Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN

ANÁLISIS DE DATOS MASIVOS (BIG DATA)

TRABAJO FIN DE MÁSTER

DETECCIÓN DE CYBERBULLYING CON NLP

NOMBRE:

MATÍAS MISAEL RACCA

CURSO 2021-2022

TÍTULO: DETECCIÓN DE CYBERBULLYING CON NLP

AUTOR: MATÍAS MISAEL RACCA

TITULACIÓN: MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS (BIG DATA)

DIRECTOR DEL PROYECTO: JOSÉ JAVIER RUIZ COBO

TUTOR DEL PROYECTO: NICOLÁS COCA LÓPEZ

FECHA: OCTUBRE de 2022

RESUMEN

Este trabajo final se encuentra enfocado en la aplicación de Natural Language Processing (NLP) como herramienta base para la detección del cyberbullying. El estudio comprende diferentes técnicas de preparación de los datos, selección de características (feature extractions) y análisis de modelos con aprendizaje supervisado. Como conclusión, en comparación con Logistic Regression, Support Vector Classifier, redes neuronales LSTM, entre otros predictores, se obtiene que Random Forest Classifier es el modelo óptimo en tareas de clasificación de texto.

Palabras clave: NLP, cyberbullying, feature extraction, modelos, clasificadores, redes neuronales, accuracy, predicción

AGRADECIMIENTOS

A Dios

A mi Familia

A mi Universidad

A mis Docentes

A mis Compañeros

A mis Amigos

Mira en tu interior, contempla. Mira a tu alrededor, contempla.

¿Puedes ver tu reflejo?

Índice

Capítulo 1. INTRODUCCIÓN

1.1 Tema

Las redes sociales virtuales son, desde hace ya muchos años, plataformas informáticas en las cuales las personas pueden interactuar compartiendo diferentes tipos de contenido, desde un simple comentario, hasta archivos multimedia de todo tipo como textos, fotografías, videos o sonidos. Es un entorno que continuamente capta la atención de millones de usuarios porque, además de ofrecer entretenimiento ilimitado, el acceso, pertenencia y uso de cualquiera de ellas es extremadamente sencillo y cautivante.

Sin embargo, además de lidiar con asuntos de seguridad informática, éstas también se han convertido en un ambiente propicio para la práctica del "Ciberacoso" o "Cyberbullying" en inglés, convirtiéndose en un gran problema, especialmente, entre los grupos sociales más vulnerables por ser considerados "débiles" o con mayor riesgo de marginación.

Este proyecto final está enfocado en el análisis y desarrollo de un modelo automático predictivo en el que, evaluando diferentes algoritmos y utilizando datasets con contenido textual de acoso, intimidación o discriminación, se logre entrenar de manera óptima para que pueda detectar, con gran eficacia, comentarios con finalidad inapropiada y tendencia nociva socialmente.

1.2 Situación problemática

El cyberbullying se ha convertido en un inconveniente a nivel mundial que, si bien involucra a todos los usuarios independientemente de la edad que posean o país en el que se encuentren, afecta principalmente a niños y adolescentes. Durante el año 2020 un estudio en Estados Unidos indicó que el 44% de usuarios de internet de dicho país había experimentado acoso cibernético, siendo la utilización de nombres ofensivos el más habitual[1].

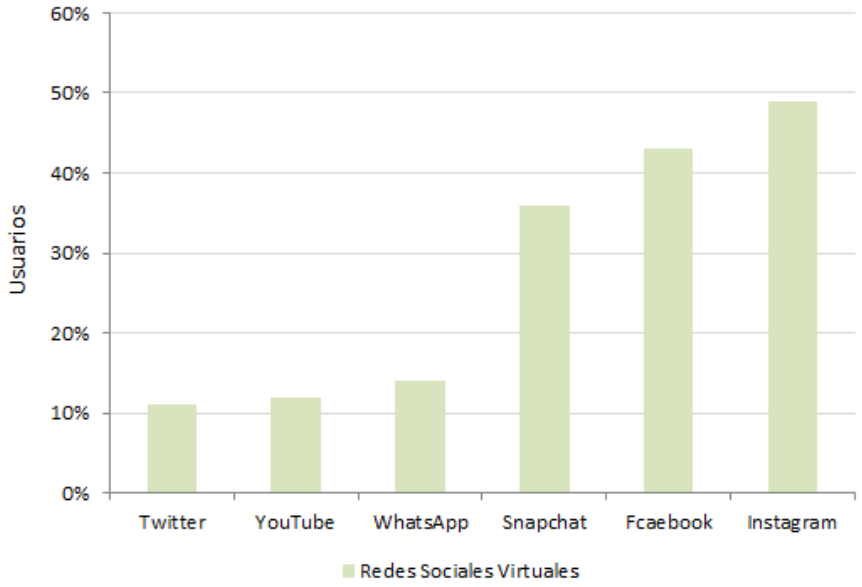


Figura 1. Porcentaje de usuarios que han sufrido cyberbullying en redes sociales virtuales [elaboración propia][2].

El cyberbullying se define como[3] el uso de herramientas de comunicación electrónicas (redes sociales virtuales, blogs, e-mails) para dañar o acosar a una persona o grupo de personas. Son acciones deliberadas y continuas, provocando

[1] Security. *Cyberbullying: Twenty Crucial Statistics for 2022* [en línea]. EE. UU.: Security, Agosto 2022. Disponible en: <https://www.security.org/resources/cyberbullying-facts-statistics/#impacts>

[2] Ditch The Label. *2017 Bullying Statistics – The Annual Bullying Survey 2017* [en línea]. Reino Unido: Ditch The Label, 2017. Disponible en: <https://www.ditchthelabel.org/research-papers/the-annual-bullying-survey-2017/>

[3] ACI, C., SARAC, E., YILDIZ, E. 2019. **Automatic Detection of Cyberbullying in FormSpring.me, MySpace and YouTube social networks.** Turquía, Turkish Journal of Engineering. Pág. 1.

agresión hacia individuos más vulnerables. El contenido de estas acciones puede ser muy diferente, siendo las más comunes:

- Sexualidad
- Diferencias de género
- Discapacidad
- Racismo / Etnias
- Terrorismo
- Carácter personal y comportamientos
- Apariencia física
- Creencias / Religión

Figura 2. Motivos del cyberbullying [elaboración propia][4].

A su vez, es importante destacar que, según la madurez emocional y fortaleza de autoestima, el individuo discriminado puede verse seriamente afectado, provocando riesgos de aislamiento, rechazo, lesiones y daños a sí mismo, como

[4] Enough is Enough. *Cyberbullying Statistics* [en línea]. EE. UU.: Enough is Enough, 2022. Disponible en: <https://enough.org/stats_cyberbullying>

hacia otras personas. La razón principal por la que el acoso puede ser tan emocionalmente y psicológicamente dañino es por su característica reiterativa[5].

Además, la pandemia COVID-19 ha mostrado evidencia de una práctica creciente de este problema, siendo los más jóvenes los principales perjudicados, dedicando gran cantidad de tiempo (académico y de ocio) a estas aplicaciones. Inconvenientes psicológicos (depresión, ansiedad, ideas suicidas) como desórdenes psicosomáticos (dolores de cabeza, fatiga) son algunos de los síntomas que sus víctimas pueden desarrollar[6].

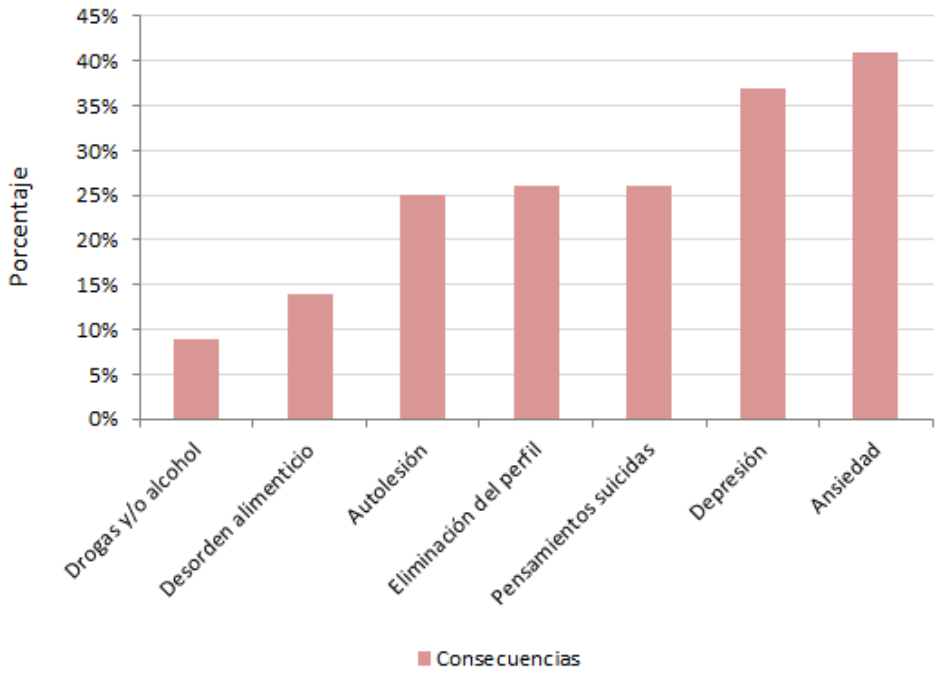


Figura 3. Consecuencias del cyberbullying [elaboración propia][7].

Un estudio realizado en la primavera del año 2020 por la Comisión Europea a 11 países de la Unión Europea, siendo España uno de los evaluados, arrojó que el 44%

[5] Cyberbullying Research Center. *What is Cyberbullying?* [en línea]. EE. UU.: Cyberbullying Research Center. Disponible en: <https://cyberbullying.org/what-is-cyberbullying>

[6] GOMEZ, C., SZTAINBERG, M., TRANA, R. 2021. **Curating Cyberbullying Datasets: a Human-AI Collaborative Approach.** EE. UU, International Journal of Bullying Prevention. Pág. 1.

[7] Ditch The Label. *2017 Bullying Statistics – The Annual Bullying Survey 2017* [en línea]. Reino Unido: Ditch The Label, 2017. Disponible en: <https://www.ditchthelabel.org/research-papers/the-annual-bullying-survey-2017/>

de niños que ya habían sufrido cyberbullying antes de la pandemia, notaron un incremento durante el confinamiento[8]:

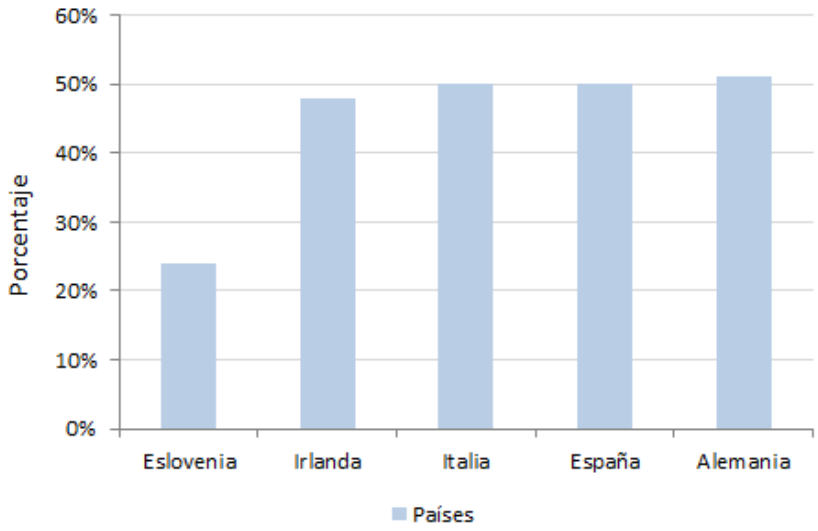


Figura 4. Crecimiento del cyberbullying en niños durante la pandemia [elaboración propia].

El acoso cibernético es un tema de gran importancia que merece especial foco de atención, considerando el notable tiempo mundialmente dedicado a utilizar aplicaciones de intercambio multimedia en internet.

Desde el lado informático, y con soporte firme de las plataformas en cuestión, aportar herramientas automáticas de análisis del lenguaje natural que permitan su rápida detección y abordaje, permitirá no sólo fomentar un entorno online socialmente más seguro e integrador, sino también, potencialmente libre de sesgos y conductas discriminatorias.

[8] Lobe, B., Velicu, A., Staksrud, E., Chaudron S., Di Gioia, R. *How children (10-18) experienced online risks during the Covid-19 lockdown - Spring 2020* [en línea]. EE. UU.: Publications Office of the European Union, 2021. Disponible en: <https://publications.jrc.ec.europa.eu/repository/handle/JRC124034>

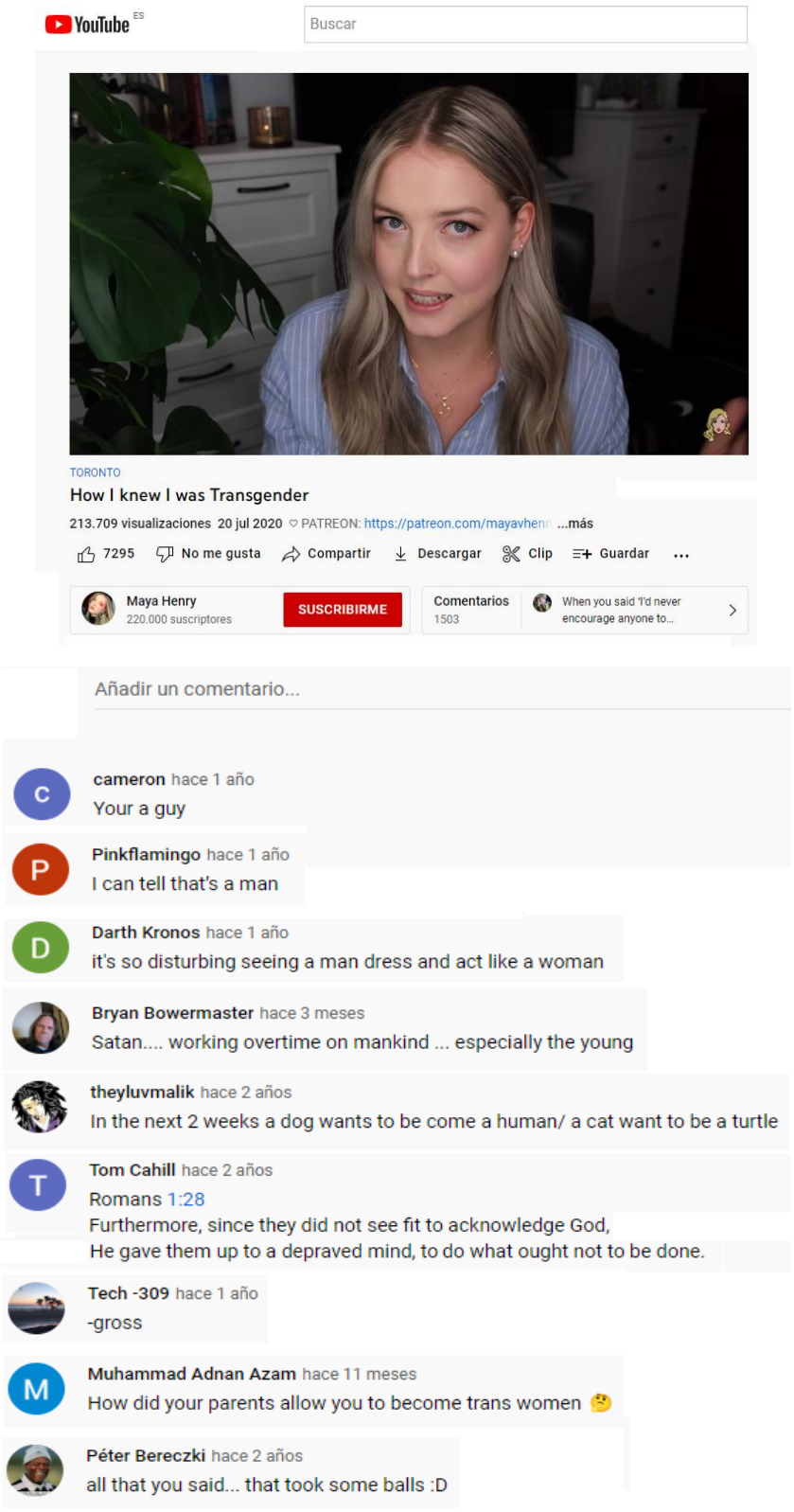


Imagen 1. Canal de YouTube "Maya Henry"[9].

[9] Henry, Maya. *How I knew I was Transgender* [en línea]. Canadá: YouTube, Julio 2020. Disponible en: <https://www.youtube.com/watch?v=iyeo5P6sWk0>

1.3 Objetivos

Como consecuencia de lo expuesto en el apartado anterior, el objetivo de este trabajo es intentar detectar y erradicar este comportamiento mediante la utilización del Procesamiento del Lenguaje Natural o "Natural Language Processing" como herramienta de análisis y proceso de textos, permitiendo transformar el lenguaje humano y haciéndolo "entendible" (accesible) por los ordenadores.

Para su implementación, se aplicará la técnica denominada "Análisis de sentimiento" que servirá como método base para arremeter la problemática que motiva esta investigación y da fin al proyecto. Así, se desarrollará un procedimiento global que incluya:

- Preprocesamiento de datos iniciales: utilizando técnicas de limpieza mediante expresiones regulares para, por ejemplo, eliminar caracteres especiales (Love😳❤) o reducir la repetición de letras en las palabras (goooood).
- Preparación de los datos: empleando los métodos de tokenización y lematización junto a las "stop words".
- Feature extraction: representando las palabras y su contexto de forma numérica aplicando "Bag of Words" y "Word Embedding" (considerando, para cada técnica, distintas variantes).
- Modelos de clasificación: Logistic Regression, Random Forest Classifier, Support Vector Classifier, Naive Bayes, Ada Boost Classifier (ensemble), Gradient Boosting Classifier, Stochastic Gradient Descent Classifier y Artificial Neural Networks (Multilayer Perceptron Classifier, Long Short Term Memory, Convolutional Neural Network). La elección de estos modelos se basa en las diversas funciones matemáticas y procedimientos que aplican, permitiendo evaluar el comportamiento de los datos en el entorno que cada algoritmo ofrece.

La combinación de todos estos procesos permitirá realizar un amplio análisis comparativo evaluando la capacidad y eficacia predictiva que cada uno podría ofrecer, de implementarse, en un entorno productivo.

1.4 Recursos utilizados

Para llevar a cabo la demostración del uso del NLP con la técnica de "Análisis de sentimiento" se utilizará:

- Jupyter Notebook como plataforma de desarrollo.
- Python como lenguaje de programación.
- Comentarios de Twitter como base de datos.

