INVESTIGACIÓN E INNOVACIÓN EN INGENIERÍAS

Cyberbullying detection in Spanish using natural language processing and machine learning

Detección de ciberacoso en español mediante procesamiento de lenguaje natural y aprendizaje automático

Angélica María Agudelo Ortiz
Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle, Cali
Laura Sofía Rodríguez Pulecio
Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle, Cali
Oscar Fernando Bedoya Leiva
Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle, Cali

OPEN ACCESS

Recibido:
24/01/2025
Aceptado:
17/07/2025
Publicado:
02/10/2025

Correspondencia:
oscar.bedoya@correounivalle.edu.co

DOI:
https://doi.org/10.17081/invinno.13.2.7922

Resumen

Objetivo: Detectar el ciberacoso en línea aplicando un enfoque interseccional mediante el uso de procesamiento de lenguaje natural (PLN) y técnicas de aprendizaje automático, específicamente dirigidas al idioma español en la red social X (anteriormente Twitter). El estudio se centra en incorporar factores interseccionales como género, raza, etnia y orientación sexual para identificar instancias de ciberacoso. **Metodología:** El proceso comenzó con la identificación de palabras clave relevantes para detectar el ciberacoso, guiado por investigaciones previas sobre los patrones lingüísticos asociados con el acoso en línea. Estas palabras clave se utilizaron para recopilar tweets, resultando en un conjunto de datos diverso que refleja diversas expresiones vinculadas al ciberacoso. Un panel de expertos anotó el conjunto de datos, categorizando los tweets como "Ciberacoso" o "No Ciberacoso," con un enfoque en la región colombiana. Posteriormente, se entrenaron tres modelos de aprendizaje automático—Naive Bayes, Máquinas de Soporte Vectorial (SVM) y Redes Neuronales—para predecir el ciberacoso. **Resultados:** Naive Bayes alcanzó un área bajo la curva ROC del 89.7%, SVM llegó al 90.9%, y las Redes Neuronales registraron el mejor rendimiento con un AUC del 91%. **Conclusiones**: Los resultados evidencian la efectividad de los modelos de aprendizaje automático, en particular de las Redes Neuronales, para detectar ciberacoso en español. La incorporación de un enfoque interseccional permitió comprender con mayor profundidad cómo se manifiestan estas expresiones dañinas en los entornos digitales.

Palabras claves: Aprendizaje automático, Ciberacoso, Procesamiento de lenguaje natural, Redes neuronales, Red social X.

Abstract

Objective: To detect online cyberbullying by applying an intersectional approach using natural language processing (NLP) and machine learning techniques, specifically targeting the Spanish language on the social network X (formerly Twitter). The study focuses on incorporating intersectional factors such as gender, race, ethnicity, and sexual orientation in identifying cyberbullying instances. **Methodology:** The process began with the identification of relevant keywords to detect cyberbullying, guided by previous research on language patterns associated with online harassment. These keywords were used to collect tweets, resulting in a diverse dataset reflecting various expressions linked to cyberbullying. A panel of experts annotated the dataset, categorizing tweets as either "Cyberbullying" or "No Cyberbullying," with a focus on the Colombian region. Subsequently, three machine learning models—Naive Bayes, Support Vector Machines (SVM), and Neural Networks—were trained to predict cyberbullying. **Results:** Naive Bayes achieved an area under the ROC curve of 89.7%, SVM reached 90.9%, and Neural Networks recorded the highest performance with an AUC of 91%. **Conclusions**: The findings demonstrate the effectiveness of machine learning models—particularly Neural Networks—for detecting cyberbullying in Spanish-language social media data. Incorporating an intersectional perspective provided deeper insights into how harmful expressions manifest in online communication.

Keywords: Machine learning, Cyberbullying, Natural Language Processing, Neural Networks, Social Network X.

Como citar (IEEE): A. M. Agudelo-Ortiz, L. S. Rodríguez-Pulecio y O. F. Bedoya-Leiva, "Cyberbullying detection in Spanish using natural language processing and machine learning," Investigación e Innovación en Ingenierías, vol. 13, no. 2, pp. 168–183, 2025, doi: https://doi.org/10.17081/invinno.13.2.7922

Introduction

Bullying is an intentional behavior aimed at harming, intimidating, or humiliating another person. This conduct can manifest in various forms, including physical, verbal, psychological, or cyber harassment, and commonly occurs in settings such as schools, workplaces, and online environments. According to the National Center for Education Statistics (NCES), approximately 20% of students aged 12 to 18 reported being victims of bullying in schools in 2019 [1]. Similarly, UNICEF reported in 2019 that around 30% of students in Latin America and the Caribbean have experienced some form of school bullying [2]. Furthermore, the European Union Agency for Fundamental Rights (FRA) found in 2020 that 54% of LGBTI youth have faced bullying or discrimination in schools due to their sexual orientation or gender identity [3]. Likewise, the Inter-American Commission on Human Rights (IACHR) has documented numerous cases of violence and discrimination against LGBTQ+ individuals in the region [4]. A 2020 report by Colombia Diversa indicated that 68% of LGBTQ+ individuals had experienced some form of violence or discrimination in educational settings [5]. Collectively, these figures highlight the prevalence of bullying and the need for targeted detection strategies.

Gender-based violence (GBV), including cyber-GBV against women and other vulnerable groups, remains a pressing global and regional issue. The World Health Organization (WHO) estimates that one in three women experiences physical or sexual violence in their lifetime [6], and in Latin America, the Economic Commission for Latin America and the Caribbean (ECLAC) report similar prevalence [7]. In Colombia, over 77,000 cases of GBV against women were reported in 2020 [8]. Online environments amplify these inequalities: between 20% and 23% of women in multiple countries have experienced cyberbullying or online harassment [9-11]. Vulnerable youth, including LGBTQ+, Indigenous, and Afro-descendant populations, are also disproportionately affected, experiencing both offline and online harassment [3,5,16-21]. In Colombia, surveys indicate that seven out of ten children have suffered cyberbullying, and digital violence against women and LGBTQ+ individuals is widespread [16–21]. These statistics underscore the urgent need for effective detection and mitigation strategies.

