterior sobre los nuevos datos de la red social.

## 1.3 Enfoque y método seguido

Se comenzará analizando los estudios actuales de lenguaje tóxico y cómo se comportan los diferentes enfoques tanto en un dataset propio creado en castellano como en el utilizado en el desarrollo del modelo de Jigsaw . El modelo en castellano, se quiere enriqueces con una librería de insultos recopilados en internet que puedan incrementar el peso del análisis.

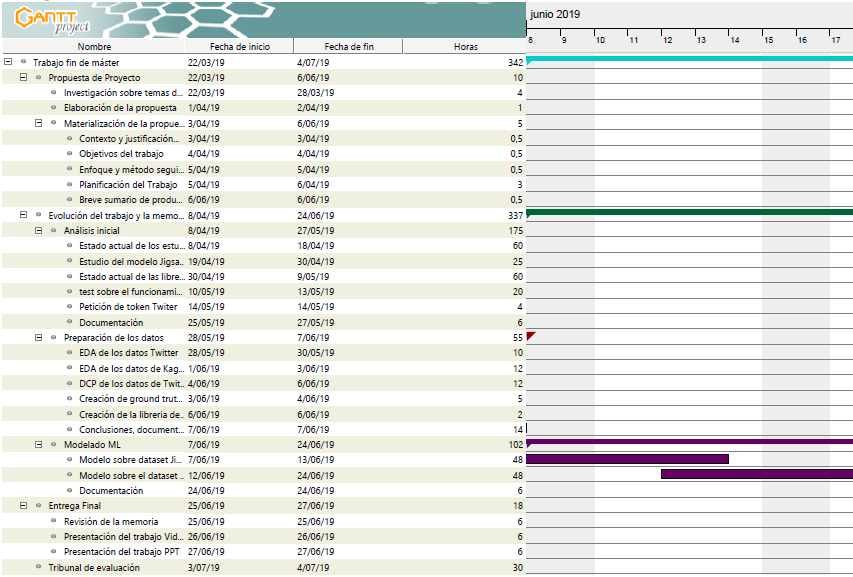
## 1.4 Planificación del Trabajo

En este proyecto debido a el volumen del dataset (train.csv: 779M) en vez de utilizar una plataforma como databricks para el uso de Pyspark, se realizará sobre a plataforma anaconda con Pandas ya que el volumen de datos no es tan grande como para necesitar un lenguaje distribuido. Además, Pandas a día de hoy sigue siendo más rápido y más fácil de implementar debido a sus múltiples librerías.

Debido a la necesidad de copiar y pegar código en la memoria final se utilizará la herramienta externa <http://www.planetb.ca/syntax-highlight-word> que permitirá mantener el formato dependiendo del lenguaje.

La forma de citar documentos, páginas web y datos se realizará como aconseja la UOC en su artículo publicado ‘[Citación bibliográfica’](http://biblioteca.uoc.edu/es/recursos/recurso/mendeley-guia-para-la-citacion-de-social-media) . Se usará la extensión de Chrome y el complemento de Word de Mendeley para el estándar IEEE.

A continuación, se presenta el Diagrama de Gantt con las tareas y los plazos que se desean realizar, se deja la semana final libre como tiempo a modo de contingencia, los plazos son asignados según los tiempos de las entregas de cada una de las PEC:



## 1.5 Breve sumario de productos obtenidos

La entrega final del proyecto incluye los elementos:

* Comparación de modelos de Machine Learning de reconocimiento de lenguaje tóxico
* Repositorio Github
* Memoria del proyecto
* Presentación del trabajo
* Librería de palabras tóxicas en español
* Corpus de tweets en español expresando emociones
* Video con una breve explicación del trabajo realizado

## 1.6 Breve descripción de los otros capítulos de la memoria

A continuación, se realizará la descripción de cada una de las fases de diseño, implementación, conclusiones y trabajo futuro.

# 2. Estado de los modelos de análisis de sentimientos.

En la actualidad encontramos numerosas librerías cuya intención es la de reconocer los sentimientos en textos escritos en la lengua inglesa. Sin embargo, no encontramos ninguno que haya sido especialmente desarrollado para la lengua española, uno de los objetivos de este trabajo es testear el comportamiento de los numerosos estudios que hay sobre lenguaje tóxico sobre tweets expresados en español.

