ISSN print 0716-0356
ISSN online 2452-5995
ISSN-L 0716-0356

TRILOGÍA

CIENCIA · TECNOLOGÍA · SOCIEDAD

ARTÍCULOS

UTEM
UNIVERSIDAD TECNOLÓGICA METROPOLITANA
del Estado de Chile

DICIEMBRE 2024

Vol. 40 • Nº 51

Óscar Magna*
Universidad Tecnológica Metropolitana, Santiago de Chile.

https://orcid.org/0000-0002-0361-3553

Igor Bustos Zurita**
Universidad Tecnológica Metropolitana, Santiago, Chile

Artículo

SUPERVISED CLASSIFIER MODEL TO IDENTIFY HATE SPEECH AND PERFORM SENTIMENT ANALYSIS IN TEXTS. USE CASE: YOUTUBE, REDDIT, AND TWITTER NETWORKS.

MODELO DE CLASIFICACIÓN SUPERVISADA PARA LA DETECCIÓN DE DISCURSO DEL ODIO EN TEXTOS. CASO DE USO: YOUTUBE, REDDIT Y TWITTER.

Recibido: 4 de octubre de 2024 | Versión final:09 de diciembre de 2024| Publicado: 31 de diciembre de 2024

Cómo citar este artículo:
Magna, O.; Bustos, I. (2024). Supervised classifier model to identify hate speech and perform sentiment analysis in texts. Use case: YouTube, Reddit, and Twitter networks. Trilogía (Santiago), 40(51), 31-51 . Santiago de Chile: Ediciones UTEM.

* Dr. & MBA, Civil Engineer in Computer Science, Department of Informatics and Computing, Universidad Tecnológica Metropolitana (UTEM), Santiago, Chile.
Correo electrónico: omagna@utem.cl
** Ingeniero Civil en Computación mención Informática, Universidad Tecnológica Metropolitana (UTEM).
Correo electrónico:igor.bustosz@utem.cl

ABSTRACT

This study presents the development and evaluation of a supervised classification model for detecting hate speech on social media platforms, specifically YouTube and Reddit. The research addresses the growing proliferation of harmful content in digital environments, proposing an automated solution based on advanced Natural Language Processing (NLP) techniques and machine learning.

The model is based on a Transformer (BERT) architecture, designed to analyze and classify various manifestations of hate speech, including defamation, threats, and advocacy. The "HateGuard" database of 15215 comments was built using web scraping techniques from YouTube and Reddit, complemented with records from the HateVal dataset from X (Twitter), and implemented with a seven-category multilabel tagging scheme to differentiate various forms of hate speech. Data preprocessing included tokenization, lemmatization, and cleaning of non-textual elements.

To improve performance and address the class imbalance, data augmentation techniques were applied, including OCR (Optical Character Recognition) to simulate errors, grapheme spelling to generate spelling errors, and backtranslation. In the initial training, the model achieved an accuracy of 82% and an F1-score of 0.83. In the comparative evaluation with models such as RoBERTa, it demonstrated a significant improvement, with an increase of 16 points in the F1-score compared to other models. The final F1-score for the hate speech classifier in HateGuard was 0.88, outperforming previous datasets such as HatEval and HaterNet. In contrast to HaterBERT, which achieved an F1-score of 0.99 in binary detection. The proposed model stands out for its multilabel classification capability, addressing more complex forms of hate speech.

The analysis of results revealed limitations, including a higher incidence of false positives in ambiguous comments and lower performance in identifying subtle threats, with a recall rate of 0.78 in this category.

This study significantly contributes to the automation of content moderation, offering a competitive and adaptable solution. Future research should focus on expanding the dataset to mitigate class imbalance and explore hybrid approaches that incorporate unsupervised learning to capture more subtle forms of hate speech. The proposed model shows considerable potential for real-time application on social platforms, contributing to improving safety and respect in digital environments

Key words: Hate Speech, Natural Language Processing, Sentiment Analysis, Machine Learning

RESUMEN

Este estudio presenta el desarrollo y evaluación de un modelo de clasificación supervisada para la detección de discurso de odio en plataformas de redes sociales, específicamente YouTube, Reddit y X (ex Twitter). La investigación aborda la creciente proliferación de contenido nocivo en entornos digitales, proponiendo una solución automatizada basada en técnicas avanzadas de Procesamiento del Lenguaje Natural (NLP) y aprendizaje automático.