Cyberbullying detection has increasingly relied on natural language processing (NLP) and machine learning (ML) techniques. Studies have applied features such as Bag-of-Words (BoW) and TF-IDF with classifiers like SVM, Naive Bayes, and Random Forests [22], dual contrastive learning for hate speech detection [23], and language-specific adaptations such as AraBERT for Arabic social media [24]. Other works integrated NLP and ML for political opinion analysis [25], misinformation detection [26], platform-specific cyberbullying identification [27], and toxic behavior in online gaming [28]. In Spanish-language contexts, recent studies have tailored NLP and ML approaches to the linguistic and cultural nuances of Twitter, demonstrating effective detection of harmful content [29,30]. Collectively, these studies highlight the need for cyberbullying detection approaches that are tailored to the language and platform involved, and that remain sensitive to the characteristics of diverse populations.

One of the challenges in detecting cyberbullying on social media is the limitation related to language and regional expressions. Cyberbullying detection systems are often designed primarily to identify patterns of harassment in languages such as English [29]. This can hinder the effective detection of cyberbullying in other languages, including regional and colloquial expressions that may vary significantly. On social media platforms like X, where short and rapid messages are shared, users frequently employ slang and expressions unique to their region. These expressions may include insults, offensive language, or abusive content that constitutes cyberbullying. Consequently, detection systems based on natural language processing algorithms may struggle to identify and understand such regional expressions, limiting their ability to effectively detect and combat cyberbullying in diverse linguistic contexts.

Moreover, regional expressions can vary even within the same language. For instance, in Colombia, linguistic variations and region-specific expressions are prevalent across different parts of the country. These regional expressions can be used in cyberbullying, further complicating detection and analysis. This limitation in detecting cyberbullying based on regional expressions can have serious consequences, as it prevents the effective protection of individuals affected by online harassment. Perpetrators who use regional or colloquial expressions to harass others may go undetected by existing detection systems, allowing harassment to continue without consequences. This technological gap poses significant challenges in protecting Spanish-speaking women from online cyberbullying and underscores the need to develop specific tools to address this issue in the Spanish-speaking world.

To address this problem, this study employed a methodology based on natural language processing and machine learning to develop cyberbullying detection models for the social network X in the Spanish language, focusing on gender, race, and sexual orientation. Keywords were selected based on aspects of intersectionality, such as gender, race, ethnicity, and sexual orientation, identified from previous studies by other authors. Tweets containing these keywords were then collected to form a dataset, which was manually annotated by a panel of experts on gender issues. Finally, machine learning models were developed and evaluated to predict whether a tweet contains cyberbullying or not.

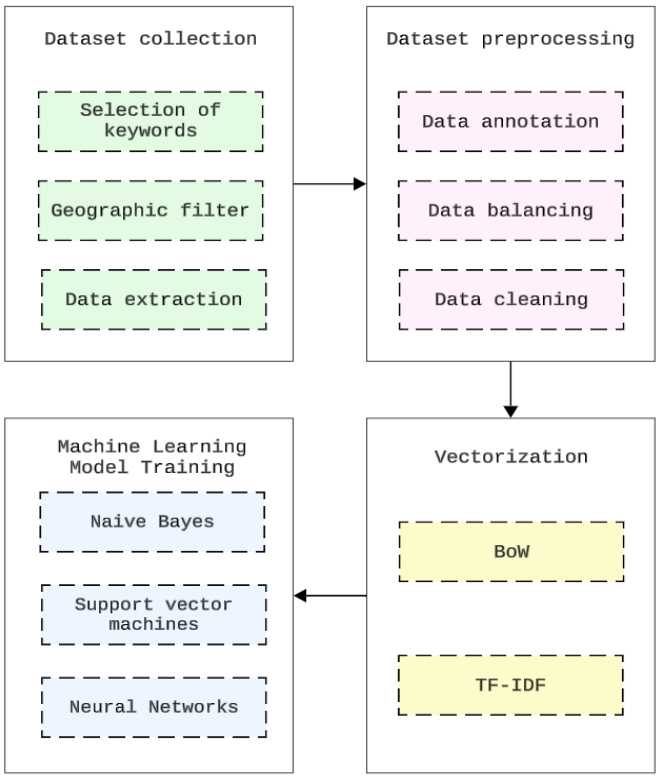
Methodology

The methodology employed in this study to develop models capable of detecting cyberbullying on social network X comprises several phases, as described below and illustrated in Figure 1. The process begins with the collection of the dataset, which involved selecting keywords used to search for relevant tweets. This selection was crucial as it allowed the incorporation of terms representative of the language commonly employed in cases of cyberbullying. An intersectional approach was adopted, considering categories such as race and gender. Additionally, the keywords were ensured to be in Spanish and reflective of everyday expressions used in Colombia. To ensure that the tweets originated from the Colombian population, a geographic filter was applied during data collection. In the dataset processing stage, the data were cleaned to remove noise and irrelevant content. Each tweet was manually annotated with the labels "Cyberbullying" and "Non-cyberbullying" by a panel of experts in gender issues, ensuring accurate classification. Subsequently, the dataset was balanced by equalizing the two classes to prevent bias in the classification models.

In the data vectorization phase, two techniques were employed: Bag-of-Words (BoW) and Term frequency – Inverse document frequency (TF-IDF). These strategies enabled the transformation of tweets into numerical feature vectors, which served as input for the classification models. Finally, in the machine learning algorithm application phase, three approaches were tested: Naive Bayes, Support Vector Machines, and Neural Networks. The results obtained from each model were analyzed to identify their advantages and limitations in detecting cyberbullying, providing a solid foundation for evaluating their effectiveness in this context. The following sections detail each task that was part of the methodology.

The experiments were conducted using Python version 3.12.0 in the Visual Studio Code IDE. The following libraries and their versions were used: NumPy 1.26.0 for numerical operations, Pandas 2.1.0 for data manipulation and analysis, and scikit-learn 1.3.0 for machine learning tasks, including data preprocessing, feature extraction, model training, and evaluation.