La base de datos de Twitter, generada por los académicos Jason Wang, Kaiqun Fu, Chang-Tien Lu para su trabajo "A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection"[10], está compuesta por comentarios clasificados en seis categorías distintas:

- Not cyberbullying: no es ciberacoso.
- Age: acoso relacionado a la edad.
- Ethnicity: acoso relacionado a la raza / etnia.
- Gender: acoso relacionado al sexo / género.
- Religion: acoso relacionado a la religión / creencias.
- Other cyberbullying: acoso general relacionado a otros aspectos.

	tweet_text	cyberbullying_type
0	In other words #katandandre, your food was cra...	not_cyberbullying
1	Why is #aussietv so white? #MKR #theblock #ImA...	not_cyberbullying
2	@XochitlSuckkks a classy whore? Or more red ve...	not_cyberbullying
3	@Jason_Gio meh. :P thanks for the heads up, b...	not_cyberbullying
4	@RudhoeEnglish This is an ISIS account pretend...	not_cyberbullying

Figura 5. Base de datos de Twitter [elaboración propia].

Es importante indicar que los datos utilizados están en idioma inglés, con lo cual, todo el análisis planteado y desarrollo realizado se encuentra enfocado a dicho idioma.

[10] Wang, J., Fu, K., Lu, CT. *SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection* [en línea]. EE. UU.: IEEE, 2020. Disponible en: <https://ieeexplore.ieee.org/document/9378065>

Capítulo 2. NATURAL LANGUAGE PROCESSING (NLP)

2.1 Definición y contexto

El Procesamiento del Lenguaje Natural o Natural Language Processing (NLP) es un área de investigación interdisciplinar que combina la inteligencia artificial y la lingüística en pos del desarrollo de programas de computadora capaces de procesar y entender discursos y textos del lenguaje humano[11]. Este es su objetivo fundamental.

Dadas sus características, esta técnica se encuentra relacionada con los siguientes conceptos[12]:

- Lingüística computacional: similar a NLP pero con fines diferentes, donde su principal objetivo es el estudio del lenguaje. En cambio, NLP está más focalizado en proporcionar nuevas funcionalidades relacionadas con el lenguaje humano.
- Aprendizaje automático: actualmente NLP se apoya en gran medida en sus algoritmos, permitiendo construir programas mucho más complejos partiendo de un conjunto de datos.
- Inteligencia artificial: para construir software con las mismas habilidades que poseen los humanos, donde la utilización del lenguaje una de las funcionalidades centrales de la inteligencia humana, NLP es un requerimiento clave para la IA.
- Ciencia de la computación: dada la naturaleza discreta y recursiva del lenguaje, su utilización en NLP permite aplicar ideas teóricas para explicar la sintaxis y semántica del mismo.
- Procesamiento digital de voz: para convertir una señal de audio en texto, el contexto del diálogo o discurso (conocimiento de las palabras) juega un

[11] GELBUKH, Alexander. 2007. **Special issue: Natural Language Processing and its Applications.** México, Instituto Politécnico Nacional, Centro de Investigación en Computación. Pág. 5

[12] EISENSTEIN, Jacob. 2019. **Introduction to Natural Language Processing.** EE. UU., The MIT Press. Pág. 1-5

papel crítico para minimizar el ruido. Aquí es donde se integra NLP con su funcionalidad de análisis de texto y, particularmente, modelos de lenguaje estadístico.

- Ética: dado que la presencia del aprendizaje automático, inteligencia artificial y NLP se ha vuelto crecientemente ubicua, entender sus beneficios y riesgos para las personas se ha vuelto crucial, principalmente en lo que respecta a ética, equidad y responsabilidad.
- Otros: NLP también se encuentra relacionada a otros campos como la ciencia social computacional, humanidades digitales, minería de textos, entre otras.

Es notoriamente sabido que la cantidad de datos que hoy en día se genera gracias a "Big Data" es recopilada y posteriormente procesada por las empresas para adquirir mayor información de la misma. La obtención de patrones e indicadores clave (KPIs) en dichos datos permite identificar su comportamiento y comprenderlos mejor. En relación a lo anterior, y teniendo en cuenta que muchas de estas empresas tienen como fuente principal de datos al lenguaje humano, la popularidad del procesamiento del lenguaje natural ha aumentado fomentando el descubrimiento de nuevos y mejorados algoritmos informáticos. Así, y desde hace ya muchos años, NLP se viene explorando y utilizando en diversas áreas y aplicativos[13]:

- Traducción automática: facilitando la comunicación de personas de todo el mundo, comenzó a recibir especial atención a principios de 1950. Es uno de los puntos que más desafíos poseen dadas las características de los lenguajes, sin embargo ha habido gran cantidad de progresos.
- Búsqueda y recuperación de información: los conocidos motores de búsqueda, herramienta utilizada por el 85% de los usuarios de internet para buscar información específica.
- Búsqueda de respuestas: proceso automático capaz de entender una pregunta formulada en lenguaje natural, devolviendo posteriormente su respuesta con la información solicitada.

[13] INDURKHYA, N., DAMERAU, F. 2010. **Handbook of Natural Language Processing.** EE. UU., Chapman & Hall/CRC. Pág. 425-661

- Extracción de la información: proceso que realiza un escaneado de textos para obtener información relevante de los mismos, extrayendo entidades, relaciones, eventos. Su análisis es mucho mayor al de sólo buscar palabras claves.
- Generador de informes: producción automática de algún material escrito solicitado por los usuarios. Por lo general, dicho material se produce a partir de datos estructurados como series temporales numéricas o una representación de contenido formal.
- Minería de texto biomédica: aplicaciones que procesan datos y tareas dentro del dominio biomédico, ofreciendo control de calidad, especialmente, a temas éticos dentro de este campo.
- Construcción ontológica: ingeniería del conocimiento ontológico apoyada principalmente en análisis de material textual.
- Análisis de sentimiento: procesamiento de información textual que expresa generalmente opiniones subjetivas (positivas o negativas) sobre los sentimientos y apreciaciones de las personas en relación a diferentes entidades y eventos.

Por otro lado, NLP también se ha estado implementando con gran énfasis en las áreas de educación y salud. Tal es así que, en el caso de organizaciones de atención médica, esta tecnología ha mejorado la prestación de cuidados a sus pacientes y avanzado en el diagnóstico de enfermedades:

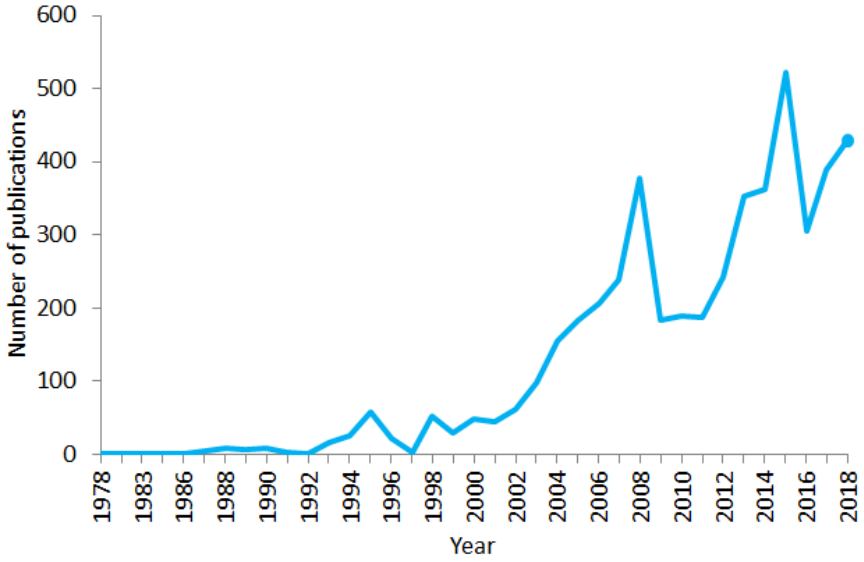


Figura 6. Nº de publicaciones con la frase "natural language processing" en PubMed[14].

Con el paso del tiempo, la investigación y evolución de NLP ha sido de indudable crecimiento. Los constantes avances en tecnología y el continuo interés en esta herramienta, han permitido progresar utilizando nuevos y diferentes enfoques, desarrollando software capaz de procesar e interpretar lenguaje natural "un poco" más cerca a como lo haría un ser humano.

[14] Lopez Yse, Diego. *Your Guide to Natural Language Processing (NLP)* [en línea]. Canadá: Towards Data Science, Enero 2019. Disponible en: <https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1>

2.2 Fundamentos base

El procesamiento del lenguaje natural es un conjunto de diferentes tareas y métodos que combinan lingüística e informática. Si bien en un principio podría creerse que la primera no sería tan necesaria en el desarrollo de aplicaciones NLP con Big Data (donde el aprendizaje se lograría directamente desde los datos), su conocimiento es importante para poder comprender cómo funciona un idioma, sus componentes, sus reglas y cómo éste es luego procesado computacionalmente. Además, aún se continúan desarrollando programas que permitan representar un lenguaje en estructuras generales "multiuso" para que luego sirvan como input en diversos procesos.

Muchos problemas de NLP se pueden representar matemáticamente en forma de optimización[15]:

$$\hat{y} = \underset{y \in \mathcal{Y}(x)}{\operatorname{argmax}} \Psi(x, y; \theta),$$

Dónde:

- x es la entrada, elemento del conjunto $\mathcal{X}$.
- y es la salida, elemento del conjunto $\mathcal{Y}(x)$.
- Ψ es la función de puntuación (el modelo) que mapea el conjunto $\mathcal{X} \times \mathcal{Y}$ a números reales.
- θ es un vector de parámetros de Ψ.
- $\hat{y}$ es la salida predicha, elegida para maximizar la función.

Luego, esta fórmula se puede aplicar a gran cantidad de situaciones como, por ejemplo, el análisis de sentimiento, donde "x" podría ser un comentario en alguna red social virtual e "y" la etiqueta clasificando la expresión de emoción del autor.

[15] EISENSTEIN, Jacob. 2019. **Introduction to Natural Language Processing.** EE. UU., The MIT Press. Pág. 7-8

De lo anterior se desprende que los algoritmos de NLP interactúan con dos tipos de módulos:

- Búsqueda: responsable de procesar la función "argmax", encontrando la salida que mayor puntuación obtenga respecto a la entrada "x". Como las salidas suelen ser valores discretos en problemas de NLP, este módulo se apoya en optimización combinatoria.
- Aprendizaje: responsable de encontrar los parámetros óptimos para la función "argmax". Como los parámetros suelen ser valores continuos, estos algoritmos se apoyan en optimización numérica, identificando vectores que optimicen el modelo.

El poder separar estos dos módulos permite la reutilización de algoritmos genéricos en diferentes tareas y modelos. Así, mucho trabajo del NLP puede enfocarse en el diseño del modelo (fenómeno lingüístico) mientras, al mismo tiempo, se beneficia del progreso de años en búsqueda, optimización y aprendizaje.

Sin embargo, también hay que tener en cuenta la expresividad de un modelo, es decir, que tan eficaz es para distinguir sutiles diferencias lingüísticas. La mayoría de los problemas de NLP requieren modelos expresivos, siendo éstos los de mayor complejidad y que más afectan el módulo de "búsqueda" (el tamaño de la entrada "x" suele ser muy grande y las búsquedas exactas son generalmente imposibles). En muchas ocasiones, este módulo precisa de utilizar un conjunto de aproximaciones heurísticas que mejoren el desempeño del modelo.

Desde el lado puramente lingüístico, hay que destacar los siguientes grandes campos relacionados estrechamente con NLP[16]:

- Gramática:
    - Morfología: estudia la estructura de las palabras y las partes que las constituyen, como los lexemas (raíces) y morfemas. Es decir, las

[16] TAHER PILEHVAR, M., CAMACHO-COLLADOS, J. 2018. **Embeddings in Natural Language Processing.** EE. UU., Morgan & Claypool. Pág. 10-11

reglas que rigen cómo se forman.

- Sintaxis: estudia la estructura, reglas y principios que indican cómo deben ordenarse las palabras dentro de las oraciones.

- Semántica: estudia el significado de las expresiones lingüísticas y sus combinaciones.

NLP lleva a cabo técnicas de proceso lingüístico que le permiten descomponer un texto en partes más pequeñas, pero siempre con el gran desafío de poder interpretar y respetar su función gramatical y significado semántico. El análisis de un texto bien escrito (respetando todas las reglas del idioma que corresponda) ya es un trabajo arduo. Además, lidiar con la ironía, subjetividad, entonación y utilización de negaciones en frases con sentido positivo, complejiza aún más las cosas. Sin embargo, hay que tener en cuenta que esto aún es más difícil con las plataformas virtuales de ocio donde, generalmente, priman la informalidad y las licencias gramaticales. Así, existen ciertas características adicionales que hay que sopesar en su análisis, como por ejemplo:

- errores ortográficos
- uso incorrecto de palabras homófonas (palabras que suenan igual pero se escriben diferentes)
- abreviaciones
- redundancia de letras
- jerga

Capítulo 3. TRABAJO DE CAMPO

3.1 Planteamiento del problema y análisis inicial

Para comenzar a trabajar sobre la generación de un clasificador capaz de detectar cyberbullying, el primer paso es la carga de los datos y entender completamente el dataset con el que se está trabajando.

En este caso, hablamos de un fichero que contiene dos variables:

- tweet_text: comentario realizado por un usuario.
- cyberbullying_type: clasificación del comentario. Las opciones son:
  - not_cyberbullying
  - age
  - ethnicity
  - gender
  - religion
  - other_cyberbullying

	tweet_text	cyberbullying_type
0	In other words #katandandre, your food was cra...	not_cyberbullying
1	Why is #aussietv so white? #MKR #theblock #ImA...	not_cyberbullying
2	@XochitlSuckkks a classy whore? Or more red ve...	not_cyberbullying
3	@Jason_Gio meh. :P thanks for the heads up, b...	not_cyberbullying
4	@RudhoeEnglish This is an ISIS account pretend...	not_cyberbullying

Figura 7. Base de datos de Twitter [elaboración propia].

Es un conjunto de datos compuesto por 2 columnas y, originalmente, 47.977 filas. Sin embargo, dado el sentido tan general que abarca, la clase “other_cyberbullying” se descarta por completo. Esto evitará afectar (innecesariamente) el entrenamiento de los modelos dado su contenido poco concreto y confuso. Así, se contemplan sólo las 40.134 filas y 5 clases restantes.

Por otro lado, también es importante comprobar la existencia de valores nulos que, en este caso, no se cumple:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 40134 entries, 0 to 47976
Data columns (total 2 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   tweet_text          40134 non-null  object
 1   cyberbullying_type  40134 non-null  object
dtypes: object(2)
memory usage: 940.6+ KB
```

Figura 8. Información general de los datos [elaboración propia].

Un aspecto clave al momento de trabajar con datos supervisados (es decir, datos de entrada ya etiquetados o clasificados), es verificar que las clases consideradas se encuentren balanceadas (que cada una posea una cantidad similar de ejemplos o registros):

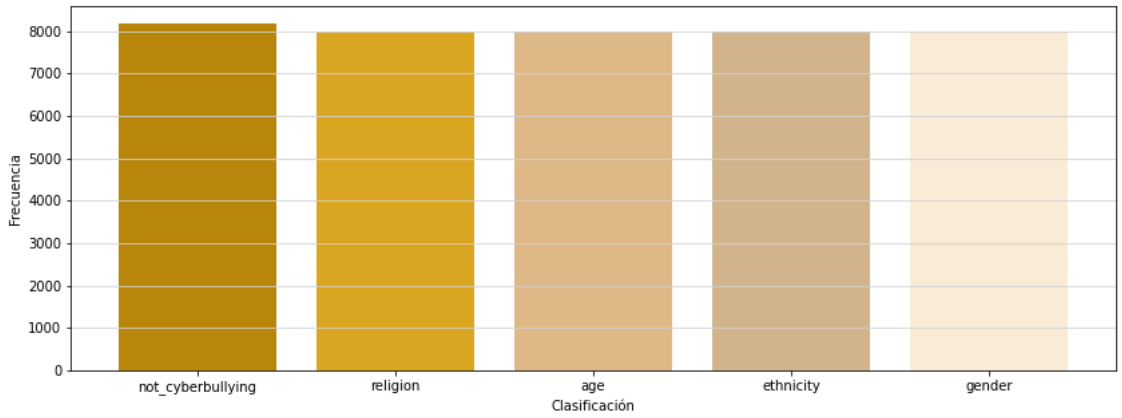


Figura 9. Gráfico de ejemplos por cada clase [elaboración propia].

Observando que esto se cumple, se puede continuar con el siguiente paso.

3.2 Preprocesado de los datos

Se observan algunos comentarios al azar para poder identificar qué técnicas de limpieza y preparación de datos podrían ser necesarias:

Figura 10. Comentarios originales de la base de datos [elaboración propia].

Según los comentarios anteriores, los tipos de preprocesados necesarios serían:

- Conversión del texto a minúsculas
- Descomposición de las contracciones típicas del idioma inglés
- Eliminación de etiquetados (hashtags)
- Eliminación de menciones y links
- Eliminación de números, caracteres especiales y de puntuación
- Eliminación de repetición de letras (dejar, como máximo, dos iguales seguidas)
- Normalización de espacios
- Eliminación de comentarios en idiomas distintos al inglés (son la minoría)

Los primeros se resuelven mediante la aplicación de expresiones regulares. Para el caso del idioma, se usa un detector que clasifica cada comentario según las palabras que lo componen. A continuación, se descartan todos aquellos distintos a “en” (english).

Luego de efectuados los cambios anteriores, los mismos comentarios se observan ahora de la siguiente manera:

```
'kids love mohamad bin zayed city'

'eek i cannot stand split keyboards does not work well with mmos'

'i was laughing at the kat amp andre hate tweets and recognized you we share the same hate'