Dentro de los sentimientos, son de especial interés aquellos que desencadenan situaciones de ira o de acoso producto de un sentimiento negativo hacia un tema o persona por ello se realizará un análisis sobre las librerías y metodologías que han sido desarrollados a día de hoy.

## 2.1. Librerias de análisis de sentimientos, modelos no supervisados.

Para la realización de las pruebas se ha construido una librería de lenguaje tóxico con 241 palabras tomadas del artículo ‘Insultos en castellano que deberías conocer y su significado’ [1] . Debido a que muchos estudios con el fin de llegar a un mayor público son realizados en inglés, en numerosas ocasiones las librerías con versión en castellano son traducciones, con esto se quiere comprobar si palabras que no son insultos comunes serían reconocidas por estas herramientas.

Se considera que frases como ‘soy lerdo’ siguen siendo tóxicas, aunque se utilicen con lenguaje jocoso. Queda fuera de este estudio el reconocimiento de la ironía, siendo cualquier palabra despectiva reconocida como tóxica independiente del contexto.

### 2.1.1 LIWC

Destacamos esta librería por ser uno de los modelos de referencia en multitud de tareas de análisis de texto, detección temprana de depresión, …

LIWC es un diccionario diseñado por humanos como sustitución al modelo de machine learning aunque en muchos estudios se utiliza como complemento a los modelos utilizando su resultado como tag de un algoritmo supervisado, convierte los textos en vectores. Sin embargo, en este proyecto, aunque se menciona porque se quiere utilizar el mismo enfoque haciendo una librería en castellano no se hará una prueba de concepto por ser de pago y no ser de utilidad en textos no formales al no incluir jergas, iniciales y emoticonos demostrado como importante en el análisis de sentimientos en textos sociales[2]. Por todo ello, lo que lo deja fuera de nuestro análisis[3].

### 2.1.2 VADER

Valence Aware Dictionary and Sentiment Reasoner es una librería de python que surgió de un modelo que mide no solo si un texto tiene sentimiento negativo o positivo sino también la intensidad de las emociones, tiene en cuenta los emoticonos, el argot, las mayúsculas, el número de signos de exclamación….El coeficiente de correlación (r=0.881)mostró como VADER se comportaba tan bien como el evaluado realizado por los humanos de forma manual (r=0.888).

Vader mantiene la vía del ‘sentiment lexicon’ como el LIWC o el basado en un glosario definido por humanos, pero está más adaptado a cualquier tipo de texto no formal, contiene listas con los emoticonos, palabras, capitalización de palabras, signos de puntuación positivos y negativos. La última de las diferencias es que es totalmente gratuito y recibe mejoras continuas de los usuarios [4].

Sin embargo, al estar basado en diccionarios definidos por humanos en español, al realizar la traducción no se comporta de forma satisfactoria:

Tabla 1 Ejemplos de funcionamiento Vader

|  |  |  |
| --- | --- | --- |
| tweet | traducción | puntuación |
| soy una mujer negra | am a black woman | {'neg': 0.0,  'neu': 1.0,  'pos':0.0,  'compound': 0.0} |
| zopenco | Stupidity | {'neg': 1.0,  'neu': 0.0,  'pos': 0.0,  'compound': -0.4404} |
| Soy cristiano | I'm Christian. | {'neg': 0.0,  'neu': 1.0,  'pos':0.0,  'compound': 0.0} |

Donde para ser aceptado positivo, negativo o neutro debe tener en ese campo un valor mayor que 0.5.. Sobre las 241 palabras de la librería anteriormente mencionada son reconocidas solamente 11.

Tabla 2 Resultados sobre 241 palabras tóxicas en castellano

|  |  |
| --- | --- |
| Número de palabras consideradas como negativas | 11 |
| Número de palabras consideradas como positivas | 3 |
| Número de palabras consideradas como neutras | 148 |

Una vez obtenidos los resultados con el traductor en el lenguaje español, se comprueba el funcionamiento en el corpus proporcionado por Kaggle

### 2.1.3 IBM Watson Sentiment Analyzer.

En este apartado se estudia una de las herramientas más utilizadas por las empresas hoy en día debido no a su precisión que se estudiará a continuación sino su versatilidad, IBM Watson

En 2014, IBM se asoció con Twitter obteniendo así toda la información que esta red social le podía proporcionar para su herramienta analítica. En contraste con VADER no se puede encontrar mucha información sobre el algoritmo o la estrategia que están utilizando.

