Revista EIA

UNIVERSIDAD EIA

Cyberbullying Detection on Twitter for the Colombian Population Using Artificial Intelligence Techniques

Felipe Mauricio Guerra Saenz[1]

Oscar Fernando Bedoya Leiva[1]

Marcela Holguín Mera[2]

1. Universidad del Valle, Colombia.
2. Universidad de San Buenaventura, Colombia.

Abstract

Cyberbullying is an increasingly relevant issue in contemporary society that can have serious consequences on the emotional and psychological well-being of victims. Currently, due to the vast exchange of digital interactions, it is challenging and labor-intensive for online platform moderators to manually detect and remove all cyberbullying comments. Therefore, there is a need for automatic models that employ artificial intelligence techniques to detect cyberbullying. This article proposes machine learning-based models and language-based models for cyberbullying detection on the social network Twitter. The machine learning techniques used are XGBoost, logistic regression, and random forests. On the language model side, a fine-tuning process was applied to a masked language model based on a transformer named roberta-base-bne. Although there are currently various models for this purpose, most of them are developed using English. In the case of Spanish, there are very few studies, and in the particular case of Colombian Spanish, there is no precedent for contributions in this area.

Additionally, this article introduces a corpus comprising tweets written in Colombian Spanish, meticulously annotated by a qualified occupational therapist. Two distinct datasets stem from this corpus. Dataset 1 is characterized by the

Revista EIA
ISSN 1794-1237
e-ISSN 2463-0950
Año XIX/ Volumen 22/ Edición N.44
Julio - diciembre 2025
Reia4402 pp. 1-37

Publicación científica semestral
Universidad EIA, Envigado, Colombia

Para citar este artículo / To reference this article /
Guerra Saenz, F. M.; Bedoya Leiva, O. F. y Holguín Mera, M.
Cyberbullying Detection on Twitter for the Colombian Population Using Artificial Intelligence Techniques

Revista EIA, 22(44), Reia4402 pp. 1-27
https://doi.org/10.24050/reia.v22i43.1757

Autor de correspondencia: Correa Bedoya Leiva, O. F.
Escuela de Ingeniería de Sistemas y Computación
Correo electrónico:
oscar.bedoya@correounivalle.edu.co

Recibido: 12-15-2023
Aceptado:10-06-2025
Disponible online: 01-07-2025

annotation of a tweet as cyberbullying and another as non-cyberbullying, both containing the same word, carefully considering the context in which each word is employed. In contrast, Dataset 2 features different words in cyberbullying and non-cyberbullying tweets. The rationale behind utilizing these two datasets lies in capturing diverse language expressions and their contextual nuances, assessing the effectiveness of the applied techniques in discerning such context. Results from dataset 1 reveal that models achieve an area under the ROC curve of 0.797, 0.796, 0.785, and 0.910 with logistic regression, random forests, XGBoost, and roberta-base-bne, respectively. Meanwhile, employing Dataset 2 yields area under the ROC curve values of 0.983, 0.978, 0.971, and 0.996, respectively. Finally, we introduce a web application named AI Cyberbullying Detector tailored for therapists, empowering them to leverage artificial intelligence in cyberbullying-related studies.

Keyworks: Artificial intelligence; Cyberbullying; Colombian Spanish; Language models; Logistic regression; Machine learning; Random forests; Transformers; Twitter; XGBoost

Detección de Ciberacoso en Twitter para la Población Colombiana Usando Técnicas de Inteligencia Artificial

Resumen

El ciberacoso es un problema cada vez más relevante en la sociedad contemporánea que puede tener consecuencias graves en el bienestar emocional y psicológico de las víctimas. Actualmente, debido al voluminoso intercambio de interacciones digitales, resulta desafiante y laborioso para los moderadores de las plataformas en línea detectar y eliminar todos los comentarios ciberacosadores de manera manual. Por lo tanto, se necesitan modelos automáticos que por medio de técnicas de inteligencia artificial detecten el ciberacoso. En este artículo se proponen modelos basados en técnicas de aprendizaje automático y en modelos del lenguaje para la detección de ciberacoso en la red social Twitter. Las técnicas de aprendizaje automático utilizadas son XGBoost, regresión logística, y bosques aleatorios. Por su parte, como modelo de lenguaje se hizo un proceso de fine-tuning al modelo de lenguaje enmascarado basado en transformer llamado roberta-base-bne. A pesar de que actualmente se tienen diferentes modelos para este mismo propósito, en su mayoría están hechos usando palabras en inglés. En el caso del español son muy

https://doi.org/10.24050/reia.v22i44.1757

pocos los trabajos propuestos y en el caso particular del español colombiano no se tiene precedente de aportes en el área.

En este artículo se propone además un corpus que contiene tweets escritos en español colombiano y que fueron anotados por una terapeuta ocupacional experta. A partir de este corpus se crean dos conjuntos de datos. El dataset 1 se caracteriza por tener, para una palabra dada, un tweet anotado como ciberacoso que contiene dicha palabra y otro tweet anotado como no ciberacoso con la misma palabra. Esto se logra gracias a que se tiene en cuenta el contexto en el que se usa cada palabra. Por su parte, el dataset 2 usa palabras diferentes en los tweets que son ciberacoso y en aquellos que no lo son. El propósito de utilizar estos dos conjuntos de datos se centra en capturar diversas manifestaciones del lenguaje y su contexto, y evaluar si las técnicas utilizadas permiten entender dicho contexto. Los resultados obtenidos muestran que los modelos propuestos con regresión logística, bosques aleatorios, XGBoost, y roberta-base-bne, alcanzan un área bajo la curva ROC en el dataset 1 de 0.797, 0.796, 0.785, y 0.910, respectivamente. Por su parte, en el dataset 2, el área bajo la curva ROC es de 0.983, 0.978, 0.971, y 0.996, respectivamente. Finalmente, se presenta una aplicación web llamada AI Cyberbullying Detector que está dirigida a terapeutas para que puedan hacer uso de la inteligencia artificial en estudios relacionados con el ciberacoso.

Palabras clave: Aprendizaje de máquina; Bosques aleatorios, Ciberacoso; Español Colombiano; Inteligencia artificial; Transformers; Modelos de lenguaje; Regresión logística; Twitter; XGBoost.

1. Introduction

Cyberbullying has been defined as a form of harassment carried out through technological means that affects millions of people worldwide each year (Hassan et al., 2023). According to Feijóo et al. (2021), cyberbullying can be described as repetitive, negative, and harmful behavior using electronic communication tools, involving a power imbalance with the less powerful individual or group being unjustly attacked. Research on cyberbullying in Latin America is limited. However, available studies indicate that Colombia has a high incidence rate of these behaviors, reaching 63%. This figure surpasses the average of 51.1% observed in 16 Latin American countries and the 29.2% reported in a comparison of 32 European nations along with the United States (Herrera-López et al., 2017). According to Marín-Cortés et al. (2020), cyberbullying can trigger severe psychosocial, affective, and academic problems in victims, and in extreme cases, it can lead to suicide, emphasizing the crucial importance of not downplaying the significance of such assaults.

According to Khan and Qureshi (2022), although major social media platforms such as Twitter, Instagram, and Facebook have implemented policies against cyberbullying, it is challenging to eliminate all offensive content in different languages considering the vast volume of data being transmitted. Within its anti-cyberbullying policies, Twitter is committed to providing a safe space for users. In an effort to promote healthy dialogue, the platform prohibits all behavior and content seeking to harass, shame, or degrade others. Examples of such behavior include selective harassment, which can manifest through the posting of numerous malicious tweets in a short period to target an individual, or mentioning/tagging users in malicious content. Twitter also prohibits behaviors that incite others to harass or attack specific individuals or groups. Additionally, the sharing of unwanted sexual content and explicit objectification of a person without consent is prohibited. Regarding insults, Twitter takes action against the use of obscene language to attack others, especially when the context involves harassment or intimidation (Aránguez, 2022). The denial of mass violent events is also a violation of this policy. In case of policy violation, Twitter implements a series of sanctions that vary based on the severity of the infringement and the offender's history. Some measures may include restricting the visibility of tweets in replies and search results, excluding tweets or accounts from email recommendations, requesting tweet removal, or even suspending accounts. To help their teams understand the context, Twitter sometimes needs to communicate directly with the recipient of the message and/or relies on other users to report malicious content. However, given the volume of interactions and the speed at which content is generated on the platform, depending on user reports or waiting for a Twitter official to detect an offensive message can prove inefficient. Therefore, it becomes imperative to implement automatic cyberbullying detection models to enable a quicker and more effective response.

Studies related to cyberbullying have predominantly focused on the social and psychological aspects of cyberbullying, such as the scope of the problem, adverse effects on victims, and methods of addressing it (Wang et al., 2024; Kee et al., 2024). However, with the emergence and rapid development of artificial intelligence, current research efforts have shifted their focus to ways of detecting cyberbullying, primarily through the use of natural language processing (NLP) and machine learning. Various studies employing machine learning techniques have been conducted to detect cyberbullying on different social networks.