El modelo se fundamenta en una arquitectura Transformer (BERT), diseñada para analizar y clasificar diversas manifestaciones de discurso de odio, incluyendo difamación, amenazas y apología. La base de datos "HateGuard" de 15215 comentarios se construyó mediante técnicas de web scraping desde YouTube y Reddit, complementados con registros del conjunto de datos HateVal de X (ex Twitter) e implementados con esquema de etiquetado multilabel de siete categorías para diferenciar las diversas formas de discurso de odio. El preprocesamiento de datos incluyó tokenización, lematización y limpieza de elementos no textuales.

Para mejorar el rendimiento y abordar el desequilibrio de clases, se aplicaron técnicas de aumento de datos, incluyendo OCR (Reconocimiento Óptico de Caracteres) para simular errores, grapheme spelling para generar errores ortográficos, y backtranslation.

En el entrenamiento inicial el modelo alcanzó una precisión del 82% y un F1-score de 0.83. En la evaluación comparativa con modelos como RoBERTa demostró una mejora significativa, con un aumento de 16 puntos en el F1-score respecto a otros modelos. El F1-score final para el clasificador de discurso de odio en HateGuard fue de 0.88, superando el rendimiento de conjuntos de datos anteriores como HatEval y HaterNet. En contraste con HaterBERT, que logró un F1-score de 0.99 en detección binaria, el modelo propuesto destaca por su capacidad de clasificación multilabel, abordando formas más complejas de discurso de odio.

El análisis de resultados reveló limitaciones, incluyendo una mayor incidencia de falsos positivos en comentarios ambiguos y un rendimiento inferior en la identificación de amenazas sutiles, con una tasa de recall de 0.78 en esta categoría.

Este estudio contribuye significativamente a la automatización de la moderación de contenidos, ofreciendo una solución competitiva y adaptable. Investigaciones futuras deberían enfocarse en ampliar el conjunto de datos para mitigar el desequilibrio de clases y explorar enfoques híbridos que incorporen aprendizaje no supervisado para capturar formas más sutiles de discurso de odio. El modelo propuesto presenta un potencial considerable para su aplicación en tiempo real en plataformas sociales, contribuyendo a mejorar la seguridad y el respeto en entornos digitales.

Palabras clave: Discurso de odio, Procesamiento de lenguaje natural, Análisis de Sentimiento, Machine Learning

1. INTRODUCTION

Hate speech on social media represents a critical challenge for contemporary digital coexistence. Defined as a form of communication that promotes, incites, or justifies discrimination, this phenomenon has proliferated significantly on online platforms (Nockleby, 2000; Davidson et al., 2017; Schmidt & Wiegand, 2017). The widespread nature of this type of discourse has been driven by growing participation in social networks, threatening the safety and well-being of users (Yi-Le Chan et al., 2023; Guiora & Park, 2017).

The identification and delimitation of hate speech presents significant challenges due to its subjective and contextual nature (Davidson et al., 2017; Schmidt & Wiegand, 2017). To effectively address this issue, it's essential to understand the context, identify the recipient, and analyze the sentiment associated with the message (Watanabe et al., 2018).

The detection and reduction of online hate speech requires the use of advanced techniques such as natural language processing (NLP) and machine learning (Qasim et al., 2022; Chiril et al., 2022). Adapting sentiment analysis to specific contexts has become an imperative need in this field (Watanabe et al., 2018).

Numerous researchers have addressed the phenomenon from various perspectives. Serrano and Díaz (2022) examined the interrelationship between misinformation and hate on social platforms. Ruiz and Sánchez (2018) mapped the dynamics of virilization of hate messages on Twitter. Da Cunha and Girona (2021) investigated machine learning techniques to identify hate content in Spanish. Fortuna and Nunes (2018), as well as Schmidt and Wiegand (2017), provided comprehensive reviews of NLP and machine learning techniques for this challenge.

The psychological and social impacts of hate speech have also been subjects of study. Gómez and Fernández (2020) evaluated the negative effects of online harassment and hate on mental health. Leracitano et al. (2024) and Castaño-Pulgarín and Suárez-Betancur (2021) addressed these implications, highlighting the need for a comprehensive perspective.

In the regulatory field, Martínez and Pérez (2019) analyzed legal frameworks and content moderation policies employed by digital platforms. Rajendran et al. (2024) offered systematic analyses of detection techniques and analysis of hate content. MacAvaney et al. (2019), Watanabe et al. (2018), and Mozafari et al. (2020) explored the use of deep learning models and knowledge transfer approaches.