. En la actualidad, está disponible en trece idiomas. En cuanto a la configuración, igual que en la API de Twitter hay que registrarse y explicar cual es el objetivo de tu proyecto para poder obtener el token necesario para hacer que el programa se comunique con su API, devuelve un resultado entre -1 y 1 que indica cuando de positivo (>0) o negativo (<0) es un comentario y dispone de un número mínimo de palabras para poder analizar la sentencia.

Han sido realizados dos experimentos:

1. Comprobación su rendimiento primero tanto con frases mal analizadas por la herramienta perspective como con riesgo de tener alto bias:

Tabla 3 Rendimiento de IBM Watson

Tabla 4 Análisis de BIAS en IBM Watson

|  |  |
| --- | --- |
| Soy una mujer gay negra | -0.647165 |
| Es que soy mujer! | 0.441942 |
| Es que soy hombre! | 0.501246 |
| Eres imbecil :) | 0.727783 |
| Eres imbécil :) | -0.897086 |
| Es que soy musulman | -0.50646 |
| Es que soy cristiano | 0 |

Queda comprobada la existencia de ese bias. Watson identifica como un mal sentimiento el ser mujer gay y negra, calcula ser hombre mejor que ser mujer y musulmán peor que cristiano.

1. Cuantas de las palabras tóxicas de la librería en español creada serían reconocidas por IBM Watson como realmente tóxicas al añadir “Cada tengo más claro que eres un”:

Tabla 5 Análisis de reconocimiento de toxicidad

|  |  |
| --- | --- |
| Número total de palabras en la librería | 241 |
| Palabras etiquetadas como positivas (Falso Positivo) | 58.5% |
| Palabras etiquetadas como negativas (Positivo) | 41.49% |

### 2.1.4 TextBlob

´

Librería de python con multitud de funciones como tokenización, corrección de ortografía y análisis de sentimiento. Devuelve dos valores polaridad que puede ir de -1 a 1 y subjetividad que va de 0 a 1 donde uno es muy subjetivo.

Se realizan las mismas pruebas que con IBM Watson

1. Análisis de las frases que han sido influenciadas por el dataset en el modelo de perspective:

Tabla 6 Análisis de Bias en TextBlow

|  |  |
| --- | --- |
| Soy una mujer gay negra | 0.416 |
| Es que soy mujer! | 0.0 |
| Es que soy hombre! | 0.0 |
| Eres imbecil :) | 0.5 |
| Eres imbécil :) | 0.5 |
| Es que soy musulman | 0.0 |
| Es que soy cristiano | 0.0 |

Se pesan también los emoticonos, obtiene los mismos resultados para las palabras con y sin acentos y no pone peso a ninguna condición.

1. A continuación, se observa si reconoce las palabras tóxicas del lenguaje castellano.

Tabla 7 Análisis de reconocimiento de toxicidad

|  |  |
| --- | --- |
| Número total de palabras en la librería | 241 |
| Palabras etiquetadas como positivas (Falso Positivo) | 0% |
| Palabras etiquetadas como negativas (Positivo) | 0% |

Se concluye que aunque TextBlow tenga un comportamiento adecuado a la hora de evaluar emoticonos y condiciones humanas no es capaz de reconocer el lenguaje tóxico en castellano.

## 3. Descripción del dataset del challenge Jigsaw Unintended Bias in Toxicity Classification

En el 2017 la plataforma llamada ‘Civil comments’ cerró su soporte y puso a disposición pública todos los comentarios que habían obtenido durante los años que estuvieron en el mercado el fruto fue este dataset.

Civil comments fue una plataforma creada como método de moderación en páginas donde se quería saber si un comentario consistía en un ataque hacia un colectivo/persona, cada vez que alguien escribía un comentario este era puntuado por el resto de usuarios antes de ser publicado. Surgió como filtro ante comentarios obscenos o de acoso pero las empresas quitaron sus foro en vez de invertir en su plataforma con lo que tuvieron que cerrar por falta de beneficios[5].

