Detection of Violent Speech Against Women in Mexican Tweets Using an Active Learning Approach

Grisel Miranda-Piña, Roberto Alejo, Eréndira Rendón-Lara, and Vicente García, *Member, IEEE*

***Abstract*—In Latin American and Caribbean States, the verbal violence against women on social networks, such as 𝕏 (formerly known as Twitter), is a serious threat that has been addressed through the implementation of social norms, public policies, and social movements. Nevertheless, a challenge is the effective and automatic real-time detection of violent tweets. In this sense, traditional machine learning algorithms have been proposed to tackle social issues where the training process is performed in a static manner. However, considering that 𝕏 is a dynamic environment where a vast number of tweets are generated each second, it requires powerful machine learning algorithms that could exploit this pool of unlabeled data to be incorporated into the model through continuous updates. This paper explores an active learning method based on uncertainty sampling, which identifies the most confusing tweets to be labeled by an expert in real-time. This focused selection prioritizes which data can be used to train a multilayer perceptron that can achieve a better performance with fewer training samples. Experimental results show that including new samples yields promising results, increasing the AUC from 0.8712 to 0.8833.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8397

***Index Terms*—Violence against women, MLP, active learning, Twitter, 𝕏, Mexican Spanish Language, speech violence detection.**

I. Introducción

La violencia contra las mujeres se reconoce como un problema de alto impacto y una clara violación de los derechos humanos. Conductas como el acoso sexual, la humillación, las amenazas sexuales y de privación de la vida son consideradas delitos que socavan la integridad y la salud de las mujeres, ya sea en el ámbito privado o público [1]. En México, debido a las diferentes realidades económicas, sociales y políticas que caracterizan al país, el número de víctimas de violencia de género ha experimentado un preocupante aumento. Según datos del Instituto Nacional de Estadística y Geografía (INEGI) del Estado Mexicano, la violencia total contra las mujeres aumentó en un 4% durante el período de 2016 al 2021 [2]. En particular, en las zonas rurales, el porcentaje pasó del 54.2% al 60.3%, mientras que en las áreas urbanas creció del 69.3% al 73.0%.

Con la llegada de la pandemia de COVID-19, la violencia digital contra mujeres y niñas tuvo un aumento significativo [3]. En el año 2022, el INEGI reportó que 9.8 millones de mujeres de 12 años en adelante (22.4%) habían experimentado alguna forma de acoso cibernético [4]. Esta problemática ha ganado una mayor atención tanto por parte de autoridades gubernamentales como de las organizaciones no gubernamentales, convirtiéndose en una prioridad en términos de salud pública y de Estado [5]. Un ejemplo de ello es el Gobierno Mexicano, que ha impulsado reformas legislativas para modificar el código penal y tipificar como delito la difusión no autorizada de videos de contenido sexual, considerándola como una forma de violación a la intimidad [6].

La violencia digital a las mujeres no se limita únicamente a actos que involucran el uso de videos, audio o imágenes, sino que también abarca la comunicación verbal escrita que refleja actitudes de desprecio, humillación y degradación hacia ellas, siendo esto también considerado como una forma de violencia [7]. Lewis *et al.* [8] señalan que Twitter, ahora llamada 𝕏, es una plataforma en la que las mujeres son objeto de violencia y abusos que lamentablemente se han vuelto comunes y normalizados. A pesar de que las redes sociales han tomado medidas para mejorar la experiencia del usuario mediante cambios en sus políticas de privacidad y uso, estas acciones aún requieren de una regulación continua [9]. Esta problemática ha dado lugar al surgimiento de iniciativas como los movimientos sociales que buscan neutralizar los efectos de la violencia, entre los que destacan *#NiUnaMas* y *#NiUnaMenos*, que se han convertido en algunas de las manifestaciones mas influyentes con el objetivo de combatir el feminicidio y crear conciencia sobre la violencia en contra de las mujeres en México [10].

Por otro lado, la comunidad científica ha llevado a cabo investigaciones encaminadas a implementar algoritmos y metodologías basadas en el aprendizaje automático (*machine learning*) y el aprendizaje profundo como parte de posibles soluciones a este problema. Por ejemplo, Contreras *et al.* [11] presentan un prototipo basado en el algoritmo SVM (*Support Vector Machine*) para detectar y clasificar las agresiones verbales en mensajes publicados en 𝕏, mientras que Prieto *et al.* [12] proponen un modelo que detecta la violencia contra las mujeres en las redes sociales en el idioma español utilizando técnicas de procesamiento del lenguaje como OM (*Opinion Mining*), DTM (*Document Term Matrix*), BoW (*Bag*

Grisel Miranda, Roberto Alejo (autor para correspondencia) and Eréndira Rendón are with Division of Postgraduate Studies and Research, National Institute of Technology of Mexico (TecNM) Campus Toluca, Metepec, Estado de Mexico, Mexico (e-mails: mm22280266@toluca.tecnm.mx, ralejoe@toluca.tecnm.mx and erendonl@toluca.tecnm.mx).

Vicente García is with Departamento de Ingeniería Eléctrica y Computación, Universidad Autonóma de Ciudad Juárez. Juarez, Chihuahua, Mexico (e-mail: vicente.jimenez@uacj.mx).

of Words) y algoritmos de aprendizaje automático. De manera similar, Salehi *et al.* [13] realizan un estudio sobre la detección automática del riesgo de violencia doméstica contra las mujeres en 𝕏 e Instagram en Irán. Rodríguez-Sánchez *et al.* [14] presentan un extenso estudio sobre la efectividad de los algoritmos de aprendizaje automático y profundo para detectar el sexismo, que abarca el odio explícito, la violencia y las expresiones sutiles dirigidas a las mujeres.

Cada día se publican aproximadamente 500 millones de tuits [15]. Esta vasta cantidad de información sin precedentes representa una gran oportunidad para construir modelos de aprendizaje profundo. Al mismo tiempo, plantea desafíos como la escalabilidad de los modelos tradicionales, la usabilidad, la adaptabilidad, entre otros [16]. Adicionalmente, un reto crítico asociado a la cantidad de datos disponibles es cómo convertirlos en datos etiquetados de alta calidad que permitan mejorar la precisión de los modelos predictivos, teniendo en cuenta que, la construcción de un conjunto de entrenamiento puede suponer un proceso con un alto costo en tiempo y recursos [17].

El aprendizaje activo se ha propuesto como una estrategia colaborativa que implica la participación humana y la asistencia de una máquina (algoritmo de aprendizaje automático), para construir un modelo predictivo y un conjunto de entrenamiento de manera incremental. Este proceso iterativo reduce el costo de obtener información etiquetada al mismo tiempo que aumenta la precisión de las predicciones de los modelos de aprendizaje automático [18]. En este tipo de interacción humano-máquina, donde un ser humano se involucra parcial o totalmente en las diversas etapas, como el entrenamiento, optimización o evaluación de modelos de aprendizaje automático, se le denomina *el humano en el bucle* o *en el circuito*. El objetivo es aprovechar las capacidades cognitivas y el conocimiento de un ser humano para mejorar el rendimiento de los modelos de aprendizaje automático. Por consiguiente, bajo este enfoque híbrido es posible incorporar la participación humana en el etiquetado de datos, donde un algoritmo elige datos sin etiqueta de clase que considera son los más relevantes para ser etiquetados por un experto humano [19].