'rt more daesh gangs killed in shengal'
```

Figura 11. Comentarios preprocesados [elaboración propia].

Quitar la parte irrelevante en los textos permite disminuir los errores en el modelado, generando clasificadores enfocados sólo en lo esencial dentro de los comentarios. Así, la cantidad de filas remanente es de 37.669.

3.3 Tokenización, Lematización y Stop Words

Como se ha indicado en la sección anterior, NLP implica aplicar diferentes procedimientos gramaticales y semánticos para trabajar con datos de tipo texto. Algunas de estas son:

- Tokenización: dado un texto, se divide en las unidades más pequeñas que tengan significado, es decir, en las palabras que lo componen. Por ejemplo, en la frase “las bonitas casas se encuentran en el bosque”, la tokenización individualizaría las palabras como:

  las | bonitas | casas | se | encuentran | en | el | bosque

- Lematización: dada una palabra en forma flexionada (en plural, en femenino, conjugada, etc.), se halla su lema o forma canónica correspondiente. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra:
  - Sustantivos: masculino singular

  "gato" es el lema de "gato/s", “gata/s”
  - Adjetivos: masculino singular

  "guapo" es el lema de “guapo/s”, "guapa/s"
  - Verbos: infinitivo

  "decir" es el lema de "dije", "diré" o "dijéramos"

  Para hacer una correcta lematización de las palabras de un texto, es importante incluir un análisis morfológico que establezca su función contextual dentro del mismo. De esta manera, se evitan inconvenientes como la desambiguación lingüística (¿a qué se refiere con “banco”?) o la mala categorización (¿”azul” funciona como sustantivo o adjetivo?).

  Continuando con el ejemplo anterior, la lematización produciría la salida:

  la | bonito | casa | se | encontrar | en | el | bosque

En el caso de las “stop words” (o palabras vacías), son aquellas que no aportan mayor sentido a un texto y que son descartadas durante el procesamiento del mismo. Son dependientes del idioma y son las más comunes dentro de éste (artículos, preposiciones, pronombres, entre otras). En relación al ejemplo anterior, finalmente quedaría como:

bonito | casa | encontrar | bosque

Continuando con el trabajo de campo, el siguiente paso es establecer un conjunto de “stop words” para el idioma inglés. Además, dado que en la plataforma Twitter se aplica una determinada jerga, dichas palabras también son consideradas para ser removidas. Por otro lado, también se han añadido ciertas abreviaciones halladas en los comentarios. Algunos ejemplos de estas “palabras vacías” son:

about, at, before, bc (because), do, had, her, lmao, plz (please), rt (retweet), the

A continuación, se aplicaron las técnicas de tokenización y lematización, dando como resultado, para cada comentario, un conjunto de palabras claves ya canonizadas. Algunos ejemplos son:

Comentario original:
mark cuban is a dumb fuck, a super pussy, what he says is wrong or meaningless, to coin a phrase he is a cowardly nigger lover.

Palabras:
mark cuban dumb fuck super pussy say wrong meaningless coin phrase cowardly nigger lover

Comentario original:
I’d never call a woman a female I got to much respect for bitches

Palabras:
would never call woman female get much respect bitch

Figura 12. Comparación entre comentarios originales y luego de tokenización y lematización [elaboración propia].

Finalmente, como algunos comentarios han quedado en blanco luego de que se aplicaran estos métodos (dada la reestructuración de las palabras y eliminación de "stop words"), se eliminan del dataset quedando un conjunto final de 37.591 filas.

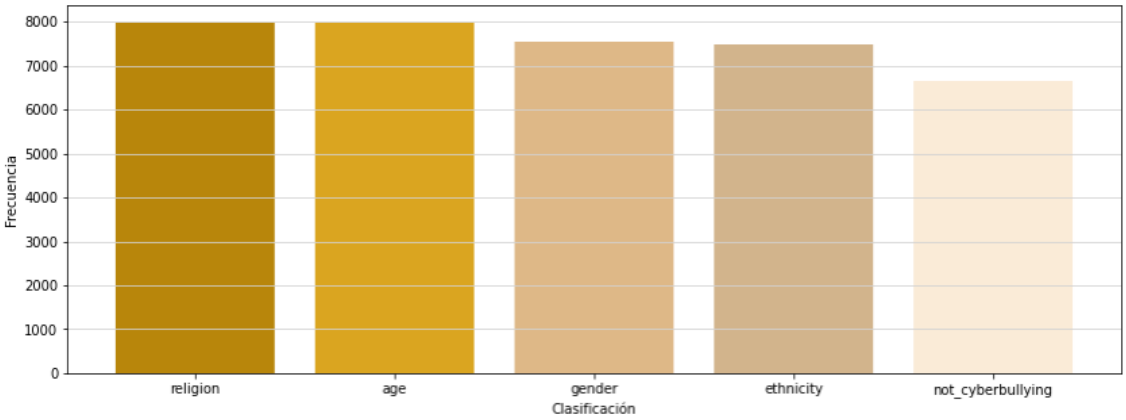


Figura 13. Ejemplos por clase luego de la preparación completa de los datos [elaboración propia].

A pesar de que, luego del procesamiento completo, existe un desnivel entre las clases (siendo la más afectada "not_cyberbullying"), no se considera demasiado acentuado y se puede continuar con su análisis sin necesidad de aplicar una técnica de igualación (como la de eliminación de ejemplos de las clases más grandes). Si hubiera existido demasiada diferencia entre ellas, sí habría sido necesario equipararlas, puesto que, de lo contrario, la "accuracy" daría valores poco confiables, pudiendo calificar a un modelo como mejor de lo que realmente es.

3.4 Análisis detallado

Del conjunto de datos ya completamente transformado, se llevaron a cabo algunos estudios específicos para poder observar patrones de comportamiento y comprender mejor qué esperar de los mismos.

- Las 20 palabras más frecuentes en cada situación (generales, no sólo insultos o nombres ofensivos):

Sin Cyberbullying

	Palabra	Frecuencia
0	bully	699
1	get	572
2	go	512
3	like	393
4	school	373
5	would	299
6	think	299
7	people	290
8	make	287
9	one	272
10	know	262
11	good	243
12	time	225
13	see	224
14	say	223
15	want	214
16	woman	196
17	need	193
18	andre	185
19	fuck	184

Sólo Cyberbullying

	Palabra	Frecuencia
0	bully	9158
1	school	8837
2	fuck	6725
3	girl	5409
4	like	5286
5	joke	5279
6	high	5226
7	nigger	5224
8	dumb	4879
9	muslim	4878
10	gay	4380
11	get	4379
12	rape	4280
13	people	4204
14	say	3908
15	call	3697
16	make	3384
17	one	3250
18	idiot	3187
19	go	3017

Figura 14. Comparación de las 20 palabras más frecuentes para cada caso [elaboración propia].

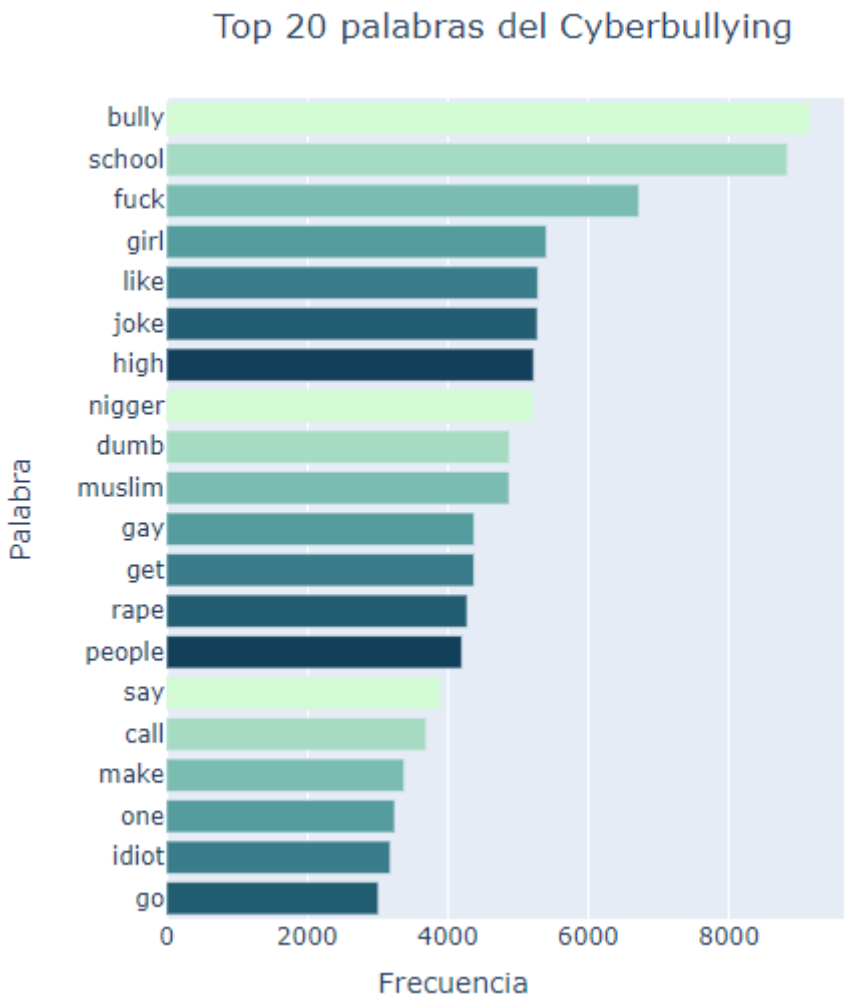


Figura 15. Las 20 palabras más frecuentes en comentarios con ciberacoso [elaboración propia].

- Palabras más usuales en cada tipo de acoso (combinadas, no sólo insultos o nombres ofensivos). Las de mayor tamaño son las más utilizadas:

Figura 16. Las palabras más frecuentes en cada tipo de ciberacoso [elaboración propia].

- Combinación de las "dos palabras" (2-grama) más utilizadas por cada tipo de acoso (combinadas, no sólo insultos o nombres ofensivos):

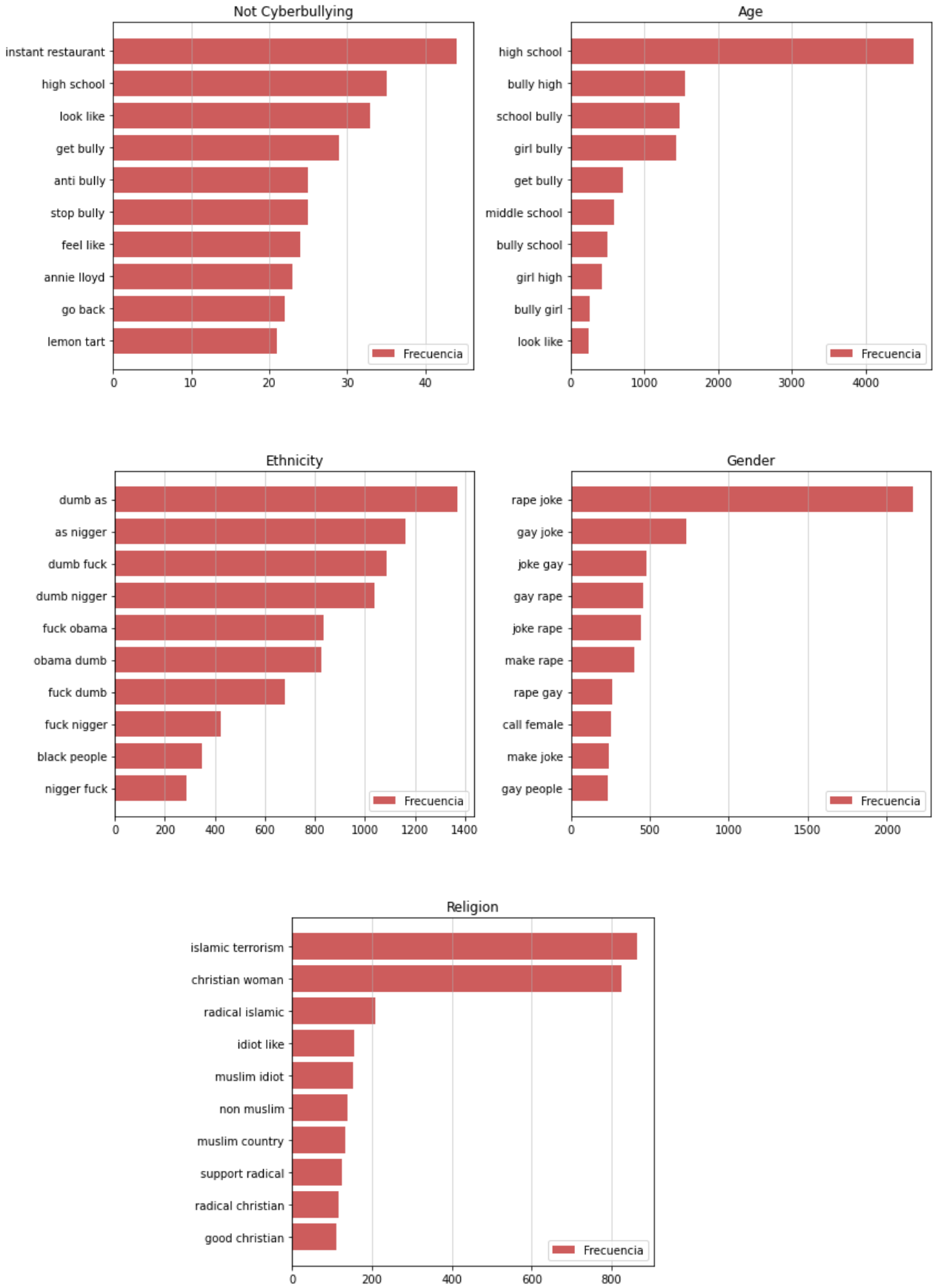


Figura 17. La combinación de “2 palabras” más frecuentes en cada tipo de ciberacoso [elaboración propia].

De lo visto hasta aquí, si bien en general se comparten algunas palabras (bully, school, people, fuck), es notable (y deseable) que cada tipo de acoso posee vocablos específicos, permitiendo al futuro clasificador identificar más fácilmente qué intención puede tener un determinado comentario.

3.5 Modelos de clasificación

Para la evaluación de modelos predictivos de clasificación, el primer paso será representar numéricamente las secuencias de palabras previamente procesadas de cada comentario. Utilizar números en vez de texto permite que estos últimos sean comprendidos por los diferentes algoritmos predictivos.

Se opta por utilizar dos técnicas de "feature extractions":

- Bag of Words (BOW): con variantes One-Hot, frecuencia y TF-IDF.
- Word Embedding (WE): con variantes Skip-Gram y Continuous Bag of Words.

"Bag of Words" (bolsa de palabras) consiste en representar un texto en vectores que indican si una palabra se encuentra contenida (valor distinto de "0") o no (valor igual a "0"). Al conjunto total de palabras a evaluar su existencia en cada comentario, se lo puede deducir del dataset total o, como alternativa, utilizar un "diccionario de palabras" o "corpus" como parámetro. Un ejemplo sencillo sería:

ID	Texto original	Palabras
1	las bonitas casas se encuentran en el bosque	bonito, casa, encontrar, bosque
2	me gusta ver llover	gustar, llover
3	estaré en casa si llueve	estar, casa, llover

Tabla 1. Ejemplo de un dataset simple habiendo aplicado tokenización y lematización [elaboración propia].

Al aplicar "bag of words" sobre esos datos se obtendría:

ID	bonito	bosque	casa	encontrar	estar	gustar	llover
1	1	1	1	1	0	0	0
2	0	0	0	0	0	1	1
3	0	0	1	0	1	0	1

Tabla 2. Ejemplo del dataset de ejemplo anterior habiendo aplicado la técnica de "bag of words" [elaboración propia].

Las variantes a aplicar con esta técnica son:

- One-Hot (OH): como se observa en el ejemplo anterior, coloca un "1" si la palabra se encuentra en el comentario y un "0" si no.

- Frecuencia (Fr): coloca el nº de veces que la palabra aparece en el comentario.
- Term frequency - Inverse document frequency (TF-IDF): su valor se calcula acompañado de una fórmula logarítmica, obteniendo la importancia de la palabra en el comentario.

"Word Embeeding" (incrustación de palabras) también representa un texto en forma de vectores como "Bag of Words", pero, en este caso, también captura su semántica en el contexto en la cual aparece (sus palabras vecinas). De esta manera, permite que vectores de palabras similares (o con usos parecidos) estén más cerca entre ellos, otorgando, inicialmente, una predicción más precisa. Cada vector está representado por los pesos dados por las redes neuronales que los entrenan. Luego, cada comentario del dataset queda representado por un vector fruto de cálculos aritméticos entre los vectores de las palabras que lo componen.

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

Las variantes a aplicar con esta técnica son:

- Skip-Gram (SG): dada una palabra, se entrena para predecir diferentes contextos. Es mejor capturando palabras diferentes morfológicamente pero que comparten algún sentido (los vectores de "perro" y "gato" estarán cerca). Aquí, el vector de dicha palabra se utiliza para predecir el contexto.
- Continuous Bag of Words (CBOW): dadas otras palabras como contexto, se entrena para predecir una palabra. Es mejor capturando palabras sintácticamente similares (los vectores de "perro" y "perros" estarán cerca). Aquí, los diferentes vectores se combinan para predecir la palabra.

Explicado lo anterior, a continuación se observa cómo han quedado representados los datos para cada "feature extraction" (para BOW se han obtenido 725 palabras como relevantes, para WE se han considerado vectores de 300 componentes):

- BOW - OH:

	able	abortion	absolutely	abt	abuse	accept	account	act	action	actual	...	ya	yall	yard	yeah	year	yes	yet	yo	young	yr
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
37586	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
37587	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
37588	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
37589	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
37590	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

37591 rows × 725 columns

Figura 18. Dataset de Twitter luego de haber aplicado BOW - OH [elaboración propia].

- BOW - Fr:

	able	abortion	absolutely	abt	abuse	accept	account	act	action	actual	...	ya	yall	yard	yeah	year	yes	yet	yo	young	yr
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	2	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
37586	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
37587	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
37588	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
37589	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
37590	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

37591 rows × 725 columns

Figura 19. Dataset de Twitter luego de haber aplicado BOW - Fr [elaboración propia].

- BOW - TF-IDF:

	able	abortion	absolutely	abt	abuse	accept	account	act	action	actual	...	ya	yall	yard	yeah	year	yes	yet	yo	young	yr
0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.344596	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
37586	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.080736	0.0	0.0	0.0
37587	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
37588	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
37589	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
37590	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0

37591 rows × 725 columns

Figura 20. Dataset de Twitter luego de haber aplicado BOW - TF-IDF [elaboración propia].

- WE - SG:

	0	1	2	3	4	5	6	7	8	9	...	290	291	292
0	0.102746	0.172226	-0.031413	-0.019200	0.097553	-0.212057	0.104507	0.360109	0.160548	-0.190730	...	-0.022867	0.143247	0.159913
1	0.012914	0.102777	-0.265883	0.192647	-0.135857	-0.051742	0.223783	0.055834	0.161703	-0.136241	...	-0.095014	0.140672	-0.078900
2	0.030633	0.108398	0.034570	-0.033109	0.017146	-0.264128	0.085610	0.289934	0.096751	-0.147135	...	-0.217256	0.118027	0.064261
3	-0.015635	0.083648	0.024387	-0.028164	0.060636	-0.214511	0.099112	0.237441	0.008079	-0.125408	...	-0.116046	0.137993	0.079417
4	-0.043072	0.253338	-0.042607	0.045007	0.001857	-0.087474	0.226114	0.196220	0.057031	-0.291368	...	0.021975	0.252098	0.134638
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
37586	-0.032895	0.092418	0.027661	0.039125	0.011352	-0.153430	0.156886	0.248192	0.040508	-0.151681	...	-0.005451	0.145054	0.109624
37587	-0.007592	0.093666	-0.014215	0.083777	0.002521	-0.149323	0.166476	0.221575	0.029326	-0.123860	...	-0.052582	0.124124	0.097567
37588	0.044777	0.104201	-0.000381	-0.033943	0.046938	-0.356957	-0.001758	0.356636	0.112543	-0.161097	...	-0.195784	0.146718	0.038342
37589	0.067152	0.139444	-0.037266	-0.001724	0.080993	-0.311706	-0.095580	0.317240	0.175502	-0.156594	...	-0.244340	0.247655	0.078161
37590	0.008557	0.138573	-0.066178	-0.026251	0.067208	-0.352415	-0.004880	0.379700	0.164733	-0.189058	...	-0.245293	0.169542	0.043758

37591 rows × 300 columns

Figura 21. Dataset de Twitter luego de haber aplicado WE - SG [elaboración propia].

- WE - CBOW:

	0	1	2	3	4	5	6	7	8	9	...	290	291	292
0	0.167330	0.449831	-0.211454	0.100899	0.024999	-0.310467	0.399437	0.727097	0.022473	-0.484169	...	-0.022665	0.365367	0.128108
1	0.121457	0.350634	-0.424165	0.180646	0.247207	-0.880225	0.383678	0.664518	-0.335196	-0.261271	...	-0.038933	0.359735	-0.101281
2	0.063133	0.155797	0.066135	0.052997	0.049785	-0.271911	0.185434	0.354647	0.090607	-0.145363	...	-0.129535	0.221746	0.094753
3	0.030519	0.184718	0.124360	0.039443	0.095263	-0.347203	0.238747	0.447154	0.083270	-0.152010	...	-0.156863	0.219247	0.173799
4	0.099925	0.351694	-0.039391	0.207931	-0.087552	-0.188626	0.495416	0.456977	-0.001152	-0.420432	...	-0.047863	0.329163	0.042000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
37586	-0.039433	0.150024	0.082550	0.281649	0.076993	-0.280438	0.264454	0.407439	-0.040369	-0.236038	...	0.006328	0.172718	0.182751
37587	-0.015817	0.130950	0.025843	0.167046	0.121306	-0.399794	0.245561	0.403428	-0.090258	-0.188204	...	-0.048783	0.154692	0.150626
37588	0.229548	0.068666	0.121720	-0.114573	0.168323	-0.511999	0.150075	0.593302	0.222740	-0.113903	...	-0.453440	0.348114	0.047458
37589	0.362990	0.083356	0.071889	-0.155234	0.178967	-0.512985	0.077169	0.656811	0.279051	-0.076503	...	-0.616604	0.476183	-0.066501
37590	0.259467	0.086850	0.098692	-0.215101	0.216147	-0.572438	0.153551	0.676416	0.217555	-0.106833	...	-0.529383	0.367874	0.017479

37591 rows × 300 columns

Figura 22. Dataset de Twitter luego de haber aplicado WE - CBOW [elaboración propia].

Luego, cada representación se dividió en dos grupos, uno para entrenar los modelos (70% de los ejemplos) y otro para evaluarlos (30% de los ejemplos).

Los modelos considerados para este estudio son los siguientes:

- Logistic Regression (LR)
- Random Forest Classifier (RFC)
- Support Vector Classifier (SVC)
- Naive Bayes (NB)
- Ada Boost Classifier (ABC)
- Gradient Boosting Classifier (GBC)
- Stochastic Gradient Descent Classifier (SGDC)
- Multilayer Perceptron Classifier (MLPC)
- Long Short Term Memory (LSTM)
- Convolutional Neural Network (CNN)

En el caso particular de LSTM y CNN, estas redes neuronales requieren una representación específica de los datos, no siendo posible evaluarlas con las enseñadas anteriormente. Así, los comentarios quedan representados por secuencias de palabras (todas las secuencias del mismo tamaño, que sería el máximo obtenido del comentario más largo) donde cada palabra queda representada por un ID único (por ejemplo, la palabra “bully” se representará en todos los comentarios con el ID “1”). El “padding”

se halla a la izquierda, el cual se completa con valores “0” para completar el largo máximo de la secuencia.