Figure 1. Methodology for cyberbullying detection.



Source: own elaboration.

Dataset collection

Five lists of derogatory terms were compiled through a review of previous studies on harassment, focusing on terms used to demean individuals based on their gender or race [31-34]. These lists were organized into two gender-based harassment categories (Women and Men) and one racism category (Race), reflecting the diversity of experiences and expressions of discrimination. A total of 71 keywords or phrases were identified for the Women category, 137 for Men, and 31 for Race.

For the Race category, three prior studies were consulted on the language used to racialize or exclude individuals. Three specific lists were compiled: Afro, Rural, and Indigenous. In total, 64 keywords or phrases were gathered, of which only 31 yielded results in the advanced search on the social network X. Table 1 provides an example of the keywords or phrases included in each of the three lists

Table 1. Keywords from the lists used in the race category

List	Keywords
Afro	pero sos muy negrito
	hay que mejorar la raza
	pelo malo
	hay que aclarar la raza
	hay que casarse con blanco para mejorar la raza
	negros brutos
	negro tenía que ser
	negro y feo
Rural	mucho montañero
	lo bajaron del monte con espejo
	eso es duro pa'l campesino

Indigenous	mucho indio
	patirrajao
	indito

Source: own elaboration.

For the gender category, two lists were created: Women and Men. These lists were based on a previous study examining language used to refer to women and men in sexual and cultural contexts [31]. A total of 566 keywords or phrases were collected, of which only 208 produced results in the advanced search on social network X. Table 2 provides an example of these lists.

Table 2. Keywords from the lists used in the gender category.

List	Keywords
Women	achapada
	alebrestada
	amachada
	amangualada
	arepera
	arequipito
Men	Abusador
	achapado
	Adúltero
	Afeminado
	Alebrestado
	analfabeto
	antipático

Source: own elaboration.

After compiling the keyword lists, a manual search for tweets was conducted on the social network X using its advanced search functionality. Geographic proximity filters were applied to ensure that the tweets were specific to the Colombian context, and language filters were used to restrict the results to Spanish. Upon completing the collection process, a total of 4,006 tweets were gathered.

Dataset Preprocessing

The dataset was generated through a rigorous process to ensure data quality and reliability. Each of the 4,006 collected tweets was assigned a specific category ("race," "men," or "women") based on the abusive keywords or phrases used, along with a unique numerical ID. Subsequently, a manual annotation process was conducted to classify the tweets according to whether they contained cyberbullying. This task was performed by a panel of 25 female experts, professionals from various fields actively working in women's organizations, with deep knowledge of gender-based violence, particularly in educational settings. The experts are members of La Red Universitaria para la Reducción de las Brechas de Género (University Network for the Reduction of Gender Gaps), a collaborative initiative among higher education institutions in Colombia aimed at promoting equality in research, teaching, and management through the exchange of best practices, policy development, creation of safe environments, and the promotion of female leadership.

The experts were asked to indicate whether each tweet constituted cyberbullying or not. In their evaluations, they considered factors such as the perceived intent to harm or distress an individual or group, and the severity of the language used, ranging from mild insults to serious threats or hate speech. Each expert applied her professional judgment and experience to classify the tweets. This expert-driven, subjective approach enabled a nuanced and reliable evaluation,

resulting in a final classification of 2,771 tweets as cyberbullying and 1,235 as non-cyberbullying. The classification was consolidated into a CSV file detailing the category and classification of each tweet, as illustrated in Table 3.

Table 3. Total number of labeled tweets by category.

Category	Total number of labeled tweets	Number of tweets labeled as cyberbullying	Number of tweets labeled as non-cyberbullying
Women	1382	1150	232
Men	2087	1155	932
Race	537	466	71
Total	4006	2771	1235

Source: own elaboration.

Table 4 presents examples of tweets labeled in both classifications, illustrating how context and language usage influence the final categorization. In the case of the "woman" category, it highlights how a characteristic word from the Colombian population, included in the keyword list for this category, can be used both in an offensive context that constitutes cyberbullying and in another completely different context that does not represent it.

Table 4. Examples of tweets labeled as cyberbullying and non-cyberbullying.

Category	Word/Phrase	Cyberbullying	Non-cyberbullying
Women	Achapada	"Mal parida bocona care simio achapada esa 👊😉✌"	"Ni muy santa, ni muy diabla. Ni tan avispada, ni tan achapada. Ni muy muy, ni tan tan👊👏✌👌😉"
Men	Drogadicto	"@lafm Jajajaja este drogadicto mitómano se cree su mentiras, sapo hijueputa!"	"@ElPensador75 Si yo fuera drogadicto, homosexual y asesino siiiii"
Race	Trabajar como negro para vivir como blanco	"@DrLuisCufre Carnal, hay q trabajar como negro para vivir como blanco"	"Que feo que es ese término "trabajar como negro para vivir como blanco" los que dicen eso ¿todo bien en casa?"

Source: own elaboration.

After the final classification, it was observed that the dataset was composed of 69.22% tweets labeled as cyberbullying. This class imbalance could lead the machine learning model to favor the majority class, negatively affecting its accuracy in correctly identifying the minority class, thereby compromising its generalization ability. To address this issue, an additional 1536 tweets representing everyday conversations, which did not contain cyberbullying, were selected. These tweets were carefully chosen to reflect a variety of common topics on the platform, ensuring they were free of offensive or harmful language. The integration of these tweets allowed for balancing the proportions between the classes within the dataset.

As a result, the final dataset was evenly balanced, consisting of a total of 5542 tweets, equally distributed between 2771 cyberbullying tweets and 2771 non-cyberbullying tweets. This balancing process was crucial for improving the performance of the machine learning model. By ensuring an equitable distribution between the classes, the model could more effectively learn the distinctive features of each category, thereby increasing its accuracy and effectiveness in detecting cyberbullying. A balanced dataset reduces the risk of bias toward the majority class and enhances the model's ability to generalize to new data, ensuring accurate detection of both cyberbullying and non-cyberbullying tweets.