For instance, Van Hee et al. (2018) employed support vector machines (SVM) for binary classification, associating each text with Cyberbullying or Non-Cyberbullying labels. The experiments, conducted on datasets of 113,698 and 78,387 posts collected from various English and Dutch social networks, respectively, achieved F1 scores of 64% for the English dataset and 61% for the Dutch dataset. Balakrishnan et al. (2020) used a dataset comprising 5,453 English tweets and the J48 decision tree algorithm to detect cyberbullying, achieving an accuracy of 91.88% and an area under the ROC curve of 97%. Similarly, Khan and Qureshi (2022) utilized a dataset comprising 7,625 tweets in Urdu and found that the Multinomial Naive Bayes (MNB) algorithm, when applied with the Bag of Words (BOW) technique, achieved a precision of 91.87%. León-Paredes et al. (2019) used a dataset of 960,578 tweets in Spanish, specifically from Chile, and found that the support vector machine technique allowed for the identification of cyberbullying cases, achieving a precision of 93%. Johari and Jaafar (2022) used logistic regression on a dataset comprising 45,580 tweets in Malay, achieving a precision of 76%. Bozyiğit et al. (2021) employed 7,000 Turkish tweets and various techniques, including K-nearest neighbors, Multinomial Naive Bayes (NBM), AdaBoost, and Random Forests, with AdaBoost being the most effective, reaching an accuracy of 90.1%. Al-garadi et al. (2016) collected English tweets from January to February 2015, specifically from the state of California. Different techniques, such as Naive Bayes, support vector machines, random forests, and K-nearest neighbors, were employed. The best model for cyberbullying detection was achieved using random forests, with an area under the ROC curve of 0.943. Chia et al. (2021) used a dataset of English tweets and techniques such as Naive Bayes, J48, and convolutional neural networks, achieving an F score of 0.883 with random forests. Additionally, Zhang et al. (2016) proposed a method based on convolutional neural networks on a dataset of 1,313 English tweets, achieving an accuracy of 0.968 and an F1-score of 0.562.

Despite the proposed models for cyberbullying detection in English and other languages, few works address cyberbullying detection in Spanish. Moreover, there is no publicly available Spanish dataset for cyberbullying detection containing words and expressions unique to Colombian Spanish. Spanish, being spoken in many countries and regions, exhibits a wide variety of dialects and variants. Colombian Spanish, for example, has peculiarities in vocabulary, pronunciation, grammar, and usage that can significantly differ from other forms of Spanish, such as Spanish from Spain, Mexico, or Argentina. For natural language processing applications, like cyberbullying detection, these differences can be crucial. For instance, in Spain, the term "gilipollas" is frequently used to insult someone, equating to calling them foolish or idiotic. This term is specific to Spanish in Spain and is not commonly used in other Spanish-speaking countries. In Colombia, an equivalent insult might be "güevón," and in Argentina, "boludo." While these terms may have other meanings depending on the context and are occasionally used in a friendly or informal manner among acquaintances, it is crucial to recognize these distinctions. Similarly, the term "pendejo" is scarcely used as an insult in Spain; however, in countries like Mexico and Colombia, it is commonly used in that manner, referring to someone as foolish or stupid. For natural language processing applications, like cyberbullying detection, these differences can be crucial. For accurate cyberbullying detection, it is essential that the model be trained on Colombian Spanish as the vocabulary and expressions can vary widely between different Spanish dialects. A word that is harmless in one country can be offensive in another. Additionally, idioms and colloquial expressions, often used in online interactions, can be unique to each region or country. Therefore, for precise cyberbullying detection, it is crucial that the model is trained with Colombian Spanish if applied in this context.

This article delves into artificial intelligence techniques, specifically machine learning and language models, for the detection of cyberbullying in Colombian Spanish. Additionally, the dataset of Colombian Spanish tweets, employed for training and evaluating the models, underwent manual annotation with the assistance of an occupational therapist. This dataset is made publicly available, enabling the exploration of additional techniques beyond those considered in this study through further research. Lastly, the article introduces a web application designed to facilitate the utilization of the proposed artificial intelligence models. The application aims to bridge the gap for therapists interested in the field of cyberbullying, providing them with a user-friendly interface for leveraging intelligent models in decision-making processes.

2. Materials and Methods

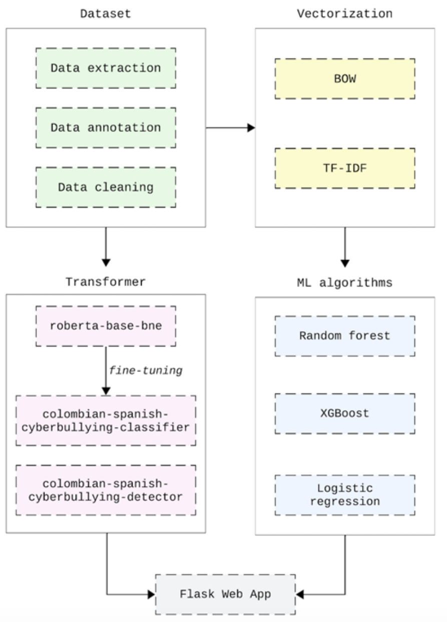
Figure 1 illustrates the proposed process for cyberbullying detection on Twitter using machine learning and language models. Initially, tweet extraction is performed. Subsequently, Natural Language Processing (NLP) is applied—a procedure involving the removal of links and special characters, conversion of all text to lowercase, and segmentation into individual words or 'tokens'. Additionally, lemmatization is employed, reducing words to their basic or root form, facilitating analysis and comprehension. Following this, vectorization techniques such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are utilized to transform textual content into numerical representations serving as input for machine learning algorithms and the language model.

Specifically, three machine learning techniques—XGBoost, Logistic Regression, and Random Forests—are employed. Furthermore, for the language model, a fine-tuning process is conducted on the roberta-base-bne language model developed by Gutiérrez et al. (2022). This fine-tuning aims to explore and enhance cyberbullying detection. Two language models are generated through this process. Finally, the best model is selected based on the area under the ROC curve and made accessible through a web application for cyberbullying detection, catering to the needs of healthcare professionals.

2.1. Dataset

For this investigation, tweets were manually collected due to recent restrictions imposed on the Twitter API. Each tweet was annotated into two categories: Cyberbullying and Non-Cyberbullying. This annotation was conducted in collaboration with an expert occupational therapist, who also provided specific keywords and phrases used for tweet selection. The Twitter search was executed using key words and phrases followed by a specific geographical location, for example, "tonto near:Colombia." The selection of key words and phrases was based on the categories outlined in the cyberbullying data annotation guide proposed by Van Hee et al. (2015). This guide meticulously describes the guidelines for cyberbullying data annotation, incorporating four categories: insult, threat, ill-wishing, and defamation. The insult category involves the use of offensive words with the intent to harm the other person, while the threat aims to damage the victim's integrity. Ill-wishing encompasses curse words or expressions wishing harm to the individual, and defamation seeks to harm the victim's reputation. These categories were chosen to capture a broad representation of the various forms in which cyberbullying can manifest.

Figure 1. *Cyberbullying Detection Process Using Machine Learning and Language Models.*



The guide mentions an additional category related to sexual harassment. However, this research did not specifically address this category, as the therapist identified that sexual harassment, considered a variant of general harassment, can manifest and be classified within the four aforementioned categories. In other words, sexual harassment may present itself to the victim in the form of a threat or insult, for example. Table 1 displays some of the keywords and phrases provided by the expert therapist that were used to search for tweets, along with two examples from each of the categories utilized in this study.

Table 1. *Examples of Tweets and Keywords by Category.*

Category	Keywords	Examples
Insult	Retrasado, enano maldito, mongólico, veneco, mugroso, mamahuevo, atolondrado, estípido, imbécil, mamaburra, analfabeta, petardo, tonta, puta, bocón, cacorro, boquisucio,bufón, gorda, perra, sapa.	1. Vendé a tu madre pa conseguir eso pues sapa hpta malnacida de mierda. 2. Perra quejona esta.
Threat	Te llegará la hora, llegará tu caída, le dan en la jeta, nos tenemos que matar, de hoy no pasas, te voy a matar, muérase, nadie te va salvar, te voy a buscar, te voy a encontrar, te voy a meter un traque.	1. Vas a volver y te voy a matar hijo de puta. 2. Mira puta, si volvés a regresar con ese man voy a publicar tus nudes y las conversaciones cochinas que tengo en mi poder.
Ill-wishing	Ojala te mueras,ojalá te violen,ojalá te maten morite, muérase, muerete, suicidate, suicídese, ahórquese, matate, cortate las venas, tírate de un balcón, peguese un tiro, tirese de un puente.	1. Nadie te quiere mamarracho y nadie te va a escuchar. 2. Tú estás peor, tu vida no vale nada, suicidate.
Defamation	Bandido, guerrillero, criminal, terrorista, paramilitar, paraco, asesino, rata, delincuente, ladrón, suplantador, impostor, prostituta, puta, violador, abusador, pedófilo, pederasta, racista, misógino.	1. Sos un delincuente, eso es lo que sos 2. Esta vieja loca está pero perdida en la droga.

In this research, two datasets were created for training and testing the models. Dataset 1, comprising a total of 3570 tweets, contains an equal number of tweets labeled as Cyberbullying and Non-Cyberbullying. What characterizes this dataset is that for a given word or phrase, there is one tweet annotated as Cyberbullying containing that word and another tweet annotated as Non-Cyberbullying with the same word. For instance, given the phrase “Te voy a buscar” (“I’m going to find you”), one might have the tweet

"Te extraño mucho y por eso te voy a buscar" ("I miss you a lot, so I'm going to find you"), which is not cyberbullying. However, in the tweet "te voy a buscar y te voy a encontrar, cerdo" ("I'm going to find you, and I will find you, pig"), this same phrase is associated with cyberbullying. This illustrates that the context of the word or phrase determines whether a tweet is classified as cyberbullying or not.

