Análisis comparativo del desempeño de diferentes métodos de clasificación para la detección de lenguaje ofensivo en idioma español aplicado a la red social X

Lissete Rosete
División de Investigación y Estudios de Posgrado
Instituto Tecnológico de Orizaba
Orizaba, Ver. México
m16011208@orizaba.tecnm.mx

Asdrúbal López
Centro Universitario UAEM Zumpango
Universidad Autónoma del Estado de México
Zumpango de Ocampo, Méx. México
alchau@uaemex.mx

Luis Ángel Reyes
División de Investigación y Estudios de Posgrado
Instituto Tecnológico de Orizaba
Orizaba, Ver. México
luis.rh@orizaba.tecnm.mx

***Resumen*— El ciberacoso, también conocido como acoso cibernético o ciberbullying, es un fenómeno creciente en la era digital que implica el uso de la tecnología, como Internet y las redes sociales, para hostigar, difamar, amenazar o molestar a otras personas. Debido a esto es esencial abordar el ciberacoso de manera responsable y tomar medidas para prevenirlo y proteger a quienes pue-dan ser víctimas de este fenómeno. Por ello, en el presente artículo se llevó a cabo el análisis del desempeño de diferentes métodos de clasificación para detectar la posible presencia de ciberacoso o lenguaje ofensivo en idioma español; concluyendo que el método de clasificación Multi-Layer Perceptron (MLP) con la vectorización de documentos basada en TF-IDF y un balanceo de clases obtienen un buen desempeño referente al valor F1-score.**

Palabras clave*— *Aprendizaje Automático, Ciberacoso, Detección de ciberacoso, Procesamiento de lenguaje natural.

I. INTRODUCCIÓN

Las redes sociales son una parte integral de la sociedad actual y tienen un impacto diverso en nuestras vidas. Pueden ser herramientas poderosas para la comunicación, la información y la organización, pero también plantean desafíos en términos de privacidad, salud mental y fenómenos como el ciberacoso, que deben ser abordados de manera responsable. Aunque el ciberacoso ha tomado diversas definiciones con el paso del tiempo ya que existen debates con respecto a la definición exacta [1], se ha realizado una conceptualización a partir de la definición del bullying tradicional; definiendo al ciberacoso como "un daño intencional y repetido infligido mediante el uso de computadoras, teléfonos celulares y otros dispositivos electrónicos" [2].

De acuerdo con el módulo sobre ciberacoso 2021 del Instituto Nacional de Estadística y Geografía (INEGI) [3] el 21.7% de la población usuaria de Internet fue víctima de ciberacoso. Además, se estima que el ciberacoso más frecuente se encuentra relacionado con el aspecto físico, a la forma de vestir y al estilo de vida. Por consiguiente, debido al incremento y perseverancia del ciberacoso es necesario crear herramientas que prevengan o detecten este tipo de situaciones, con la finalidad de disminuir la cantidad de personas afectadas.

Se han realizado trabajos para detectar mensajes ofensivos, principalmente enfocados en otros idiomas diferentes al español, tal como en [4] donde detectan comentarios de ciberacoso en lenguaje árabe, basándose en un corpus de palabras clave de acoso y agresión, su esquema propuesto clasifica comentarios en diferentes clases, según su intensidad (leve, medio y fuerte) utilizando una función ponderada; dicho esquema fue evaluado utilizando un conjunto de datos reales obtenido de las plataformas Youtube y Twitter, identificando la mayoría de los comentarios con posibles casos de ciberacoso.

Por otro lado, en el trabajo propuesto por Chen et al. [5] se presentó una arquitectura de característica sintáctica léxica para detectar contenido ofensivo e identificar posibles usuarios agresores a partir de reglas sintácticas escritas a mano; sus resultados mostraron una precisión del 98.24% y recall de 94.34% en la detección de sentencias ofensivas en lenguaje inglés.

