Universidad Nacional de
Educación a Distancia

Escuela Técnica Superior de
Ingeniería Informática

Detección de lenguaje ofensivo en redes sociales

Paula Ezquerro Abad

Director: Álvaro Rodrigo Yuste

Co-director: Jorge Pérez Martín

Trabajo de Fin de Máster

Máster Universitario en
Ingeniería y Ciencia de Datos

Septiembre 2023

Resumen

Desde su aparición a principios de la década de los 2000, el uso de las redes sociales se ha ido incrementando entre usuarios de todas las edades y procedencias del mundo. Este crecimiento masivo ha llevado a un aumento significativo en la interacción y la comunicación en línea. Las redes sociales son la principal herramienta de comunicación y una de las principales fuentes de información entre la mayor parte de la población mundial.

Sin embargo, a pesar de traer consigo una amplia gama de beneficios, recientemente han surgido desafíos relacionados con el lenguaje ofensivo y el discurso de odio en ellas.

Todas las redes sociales disponen de sistemas de moderación, algunos de los cuales son manuales y bastante tediosos. Hoy en día, sigue siendo un reto conseguir un sistema que sea rápido, eficaz y que aprenda igual de rápido que evoluciona el lenguaje humano. Las redes sociales serían un ecosistema mucho más seguro y sano si se consiguieran detectar este tipo de mensajes y frenar su publicación en tiempo real.

En este trabajo se estudia la idoneidad de los modelos basados en Transformers, considerados el estado del arte en el campo del procesamiento de lenguaje natural, para detectar mensajes ofensivos en redes sociales. Se investigan los modelos más actuales de esta área para encontrar aquel que esté más cerca de lograr el objetivo. Se han realizado experimentos entrenando con fuentes de datos diferentes a las que fueron entrenados para probar si mejora su generalización.

Una vez realizados los experimentos que se detallan en este trabajo se ha llegado a la conclusión de que actualmente se dispone de una serie de modelos pre-entrenados que pueden suponer una muy buena base para el desarrollo de estos sistemas. Estos modelos requieren de una correcta fase de entrenamiento que permita mejorar sus métricas y generalizarlo a todas las redes sociales y usuarios del mundo.

Abstract

Since its appearance in the early 2000s, the use of social networks has increased among users of all ages and backgrounds around the world. This massive growth has led to a significant increase in online interaction and communication. Social networks are the main communication tool and one of the main sources of information among most of the world's population.

However, despite bringing a wide range of benefits, challenges related to offensive language and hate speech in them have recently emerged.

All social networks have moderation systems, some of which are manual and quite tedious. Today, it remains a challenge to achieve a system that is fast, efficient and that learns as quickly as human language evolves. Social networks would be a much safer and healthier ecosystem if these types of messages could be detected and their publication stopped in real time.

This project studies the suitability of Transformers-based models, considered the state of the art in the field of natural language processing, to detect offensive messages on social networks. The most current models in this area are investigated to find the one that is closest to achieving the objective. Experiments have been carried out training with data sources different from those on which they were trained to test if their generalization improves.

Once the experiments detailed in this work have been carried out, it has been concluded that there is currently a series of pre-trained models that can provide a very good basis for the development of these systems. These models require a correct training phase that allows them to improve their metrics and generalize them to all social networks and users in the world.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1. Introducción

En este capítulo se explicará la situación y necesidades actuales que motivaron la realización de este proyecto, así como los objetivos que se persiguen y la distribución que se ha escogido para la redacción de la memoria del proyecto.

1.1 Motivación

La Real Academia Española [1] define ofender como humillar o herir el amor propio o la dignidad de alguien o ponerlo en evidencia con palabras o con hechos. La aparición de Internet y las redes sociales supuso un cambio de paradigma en lo que a la comunicación entre personas se refiere, permitiendo un contacto fácil y anónimo entre todo tipo de personas, lo que ha propiciado el aumento de las acciones dirigidas a ofender a otros usuarios o colectivos.

Las nuevas tecnologías nos han dado la posibilidad de permanecer en contacto sin límites temporales, geográficos ni culturales. Pero también ha dado lugar a un nuevo método de difamación con mucho más alcance que los tradicionales. Cualquier usuario tiene poder suficiente para ofender de forma pública a cualquier persona del mundo, incluso manteniendo

Hemos desarrollado herramientas suficientes para hacer de Internet un espacio virtual común, pero nos falta tener suficiente tecnología regulatoria para que este vaya en paralelo con los valores básicos de la humanidad. (Kaufman, 2016) lo compara con un jardín, en el que todos queremos ver flores maravillosas, pero nadie quiere que haya un jardinero que decida qué flores merecen crecer y cuales merecen ser cortadas. Pero sin este rol Internet se está convirtiendo en una mezcla de jardines y tierras maltratadas que coexisten en el espacio virtual y con las que nos encontramos día a día. El autor afirma que Internet contiene lo mejor y lo peor de lo que ha producido la humanidad.

Según el estudio anual de redes sociales realizado por la IAB[2] (Interactive Adversiting Bureau) en 2022 el 85% de los internautas españoles de entre 12 y 70 años usaron redes sociales. La Figura 1 muestra las 5 redes sociales más usadas entre los encuestados.

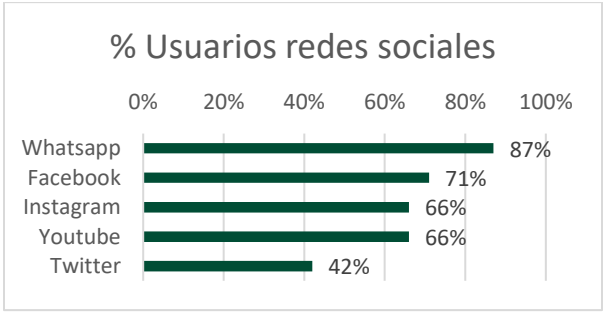

Figura 1. Porcentaje de usuarios por red social en 2022 según la IAB

[1] https://dle.rae.es/
[2] https://iabspain.es/

La existencia de este gran número de usuarios ha obligado a las redes sociales a adoptar diferentes estrategias de moderación de contenidos:

- Whatsapp: Permite denunciar a un contacto que esté enviando contenido inadecuado. En ese momento los moderadores reciben los últimos 5 mensajes del contacto[3], saltándose en este caso el cifrado de extremo a extremo del que disponen los chats. También puede reportarse únicamente un mensaje. En ambos casos, los moderadores pueden eliminar al usuario reportado. Además, cualquier usuario puede bloquear a otro en cualquier momento.

- Facebook: Dispone de entre 800 y 1000 moderadores en todo el mundo, los cuales reciben las denuncias que los usuarios hacen de otros usuarios o de las publicaciones que consideran inapropiadas. Tras un estudio por un equipo de moderadores se decide si se elimina la publicación, si se inactiva la cuenta del usuario, si no se hace nada o si se escala a las autoridades. Permite bloquear a otro usuario.

- Instagram: Permite bloquear la cuenta de un usuario o todas las cuentas que un usuario tenga y pueda tener a futuro. Además, desde hace un par de años dispone de la funcionalidad "Hidden Words"[4] que permite, a los usuarios que la activen, no visualizar comentarios o mensajes que sean clasificados como ofensivos por la herramienta. Adicionalmente los usuarios pueden decidir estos filtros a mano, de forma que no visualizan mensajes o comentarios que contengan las frases o palabras que ellos decidan. Los usuarios que realizan este tipo de acciones de forma recurrente acaban por ser bloqueados de la red social, impidiéndole generar nuevas cuentas ni usar las existentes.

- Youtube: Permite denunciar un vídeo, un canal o un comentario al equipo de moderadores, que decidirán si el recurso incumple o no las normas de la comunidad. Dispone de un sistema de faltas, según el cual un usuario puede acumular hasta 2 faltas[5] sin que se le borre la cuenta.

- Twitter: Permite bloquear usuarios para evitar que interactúen con nosotros, además de reportar contenido inadecuado para que sea revisado por la compañía[6].

La mayor parte de estas medidas se engloban en el tipo de moderación reactiva[7]. Lo que significa que los usuarios son los responsables de denunciar los contenidos que ya se han publicado. Esto implica que ya se ha realizado la ofensa, que el receptor ya ha sufrido las consecuencias y que esta puede haber sido vista por un gran número de personas.

Existen trabajos en el campo de la detección de lenguaje ofensivo en redes sociales mediante sistemas automáticos, algunos de los cuales pueden conocerse en el Capítulo 3. Sin embargo, estos se entrenan con datos muy similares a los de evaluación, lo que dificulta su generalización. La idea de este trabajo es seguir

[3] https://www.europapress.es/portaltic/socialmedia/noticia-whatsapp-accedera-ultimos-mensajes-chat-verificar-denuncias-20210902152906.html
[4] https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse
[5] https://support.google.com/youtube/answer/2802032?hl=es
[6] https://help.twitter.com/es/rules-and-policies/twitter-report-violation
[7] https://www.cyberclick.es/numerical-blog/moderacion-de-contenidos-que-es-y-mejores-herramientas

avanzando en las moderaciones automáticas probando a entrenar y probar modelos con diferentes conjuntos de datos, estudiando así su generalización.