The cleaning of the 5542 tweets involved several key stages to ensure data quality and consistency: the data was imported from a CSV file, and usernames,

links, emojis, hashtags, and punctuation were removed to prevent the influence of non-textual or irrelevant elements in the analysis. The text was then tokenized, unnecessary common words were discarded, and lemmatization was applied to unify similar terms. This process allowed for the cleaning and simplification of the tweets, leaving only meaningful words for subsequent analysis and modeling. Finally, the preprocessed text was stored in the same CSV file for analysis. Table 5 shows an example for each category of how an unprocessed message looks versus the processed message.

Table 5. Examples of tweets cleaning.

Category	Unprocessed tweet	Preprocessed tweet
Women	"@JuanJo_12x @juan_alesander @caritodtorre @Veronicalcocerg Y a mi que me importa donde viva el susodicho, simplemente esta poniendo una queja y esa aparece en la imagen, ahí se ve una brincona alebrestada haciendo el ridículo, donde nadie la a llamado y si yo soy un 🤡 usted es un 🐽 de 💩 qué da 🤮"	me importa viva susodicho simplemente poniendo queja aparece imagen ahí ve brincona alebrestada haciendo ridículo nadie llamado si usted da
Men	"El dueño de este canal: el hambriento ángulo financió partidos políticos en las elecciones pasadas para que tumbaran la Reforma Laboral, ahora tampoco deja ver los partidos de fútbol sin pagar socio. Avaro a morir!"	dueño canal hambriento ángulo financió partidos políticos elecciones pasadas tumbaran reforma laboral ahora tampoco deja ver partidos fútbol pagar socio avaro morir
Race	"Queman a la persona que más alegrías les dió desde hace 2 décadas a todos esos negros brutos. No les quedan neuronas disponibles. Son unos desagradecidos."	queman persona alegrías dió hace décadas toda negros brutos quedan neuronas disponibles desagradecidos

Source: own elaboration.

Vectorization

Two text vectorization techniques, CountVectorizer and TfidfVectorizer, were employed using the sklearn.feature_extraction.text library to convert the tweets into numerical representations that the machine learning models could process. With a configuration of min_df=2 to include only terms present in at least two documents and a ngram_range= (1,4) to capture unigrams, bigrams, trigrams, and four grams, these techniques transformed the raw text into a structured matrix. CountVectorizer creates a term count matrix (Bag of Words), while TfidfVectorizer weighs the terms based on their frequency in the document and the entire corpus, providing a refined representation of the terms' relevance.

Both techniques generated a vocabulary of 7507 unique terms, ensuring that the vectorized representations captured the same amount of information. This consistent approach across the two techniques provides a solid foundation for training and evaluating machine learning models, maximizing the understanding of the patterns present in the tweets. Table 6 presents an example of one of the tweets represented with BoW and TF-IDF.

Table 6. Representation with BoW and TF-IDF.

Unprocessed tweet	Preprocessed tweet	BoW (count)	TF-IDF (weight)
@HoyPalestina El idiota de Petrocacas, defensor a ultranza de	Idiota petrocacas defensor ultranza	idiota: 1 defensor: 1 palestina :1	idiota: 0.3174 defensor: 0.4197 palestina :0.4197

Palestina, ni siquiera sabe que la bandera está al revés.	palestina siquiera sabe bandera revés	siquiera: 1 sabe: 1 bandera: 1 revés: 1	siquiera: 0.3597 sabe: 0.2765 bandera: 0.4354 revés: 0.3891

Source: own elaboration.

Machine learning model training

After preprocessing the data, it was split into training and test sets, with 80% of the data assigned to training and the remaining 20% to testing. The dataset was divided randomly, maintaining the aforementioned proportion between the training and test sets. This random division helped ensure that the datasets were representative and that the model was not biased toward a particular set of data.

For the prediction model using Naive Bayes, the GaussianNB function from the sklearn library was employed. To optimize its performance, GridSearchCV was used to conduct an exhaustive search for hyperparameters, focusing on variance smoothing (var_smoothing) with a range of values from 1e-12 to 35, with increments of 1.3. Specifically, the goal was to maximize the area under the ROC curve, and five-fold cross-validation was used. A total of 27 different configurations were tested for each vectorizer (CountVectorizer and TfidfVectorizer), allowing the model to be fine-tuned and predictions to be made on the test set. Key metrics such as accuracy, precision, recall, F1 score, specificity, and the area under the ROC curve were evaluated.

For the prediction model using support vector machines, the SVC function from the sklearn library was employed. An exhaustive search for hyperparameters was conducted through GridSearchCV, exploring configurations for the kernel (linear, rbf, poly, sigmoid), gamma (scale, auto), and the regularization parameter C (values between 0.1 and 1.0 with increments of 0.1). A total of 80 hyperparameter combinations were tested.

Finally, for the prediction model using neural networks, the MLPClassifier function from sklearn was employed. The hyperparameter search included 900 combinations, exploring options for the hyperparameters activation (identity, logistic, tanh, relu), solver (lbfgs, sgd, adam), alpha (values between 0.1 and 1.0 in increments of 0.2), and hidden_layer_sizes (ranging from one to three layers, and from one to five neurons per layer).

Results and discussion

Metrics

Evaluating the performance of predictive models is essential to determine their accuracy and consistency. The evaluation metrics, derived from the confusion matrix, allow quantification of the models' ability to make correct predictions. For cyberbullying detection, it is crucial to correctly distinguish between bullying cases (true positives) and non-bullying cases (true negatives), as well as minimize classification errors (false positives and false negatives). The metrics used were accuracy, precision, sensitivity, specificity, and F1-score [35]. Additionally, the area under the ROC curve (AUC) was used to measure the model's ability to discriminate between classes. A higher AUC indicates better performance, with 1.0 representing a perfect model and 0.5 equivalent to random performance.

The evaluation of machine learning models involved an exhaustive exploration of multiple hyperparameter combinations and the subsequent calculation of various key metrics. The models were ranked based on the area under the ROC curve. This metric was chosen due to its ability to provide a comprehensive assessment of the models' performance, considering both sensitivity and specificity, which is particularly relevant in problems where balancing the detection of positive and negative cases is critical.