Para la detección de lenguaje ofensivo presente en textos es posible utilizar métodos de procesamiento del lenguaje natural o algoritmos con diferentes técnicas de clasificación, en [6] se llevó a cabo una investigación en la que se compararon las mejores técnicas de aprendizaje automático para detectar lenguaje ofensivo en tweets, posteriormente fueron seleccionados los algoritmos Linear SVM y Naive Bayes para las pruebas realizadas, obteniendo buenos resultados con este último algoritmo mencionado utilizado con un conjunto de datos público que contiene tweets clasificados en comentarios de odio, lenguaje ofensivo y lenguaje normal.

Consecuente al menor enfoque del lenguaje español para la detección de lenguaje ofensivo presente en textos de redes sociales, se reduce la cantidad de conjuntos de datos referentes al ciberacoso, ya que la mayoría de estos conjuntos de datos se encuentran dirigidos al idioma inglés, al igual que los métodos o algoritmos de clasificación. Por consiguiente, el presente trabajo describe un análisis comparativo de tres métodos de aprendizaje supervisado para la detección de la posible presencia de ciberacoso; el cual contemplo un corpus de 2500 documentos descargados de la plataforma X (anteriormente llamada Twitter). Los algoritmos de clasificación seleccionados para la comparación, fueron Maquina de Vector Soporte (SVM del inglés Support Vector Machine), Bosque aleatorio (RF del inglés Random Forest) y Perceptrón Multicapa (MLP del inglés Multilayer Perceptron).

La estructura de este artículo es la siguiente: la sección II abarca la metodología que se llevó a cabo para la recolección y el etiquetado manual de los documentos, así como el proceso para preparar los documentos; en la sección III especifica la

etapa de experimentación y evaluación; en la sección IV se presenta el trabajo a futuro y finalmente en la sección V, las conclusiones.

II. METODOLOGÍA

En esta sección se presenta la metodología que se estableció para llevar a cabo el análisis de los modelos predictivos. Dicha metodología se ve reflejada gráficamente en la Fig. 1 y se compone de las etapas de recolección de datos, generación de un corpus, etiquetado manual de documentos, preprocesamiento de los documentos, extracción de características, ajuste de parámetros para los modelos predictivos y la experimentación junto con la evaluación de los modelos. A continuación, se explican detalladamente cada una de las etapas mencionadas.

Fig. 1. Metodología para el análisis de los modelos predictivos.

A. *Recolección de datos*

En esta etapa, se obtuvieron datos que se descargaron de la plataforma X. Para dicha recolección se utilizó un script escrito en Python que emplea la biblioteca snscrape; tomando en cuenta que el tema de estudio es la detección de ciberbullying, el procedimiento para recolectar los documentos (tweets, ahora llamados posts), fue el siguiente:

- Se realizó una consulta principal con las palabras clave en la presencia del ciberacoso, dichas palabras fueron tomadas de un diccionario de palabras recabado anteriormente en [7], el cual se construyó a través de un análisis cualitativo para determinar la repetitividad de las palabras y contemplar aquellas más comunes en el lenguaje verbal violento empleado en el grupo poblacional comprendido por jóvenes en el ámbito del ciberacoso; dicho diccionario se basó en el español mexicano, sin embargo como trabajo futuro se espera que el mismo pueda crecer con la finalidad de tener un lexicón más amplio e implementarlo en la detección de lenguaje ofensivo. Además, se realizaron consultas de posts aleatorios.
- Las consultas realizadas se limitaron a recabar posts solamente escritos en idioma español, cabe mencionar que debido a la esencia de las palabras clave utilizadas, la mayoría de los posts contemplaron agresiones/blasfemias más generales en el habla hispana y así mismo, propias del español mexicano.
- Los documentos descargados se encontraban en formato JSON (acrónimo de JavaScript Object Notation, notación de objeto de JavaScript) y posteriormente se almacenaron en archivos de valores separados por comas, obteniendo una cantidad de 10 mil documentos o posts de acuerdo a la plataforma X.