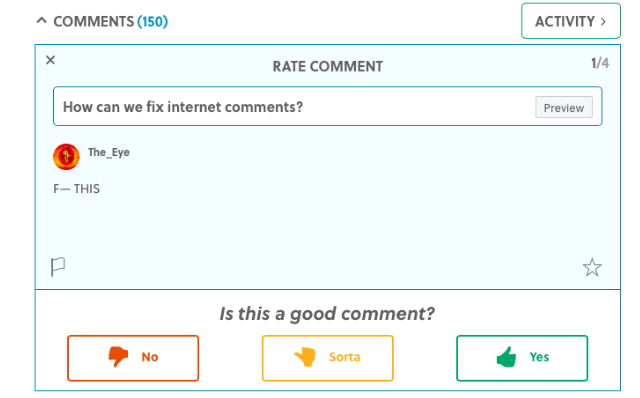


Ilustración 3. Pantalla usada po la plataforma Civil Comments

Google Ideas (**Jigsaw**) decidió utilizar este dataset y junto con ‘The times’ se propusieron lanzar una herramienta de machine learning que pudiera clasificar los comentarios en tiempo real, hasta entonces the times disponía de un equipo que censuraba manualmente[6]. El año pasado lanzaron un challenge en Kaggle que terminó en la herramienta Perspective desarrollada con Tensorflow, Sin embargo, debido a los comentarios que tuvieron en cuenta algunas frases no fueron bien clasificadas y lanzaron este año un nuevo reto cuyo contenido es el utilizado en este proyecto y que pretende eliminar la desviación producida por el ‘bias’.

## 3.1 Descripción de los datos del dataset

Nos Descargamos los datos del challenge (<https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data>) y nos encontramos con 3 archivos de datos.. Para hacer una inspección inicial de los datos utilizamos bash:

1. sample\_submission.csv

Este archivo muestra cómo debemos subir los archivos en la plataforma kaggle para que nuestro algoritmo sea analizado. Será el id del comentario con el grado de toxicidad que hayamos calculado.. Las columnas tendrán los nombre id

$ head -2 sample\_submission.csv

id,prediction

7000000,0.0

1. test.csv

Este documento es el núcleo del análisis donde tenemos el campo ‘id’ con el identificador del comentario y el campo ‘comment\_text’ con el contenido del comentario analizar.

1. train.csv

$ head -2 test.csv

id,comment\_text

7000000,Jeff Sessions is another one of Trump's Orwellian choices. He believes and has believed his entire career the exact opposite of what the position requires.

En este archivo ya se ha formado el archivo que será insertado en el procesamiento, tiene 45 características o columnas

$ head -2 train.csv

id,target,comment\_text,severe\_toxicity,obscene,identity\_attack,insult,threat,asian,atheist,bisexual,black,buddhist,christian,female,heterosexual,hindu,homosexual\_gay\_or\_lesbian,intellectual\_or\_learning\_disability,jewish,latino,male,muslim,other\_disability,other\_gender,other\_race\_or\_ethnicity,other\_religion,other\_sexual\_orientation,physical\_disability,psychiatric\_or\_mental\_illness,transgender,white,created\_date,publication\_id,parent\_id,article\_id,rating,funny,wow,sad,likes,disagree,sexual\_explicit,identity\_annotator\_count,toxicity\_annotator\_count

59848,0.0,"This is so cool. It's like, 'would you want your mother to read this??' Really great idea, well done!",0.0,0.0,0.0,0.0,0.0,,,,,,,,,,,,,,,,,,,,,,,,,2015-09-29 10:50:41.987077+00,2,,2006,rejected,0,0,0,0,0,0.0,0,4

,jewish,latino,male,muslim,other\_disability,other\_gender,other\_race\_or\_ethnicity,other\_religion,other\_sexual\_orientation,physical\_disability,psychiatric\_or\_mental\_illness,transgender,white,created\_date,publication\_id,parent\_id,article\_id,rating,funny,wow,sad,likes,disagree,sexual\_explicit,identity\_annotator\_count,toxicity\_annotator\_count

59848,0.0,"This is so cool. It's like, 'would you want your mother to read this??' Really great idea, well done!",0.0,0.0,0.0,0.0,0.0,,,,,,,,,,,,,,,,,,,,,,,,,2015-09-29 10:50:41.987077+00,2,,2006,rejected,0,0,0,0,0,0.0,0,4

## 3.2 EDA

Exploratory Data Analysis es el proceso en el que se realizan las primeras investigaciones en los dataset para detectar anomalías, descubrir patones, realizar pruebas que ayuden en la realización del modelo futuro , mostrar resultados de forma gráfica y encontrar posibles correlaciones[7].