In this dataset, Non-Cyberbullying tweets mostly contain offensive words that, in their context, do not correspond to cyberbullying, for example, "Marica, se me olvidó ver el partido" ("Dude, I forgot to watch the game"). Similarly, the Non-Cyberbullying category includes, to a lesser extent, tweets obtained from the trending topics in the Colombian region. Twitter trends reflect the most popular topics and conversations in a specific region at a given time, in other words, they provide a snapshot of what people are discussing and sharing online in that geographical area. Trending tweets were used in cases where it was not possible to obtain Non-Cyberbullying tweets with a specific offensive word or phrase, such as "ojalá te violen" ("I hope you get raped").

Regarding the distribution of tweets in Dataset 1, for both Cyberbullying and Non-Cyberbullying classes, there are 968 tweets in the Insult category, 128 tweets in the Threat category, 187 related to Ill-wishing, and 502 tweets in the Defamation category. This implies that for each category, there is an equal number of tweets labeled as Cyberbullying and Non-Cyberbullying. For instance, in the Insult category, there are 968 tweets that are Cyberbullying and 968 that are not. In this dataset, it is observed that the Insult and Defamation categories have significantly more tweets than the Threat and Ill-wishing categories. This is attributed to the fact that certain types of behaviors, such as insults, can be much more common in online communication than others (Wang et al., 2014), which may impact the dataset distribution.

In Dataset 2, comprising 2566 tweets, a balanced distribution was also maintained between Cyberbullying and Non-Cyberbullying tweets. However, in this case, Non-Cyberbullying tweets are characterized by the absence of obscene language. Specifically, these tweets were extracted from Colombian Twitter accounts with over ten thousand followers that regularly and diversely publish content, providing a broad spectrum of discourses and topics. Examples of such accounts include @diegoalejocm, @CristoAtado, and @ LeMonda__. The 1283 tweets in the Cyberbullying class in this dataset were randomly selected from the 1785 tweets in Dataset 1. Concerning the distribution of Cyberbullying class tweets in Dataset 2, there are 688 tweets in the Insult category, 101 tweets in the Threat category, 137 related to Ill-wishing, and 357 tweets in the Defamation category. In this dataset, the distinction between Cyberbullying and Non-Cyberbullying tweets is clearer. For instance, a tweet that is not cyberbullying is "Otra vela por Wilson y cuatro perritos que aparezcan" ("Another candle for Wilson and four little dogs that show up"), whereas a tweet considered cyberbullying is "Te voy a violar maldita perra, me tenés mamado con tu ideología de género maldita enferma" ("I'm going to rape you, damn bitch, I'm fed up with your damn sick gender ideology").

The decision to use these two datasets is centered on the intention to capture diverse manifestations of language and their context on social media platforms, specifically Twitter. Each dataset possesses distinctive characteristics that can contribute to enhancing the robustness and generalization of predictive models. Dataset 1 allows an examination of whether machine learning models and language models can identify cyberbullying not solely based on the use of obscene or strong words but also within the context in which they are employed. This is crucial because the mere presence of offensive language does not necessarily determine cyberbullying. Moreover, tweets labeled as Cyberbullying may lack words considered strong or obscene, underscoring the importance of context. Therefore, while this dataset provides a more nuanced representation of cyberbullying, it may pose challenges for predictive models employed in this research.

On the other hand, Dataset 2 is particularly useful for working with vectorization techniques such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), as well as classification techniques like XGBoost, logistic regression, and random forests. Since this dataset maintains a more pronounced distinction between Cyberbullying and Non-Cyberbullying tweets, it may be easier for models to detect these classes. It is important to highlight that the inclusion of typical Colombian Spanish words or phrases in both datasets, such as “gonorrea,” “mamaburra,” or “te voy a meter un traque,” adds an additional layer of precision to the model in identifying cyberbullying. This is because these expressions may have specific connotations and uses in the Colombian context that might not be recognized using a more generic approach. Integrating these linguistic peculiarities allows the model to be more sensitive and effective in capturing local manifestations of cyberbullying. The datasets used in this article, along with the annotation provided by the expert therapist, are made available for further research to explore additional techniques beyond those considered in this study (Guerra, 2023c; Guerra, 2023d).

2.2 Data Preprocessing

The data preprocessing involved a series of steps to prepare the tweets for subsequent modeling. Firstly, emojis were removed, and all text was converted to lowercase to ensure uniformity in the data. Mentions, links, and special characters were also eliminated to reduce noise in the tweets and focus on relevant textual content. Next, tokenization was performed, involving the division of text into individual words or “tokens,” which are the basic unit of analysis in natural language processing. Subsequently, stop words, common words that do not contribute meaning to the text, such as “the,” “and,” “in,” among others, were removed. To further refine the text, lemmatization was carried out. Lemmatization involves reducing words to their lemma or base form, considering the linguistic context. These steps allow grouping similar words and reducing the dimensionality of the feature space.

To represent the data in a format suitable for processing by machine learning algorithms, two techniques were employed: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). BoW converts the text into a matrix where each row represents a document, and each column represents a unique word in the corpus. The values in the matrix denote the frequency of each word in the document. On the other hand, TF-IDF not only considers word frequency but also penalizes words that appear too frequently in the corpus, as such words may be less informative. The size of the Bag of Words (BoW) in dataset 1 is 6892 terms, and in dataset 2, it is 5416 terms. The term sizes for TF-IDF are likewise 6892 terms for dataset 1 and 5416 terms for dataset 2. The term counts in each dataset represent the number of unique features (words) considered in the analysis for both methods. The equality in the number of terms between BoW and TF-IDF in each particular dataset is because both methods analyze the same set of words, although the way these words are weighted and used for analysis differs between the two techniques. In the BoW technique, text is represented as a set of unique words and the number of times each word occurs in the text. In contrast, the TF-IDF technique reflects the importance of a word in a document compared to the entire corpus. The vector representations of BoW and TF-IDF are sparse matrices, as each tweet is represented as a vector in a multi-dimensional space, where each dimension corresponds to a unique word in the dataset. Most dimensions will be zero since a given tweet will only contain a small fraction of the possible words. Table 2 displays the representation of one of the tweets with its respective counts (in the case of BoW) or weights (in the case of TF-IDF). In the TF-IDF representation, a higher weight indicates that the word is more important or relevant to the specific document within the total set of documents.

Table 2. *Representation of a Tweet using BoW and TF-IDF.*

Preprocessed Tweet	BoW (counts)	TF-IDF (weights)
odar sapo catretahijueputa malparir asqueroso gonorreo inmundar rata malparida	odar : 1 sapo : 1 catretahijueputa : 1 malparir : 1 asqueroso : 1 gonorreo : 1 inmundar : 1 rata : 1 malparida : 1	TF-IDF: odar : 0.31755517506076486 sapo : 0.36325667926605104 catretahijueputa : 0.47512946662948546 malparir : 0.25080021947838876 asqueroso : 0.2652485975857141 gonorreo : 0.29241921119211706 inmundar : 0.38260088536904013 rata : 0.23832091191611024 malparida : 0.3456999095654974

2.3. Application of Artificial Intelligence Techniques

This research employed various models for cyberbullying detection, utilizing both machine learning algorithms and language models. The scikit-learn tool in the Python programming language was employed for machine learning algorithms. Additionally, the roberta-base-bne, a masked language model based on transformers, was used for language models. Each technique involves a set of hyperparameters that require tuning through experimentation to ascertain the model capable of making predictions with the highest accuracy.

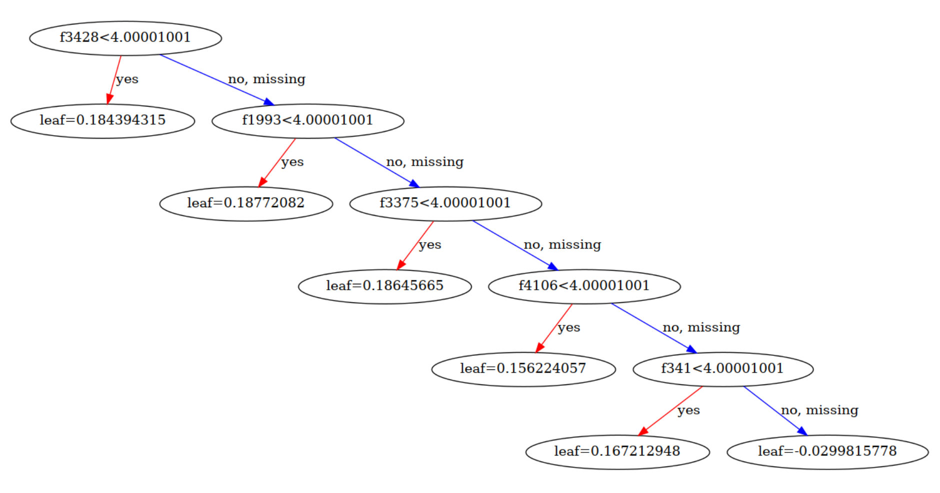
2.3.1. Cyberbullying Detection Models with XGBoost

In this study, models were developed using the XGBoost technique implemented through the xgboost library in Python. The hyperparameters explored during experimentation included the number of estimators (n_estimators), maximum tree depth (max_depth), learning rate (learning_rate), and the random seed for algorithm initialization (random_state). The combinations of these hyperparameters resulted in a total of 72 configurations.