En este trabajo, exploramos el uso del aprendizaje activo para construir un perceptrón multicapa (MLP, *Multilayer Perceptron*) para la detección de lenguaje violento contra las mujeres en 𝕏. En concreto, el proceso comienza con la construcción del MLP utilizando un conjunto de datos inicial etiquetado por voluntarios y validado por especialistas en el tema de la violencia de género. Los voluntarios fueron capacitados previamente, presentándoles la definición establecida por la Organización de las Naciones Unidas en 1993: "La violencia contra la mujer se entiende todo acto de violencia basado en la pertenencia al sexo femenino que tenga o pueda tener como resultado un daño o sufrimiento físico, sexual o sicológico para la mujer, así como las amenazas de tales actos, la coacción o la privación arbitraria de la libertad, tanto si se producen en la vida pública como en la vida privada" [1]. Dicha violencia fue acotada a expresiones escritas en tuits.

Posteriormente, a partir de la obtención de tuits utilizando una API de *streaming* conectada a 𝕏 por 6 horas, el MLP entrenado estima la probabilidad de que cada tuit pertenezca a una de las dos clases: violenta y no violenta. Estas probabilidades se utilizan para calcular la incertidumbre de cada tuit, definida como la diferencia entre estas dos probabilidades. Si este valor es menor que un umbral establecido, se solicita la etiqueta a un experto y se incorpora al conjunto de entrenamiento. Finalmente, el MLP se entrena nuevamente utilizando el conjunto de entrenamiento actualizado que incluye los datos previos y los nuevos incorporados. Dado que el número de muestras por clase no está balanceado, se aplicó una estrategia de sobremuestreo antes de entrenar el MLP para balancear el conjunto de datos. El MLP es una poderosa herramienta para tareas de clasificación o predicción, incluidas aquellas con aprendizaje profundo [20]. Además, puede formar parte importante en otras redes neuronales más especializadas, como las Convolucionales o las más recientes Transformer. El MLP es ideal para este estudio, ya que los resultados obtenidos podrían transferirse directamente a otros modelos de aprendizaje profundo más complejos en sus capas densas o completamente conectadas.

A partir de lo anterior, este artículo ofrece una doble contribución: primero, muestra cómo un algoritmo de aprendizaje automático es una herramienta útil para abordar un problema social complejo a nivel mundial; y segundo, presenta una estrategia de tipo hombre-máquina que aprovecha la gran cantidad de tuits al seleccionar instancias informativas para que sean etiquetadas por un ser humano. Esto permite, la creación de un conjunto de datos reducido y de alta calidad, que se traduce en un modelo con predicciones precisas.

II. Trabajos Relacionados

En las últimas décadas la computación social, la cual se define como cualquier tipo de aplicación informática en la que el software sirve como intermediario o centro de una relación social [21], se ha desarrollado rápidamente debido a los importantes avances tecnológicos. Este campo tiene aplicaciones en diversas áreas, como el *marketing* y la atención al cliente, donde se utiliza para conocer las opiniones de los usuarios sobre productos y para analizar sus emociones en publicaciones en redes sociales [22]. Cada vez más, la computación social se basa en algoritmos de aprendizaje automático que pueden aprender de los datos. Además, implica el uso de técnicas de Procesamiento de Lenguaje Natural e Inteligencia Artificial para abordar problemas sociales en entornos como las redes socio-digitales. Un ejemplo de esto es la investigación de Cavaliere *et al*. [23] que propone un marco de trabajo para analizar las reacciones emocionales de los usuarios de 𝕏 a lo largo del tiempo durante el periodo de confinamiento por COVID-19. Su objetivo es descubrir tendencias importantes y rastrear la propagación de noticias falsas sobre las vacunas en este contexto de salud pública.

En el ámbito de la detección de violencia en redes sociales, se emplean comúnmente algoritmos tradicionales de aprendizaje automático como SVM, RF (*Random Forest*), NB (*Naive Bayes*), entre otros. Por ejemplo, García-Díaz *et al.* [24] y Gutiérrez-Esparza *et al*. [25] demostraron que al implementar diversos algoritmos, como SVM y RF, para identificar misoginia, racismo, violencia basada en la orientación sexual

y violencia contra la mujer en 𝕏 en español castellano y latinoamericano, lograron valores de exactitud entre el 50% y el 85%. De manera similar, en la clasificación de sentimientos, Masruroh *et al.* [26] evaluaron los clasificadores SVM y NB para analizar los mensajes de 𝕏 relacionados con un proyecto de ley sobre la eliminación de la violencia sexual en Indonesia. Los resultados fueron prometedores, alcanzando valores de exactitud entre el 94% y el 97% al clasificar los tuits en positivos, negativos o neutros en relación con la referida ley.

En años recientes, ha ganado popularidad el uso de algoritmos basados en aprendizaje profundo (*deep learning*) o redes neuronales profundas para tareas de clasificación de sentimientos en texto [27], [28]. Por ejemplo, Aragón *et al.* [29] y Frenda *et al.* [30], emplearon algoritmos como CNN (*Convolutional Neural Network*) y LSTM (*Long-Short Term Memory*), para detectar discursos de odio y agresión en tuits mexicanos (clasificándolos en agresivos y no agresivos), obteniendo resultados por debajo del 60% de exactitud. Por otro lado, Algaradi *et al.* [31] estudiaron varios algoritmos de aprendizaje automático y profundo, incluyendo los actuales *transformers*, para identificar automáticamente si se denuncia la violencia de pareja o no, en 𝕏, logrando resultados entre el 89% y el 95% de exactitud.

Vallecano *et al.* [32] abordaron la detección automática de discursos, escritos en español de España en 𝕏, en dos categorías: odio y no odio. Para ello, diseñaron un algoritmo basado en la red transformadora BERT (*Bidirectional Encoder Representations from Transformers*) denominado HaterBERT (*HaterNet + BERT*), logrando mejoras de entre el 3% y el 27% en comparación con los clasificadores actuales. Díaz *et al.* [33] presentan un sistema de alerta anti-sexista que emplea redes neuronales profundas preentrenadas y técnicas de Procesamiento de Lenguaje Natural para analizar comentarios públicos en español castellano que provienen de periódicos, videos o tuits. Este sistema determina si los comentarios pueden considerarse o no sexistas, obteniendo tasas de efectividad de 0.75 en términos de *F1, precision* (Precisión) y *recall* (Sensibilidad). Por su parte, Li *et al.* [34] investigan hasta qué punto el problema de la misoginia puede explicarse mediante el aprendizaje profundo, utilizando GAT (*Graph Attention Networks*) con aprendizaje semi-supervisado. Su estudio se enfoca en comentarios de redes sociales obtenidos de 𝕏 y Facebook, en los idiomas español, hindi, e inglés. El modelo propuesto obtuvo resultados aproximados al 85% de exactitud en la clasificación de lenguaje misógino y no misógino.