1.2 Propuesta y objetivos

El objetivo general de este trabajo es estudiar la generalización de modelos basados en Transformers en la tarea de detección de mensajes ofensivos, con el fin de poder aplicarlo en la moderación automática de las redes sociales.

Con la realización de este trabajo se persiguen los siguientes objetivos:

- Comparar los resultados de distintos modelos entrenados sobre colecciones distintas a aquellas utilizadas para su evaluación.
- Estudiar el uso de distintas estrategias de entrenamiento.

1.3 Estructura del documento

La memoria de este proyecto presenta las siguientes secciones:

- **Introducción:** Motivación de realizar el proyecto, objetivos que se persiguen con él y estructura de este documento.
- **Estado del Arte:** Tareas y proyectos que se están realizando actualmente en materia de detección de lenguaje ofensivo en redes sociales mediante aprendizaje automático.
- **Materiales y métodos**: Entornos usados para desarrollar el proyecto, conjuntos de datos usados y métricas de evaluación seleccionadas para comparar los resultados de las diferentes pruebas del proyecto.
- **Resultados:** Presentación de las pruebas realizadas y resultados devueltos por estas según las métricas escogidas.
- **Conclusiones y trabajos futuros:** Conclusiones de la tarea tras la realización del proyecto y posibles próximos proyectos para avanzar en la materia.
- **Bibliografía y referencias:** Recursos utilizados para la realización del proyecto.

Capítulo 2. Marco teórico

En este apartado se van a exponer las definiciones y conceptos que se nombran a lo largo de las secciones de este documento y conviene conocer antes de leerlo para comprenderlo en su totalidad y tomar conciencia de la evolución de los modelos en el marco de la tarea.

2.1 Aprendizaje automático

El aprendizaje automático según (Judith Sandoval, 2018) es la rama de la Inteligencia Artificial mediante el cual se generan modelos que generan predicciones a partir de unos datos de entrada. Estos modelos están basados en algoritmos matemáticos que identifican patrones en los datos recibidos y van aprendiendo y mejorando su rendimiento cuantos más datos van consumiendo.

Dentro del aprendizaje automático se distinguen dos categorías:

- Aprendizaje supervisado: Modelos que han sido entrenados con un conjunto de datos etiquetados. Han requerido de la intervención humana para preparar los datos de entrenamiento.
- Aprendizaje no supervisado: Modelos que han sido entrenados con datos sin etiquetar. La intervención humana es necesaria únicamente para establecer los parámetros del modelo.

2.2 Aprendizaje profundo

El aprendizaje profundo según (Bengio et al., 2012) es un tipo de aprendizaje automático que busca simular la forma en la que el cerebro humano obtiene los conocimientos. Para ello se relacionan los nuevos aprendizajes con los ya adquiridos. Se dispone de una serie de algoritmos organizados de forma jerárquica en capas por los que los datos de entrada van pasando hasta que se consigue un nivel de precisión aceptable.

Este tipo de procesamiento recibe el nombre de **red neuronal**, ya que esta estructura de capas conectadas tiene cierta similitud con las conexiones neuronales del cerebro de los seres vivos. Están formadas por un conjunto de neuronas conectadas entre sí que trabajaban en conjunto. Se dispone de una capa de entrada (recibe los datos), una o varias capas ocultas (generan las predicciones) y una capa de salida (devuelve el resultado).

La Figura 2 muestra un esquema de este tipo de modelo. En esta figura la capa de entrada se representa con círculos verdes, hay dos capas ocultas de color naranja y capa de entrada es la de color azul.

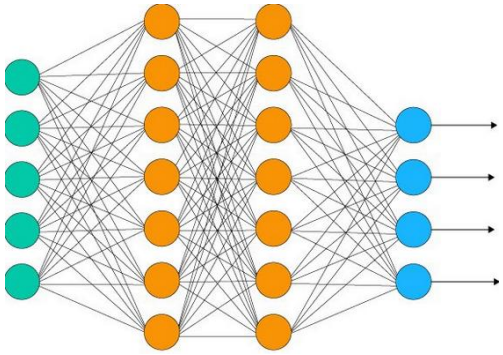

Figura 2. Esquema red neuronal

Uno de los modelos más conocidos de este tipo de aprendizaje son las **máquinas de vectores de soporte (SVM)**. Estos modelos representan los datos como puntos en el espacio y generan un hiperplano que los separa de la forma más amplia posible. Las nuevas muestras se clasifican según la parte del hiperplano en la que se representen. Los vectores de soporte representan la distancia mínima que deben tener las muestras al hiperplano, como puede observarse en la Figura 3. Máquinas de vectores de soporte.

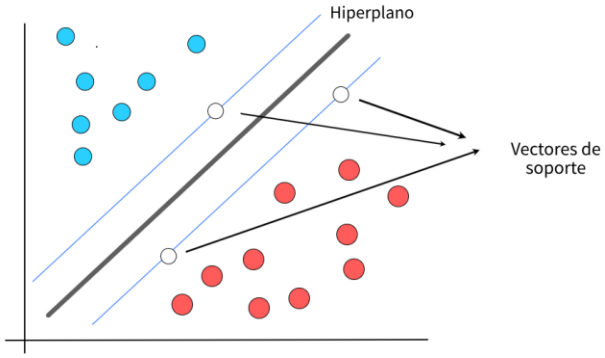

Figura 3. Máquinas de vectores de soporte

Para los casos en los que no se puede realizar una separación lineal existe el modelo **perceptrón multicapa (MLP)**, que posee una capacidad de generalización muy alta. Sin embargo, al tener una estructura más compleja puede resultar computacionalmente más costoso y requerir de un conjunto de entrenamiento muy extenso y una elección rigurosa de los parámetros.

Dentro de las redes neuronales podemos encontrar las **redes neuronales recurrentes (RNN)**. Mientras que las redes neuronales tradicionales asumían que las entradas y salidas son independientes entre sí, las RNN tienen "memoria". Es decir, almacenan información de entradas anteriores para influir en posteriores entradas y salidas.

Existen las llamadas redes **LSTM**, capaces de "recordar" información más a largo plazo que las RNN, que solo recuerdan relaciones cercanas en el tiempo, las cuales eran consideradas el estado del arte antes de la aparición de los Transformers.

2.2.1 Transformers

Hace pocos años de presentaron los **Transformers** (Vaswani et al., 2017). Estos modelos surgieron a partir de las redes neuronales, manteniendo el proceso de codificador-decodificador, pero simplificando la estructura al sustituir las capas de recurrencia por un mecanismo de atención para generar dependencias. La atención permite detectar y tener en cuenta las relaciones entre palabras de una oración, con el fin de comprender mejor su significado. De esta manera cuando se está procesando una palabra se sabe con que otra u otras palabras de la entrada completa está relacionada.

La llegada de los Transformers supuso grandes ventajas sobre las RNN. Mientras que las RNN requieren que los datos de entrada estén ordenados (van procesando las relaciones entre las entradas de forma secuencial), los Transformers no necesitan que se siga ningún orden (pueden procesar todas las entradas al mismo tiempo), lo que les otorga una mayor capacidad de paralelización.

La Figura 4 obtenida de (Alammar, 2018) muestra la estructura de este tipo de modelos.

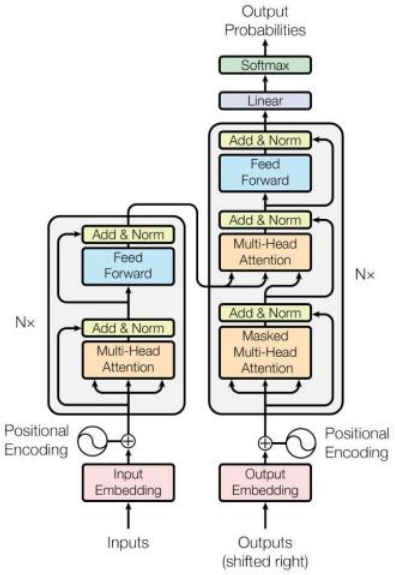

Figura 4. Estructura Transformers

La estructura es la siguiente:

- Codificador: Existe una pila de N=6 codificadores cada uno de los cuales tiene dos subcapas (auto-atención y red de retroalimentación). Cada capa ejecuta su función y envía su salida a la siguiente. Para facilitar estas conexiones todas las capas producen salidas de dimensión 512. Los codificadores se encargan de modificar los datos de entrada buscando una representación que el modelo pueda interpretar. El hecho de disponer de una capa de auto-atención permite que la codificación se mejore teniendo en cuenta las relaciones entre la entrada que se está procesando y las demás.

- Decodificador: También se compone de una pila de N=6 capas idénticas. Además de las dos subcapas descritas anteriormente tiene una tercera subcapa de atención entre las dos anteriores. Los decodificadores reciben las salidas de la codificación y realizan las predicciones. Esta salida es enviada a una capa lineal y posteriormente a la capa softmax, que devuelve las probabilidades de cada salida.

La Figura 5 muestra de forma esquemática la codificación y decodificación.