The hyperparameter n_estimators determines the number of individual decision trees in the ensemble model. Adjusting this hyperparameter allows an evaluation of how the quantity of trees influences model performance. While a higher number of trees may enhance accuracy, it could also lead to overfitting if the number is excessively high (Peláez & Lena, 2021). Hence, values of 100, 200, 400, and 500 were tested. The max_depth hyperparameter sets the maximum depth of the trees, controlling the model's complexity and the number of features that can be used for prediction. In this case, values of 5, 8, and 10 were tested to strike a balance between capturing complex interactions in the data and preventing overfitting. The learning_rate hyperparameter affects the speed at which the model learns. A lower learning rate may require more estimators to converge to a solution but can also provide a more robust solution. Two values, 0.01 and 0.1, were tested. Lastly, the random_state hyperparameter was used to ensure reproducibility of model training results and to explore how different random initializations may influence outcomes. Values of 10, 42, and 250 were employed for this hyperparameter.

For each set of hyperparameters, a grid search was conducted using the GridSearch function, employing a five-fold crossvalidation. This approach enabled the exploration of a broad range of configurations by optimizing performance based on the area under the ROC curve. The tree depicted in Figure 2 is one instance of a decision tree within the XGBoost model. In the graph, each node (except the leaves) represents a split based on a feature value. This is indicated by the statement in the node, such as "f3428<4.00001001," signifying that the node will split the data into two groups based on whether the value of feature 1 (f3428) is less than 4.00001001 or not. In this specific case, "f3428" refers to the word "rata" (rat) from the bag of words. This word was found in several tweets labeled as cyberbullying, such as "disgusting rat bastard." The tree node is leveraging this feature to make a splitting decision. The condition "f3428<4.00001001" is a split decision generated by the XGBoost algorithm. In this context, it could be interpreted as an indication of how many times the word "rat" needs to appear in a text for a decision path in the decision tree to be taken.

Figure 2. *Instance of a Decision Tree within the XGBoost Model.*



In the BoW technique, each word in the corpus is associated with a count indicating how many times it appears in each text. Thus, the number 4.00001001 is the threshold that XGBoost has identified for this tree node during the training process. If the value of the feature "f3428" (the frequency of the word "rat" in this case) for a particular instance is less than 4.00001001, then the instance follows the branch that says "yes" (the left arrow). If it is greater than or equal to 4.00001001, or if the value is unknown ("missing"), the instance follows the branch that says "no, missing" (the right arrow).

When the instance follows the "yes" branch, it reaches a leaf of the tree. The leaf has a value of "leaf=0.184394315." This is the value assigned to this instance in this tree. In the case of a binary classification problem, this value is aggregated with the values of all other leaves in all other trees in the XGBoost model. Then, a link function is applied to obtain the final probability of belonging to the positive class. If the instance follows the "no, missing" branch, it encounters another decision based on the feature "f1993" and a threshold of 4.00001001, and the process continues until reaching a leaf. This is just one of the many trees in the XGBoost model. To make a prediction, XGBoost uses all the trees in the model and sums all the values of the corresponding leaves. The interpretation of the leaves in terms of whether they correspond to Cyberbullying or Non-Cyberbullying cannot be determined from a single tree; it depends on the values of the leaves from all the trees in the model.

2.3.2 Cyberbullying Detection Models with Logistic Regression

In developing prediction models using the logistic regression technique, the LogisticRegression class from scikit-learn was employed. The hyperparameters explored during experimentation included the inverse regularization parameter (C), the type of penalty (penalty), the random seed for algorithm initialization (random_state), the optimization algorithm used (solver), and, in some cases, the ratio between l1 and l2 penalties (l1_ratio). The combinations of these hyperparameters resulted in a total of 135 models.

The inverse regularization parameter C is crucial in logistic regression as it determines the amount of regularization applied. Regularization is a technique to prevent overfitting by reducing the model's complexity, and the value of C controls the strength of this regularization. A lower C value means more regularization and a simpler model, while a higher C value means less regularization and a more complex model. Therefore, varying C allows balancing the model's bias and variance (Salehi et al., 2019). The values 1, 10, 50, 100, and 200 were tested for this hyperparameter. The type of penalty, specified by the penalty hyperparameter, is also significant in logistic regression. The l1 penalty (also known as LASSO regularization) tends to make some feature weights exactly zero, meaning those features do not contribute at all to the model, reducing the number of features the model is using. On the other hand, the l2 penalty tends to distribute model weights more evenly among features. The 'elasticnet' penalty type combines both penalties, allowing a balance between generating sparse models and weight distribution. The ratio between l1 and l2 penalties, specified by the 'l1_ratio' hyperparameter, was tested with the values 0.1, 0.5, and 0.9.

The random_state hyperparameter was used to ensure reproducible results. Varying this hyperparameter allows exploring the model's sensitivity to different random initializations. The values 10, 42, and 250 were used. Lastly, the solver hyperparameter specifies the algorithm used for optimization. 'SAGA' was chosen as it supports all penalties, which is useful for this hyperparameter exploration, and 'newton-cg' and 'liblinear' were also chosen, which work only with l2 penalties in the case of 'newton-cg' and l1 and l2 penalties in the case of 'liblinear.'

2.3.3 Cyberbullying Detection Models with Random Forests

Prediction models using the random forest technique were obtained using the scikit-learn library in Python. The hyperparameters tested included the number of trees in the forest (n_estimators), the maximum depth of the tree (max_depth), the minimum number of samples required to split an internal node (min_samples_split), the minimum number of samples required to be in a leaf node (min_samples_leaf), the criterion used to measure the quality of a split (criterion), and the seed for the random number generator used by the model (random_state). The combinations used resulted in a total of 384 models.

The number of trees in the forest (n_estimators) determines the quantity of trees comprising the model. In this research, 100, 200, 400, and 500 trees were tested. A higher number of trees can increase the model's accuracy but may also lead to overfitting. The maximum depth of the tree (max_depth) in these models was kept as None, meaning nodes expand until all leaves are pure or until all leaves contain fewer samples than min_samples_split. The minimum number of samples required to split an internal node (min_samples_split) was tested with values 2, 10, 30, and 50. This hyperparameter prevents the model from overfitting to the training data by disallowing splits that result in nodes with very few samples. The minimum number of samples required to be in a leaf node (min_samples_leaf) was tested with values 1, 3, 5, and 10, allowing greater flexibility in leaf creation. The criterion for measuring the quality of a split (criterion) was varied between 'gini' and 'entropy.' These two methods provide different ways to measure node impurity and may lead to different splits. Finally, the seed for the random number generator (random_state) was set to specific values (10, 42, 250) to ensure reproducible results and explore how different random initializations may influence outcomes.

Figure 3. *Decision Tree Generated by the Random Forest Technique.*

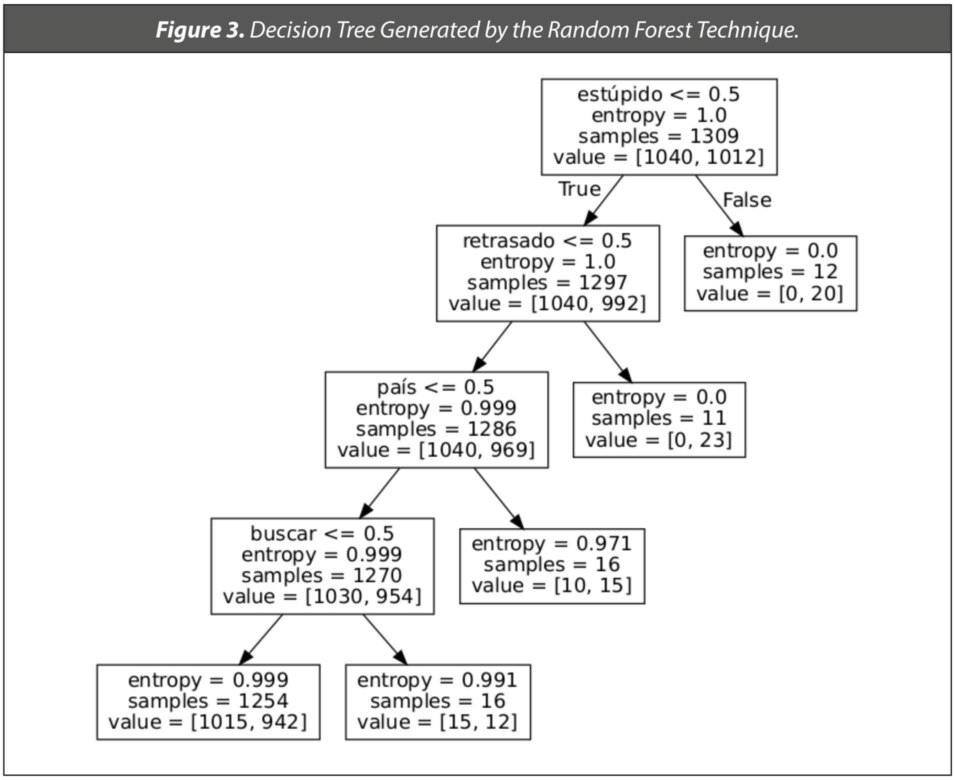


Figure 3 displays a segment of the best decision tree generated by the random forest technique. Each node in the tree represents a test applied to a data sample. Generally, the decision tree is constructed by splitting nodes based on features and values that best separate samples according to some criterion, such as entropy. Nodes closer to the root of the tree are more crucial features for decision-making, while nodes closer to the leaves contain more specific information about sample classes. For example, in the following tree, the word “stupid” plays a fundamental role at the root for cyberbullying detection. If the presence of “stupid” in a text is less than or equal to 0.5, the tree branches toward decisions based on other keywords, such as “retarded,” maintaining an entropy of 1.0, indicating high uncertainty at this level. This analysis pattern repeats with each feature, where each node evaluates a specific word against a set threshold. With 1309 samples at the root node, the tree reflects a diverse data distribution, as shown in the values. As you progress through the tree branches, other nodes consider terms like “country” and “search,” each contributing to the final classification of texts into Cyberbullying or Non-Cyberbullying categories. This stratified approach allows the model to discern potentially abusive language patterns more accurately, using linguistic features as key indicators.