Se empieza analizando lo más básico del dataset como es de cuantas columnas (tags) dispone y el tamaños del dataset:

Tabla 8 Tamaño de los dataset

|  |  |
| --- | --- |
| Dimensiones del dataset de entrenamiento | 1804874 |
| Dimensión deldataset de test | 97320 |
| Columnas del archivo de entrenamiento: | Index(['id', 'target', 'comment\_text', 'severe\_toxicity', 'obscene',  'identity\_attack', 'insult', 'threat', 'asian', 'atheist', 'bisexual',  'black', 'buddhist', 'christian', 'female', 'heterosexual', 'hindu',  'homosexual\_gay\_or\_lesbian', 'intellectual\_or\_learning\_disability',  'jewish', 'latino', 'male', 'muslim', 'other\_disability',  'other\_gender', 'other\_race\_or\_ethnicity', 'other\_religion',  'other\_sexual\_orientation', 'physical\_disability',  'psychiatric\_or\_mental\_illness', 'transgender', 'white', 'created\_date',  'publication\_id', 'parent\_id', 'article\_id', 'rating', 'funny', 'wow',  'sad', 'likes', 'disagree', 'sexual\_explicit',  'identity\_annotator\_count', 'toxicity\_annotator\_count'],  dtype='object') |

La columna target es la que el modelo debe predecir. Este campo representa la fracción de personas que creyeron que ese comentario era tóxico (>=0.5). Existen 1660540 etiquetados como no tóxicos y 97320 como tóxicos, por lo que se concluye que la mayoría de los campos no son tóxicos

.

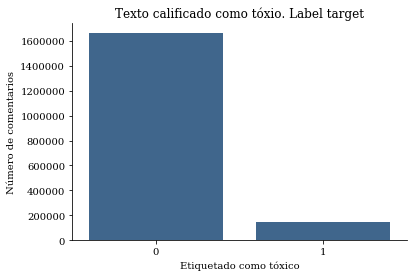


Ilustración 4 Cantidad de comentarios tóxicos

El 98.66% del dataset es de comentarios únicos, 24051 de los comentarios aparecen más de una vez, aparecen 24051 comentarios repetidos por lo que únicos solo tenemos el 98.66% del dataset.

En cuanto a la longitud de los comentarios la mayoría contienen entre 5 y 15 palabras a analizar, aunque puede haber hasta 205.

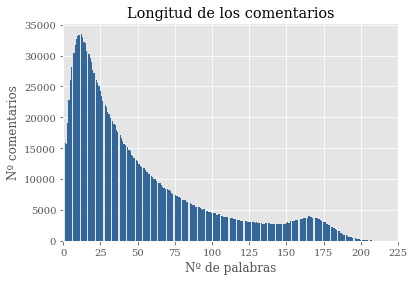


Ilustración 5 Número de palabras por comentario

Por último, se analiza una de las cosas más importantes a la hora de entrenar un dataset, si el dataset no está balanceado. Un dataset no está balanceado cuando las clases no están representadas de la misma forma.

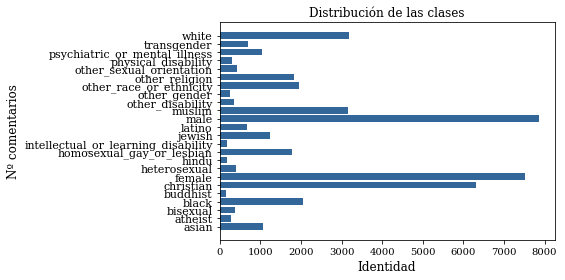


Ilustración 6 Distribución de los comentarios tóxicos

Efectivamente el dataset no está balanceado presenta categorías con una diferencia mayor de 4:1 en cuanto al volumen presente de comentarios etiquetados. En este caso, no se debe tomar el grado de acierto o exactitud como medida de evaluación del modelo. Se debe tomar alguna de las siguientes[8]:

* Matriz de confusión: Que muestra la cantidad de positivos falsos positivos, negativos y falsos negativos de modelo en forma de tabla.



Ilustración 7 Matriz de confusión

* Precisión: Cantidad de positivos reales entre todos los positivos detectados.

* Recall: La cantidad de positivos detectados entre todos los positivos del dataset.

* F1 score o F score: Media entre las medidas precisión y recall, su valor perfecto es 1 y el pero el 0.