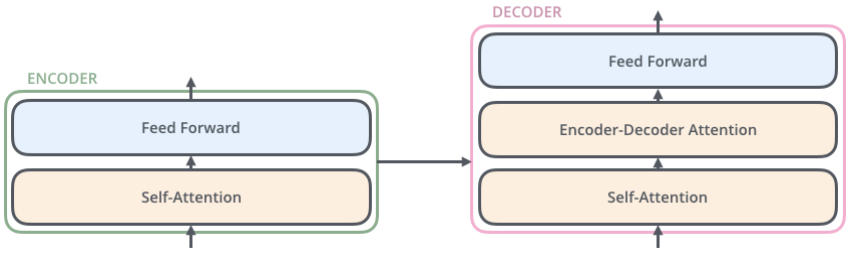

Figura 5. Codificador y decodificador de un Transformer

Algunos de los modelos basados en Transformers más usado son:

- **BERT (Bidirectional Encoder Representations from Transformers):** Modelo de aprendizaje profundo desarrollado por Google en 2018 y presentado en (Devlin et al., 2018). Este modelo ha demostrado ser eficaz en amplia gama de tareas entre las que se encuentran, por ejemplo: análisis de sentimientos, resumen de textos, reconocimiento de entidades nombradas, sistemas de preguntas y respuestas, traducción, etc. Para el uso en español surgió el modelo BETO[8], entrenado con un conjunto de datos en español.

- **GPT-4:** Modelo desarrollado en marzo de 2023 por OpenAI[9]. Este modelo sigue a sus predecesores, GPT (2018), GPT-2 (2019), GPT-3 (2020) y GPT-3.5 (2022), desarrollados también por la misma compañía. GPT-4 (OpenAI, 2023) está basado en Transformer y permite entradas de imágenes y texto para producir salidas de textos. Supone una mejora sustancial respecto a sus anteriores modelos al conseguir captar en mayor medida la intención del usuario. Su principal aplicación es ChatGPT [10] , un sistema de chat que permite mantener una conversación en la que se le solicita y pregunta al modelo cualquier tipo de información.

[8] https://github.com/dccuchile/beto
[9] https://openai.com/
[10] https://chat.openai.com

- **T5 (Text to Text Transfer Transformer)**[11]: Modelo basado en Transformers lanzado por Google en 2019 presentado en el artículo (Raffel et al., 2019). Se basa en la idea de recibir un texto de entrada y producir un texto de salida. Este paradigma permite simplificar el proceso al aplicar el mismo modelo, proceso de entrenamiento y proceso de decodificación para todas las tareas. Fue entrenado con el conjunto de datos de C4, recopilado por Common Crawl[12] a través del rastreo de contenidos en la web.

2.2.2 Transfer learning

El aprendizaje por transferencia se define en (Pan & Yang, 2010) como un método de aprendizaje automático en el que un modelo desarrollado para una tarea concreta se puede usar como punto de partida para el desarrollo de otro modelo para una tarea distinta. Utilizar un modelo pre-entrenado como punto de partida permite desarrollar modelos eficaces de forma rápida y ahorrando muchos recursos.

Este método se basa en explorar la generalización de los modelos, utilizar el conocimiento que un modelo ha aprendido con una tarea con muchos datos de entrenamiento para una tarea de la que no se tienen tantos datos.

Se distinguen 3 tipos de aprendizaje por transferencia:

- **Aprendizaje por transferencia inductivo:** Las tareas fuente y objetivo son distintas, pero los dominios de origen y destino son el mismo. El modelo previamente entrenado ya conoce las características del dominio, lo que le da ventaja respecto a si se entrenara de cero.
- **Aprendizaje por transferencia no supervisado:** Similar al anterior. La tarea objetivo es distinta de la tarea fuente, pero está relacionada con ella.
- **Aprendizaje por transferencia transductivo**: Las tareas de origen y destino son las mismas, mientras que los dominios de origen y destino son diferentes. No hay datos etiquetados en el dominio de destino, pero sí en el de origen.

Existen diversas tareas en las que se estudia la eficacia de este método de aprendizaje, algunas de las cuales se detallan a continuación:

- **Respuesta a preguntas mediante transferencia de aprendizaje a partir de grandes datos de supervisión detallados** (Min et al., 2017): Utilizan el modelo el aprendizaje por transferencia en la tarea de "Preguntas y respuestas". El experimento consiste en entrenar el modelo BiDAF (Seo et al., 2016) en el conjunto de datos SQuAD (Rajpurkar et al., 2016) y, posteriormente evaluarlo en los conjuntos de datos WikiQA (Yang et al., 2015) y SemEval2016 (Nakov Lluís et al., 2016). La tarea demostró que los resultados del modelo mejoraban hasta en un 8% tras este re-entrenamiento y aplicación a estos conjuntos de datos.

- **Ajuste del modelo de lenguaje universal para la clasificación de textos** (Howard & Ruder, 2018)**:** Aportan el Modelo de Lenguaje Universal Ajuste Fino, un método que permite lograr aprendizaje por transferencia muy eficiente para cualquier tarea de procesamiento de lenguaje

[11] https://github.com/google-research/text-to-text-transfer-transformer
[12] https://commoncrawl.org

natural. Una de las pruebas que realizan para evaluar su método es el análisis de sentimientos en opiniones de películas en IMDb[13] y Yelp[14]. Comprueban que este método mejora los resultados de sus antecesores, reduciendo las tasas de error.

- **Uso de millones de apariciones de emoji para detectar sentimientos, emociones y sarcasmo** (Felbo et al., 2017): Aplican la interpretación de los emojis para la comprensión de emociones de los Tweets. Para ello realizan una correlación entre tweets y emojis que permite identificar los sentimientos contenidos en estos. La clasificación se realiza con un modelo LSTM, al cual le aplican un método de "descongelación en cadena" que permite ir descongelando secuencialmente y afinando una sola capa a la vez.

[13] https://www.imdb.com/
[14] https://www.yelp.es/

Capítulo 3. Tareas de evaluación de lenguaje ofensivo

En esta sección se muestran algunas de las tareas realizadas en los últimos años con estos modelos en la tarea de detección de lenguaje de ofensivo.

3.1 OffensEval[15]

Conjunto de tareas compartidas sobre identificación y categorización de lenguaje ofensivo organizadas en el Taller Internacional sobre Evaluación Semántica (SemEval[16]). Las tareas se realizaron en dos ediciones, las cuales se detallan en las siguientes secciones.

3.1.1 OffensEval 2019

En OffensEval 2019 (Zampieri et al., 2019b) se utilizó el conjunto de datos OLID, el cual está formado por 14000 tweets en inglés. Con este se realizaron 3 subtareas cada una de las cuales está basada en la anterior:

- **Subtarea A:** Su objetivo era la clasificación de las publicaciones en ofensivas y no ofensivas.
- **Subtarea B:** Su objetivo era clasificar el contenido ofensivo según si era dirigido (insulta o amenaza a un individuo o grupo) o no dirigido (contienen palabras ofensivas, pero no están dirigas a nadie).
- **Subtarea C:** Su objetivo era clasificar las publicaciones ofensivas y dirigidas según su objetivo en individuales (ofensa dirigida a una persona concreta), grupo (ofensa dirigida a un colectivo político, religioso, étnico, etc) u otros (ofensa dirigida a situaciones, eventos, sucesos,etc).

En la primera subtarea participaron 104 equipos. Entre los 10 que mejores resultados obtuvieron se encuentran 7 que utilizaron BERT, con variaciones en los parámetros usados y los pasos de preprocesamiento. De estos el que consiguió un mejor rendimiento fue el equipo NULI (P. Liu et al., 2019), que utilizó BERT-base-uncased con parámetros predeterminados, una longitud máxima de 64 y entrenado para 2 épocas. Este equipo consiguió un valor de 0,829 para F1.

En la segunda subtarea participaron 71 equipos. A diferencia de la subtarea anterior, en esta 5 de los 10 mejores equipos usaron modelos basados en ensamblaje de clasificadores. El mejor entre estos, el equipo jhan014 (Han et al., 2019) utilizó en enfoque basado en reglas con un filtro de palabras clave basado en un diccionario de palabras clave de Twitter. Mediante este enfoque obtuvo un valor de 0,755 para F1.

En la tercera subtarea participaron 66 equipos. Al igual que en la primera subtarea el equipo que mejores resultados logró usó BERT, modelo con el que consiguieron un valor de 0,64 para F1. Este equipo fue vradivchev anikolov (Radivchev & Nikolov, 2019). Usaron El segundo mejor equipo utilizó varios modelos como OpenAI, Finetune y Transformers.

[15] https://sites.google.com/site/offensevalsharedtask/home
[16] https://semeval.github.io/

3.1.2 OffensEval 2020

En OffensEval 2020 (Zampieri et al., 2020) se realizaron las mismas subtareas que en la edición anterior. Sin embargo, se usó un conjunto de datos multilenguaje en inglés, árabe, danés, griego y turco.

Aunque participaron 528 equipos únicamente 145 de ellos presentaron resultados. 6 equipos presentaron resultados en los 5 idiomas, 19 de ellos lo hicieron en 4, 11 en 3 y los 96 restantes se centraron en un único idioma.

En la subtarea A el mejor equipo fue UHH-LT(Wiedemann et al., 2020) logrando un valor de F1 de 0,9204 mediante un conjunto de modelos ALBERT (Lan et al., 2019). Estos están basados en BERT pero limitan los problemas que tiene este en cuanto a memoria y tiempo de entrenamiento. El segundo utilizó RoBERTa-large (Y. Liu et al., 2019).