Para la generación del corpus, de los 10 mil posts recolectados se tomó una muestra aleatoria de 2500, que conformarían los documentos a etiquetar.

B. *Etiquetado*

El aprendizaje supervisado es un tipo de técnica de aprendizaje automático en la que se entrena a un modelo utilizando un conjunto de datos etiquetado, es decir, un conjunto de ejemplos de entrada junto con las respuestas deseadas. Por lo que, en este trabajo, se llevó a cabo el etiquetado manual de los documentos que conforman el corpus, se dio lectura detallada a cada uno y se le asigno la etiqueta o categoría que mejor correspondiera con lo reflejado en el texto. Las etiquetas que podían ser asignadas fueron las siguientes: “Ciberbullying”, “No Ciberbullying” y “No Relacionado”. Dichas etiquetas o categorías quedaron sujetas a un conjunto de reglas que fueron establecidas, las cuales se exponen a continuación:

- Cyberbullying. Se coloca esta etiqueta a un documento, si contiene:
  - Por lo menos una grosería dirigida a una persona o varias personas.
  - Mensajes que contengan palabras hirientes dirigidas a una o varias personas.
  - Mensajes que inciten a la discriminación.
  - Mensajes que insulten a una persona o varias personas.
  - Mensajes que amenacen a una persona o varias personas.
  - Mensajes que denigren a un sexo en específico.
  - Maldiciones o burlas en contra de una persona o varias personas.
  - Mensajes que contengan un apodo despectivo hacia un individuo o individuos.
  - Mensajes que contengan emojis relacionados con el ciberacoso, tales como:

    (🤢🤮🤡💩💀☠️🤦‍♀️🤦‍♂️🐷🐭🤬🐍🦏🍑🍆🤬👊🖕)
- No Cyberbullying. Se coloca esta etiqueta a un documento, si contiene:
  - Mensajes que contengan una o varias groserías que no estén dirigidas hacia ninguna persona.
  - Mensajes que adulen a una o varias personas.
  - Mensajes que demuestren apoyo hacia una o varias personas.
  - Mensajes cuya finalidad sea alentar o animar a uno o varios individuos.
  - Mensajes que tengan como objetivo defenderse o defender a otra/as personas; sin intensión de insultar.
- No relacionado. Se coloca esta etiqueta a un documento, si contiene:
  - Mensajes que no tengan relación con los puntos anteriores, como noticias, campañas de marketing o publicidad.
  - Mensajes donde únicamente coloquen hashtags o se etiqueten personas

- Mensajes que realicen invitaciones a conferencias, concursos, programas de radio, entre otros.
- Mensajes que contengan imágenes, videos o enlaces.

Una vez realizado el etiquetado de los 2500 documentos seleccionados, la cantidad de documentos categorizados en las diferentes clases concluyó de la siguiente manera:

- Ciberbullying: 616
- No ciberbullying: 495
- No relacionado: 1389

C. Preprocesamiento de textos

Prosiguiendo con la etapa de procesamiento de documentos, ya que los documentos que fueron descargados contienen diversos elementos en el texto, como símbolos, emoticones, hashtags, imágenes, videos, enlaces, entre muchos otros; es necesaria la eliminación de elementos que no contribuyen o que no aportan información relevante para el tema de estudio, en este caso, identificar mensajes ofensivos. Un ejemplo claro de los elementos no relevantes son las palabras vacías (Stop Words), que como se explica en el trabajo de [8] son palabras que se encuentran comúnmente en textos sin dependencia de un tema en particular (por ejemplo, conjunciones, preposiciones, artículos, etc.).

Después de analizar los elementos de importancia en los documentos, se desarrollaron scripts en Python que permitieron la limpieza de cada uno de los documentos; se eliminaron o removieron los siguientes elementos:

- Se eliminaron los acentos y números.
- Caracteres especiales tales como ”·$%&/()=?¿.
- Direcciones de internet, las cuales comienzan con el texto “http”.
- Emoticones no relacionados con el ciberbullying.
- Stopwords o palabras vacías (la, que, el, en, y, a, los, del).