```
[[   0    0    0 ...    0   79  591]
 [   0    0    0 ...    0    0   28]
 [   0    0    0 ... 1888 1018  695]
 ...
 [   0    0    0 ...  499  212    3]
 [   0    0    0 ...    3   11    8]
 [   0    0    0 ...   11    8   27]]
(37591, 37)
```

Figura 23. Dataset de Twitter luego de secuenciar los comentarios [elaboración propia].

Por otro lado, esta enumeración de palabras también requiere de una matriz de pesos para cada una. Dicha matriz se obtiene utilizando el mismo proceso que el que fue utilizado para codificar WE - SG.

```
[[ 0.          0.          0.         ...  0.          0.
   0.        ]
 [ 0.02581957  0.1515667   0.28516608 ... -0.23586971  0.17209819
  -0.47069734]
 [-0.17666867  0.12785622  0.02182264 ... -0.38539878  0.22085629
  -0.17294759]
 ...
 [ 0.          0.          0.         ...  0.          0.
   0.        ]
 [ 0.          0.          0.         ...  0.          0.
   0.        ]
 [ 0.          0.          0.         ...  0.          0.
   0.        ]]
(27170, 300)
```

Figura 24. Matriz de pesos de 27.170 palabras halladas en el dataset [elaboración propia].

De esta manera, para LSTM y CNN sólo se estudiará un “feature extraction” (con características similares a WE - SG) también dividido en grupos de training (70% de los ejemplos) y test (30% de los ejemplos). Para dichos modelos, el número de épocas de entrenamiento fue de 10 vueltas (en general, para la mayoría de modelos un número

mayor a 12 produciría sobre-entrenamiento, haciendo que la función de pérdida sea casi cero pero generando predicciones erróneas para datos no vistos por el modelo[17]).

En relación al primer conjunto de predictores (LR, RFC, SVC, NB, ABC, GBC, SGDC y MLPC), cabe destacar que los mismos se han evaluado utilizando un "cross-validation" de 5 vueltas (generalmente se utilizan valores entre 5 y 10, según tamaño de dataset y recursos computacionales disponibles[18]), promediando luego los resultados obtenidos. A su vez, para cada uno se obtuvo la gráfica "ROC Curves" (graficando una curva por clase, empleando la estrategia "One vs. Rest") para medir el desempeño del "feature extraction" más efectivo.

La curva ROC (Receiver Operating Characteristics), es un gráfico de probabilidad que enseña el rendimiento de un modelo para diferentes umbrales de clasificación (thresholds). Dicha curva queda definida por los parámetros "True Positive Rate" (TPR) y "False Positive Rate" (FPR), acompañada normalmente del valor AUC (Area Under The Curve) que mide la capacidad de un predictor para diferenciar las clases en todos los umbrales de clasificación posibles. Se puede pensar como el cálculo de una integral que mide la separabilidad de clases, donde a mayor valor de AUC, mejor es el modelo prediciendo "K" como "K" (verdadero positivo "TP") y "no K" como "no K" (verdadero negativo "TN")[19].

[17] GeeksforGeeks. *Choose optimal number of epochs to train a neural network in Keras* [en línea]. India: GeeksforGeeks, 2022. Disponible en: <https://www.geeksforgeeks.org/choose-optimal-number-of-epochs-to-train-a-neural-network-in-keras/>

[18] Lyashenko, V., Jha, A. *Cross-Validation in Machine Learning: How to Do It Right* [en línea]. Polonia: Neptune, 2022. Disponible en: <https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right>

[19] Google Developers. *Clasificación: Curva ROC y AUC* [en línea]. EE.UU.: Google Developers, Septiembre 2022. Disponible en: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es_419 >

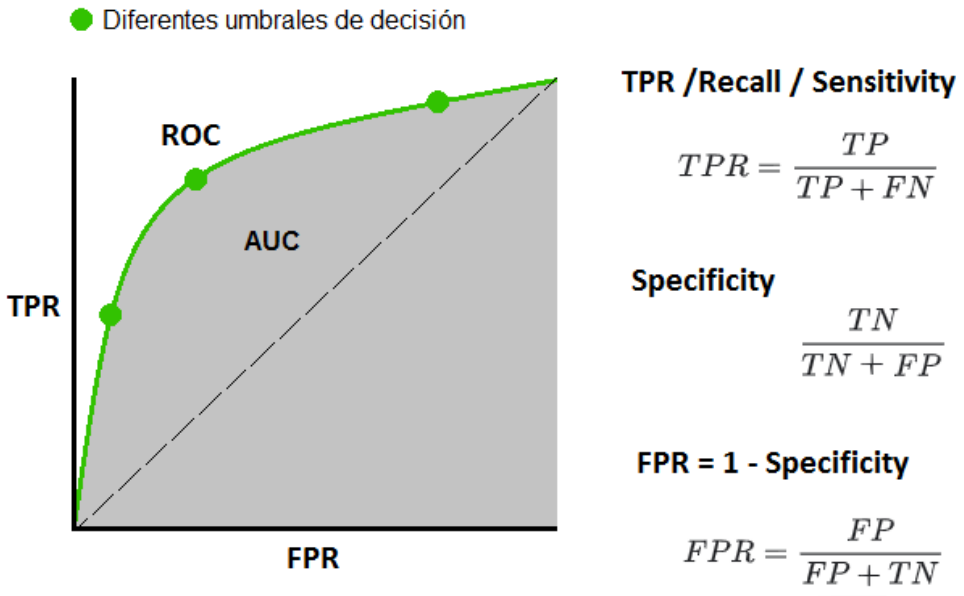


Figura 25. Gráfica explicativa sobre "ROC Curve" y "AUC"[20].

Al disminuir el valor del umbral, se clasifican más elementos como positivos, tanto los verdaderos positivos "TP" (disminuyendo los falsos negativos "FN" y aumentando "TPR" o "Sensitivity") como los falsos positivos "FP" (disminuyendo "Specificity" y aumentando "FPR"). Así, como regla:

⬆ TPR/Sensitivity ⬇ Specificity ⬆ FPR

Por ejemplo, un umbral de "0,2" implica que cuando el predictor obtenga una probabilidad >= 0,2, un ejemplo se clasifica como "K", lo cual puede ser cierto (TP) o falso (FP). Si se aumenta dicho umbral, disminuyen los falsos positivos, pero también la capacidad de detectar los verdaderos. La definición de un umbral depende del problema que se esté estudiando y de qué es más importante (con mayor peso y coste) al momento de clasificar. Por ejemplo, en detección de enfermedades, es preferible maximizar un "TPR" a costa de un también mayor "FPR", ya que es más grave un "FN" que un "FP". Es decir, es peor indicar que se está sano cuando en realidad existe la enfermedad, que indicar que se está enfermo cuando en realidad no.

[20] Narkhede, Sarang. *Understanding AUC - ROC Curve* [en línea]. Canadá: Towards Data Science, Junio 2018. Disponible en: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

Aquí la situación ideal es “TPR = 1” (FN = 0) y “FPR = 0” (FP = 0). De esta manera, AUC obtendría su valor óptimo de “1”, logrando una perfecta distinción de clases. A continuación se observan diferentes casos de distribuciones de probabilidades entre las categorías y su curva ROC y medida AUC asociadas:

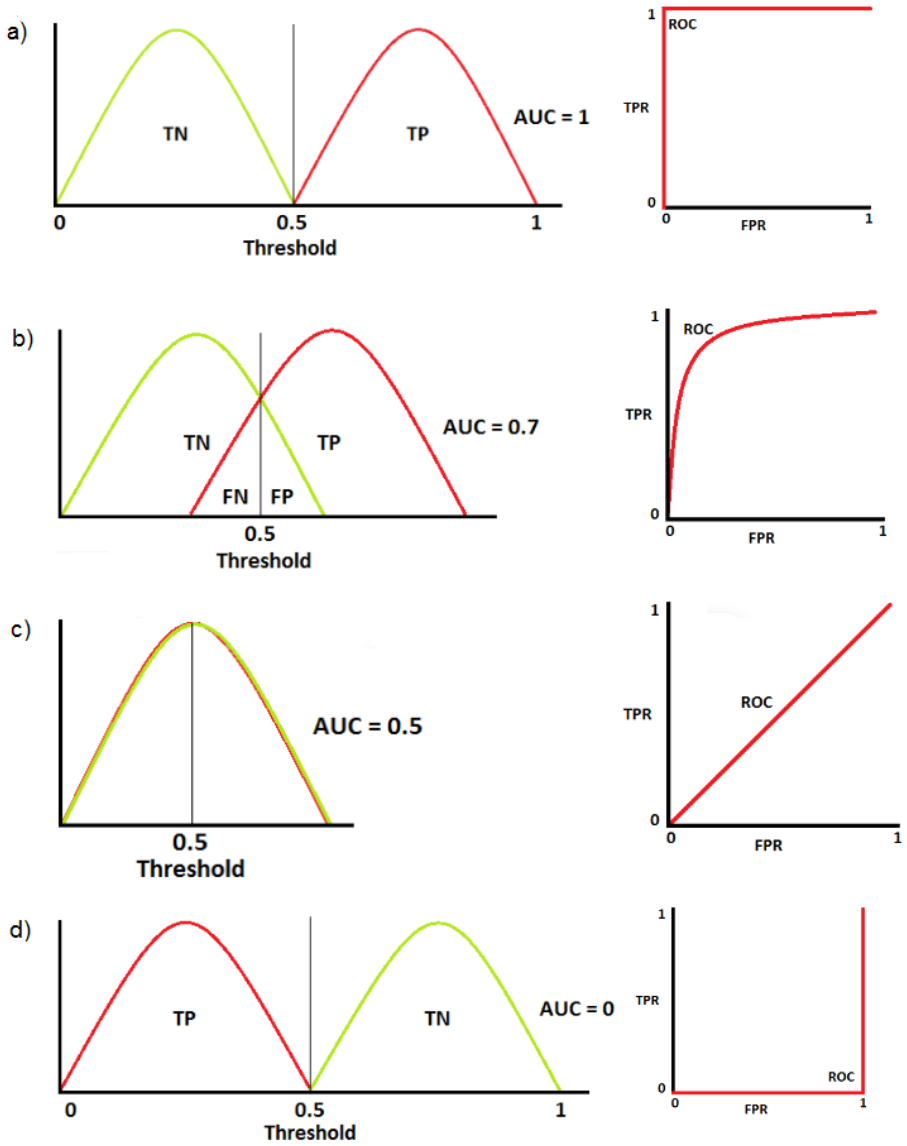


Figura 26. Diferentes distribuciones de probabilidad de categorías y su “ROC Curve” y AUC asociadas[21].

a) es la situación ideal, las curvas no se superponen y hay una distinción de clases perfecta.

[21] Narkhede, Sarang. *Understanding AUC - ROC Curve* [en línea]. Canadá: Towards Data Science, Junio 2018. Disponible en: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

b) la situación más común, donde las curvas se superponen introduciendo errores de tipo I (falso positivo) y tipo II (falso negativo). Estos se pueden controlar mediante la selección del umbral de decisión. En este ejemplo hay un 70% de posibilidades de que se distingan las clases.
c) la peor situación donde el predictor no logra distinguir las clases, es un clasificador completamente aleatorio.
d) situación en que el modelo predice las clases completamente al revés.

Por otro lado, dentro del análisis también se ha incluido, para todos los modelos y su mejor "feature extraction" (aquella con mayor "accuracy"), un reporte de clasificación que enseña métricas adicionales (precision, recall, f1-score) sobre cada categoría. Estos valores ofrecen mayor comprensión sobre la calidad predictora que tendrá un clasificador para cada clase.

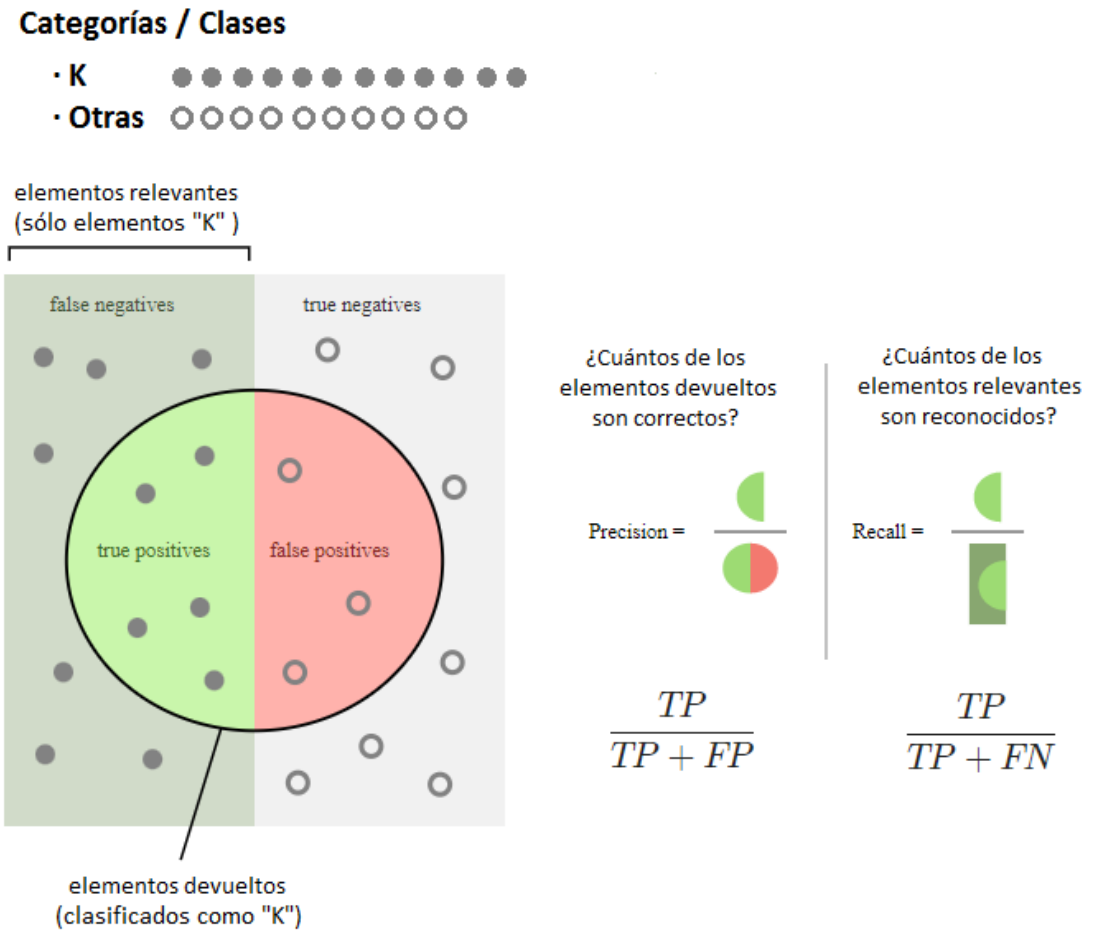


Figura 27. Gráfica explicativa sobre "precision" y "recall"[22].

La "precision" mide cuántos de los ejemplos clasificados como categoría "K", son realmente categoría "K". Aquí lo ideal es que "FP = 0" (precision = 1). Caso contrario, si "FP > 0" (precision < 1), indica que ejemplos de otras categorías se clasificaron erróneamente como "K".

En el caso de "recall", mide, de todos los ejemplos pertenecientes a la categoría "K", cuántos son correctamente clasificados cómo "K". Aquí lo ideal es que "FN = 0" (recall = 1). Caso contrario, si "FN > 0" (recall < 1), significa que ejemplos de "K" se clasificaron erróneamente como otras clases.

[22] Riggio, Christopher. *What's the deal with Accuracy, Precision, Recall and F1?* [en línea]. Canadá: Towards Data Science, Noviembre 2019. Disponible en: <https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021>

Dicho lo anterior, generalmente estas métricas se evalúan juntas, ya que, si sube el valor de una, baja el de la otra, con lo que se busca un balance entre ambas. Esto lo otorga "f1-score", cuya fórmula comprende (y da importancia) a las dos métricas previamente vistas:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

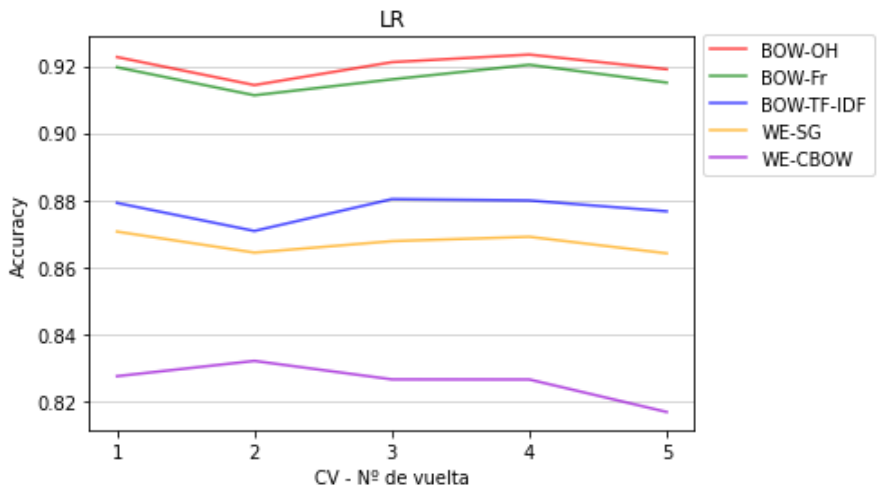
Aquí, si "f1-score" posee un valor alto, "precision" y "recall" también poseen valor alto, implicando que el modelo ha realizado un buen desempeño. Por otro lado, la diferencia de esta métrica con "accuracy", cuya fórmula es:

$$accuracy = \frac{\text{Nº de predicciones correctas}}{\text{Nº total de predicciones}} = \frac{TP + TN}{TP + TN + FP + FN}$$

es que esta última es más general en significado, dando una idea global de cómo clasifica el predictor. Sin embargo, si es necesario más detalle o dar más peso a "falsos positivos" y "falsos negativos" (por ejemplo, es más grave indicar que un paciente no corre riesgo en tener cáncer cuando en realidad sí), esto se consigue con "f1-score" (ya que no permite que un gran número de "verdaderos negativos", usualmente no tan importantes, influya demasiado en la puntuación). Adicionalmente, ésta también es más apta en situaciones de clases desbalanceadas, ya que "accuracy" es más propensa a calificar un modelo como "bueno", cuando en realidad no lo es tanto.

Luego de esta explicación, a continuación se enseñan los resultados obtenidos durante la evaluación de los modelos y una interpretación de los mismos.

A) Logistic Regression (LR)



LR	
BOW-OH	0.920116
BOW-Fr	0.916467
BOW-TF-IDF	0.877361
WE-SG	0.867176
WE-CBOW	0.825865

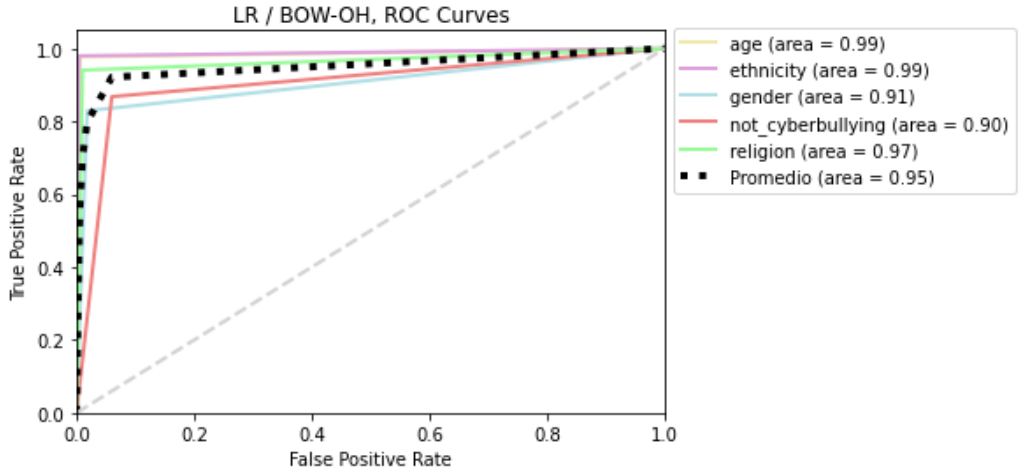


Figura 28. Análisis general de LR [elaboración propia].

Su "accuracy" máxima, un poco más de 0,92, se logra utilizando BOW - OH. Para este caso, las curvas ROC de "gender" y "not_cyberbullying" son las que menos área comprenden, lo que significa que el clasificador presenta algunos inconvenientes con estas clases. Para el primer caso, algunos comentarios no se están clasificando como "gender", aunque tampoco hay otras clases erróneas que se clasifiquen como tal. En "not_cyberbullying", sí se observa que no detecta bien los propios comentarios y que otras clases también se están asignando aquí.

LR / BOW-OH, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.98	0.98	0.98	2391
ethnicity	0.98	0.98	0.98	2234
gender	0.92	0.83	0.87	2263
not_cyberbullying	0.76	0.87	0.81	1998
religion	0.96	0.94	0.95	2392
accuracy			0.92	11278
macro avg	0.92	0.92	0.92	11278
weighted avg	0.93	0.92	0.92	11278

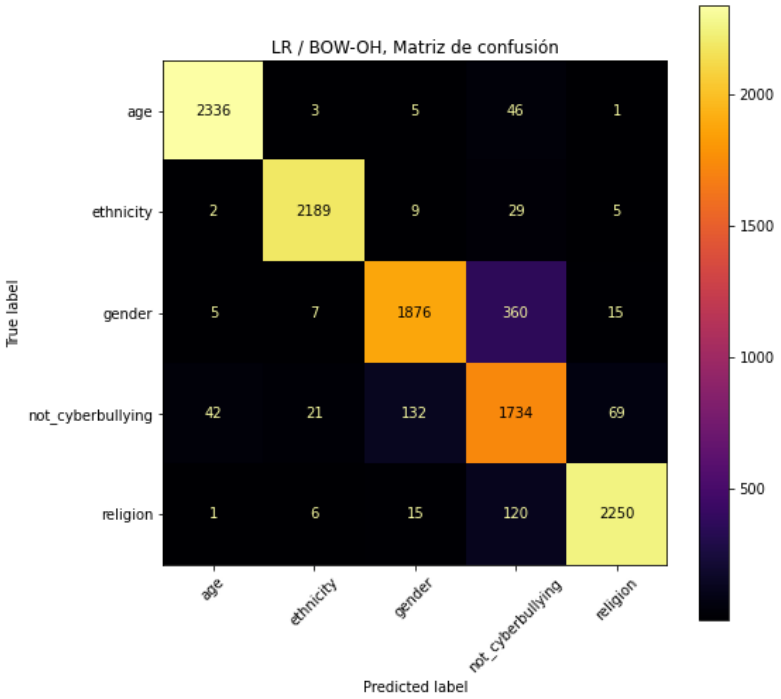
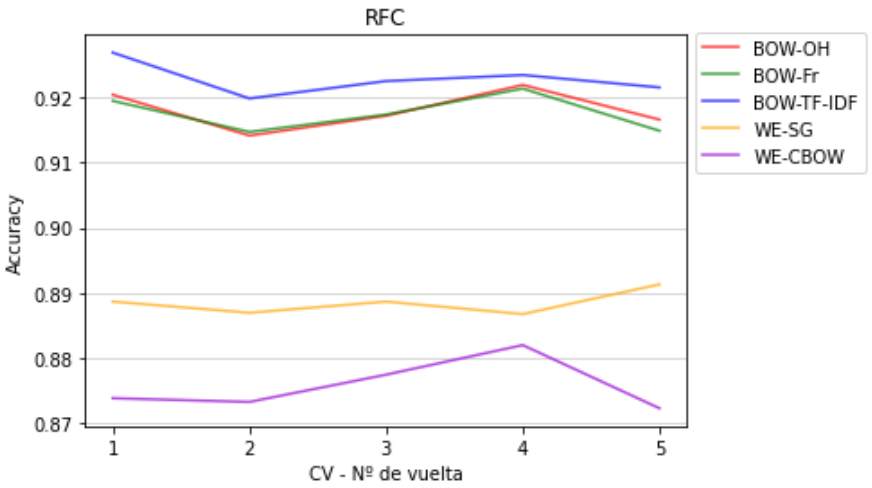


Figura 29. Análisis detallado de LR [elaboración propia].

Al observar el reporte de clasificación y matriz de confusión asociadas, se observa que las clases previamente nombradas son las más complicadas de atinar. En “not_cyberbullying”, a pesar de que su “recall” (acierto de los propios ejemplos) ha logrado un aceptable 0,87, posee una “precision” menor de 0,76, indicando que algunos comentarios han sido clasificados como tal cuando en realidad no lo son, como se puede ver para “gender” dentro de la matriz. Para el resto de clases estas medidas son altas y coincidentes entre sí, lo que quiere decir que, generalmente, han predicho correctamente sus propios comentarios y no han recibido tantas clasificaciones erróneas del resto de categorías.

B) Random Forest Classifier (RFC)



RFC	
BOW-OH	0.918025
BOW-Fr	0.917531
BOW-TF-IDF	0.922814
WE-SG	0.888458
WE-CBOW	0.875765

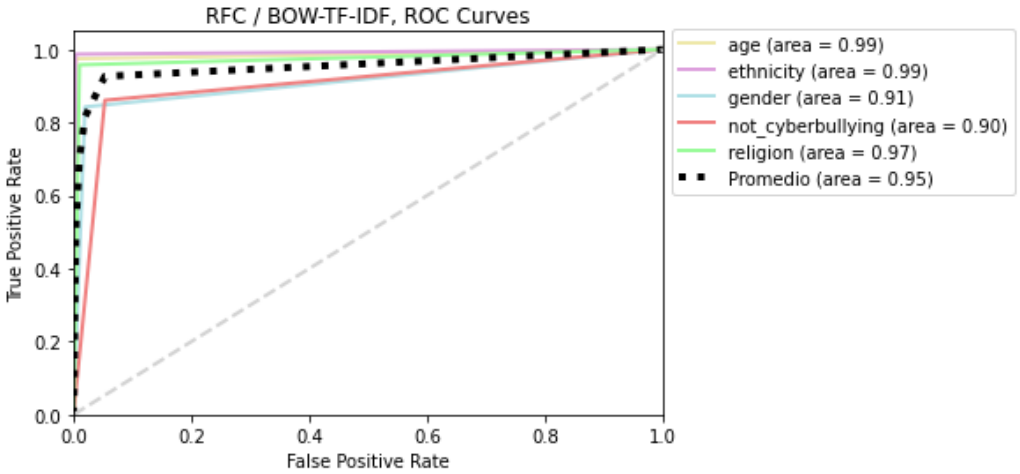


Figura 30. Análisis general de RFC [elaboración propia].