En la subtarea B el equipo con mejores resultados fue Galileo (Wang et al., 2020). Este equipo optó por ALBERT-XXLarge, logrando unas puntuación F1 de 0,7462. Los dos siguientes mejores equipos usaron BERT.

En la subtarea C se siguió el mismo patrón que en el caso de la anterior, obteniendo los mejores resultados con ALBERT-XXLarge y el segundo y tercero con BERT. Al igual que en la subtarea B el equipo que mejor puntuación logró fue Galileo, llegando a una puntuación F1 de 0,7145. Tant

3.2 IberLeF[17]

Campaña de evaluación compartida de sistemas de Procesamiento del Lenguaje Natural en español y otras lenguas ibéricas.

En las siguientes subsecciones se muestran algunas de las tareas de este foro más relevantes para el marco de este proyecto.

3.2.1 MEX-A3T 2018

El objetivo principal de MEX-A3T en 2018 (Alvarez-Carmona et al., 2018) es promover el procesamiento de lenguaje natural en redes sociales en la variedad mexicana del español, ya que las tareas realizadas hasta el momento se centran en inglés y, en algunos casos en el español peninsular.

La tarea se realiza sobre un corpus con 11 mil tweets etiquetados de forma manual como agresivos o no agresivos.

El equipo ganador fue INGEOTEC (Graff, Miranda-Jiménez, Tellez, Moctezuma, et al., 2018), que usó EvoMSA (Graff, Miranda-Jiménez, Tellez, & Moctezuma, 2018)y un lexicón de palabras agresivas. EvoMSA es un sistema de análisis de sentimientos que combina diferentes clasificadores de texto.

[17] http://sepln2023.sepln.org/iberlef/

3.2.2 MEX-A3T 2019

La tarea MEX-A3T en la edición de IberLeF 2019 (Aragón et al., 2019) persigue los mismos objetivos que el año anterior. Para ello se construyó un nuevo corpus de tweets de usuarios mexicanos.

El equipo ganador fue UACh (Casavantes et al., 2019). Debido a las limitaciones de tiempo y memoria, este equipo optó por un enfoque sencillo, usando un perceptrón multicapa entrenado únicamente con n-gramas de rango [3,4]. A pesar de ello, lograron una precisión de 0,73 y un valor de 0,65 para F1.

3.2.3 MEX-A3T 2020

El objetivo de la tercera edición de MEX-A3T (Aragón et al., 2020) es mejorar aún más la investigación de procesamiento de lenguaje natural en mexicano, para lo que se revisaron los criterios de identificación de agresividad y se generó un conjunto de datos mejorado.

El equipo con mejores resultados fue CIMAT (Guzman-Silverio et al., 2020). Este equipo aplicó un conjunto de modelos BERT y aumentó el conjunto de datos, de forma que el modelo crea una nueva entrada para cada oración mal clasificada. A pesar de haber obtenido buenos resultados (el valor de F1 fue 0,8596), el coste computacional de la generación de datos es muy elevado.

3.2.4 MeOffendEs 2021

El objetivo principal de MeOffendEs (Plaza-Del-Arco, Casavantes, et al., 2021) es promover la detección del lenguaje ofensivo en las variantes del español. Para ello se usa un conjunto de datos compuesto por comentarios de usuarios en Twitter, Youtube e Instagram.

Se compone de 4 tareas:

- **Subtarea 1:** Clasificación de los comentarios en una de las cuatro categorías (ofensivos hacia una persona, ofensivos hacia un grupo, no ofensivos con lenguaje malsonante y ofensivos.).
- **Subtarea 2:** Tiene el mismo objetivo que la subtarea 1 pero tiene en cuenta los metadatos.
- **Subtarea 3:** Clasificación de tweets en mexicano en ofensivos y no ofensivos.
- **Subtarea 4:** Tiene el mismo objetivo que la subtarea 3 pero tiene en cuenta los metadatos.

Los mejores resultados para las subtareas 1 y 3 fueron obtenidos por un equipo que uso XLM-RoBERTa pre-entrenado en textos de Twitter y datos de análisis de sentimiento.

El mejor resultado para la subtarea 1 fue obtenido por el equipo NLP-CIC (Taofeek Aroyehun & Gelbukh, 2021) logrando un valor para F1 de 0,8815 con el modelo XLM-RoBERTa.

En la subtarea 2 el equipo con mayor puntuación fue UMUTeam (Antonio García-Díaz, 2021), con una puntuación de F1 de 0,8782. Para ello usaron un modelo BETO combinado con el análisis de características lingüísticas.

3.3 OSACT4

El estudio OSACT4 (Husain, 2020) tiene como objetivo investigar el impacto de la fase de preprocesamiento sobre la clasificación de textos en lo que tiene que ver con lenguaje ofensivo y la clasificación del lenguaje de odio para textos en árabe.

Para ello se utiliza un conjunto de tweets en árabe y se realizan dos subtareas:

-Clasificar la publicación como ofensiva o no ofensiva.

-Determinar si las publicaciones ofensivas contienen o no discurso de odio.

En ambos casos el modelo que obtuvo mejores resultados fue un SVM el cual clasificó tweets que previamente habían pasado por un preprocesado muy completo.

3.4 Conclusiones

Al estudiar las tareas realizadas en detección de lenguaje ofensivo en los últimos años se observa el crecimiento de la popularidad de los modelos basados en Transformers, convirtiéndose en el modelo más usado en las principales tareas en pocos años desde su publicación.

En las primeras convocatorias de las tareas mostradas se llegaba a los mejores resultados mediante combinaciones de otros modelos, diccionarios, redes neuronales, máquinas de soporte vectorial y otras estrategias, mientras que en las convocatorias más recientes los modelos basados en Transformers ocupan el podio, especialmente aquellos basados en BERT y más concretamente en BETO al tratarse de tareas en español.

Por otro lado, se observa un gran esfuerzo por parte de la comunidad hispano hablante por promover los modelos y tareas de detección de lenguaje ofensivo en español y sus dialectos, ya que la mayor parte de estas está orientada a publicaciones en inglés. Sin embargo, todos ellos se entrenan y evalúan sobre colecciones similares, lo que motiva la realización de este trabajo.

Por ello en este trabajo se quiere continuar con esta corriente y estudiar la idoneidad de modelos basados en Transformers y pre-entrenados en español para la detección del lenguaje ofensivo en publicaciones de redes sociales redactadas en español y sus dialectos, aunque lo ideal sería poder generalizarlas también a las redactadas en inglés, evitando así la polaridad actual entre inglés-español.

Capítulo 4. Entorno de trabajo

4.1 Gestión documental

Para el almacenamiento de la documentación generada durante el proyecto se ha usado Microsoft OneDrive[18]. Siguiendo la apuesta de la UNED por Microsoft y, con el fin de aprovechar los recursos que aporta la licencia ofrecida se ha decidido usar esta herramienta.

OneDrive permite la creación de documentos de Office en línea y el posterior almacenamiento y compartición de estos con los usuarios deseados.

4.2 Entorno de desarrollo

El código generado para el proyecto se ha desarrollado en Google Colab[19], la herramienta de Google para escribir y ejecutar código Python de forma colaborativa en el navegador. Está basado en el proyecto de código abierto Jupyter.

Al igual que Jupyter Notebook[20], se basa en la creación de cuadernos, en los que se puede escribir tanto código de Python como texto enriquecido y que permiten trabajar con imágenes, gráficos, tablas, etc. Este código se ejecuta en máquinas virtuales de Google, cuya capacidad dependerá de la versión adquirida.

Para la realización de este proyecto se ha optado por la versión gratuita de Google Colab, ya que proporcionaba recursos suficientes para la ejecución del código desarrollado.

[18] https://www.microsoft.com/es-es/microsoft-365/onedrive/online-cloud-storage
[19] https://colab.research.google.com/?hl=es
[20] https://jupyter.org/

4.3 Librerías

Durante el desarrollo del código del proyecto se han usado las siguientes librerías de Python:

- **Pandas[21]:** Permite la manipulación y análisis de datos en Python. Permite realizar operaciones sobre los datos de forma sencilla, lo que la convierte en la biblioteca más popular en Ciencia de Datos.
- **Huggingface[22]**: Plataforma dedicada a todo lo relacionado con el procesamiento de lenguaje natural. En ella puede obtenerse acceso a modelos pre-entrenados, conjuntos de datos, ayuda, etc. Mediante el uso de su librería pueden usarse sus modelos.
- **Keras/Tensorflow[23]**: Tensorflow es una librería para Machine Learning que permite construir y entrenar redes neuronales. Keras es la API de TensorFlow para construir y entrenar modelo de aprendizaje profundo en Python.
- **Pysentimiento[24]:** Biblioteca de código abierto para procesamiento de lenguaje natural en Python publicada en (Pérez, Giudici, et al., 2021) que permite realizar tareas de minería de opinión (análisis de sentimientos, de emociones, detección de discursos de odio, detección de ironias) en Python en cuatro idiomas: español, inglés, italiano y portugués.
  Al estar entrenados con diferentes conjuntos de datos según el idioma, los modelos pueden presentar algunas diferencias. Según el idioma se distinguen los modelos mostrados en la Tabla 1.