En el contexto del aprendizaje continuo, aún son escasas las investigaciones centradas en la detección automática de violencia en 𝕏. Para este trabajo, se realizó una búsqueda en *Google Scholar, Web of Science* y *Scopus*, utilizando palabras clave como *violence against women, machine learning, continuous*, *active learning*, *Twitter* y 𝕏. Además, se consideró la relevancia de los trabajos y su temporalidad, dando preferencia aquellos con fechas de publicación más actuales. Uno de los trabajos encontrados es el de Elshakankery *et al.* [35], donde se presenta un sistema de aprendizaje semiautomático para analizar los sentimientos en tuits en árabe utilizando el algoritmo *SVM* y el aprendizaje incremental. Este sistema es capaz de actualizar el léxico a medida que se producen cambios lingüísticos. Los resultados muestran que el modelo alcanza un exactitud del 73% para la clasificación en 3 clases (tuits positivos, negativos y neutros) y un 83% para la clasificación en 2 clases (tuits positivos y negativos). Quian *et al.* [36] proponen un enfoque de aprendizaje continuo para clasificar discursos de odio (por grupo al cual pertenecen), en redes sociales, utilizando un Aprendizaje de Representación Variacional con un módulo de memoria basado en LB-SOINN (*Load-Balancing Self-Organizing Incremental Neural Network*). Esto último evita que el algoritmo olvide lo previamente aprendido. Con ello se logran mejores resultados en comparación con las técnicas tradicionales de aprendizaje estático del estado del arte.

Considerando lo discutido anteriormente, se puede notar que la mayoría de los trabajos se enfocan en contextos estáticos, *i.e.*, donde los datos utilizados para construir el clasificador no cambian a medida que se ejecuta el modelo. Sin embargo, diversas fuentes de información cambian con el tiempo, por lo que un modelo estático, aunque pudiera funcionar bien algunos meses o años, no sería capaz de mantener su eficacia, rendimiento y relevancia en problemas en tiempo real, ni adaptarse a los cambios producto de nuevas tendencias, patrones, al mismo tiempo que se optimiza el uso de los recursos.

III. Fundamento Teórico

A. Perceptrón Multicapa de Aprendizaje Profundo

Un MLP de aprendizaje profundo se caracteriza por su arquitectura, que consta de dos o más capas ocultas entre las capas de entrada y salida. No tiene conexiones hacia atrás ni entre neuronas de la misma capa. En este modelo, cada capa, excepto la final (de salida), debe alimentar a las neuronas de la siguiente capa hasta llegar a la final. La adición de muchas capas ocultas implica un problema más complejo y computacionalmente costoso; sin embargo, una de las ventajas del aprendizaje profundo es que la complejidad del modelo se puede resolver reduciendo el número de nodos en las capas ocultas [20].

El entrenamiento de un algoritmo de aprendizaje profundo se puede expresar como un problema de optimización. Una de las metodologías predominantes para el entrenamiento de redes neuronales es el uso de métodos basados en el descenso de gradiente estocástico (SDG, *Stochastic Gradient Descent*). En el contexto del entrenamiento de redes neuronales profundas el algoritmo Adam [37] es uno de los más ampliamente utilizados. Adam minimiza la función de pérdida $f(\mathbf{X}, \mathbf{W})$, donde $\mathbf{X}$ es el conjunto de muestras etiquetadas y $\mathbf{W}$ son los hiperparámetros de los cuales depende $f(\cdot)$, a través de la estimación de razones de aprendizaje adaptativas individuales para los diferentes parámetros [38].

Las redes neuronales artificiales profundas están inspiradas en el funcionamiento de las redes biológicas. En estas últimas, es común el aprendizaje continuo, es decir, adquieren nuevos conocimientos a lo largo de la vida. El aprendizaje continuo tiene como objetivo entrenar múltiples tareas en secuencia para que el modelo sea capaz de recordar tareas previamente aprendidas mientras aprende otras nuevas [39].

El aprendizaje continuo dentro del paradigma del aprendizaje automático ha sido estudiado desde hace ya varias décadas (por ejemplo, véase McCloskey y Cohen [40]). Sin embargo, es común seguir un procedimiento tradicional donde no se adquiere nuevo conocimiento, sino que los datos se ajustan a lo aprendido previamente de forma estática. En este mismo sentido, lo habitual es que los modelos de aprendizaje profundo se entrenen en una etapa y posteriormente, en otra etapa, ejecuten las tareas para los cuales fueron desarrollados. Generalmente, ambas etapas no se retroalimentan entre sí.

B. Aprendizaje Activo

A diario se generan grandes volúmenes de datos que podrían contener información nueva o desconocida, lo que abre la oportunidad para desarrollar modelos de aprendizaje automático eficientes y precisos capaces de adaptarse a nuevas circunstancias. Sin embargo, la calidad y disponibilidad de datos etiquetados requiere de un proceso que es llevado a cabo por uno o varios humanos, resultando costoso en términos de tiempo y recursos. No obstante, si el proceso de etiquetado se lleva a cabo selectiva y repetitivamente solo en ejemplos relevantes o ambiguos, se obtienen ventajas significativas como: maximizar el valor de cada etiqueta al construir modelos con un rendimiento similar o superior utilizando menos datos en comparación con enfoques tradicionales, reducir el costo asociado al etiquetado manual al requerir menos intervención humana, mejorar progresiva y continuamente el modelo, y a diferencia de estrategias estáticas, no se requiere de etiquetar todos los datos disponibles. Por estas razones, se ha propuesto el uso del aprendizaje activo como una estrategia que aprovecha la interacción entre humanos y máquinas para explotar las fortalezas de ambos.

El aprendizaje activo tiene como objetivo reducir el costo temporal y espacial de un proceso de etiquetado, al mismo tiempo que mejora el rendimiento del clasificador al entrenar dicho modelo a partir de un conjunto de entrenamiento que se enriquece iterativamente de datos etiquetados por un humano experto. En este sentido, Cui *et al*. [41] mencionan que no existen métricas que permitan evaluar el desempeño del experto ni comprender la decisión que toma al asignar una de las etiquetas. No obstante, se ha determinado que factores como un exceso de esfuerzo, la frustración y una carga mental y física considerable afectan la tarea del experto.

El etiquetar un dato requiere que primero la máquina, en este caso el clasificador, realice una selección de forma secuencial de aquellas instancias que pudieran ser relevantes, a través de consultas a conjuntos de datos no etiquetados (CDNE) que se asume contienen datos de todas las clases. Para llevar a cabo estas consultas se han propuesto diferentes marcos de trabajo [19]: (a) *Uncertainty Sampling* es el más simple y utilizado, se basa en seleccionar el dato de acuerdo con una probabilidad y un umbral, utilizando alguna métrica de cuantificación de la incertidumbre; (b) *Query-By-Committee* involucra un comité de n modelos con diferentes hipótesis y entrenados en el conjunto etiquetado; cada modelo asigna etiquetas de clase a cada uno de los datos de consulta y selecciona aquellos donde haya más desacuerdo en la asignación de clase realizada por el comité; (c) *Expected Model Change* selecciona los datos con mayor probabilidad de generar cambios significativos en las predicciones del modelo; (d) *Variance Reduction and Fisher Information Ratio*, se seleccionan los datos con mayor varianza en el CDNE; (e) *Estimated Error Reduction*, elige el dato más probable para reducir al máximo el error en un CDNE; y (f) *Density-Weighted Methods*, se basa en la premisa que los datos informativos no solo están sujetos a aquellos que presentan incertidumbre, sino también a aquellos que se encuentran en regiones densas del espacio de entrada.