Results obtained using Naive Bayes

Table 7 presents the five best configurations selected from the 27 evaluated for the Naive Bayes technique using the BoW vectorization strategy, as well as the five best configurations among the 27 evaluated with the TF-IDF strategy. The results are organized according to the area under the ROC curve, highlighting the configurations with the best performance in each case.

In the case of vectorization with BoW, the model achieves an accuracy of 75.8%, with precision ranging from 67.2% to 68.3%. The sensitivity (recall) is notably high, reaching 96.0%; however, the specificity is relatively low, ranging from 55.9% to 61.3%. This indicates that the model tends to overestimate positive instances, which affects the correct identification of true negatives. On the other hand, when using TF-IDF vectorization, higher accuracy is achieved, ranging from 78.8% to 81%, along with precision also greater than that obtained with BoW, ranging from 72.2% to 78.2%. Furthermore, specificity shows greater consistency, reaching values between 64% and 76.1%, reflecting a better ability to identify both true positives and true negatives. The area under the ROC curve reaches 89.7%, highlighting superior performance in differentiating between classes.

Table 7. Configurations with the highest AUC using Naive Bayes.

Vectorization	var_smoothing	Accuracy	AUC	Precision	Sensitivity	F1-score	Specificity
BoW	0.01	0.758	0.769	0.683	0.960	0.798	0.556
	0.001	0.755	0.756	0.699	0.897	0.785	0.613
	0.0001	0.741	0.742	0.683	0.900	0.777	0.583
	0.00001	0.734	0.734	0.675	0.904	0.773	0.565
	0.000001	0.732	0.732	0.672	0.906	0.772	0.559
TF-IDF	0.8	0.788	0.897	0.722	0.936	0.816	0.640
	1.0	0.788	0.897	0.722	0.936	0.816	0.640
	2.0	0.789	0.895	0.727	0.927	0.815	0.651
	5.0	0.807	0.894	0.751	0.920	0.827	0.694
	15.0	0.810	0.894	0.782	0.859	0.819	0.761

Source: own elaboration.

Results obtained using Support Vector Machines

Table 8 presents the five best configurations selected from the 80 evaluated for the support vector machine technique using the BoW vectorization strategy, as well as the five best configurations among the 80 evaluated with the TF-IDF strategy. As previously indicated, the results are organized according to the area under the ROC curve, highlighting the configurations with the best performance in each case.

According to the results, when using the Bag of Words technique, accuracy ranges from 81.5% to 82.8%, indicating a good overall performance. Precision ranges from 79.5% to 80.9%, while sensitivity fluctuates between 82.6% and 88%, demonstrating the model's ability to correctly identify true positives. Specificity ranges from 77.6% to 80.5%, reflecting a good ability to predict true negatives. On the other hand, when using TF-IDF vectorization, accuracy ranges from 81.7% to 82.4%, with these values being nearly identical to those obtained with BoW. Precision showed little variation, remaining at 80.6% in most cases. Sensitivity reaches a maximum of 85%. Meanwhile, specificity, similar to BoW, falls within the range of 79.6% to 79.9%. These results demonstrate that the SVM models perform robustly with both BoW and TF-IDF in the cyberbullying detection task.

Table 8. Configurations with the highest AUC using Support Vector Machines.

Vectorization	kernel	gamma	C	Accuracy	AUC	Precision	Sensitivity
BoW	linear	scale	0.1	0.823	0.905	0.800	0.861
	linear	auto	0.1	0.823	0.905	0.800	0.861
	rbf	scale	1.0	0.815	0.903	0.809	0.826

	sigmoid	scale	0.8	0.828	0.902	0.797	0.880
	sigmoid	scale	0.9	0.822	0.902	0.795	0.868
TF-IDF	sigmoid	scale	0.4	0.817	0.909	0.806	0.835
	linear	scale	0.4	0.820	0.908	0.806	0.843
	linear	auto	0.4	0.820	0.908	0.806	0.843
	sigmoid	scale	0.5	0.822	0.908	0.806	0.848
	linear	auto	0.5	0.824	0.908	0.809	0.850

Source: own elaboration.

Results obtained using Neural Networks

Table 9 presents the five best configurations selected from the 900 evaluated for the support vector machine technique using the BoW vectorization strategy, as well as the five best configurations among the 900 evaluated with the TF-IDF strategy. The results are organized according to the area under the ROC curve.

When using the Bag of Words technique, accuracy ranges from 81.5% to 82.3%, suggesting a stable ability to correctly predict the class to which the tweets belong. The area under the ROC curve, which measures the model's ability to distinguish between classes, remained stable at 90.3%, indicating good discriminative power. Precision fluctuated between 80.5% and 81.1%, reflecting that a high proportion of positive predictions were correct. Sensitivity ranged from 82.3% to 84.6%, demonstrating the effectiveness of the models in capturing cyberbullying tweets. Finally, specificity, which measures the model's ability to correctly identify non-cyberbullying tweets, showed little variation, with a value of 80%. Taken together, the results suggest that the neural network models with BoW exhibit solid and stable performance in correctly classifying tweets.

Table 9. Configurations with the highest AUC using Neural Networks.

Vectorization	activation	solver	alpha	hidden_layer_sizes	Accuracy	AUC	Precision
BoW	relu	sgd	0.9	(1,)	0.821	0.903	0.811
	relu	sgd	0.7	(1,)	0.816	0.903	0.808
	relu	sgd	0.5	(1,)	0.815	0.903	0.809
	relu	sgd	0.7	(2,)	0.818	0.903	0.805
	relu	sgd	0.9	(2,)	0.823	0.903	0.808
TF-IDF	relu	adam	0.7	(2,)	0.817	0.910	0.806
	relu	adam	0.7	(3,)	0.817	0.909	0.804
	tanh	sgd	0.3	(3,)	0.816	0.909	0.803
	identity	adam	0.7	(3,)	0.816	0.909	0.804
	tanh	adam	0.7	(3,)	0.816	0.909	0.804

Source: own elaboration.