Su "accuracy" máxima de 0,9228 se logra utilizando BOW - TF-IDF. Para este caso, las curvas ROC de "gender" y "not_cyberbullying" presentan un comportamiento similar al visto anteriormente, aunque con una leve mejoría para el caso de la primera.

RFC / BOW-TF-IDF, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.98	0.97	0.98	2391
ethnicity	0.99	0.99	0.99	2234
gender	0.92	0.84	0.88	2263
not_cyberbullying	0.78	0.86	0.82	1998
religion	0.96	0.96	0.96	2392
accuracy			0.93	11278
macro avg	0.93	0.92	0.92	11278
weighted avg	0.93	0.93	0.93	11278

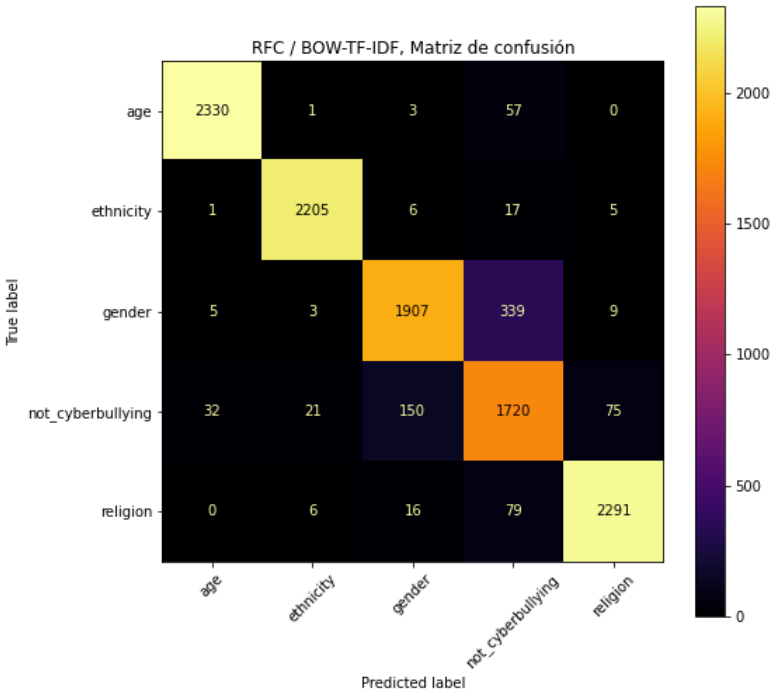
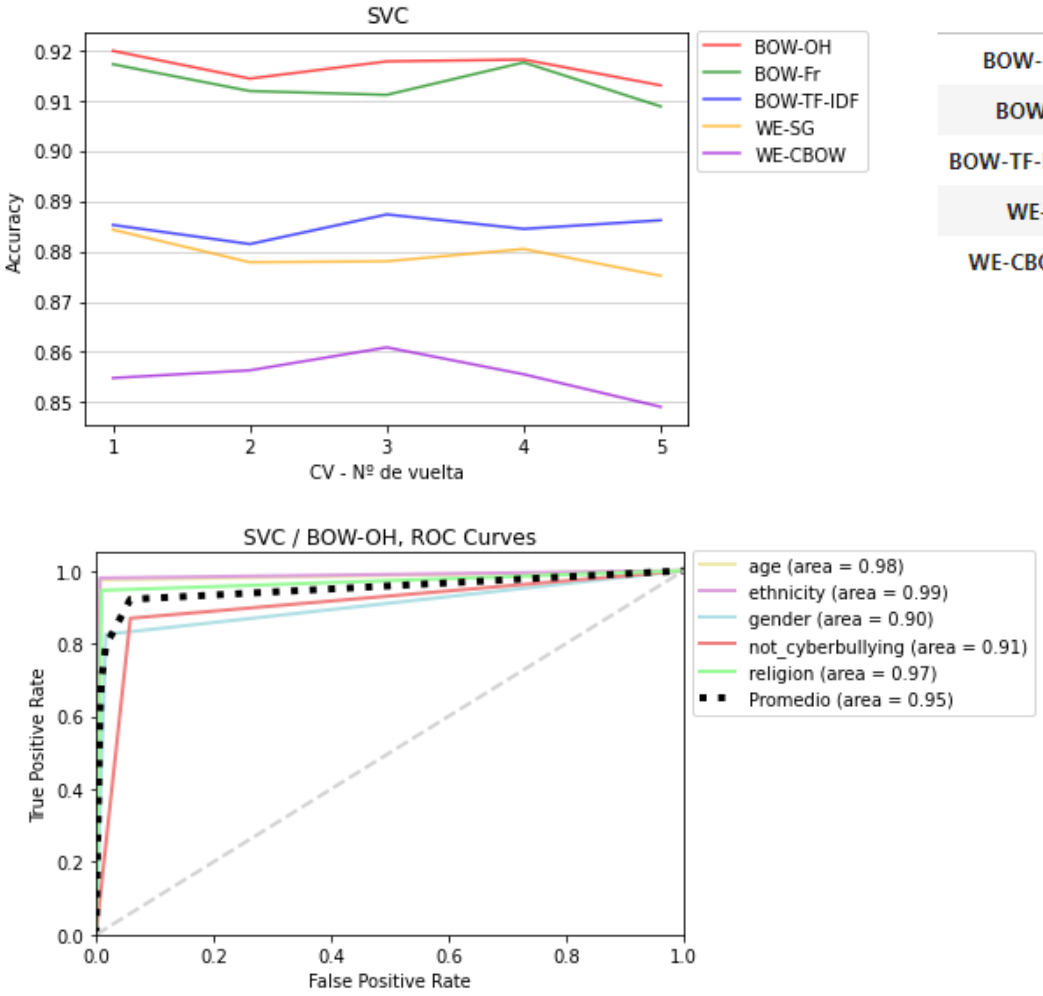


Figura 31 Análisis detallado de RFC [elaboración propia].

En general, se observan mejores resultados que el anterior modelo y destaca "ethnicity" por ser la mejor clasificada. "gender" y "not_cyberbullying" siguen siendo las más complicadas dentro de la tarea, donde su mayor confusión se produce entre ellas mismas.

C) Support Vector Classifier (SVC)



	SVC
BOW-OH	0.916581
BOW-Fr	0.913275
BOW-TF-IDF	0.884924
WE-SG	0.879147
WE-CBOW	0.855356

Figura 32. Análisis general de SVC [elaboración propia].

Su "accuracy" máxima de 0,9166 se logra utilizando BOW - OH. Observándose ya una tendencia, las curvas ROC de "gender" y "not_cyberbullying" continúan siendo las más afectadas en la clasificación.

SVC / BOW-OH, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.98	0.97	0.97	2391
ethnicity	0.97	0.98	0.98	2234
gender	0.93	0.83	0.87	2263
not_cyberbullying	0.76	0.87	0.81	1998
religion	0.96	0.95	0.96	2392
accuracy			0.92	11278
macro avg	0.92	0.92	0.92	11278
weighted avg	0.93	0.92	0.92	11278

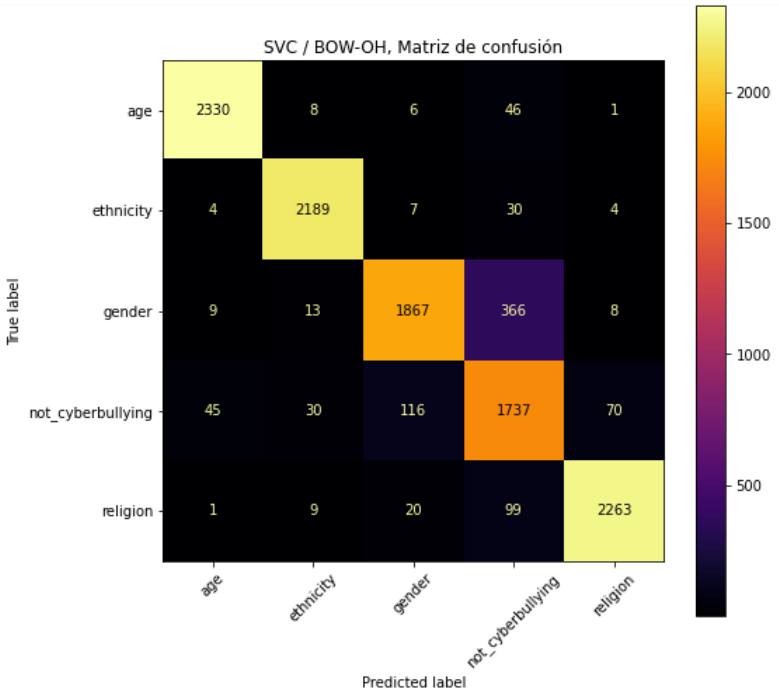
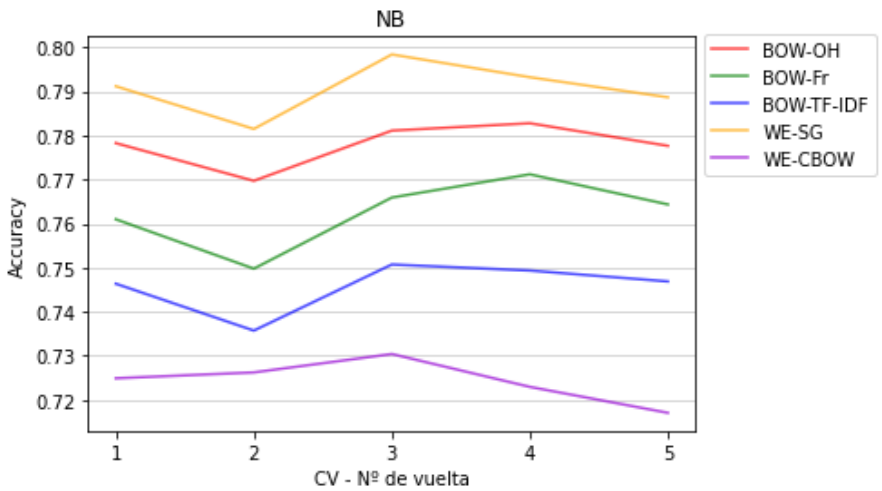


Figura 33. Análisis detallado de SVC [elaboración propia].

Las métricas se han reducido un poco para ciertas clases. Sin embargo, "not_cyberbullying" es la que mejor respuesta ha dado, hasta ahora, en "recall".

D) Naive Bayes (NB)



	NB
BOW-OH	0.777905
BOW-Fr	0.762437
BOW-TF-IDF	0.745791
WE-SG	0.790598
WE-CBOW	0.724280

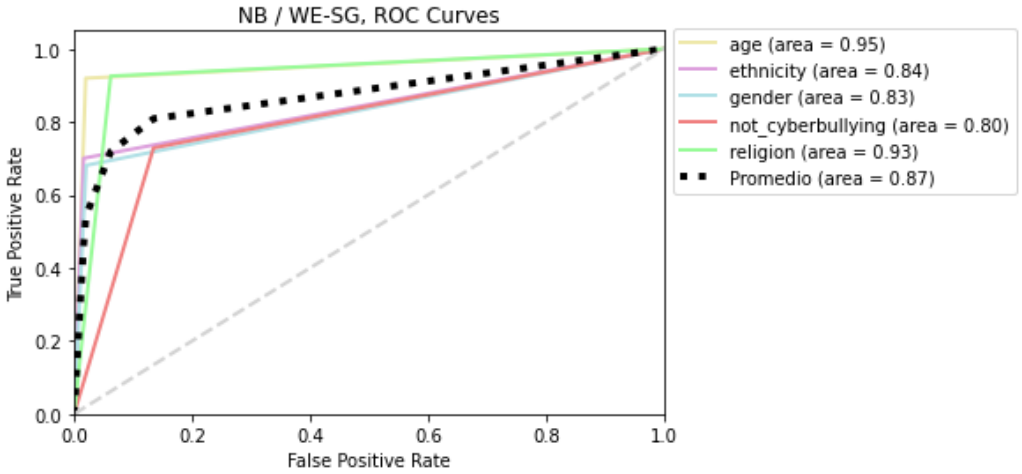


Figura 34. Análisis general de NB [elaboración propia].

Ha logrado su "accuracy" máxima con WE - SG, aunque alcanzando poco más de 0,79. Para este caso las curvas ROC son todas inferiores al resto de modelos (existen mayores errores en la clasificación de los comentarios), aunque sorprende cómo "ethnicity" ha decrecido tanto.

NB / WE-SG, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.93	0.92	0.92	2391
ethnicity	0.92	0.70	0.79	2234
gender	0.89	0.68	0.77	2263
not_cyberbullying	0.54	0.73	0.62	1998
religion	0.80	0.93	0.86	2392
accuracy			0.80	11278
macro avg	0.81	0.79	0.79	11278
weighted avg	0.82	0.80	0.80	11278

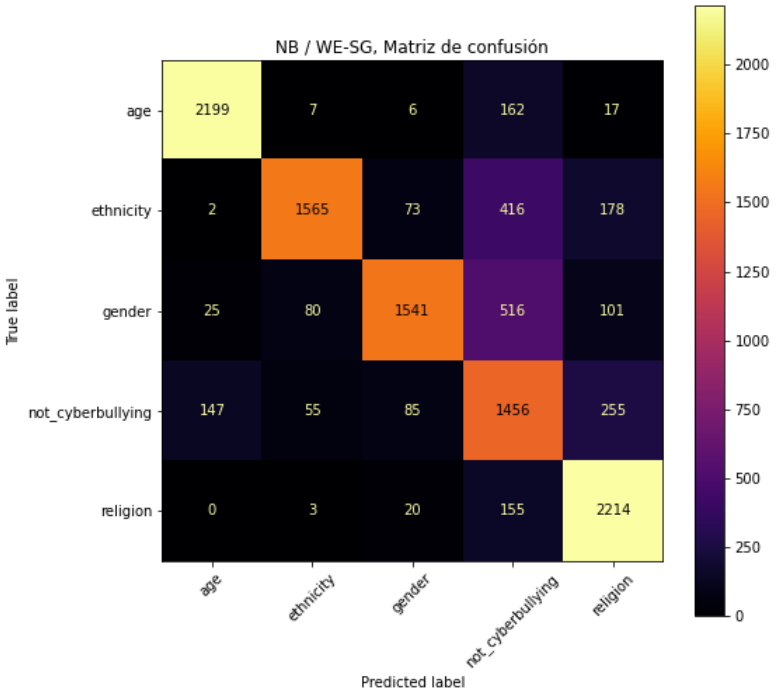
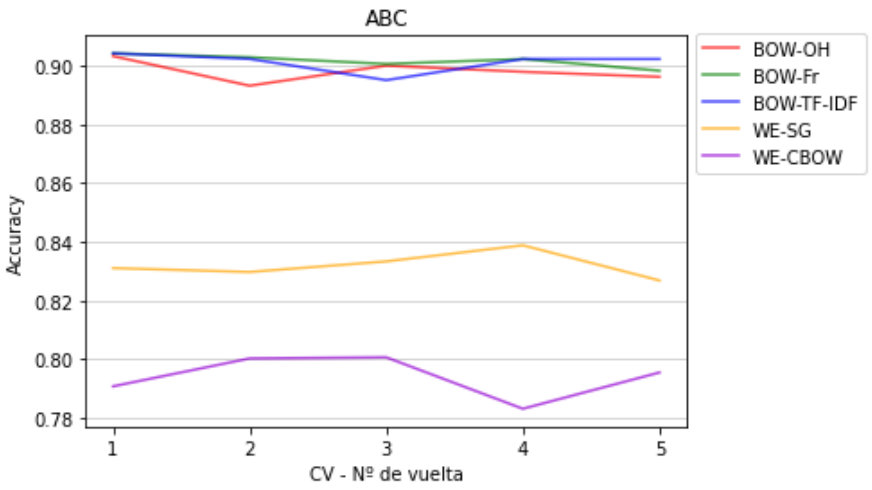


Figura 35 Análisis detallado de NB [elaboración propia].

Todas las clasificaciones han empeorado, incluso se observan asignaciones que antes no se producían, como el aumento de ejemplos de ”age”, “ethnicity” y “religion” a “not_cyberbullying”. Bastante malos resultados al ya haberse analizado los modelos previos.

E) Ada Boost Classifier (ABC)



	ABC
BOW-OH	0.898149
BOW-Fr	0.901683
BOW-TF-IDF	0.901266
WE-SG	0.831984
WE-CBOW	0.794094

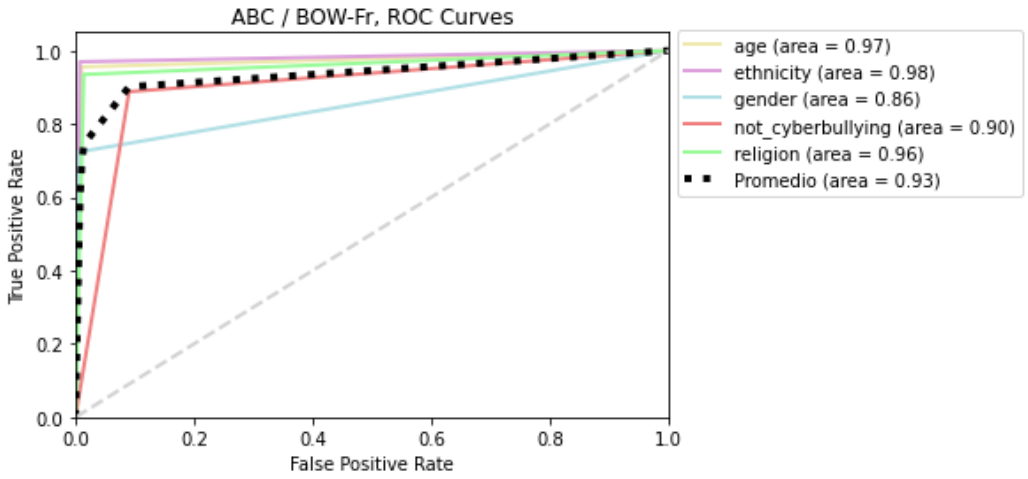


Figura 36. Análisis general de ABC [elaboración propia].

Con BOW - Fr ha logrado se máxima "accuracy" de 0,9016. Para este caso, y considerando otros resultados similares, lo que más resalta es el desplome de curva ROC de "gender", disminuyendo la clasificación correcta de sus propios comentarios.

ABC / BOW-Fr, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.96	0.96	0.96	2391
ethnicity	0.97	0.97	0.97	2234
gender	0.97	0.72	0.83	2263
not_cyberbullying	0.68	0.89	0.77	1998
religion	0.95	0.93	0.94	2392
accuracy			0.90	11278
macro avg	0.90	0.89	0.89	11278
weighted avg	0.91	0.90	0.90	11278

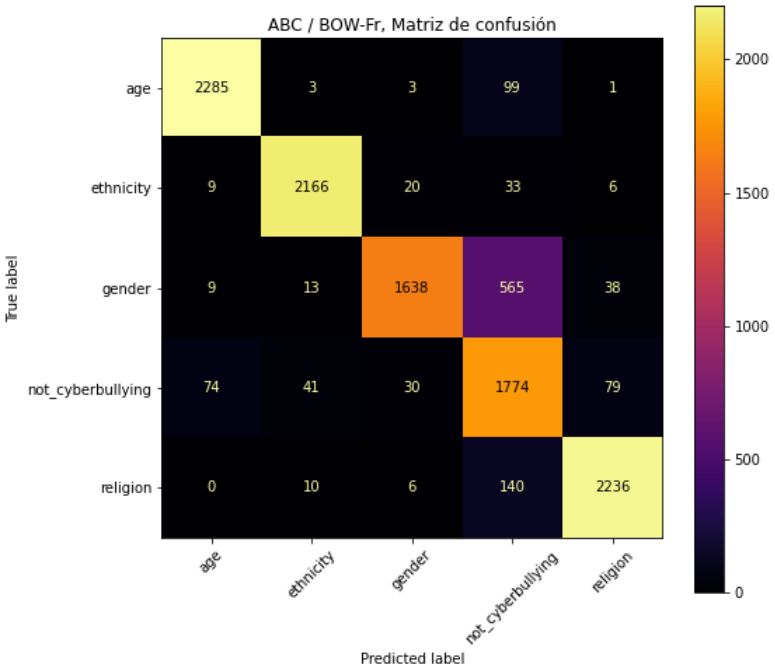
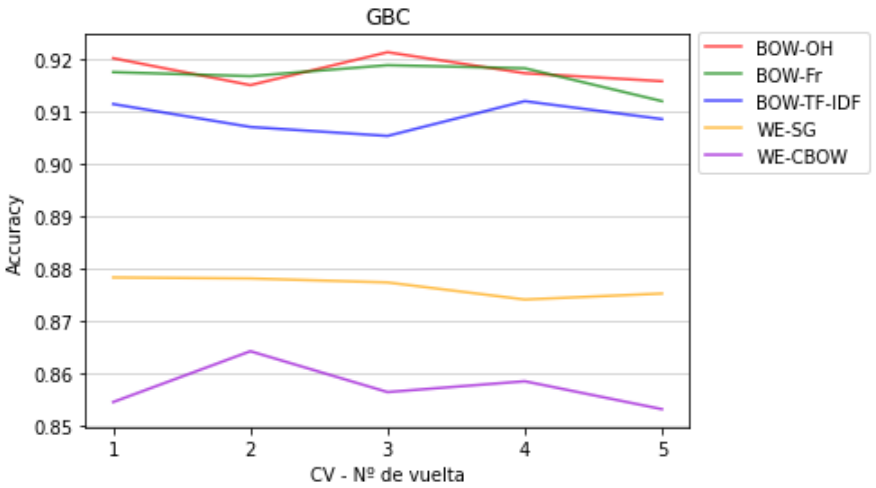


Figura 37 Análisis detallado de ABC [elaboración propia].

En este caso, el valor de "recall" para "not_cyberbullying" ha sido el mayor hasta el momento, indicando gran acierto entre sus propios ejemplos. Sin embargo, "gender" ha obtenido sólo un 0,72, clasificando gran parte de sus comentarios como "not_cyberbullying" y generando una gran reducción de "precision" de esta última.

F) Gradient Boosting Classifier (GBC)



	GBC
BOW-OH	0.917949
BOW-Fr	0.916695
BOW-TF-IDF	0.908904
WE-SG	0.876715
WE-CBOW	0.857485

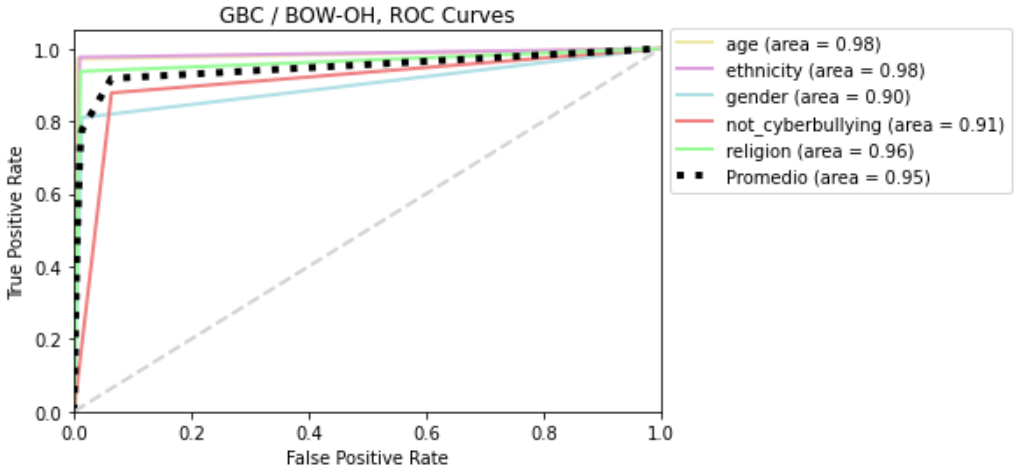


Figura 38. Análisis general de GBC [elaboración propia].

Ha alcanzado su valor máximo de "accuracy" con BOW - OH, obteniendo un 0,9179. Para este caso, se tienen curvas ROC similares a SVC / BOW - OH, con lo que se espera un patrón de comportamiento similar a dicho modelo.

GBC / BOW-OH, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.98	0.97	0.98	2391
ethnicity	0.96	0.98	0.97	2234
gender	0.94	0.81	0.87	2263
not_cyberbullying	0.75	0.88	0.81	1998
religion	0.96	0.94	0.95	2392
accuracy			0.92	11278
macro avg	0.92	0.91	0.91	11278
weighted avg	0.92	0.92	0.92	11278

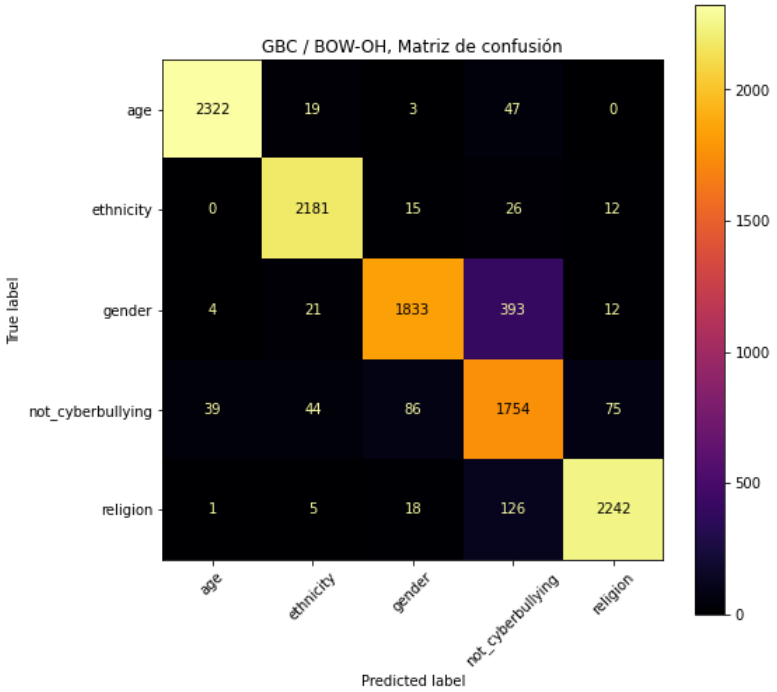
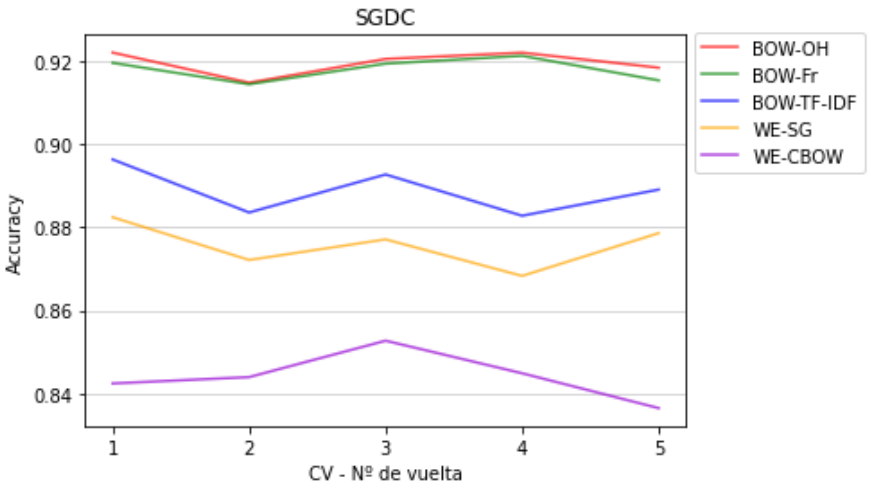


Figura 39 Análisis detallado de GBC [elaboración propia].

Métricas parecidas a SVC, aunque con una leve mejoría en la clase de "not_cyberbullying" y leve empeoramiento de "gender" y "religion".

G) Stochastic Gradient Descent Classifier (SGDC)



	SGDC
BOW-OH	0.919432
BOW-Fr	0.917873
BOW-TF-IDF	0.888838
WE-SG	0.875689
WE-CBOW	0.844145

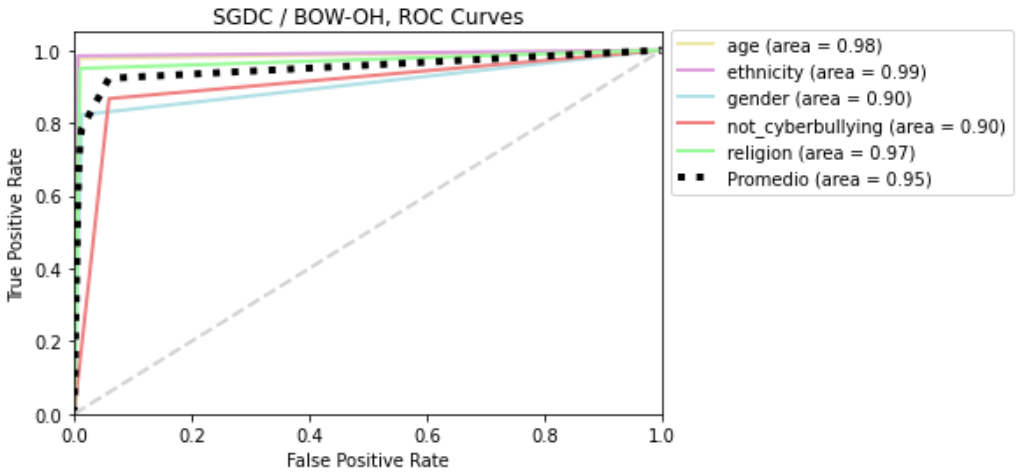


Figura 40. Análisis general de SGDC [elaboración propia].

Máximo "accuracy" de 0,9194 con, una vez más, BOW - OH. Las curvas ROC, para este caso, también son parecidas al modelo SVC / BOW - OH.

SGDC / BOW-OH, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.97	0.98	0.97	2391
ethnicity	0.97	0.98	0.98	2234
gender	0.94	0.82	0.88	2263
not_cyberbullying	0.76	0.87	0.81	1998
religion	0.96	0.95	0.96	2392
accuracy			0.92	11278
macro avg	0.92	0.92	0.92	11278
weighted avg	0.93	0.92	0.92	11278

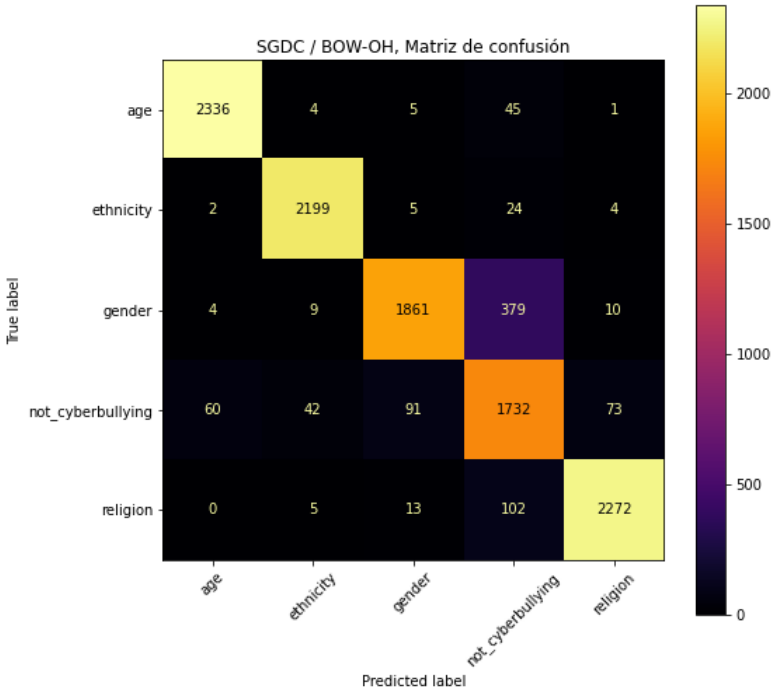
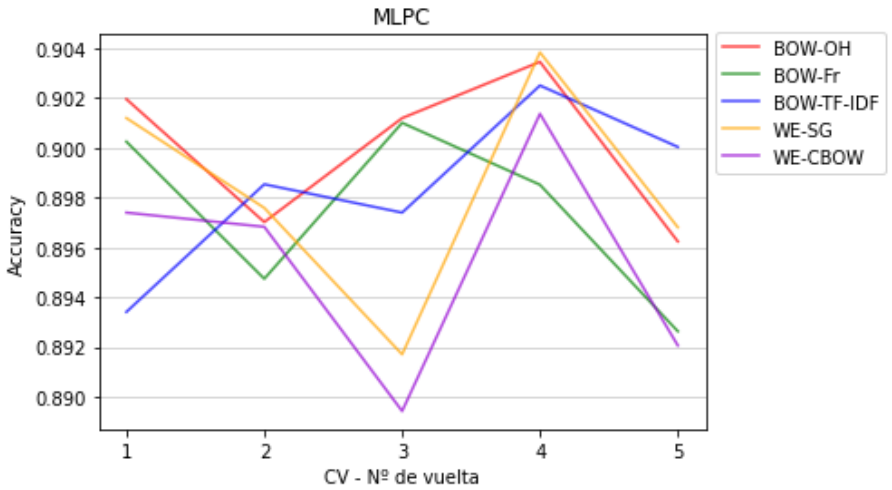


Figura 41. Análisis detallado de SGDC [elaboración propia].

En relación a SVC, ha clasificado de mejor manera "age" aunque, al igual que GBC, "gender" ha decrecido en "recall".

H) Multilayer Perceptron Classifier (MLPC)



	MLPC
BOW-OH	0.899973
BOW-Fr	0.897427
BOW-TF-IDF	0.898377
WE-SG	0.898225
WE-CBOW	0.895413

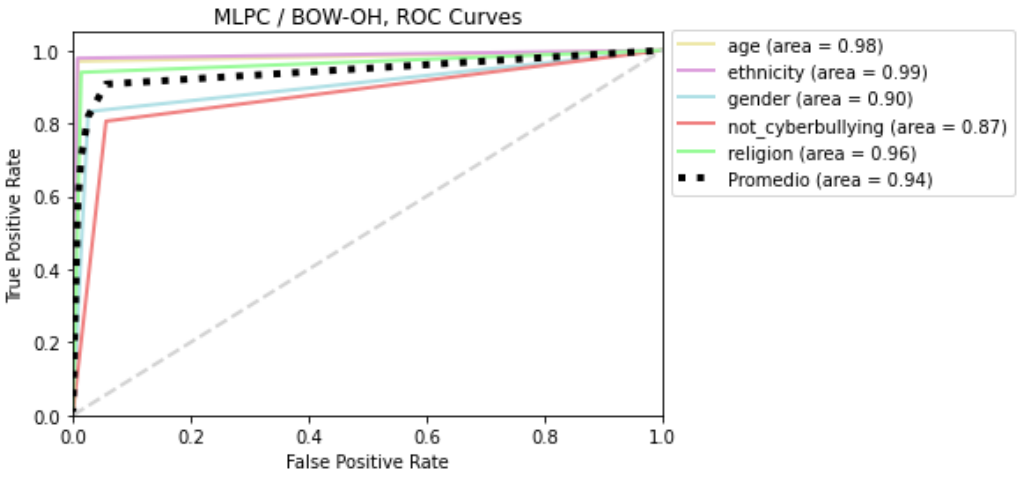


Figura 42. Análisis general de MLPC [elaboración propia].

Con unas "accuracies" gráficamente caóticas e impredecibles, ha alcanzado el valor máximo de casi 0,9 con BOW - OH. Sin embargo, la diferencia numérica entre éste y el menor valor (WE - CBOW) es muy pequeña, con lo que la disparidad de las curvas podría ser efecto del ruido en los datos. Aun así, llama la atención la línea de BOW - TF-IDF que parece que tuviera una tendencia estacional creciente con el aumento de vueltas de CV. De cualquier modo, y haciendo enfoque en la mejor métrica, es la primera vez que se observa que la curva ROC de "gender" encierra completamente la de "not_cyberbullying", indicando que las clasificaciones a su propia categoría serán peores.

MLPC / BOW-OH, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.97	0.97	0.97	2391
ethnicity	0.97	0.98	0.97	2234
gender	0.89	0.83	0.86	2263
not_cyberbullying	0.76	0.81	0.78	1998
religion	0.95	0.94	0.94	2392
accuracy			0.91	11278
macro avg	0.91	0.91	0.90	11278
weighted avg	0.91	0.91	0.91	11278

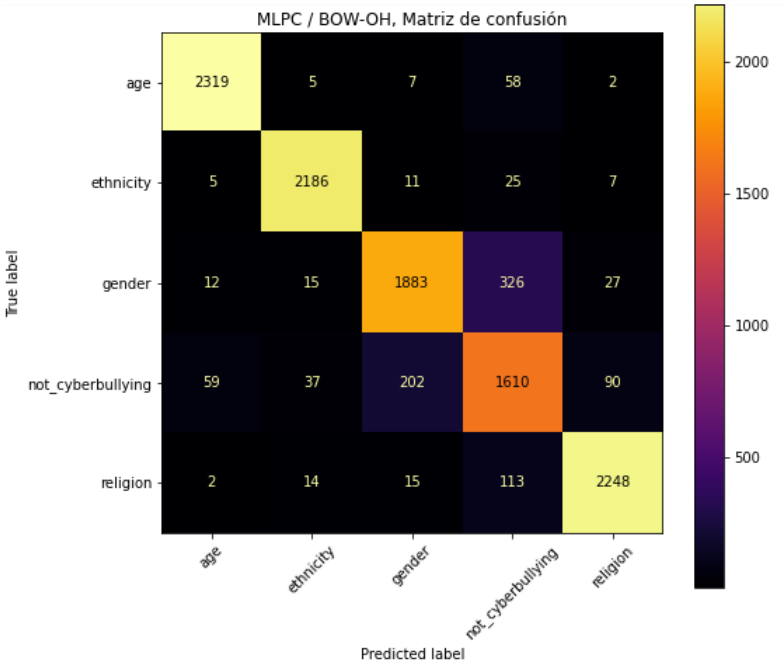


Figura 43. Análisis detallado de MLPC [elaboración propia].

Como se ha indicado previamente, la situación más notoria es la de "not_cyberbullying" desmejorando su desempeño y asignando gran cantidad de sus comentarios a la categoría de "gender".

Antes de continuar con LSTM y CNN, se presentarán resultados parciales de los modelos hasta ahora analizados ya que su forma de evaluación ha sido idéntica:

	LR	RFC	SVC	NB	ABC	GBC	SGDC	MLPC
BOW-OH	0.920116	0.918025	0.916581	0.777905	0.898149	0.917949	0.919432	0.899973
BOW-Fr	0.916467	0.917531	0.913275	0.762437	0.901683	0.916695	0.917873	0.897427
BOW-TF-IDF	0.877361	0.922814	0.884924	0.745791	0.901266	0.908904	0.888838	0.898377
WE-SG	0.867176	0.888458	0.879147	0.790598	0.831984	0.876715	0.875689	0.898225
WE-CBOW	0.825865	0.875765	0.855356	0.724280	0.794094	0.857485	0.844145	0.895413

Figura 44. Comparativa parcial de "accuracies" de los modelos [elaboración propia].

En la tabla de "accuracies", de los mejores resultados resaltados en color verde, el mayor valor corresponde al clasificador RFC / BOW - TF-IDF alcanzando un 0,9228.

También es interesante destacar cómo la "feature extraction" WE generalmente tiene los peores valores (en especial la variante CBOW poseyendo, siempre, las "accuracies" más bajas). Aunque pudiera sorprender por el auge del enfoque de aprendizaje del contexto, en realidad hay veces que WE puede funcionar peor que BOW, más que nada en situaciones en que palabras con distinto significado poseen contextos parecidos (por ejemplo palabras vecinas que se repiten en frases de diferentes acosos), haciendo que sus vectores sean similares y sea difícil distinguir su intención (clase).

	PRECISION		RECALL		F1-SCORE	
age	0.98	LR / BOW-OH RFC / BOW-TF-IDF SVC / BOW-OH GBC / BOW-OH	0.98	LR / BOW-OH SGDC / BOW-OH	0.98	LR / BOW-OH RFC / BOW-TF-IDF GBC / BOW-OH
ethnicity	0.99	RFC / BOW-TF-IDF	0.99	RFC / BOW-TF-IDF	0.99	RFC / BOW-TF-IDF
gender	0.97	ABC / BOW-Fr	0.84	RFC / BOW-TF-IDF	0.88	RFC / BOW-TF-IDF SGDC / BOW-OH
not_cyberbullying	0.78	RFC / BOW-TF-IDF	0.89	ABC / BOW-Fr	0.82	RFC / BOW-TF-IDF
religion	0.96	LR / BOW-OH RFC / BOW-TF-IDF SVC / BOW-OH GBC / BOW-OH SGDC / BOW-OH	0.96	RFC / BOW-TF-IDF	0.96	RFC / BOW-TF-IDF SVC / BOW-OH SGDC / BOW-OH

Figura 45. Comparativa parcial entre categorías y mejores métricas de clasificación indicando los modelos con igual desempeño [elaboración propia].

En la tabla anterior se enseñan resumidas las mejores métricas de los reportes de clasificación (valores en color verde) indicando, para cada categoría, a cuáles de los modelos con mayor "accuracy" corresponden. Concentrándose en "f1-score" que, como se ha visto previamente, valora tanto una alta "precision" como una alta "recall", es el clasificador RFC / BOW - TF-IDF el único que se encuentra en todas las categorías, significando que es el que mejor responde a la tarea (conjuntamente es el que más se repite, por eso está realzado en color azul).

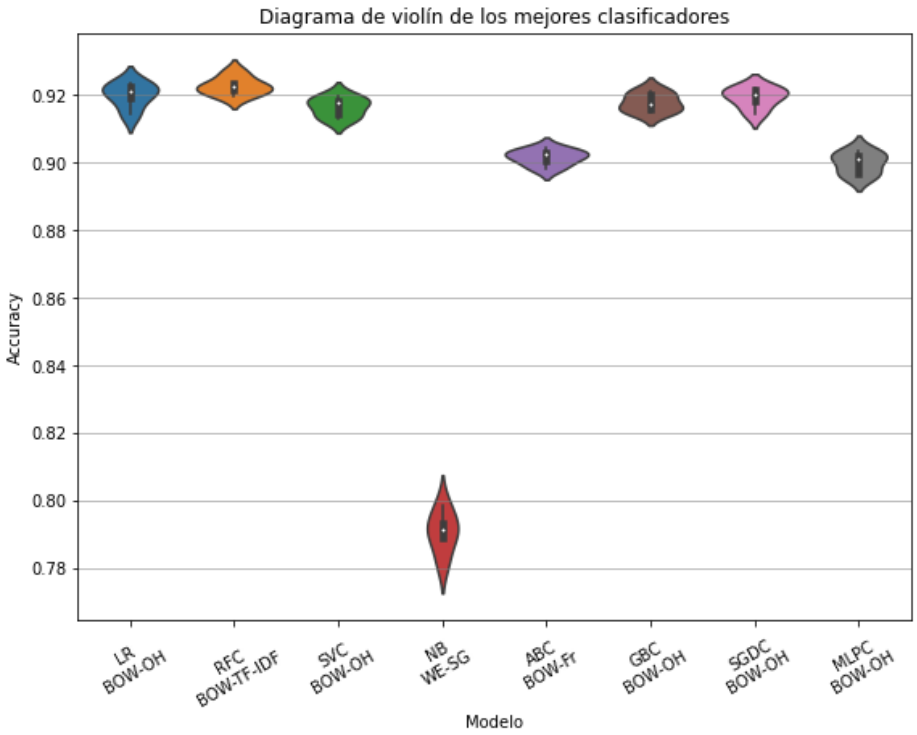
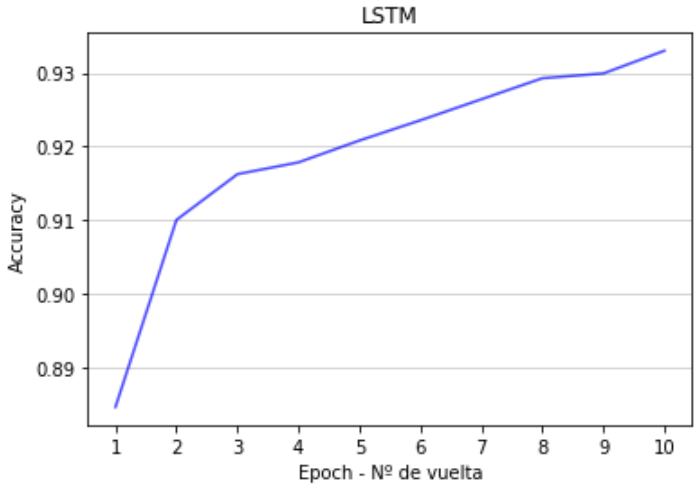


Figura 46. Diagrama de violín de "accuracies" de los modelos [elaboración propia].

De lo estudiado hasta aquí, se comprueba que el mejor clasificador es RFC / BOW - TF-IDF. Además, el diagrama de violín que enseña la distribución de las "accuracies" de las vueltas de CV, enseña que dicho modelo concentra sus valores en un rango pequeño y de alta eficacia. Así, será el elegido para continuar con la evaluación de los clasificadores restantes.

I) Long Short Term Memory (LSTM)