La definición de lo que se considera como dato relevante dependerá del método de consulta. En el caso de *uncertainty sampling*, el clasificador selecciona aquellas instancias en las que su predicción es altamente incierta, *i.e.*, que podrían localizarse en regiones ambiguas donde la probabilidad de pertenencia a dos o más clases es aproximadamente igual [19]. Esta incertidumbre se puede interpretar como una falta de conocimiento por parte del algoritmo de clasificación, la cual se verá reducida a medida que se agregan más datos al conjunto de entrenamiento [42].

C. Método Tf-idf Vectorizer

Los algoritmos de aprendizaje automático y profundo requieren datos, para su construcción y optimización, en un formato específico. En caso del MLP deben ser vectores numéricos y etiquetados. Para ello, en el contexto de la clasificación de texto (en este trabajo se trata de tuits), es necesario llevar a cabo un proceso conocido como "extracción de características", el cual utiliza una lista de palabras para transformarlas en un conjunto de características únicas [43]. Se han empleado diversas técnicas computacionales para llevar a cabo esta tarea, y una de las más sencillas es Tf-idf (*Term frequency-inverse document frequency*) [44], que es un enfoque comúnmente utilizado en el análisis de sentimientos. Su función principal es reflejar la relevancia de una palabra en un documento dentro de una colección de documentos.

El método Tf-idf Vectorizer utiliza la matriz CountVectorizer [44] y calcula la frecuencia inversa de cada documento (IDF) (Ec. (1)) para normalizar los valores TF-IDF de cada palabra i en el documento D en un rango de 0 a 1 en el vector numérico resultante.

$$z_i = TF_i * IDF_i, \tag{1}$$

donde el peso z_i es una función de TF_i (frecuencia de términos), es decir, el número de veces que aparece la palabra i en un documento D, y la función IDF_i (frecuencia de documento inversa) se encuentra dada por:

$$IDF_i = log(\text{Total de documentos}/DF_i), \tag{2}$$

siendo DF_i (frecuencia de documento) la cantidad de documentos en los que la palabra i aparece al menos una vez. Al aplicar la función IDF_i, se reduce el peso de las palabras de alta frecuencia que no son significativas, como conjunciones, preposiciones o palabras comunes. Esto se debe a que este tipo de palabras aparecerán en varios documentos y, por lo tanto, tendrán un peso mayor que las palabras que pueden ser relevantes en el documento, pero que aparecen menos veces en

los documentos. Para obtener más detalles, consulte el trabajo de Waykole y Thakare [43].

IV. Metodología

A. Conjunto de Datos Inicial

El proceso de aprendizaje activo requiere contar con un conjunto inicial de entrenamiento etiquetado, que se utiliza para inicializar el clasificador. Por lo tanto, se recopilaron un total de 31,240 tuits, de los cuales 28,331 se descargaron a través de la API de *streaming* de $\mathbb{X}$, y 2,909 se extrajeron manualmente de cuentas dedicadas a agredir a las mujeres. Para identificar las cuentas que se dedicaban a agredir a las mujeres, se realizó una búsqueda de tuits utilizando palabras clave (palabras o expresiones usadas para denigrar mujeres en México). A partir de esto, solo se consideraron las cuentas de usuarios que presentaban una alta incidencia de tuits con contenidos violentos hacia la mujer. Luego, al igual que en otros trabajos, se llevó a cabo un etiquetado manual de cada tuit [45], utilizando tres voluntarios previamente capacitados en el tema de violencia de género. Cada participante asignó según su conocimiento y criterio un valor de 0 al tuit que consideraron que debería pertenecer a la clase no violenta (MNV) o un valor de 1 en caso contrario (clase violenta (MV)). Para evitar posibles sesgos en el etiquetado, se empleó un proceso de voto mayoritario que permitió asignar la etiqueta final en función de la que se presentó con más frecuencia entre los tres voluntarios. Al finalizar este proceso, se obtuvo un conjunto de datos base con un desbalance de clases, *i.e.*, con más muestras de la clase MNV que de la MV, lo cual puede afectar el rendimiento del clasificador, como se ha evidenciado en las investigaciones de Abdi y Hashemi [46] y de Rendon *et al*. [47]. A este conjunto de datos inicial se le denominó BD30 y se dividió en dos subconjuntos disjuntos: el conjunto de entrenamiento (CE), que se utiliza para construir y optimizar el modelo, y el conjunto de prueba (CP), que se emplea para medir la efectividad de la propuesta. Ambos conjuntos contienen el 70% y 30% del total de datos, respectivamente, donde $\text{CE} \cap \text{CP} = \emptyset$. Antes de ingresar los datos al modelo neuronal, se aplicó un proceso de limpieza que eliminó información irrelevante para la clasificación de tuits, como caracteres especiales, direcciones URL, emojis, menciones a usuarios (@usuario), etiquetas (#hashtag), artículos, preposiciones y conjunciones, de la misma forma que se hace en trabajos relacionados del estado del arte. Para este trabajo, no se consideraron relevantes los *hashtags*; no obstante, no se descarta la posibilidad de trabajar más adelante con estos elementos y analizar si presentan información valiosa para el análisis de tuits. Posteriormente, se utilizó el método de extracción de características Tf-idf [44] para convertir los datos cualitativos en datos cuantitativos, permitiendo así que el modelo neuronal sea capaz de procesarlos. Por último, se empleó el método de sobremuestreo SMOTE (*Synthetic Minority Over-sampling Technique*) [48], únicamente al CE para lograr un balance en las clases y evitar que el desempeño del clasificador esté sesgado por la clase mayoritaria. SMOTE cuyo objetivo principal es generar muestras sintéticas de la clase minoritaria mediante la interpolación de muestras existentes y cercanas entre sí, es uno de los algoritmos ampliamente utilizados para el tratamiento del desbalance de clases que ha demostrado su efectividad en diversos escenarios y contextos [47]. En este trabajo, se utilizó SMOTE con los parámetros habituales: distancia euclidiana, 5 vecinos más cercanos de la clase minoritaria y un sobremuestreo de dicha clase hasta lograr un balance entre las clases. A este conjunto de datos balanceado se le añadieron las muestras que el clasificador consideró las más informativas.

B. Configuración de la Red Neuronal

El modelo de red seleccionado fue el MLP, desarrollado en la plataforma Google Colaboratory con la version 2.10 para Tensorflow y Keras. Ambas bibliotecas se importaron directamente desde Tensorflow, por lo que comparten la misma versión. Asimismo se usó scikit-learn 1.1.2.

La arquitectura del MLP está compuesta por tres capas ocultas con 10, 5 y 3 neuronas, respectivamente, y una tasa de aprendizaje establecida en 0.0006, con un criterio de parada de 30 iteraciones. Estos valores se determinaron mediante un proceso de prueba y error, en el cual se utilizó el 70% y 30% del CE para entrenar y probar diferentes valores de la tasa de aprendizaje, e identificar el número más apropiado de neuronas e iteraciones basados en los resultados del área bajo la curva ROC (*Receiver Operating Characteristic*) de cada prueba (ver Sección IV-D). Una vez identificados los parámetros que generaron los mejores resultados, se volvió a entrenar el modelo con esos parámetros utilizando el CE completo. Para la evaluación, se usó el CP, el cual no fue usado durante el proceso de prueba y error. En la capa de salida, se utilizó una función softmax [20] con dos neuronas para la clasificación, y se optó por el método de optimización Adam [37].