2.3.4 Cyberbullying Detection Model Using Language Models

In this study, a fine-tuning process was conducted on the roberta-base-bne language model developed by Gutiérrez et al. (2022) to explore and enhance cyberbullying detection. The roberta-base-bne model, available on the Hugging Face platform, is a masked language model based on transformers specifically designed and optimized for the Spanish language. However, it is also designed for non-generative tasks such as question answering, text classification, and named entity recognition, expanding its applicability and efficiency in various scenarios and contexts. The roberta-base-bne model was pre-trained using the largest known Spanish corpus to date, comprising a total of 570 GB of clean and processed text compiled from web crawls conducted by the National Library of Spain from 2009 to 2019. This extensive database provides the model with a profound and contextual understanding of the Spanish language, crucial for context-sensitive tasks like cyberbullying detection. The effectiveness of roberta-base-bne is largely attributed to its RoBERTa-based architecture, an optimized variant of the BERT transformer

(Devlin et al., 2019). RoBERTa employs a bidirectional architecture, processing information from all words in a text simultaneously, which is crucial for understanding the context in which words are used and vital for the proper identification and categorization of instances of cyberbullying.

The fine-tuning process of roberta-base-bne using dataset 1 resulted in the colombian-spanish-cyberbullying-classifier model, available on the Hugging Face platform (Guerra, 2023a). The configuration of training hyperparameters was meticulously adjusted to optimize the model's performance. To achieve this, the Optuna library (Takuya et al., 2019), an advanced tool for hyperparameter optimization, was utilized along with the TrainingArguments class. Considered hyperparameters included the learning rate, the number of epochs, and batch sizes for training. The learning rate range was set between 0.000003 and 0.01, optimized on a logarithmic scale to explore a broad range of values. The number of epochs varied between 1 and 5, providing enough flexibility to assess model convergence and the possibility of overfitting. Batch sizes for training and evaluation were logarithmically adjusted between 4 and 32, ensuring a balance between computational efficiency and training quality. Additionally, hyperparameters such as weight_decay and warmup_steps were optimized. Weight_decay, essential for model regularization and overfitting mitigation, was adjusted in a range from 0.005 to 0.02. Warmup_steps, associated with training stability during initial phases, was explored in a range from 100 to 1000.

On the other hand, the fine-tuning process of roberta-base-bne using dataset 2 resulted in the colombian-spanish-cyberbullying-detector model, available on the Hugging Face platform (Guerra, 2023b). For this model, the optimization of training hyperparameters was adjusted similarly using the Optuna library, considering the same number of hyperparameters and experimentation ranges as in the model with dataset 1.

3. Results and Discussion

The evaluation and comparison of prediction model performance were initially conducted using the confusion matrix. The parameters calculated in each trial include true positives (TP), corresponding to the number of cyberbullying-related tweets correctly classified by the model as cyberbullying; true negatives (TN), indicating the number of non-cyberbullying tweets correctly classified by the model as non-cyberbullying; false positives (FP), representing the number of non-cyberbullying tweets incorrectly classified by the model as cyberbullying; and finally, false negatives (FN), corresponding to the number of cyberbullying-related tweets incorrectly classified by the model as non-cyberbullying. Subsequently, using the values from the confusion matrix, accuracy ((TP+TN)/(TP+TN+FN+FP)) was calculated based on the number of correct positive predictions, precision (TP/(TP+FP)) allowing insight into the fraction of true positives among positive cases, sensitivity defined according to the formula TP/(TP+FN), indicating the rate of true positives, specificity ((TN/(TN+FP)), and the F1-score (2*(precision*recall)/(precision + recall)). Additionally, the area under the Receiver Operating Characteristic (ROC) curve was computed for each model.

3.1 Results of XGBoost Prediction Models

Table 3 presents the results obtained by the top 10 models based on the area under the ROC curve (AUROC) using the XGBoost algorithm along with the Bag of Words (BoW) technique on Dataset 1. Overall, the highest area under the ROC curve is 0.778, indicating that the model's predictions are mostly accurate but have limitations. In other words, there is still a significant proportion of cases where the model struggles to distinguish correctly between Cyberbullying and Non-Cyberbullying classes, leading to classification errors, both in terms of false positives and false negatives. Table 3 highlights that the model's primary challenge lies in its sensitivity or recall, reflecting a limitation in consistently identifying all true cases of cyberbullying. This deficiency is crucial in cyberbullying contexts, where comprehensive detection is paramount. Additionally, although the model's precision is moderately high, a significant margin of false positives is still present, suggesting the need to fine-tune the balance between accurately identifying cyberbullying cases and reducing incorrect alerts.

Results obtained using the XGBoost algorithm along with the TF-IDF technique on Dataset 1 reveal a precision ranging from 0.7073 to 0.7286, while recall fluctuates between 0.624 and 0.699. Specificity also exhibits variability, with values ranging from 0.683 to 0.730. Accuracy, on the other hand, falls within the range of 0.675 to 0.704. The F1 score and AUROC follow a similar trend, with values from 0.667 to 0.710 and 0.750 to 0.778, respectively.

Table 3. *Results obtained by the top 10 models using the XGBoost technique with Bag of Words (BoW) on Dataset 1.*

Estimators	Max depth	Learning rate	Random state	Precision	Recall	Specificity	Accuracy	F1 score	AUROC
500	10	0.1	10	0.723	0.694	0.709	0.701	0.708	0.778
500	8	0.1	250	0.721	0.694	0.706	0.700	0.707	0.777
500	5	0.1	42	0.727	0.694	0.715	0.704	0.710	0.775
400	5	0.1	42	0.721	0.680	0.712	0.696	0.700	0.770
200	10	0.1	10	0.707	0.699	0.683	0.691	0.703	0.768
200	8	0.1	250	0.719	0.672	0.712	0.691	0.695	0.768
200	5	0.1	42	0.728	0.662	0.730	0.694	0.693	0.765
100	8	0.1	250	0.718	0.656	0.718	0.686	0.686	0.765
100	10	0.1	10	0.716	0.678	0.706	0.691	0.696	0.759
100	5	0.1	42	0.716	0.624	0.730	0.675	0.667	0.750

The results from models evaluated using the XGBoost algorithm, both with BoW and TF-IDF, applied to Dataset 1, reveal certain limitations in their classification capability. The distinctive nature of Dataset 1 offers explanations for this performance, as this dataset presents a language complexity that might have challenged the models. Tweets labeled as Non-Cyberbullying in Dataset 1 contain obscenities that, in their context, do not correspond to cyberbullying. An example of a false positive in this dataset is "El país vuelto mierda y usted anda mamando gallo" ("The country turned to shit, and you're fooling around"). In this case, the use of the word "mierda" ("shit") makes the model incorrectly classify this tweet as cyberbullying, as the context in which the word is used does not make it cyberbullying. On the other hand, a false negative presented by this model is the tweet "Vieja lesbiana y menopáusica dejé de hablar basura" ("Old lesbian and menopausal, stop talking garbage"). In particular, it can be observed that the model is unable to capture the context of the phrase to classify it as cyberbullying. In conclusion, the models show limited ability to navigate the subtleties and nuances of language in Dataset 1. The presence of obscenities in a non-offensive context and the variability in language used in tweets labeled as Cyberbullying may have contributed to this moderate performance.

Next, the results obtained using Dataset 2 are presented. Table 4 shows the results of the top ten models organized according to the value obtained in the area under the ROC curve when using the XGBoost algorithm and the BoW technique. The standout model achieved an area under the ROC curve of 0.971 and a precision of 0.966. These metrics, along with equally high recall, suggest that the model is not only effective in correctly identifying cyberbullying cases but also in minimizing false positives. This improvement in the balance between precision and sensitivity is crucial for real-world applications where both accurate detection and minimizing false positives are essential. According to the results, it can be observed that tweets in Dataset 2 are easier to classify. This is due to the nature of this dataset, where there is a significant difference in the language used in tweets that are cyberbullying and those that are not.