In contrast, when using TF-IDF vectorization, a stable accuracy of 81.7% is achieved. The area under the ROC curve was slightly higher than when using BoW, at 91%. Precision showed no significant variations, reaching 80.6%, while sensitivity reached 83.9%.

When comparing the results of neural networks with BoW and TF-IDF, both approaches demonstrate consistent and reliable performance for classifying both cyberbullying and non-cyberbullying tweets. Although the TF-IDF models have a slight advantage in the area under the ROC curve, the BoW models present comparable metrics. This suggests that both BoW and TF-IDF are suitable options for cyberbullying detection, with solid performance in the key classification metrics.

Comparison of results

Table 10 presents a summary of the best configuration for each of the machine learning techniques employed in this research to address the problem of cyberbullying detection on social network X. It is observed that neural networks

achieved the best performance, reaching an area under the ROC curve of 91%. The table also includes the optimal values of the hyperparameters identified during this work. These results are especially relevant, as they provide a starting point for future hyperparameter tuning, based on a rigorous experimentation process. Additionally, it is noteworthy that the TF-IDF vectorization technique is present in the most effective configurations across all three evaluated techniques, highlighting its effectiveness in transforming textual data into useful numerical representations for the models

Table 10. Best parameters found by technique.

Technique	Best parameters found	AUC
Naive Bayes	var_smoothing=0.8 vectorization technique=TF-IDF	0.897
SVM	kernel=sigmoid gamma=scale C=0.4 vectorization technique=TF-IDF	0.909
Neural Networks	activation=ReLU solver=Adam alpha=0.7 topology= (2,) vectorization technique=TF-IDF	0.910

Source: own elaboration.

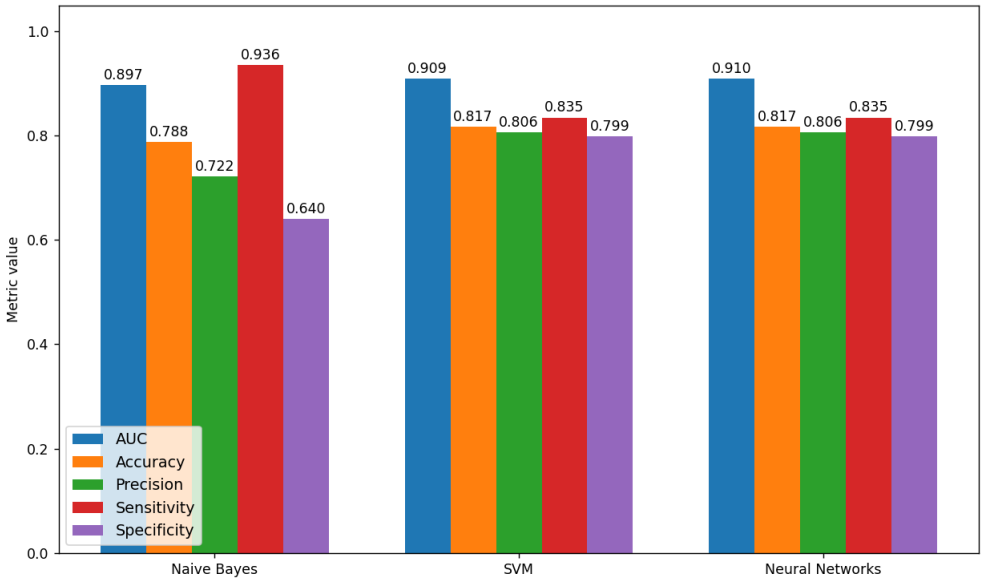
To further illustrate the performance differences among the models, Figure 2 presents a grouped bar chart comparing the key evaluation metrics—AUC, accuracy, precision, sensitivity, and specificity—across the three machine learning techniques: Naive Bayes, Support Vector Machines, and Neural Networks. The figure clearly shows that the Neural Network model consistently achieves the highest AUC, indicating its superior ability to discriminate between cyberbullying and non-cyberbullying tweets. While accuracy and precision are comparable across models, variations in sensitivity and specificity reveal differences in how effectively each model identifies positive and negative cases. This visualization provides a concise and intuitive summary of model performance, complementing the information presented in Table 10.

Upon evaluating the models and configurations tested, it was determined that the neural network model with a ReLU activation function, the Adam solver, an alpha value of 0.7, a hidden layer with 2 neurons, and the TF-IDF vectorization technique achieved the best performance, attaining the highest area under the ROC curve. This model excelled in multiple evaluation metrics, achieving an accuracy of 81.7%, a sensitivity of 83.5%, and a specificity of 79.9%. These metrics reflect its outstanding ability to effectively classify tweets as cyberbullying or not, positioning it as the most robust and reliable option for this task.

The proposed model can be used as an automated tool for the early detection of cyberbullying on social networks, particularly in Spanish-speaking contexts such as Colombia, addressing the limitations highlighted in previous studies regarding the lack of systems adapted to languages other than English. In a practical implementation, social media platforms could integrate this model into their moderation systems, enabling the identification of messages with potential cyberbullying content based on keywords, regional expressions, and linguistic patterns specific to Colombian Spanish. This would facilitate the automatic classification of offensive messages and assist human moderators in prioritizing the review of posts that pose a higher risk, optimizing resources to combat online harassment. Additionally, the model could be used in digital awareness campaigns, highlighting how certain linguistic expressions commonly used in the digital environment may contribute to perpetuating gender, racial, or sexual orientation-

based violence, as described in the introduction. In this way, not only would cyberbullying be identified, but awareness would also be raised and responsible language use in virtual environments promoted

Figure 2. Comparison of evaluation metrics across machine learning models for cyberbullying detection.



Source: own elaboration.