C. Método Propuesto

En un problema de clasificación binaria, la capa de salida genera un vector numérico con dos valores de probabilidad entre 0 y 1. A partir de este vector, se extrae el valor de probabilidad más alto para asignar la etiqueta de clase a las muestras. No obstante, cuando cada elemento del vector de probabilidad tiene valores cercanos a 0.5, la asignación de la etiqueta de clase puede ser confusa, lo que podría llevar al clasificador a asignar una etiqueta incorrecta [49]. Considerando lo anterior y con base en la idea presentada por Alejo *et al.* [49], se propone un esquema de trabajo que se basa en el uso de un umbral de decisión (μ), para identificar aquellas muestras que podrían ser clasificadas incorrectamente por el clasificador, a través del análisis de las probabilidades obtenidas en la capa de salida. La Fig. 1 presenta el flujo de trabajo propuesto, en el cual se usa μ para determinar si el clasificador debe asignar una etiqueta de clase o abstenerse, evitando así una posible clasificación errónea.

En la primera etapa (Fig. 1, sección a), el CE y el CP provenientes de la base de datos original BD30 o primera iteración (vea sección IV-A) se ingresan al modelo para entrenar y evaluar su efectividad en la clasificación. Posteriormente, en la segunda etapa se presenta el enfoque de umbralización

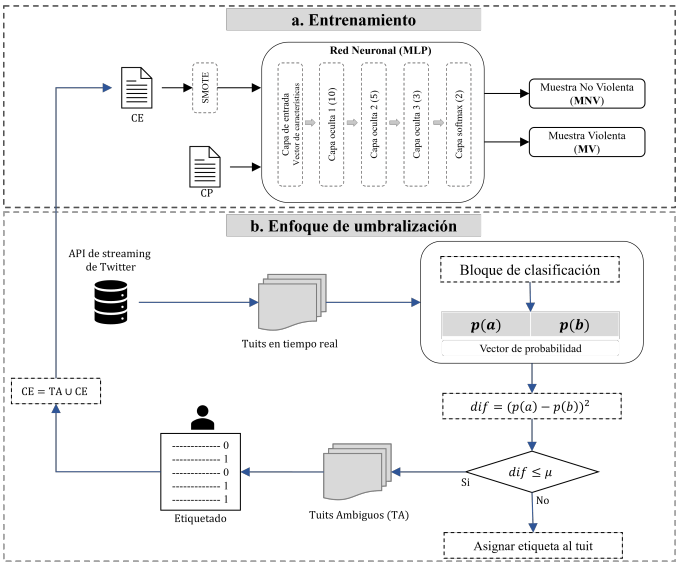

Fig. 1. Esquema de trabajo para la aplicación del umbral de decisión en la clasificación y evitar posibles errores en el etiquetado de la clase.

(Fig. 1, sección b), el cual establece una conexión con la API de *streaming* de $\mathbb{X}$ para acceder a los tuits que están siendo publicados en el territorio mexicano e ir asignado su etiqueta de clase en tiempo real, basándose en lo que el algoritmo aprendió durante su primer entrenamiento (Fig. 1, sección a). En este punto de asignación de etiqueta, los valores de cada elemento del vector de probabilidad generado por cada tuit se extraen y asignan a las variables $p(a)$ y $p(b)$, respectivamente. Una vez realizado este proceso, se obtiene la diferencia de probabilidades (Ec. 3), y este valor se eleva al cuadrado para asegurar que siempre sea positivo, evitando que se interprete como ambiguo debido a tener una diferencia negativa, i.e., un valor inferior al umbral establecido,

$$dif = (p(a) - p(b))^2, \tag{3}$$

donde $p(a)$ corresponde a la probabilidad de la primera posición del vector y $p(b)$ a la probabilidad de la posición 1 del vector, es decir, a su último elemento.

Posterior a la obtención de la diferencia de probabilidades (Ec. 3) y considerando que en una tarea de clasificación binaria el vector de salida deseado (o ideal) debería contener valores cercanos al 0 o 1. En caso contrario, el riesgo de una asignación de clase errónea se incrementa, sobre todo cuando se tiene un vector con probabilidades muy cercanas al 0.5, i.e., cuando no está claramente definida la clase de la muestra evaluada [49]. Por lo tanto, para reducir asignaciones incorrectas se puede usar μ, para que el clasificador se abstenga de asignar una etiqueta cuando tenga duda sobre la etiqueta de clase real de la muestra evaluada. El valor de μ propuesto para este trabajo se estableció en 0.1, y se obtuvo mediante un proceso de prueba y error, similar al mostrado en la Sección IV-B, es decir, se probaron diferentes valores de μ con la configuración óptima para el modelo neuronal. En este procedimiento solo se utilizó el CE, y nunca el CP para evitar resultados sesgados. A continuación, se evalúan todos los tuits obtenidos en un periodo de 9 días; durante los cuales se realizó el entrenamiento y la conexión a la API de *streaming* de $\mathbb{X}$ durante 6 horas cada día. Si el valor de la diferencia obtenida es menor o igual a μ (Ec. 4), entonces el tuit se extrae y almacena en un nuevo conjunto de datos (TA, Tuits Ambiguos), con el fin de que todos los mensajes confusos para el MLP sean analizados y etiquetados manualmente por un experto o grupo de expertos humanos. Luego, el TA (ya etiquetado) se añade al CE = CE ∪ TA, y se sobremuestrea a través de SMOTE para volver a entrenar y optimizar el clasificador, considerando nuevo conocimiento y de esta forma mejorar su efectividad en el proceso de clasificación.

$$dif_k <= \mu \tag{4}$$

La cantidad de muestras recolectadas en cada iteración (día) se presentan en la Tabla I, en donde la primera columna representa los diferentes conjuntos de datos (CD) que se utilizaron en cada iteración y que se distinguen por el nombre original y la versión (v) de acuerdo al día en que se probó el modelo. Las siguientes columnas indican el total de características y muestras con las que se entrenó el modelo, así como los tuits que se agregaron al CE en cada iteración (incremento), respectivamente.