* Curva ROC: Es la representación del ratio de verdaderos positivos contra el ratio de falsos positivos o lo que es lo mismo la representación del recall contra (1-especificidad)

Por último, se ha analizado el rango temporal en el que aparecen más comentarios tóxicos dirigidos a un grupo u otro de las identidades

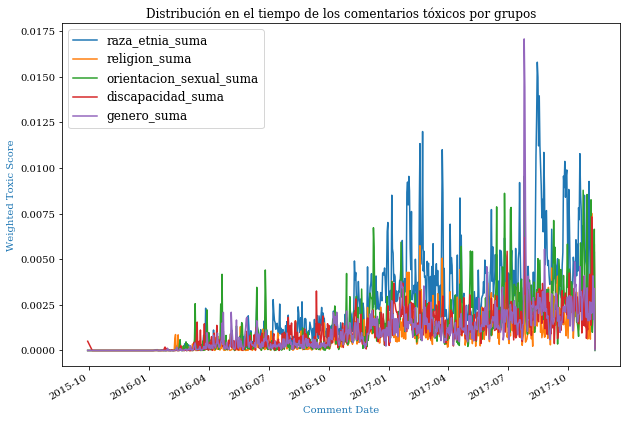


Ilustración 8 Distribución temporal por grupo de identidades

Con esto podemos analizar en que rango temporal fueron obtenidos estos datos y cuando un grupo fue atacado más que otro. Además, si fuera necesario podría ser el corte para balancear los datos y tomar un rango de tiempo en el que se tiene aproximadamente el mismo número de comentarios para cada uno de estos grupos.

# 4. Descripción del dataset de twitter

## 4.1 Creación del corpus de twitter

Debido a que no se ha encontrado un dataset en español con lenguaje tóxico ha sido necesaria la realización de uno propio.

El desarrollo para la obtención se ha realizado con Python con llamadas a la API de Twitter, para poder obtener las credenciales de autenticación que posibilitan su uso se ha de justificar y rellenar numerosos campos en los que el usuario indica cual es el propósito de esa extracción de datos.

El código se puede obtener en el repositorio y en los anexos.

## 4.2 Preprocesamiento de los datos de twitter

Como se muestra en el estudio realizado por CrowdFlower, el 80% del tiempo de un científico de datos, se emplea en la limpieza, procesamiento y análisis de los datos, un 20% en el estudio del modelo que se empleará en la realización del algoritmo.

Las buenas prácticas dividen este paso en cuatro pasos que nos ayudan a entender la importancia y naturaleza de cada uno de los campos de nuestros datos:

1. Limpieza de los datos: Eliminación de caracteres sin información como son los signos de puntuación, la estandarización de los datos en un mismo formato como es la eliminación de mayúsculas (teniendo especial cuidado con las palabras como US, us que cambian el significado al cambiar su morfología) y la sustitución de contracciones por sus palabras completas.
2. Esquematización de los datos: Etiquetación, división de los datos en campos y nombres que el algoritmo pueda entender y utilizar.
3. Normalización: Uso de técnicas como son las que convierten las palabras derivadas a su raíz (escribiendo se dividirá en raíz: escribir y sufijo: -endo ) o el uso de diccionarios con el significado morfológico de cada palabra (la palabra escritura en la [RAE](https://dle.rae.es/?id=GKXHvDj) significa ‘Acción y efecto de escribir.’ el Lema devuelto será: escribir)
4. Análisis: Realización de estadísticas indicativas de los campos con más dependencias, los que no tienen valor para la realización posterior de un análisis de características.

## 4.2 Preprocesamiento de los datos de twitter

En el corpus se han dejado fuera los retweets por no querer información duplicada, se ha activado la opción extended para poder obtener los nuevos tweets, con más contenido que en el pasado.

A continuación, se realiza como hemos comentado la limpieza de los datos, debido a que en redes sociales hay contracciones, emoticonos…,El proceso de limpieza ha sido largo hasta poder sacar información de valor. Se han eliminado las cuentas a las que hace referencia (@XXXX), aglutinado múltiples espacios en uno, eliminación de las palabras que comiencen por \_ y eliminación del símbolo #, así como de las comillas que no aportaban información en el análisis.

En cuanto a la esquematización de los datos. Además de la etiqueta que indica si un comentario es o no tóxico, se añaden indicativos de énfasis como el número de admiraciones, de emoticonos, número de palabras en mayúsculas , cantidad de palabras en mayúsculas. La mayoría del lenguaje tóxico está dirigida hacia otras personas por lo que también se ha creado un label indicando que el comentario está dirigido a una tercera persona. La última de las etiquetad creadas se realiza con la ayuda del artículo ‘Emoji Sentiment Ranking’ [9] donde hemos extraído otra librería con los emoticonos que aparecen en este corpus y el nivel de sentimiento negativo que tenía cada uno de ellos.