Idioma	Tarea	Tipo	Clases
es, en, pt	Sentimiento	Multiclase	POS, NEG, NEU
es, en	Emoción	Multiclase	Ira, alegría, tristeza, miedo, sorpresa, disgusto, neutral
es, en	Discurso de odio	Multietiqueta	Odioso, dirigido, agresivo
es, en, it, pt	Ironía	Binario	Irónico, no irónico
it	Sentimiento	Multietiqueta	Positivo, negativo
it	Emoción	Multiclase	Alegría, ira, tristeza, miedo

Tabla 1. Tareas de Pysentimiento por idioma

Dentro de estos modelos los que más interés tienen dentro del marco de este proyecto son BERTin y RoBERTuito, de los cuales se puede encontrar información en las secciones 3.6.1.1 y 3.6.1.2 respectivamente.

[21] https://pandas.pydata.org/
[22] https://huggingface.co/
[23] https://www.tensorflow.org/
[24] https://github.com/pysentimiento/pysentimiento

4.4 Métricas de evaluación

Para medir la calidad y comparar entre las distintas pruebas realizadas se usarán las siguientes métricas que se detallan en esta sección. Para especificar las fórmulas de estas se usarán las nomenclaturas de la matriz de confusión mostrada en la Tabla 2. Matriz de confusion:

- TN (True negative): Verdaderos negativos.
- TP (True positive): Verdaderos positivos.
- False positive (FP): Falsos positivos.
- False negative (FN): Falsos negativos.

		Predicción	
		0	1
Realidad	0	TN	FP
	1	FN	TP

Tabla 2. Matriz de confusion

- **Precisión:** Mide el porcentaje de casos que ha acertado el modelo. Se calcula dividiendo el número de predicciones correctas entre el total de predicciones.

$$precisión = \frac{TP}{TP + FP}$$

- **Exhaustividad:** Mide el porcentaje de positivos que el modelo es capaz de detectar. Se calcula dividiendo los positivos clasificados por el modelo entre el número de positivos real.

$$exhaustividad = \frac{TP}{TP + FN}$$

- **F1:** Combina las medidas de precisión y exhaustividad en un solo valor. Se calcula multiplicando por 2 el resultado de dividir el producto de la precisión y la exhaustividad entre la suma de estos.

$$F1 = 2 \cdot \frac{precisión \cdot exhaustividad}{precisión + exhaustividad}$$

La elección de estás métricas se ha debido principalmente a que son las más usadas en otros trabajos similares y, por lo tanto, permiten realizar una comparación con estos de forma sencilla.

El cálculo de las métricas se realiza mediante el método micro. Es decir, se calculan las métricas de forma global mediante el recuento total de TP, FP y FN.

4.5 Conjuntos de datos

Para la realización del proyecto se han usado 3 conjuntos de datos, los cuales se detallan en las siguientes subsecciones.

La elección de estos conjuntos de datos en concreto viene motivada por varios motivos:

-Conjuntos procedentes de tareas recientes y punteras en el procesamiento de lenguaje natural.

-Cada uno de ellos es de un lenguaje/dialecto distinto.

-Diversidad entre ellos. Los conjuntos de datos son diferentes en cuanto a extensión, proporción texto plano/emojis/URL.

4.5.1 OffendES

Conjunto de datos en español obtenido de la tarea (Plaza-Del-Arco, Montejo-Ráez, et al., 2021). Contiene 30416 comentarios provenientes de Twitter, Youtube e Instagram, al ser estas las redes sociales más usadas entre la población joven. Fueron recopilados de las cuentas en estas plataformas de 12 influencers polémicos.

Los comentarios están etiquetados en las siguientes categorías:

- NO: No es ofensivo ni contiene lenguaje vulgar.
- NOM: No es ofensivo, pero contiene lenguaje vulgar, aunque con una connotación positiva.
- OFP: Ofensivo y dirigido a un individuo específico.
- OFG: Ofensivo y dirigido a un grupo de personas o colectivo.

El total de comentarios está repartido en los siguientes subconjuntos:

→ 100 comentarios de validación
→ 16710 comentarios de entrenamiento
→ 13606 comentarios de pruebas

En la Tabla 3 se muestra la distribución de las etiquetas en cada uno de los subconjuntos:

Etiqueta	Validación	Entrenamiento	Pruebas
NO	64	13212	9651
NOE	22	1235	2340
OFP	10	2051	1404
OFG	4	212	211

Tabla 3. Distribución etiquetas en OffendES

4.5.2 MEX-A3T

Publicado en (Guzman-Silverio et al., 2020b). Conjunto de tweets recopilados durante 3 meses teniendo en cuenta su geolocalización. Se tomó la Ciudad de México como centro y se usó un radio de 500 km. Tras la recopilación inicial se realizó un filtrado usando como criterios las palabras catalogadas como vulgares en el Diccionario de Mexicanismos de la Academia Mexicana de la Lengua, así como las palabras y hashtags identificados por el Instituto Nacional de las Mujeres en relación con la violencia y el acoso sexual.

Los tweets están etiquetados en las siguientes categorías:

- 0: No agresivo
- 1: Agresivo

Tras estas tareas iniciales se obtuvieron 10475 tweets repartidos en los siguientes subconjuntos:

→ 7332 tweets de entrenamiento
→ 3143 tweets de pruebas

En la Tabla 4 se muestra la distribución de las etiquetas en cada uno de los subconjuntos:

Etiqueta	Entrenamiento	Pruebas
No agresivo	5222	2238
Agresivo	2110	905

Tabla 4. Distribución de etiquetas en MEX-A3T

4.5.3 OLID

Publicado en (Zampieri et al., 2019a). Conjunto de tweets en inglés recuperados de la red social buscando por palabras y construcciones que son usadas frecuentemente en comentarios ofensivos y añadiendo aquellos dirigidos a las cuentas de noticias New Yorker[25] y BreitBartNews[26] con ideales políticos extremos, ya que tienden a tener comentarios ofensivos sobre política. Durante el proceso de construcción del conjunto de datos fueron afinándose los criterios para conseguir mantener la proporción del 30% ofensivos.

Usa un sistema de anotación basado en 3 niveles:

- Subtarea A: Distingue si el comentario es ofensivo o no. Se distinguen los no ofensivos (NOT) y los ofensivos (OFF).
- Subtarea B. Tipo de ofensivo. Distingue entre insulto dirigido a personas o grupos concretos (TIN), insulto no dirigido a ningún colectivo o persona en concreto (UNT).
- Subtarea C. Objetivo de la ofensa. Distingue entre individual (IND), grupo (GRP) y otros (OTH).

En nuestro caso, al querer identificar únicamente si se trata de comentarios ofensivos, independientemente de a quien se dirijan nos interesa la subtarea A.

En la Tabla 5 se muestra la distribución de las etiquetas en cada uno de los subconjuntos:

Etiqueta	Entrenamiento	Pruebas
No ofensivo	8840	620
Ofensivo	4400	240

Tabla 5. Distribución de etiquetas en OLID

[25] https://www.newyorker.com/
[26] https://www.breitbart.com/

4.6 Modelos

Desde un primer momento se decidió explorar los modelos basados en Transformers. Este tipo de modelos son considerados el estado del arte y constituyen la base de la mayor parte de los modelos que se implementan en el ámbito del procesamiento del lenguaje natural. Por ello la elección del modelo para la realización de los experimentos se centró en estos.

Dentro de este tipo de modelos los más interesantes para la tarea son los basados en BERT, ya que se centra en el procesamiento de lenguaje natural. Entre los modelos basados en BERT se escogieron los que mejores resultados devolvían para la detección de odio que en orden de mayor a menor puntuación son RoBERTuito, RoBERTa y BERTin, como puede observarse en la Tabla 6. Comparativa modelos basados en BERT. En ella pueden observarse los resultados expresados como puntuación media de Macro F1 en 10 ejecuciones de aplicar los modelos en los siguientes conjuntos de datos:

-Detección de odio: HatEval (Basile et al., 2019).

-Análisis de sentimientos: TASS 2020 Task A (García-Vega et al., 2020).

-Análisis de emociones: TASS 2020 Task B (García-Vega et al., 2020).

-Detección de ironía: IrosVA 2019 (Aragón et al., 2019).

Modelo	Detección de odio	Análisis de sentimientos	Análisis de emociones	Detección de Ironía	Puntuación
RoBERTuito	80,1	70,7	55,1	73,6	69,9
RoBERTa	76,6	66,9	53,3	72,3	67,3
BERTin	76,7	66,5	51,8	71,6	66,7
BETO	76,8	66,5	52,1	70,6	66,5

Tabla 6. Comparativa modelos basados en BERT

Estos son los modelos que se explican en las siguientes secciones.

4.6.1 BERTin

Modelo basado en BERT publicado en (De la Rosa et al., 2022) que busca simplificar el proceso tradicional de entrenamiento de modelos. Para ello desarrollan una técnica llamada muestreo de perplejidad que permite el pre-entrenamiento de los modelos con la mitad de los pasos y una quinta parte de los datos normalmente necesarios.