However, it is important to mention the model's limitations. Firstly, although a significant effort was made to construct a balanced and representative dataset for the Colombian context, it is possible that some regional expressions or linguistic variations were not included. This could limit the model's ability to identify cyberbullying cases that use language different from that included in the training data. On the other hand, the vectorization techniques employed, such as Bag-of-Words and TF-IDF, while useful for transforming text into numerical representations, have limitations in capturing deep semantic relationships in language. These techniques fail to understand the context or ambiguity of certain terms, which could affect the model's performance in more complex situations. Furthermore, although the evaluation metrics show a high level of sensitivity, specificity is relatively low. This implies that, while the model is effective in identifying cyberbullying cases, it struggles to accurately distinguish messages that do not contain offensive content, increasing the likelihood of false positives. This imbalance in the key metrics may limit its applicability in scenarios where avoiding incorrect classifications is essential.

The results of this research show competitive performance in cyberbullying detection compared to previous works that have addressed the same problem. For example, the study by López-Martínez [29] proposed an approach based on natural language processing and machine learning to detect cyberbullying on Twitter, achieving an accuracy close to 85% in a general Spanish language context. While these results are significant, the approach does not address specific cultural and linguistic peculiarities, such as Colombian Spanish. In contrast, the model developed in this work incorporates local expressions and contexts, an advantage that improves the relevance and regional applicability of the results.

Moreover, research such as that by Islam et al. [22] highlights the use of features such as Bag-of-Words (BoW) and TF-IDF for detecting abusive messages on social media in English, with metrics ranging from 80% to 86% accuracy, depending on the model used. Although these vectorization techniques were also employed here, our research achieved an area under the ROC curve (AUC) of 91% using neural networks with TF-IDF, surpassing the discriminative capability

observed in these previous studies. However, a shared limitation is that the vectorization techniques used, although effective, do not capture the deep semantic relationships in the text. This aspect could be overcome by using pretrained models, such as those proposed by Müller et al. [26] with Covid-twitter-BERT, which demonstrate a greater ability to interpret the context and subtleties of language.

A key differentiating aspect of this work is the dataset employed. While León-Paredes et al. [30] worked with pre-labeled datasets without an intersectional approach, this work made a significant effort to construct its own dataset, carefully balanced and adapted to Colombian Spanish. This process, although resource- and time-intensive, aims to make the model more relevant to a specific cultural context. This specificity represents an important advantage, although it also limits the generalization of the model to other contexts and dialects.

Conclusions

This paper evaluated various machine learning techniques, including Naive Bayes, Support Vector Machines (SVM), and Neural Networks, for detecting cyberbullying on social networks, with a focus on Spanish and the Colombian context. Our findings highlight the effectiveness of neural network-based models combined with vectorization techniques such as TF-IDF for representing text data. The study also identified limitations regarding generalization to new linguistic and temporal contexts, as well as the semantic constraints of traditional vectorization methods, suggesting the need for more advanced approaches using pre-trained language models adapted to Spanish.

Several avenues for future research emerge from this work, including the continuous expansion and updating of datasets to capture emerging expressions and dialects, the use of contextual embeddings from Transformer-based models, the exploration of deeper neural network architectures, and strategies to optimize model performance while minimizing errors. Extending this approach to other social platforms and languages could further assess its generalization capability and practical impact.

Overall, this work establishes an effective and replicable approach for detecting cyberbullying in Spanish, advancing the field by integrating machine learning and natural language processing within culturally and linguistically specific contexts. By providing a robust methodological foundation and demonstrating practical applicability, this study not only addresses online harassment in the Colombian and Spanish-speaking context but also offers a model that can be adapted and extended to other platforms, dialects, and multicultural settings.

References

[1] K. Wang, Y. Chen, J. Zhang, and B. A. Oudekerk, "Indicators of School Crime and Safety: 2019. NCES 2020-063/NCJ 254485," National Center for Education Statistics, 2020.

[2] UNICEF, "Violencia contra niños, niñas y adolescentes en América Latina y el Caribe 2015-2021," 2021. [Online]. Available: https://www.unicef.org/lac/media/29031/file/Violencia-contra-ninos-ninas-y-adolescentes-en-America-Latina-y-el-Caribe-2015-2021.pdf

[3] GLEN, "'Being LGBT in School'A Resource for Post-primary Schools to Prevent Homophobic and Transphobic Bullying and Support LGBT Students," ed: GLEN and The Department of Education and Science Dublin, OH, 2016.

[4] Comisión Interamericana de Derechos Humanos, "Violencia contra personas lesbianas, gays, bisexuales, trans e intersex en América," 2015.

[5] J. J. Jaramillo and J. E. R. Pineda, "Padres y madres homosexuales y bisexuales en Colombia. Experiencias de discriminación y estrategias de afrontamiento," RES. Revista Española de Sociología, vol. 28, no. 1, pp. 95-112, 2019. doi: http://dx.doi.org/10.22325/fes/res.2018.62.

[6] Asamblea Mundial de la Salud, 69. (2016). Plan de acción mundial de la OMS para fortalecer la función del sistema de salud en el marco de una respuesta nacional multisectorial para abordar la violencia interpersonal, en particular contra las mujeres y las niñas, y contra los niños en general. Organización Mundial de la Salud. https://iris.who.int/handle/10665/253191.

[7] N. CEPAL, "Enfrentar la violencia contra las mujeres y las niñas durante y después de la pandemia de COVID-19 requiere FINANCIAMIENTO, RESPUESTA, PREVENCIÓN Y RECOPILACIÓN DE DATOS," 2020.

[8] Instituto Nacional de Medicina Legal y Ciencias Forenses, "Forensis, Datos para la vida. Número 1. ISSN 2145-0250," Bogotá, D.C. República de Colombia, vol. 23, 2021.

[9] N. Gherardi, "Otras formas de violencia contra las mujeres que reconocer, nombrar y visibilizar," 2016.

[10] K. Barker and O. Jurasz, "Online violence against women as an obstacle to gender equality: A critical view from Europe," European Equality Law Review, vol. 2020, no. 1, pp. 47-60, 2020.

[11] N. Giant, Ciberseguridad para la i-generación: Usos y riesgos de las redes sociales y sus aplicaciones. Narcea Ediciones, 2016.

[12] S. L. Price, "Differences in School Discipline Efforts and Cyberbullying by School Level: A National Analysis," Sam Houston State University, 2021.