El objetivo de este paso es mejorar la calidad de las características y al mismo tiempo reducir la dificultad del proceso de clasificación [9].

D. Extracción de características

La extracción de características en documentos se refiere al proceso de identificar y representar las características clave o información relevante contenida en un documento de texto. Para ello se requiere que los documentos se encuentren como datos estructurados, es decir, que se vectoricen. La vectorización de un documento implica convertir un documento de texto en una representación numérica (vector). Hay varias técnicas para llevar a cabo la vectorización de documentos.

En el presente trabajo, se empleó la técnica TF-IDF (TF-IDF del inglés Term Frequency – Inverse Document Frequency, Frecuencia de Término – Frecuencia Inversa), la cual elimina los términos más comunes y extrae solo los términos más relevantes del corpus [10]. En esta técnica el valor de cada token aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección de documentos, lo que permite manejar el hecho de que algunas palabras son generalmente más comunes que otras.

Aplicando la técnica TF-IDF sobre el corpus, se obtuvo una matriz dispersa de 2500 documentos y un total de 11191 palabras.

E. Ajuste de parámetros para los modelos predictivos

El proceso de utilizar algoritmos de aprendizaje automático para construir modelos a partir de datos se llama aprendizaje o entrenamiento y el objetivo del aprendizaje automático es encontrar o aproximar la verdad fundamental [11]. Existen varios métodos de aprendizaje automático, para el presente trabajo se contemplaron métodos de aprendizaje supervisado, en el que se entrena a un modelo utilizando un conjunto de datos etiquetado. Cada ejemplo de entrenamiento consta de una entrada (características) y una etiqueta (la salida deseada).

Las técnicas de aprendizaje automático supervisado son aplicables en numerosos ámbitos. Para la identificación de la posible presencia de ciberacoso, se consideraron tres diferentes tipos de métodos de aprendizaje supervisado: Maquina de Vector Soporte (SVM del inglés Support Vector Machine), Bosque aleatorio (RF del inglés Random Forest) y Perceptrón Multicapa (MLP del inglés Multilayer Perceptron); generalmente, las SVM y las redes neuronales tienden a funcionar mucho mejor cuando se trata de características multidimensionales y continuas [12].

Debido a que cada uno de los métodos contienen sus propios parámetros, en lugar de adivinar cuáles son los mejores valores para estos hiperparámetros se optó por emplear el método de búsqueda por rejilla, también conocido como búsqueda en cuadrícula o "grid search" en inglés; lo que permitió encontrar la mejor combinación de hiperparámetros para el modelo, en la Tabla I se muestra la configuración acorde a cada método de clasificación.

TABLA I. CONFIGURACIONES DE HIPERPARÁMETROS A TRAVÉS DEL MÉTODO DE BÚSQUEDA POR REJILLA.

Clases	**Clasificador**		
	SVM	***RF***	***MLP***
TF-IDF	kernel: 'rbf' gamma: 2.3734	criterion: 'gini' max_depth:100 n_estimators: 150	activation: 'tanh' alpha: 0.05 hidden_layer_sizes: (100,) learning_rate: 'adaptive' solver: 'adam'

III. EXPERIMENTACIÓN Y EVALUACIÓN

En esta etapa se llevó a cabo la comparación del desempeño de los diferentes métodos de clasificación empleando la vectorización TF-IDF. Las métricas a tomar en cuenta fueron la precisión (precision), la exhaustividad (recall) y F1-score. Los resultados obtenidos se exponen en la Tabla II.

TABLA II. DESEMPEÑO DE LOS CLASIFICADORES PARA LA VECTORIZACIÓN TF-IDF.