Para la realización del modelo partieron del conjunto de datos de Common Crawl en español, el cual contiene unos 416 millones de documentos y 235 billones de palabras y lo dividieron en subconjuntos más pequeños, de una quinta parte de su tamaño aproximadamente, usando la técnica de muestreo de perplejidad, que permite obtener subconjuntos representativos que permiten entrenar al modelo.

4.6.2 RoBERTuito

Modelo basado en BERT publicado en (Pérez, Furman, et al., 2021). Surge de la necesidad de disponer modelos especializados en español. En inglés existen modelos para tareas particulares (artículos científicos, documentos médicos, etc) los cuales mejoran significativamente el rendimiento. Sin embargo, en otros idiomas no existen tantos modelos de este tipo.

Se centra en textos generados por los usuarios. Por ello el modelo fue entrenado con un conjunto de alrededor de 500 millones de tweets escritos en español.

Como puede observarse en la Tabla 6. Comparativa modelos basados en BERT el modelo es muy preciso detectando odio en mensajes escritos por usuario, lo que va en línea con la finalidad de este proyecto. Por ello se decidió el uso de este modelo para la realización del proyecto.

Otra de las motivaciones de RoBERTuito es que el modelo fuera fácilmente accesible por cualquier persona, por ello lo pusieron a disposición en HuggingFace[27].

4.6.3 RoBERTa

Modelo basado en BERT publicado en (Y. Liu et al., 2019). Surge tras detectar carencias en el entrenamiento de BERT, que impiden obtener el máximo rendimiento posible de este. RoBERTa se presenta como una versión mejorada de este, basada en mejorar algunas condiciones del entrenamiento. Algunas de estas son: disponer de un conjunto de entrenamiento más grande, aumentar el tiempo de entrenamiento o eliminar la tarea de predicción de la siguiente oración.

[27] https://huggingface.co/pysentimiento/robertuito-hate-speech

4.6.4 Elección del modelo

Dentro de los modelos basados en Transformers, era fundamental escoger uno que devolviera buenos resultados con textos en español y sus dialectos, aunque lo ideal era un modelo multilenguaje, para obtener una solución universal. Por ello el proyecto se ha desarrollado usando un conjunto de datos en español, uno con su dialecto mexicano y otro en inglés.

Durante la investigación de las opciones disponibles fue crucial el descubrimiento de Pysentimiento, al ser una biblioteca muy extendida cuya finalidad estaba alineada con la de este proyecto. Dentro de esta se realizaron pruebas combinando las salidas de BERTin y RoBERTa, que no devolvieron buenos resultados, como puede observarse en la Tabla 7 y la Tabla 8. Sin embargo, las pruebas con RoBERTuito devolvieron los resultados esperados y se decidió explorar esta vía.

Conjunto de datos pruebas	**Precisión**	**Exhaustividad**	**F1**
OffendES	0,5226	0,6138	0,5646
MEX-A3T	0,4693	0,7005	0,5621
OLID	0,1672	0,0955	0,1216

Tabla 7. Resultados BERTin+RoBERTa

Conjunto de datos pruebas	**Precisión**	**Exhaustividad**	**F1**
OffendES	0,7884	0,6138	0,6902
MEX-A3T	0,2672	0,7005	0,3868
OLID	0,1456	0,353	0,2062

Tabla 8. Resultados RoBERTa+ BERTin

Capítulo 5. Experimentos

En este apartado se explicarán los experimentos realizados, desde la preparación y limpieza de datos para mejorar el rendimiento y resultados del modelo, como la elección del modelo a estudiar y los parámetros con los que realizar las pruebas.

5.1 Preprocesado

5.1.1 Adaptación de etiquetas

Dado que el modelo devuelve los datos en formato numérico (0 si no se detecta odio, 1 si se detecta odio) ha sido necesario realizar una transformación en las etiquetas de cada uno de los conjuntos para adaptarlas a esta salida y poder comparar mejor los resultados.

En el caso de los datos de OffendES se han realizado las siguientes equivalencias:

- OFP (Ofensivo dirigido a una persona) → 1
- NOM (No ofensivo con lenguaje malsonante) → 0
- OFG (Ofensivo dirigido a un grupo) → 1
- NO (No ofensivo) → 0

En los datos procedentes de MEX-A3T no ha sido necesario realizar ninguna modificación ya que sus etiquetas ya eran 0-1.

En los datos procedentes de OLID se han realizado las siguientes equivalencias:

- NOT (No ofensivo) → 0
- OFF (Ofensivo) → 1

5.1.2 Preprocesado Pysentimiento

Se ha utilizado el preprocesador de Pysentimiento, que realiza una normalización de los datos, facilitando la tarea al modelo y mejorando los resultados de este.

Algunas de las características que presenta este módulo son:

→ **Reemplazo de identificadores de usuario y URL por tokens especiales:** Se sustituyen las menciones a usuarios por @USUARIO y los enlaces por el literal url. De esta manera se evita confundir al modelo y que no procese nombres de usuario ni palabras dentro de los enlaces.

→ **Acortar caracteres repetidos:** Reduce el número de caracteres repetidos seguidos. De esta manera se evita que el modelo no reconozca palabras por tener vocales repetidas, lo que es muy frecuente en los textos de redes sociales.

→ **Normalizar las risas:** Detecta las expresiones de risa y las sustituye por el literal jaja que es el que los modelos asocian con expresiones de risa.

→ **Manejar hashtags:** Integra los hashtags dentro de la propia sentencia. Algunos de los hashtags aportan significado a las oraciones, por lo que este preprocesamiento permite que los modelos puedan analizarlos.
→ **Manejar emojis:** Sustituye los emojis por el literal emoji seguido de su descripción, lo que permite que el modelo interprete la intención del usuario al usar el emoji.

5.2 Planificación de experimentos

Se dispone del modelo robertuito-hate-speech[28] pre-entrenado con tweets en español el cual se entrenará con cada uno de los conjuntos de entrenamiento de los conjuntos de datos explicados en la sección 4.5 y se probará en los conjuntos de pruebas de estos.

Con el fin de obtener los mejores resultados posibles se realizará un entrenamiento del modelo completo con ajuste fino de los parámetros.

5.3 Hiperparámetros

Se usó el mismo valor para los hiperparámetros en todos los experimentos realizados, con el fin de compararlos en las mismas condiciones. Fueron los siguientes:

- learning_rate = 5E-5
- optimizer: Adam[29]
- epsilon=1E-07
- tam_batch: 12
- num_epochs: 1
- loss=SparseCategoricalCrossentropy (from_logits=True)

[28] https://huggingface.co/pysentimiento/robertuito-hate-speech
[29] https://keras.io/api/optimizers/adam/

Capítulo 6. Resultados

En esta sección se expondrán los resultados obtenidos con la realización de cada experimento con las métricas escogidas, así como una reflexión general sobre estos.

6.1 Modelo pre-entrenado

Con el fin de poder evaluar los resultados de los experimentos y verificar si el re-entrenamiento con los conjuntos de datos propuestos consigue mejorar la generalización del modelo se ha realizado la evaluación del modelo pre-entrenado sin ningún ajuste en estos, cuyos resultados pueden consultarse en la Tabla 9.

Conjunto de datos pruebas	Precisión	Exhaustividad	F1
OffendES	0,7922	0,561	0,6568
MEX-A3T	0,6958	0,347	0,4604
OLID	0,5977	0,6	0,5989

Tabla 9. Resultados modelo pre-entrenado

Los resultados devueltos por el modelo demuestran que el modelo está pre-entrenado en español, al devolver métricas superiores para el conjunto de datos en este idioma. La generalización a los dialectos e idiomas en los que no está pre-entrenados no es muy buena, especialmente en el caso del mejicano en el que el modelo no es capaz de detectar la mayor parte de los positivos.

6.2 Experimento 1: Modelo entrenado con datos de OffendES

El entrenamiento con los datos de OffendES y pruebas con los diferentes conjuntos devolvió los resultados mostrados en la Tabla 10.

Conjunto de datos pruebas	Precisión	Exhaustividad	F1
OffendES	0,9265	0,9071	0,9167
MEX-A3T	0,7993	0,9448	0,866
OLID	0,7445	0,0932	0,1657

Tabla 10. Resultados entrenando con OffendES

A la vista de los resultados se demuestra que al haberse usado un modelo pre-entrenado con tweets en español y re-entrenarlo con un conjunto de datos en español no es aplicable a otros idiomas, dado los resultados obtenidos con el conjunto de datos OLID.

En este proyecto, dirigido a estudiar la generalización de un modelo para la tarea de detección de lenguaje ofensivo para su uso en moderación de contenido en redes sociales resulta de gran relevancia el resultado de la exhaustividad. Es importante que el modelo sea preciso detectando los positivos para evitar que se lleguen a publicar mensajes ofensivos. Como puede observarse en la tabla este resultado es muy bajo para las pruebas con el conjunto de datos OLID, detectando únicamente un 10% de estos.

El re-entrenamiento del modelo con el conjunto de datos OffendES ha mejorado los resultados para los conjuntos de datos en español. Con este experimento se consiguen valores de precisión y exhaustividad muy altos en español y mejicano. El gran tamaño del conjunto de datos de pruebas de OffendES es un elemento clave para justificar estos resultados.