[13] J. van Tiel, "Cyberbullying, an overlooked and ever growing danger to the development of children," Technical report, KidsRights, 2020.

[14] ONU Mujeres, "Frequently Asked Questions: Types of Violence against Women and Girls," 2021.

[15] U. Interparlamentaria, "Sexismo, acoso y violencia contra las mujeres parlamentarias," Issues Brief octubre, 2016.

[16] G. Mura and D. Diamantini, "Cyberbullying among Colombian students: An exploratory investigation," European Journal of investigation in health, Psychology and Education, vol. 3, no. 3, pp. 249-256, 2013. https://doi.org/10.3390/ejihpe3030022.

[17] ONG Internacional Bullying Sin Fronteras para América, Asia, Oceanía y África, "Estadísticas Mundiales de Bullying 2023. Colombia Noveno lugar. 41.500 casos." [Online]. Available: https://bullyingsinfronteras.blogspot.com/2018/11/estadisticas-de-bullying-en-colombia.html

[18] L. M. Munar, A. A. R. Diaz, and R. d. J. R. García, "Violencia de género en línea en Colombia: desafíos y acciones para un entorno digital seguro," Informática y Derecho. Revista Iberoamericana de Derecho Informático (2. ª época), no. 14, pp. 59-71, 2023.

[19] Colombia Diversa, "La violencia no nos impide ser y amar," Informe: Situación de Derechos Humanos de personas LGBT, 2021.

[20]Comisión Interamericana de Derechos Humanos, "Derechos económicos, sociales, culturales y ambientales de personas afrodescendientes," 2021. [Online]. Available: http://www.oas.org/es/cidh/informes/pdfs/DESCA-Afro-es.pdf

[21] J. Martínez-Santiago, I. Zych, and A. J. Rodríguez-Hidalgo, "Bullying personal y étnico-cultural en la Amazonía peruana: prevalencia, solapamiento y predictores," Revista de Psicodidáctica, vol. 28, no. 2, pp. 153-163, 2023. https://doi.org/10.1016/j.psicod.2023.05.003.

[22] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, "Cyberbullying detection on social networks using machine learning approaches," in 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6. doi:10.1109/CSDE50874.2020.9411601

[23]J. Lu et al., "Hate speech detection via dual contrastive learning," IEEE/ACM Trans. Audio Speech Lang. Process., 2023. https://doi.org/10.48550/arXiv.2307.05578.

[24]T. Kanan et al., "An Intelligent Health Care System for Detecting Drug Abuse in Social Media Platforms Based on Low Resource Language," IEEE/ACM Trans. Audio Speech Lang. Process., 2023. https://doi.org/10.1109/TASLP.2023.3294699.

[25] D. Maynard and A. Funk, "Automatic detection of political opinions in tweets," in The Semantic Web: ESWC 2011 Workshops: ESWC 2011 Workshops, Heraklion, Greece, May 29-30, 2011, Revised Selected Papers 8, 2012: Springer, pp. 88-99. https://doi.org/10.1007/978-3-642-25953-1_8.

[26]M. Müller, M. Salathé, and P. E. Kummervold, "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter," Frontiers in artificial intelligence, vol. 6, p. 1023281, 2023. https://doi.org/10.48550/arXiv.2005.07503.

[27] Ç. Acı, E. Çürük, and E. S. Eşsiz, "Automatic detection of cyberbullying in formspring. me, myspace and Youtube social networks," Turkish

Journal of Engineering, vol. 3, no. 4, pp. 168-178, 2019. https://doi.org/10.31127/tuje.554417.

[28]H. Kwak, J. Blackburn, and S. Han, "Exploring cyberbullying and other toxic behavior in team competition online games," in Proceedings of the 33rd annual ACM conference on human factors in computing systems, 2015, pp. 3739-3748. https://doi.org/10.1145/2702123.2702529.

[29]A. López-Martínez, J. A. García-Díaz, R. Valencia-García, and A. Ruiz-Martínez, "CyberDect. A novel approach for cyberbullying detection on twitter," in Technologies and Innovation: 5th International Conference, CITI 2019, Guayaquil, Ecuador, December 2–5, 2019, Proceedings 5, 2019: Springer, pp. 109-121. https://doi.org/10.1007/978-3-030-34989-9_9.

[30]G. A. León-Paredes et al., "Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the Spanish language," in 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 2019: IEEE, pp. 1-7. https://doi.org/10.1109/CHILECON47746.2019.8987684.

[31]M. Rosero and L. Saavedra, "Estudio comparativo de la creación léxica en estudiantes de la Universidad del Valle en Santiago de Cali y Cartago para referirse a las mujeres en el ámbito sexual y cultural". Tesis. Licenciatura en Lenguas Extranjeras. Escuela de Ciencias del Lenguaje. Universidad del Valle. 2014.

[32]M. Moreno Mosquera, S. Reyes Mosquera, and L. Rivas Portocarrero, "Internalización de la estigmatización y la discriminación racial entre estudiantes de raza negra de la jornada diurna en la Universidad del Valle Sede Pacifico ¿endorracismo?". Tesis. Escuela de Trabajo Social y Desarrollo Humano. Facultad de Humanidades. Universidad del Valle. 2020.

[33]M. D. Castellano Ascencio, "Análisis pragmático de la función de los tratamientos nominales en actos de habla descorteses en Medellín (Colombia)," Forma y función, vol. 30, no. 2, pp. 139-162, 2017. https://doi.org/10.15446/fyf.v30n2.65794.

[34]O. J. Cuesta-Moreno and A. Gómez-Melo, "Frases que racializan, excluyen y minimizan al sujeto en el lenguaje cotidiano de un grupo de jóvenes de Bogotá," PROSPECTIVA. Revista de Trabajo Social e intervención social, pp. 143-166, 2014. https://doi.org/10.25100/prts.v0i19.970.

[35]Powers, D. M. W. "Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness & correlation." Journal of Machine Learning Technologies, 2011, vol 2, 37–63. https://doi.org/10.48550/arXiv.2010.16061.