A significant advance has been the adoption of Transformer-based architectures, such as BERT (Devlin et al., 2019). Plaza-del-Arco et al. (2021) and Castillo-López et al. (2022) explored the potential of BERT and other pre-trained models for hate detection in different linguistic contexts. Croce et al. (2020) proposed the GAN-BERT model to improve performance on limited datasets, while Conneau et al. (2019) introduced RoBERTa, an optimized version of BERT.

In the context of Spanish, Plaza del Arco et al. (2021) evaluated multilingual models such as XLM and mBERT. Zhang et al. (2018) pointed out the need for research on Spanish variants in hate speech detection. Philippy et al. (2023) and Castillo-López et al. (2022) explored cross-linguistic approaches and their extension to variants within the same language.

Despite advances, there are still pending challenges. More comprehensive approaches are required that consider the intersection of hate speech with other forms of discrimination and cultural and linguistic diversity. In addition, greater collaboration between academia, in-

dustry, and regulatory bodies is necessary to develop more effective and ethical solutions. In this context, the main objective of this study is the development of a supervised classifier model and the construction of a functional prototype that makes predictions about the categorization of messages as 'Hate speech' with an accuracy level above 80%. This work aims to contribute to the development of effective tools for the automated identification of hate speech on social media platforms such as YouTube, Reddit, and Twitter, thereby contributing to the creation of safer and more respectful digital environments.

The importance of this study lies in its potential to address a critical problem in contemporary digital society. Online hate speech not only affects specific individuals and communities but also undermines the principles of mutual respect and peaceful coexistence in the digital space. By developing a supervised classifier model with high accuracy, this study seeks to provide a valuable tool for early detection and mitigation of hate speech on widely used social media platforms.

The hypothesis underlying this research is that, through the application of advanced NLP and machine learning techniques, it's possible to develop a multilabel classifier model capable of identifying hate speech in social networks with an accuracy of over 80%. This hypothesis is based on recent advances in the field of AI and NLP, as well as promising results obtained in previous studies on automated detection of offensive content online.

Specific objectives include the development of a supervised classifier model, the construction of a functional prototype, the evaluation of performance on different platforms (YouTube, Reddit, and X or ex Twitter), and the analysis of its effectiveness in detecting various forms of hate speech, considering linguistic and cultural variations.

The distinctive feature of this study is its comprehensive approach, combining advanced NLP techniques with a deep analysis of the social and linguistic dynamics of online hate speech. By addressing multiple platforms and considering linguistic variations within Spanish, this research seeks to provide a robust and adaptable solution to different digital contexts. Finally, this work aims to contribute significantly to the field of automated hate speech detection, offering not only a technically advanced model but also valuable insights into the manifestations and propagation of hate in digital environments. The results have the potential to inform the development of more effective policies for online content moderation and promote a more inclusive and respectful digital culture.

2. MATERIALS AND METHODS

2.1. Dataset design

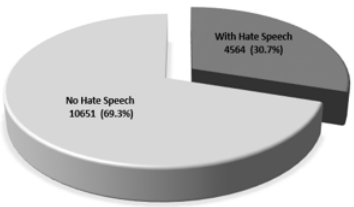
The dataset design process is divided into three phases: planning and data collection, preprocessing, and model development. This systematic structure ensures a comprehensive approach that addresses the inherent challenges in detecting hate speech in digital environments.

The requirements and technologies necessary for model development were defined in the planning and collection phase. Data collection was carried out using web scraping techniques, covering various sources: YouTube and Reddit comments and Twitter comments extracted from the HatEval-2019 dataset. This multi-platform approach ensures a broad representation of the different manifestations of hate speech in the digital ecosystem.

The final dataset created and named "Hateguard" comprises a total of 15215 comments (Figure 1), of which 4,564 (30.7%) are labeled as hate speech and 10651 (69.3%) do not have this label. This distribution in the social networks here studied reflects the phenomenon prevalence and provides a solid basis for model training.