Sin embargo, no demuestra ser útil para generalizar el modelo al inglés ya que llega incluso a empeorar el valor de la exhaustividad para este idioma. El re-entrenamiento con estos datos aumenta el ajuste al español, reduciendo la generalización del modelo inicial.

6.3 Experimento 2: Modelo entrenado con datos de MEX-A3T

El entrenamiento con los datos de MEX-A3T y pruebas con los diferentes conjuntos devolvió los resultados mostrados en la Tabla 11 .

Conjunto de datos pruebas	Precisión	Exhaustividad	F1
MEX-A3T	0,8622	0,7746	0,7331
OffendES	0,8967	0,6358	0,744
OLID	0,7532	0,1229	0,2112

Tabla 11. Resultados entrenando con MEX-A3T

Las conclusiones son similares a las del experimento anterior. No se obtienen buenos resultados entrenando el modelo en español y probándolo en OLID. Los resultados en este conjunto de datos son ligeramente mejores que en el experimento anterior, pero siguen sin ser los esperados para ser aplicados a una red social internacional con usuarios que hablen diferentes idiomas.

En general, se observan peores resultados que en el experimento 1. En el único caso que mejoran es en el de MEX-A3T debido a que el entrenamiento se ha realizado con este conjunto de datos. Por lo que podría afirmarse que el modelo generaliza mejor cuando se entrena con un conjunto de datos en español que con el dialecto mexicano.

6.4 Experimento 3: Modelo entrenado con datos de con OLID

El entrenamiento con los datos de OLID y pruebas con los diferentes conjuntos devolvió los resultados mostrados en la Tabla 12. Resultados entrenando con OLID.

Conjunto de datos pruebas	Precisión	Exhaustividad	F1
OLID	0,8229	0,8229	0,7982
OffendES	0,7755	0,937	0,8486
MEX-3AT	0,629	0,8343	0,7172

Tabla 12. Resultados entrenando con OLID

En este caso, en comparación con los anteriores se han obtenido mejores resultados.

Al ser un modelo pre-entrenado con tweets en español y entrenarlo con un conjunto de datos en inglés se consigue que generalice mejor a ambos idiomas, aunque los resultados son notablemente mejores en el conjunto de datos en español que en el del dialecto mexicano.

A pesar de ser mejores resultados, en las 3 pruebas se obtienen muchos falsos positivos. En la práctica, los falsos positivos son más favorables que los falsos negativos. Esto se podría traducir en que el modelo moderara automáticamente comentarios no ofensivos como ofensivos, se les notificara a los usuarios afectados, estos reclamaran y en ese momento una persona revisaría la clasificación y retiraría la etiqueta de no ofensivo. De esta forma no llegarían a publicarse mensajes ofensivos, pero habría que valorar la carga de trabajo manual que esto supondría para los moderadores humanos.

Tras obtener resultados similares con los entrenamientos realizados con cada conjunto de datos se ha considerado oportuno realizar entrenamientos con combinaciones de los conjuntos de datos anteriores. Los resultados de estas pruebas se detallan en los apartados siguientes.

6.5 Experimento 4: Modelo entrenado con datos de OLID+MEX-A3T

El entrenamiento con los datos de OLID y MEX-A3T de forma de conjunta y pruebas con los diferentes conjuntos devolvió los resultados mostrados en la Tabla 13.

Conjunto de datos pruebas	Precisión	Exhaustividad	F1
OLID+MEX-A3T	0,8316	0,752	0,79
OLID	0,7302	0,0412	0,079
OffendES	0,9104	0,639	0,7504
MEX-3AT	0,8549	0,8961	0,875

Tabla 13. Resultados entrenando con OLID+MEX-A3T

Los resultados obtenidos con la combinación de estos dos conjuntos de datos siguen sin ser lo satisfactorios que se buscaba. A pesar de que mejoran las métricas del modelo pre-entrenado para el español y el dialecto mexicano sigue sin conseguirse su aplicación al inglés. Al ser un modelo pre-entrenado en español y volverlo a entrenar en español se consigue un sobre ajuste a este idioma que le impide generalizar a otros.

Además, estos resultados son muy similares a los obtenidos en los entrenamientos con los conjuntos de datos de forma individual, a pesar de ser un conjunto de entrenamiento superior y, por lo tanto, requerir más recursos para el entrenamiento. Se puede determinar que no compensa este esfuerzo con los datos obtenidos.

Se sigue si encontrar ninguna combinación que consiga generalizar para OLID si no es entrenando con su propio conjunto de datos de entrenamiento. En el caso de OffendES se suelen conseguir buenos resultados al ser el mismo idioma que el del conjunto de datos de pre-entrenamiento por lo que este entrenamiento no aporta ningún beneficio sobre esto. Ocurre algo similar en el caso de MEX-3AT, los resultados obtenidos con esta combinación ya se consiguieron entrenando con OffendES y con OLID, por lo que este entrenamiento combinado no resulta efectivo.

6.6 Experimento 5: Modelo entrenado con datos de OLID+OFFENDES

El entrenamiento con los datos de OLID y OFFENDES de forma de conjunta y pruebas con los diferentes conjuntos devolvió los resultados mostrados en la Tabla 14.

Conjunto de datos pruebas	Precisión	Exhaustividad	F1
OLID+OFFENDES	0,9067	0,7044	0,7929
OLID	0,7221	0,0042	0,0083
OffendES	0,9189	0,777	0,842
MEX-3AT	0,7967	0,968	0,874

Tabla 14. Resultados entrenando con OLID+OFFENDES

Los resultados con esta segunda combinación de resultados son ligeramente mejores que los del apartado anterior. Este entrenamiento consigue que el modelo tenga una exhaustividad superior, detectando comentarios ofensivos en español y su dialecto mexicano, aunque se sigue sin lograr la generalización al idioma inglés.

En este punto, es importante tener en cuenta que el tamaño del conjunto de datos con el que se pre-entrenó el modelo y el de entrenamiento de OffendES son muy superiores al de OLID, por lo que esta proporción dificulta que el modelo generalice a ambos idiomas. Por otra parte, el conjunto de datos OLID es el que más elementos tiene (hashtags, menciones a usuarios, emoticonos) mientras que los demás tienen una mayor proporción de texto plano. A pesar de estar usando el módulo de pre-procesado de Pysentimiento, puede que estos elementos no estén siendo interpretados correctamente y estén implicando una incorrecta clasificación de las publicaciones.

6.7 Experimento 6: Modelo entrenado con datos de OFFENDES+MEX-A3T

El entrenamiento con los datos de OLID y MEX-A3T de forma de conjunta y pruebas con los diferentes conjuntos devolvió los resultados mostrados en la Tabla 15.

Conjunto de datos pruebas	Precisión	Exhaustividad	F1
OFFENDES+MEX-A3T	0,9119	0,4873	0,6352
OLID	0,7314	0,0542	0,1009
OffendES	0,9189	0,777	0,842
MEX-3AT	0,8549	0,7856	0,8188

Tabla 15. Resultados entrenando con OFFENDES+MEX-A3T

En esta prueba combinando conjuntos de datos para el entrenamiento tampoco se han conseguido mejorar los resultados individuales.

Se sigue la tendencia anterior de las mejoras en los conjuntos de datos en español, pero con este entrenamiento se sigue sin lograr la generalización al inglés. Sin embargo, cabe destacar que introducir datos en el dialecto mexicano mejora en gran medida los resultados en MEX-3AT. A la vista de los experimentos podemos deducir que el modelo pre-entrenado estaba muy ajustado al español y no

generalizaba a sus dialectos, problema que se ve corregido al usar datos de entrenamiento en este dialecto, a pesar de que su extensión es inferior a los conjuntos de datos en español.

6.8 Experimento 7: Modelo entrenado con datos de OFFENDES+MEX-A3T+OLID

En entrenamiento combinando los datos de entrenamiento de todos los conjuntos de datos y pruebas con el resto de los conjuntos de datos devolvió los resultados mostrados en la Tabla 16.

Conjunto de datos pruebas	Precisión	Exhaustividad	F1
OFFENDES+MEX-A3T+OLID	0,8993	0,793	0,8428
OLID	0,8453	0,7839	0,8135
OffendES	0,9128	0,7603	0,8296
MEX-3AT	0,8549	0,8829	0,8687

Tabla 16. Resultados entrenando con OFFENDES+MEX-A3T

Esta última prueba, combinando todos los conjuntos de datos de entrenamiento ha devuelto los mejores resultados. Las métricas tienen valores similares en todos los conjuntos de datos, por lo que se ha conseguido una buena generalización entrenando con todos los datos. Por otro lado, se ha conseguido que la precisión del modelo aumente, aunque podría mejorarse su exhaustividad.

El entrenamiento realizado en este último experimento resulta ser el más idóneo para conseguir un modelo que generalice bien al español, al mejicano y al inglés. Las métricas devuelven valores superiores en todos los casos al del modelo pre-entrenado.

Sin embargo, para la realización de este experimento se han necesitado una gran cantidad de recursos

Capítulo 7. Conclusiones y trabajos futuros

Durante el proyecto se han realizado 6 experimentos para estudiar la capacidad de generalización del modelo RoBERTuito pre-entrenado a diferentes conjuntos de datos y si esta puede ser mejorada mediante el re-entrenamiento de este modelo con diferentes conjuntos de datos y sus combinaciones. Algunas de los experimentos han demostrado la idoneidad de estos re-entrenamientos para la tarea de detectar lenguaje ofensivo en redes sociales, incluso las internacionales, en las que usuarios de diferentes procedencias realizan publicaciones en sus propios idiomas y dialectos.