Clases	Clasificador	Precision	Recall	F1-Score
Ciberbullying	SVM	0.57	0.37	0.45
	RF	0.74	0.27	0.39
	MLP	0.59	0.35	0.44
No ciberbullying	SVM	0.66	0.42	0.52
	RF	0.71	0.53	0.61
	MLP	0.62	0.36	0.45
No relacionado	SVM	0.74	0.94	0.83
	RF	0.71	0.98	0.82
	MLP	0.69	0.93	0.79

Se observa que el desempeño de los clasificadores es bajo para las clases ciberbullying y no ciberbullying, debido al desbalance de las clases, pues como previamente se mencionó, dichas clases presentaban una minoría de documentos en comparación con la clase “No relacionado”. El desbalanceo puede afectar negativamente el rendimiento del modelo, ya que el algoritmo puede tener dificultades para aprender la clase minoritaria debido a la falta de instancias.

Existen algunas estrategias comunes para abordar el desbalanceo de clases en el aprendizaje automático, para este caso en particular se optó por emplear SMOTE que es un acrónimo que significa "Synthetic Minority Oversampling Technique". SMOTE es una técnica de sobre muestreo que se utiliza para generar ejemplos sintéticos de la clase minoritaria, lo que ayuda a equilibrar las proporciones de clases en el conjunto de datos sin necesidad de reducir el tamaño de la clase mayoritaria.

Después de aplicar la técnica correspondiente para balancear las clases, nuevamente se llevó a cabo la clasificación para cada uno de los diferentes métodos y los resultados obtenidos se reflejan en la Tabla III; observando que se obtuvieron mejores desempeños en las métricas contempladas.

TABLA III. DESEMPEÑO DE LOS CLASIFICADORES PARA LA VECTORIZACIÓN TF-IDF CON TÉCNICA SMOTE.

Clases	Clasificador	Precision	Recall	F1-Score
Ciberbullying	SVM	1.00	0.73	0.84
	RF	0.95	0.76	0.84
	MLP	0.93	0.91	0.92
No ciberbullying	SVM	0.99	0.84	0.91
	RF	0.88	0.90	0.89
	MLP	0.94	0.95	0.94
No relacionado	SVM	0.70	0.99	0.82
	RF	0.80	0.94	0.87
	MLP	0.90	0.91	0.91

El clasificador que obtuvo mejor desempeño para la vectorización TF-IDF fue MLP, pues presenta mejores métricas y los métodos de clasificación SVM y RF, presentan resultados similares entre sí.

IV. TRABAJO FUTURO

Como trabajo a futuro se contempla añadir el uso de un lexicón que apoye en la identificación de mensajes ofensivos, proporcional a la constante evolución del vocabulario. Además, el uso del algoritmo para la clasificación de posibles incidencias de ciberacoso, será implementado en un módulo de una aplicación móvil en la que se detecten mensajes ofensivos provenientes de la plataforma X.

V. CONCLUSIONES

El ciberacoso es un problema grave y creciente en la era digital, con impactos significativos en la salud mental, la seguridad y el bienestar de las personas, especialmente de los jóvenes. Las redes sociales son un medio común para el ciberacoso debido a su amplio alcance y facilidad de acceso.

Actualmente existen métodos que permiten clasificar mensajes ofensivos; sin embargo, la mayoría están enfocados a otros idiomas lo que conlleva que el lenguaje sea diferente al español, cambiando también las palabras o términos que se suelen usar para ofender o humillar a una persona.

El objetivo principal de este artículo fue realizar el análisis del desempeño de diferentes métodos de aprendizaje supervisado. Se observó que tras una primera clasificación se obtuvieron desempeños bajos en las métricas de las clases “Ciberbullying” y “No ciberbullying”, atribuido al desbalance de clases, por lo que como estrategia para mejorar los desempeños se empleó la técnica SMOTE y una vez realizado el balanceo de clases, los desempeños de los clasificadores mejoraron significativamente. Concluyendo que, para el conjunto de datos presentado, la combinación de TF-IDF y MLP demuestra un mejor desempeño, presentando un puntaje alto en diferentes métricas.