TABLA I
CARACTERÍSTICAS, NÚMERO DE MV Y MNV UTILIZADAS CON EL ENFOQUE DE UMBRAL PROPUESTO, DONDE CD E INCR. SON LOS ACRÓNIMOS DE CONJUNTO DE DATOS E INCREMENTO

CD	Características	MV	Incr.	MNV	Incr.	Total
BD30v1	11204	2080	80	20512	512	22592
BD30v2	11402	2153	73	21009	497	23162
BD30v3	11602	2189	36	21521	512	23710
BD30v4	11787	2236	47	21977	456	24213
BD30v5	11984	2298	62	22309	332	24607
BD30v6	12418	2378	80	22854	545	25232
BD30v7	12499	2405	27	23084	230	25489
BD30v8	12706	2513	108	23533	449	26046
BD30v9	12816	2520	7	23950	417	26470

D. Rendimiento del Clasificador

Para evaluar la efectividad del modelo estudiado, se utilizó el área bajo la curva ROC, también conocida como AUC (*Area Under the Curve ROC*). Es una métrica comúnmente empleada en problemas de clasificación cuando existe un desbalance de clases [46]. Se caracteriza por tener en cuenta el valor de exactitud de cada clase, evitando el sesgo hacia la clase mayoritaria. La definición formal de esta métrica viene dada por la Ec. (5) y sus variables por las Ecs. (6) y (7),

$$AUC = \frac{sensibilidad + especificidad}{2} \tag{5}$$

$$sensibilidad = \frac{vp}{vp + fn} \tag{6}$$

$$especificidad = \frac{vn}{vn + fp} \tag{7}$$

donde la sensibilidad se define como la proporción de datos violentos correctamente identificados (verdaderos positivos,

vp) y la especificidad como la proporción de datos no violentos correctamente etiquetados (verdaderos negativos, vn), las cuales son obtenidas a partir de una matriz de confusión (vea Tabla II).

TABLA II
MATRIZ DE CONFUSIÓN PARA CLASIFICACIÓN BINARIA

		Predicción	
		Positivo	Negativo
Real	Positivo	Verdaderos Positivos (vp)	Falsos Negativos (fn)
	Negativo	Falsos Positivos (fp)	Verdaderos Negativos (vn)

V. RESULTADOS Y DISCUSIÓN

La Tabla I muestra claramente que el número de tuits detectados utilizando el mecanismo de μ es mucho menor para la clase MV que para la clase MNV. Existe una diferencia aproximada de 365 tuits, lo que continúa generando un CD con clases no balanceadas. Por otro lado, de una iteración a otra no se observa una tendencia en el incremento del número de tuits identificados por la Ec. 4.

Los valores de especificidad, sensibilidad y AUC obtenidos, tanto con la base de datos original BD30 como con los generados en cada uno de los 9 experimentos realizados empleando el umbral de decisión, se presentan en la Tabla III. Según los resultados mostrados en esta tabla, se observa una tendencia al aumento en los valores de AUC en cada prueba o iteración, lo que indica que este enfoque permite mejorar el rendimiento de clasificación, reduciendo la confusión del algoritmo al asignar una etiqueta a un tuit. Asimismo, los resultados sugieren que Tf-idf es adecuada para construir el CD con el cual se entrenará el clasificador para identificar lenguaje violento contra las mujeres en tuits en el idioma español de México.

Por otra parte, se observa que a medida que el número de características aumenta, también lo hace el valor de AUC, lo que podría interpretarse como que un mayor número de características proporciona más información a la red para aprender a realizar una clasificación binaria más precisa.

En lo que respecta a la efectividad sobre la clase de interés, MV, se observa un ligero incremento en los valores de sensibilidad de una iteración a otra. Por ejemplo, desde la iteración 0 (con el CD original) hasta la iteración 6, se registran incrementos en los valores de sensibilidad en promedio de 0.0018. Sin embargo, de la iteración 6 a la 7, se observa una disminución de 0.022, para después volver a aumentar en la 8 y disminuir en la 9. En otras palabras, aunque el enfoque de umbralización con aprendizaje activo permite mejorar los valores de AUC, la sensibilidad no muestra un comportamiento consistente, pero tampoco empeora de forma importante la efectividad del clasificador en la clase MV.

En cuanto a la clase mayoritaria, MNV, se nota una tendencia al aumento en los valores de especificidad, aunque en algunas iteraciones disminuyen, como en la transición de la iteración 4 a la 5. Sin embargo, estos decrementos no son considerables. En otras palabras, existe la limitante

TABLA III
VALORES DE ESPECIFICIDAD, SENSIBILIDAD Y AUC CORRESPONDIENTES AL CD ORIGINAL (BD30), Y A LOS 9 EXPERIMENTOS REALIZADOS CON EL ENFOQUE DE APRENDIZAJE ACTIVO

CD	Especificidad	Sensibilidad	AUC
BD30	0.8953	0.8471	0.8712
BD30v1	0.8982	0.8493	0.8737
BD30v2	0.8947	0.8539	0.8743
BD30v3	0.8974	0.8526	0.8750
BD30v4	0.9003	0.8526	0.8764
BD30v5	0.8982	0.8548	0.8765
BD30v6	0.8995	0.8581	0.8788
BD30v7	0.9219	0.8361	0.8790
BD30v8	0.9045	0.8539	0.8792
BD30v9	0.9282	0.8383	0.8833

de que la clase que se ve más beneficiada por el enfoque propuesto es la MNV, cuando lo ideal sería el aumento de los valores de sensibilidad en la clase MV. A pesar de esto, los incrementos en los valores de AUC respaldan la validez del método propuesto. No obstante, la principal amenaza a la confiabilidad y validez del estudio está relacionada con el cálculo del umbral μ. Este valor debe ser obtenido de manera precisa, ya que es fundamental para la identificación de tuits que podrían ser clasificados incorrectamente. Además de la necesidad de la intervención humana, la cual retroalimentará al modelo para trabajar de una manera más certera.

La Tabla IV muestra un resumen comparativo de algunos trabajos del estado del arte enfocados en la detección de lenguaje violento contra la mujer en comentarios de diversas redes sociales.

Al comparar los resultados obtenidos en esta investigación con otros trabajos(ver Tabla IV), se observan diferencias que limitan la realización de una comparación justa, como las métricas de evaluación empleadas o el conjunto de datos usado, lo que puede causar diferencias en los resultados del clasificador. Sin embargo, se pueden discutir algunas de ellas, por ejemplo, en el trabajo de Prieto Cruz y Montoya Vasquez [12], el tamaño de la clase MV predomina sobre la clase MNV, lo que resulta en tasas de clasificación sobre la clase de interés por encima del 90% en términos de sensibilidad. Otro ejemplo, es el presentado por Al-Garadi *et al*. [31], que muestra valores de exactitud entre 85% y 95%, pero un desempeño en la clase minoritaria (con puntaje F1) de 0.76 y un conjunto de datos que incluye solo 7016 tuits, i.e., el desempeño en la clase principal es menor que el obtenido por nuestro método propuesto. En los trabajos de Salehi *et al.* [13], Frenda *et al.* [30] y Díaz *et al.* [33] el tamaño del conjunto de datos también es menor (el de mayor tamaño tiene 11,000 comentarios aproximadamente). Además, estos conjuntos de datos muestran un menor nivel de desbalance en comparación con este trabajo, que utilizó 31,240 comentarios con un alto nivel de desbalance (consulte la Sección IV-A). En las investigaciones de García-Díaz *et al.* [24] y Gutiérrez-Esparza *et al*. [25], el conjunto de datos usado es balanceado y su tamaño es inferior al estudiado aquí. A pesar de las

TABLA IV
TRABAJOS REPRESENTATIVOS DEL ESTADO DEL ARTE EN LA DETECCIÓN DE LENGUAJE VIOLENTO CONTRA LA MUJER EN COMENTARIOS EN DIFERENTES REDES SOCIALES-ONLINE. NC ES EL NÚMERO DE COMENTARIOS O EN SU CASO TUITS, Y RD=MV/MNV ES LA RAZÓN DE DESBALANCE (ENTRE MÁS CERCA DE CERO ESTÉ RD, MÁS ALTO ES EL NIVEL DE DESBALANCE)