Para la normalización solo ha sido necesario el poner tanto el texto a analizar como la librería con el mismo formato, en minúsculas y sin tildes Esto se hace debido a que en las redes sociales cada vez más no se respetan las reglas de ortografía. Con el objetivo de no confundir palabras como moño con mono, la ñ se mantiene en el texto.

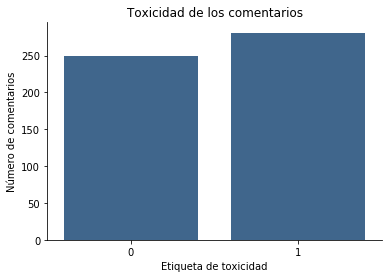
En cuanto a los resultados, finalmente fue construido un corpus con 530 comentarios tóxicos y no tóxicos obtenidos de twitter en donde la información que nos interesaba era si eran o no eran tóxicos, las etiquetas fueron añadidas manualmente por lo que no hay ningún elemento a nulo en esas dos columnas iniciales.

Tabla 9 Nuevas columnas del dataframe de Twitter

|  |  |
| --- | --- |
| C:\Users\Esperanza\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\D8DD2D3A.tmp | C:\Users\Esperanza\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\C10D1958.tmp |
| C:\Users\Esperanza\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9A1AC126.tmp | |

Mediante estas representaciones se ha querido observar si había una población realmente representativa de cada una de las categorías tomadas como medios de exaltación de sentimientos y efectivamente se determina que la hay.

El siguiente paso es igual que con el anterior dataset, determina si el corpus está balanceado para la futura toma de decisiones, para que sea considerado como no balanceado debe de haber una diferencia mayor de 4:1 algo que no ocurre como se puede ver en la figura siguiente:



Para analizar las principales palabras que aparecen y hacer un resumen de los datos se utiliza un wordcloud. Este método proporciona una forma visual de ver las palabras más relevantes para cada una de las categoría, en el caso de los comentarios categorizados como tóxicos

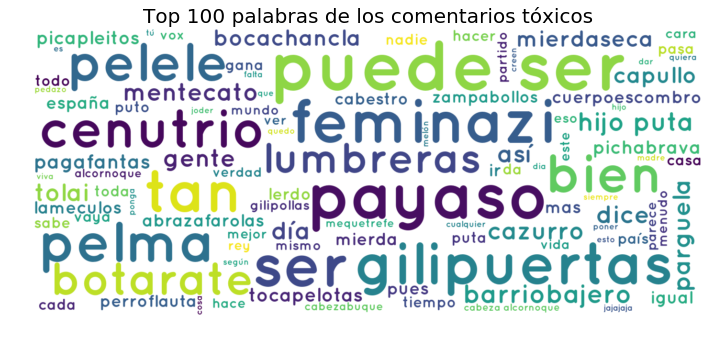


Ilustración 9 100 palabras consideradas como tóxicas



Ilustración 10 100 palabras consideradas como no tóxicas

Con estas dos ilustraciones se observa como sin ninguna intención no solo aparecen palabras que se consideran como tóxicas independientemente del contexto, sino que también hay alguna contextualizada por el tiempo en el que se amplió este corpus para unos mejores resultados.

Además, se utilizaron técnicas de análisis que no llevaron a conclusiones relevantes como fue el estudio de la relación entre palabra conocido como Bigram, este busca en todo el texto manejado cuales son el grupo de palabras que más se repiten alrededor de las demás.

Tabla 10 Resultados Bigram

|  |  |
| --- | --- |
| ('no', 'se', 'puede') | 12 |
| ('se', 'puede', 'ser') | 12 |
| ('puede', 'ser', 'más') | 10 |
| ('', '', '') | 10 |
| ('data', 'science', '') | 4 |
| ('', 'ofertas', 'trabajo') | 3 |

El siguiente factor tomado en el análisis es el del Coeficiente de correlación de Pearson, coeficiente de correlación más usado para medir la asociación entre variables. Cuantifica el grado en el que la relación entre dos variables puede ser descrita por una línea, su formulación es la siguiente[10]:

Los resultados fueron los esperados, el número de insultos efectivamente tiene una alta relación con que el comentario sea tóxico y el nivel de toxicidad depende en gran medida con el número de emoticonos a pesar de que se haga una media entre sus valores, a mayor número de emoticonos más expresividad se le añade al comentario.