Table 4. *Results obtained by the top 10 models using the XGBoost technique with Bag of Words (BoW) on Dataset 2.*

Estimators	Max depth	Learning rate	Random state	Precision	Recall	Specificity	Accuracy	F1 score	AUROC
500	5	0.1	42	0.966	0.878	0.968	0.922	0.920	0.971
400	5	0.1	42	0.966	0.867	0.968	0.916	0.914	0.970
200	10	0.1	10	0.966	0.875	0.968	0.920	0.918	0.970
500	8	0.1	250	0.963	0.893	0.964	0.928	0.927	0.967
200	8	0.1	250	0.966	0.863	0.968	0.914	0.912	0.966
500	10	0.1	10	0.959	0.890	0.960	0.924	0.923	0.965
200	5	0.1	42	0.955	0.818	0.960	0.887	0.881	0.957
100	10	0.1	10	0.955	0.821	0.960	0.889	0.883	0.954
100	8	0.1	250	0.954	0.795	0.960	0.875	0.867	0.948
100	5	0.1	42	0.953	0.776	0.960	0.865	0.855	0.933

Results obtained using the XGBoost algorithm along with the TF-IDF technique on Dataset 2 reveal precision ranging from 0.934 to 0.961, while recall fluctuates between 0.753 and 0.871. Specificity also exhibits variability, with values ranging from 0.936 to 0.968. Accuracy, on the other hand, falls within the range of 0.857 to 0.910. The F1 score and the area under the ROC curve follow a similar trend, with values from 0.845 to 0.905 and 0.935 to 0.968, respectively. The main difference between the best results of each approach is that, although both have very close AUROC values, the BoW-based model outperforms the TF-IDF model in precision and recall. This indicates that, for this dataset and specific task, BoW is more effective in accurately classifying cyberbullying cases and identifying a higher proportion of these actual cases, despite the overall class discrimination ability (AUROC) being similar in both models.

In Dataset 2, it is observed that the set of hyperparameters and vectorization techniques employed significantly influenced the results obtained by the models. The BoW vectorization method achieved the best performance, with the hyperparameter configuration ‘n_estimators’: 500, ‘max_depth’: 5, and ‘learning_rate’: 0.1 obtaining the highest AUROC of 0.971. The selected hyperparameters allowed the model to handle the data variability adequately, with a tree depth high enough to capture complex patterns and sufficient trees to stabilize predictions. Similarly, with TF-IDF vectorization, the same hyperparameter configuration obtained the highest AUROC of 0.968. BoW vectorization was more effective, suggesting that the presence of words is more predictive for this specific problem. Regarding other metrics, similar trends can be observed. Configurations with ‘max_depth’: 5 tend to achieve higher precision, indicating that these models are less prone to generating false positives. On the other hand, the configuration with ‘max_depth’: 10 tends to have higher recall, meaning that these models are good at capturing most true positives but at the cost of issuing more false positives.

3.2 Results of Logistic Regression Prediction Models

Table 5 presents the results obtained by the top 10 models based on the area under the ROC curve using Logistic Regression along with the Bag-of-Words (BoW) technique on Dataset 1. According to the results, it is observed that this technique exhibits similar limitations to those of models obtained with the XGBoost algorithm. A precision of 0.713 indicates that, although the model can correctly identify a significant portion of the cases, there is still a considerable proportion of tweets that it does not classify adequately. This is particularly relevant in the context of cyberbullying detection.

Table 5. *Results obtained by the top 10 models using Logistic Regression with Bag-of-Words (BoW) technique on Dataset 1.*

C	Penalty	Solver	Random State	Precision	Recall	Specificity	Accuracy	F1 score	AUROC
10	l2	saga	250	0.713	0.774	0.659	0.719	0.742	0.795
50	l2	saga	250	0.712	0.745	0.671	0.710	0.728	0.783
50	elasticnet	saga	10	0.709	0.745	0.665	0.707	0.726	0.779
100	l2	saga	250	0.709	0.734	0.671	0.704	0.722	0.778
100	elasticnet	saga	10	0.709	0.731	0.671	0.703	0.720	0.778
1	l1	saga	42	0.708	0.742	0.665	0.705	0.725	0.774
10	elasticnet	saga	10	0.695	0.747	0.642	0.697	0.720	0.772
10	l1	saga	42	0.696	0.739	0.648	0.696	0.717	0.766
200	l2	newton-cg	42	0.702	0.734	0.659	0.698	0.718	0.765
50	l1	liblinear	42	0.686	0.721	0.639	0.682	0.703	0.746

The results using Logistic Regression along with the TF-IDF technique on Dataset 1 exhibit a precision ranging from 0.686 to 0.749, while recall fluctuates between 0.638 and 0.713. Specificity ranges from 0.671 to 0.739. Additionally, accuracy falls within a range of 0.663 to 0.725. The F1 score and the area under the ROC curve achieve values of 0.671 to 0.730 and 0.738 to 0.797, respectively.

Models evaluated using the Logistic Regression algorithm with both BoW and TF-IDF, applied to Dataset 1, show notable differences in their performance. TF-IDF proves to be superior under these conditions, with values generally outperforming those obtained with BoW. However, there is still room for improvement, as the values do not reach excellent thresholds due to the particular nature of Dataset 1 and the limitations of classification algorithms. According to the results from Dataset 1, both hyperparameter selection and the vectorization method play a crucial role in the performance of the

Logistic Regression model. Overall, TF-IDF has proven to be better than BoW for this task, possibly reflecting a greater sensitivity to the relevance of words in the corpus. Regarding hyperparameters, it is observed that the regularization constant (C), the type of penalty, and the solver influence the model metrics differently.

Next, the results obtained using Dataset 2 will be presented. Table 6 shows the results of the top ten models using BoW, ordered according to the area under the ROC curve. Once again, similar to XGBoost, a significant improvement has been achieved in all key metrics. The enhancement in the area under the ROC curve indicates superior ability to discriminate between the Cyberbullying and Non-Cyberbullying classes.

Table 6. *Results obtained by the top 10 models using Logistic Regression with Bag-of-Words (BoW) technique on Dataset 2.*

C	penalty	solver	Random state	Precision	Recall	Specificity	Accuracy	F1 score	AUROC
10	l1	saga	42	0.956	0.912	0.956	0.933	0.934	0.980
10	elasticnet	saga	10	0.956	0.905	0.956	0.929	0.929	0.980
50	l1	liblinear	42	0.952	0.916	0.952	0.933	0.934	0.980
1	l1	saga	42	0.967	0.893	0.968	0.929	0.929	0.975
200	l2	newton-cg	42	0.944	0.905	0.944	0.924	0.924	0.974
10	l2	saga	250	0.948	0.901	0.948	0.924	0.924	0.971
50	l2	saga	250	0.944	0.909	0.944	0.926	0.926	0.971
100	l2	saga	250	0.948	0.905	0.948	0.926	0.926	0.970
100	elasticnet	saga	10	0.948	0.905	0.948	0.926	0.926	0.970
50	elasticnet	saga	10	0.956	0.909	0.956	0.931	0.932	0.969

The results obtained using the Logistic Regression method along with the TF-IDF technique on Dataset 2 reveal a precision ranging from 0.931 to 0.969, while recall oscillates between 0.844 and 0.931. Specificity also exhibits variability, with values ranging from 0.928 to 0.972. Accuracy, on the other hand, falls within a range of 0.906 to 0.943. The F1 score and AUROC follow a similar trend, with values from 0.902 to 0.944 and 0.973 to 0.983, respectively.

In Dataset 2, the selection of hyperparameter sets and the vectorization technique had a notable impact on the results of AUROC, precision, recall, specificity, accuracy, and F1-score metrics for each trained Logistic Regression model. In the BoW vectorization technique, the hyperparameter configuration (C=10, penalty='l1', solver='saga') yielded the highest AUROC of 0.982. This set of hyperparameters allowed the model to effectively handle data variability. Conversely, with the TF-IDF vectorization technique, an even higher AUROC of 0.983 was achieved with the same hyperparameter configuration. In this case, TF-IDF outperformed BoW, suggesting that both the presence and frequency of words throughout the corpus are more predictive for this specific problem.

3.3 Results of Random Forest Models

Table 7 presents the results of the top 10 models based on the area under the ROC curve using the Random Forest algorithm along with Bag of Words (BoW) on Dataset 1. The obtained values reveal that the model exhibits varying capacity to identify cases of cyberbullying and non-cyberbullying. The recall demonstrates the model's reasonable ability to identify the majority of true positives, i.e., actual cases of cyberbullying. On the other hand, although moderate, specificity indicates the model's ability to recognize true negatives or situations that do not constitute cyberbullying. The model's effectiveness in both aspects is crucial for balanced performance, aiming to maximize cyberbullying detection while simultaneously minimizing misclassification of harmless content. The observed variability in these metrics suggests that fine-tuning hyperparameters can have a significant impact on the model's ability to differentiate between classes.

Table 7. Results obtained by the Top 10 Prediction Models using the Random Forest Technique with BoW on Dataset 1.