```
Epoch 10/10
823/823 [==============================] - 121s 148ms/step - loss: 0.1804 - accuracy: 0.9330
```

Figura 47. Análisis general de LSTM [elaboración propia].

Para unas “epochs” de 10, su valor máximo de “accuracy” es de 0,933.

LSTM / WE-SG, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.99	0.97	0.98	2391
ethnicity	0.98	0.98	0.98	2234
gender	0.97	0.80	0.88	2263
not_cyberbullying	0.75	0.91	0.82	1998
religion	0.96	0.95	0.96	2392
accuracy			0.93	11278
macro avg	0.93	0.92	0.92	11278
weighted avg	0.93	0.93	0.93	11278

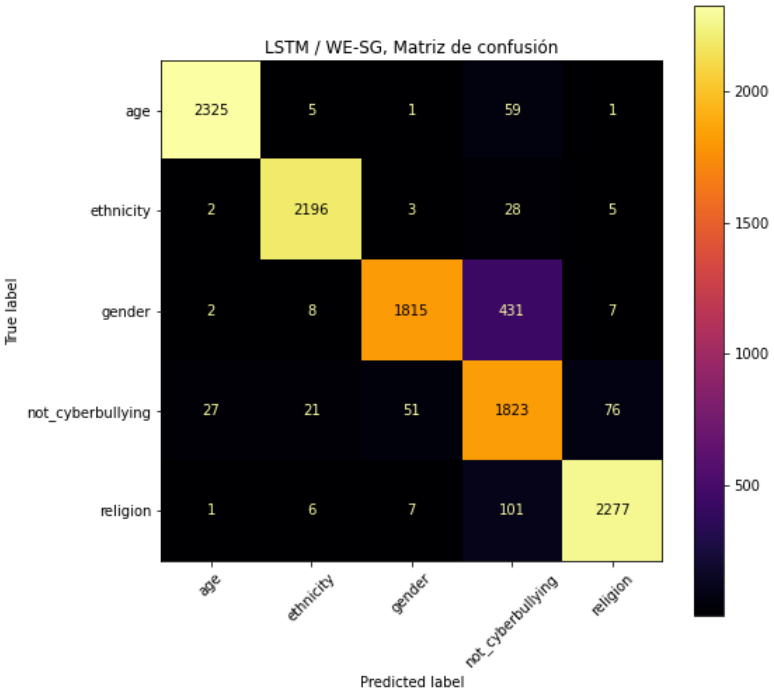
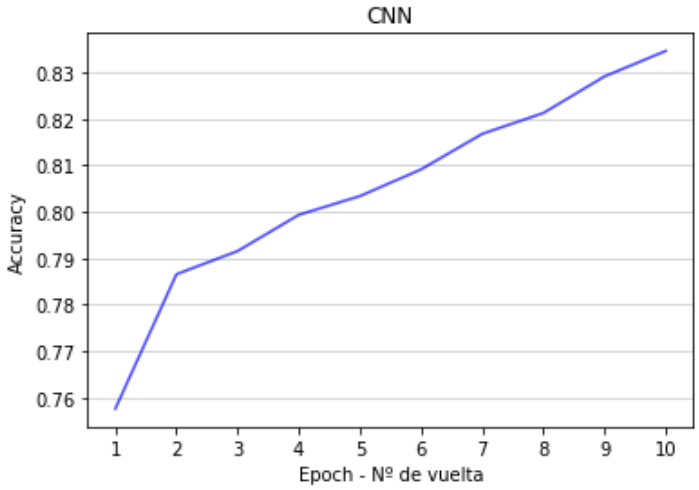


Figura 48. Análisis detallado de LSTM [elaboración propia].

Lo más interesante es la mejora de clasificación para los comentarios de "not_cyberbullying", alcanzando un valor de "recall" de 0,91, el mejor de lo observado hasta el momento. Sin embargo "gender" ha reducido su eficacia, generando peores resultados.

J) Convolutional Neural Network (CNN)