A pesar de que el modelo ya estaba pre-entrenado con publicaciones de redes sociales en español para una tarea similar ha resultado crucial el entrenamiento posterior, la elección correcta de los datos de entrenamiento y el ajuste fino de los hiperparámetros, como puede observarse en la diversidad entre los resultados devueltos en los distintos experimentos.

Para mejorar los resultados sería interesante disponer de más datos y que el reparto en idiomas y dialectos sea más proporcionado, ya que el 50% de los datos de entrenamiento y el 77% de los de pruebas corresponden a OffendES, cuyos resultados podrían deberse a un sobreajuste. Por otro lado, el conjunto de pruebas de OLID está compuesto únicamente por 860 tweets, lo que podría no ser suficiente para tener en consideración las métricas sobre este.

Además de su mayor proporción y tamaño, convendría disponer de un conjunto de datos de entrenamiento y pruebas que dispusiera de una mayor variedad en cuanto a menciones, hashtags, enlaces y expresiones en todos los idiomas.

Por otro lado, a pesar de haber conseguido buenos resultados el entrenamiento con todos los datos de pruebas ha sido costoso en tiempo y recursos.

Los futuros experimentos podrían ir destinados a aplicar los experimentos a más idiomas y dialectos de estos. Un primer paso podría ser añadir al entrenamiento un conjunto de datos en inglés británico, para asegurar la generalización a los dos dialectos más extendidos del inglés. Posteriormente podría continuarse con las pruebas a los dialectos sudamericanos del español y, una vez finalizados estos a otros idiomas.

Capítulo 8. Bibliografía

Alammar, J. (2018). *The Illustrated Transformer*. https://jalammar.github.io/illustrated-transformer/

Alvarez-Carmona, M. A. ´, Guzmán-Falcón, E., Montes-Y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). *Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets*. https://pan.webis.de/clef18/pan18-web/index.html

Antonio García-Díaz, J. (2021). *UMUTeam at MeOffendEs 2021: Ensemble Learning for Offensive Language Identification using Linguistic Features, Fine-grained Negation, and Transformers*. https://github.com/pendrag/MeOffendEs.

Aragón, M. E., Álvarez-Carmona, M. Á., Montes-Y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., & Moctezuma, D. (2019). *Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets*.

Aragón, M. E., Jarquín-Vásquez, H., Montes-Y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Gómez-Adorno, H., Posadas-Durán, J.-P., & Bel-Enguix, G. (2020). *Overview of MEX-A3T at IberLEF 2020: Fake News and Aggressiveness Analysis in Mexican Spanish*. http://ceur-ws.org

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., & Sanguinetti, M. (2019). *SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter*. http://evalita.org

Bengio, Y., Courville, A., & Vincent, P. (2012). *Representation Learning: A Review and New Perspectives*. http://arxiv.org/abs/1206.5538

Casavantes, M., López, R., & Carlos González, L. (2019). *UACh at MEX-A3T 2019: Preliminary Results on Detecting Aggressive Tweets by Adding Author Information Via an Unsupervised Strategy*.

De la Rosa, J., Ponferrada, E. G., Villegas, P., De Prado Salas, P. G., Romero, M., & Grandury, M. (2022). BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento Del Lenguaje Natural*, *68*, 13–23. https://doi.org/10.26342/2022-68-1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). *Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm*. https://doi.org/10.18653/v1/D17-1169

García-Vega, M., Carlos Díaz-Galiano, M., García-Cumbreras, M. Á., Plaza Del Arco, F. M., Montejo-Ráez, A., María Jiménez-Zafra, S., Cámara, E. M., Aguilar, A., Antonio, M., Cabezudo, S., Chiruzzo, L., & Moctezuma, D. (2020). *Overview of TASS 2020: Introducing Emotion Detection*. https://www.ujaen.es/

Graff, M., Miranda-Jiménez, S., Tellez, E. S., & Moctezuma, D. (2018). *EvoMSA: A Multilingual Evolutionary Approach for Sentiment Analysis*. https://doi.org/10.1109/MCI.2019.2954668

Graff, M., Miranda-Jiménez, S., Tellez, E. S., Moctezuma, D., Salgado, V., Ortiz-Bejar, J., & Sánchez, C. N. (2018). *INGEOTEC at MEX-A3T: Author profiling and aggressiveness analysis in Twitter using μTC and EvoMSA*. https://mexa3t.wixsite.com/home

Guzman-Silverio, M., Balderas-Paredes, Á., & Pastor López-Monroy, A. (2020a). *Transformers and Data Augmentation for Aggressiveness Detection in Mexican Spanish*. https://www.cimat.mx/es/adri

Guzman-Silverio, M., Balderas-Paredes, Á., & Pastor López-Monroy, A. (2020b). *Transformers and Data Augmentation for Aggressiveness Detection in Mexican Spanish*. https://www.cimat.mx/es/adri

Han, J., Wu, S., & Liu, X. (2019). *jhan014 at SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media*.

Howard, J., & Ruder, S. (2018). *Universal Language Model Fine-tuning for Text Classification*. http://arxiv.org/abs/1801.06146

Husain, F. (2020). *OSACT4 Shared Task on Offensive Language Detection: Intensive Preprocessing-Based Approach*.

Judith Sandoval, L. (2018). *ENERO-DICIEMBRE 2018 Derechos Reservados • Escuela Especializada en Ingeniería ITCA-FEPADE* (Vol. 11).

Kaufman, G. A. (2016). *Odium Dicta - Libertad De Expresión Y Protección De Grupos Discriminados En Internet*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2686171

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. http://arxiv.org/abs/1909.11942

Liu, P., Li, W., & Zou, L. (2019). *NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers*. http://www.hatebase.org

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. http://arxiv.org/abs/1907.11692

Min, S., Seo, M., & Hajishirzi, H. (2017). Question answering through transfer learning from large fine-grained supervision data. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *2*, 510–517. https://doi.org/10.18653/v1/P17-2081

Nakov Lluís, P. M., Alessandro Moschitti Walid Magdy Hamdy Mubarak Abed Alhakim Freihat, arquez, Glass, J., & Randeree Qatar Living, B. (2016). *SemEval-2016 Task 3: Community Question Answering*. http://alt.qcri.org/semeval2015/task3

OpenAI. (2023). *GPT-4 Technical Report*. http://arxiv.org/abs/2303.08774

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. In *IEEE Transactions on Knowledge and Data Engineering* (Vol. 22, Issue 10, pp. 1345–1359). https://doi.org/10.1109/TKDE.2009.191

Pérez, J. M., Furman, D. A., Alemany, L. A., & Luque, F. (2021). *RoBERTuito: a pre-trained language model for social media text in Spanish*. http://arxiv.org/abs/2111.09453

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. http://arxiv.org/abs/2106.09462

Plaza-Del-Arco, F. M., Casavantes, M., Escalante, H. J., Martín-Valdivia, M. T., Montejo-Ráez, A., Montes-Y-Gómez, M., Jarquín-Vásquez, H., & Villaseñor-Pineda, L. (2021). Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants. *Procesamiento Del Lenguaje Natural*, *67*, 183–194. https://doi.org/10.26342/2021-67-16

Plaza-Del-Arco, F. M., Montejo-Ráez, A., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). OffendES: A New Corpus in Spanish for Offensive Language Research. *International Conference Recent Advances in Natural Language Processing, RANLP*, 1096–1108. https://doi.org/10.26615/978-954-452-072-4_123

Radivchev, V., & Nikolov, A. (2019). *Nikolov-Radivchev at SemEval-2019 Task 6: Offensive Tweet Classification with BERT and Ensembles*. https://sites.google.com/view/trac1/

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. http://arxiv.org/abs/1910.10683

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. https://stanford-qa.com,

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). *Bidirectional Attention Flow for Machine Comprehension*. http://arxiv.org/abs/1611.01603

Taofeek Aroyehun, S., & Gelbukh, A. (2021). *Evaluation of Intermediate Pre-training for the Detection of Offensive Language*. https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Wang, S., Liu, J., Ouyang, X., & Sun, Y. (2020). *Galileo at SemEval-2020 Task 12: Multi-lingual Learning for Offensive Language Identification using Pre-trained Language Models*. Online. http://research.baidu.com/Blog/index-view?id=128

Wiedemann, G., Yimam, S. M., & Biemann, C. (2020). *UHH-LT at SemEval-2020 Task 12: Fine-Tuning of Pre-Trained Transformer Networks for Offensive Language Detection*. Online. https://commoncrawl.org

Yang, Y., Yih, W.-T., & Meek, C. (2015). *WIKIQA: A Challenge Dataset for Open-Domain Question Answering*. Association for Computational Linguistics. http://aka.ms/WikiQA.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019a). *Predicting the Type and Target of Offensive Posts in Social Media*. http://arxiv.org/abs/1902.09666

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). *SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)*. http://competitions.codalab.org/

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). *SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)*. http://arxiv.org/abs/2006.07235