Por otro lado, se puede determinar que la calidad del etiquetado de documentos puede influir en el desempeño de los métodos de clasificación; dicha calidad se refiere a la precisión y consistencia con la que se asignan etiquetas o categorías a los documentos dentro de un conjunto de datos etiquetado. La calidad del etiquetado de documentos puede afectar significativamente la efectividad de los modelos de aprendizaje automático entrenados en esos datos.

AGRADECIMIENTOS

Se agradece al Tecnológico Nacional de México por el apoyo otorgado, mencionando al Instituto Tecnológico de Orizaba por ser el anfitrión del desarrollo de este proyecto.

REFERENCIAS

[1] J. S. Chun, J. Lee, J. Kim, y S. Lee, “An international systematic review of cyberbullying measurements”, Comput Human Behav, vol. 113, núm. July, p. 106485, 2020, doi: 10.1016/j.chb.2020.106485.

[2] S. Hinduja y J. W. Patchin, “Cyberbullying: An exploratory analysis of fac-tors related to offending and victimization”, Deviant Behav, vol. 29, núm. 2, pp. 129–156, 2008, doi: 10.1080/01639620701457816.

[3] INEGI, “Módulo sobre ciberacoso, Comunicado de prensa num. 364, el 13 de julio de 2022”, 2022. [En línea]. Disponible en: https://www.inegi.org.mx/programas/mociba/2021/

[4] D. Mouheb, R. Ismail, S. Al Qaraghuli, Z. Al Aghbari, y I. Kamel, “Detection of Offensive Messages in Arabic Social Media Communications”, Proceedi-ngs of the 2018 13th International Conference on Innovations in Informa-tion Technology, IIT 2018, pp. 24–29, 2019, doi: 10.1109/INNOVATIONS.2018.8606030.

[5] Y. Chen, Y. Zhou, S. Zhu, y H. Xu, “Detecting offensive language in social media to protect adolescent online safety”, Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, Social-Com/PASSAT 2012, pp. 71–80, 2012, doi: 10.1109/SocialCom-PASSAT.2012.55.

[6] G. A. De Souza y M. Da Costa-Abreu, “Automatic offensive language detection from Twitter data using machine learning and feature selection of meta-data”, Proceedings of the International Joint Conference on Neural Net-works, 2020, doi: 10.1109/IJCNN48605.2020.9207652.

[7] L. Rosete, L. A. Reyes, B. A. Olivares, I. Lopez, y L. A. Décaro, “Diseño de un módulo para la detección de ciberbullying en la red social Twitter utilizando lenguaje natural en una aplicación móvil

Android implementado en un entorno universitario", Research in Computing Science, vol. 152, núm. 7, pp. 101–114, 2023, [En línea]. Disponible en: https://www.rcs.cic.ipn.mx/2023_152_7/

[8] A. K. Uysal y S. Gunal, "The impact of preprocessing on text classification", Inf Process Manag, vol. 50, núm. 1, pp. 104–112, 2014, doi: 10.1016/j.ipm.2013.08.006.

[9] A. I. Kadhim, Y. N. Cheah, y N. H. Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering", Pro-ceedings - 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, ICAIET 2014, pp. 69–73, 2015, doi: 10.1109/ICAIET.2014.21.

[10] P. Bafna, D. Pramod, y A. Vaidya, "Document clustering: TF-IDF approach", International Conference on Electrical, Electronics, and Optimization Tech-niques, ICEEOT 2016, pp. 61–66, 2016, doi: 10.1109/ICEEOT.2016.7754750.

[11] Z.-H. Zhou, Machine Learning, 1a ed. Springer Singapore, 2021. doi: https://doi.org/10.1007/978-981-15-1967-3.

[12] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, y J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison", International Journal of Computer Trends and Technology, vol. 48, núm. 3, pp. 128–138, 2017, doi: 10.14445/22312803/ijctt-v48p126.