```
Epoch 10/10
823/823 [==============================] - 41s 50ms/step - loss: 0.4539 - accuracy: 0.8346
```

Figura 49. Análisis general de CNN [elaboración propia].

Para unas “epochs” de 10, su valor máximo de “accuracy” es de 0,8346.

CNN / WE-SG, Reporte de clasificación:

	precision	recall	f1-score	support
age	0.92	0.79	0.85	2391
ethnicity	0.88	0.81	0.84	2234
gender	0.85	0.63	0.72	2263
not_cyberbullying	0.50	0.82	0.62	1998
religion	0.91	0.83	0.87	2392
accuracy			0.78	11278
macro avg	0.81	0.78	0.78	11278
weighted avg	0.82	0.78	0.79	11278

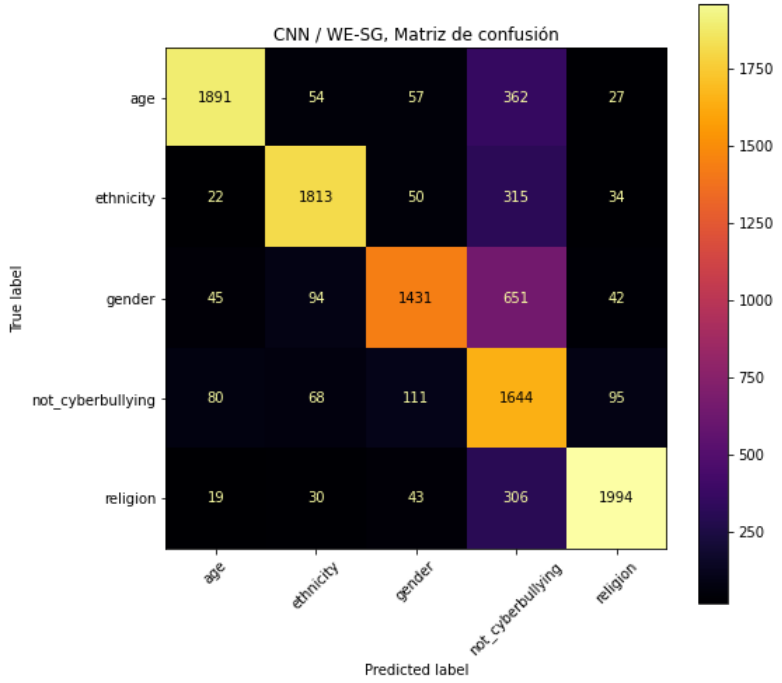


Figura 50. Análisis detallado de CNN [elaboración propia].

A pesar de tener un mayor valor de “accuracy” que NB, en general, produce muy malas asignaciones, incluso casi peores que éste. “not_cyberbullying” es la etiqueta que más errores de clasificación se lleva, quedando reflejado en su baja “precision”.

Habiéndose ya estudiado estos últimos dos modelos, queda realizar la comparativa final:

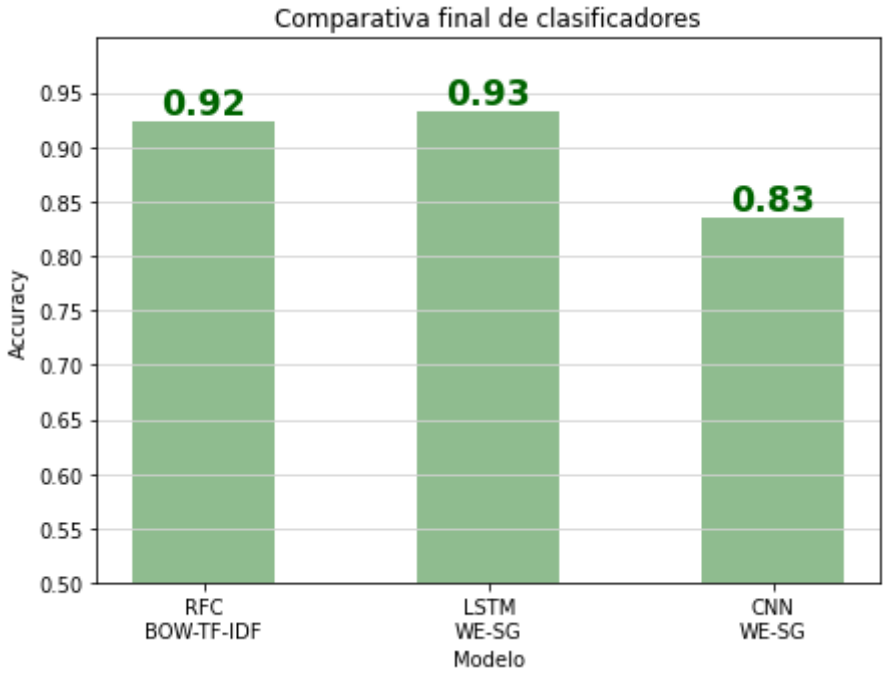


Figura 51. Comparativa final de "accuracies" de los modelos [elaboración propia].

	PRECISION		RECALL		F1-SCORE	
age	0.99	LSTM / WE-SG	0.97	RFC / BOW-TF-IDF LSTM / WE-SG	0.98	RFC / BOW-TF-IDF LSTM / WE-SG
ethnicity	0.99	RFC / BOW-TF-IDF	0.99	RFC / BOW-TF-IDF	0.99	RFC / BOW-TF-IDF
gender	0.97	LSTM / WE-SG	0.84	RFC / BOW-TF-IDF	0.88	RFC / BOW-TF-IDF LSTM / WE-SG
not_cyberbullying	0.78	RFC / BOW-TF-IDF	0.91	LSTM / WE-SG	0.82	RFC / BOW-TF-IDF LSTM / WE-SG
religion	0.96	RFC / BOW-TF-IDF LSTM / WE-SG	0.96	RFC / BOW-TF-IDF	0.96	RFC / BOW-TF-IDF LSTM / WE-SG

Figura 52. Comparativa final entre categorías y mejores métricas de clasificación indicando los modelos con igual desempeño [elaboración propia].

En primera instancia, queda en evidencia que CNN queda completamente descartado. Por el lado de RFC y LSTM, ambos resultados son muy buenos, alcanzando gran eficacia de clasificación. Sin embargo, a pesar de que LSTM posee mejor "accuracy", es nuevamente RFC el que destaca en la métrica "f1-score", siendo ésta más confiable a la hora de desempatar, indicando que dicho modelo responde mejor a la asignación de

acosos. Esto implica, que de tener que implementar una solución final, sería la combinación RFC / BOW - TF-IDF la escogida.

Igual, otro tema importante aquí es cómo mejorar las métricas de "not_cyberbullying", principalmente "precision". En este caso, los falsos positivos "FP" son especialmente importantes, puesto que clasificar un comentario de acoso "gender" como inofensivo o inocente, es más grave (con mayor coste) que al revés. Es decir, lo relevante es encontrar la forma de minimizar el "FPR" para esta clase, aunque a costa de un también menor "TPR / Recall" (que aquí sería aceptable).

Utilizar un mismo modelo multinomial que incluya también "not_cyberbullying" como posibilidad es complicado, ya que si se desea establecer un mayor umbral de clasificación (threshold) para dicha categoría, existen aún cuatro probabilidades más sobre las que decidir ("age", "ethnicity", "gender", "religion").

Para esto, podría ser interesante entrenar previamente un nuevo modelo general (incluso otro RFC) donde la salida de "cyberbullying_type" sea binaria: "0" para "not_cyberbullying" y "1" para las demás. Una vez entrenado el clasificador, para establecer "manualmente" el umbral (que por defecto es 0,5), se evaluaría la probabilidad asignada a la clase "0" y se definiría allí mismo si es ciberacoso o no:

```
umbral = 0.8 # sólo será "not_cyberbullying" de tener gran probabilidad de serlo
y_pred_proba = RFC.predict_proba(X_test)
y_pred = []
for x in y_pred_proba[:,0]: # se evalúan las probabilidades de "0" (not_cyberbullying)
    if x >= umbral:
        y_pred.append(0)
    else:
        y_pred.append(1)
```

Así, aquellos comentarios clasificados como "1", pasarían al segundo predictor RFC / BOW - TF-IDF que indicaría, en una segunda fase, el tipo de acoso. Vale aclarar que

este segundo modelo debería ser nuevamente entrenado descartando, esta vez, la categoría “not_cyberbullying” evaluada en la primera fase. Además, esta solución mejoraría notablemente las métricas de “gender”, categoría que más causaba conflictos en la “precision” de “not_cyberbullying”.

Capítulo 4. CONCLUSIÓN

La necesidad de aplicar medidas de protección frente a lo que viene siendo un problema que, principalmente en los últimos años, se ha ido incrementando, ha ocasionado gran interés por el desarrollo de herramientas que alerten la intención malintencionada de un texto.

Diferentes organizaciones y muchos investigadores a nivel mundial, han decidido poner manos a la obra en crear técnicas cada vez más refinadas de NLP y la detección del cyberbullying. Realmente, son muchas las mejoras alcanzadas y el progreso en el área es indudablemente un hecho. Incluso las mismas plataformas de redes sociales virtuales intentan contribuir para ofrecer un servicio más robusto y seguro.

Como trabajo de análisis futuro, se podrían aplicar técnicas adicionales para incrementar los resultados de los clasificadores. Por ejemplo, la aplicación de "transformers", redes neuronales con gran renombre en los avances de NLP, que utilizan técnicas matemáticas de "atención" para aprender el contexto y dependencia de las palabras. O incluso, añadir al estudio de los modelos la búsqueda de hiperparámetros "Grid Search" para encontrar las configuraciones óptimas.

De cualquier modo, aún queda bastante por seguir avanzando en este aspecto para generar un entorno virtual socialmente igualitario, porque no sólo depende de los avances en el procesamiento del lenguaje natural en este aspecto, sino también, del progreso en la tolerancia a nivel poblacional que aún dista mucho de aquella que debería ser la considerada "común o normal".

BIBLIOGRAFÍA

- **Libros**
  - EISENSTEIN, Jacob. 2019. **Introduction to Natural Language Processing.** EE. UU., The MIT Press.
  - GELBUKH, Alexander. 2007. **Special issue: Natural Language Processing and its Applications.** México, Instituto Politécnico Nacional, Centro de Investigación en Computación.
  - INDURKHYA, N., DAMERAU, F. 2010. **Handbook of Natural Language Processing.** EE. UU., Chapman & Hall/CRC.
  - TAHER PILEHVAR, M., CAMACHO-COLLADOS, J. 2018. **Embeddings in Natural Language Processing.** EE. UU., Morgan & Claypool. Pág. 10-11

- **Artículos científicos, investigaciones académicas, trabajos de campo**
  - ACI, C., SARAC, E., YILDIZ, E. 2019. **Automatic Detection of Cyberbullying in FormSpring.me, MySpace and YouTube social networks.** Turquía, Turkish Journal of Engineering.
  - GOMEZ, C., SZTAINBERG, M., TRANA, R. 2021. **Curating Cyberbullying Datasets: a Human-AI Collaborative Approach.** EE. UU., International Journal of Bullying Prevention.

- **Portales Web**
  - Cyberbullying Research Center. *What is Cyberbullying?* [en línea]. EE. UU.: Cyberbullying Research Center. Disponible en: <https://cyberbullying.org/what-is-cyberbullying>
  - Ditch The Label. *2017 Bullying Statistics – The Annual Bullying Survey 2017* [en línea]. Reino Unido: Ditch The Label, 2017. Disponible en: <https://www.ditchthelabel.org/research-papers/the-annual-bullying-survey-2017/>
  - Enough is Enough. *Cyberbullying Statistics* [en línea]. EE. UU.: Enough is Enough, 2022. Disponible en: <https://enough.org/stats_cyberbullying>
  - GeeksforGeeks. *Choose optimal number of epochs to train a neural*

network in Keras [en línea]. India: GeeksforGeeks, 2022. Disponible en: <https://www.geeksforgeeks.org/choose-optimal-number-of-epochs-to-train-a-neural-network-in-keras/>

- Google Developers. *Clasificación: Curva ROC y AUC* [en línea]. EE.UU.: Google Developers, Septiembre 2022. Disponible en: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es_419 >
- Henry, Maya. *How I knew I was Transgender* [en línea]. Canadá: YouTube, Julio 2020. Disponible en: <https://www.youtube.com/watch?v=iyeo5P6sWk0>
- Lobe, B., Velicu, A., Staksrud, E., Chaudron S., Di Gioia, R. *How children (10-18) experienced online risks during the Covid-19 lockdown - Spring 2020* [en línea]. EE. UU.: Publications Office of the European Union, 2021. Disponible en: <https://publications.jrc.ec.europa.eu/repository/handle/JRC124034>
- Lopez Yse, Diego. *Your Guide to Natural Language Processing (NLP)* [en línea]. Canadá: Towards Data Science, Enero 2019. Disponible en: <https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1>
- Lyashenko, V., Jha, A. *Cross-Validation in Machine Learning: How to Do It Right* [en línea]. Polonia: Neptune, 2022. Disponible en: <https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right>
- Narkhede, Sarang. *Understanding AUC - ROC Curve* [en línea]. Canadá: Towards Data Science, Junio 2018. Disponible en: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
- Riggio, Christopher. *What's the deal with Accuracy, Precision, Recall and F1?* [en línea]. Canadá: Towards Data Science, Noviembre 2019. Disponible en: <https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021>
- Security. *Cyberbullying: Twenty Crucial Statistics for 2022* [en línea]. EE. UU.: Security, Agosto 2022. Disponible en:

<https://www.security.org/resources/cyberbullying-facts-statistics/#impacts>

- Wang, J., Fu, K., Lu, CT. *SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection* [en línea]. EE. UU.: IEEE, 2020. Disponible en: <https://ieeexplore.ieee.org/document/9378065>