Ref.	NC	RD	Modelo	Métrica	Valor
[12]	1507	0.6487	SVM	Sensibilidad	0.9398
			NB		0.8032
			DT		0.9726
[13]	1611	0.1657	SVM	Exactitud	0.7500
			RF		0.7355
			LR		0.8646
			DT		0.6900
			NB		0.8677
[24]	7682	1	RF	Exactitud	0.7930
			SMO		0.8517
			LSVM		0.8288
[25]	2000	0.5384	RF	Exactitud	0.8423
			OneR		0.9564
[30]	10856	0.5483	CNN	F1-score	0.4400
				Sensibilidad	0.5300
				Precisión	0.3700
[31]	7016	0.1176	DT	Exactitud	0.8900
			SVM		0.9300
			NN		0.9100
			BiLSTM		0.9100
			BERT		0.9400
			RoBERTa		0.9500
[33]	3794	1	Transformer	Sensibilidad	0.7500
				Precisión	0.7500
				F1-score	0.7500
Propuesta	35710	0.1062	MLP	AUC	0.8833
				Sensibilidad	0.8383
				Especificidad	0.9282

diferencias en objetivos, datos usados, configuraciones en estos estudios, la propuesta presentada en este trabajo muestra resultados altamente competitivos.

VI. CONCLUSIONES

La detección de violencia contra las mujeres en mensajes de 𝕏 sigue siendo un desafío actual, ya que es de suma importancia determinar si existen o no expresiones violentas escritas dirigidas hacia las mujeres. En este trabajo, se abordó este problema desde la perspectiva del aprendizaje activo, en el cual se presentó una técnica para evitar que el clasificador se abstenga de asignar una etiqueta cuando existe incertidumbre acerca de la veracidad de la etiqueta que está asignando. Para ello, se utilizó un umbral para determinar qué tuits deberían ser enviados a un experto o grupo de expertos para su etiquetado.

Una limitación de este estudio es que no exploramos otras métricas para cuantificar la incertidumbre, como la entropía, así como la falta de un análisis teórico. No obstante, los resultados experimentales en este trabajo muestran que la estrategia propuesta aumenta la efectividad del clasificador a medida que se añaden más muestras etiquetadas (por un experto o grupo de ellos) al conjunto de entrenamiento. Asimismo, en términos de sensibilidad no se observa un claro incremento en sus valores, lo cual podría estar influenciado por la efectividad del método de manejo de desbalance empleado para abordar este problema. Por otro lado, en cuanto a la especificidad, es evidente la tendencia a incrementar sus valores, lo cual está relacionado con el hecho de que el número de muestras de la clase MNV agregadas al conjunto de entrenamiento es considerablemente mayor que el de la clase MV.

Actualmente se trabaja en mejorar esta propuesta, buscando mecanismos más apropiados para aumentar el número de muestras en la clase minoritarias y abordar el desbalance. Esto es especialmente relevante debido a la naturaleza del tema estudiado, que implica el análisis de textos asociados a problemáticas sociales como la detección de lenguaje violento contra la mujer, la identificación de misoginia o xenofobia, entre otros. Además, se está considerando el uso de algoritmos de clasificación más sofisticados, como las redes *transformer*, y técnicas más avanzadas de procesamiento de lenguaje natural (como la tecnología *Word Embeddings*, que permite codificar la semántica como la relación de las palabras entre sí, otorgando contexto a las palabras analizadas), que puedan contribuir a mejorar la efectividad de nuestro enfoque. Finalmente, se pretende explotar la información que los *hashtags* pueden proveer al ser utilizados como parte gramatical de la oración. Esto posiblemente complementaría el tuit al ofrecer información semántica, contextual y organizativa.

AGRADECIMENTOS

Esta investigación fue financiada parcialmente por el proyecto 10880.21-P del TecNM, y por el CONAHCyT (México) con la beca número 1171228.

REFERENCES

[1] ONU, "Declaration on the elimination of violence against women," *UN General Assembly: New York, NY, USA*, 1993.
[2] INEGI, "Violencia contra las mujeres en méxico." http://tinyurl.com/2awdwpkk, 2023. Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH). Ediciones 2016 y 2021.
[3] U. Women, "Covid-19 and ending violence against women and girls." http://tinyurl.com/dua9vj2p, 2020. Policy brief no. 17.
[4] INEGI, "Comunicado de prensa núm. 404/23." http://tinyurl.com/mryvjbm7, 2023. Módulo sobre ciberacoso 2022.
[5] WHO, "Violence against women prevalence estimates, 2018." http://tinyurl.com/4k4mnnun, 2021. Accessed: 20-07-2023.
[6] O. J. Nacional, "Ficha Técnica - Ley Olimpia." http://tinyurl.com/mwpxmv85. Accessed: 16-09-2023.
[7] V. Castro, C. L. Vidal, and R. S. Riquelme, "Detección de violencia verbal hacia las mujeres en redes sociales mediante técnicas de aprendizaje automático," *Repositorio Digital Sistema de Bibliotecas Universidad del Bio-Bio (SIBUBB)*, 2019.
[8] R. Lewis, M. Rowe, and C. Wiper, "Online abuse of feminists as an emerging form of violence against women and girls," *British journal of criminology*, vol. 57, no. 6, pp. 1462–1481, 2017.
[9] G. M. Abaido, "Cyberbullying on social media platforms among university students in the united arab emirates," *International Journal of Adolescence and Youth*, vol. 25, no. 1, pp. 407–420, 2020.
[10] A. Prusa, B. G. Nice, and O. Soledad, "Not one women less, not one more death: Feminist activism and policy responses to gender-based violence in latin america." http://tinyurl.com/yck644xn. Accessed: 30-07-2023.
[11] M. E. R. Contreras and J. V. Alvarez, *Reconocimiento de agresión verbal en Twitter con el uso de patrones lingüísticos*. PhD thesis, Pontificia Universidad Católica de Valparaíso, 2017.
[12] G. A. P. Cruz and E. E. M. Vasquez, "Modelo de detección de violencia contra la mujer en redes sociales en español, utilizando opinion mining," bachelor's thesis, Universidad Tecnológica de Perú, 2020.

[13] M. Salehi, S. Ghahari, M. Hosseinzadeh, and L. Ghalichi, "Domestic violence risk prediction in Iran using a machine learning approach by analyzing Persian textual content in social media," *Heliyon*, vol. 9, no. 5, p. e15667, 2023.

[14] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, and L. Plaza, "Automatic classification of sexism in social networks: An empirical study on twitter data," *IEEE Access*, vol. 8, pp. 219563–219576, 2020.

[15] J. M. Lane, D. Habib, and B. Curtis, "Linguistic methodologies to surveil the leading causes of mortality: Scoping review of twitter for public health data," *J Med Internet Res*, vol. 25, p. e39484, 2023.

[16] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[17] IBM, "¿qué es el etiquetado de datos?." https://www.ibm.com/es-es/topics/data-labeling, 2023. Accessed: 20-07-2023.