Estimators	Criterion	min samples split	min samples leaf	Random State	Precision	Recall	Specificity	Accuracy	F1 Score	AUROC
200	entropy	10	1	42	0.689	0.774	0.618	0.700	0.729	0.796
500	entropy	10	1	250	0.685	0.788	0.604	0.700	0.733	0.794
200	entropy	10	1	10	0.691	0.788	0.615	0.705	0.736	0.793
100	entropy	10	1	250	0.683	0.782	0.604	0.697	0.730	0.790
100	gini	10	1	42	0.693	0.772	0.627	0.703	0.730	0.788
500	gini	2	3	42	0.722	0.731	0.692	0.712	0.727	0.787
100	gini	10	1	10	0.689	0.761	0.624	0.696	0.723	0.774
500	gini	30	5	10	0.712	0.731	0.677	0.705	0.722	0.772
200	gini	50	10	250	0.719	0.699	0.700	0.700	0.709	0.752
400	entropy	10	10	42	0.702	0.678	0.686	0.682	0.690	0.751

The results obtained using the Random Forest algorithm along with the TF-IDF technique on Dataset 1 reveal precision ranging from 0.719 to 0.750, while recall varies between 0.664 and 0.694. Specificity ranges from 0.709 to 0.750. Accuracy, on the other hand, falls within the range of 0.689 to 0.715. The F1 score and AUROC exhibit values from 0.693 to 0.715 and 0.760 to 0.795, respectively.

The comparison between BoW and TF-IDF does not yield a clear superiority of one technique over the other in this dataset, with some values favoring BoW and others favoring TF-IDF. Concerning the AUROC metric, the difference between both techniques is minimal, with BoW slightly outperforming TF-IDF. According to the results of Dataset 1, it can be asserted that the choice of hyperparameters and the vectorization method is crucial in the performance of the model obtained with the Random Forest technique. In general, both BoW and TF-IDF have demonstrated relatively similar performances in this classification task. However, upon closer analysis, key differences can be identified, especially when considering how certain hyperparameters influence performance metrics.

The obtained results indicate that, in both approaches, hyperparameters such as the number of estimators, the criterion, and sampling parameters have a notable impact on the metrics. For instance, changes in these hyperparameters can lead to variations in the model's ability to correctly identify cyberbullying cases (recall) and adequately distinguish non-cyberbullying cases (specificity). This analysis suggests that, although both methods have comparable capabilities overall, the choice and adjustment of specific hyperparameters are crucial to optimizing the model's performance in different aspects of classification. This highlights the importance of careful hyperparameter selection to enhance the model's effectiveness in specific tasks.

Now, the results obtained using Dataset 2 are presented. Table 8 displays the outcomes of the top ten models employing BoW, organized according to the area under the ROC curve.

Table 8. *Results obtained by the Top 10 Prediction Models using the Random Forest Technique with BoW on Dataset 2.*

Estimators	Criterion	min samples split	min samples leaf	Random state	Precision	Recall	Specificity	Accuracy	F1 Score	AUROC
500	entropy	10	1	250	0.949	0.928	0.948	0.937	0.938	0.973
100	entropy	10	1	250	0.949	0.928	0.948	0.937	0.938	0.972
200	entropy	10	1	10	0.949	0.928	0.948	0.937	0.938	0.972
200	entropy	10	1	42	0.949	0.928	0.948	0.937	0.938	0.971
500	gini	2	3	42	0.946	0.863	0.948	0.904	0.902	0.971
100	gini	10	1	10	0.946	0.931	0.944	0.937	0.938	0.970
100	gini	10	1	42	0.945	0.924	0.944	0.933	0.934	0.970
500	gini	30	5	10	0.942	0.803	0.948	0.873	0.867	0.964
400	entropy	10	10	42	0.935	0.708	0.948	0.824	0.806	0.923
200	gini	50	10	250	0.952	0.685	0.964	0.821	0.797	0.917

In the case of Dataset 2, it is evident how the combination of selected hyperparameters and vectorization techniques plays a crucial role in improving metrics. It is noteworthy that the Bag of Words (BoW) vectorization method demonstrated robust performance using 500 estimators, entropy as the criterion, a min_samples_split of 10, and min_samples_leaf of 1, achieving the highest AUROC of 0.973. This configuration enabled the model to handle data variability, ensuring a sufficient number of trees to stabilize predictions. However, the TF-IDF vectorization technique outperformed BoW in terms of performance. With only a difference in the number of estimators, using a value of 200, TF-IDF achieved an even higher AUROC of 0.978. This suggests that, for this specific problem, the TF-IDF approach, considering both word frequency in individual documents and throughout the corpus, serves as a stronger predictor.

3.4 Results of Language Model Prediction Models

Table 9 displays the outcomes of the experiments conducted on the colombian-spanish-cyberbullying-classifier language model proposed in this article. This model was derived through fine-tuning of roberta-base-bne using Dataset 1. In this case, the results are encouraging. It is noteworthy that the values achieved with this approach significantly surpassed those obtained using traditional machine learning algorithms on the same dataset. Machine learning algorithms such as Logistic Regression, Random Forests, and XGBoost yielded AUROC scores of 0.797, 0.796, and 0.785, respectively, while the best language model reached an AUROC of 0.910. This remarkable improvement stems from the inherent ability of transformer-based models to capture and analyze the complete context of sentences.

The effectiveness of transformers in this context is attributed to their bidirectional nature, enabling them to consider the entire context of each word in a sentence, both preceding and following. This is crucial for understanding nuances and connotations in language, essential for the precise detection of cyberbullying. The capability to comprehend context and semantics in a deeper and more comprehensive manner distinguishes transformer-based models from more traditional machine learning methods, which may lack such sophistication in natural language processing. The implementation of transformers, therefore, represents a significant advancement in the accurate and effective identification of cyberbullying.

Table 9. *Results obtained using Language Models on Dataset 1.*

Learning rate	Weight decay	Train batch size	Train loss	Eval Loss	Precision	Accuracy	F1 score	AUROC
0.00001	0.01	8	0.478	0.416	0.859	0.836	0.845	0.910
0.00002	0.02	16	0.511	0.396	0.882	0.836	0.840	0.909
0.00003	0.01	16	0.486	0.387	0.847	0.824	0.834	0.908
0.00002	0.01	8	0.453	0.531	0.832	0.816	0.828	0.902
0.00003	0.02	8	0.458	0.525	0.843	0.823	0.834	0.901
0.00003	0.02	32	0.570	0.403	0.807	0.820	0.840	0.900
0.00001	0.02	16	0.559	0.410	0.859	0.820	0.827	0.897
0.00005	0.01	16	0.458	0.438	0.790	0.810	0.833	0.895
0.00002	0.01	32	0.601	0.418	0.802	0.810	0.830	0.894
0.00005	0.02	8	0.456	0.434	0.796	0.799	0.819	0.882

Some of the false negatives presented by traditional machine learning models in Dataset 1, but correctly identified as cyberbullying by the proposed colombian-spanish-cyberbullying-classifier model, include phrases like “Al menos abrí el link de la noticia y leé, bobo hp, analfabeta” (“At least open the link to the news and read, stupid jerk, illiterate”, “Le pagan por ser un estupido de tiempo completo?” (“Do they pay you to be a full-time idiot?”), and “Otra hija de puta que trae veneno en su alma” (“Another bitch carrying venom in her soul”). While machine learning techniques rely on the presence or absence of specific words or word combinations to detect cyberbullying, language models take into account the context and meaning of the entire phrase. This is because these models are trained on vast text corpora and can capture the subtleties of language, including irony, sarcasm, and double entendre, which are often common in cyberbullying.

Table 10 displays the results obtained with the language model using Dataset 2. The achieved values are excellent, even surpassing the Logistic Regression model, which obtained an AUROC of 0.983. As observed, the best language model resulting from the fine-tuning process with Dataset 2 achieved an AUROC of 0.996, indicating an outstanding ability to distinguish between Cyberbullying and Non-Cyberbullying classes.

***Table 10.** Results obtained using Language Models on Dataset 2.*

Learning rate	Weight decay	Train batch size	Train loss	Eval Loss	Precision	Accuracy	F1 score	AUROC
0.00005	0.02	8	0.183	0.154	0.954	0.968	0.968	0.996
0.00005	0.01	16	0.211	0.138	0.942	0.961	0.961	0.996
0.00002	0.01	8	0.187	0.161	0.953	0.959	0.958	0.995
0.00003	0.01	16	0.223	0.155	0.957	0.962	0.962	0.995
0.00003	0.02	8	0.181	0.220	0.939	0.957	0.957	0.994
0.00002	0.02	16	0.255	0.132	0.968	0.966	0.966	0.994
0.00001	0.01	8	0.221	0.159	0.953	0.962	0.962	0.993
0.00003	0.02	32	0.341	0.155	0.960	0.957	0.956	0.992
0.00001	0.02	16	0.323	0.158	0.949	0.959	0.958	0.991
0.00002	0.01	32	0.396	0.142	0.945	0.953	0.952	0.991

3.5. Development of a Web Application for Cyberbullying Detection

As part of this research, the AI Cyberbullying Detector application was developed, designed for use by mental health professionals and created based on the requirements provided by the occupational therapist involved in the project. Following the therapist's recommendation, the application stores additional information about the author of each tweet, such as gender, sexual orientation, age group, socio-economic stratum, disabilities, whether they were a victim of conflict, and if they have a support network. The purpose of this is to enable the analysis of different sociodemographic components related to cyberbullying. This will facilitate informed decision-making and the development of effective strategies in the fight against cyberbullying.