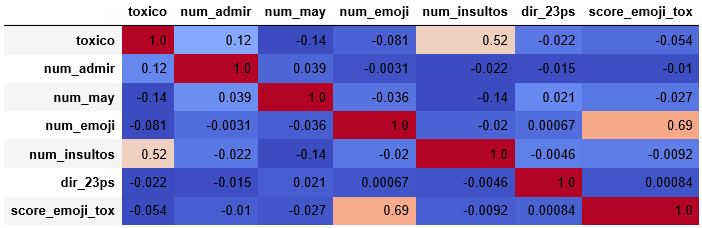


Ilustración 11 Coeficiente de correlación de Pearson

# 5. Estudio de modelos de Machine Learning

Una vez analizado el desempeño de las técnicas presentes en el mercado se decide investigar sobre la construcción de modelos de Deep learning que puedan mejorar estos resultados.

## 5.1 Pasos iniciales

Step1: Get data.

Step 2: Generate embeddings

Step 3: Model architecture

Step 4: Model Parameters

Step 5: Train and test the model

Step 6: Run the model

Word embedding es una técnica de ingeniería de características donde cada palabra es representada como un vector. Generalmente se utiliza para que palabras con significados similares sean agrupadas juntas en el hiper plano y las opuestas en una posición lejos de este.

# 3. Conclusiones

Este capítulo tiene que incluir:

* Una descripción de las conclusiones del trabajo: ¿Qué lecciones se han aprendido del trabajo?.
* Una reflexión crítica sobre el logro de los objetivos planteados inicialmente: ¿Hemos logrado todos los objetivos? Si la respuesta es negativa, por qué motivo?
* Un análisis crítico del seguimiento de la planificación y metodología a lo largo del producto: ¿Se ha seguido la planificación? ¿La metodología prevista ha sido la adecuada? Ha habido que introducir cambios para garantizar el éxito del trabajo? ¿Por qué?
* Las líneas de trabajo futuro que no se han podido explorar en este trabajo y han quedado pendientes.

# 4. Glosario

Definición de los términos y acrónimos más relevantes utilizados dentro de la Memoria.

# 5. Bibliografía

[1] «Lista: Insultos en castellano que deberías conocer y su significado». [En línea]. Disponible en: https://listas.20minutos.es/lista/insultos-en-castellano-que-deberias-conocer-y-su-significado-393340/. [Accedido: 10-abr-2019].

[2] D. Davidov, O. Tsur, y A. Rappoport, «Enhanced Sentiment Learning Using Twitter Hashtags and Smileys», 2010.

[3] Y. Gong, C. Poellabauer, y K. Shin, «Improving LIWC Using Soft Word Matching Speech and Language Processing for Health View project Improving LIWC Using Soft Word Matching», 2018.

[4] C. J. Hutto y E. Gilbert, «VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text», 2014.

[5] Bogdanoff Aja, «Saying goodbye to Civil Comments», *Medium*, 2017.

[6] «The Times is Partnering with Jigsaw to Expand Comment Capabilities», *New York Times Co.*, 2016.

[7] Prasad Patil, «What is Exploratory Data Analysis? – Towards Data Science». [En línea]. Disponible en: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15. [Accedido: 28-may-2019].

[8] N. V. Chawla, «Data Mining for Imbalanced Datasets: An Overview», *Data Mining and Knowledge Discovery Handbook*, 2009. [En línea]. Disponible en: http://link.springer.com/10.1007/978-0-387-09823-4\_45. [Accedido: 29-may-2019].

[9] «Emoji Sentiment Ranking v1.0». [En línea]. Disponible en: http://kt.ijs.si/data/Emoji\_sentiment\_ranking/. [Accedido: 05-jun-2019].

[10] Ruslana Dalinina, «Introduction to Correlation», *Oracle + DataScience.com*, 2017.

<https://civic.mit.edu/2019/01/24/how-automated-tools-discriminate-against-black-language/>

<https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjHufrv87nhAhXsx4UKHdfYBqsQjhx6BAgBEAM&url=http%3A%2F%2Famericablog.com%2F2014%2F06%2Fonline-harassment-study-facebook-worst-women-targeted-men.html&psig=AOvVaw3RE9Fk7vhmRzLbfYom2uUr&ust=1554586334597939>

<https://www.perspectiveapi.com/#/>

# 6. Anexos