[18] J. Bengar, J. van de Weijer, B. Twardowski, and B. Raducanu, "Reducing label effort: Self-supervised meets active learning," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1631–1639, 2021.

[19] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2023.

[20] Y. B. I. Goodfellow and A. Courville, *Deep Learning*. MIT Press, 2016.

[21] D. Schuler, "Social computing," *Communications of the ACM*, vol. 37, no. 1, pp. 28–29, 1994.

[22] M. Riveni, T.-D. Nguyen, M. S. Aktas, and S. Dustdar, "Application of provenance in social computing: A case study," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 3, p. e4894, 2019.

[23] D. Cavaliere, G. Fenza, V. Loia, and F. Nota, "Emotion-aware monitoring of users' reaction with a multi-perspective analysis of long- and short-term topics on twitter," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. In Press, no. In Press, pp. 1–10, 2023.

[24] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García, "Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings," *Future Generation Computer Systems*, vol. 114, pp. 506–518, 2021.

[25] G. O. Gutiérrez-Esparza, M. Vallejo-Allende, and J. Hernández-Torruco, "Classification of cyber-aggression cases applying machine learning," *Applied Sciences*, vol. 9, no. 9, p. 1828, 2019.

[26] S. U. Masruroh, D. Z. A. Utami, D. Khairani, M. Azhari, M. I. Helmi, and R. A. Putri, "Sentiment analysis on twitter towards the ratification of a bill on the elimination of sexual violence in Indonesia using machine learning," in *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–5, 2022.

[27] P. Kapil, A. Ekbal, and D. Das, "Investigating deep learning approaches for hate speech detection in social media," *arXiv preprint arXiv:2005.14690*, 2020.

[28] S. Adeeba, K. Banujan, B. T. G. S. Kumara, and Z. Li, "Twitter mining for detecting home violence," in *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, pp. 142–147, 2023.

[29] M. E. Aragón and A. P. López-Monroy, "Author profiling and aggressiveness detection in spanish tweets: Mex-a3t 2018.," in *IberEval SEPLN*, pp. 134–139, 2018.

[30] S. Frenda, S. Banerjee, P. Rosso, and V. Patti, "Do linguistic features help deep learning? the case of aggressiveness in mexican tweets," *Computación y Sistemas*, vol. 24, no. 2, pp. 633–643, 2020.

[31] M. A. Al-Garadi, S. Kim, Y. Guo, E. Warren, Y.-C. Yang, S. Lakamana, and A. Sarker, "Natural language model for automatic identification of intimate partner violence reports from twitter," *Array*, vol. 15, p. 100217, 2022.

[32] G. del Valle-Cano, L. Quijano-Sánchez, F. Liberatore, and J. Gómez, "Socialhaterbert: A dichotomous approach for automatically detecting hate speech on twitter through textual analysis and user profiles," *Expert Systems with Applications*, vol. 216, p. 119446, 2023.

[33] R. P. Díaz Redondo, A. Fernández Vilas, M. Ramos Merino, S. M. Valladares Rodríguez, S. Torres Guijarro, and M. M. Hafez, "Anti-sexism alert system: Identification of sexist comments on social media using ai techniques," *Applied Sciences*, vol. 13, no. 7, pp. 1–14, 2023.

[34] K. Li, "An evaluation of automation on misogyny identification(ami) and deep-learning approaches for hate speech -highlight on graph convolutional networks and neural networks," in *2022 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, pp. 239–244, 2022.

[35] K. Elshakankery and M. F. Ahmed, "Hilatsa: A hybrid incremental learning approach for arabic tweets sentiment analysis," *Egyptian Informatics Journal*, vol. 20, no. 3, pp. 163–171, 2019.

[36] J. Qian, H. Wang, M. ElSherief, and X. Yan, "Lifelong learning of hate speech classification on social media," *arXiv preprint arXiv:2106.02821*, 2021.

[37] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, p. 1–14, 2016.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, p. 1–15, 2017.

[39] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "Ernie 2.0: A continual pre-training framework for language understanding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 8968–8975, 2020.

[40] M. McCloskey and N. J. Cohen, *Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem*, vol. 24, pp. 109–165. Academic Press, 1989.

[41] Y. Cui, P. Koppol, H. Admoni, S. Niekum, R. Simmons, A. Steinfeld, and T. Fitzgerald, "Understanding the relationship between interactions and outcomes in human-in-the-loop machine learning," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4382–4391, 2021.

[42] V.-L. Nguyen, M. H. Shaker, and E. Hüllermeier, "How to measure uncertainty in uncertainty sampling for active learning," *Machine Learning*, vol. 111, no. 1, pp. 89–122, 2022.

[43] R. N. Waykole and A. D. Thakare, "A review of feature extraction methods for text classification," *Int. J. Adv. Eng. Res. Dev*, vol. 5, no. 04, pp. 351–354, 2018.

[44] U. Sharma and J. Singh, "Review of feature extraction techniques for fake news detection," in *Advances in Information Communication Technology and Computing*, (Singapore), pp. 389–399, 2023.

[45] S. Arroni, Y. Galán, X. Guzmán-Guzmán, E. R. Nuñez-Valdez, and A. Gómez, "Sentiment analysis and classification of hotel opinions in twitter with the transformer architecture," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 53–63, 2023.

[46] L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 238–251, 2016.

[47] E. Rendon, R. Alejo, C. Castorena, F. J. Isidro-Ortega, and E. E. Granda-Gutierrez, "Data sampling methods to deal with the big data multi-class imbalance problem," *Applied Sciences*, vol. 10, no. 4, p. 1276, 2020.

[48] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[49] R. Alejo, J. Monroy-de Jesús, J. H. Pacheco-Sánchez, E. López-González, and J. A. Antonio-Velázquez, "A selective dynamic sampling back-propagation approach for handling the two-class imbalance problem," *Applied Sciences*, vol. 6, no. 7, p. 200, 2016.

Grisel Miranda-Piña is a computer Systems Engineer graduated in 2021 from the Technological Institute of Higher Studies of Jocotitlan. She is pursuing a Master's degree in Engineering Sciences at the Technological Institute of Toluca (Mexico); her research interests focus on the applications of artificial intelligence and artificial neural networks in solving real-world problems within a big data context.

Roberto Alejo received a Ph.D. in Advanced Computer Systems from the Universitat Jaume I (Spain) and a full-time professor at the Technological Institute of Toluca, National Technological Institute of Mexico, with a profound scientific interest in applying artificial intelligence to real-world problem-solving. It specializes in artificial neural networks, machine learning, and data mining.

Eréndira Rendón-Lara received a Ph.D. in Computer Science from the Technological Institute of Toluca. Currently serving as a professor-researcher in the Division of Graduate Studies and Research at the National Technological Institute of Mexico, Toluca campus. Her primary academic interests focus on data mining and, more recently, on Material Informatics.

Vicente García received a Ph.D. in Advanced Computer Systems from Universitat Jaume I (Spain), in 2010. He is a full-time professor in the Department of Electrical Engineering and Computer Science at the Autonomous University of Ciudad Juarez. His research interests include data preprocessing methods, data complexity, non-parametric classification, performance evaluation, and big data.