The AI Cyberbullying Detector application is developed in Python and uses the Flask library to create a web server that manages a PostgreSQL database. It provides a user interface allowing the storage of tweets manually entered by the healthcare professional. The application offers various user interaction paths, including adding, deleting, and editing tweets, as well as bulk data loading through a CSV file. This file should not only contain tweets from a group of people to be analyzed but also their corresponding sociodemographic information. According to the therapist's recommendations, it is advisable for the tweets stored in the database, as well as the associated sociodemographic information or the data analyzed from the CSV file, to come from a diverse set of individuals rather than a single person. This diversity will facilitate robust comparative analyses, leading to more effective campaigns in the fight against cyberbullying.

In addition to providing an interface to interact with the tweet database, the application also offers text analysis and visualization functionalities. Tweets are preprocessed using natural language processing techniques and vectorized using a pre-trained model, providing an indication of whether the tweet contains cyberbullying or not. The application also features a visualization interface based on Dash for exploring the data. This interface displays bar charts and a table of the data, making it easy for users to explore relationships between the sociodemographic components mentioned earlier and view the results of cyberbullying analysis.

4. Conclusions

In this study, machine learning techniques and language models were employed to detect whether a tweet is associated with cyberbullying in the Colombian population. Among the machine learning algorithms tested, the model obtained with Logistic Regression achieved an area under the ROC curve of 0.797 in dataset 1. Meanwhile, the AUROC of models obtained with Random Forests and XGBoost was 0.796 and 0.785, respectively. In dataset 2, the Logistic Regression technique achieved an AUROC of 0.983, while Random Forests and XGBoost achieved AUROC values of 0.978 and 0.971, respectively. These results demonstrate that Logistic Regression, especially when combined with TF-IDF vectorization, offers greater capability in identifying cyberbullying on Twitter compared to Random Forests and XGBoost techniques.

However, the results highlight the challenge faced by machine learning techniques in dataset 1, as they struggle to capture the context in which words associated with cyberbullying are used. To address this, a fine-tuning process was applied to the masked language model based on transformers called roberta-base-bne. This led to the proposal of two language models in this article, achieving an AUROC of 0.910 in dataset 1 and 0.996 in dataset 2.

The models proposed in this article not only achieve similar precision and AUROC values to those reported in other studies but, in some cases, surpass them. For instance, while the model proposed by León-Paredes et al. (2019) achieved a precision of 0.93 using Support Vector Machines, this article exceeded that margin with 0.983 using Logistic Regression in dataset 2. Similarly, Khan & Qureshi (2022) reported a precision of 0.939 with Logistic Regression, whereas Balakrishnan et al. (2020) reached an AUC-ROC of 0.97. All these values are surpassed by the best machine learning models and, to a greater extent, by the language models proposed in this article. However, it should be noted that the performance of the models may vary due to various factors such as the quality and quantity of training data, language complexity, hyperparameter selection, and tuning, among others.

Finally, it is important to mention that the proposed models have been made available to healthcare professionals through a web application. The AI Cyberbullying Detector application allows the use of the artificial intelligence models proposed in this research through a user-friendly graphical interface, providing additional features to aid therapists in cyberbullying analysis. This is a key aspect since literature findings do not enable the use of proposed models by healthcare professionals.

5. References

l-garadi, M.A.; Varathan, K.D.; Ravana, S.D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, 63, pp. 433–443. https://doi.org/10.1016/j.chb.2016.05.051

Aránguez Sánchez, T. (2022). Límites legales a la libertad de expresión en Twitter. Un análisis feminista. *Algoritmos, teletrabajo y otros grandes temas del feminismo digital*, pp. 249–267. ISBN 978-84-1122-494-9

Balakrishnan, V.; Khan, S.; Arabnia, H. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security*, 90. https://doi.org/10.1016/j.cose.2019.101710

Bozyiğit, A.; Semih, U.; Nasiboğlu, E. (2021). Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, 179. https://doi.org/10.1016/j.eswa.2021.115001

Chia, Z.; Ptaszynski, M.; Masui, F.; Leliwa, G.; Wroczynski, M. (2021). Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. *Information Processing & Management*, 58(4). https://doi.org/10.1016/j.ipm.2021.102600

Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.1810.04805

Feijóo, S.; Foody, M.; O'Higgins Norman, J.; Pichel, R.; Rial, A. (2021). Cyberbullies, the Cyberbullied, and Problematic Internet Use: Some Reasonable Similarities. *Psicothema*, 33(2), pp. 198–205. https://doi.org/10.7334/psicothema2020.209

Guerra, F. (2023a). Colombian-spanish-cyberbullying-classifier. Disponible en: https://huggingface.co/FelipeGuerra/colombian-spanish-cyberbullying-classifier

Guerra, F. (2023b). Colombian-spanish-cyberbullying-detector. Disponible en: https://huggingface.co/FelipeGuerra/colombian-spanish-cyberbullying-detector

Guerra, F. (2023c). Colombian_Spanish_Cyberbullying_Dataset_1. Disponible en: https://huggingface.co/datasets/FelipeGuerra/Colombian_Spanish_Cyberbullying_Dataset_1

Guerra, F. (2023d). Colombian_Spanish_Cyberbullying_Dataset_2. Disponible en: https://huggingface.co/datasets/FelipeGuerra/Colombian_Spanish_Cyberbullying_Dataset_2

Gutiérrez, A.; Armengol, J.; Pàmies, M.; Llop, J.; Silveira, J.; Pio, C.; Armentano, C.; Rodriguez, C.; Gonzalez, A.; Villegas, M. (2022). MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68. https://doi.org/10.26342/2022-68-3

Hassan, S.A.; Khairalla, M.A.; Fakhrou, A. (2023). The crime of cyberbullying and its relationship to addiction to social networking sites: A study at the law college Prince Mohammad Bin Fahd University. *Computers in Human Behavior Reports*, 12. https://doi.org/10.1016/j.chbr.2023.100346

Herrera-López, M.; Romera, E.; Ortega-Ruiz, R. (2017). Bullying y cyberbullying en Colombia; coocurrencia en adolescentes escolarizados. *Revista Latinoamericana de Psicología*, 49(3), pp. 163–172. https://doi.org/10.1016/j.rlp.2016.08.001

Johari, N.; Jaafar, J. (2022). A Malay Language Cyberbullying Detection Model on Twitter using Supervised Machine Learning. *International Visualization, Informatics and Technology Conference (IVIT)*, Kuala Lumpur, Malaysia, IEEE, pp. 325–332. https://doi.org/10.1109/IVIT55443.2022.10033395

Kee, D.; Anwar, A.; Vranjes, I. (2024). Cyberbullying victimization and suicide ideation: The mediating role of psychological distress among Malaysian youth. *Computers in Human Behavior*, 150. https://doi.org/10.1016/j.chb.2023.108000

Khan, S.; Qureshi, A. (2022). Cyberbullying Detection in Urdu Language Using Machine Learning. *International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)*, Lahore, Pakistan, IEEE, pp. 1–6. https://doi.org/10.1109/ETECTE55893.2022.10007379

León-Paredes, G.; Palomeque, W.; Gallegos, P.; Vintimilla, P.; Bravo, J.; Barbosa, L.; Paredes, M. (2019). Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language. *IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, Valparaíso, Chile, IEEE, pp. 1–7. https://doi.org/10.1109/CHILECON47746.2019.8987684

Marín-Cortés, A.; Franco-Bustamante, S.; Betancur-Hoyos, E.; Vélez-Zapata, V. (2020). Miedo y tristeza en adolescentes espectadores de cyberbullying. Vulneración de la salud mental en la era digital. *Revista Virtual Universidad Católica del Norte*, 61, pp. 66–82. https://doi.org/10.35575/rvucn.n61a5

Peláez, G.; Lena, F. (2021). Árboles de decisión y bosques aleatorios en sistemas expertos: un enfoque fundamental. *Advances in education, ICT and innovation: issues for business and social enhancing*, Madrid. https://doi.org/10.14679/1243

Salehi, F.; Abbasi, E.; Hassibi, B. (2019). The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems*, 32. https://doi.org/10.48550/arXiv.1906.03761

Takuya, A.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.48550/arXiv.1907.10902

Van Hee, C.; Jacobs, G.; Emmery, C.; Desmet, B.; Lefever, E.; Verhoeven, B.; De Pauw, G.; Daelemans, W.; Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLoS ONE*, 13(10): e0203794. https://doi.org/10.1371/journal.pone.0203794

Van Hee, C.; Verhoeven, B.; Lefever, E.; De Pauw, G.; Daelemans, W.; Hoste, V. (2015). *Guidelines for the Fine-Grained Analysis of Cyberbullying*, version 1.0. Technical Report LT3 15-01, LT3, Language and Translation Technology Team – Ghent University.

Wang, L.; Jiang, S.; Zhou, Z.; Fei, W.; Wang, W. (2024). Online disinhibition and adolescent cyberbullying: A systematic review. *Children and Youth Services Review*, 156. https://doi.org/10.1016/j.childyouth.2023.107352

Wang, W.; Chen, L.; Thirunarayan, K.; Sheth, A. (2014). Cursing in English on twitter. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14)*, Association for Computing Machinery, New York, NY, USA, pp. 415–425. https://doi.org/10.1145/2531602.2531734

Zhang, X.; Tong, J.; Vishwamitra, N.; Whittaker, E.; Mazer, J.; Kowalski, R.; Hu, H.; Luo, F.; Macbeth, J.; Dillon, E. (2016). Cyberbullying detection with a pronunciation based convolutional neural network. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 740–745. https://doi.org/10.1109/ICMLA.2016.0132