Figure 1. Number of comments in Dataset

Dataset	Total	%	with HS	%	Non HS	%
YouTube	4048	26,6%	1316	28,8%	2732	25,65%
Reddit	4659	30,6%	1062	23,3%	3597	33,77%
Twitter - actual X (HatEval)	6508	42,8%	2186	47,9%	4322	40,58%
HateGuard	15215	100,00%	4564	30,70%	10651	69,30%



Source: Author's elaboration

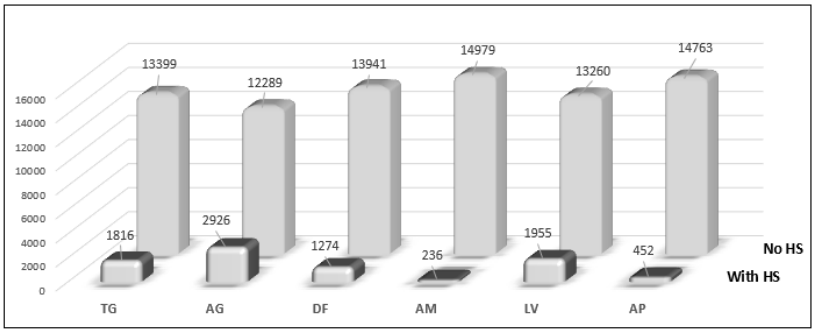
The labeling process is fundamental for messages analysis and classification , as it allows for the identification of terms and expressions that indicate negative concepts. The target identification is essential to recognize references or expressions that can harm specific individuals or groups. Furthermore, this process is crucial for protecting affected individuals and promoting responsibility among content creators.

The dataset includes additional labels such as Defamation (DF), Threat (AM), Vulgar Language (LV), and Apology (AP), which expands the characteristics of text related to hate speech. The data collection is performed using APIs provided by the mentioned social platforms, ensuring efficient access to relevant comments.

Once the data is collected, the preprocessing stage begins, where tools like Python are used to clean the texts by removing irrelevant elements such as HTML tags and usernames. Lemmatization is also applied to reduce words to their base form, and stopwords are removed to avoid noise in the analysis. This process is essential to ensure the data is consistent and relevant for model training.

The exploratory data analysis (EDA) revealed significant patterns in comments labeled as hate speech (Figure 2). On YouTube, 28.8% of 4048 comments were identified as hate speech, with racism being the most prevalent category. On Reddit, 23.3% of 4659 comments were classified as hate speech, also with a high prevalence of racism. The HatEval dataset (Twitter) showed 47.9% of comments labeled as hate speech, with an emphasis on vulgar and aggressive language. This quantitative analysis reveals not only the magnitude of the problem but also the specific nature of offensive content present on studied platforms.

Figure 2. Proportion of labels in the final dataset.



Source: Author's elaboration

To address the imbalance in labels of the 'Hateguard' dataset, data augmentation techniques were implemented. These included: assigning weights to classes during training, artificial increase through techniques such as OCR to simulate typographical errors, and backtranslation to five different languages and subsequent translation back to Spanish.

These techniques allowed the dataset to be expanded to 39,102 labeled comments, significantly improving the representativeness and balance of the dataset.

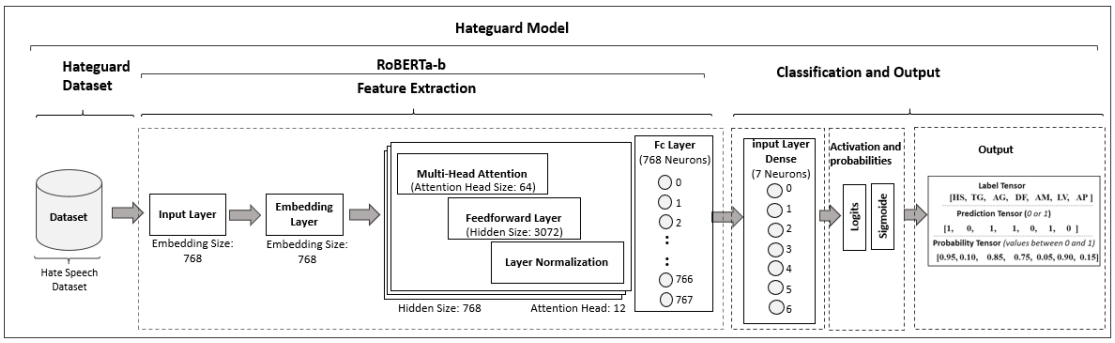
The model development was carried out in Google Colab, using libraries such as Pytorch and Scikit-learn. A model based on RoBERTa-b was implemented, an advanced architecture specifically designed for NLP tasks. The choice of this architecture is based on its demonstrated ability to effectively process text sequences and generate accurate predictions in natural language classification tasks (Acheampong et al., 2020; Cem ÖZKURT, 2024).

The model training was performed using a split of the dataset into training (80%) and test (20%) sets. Standard metrics such as precision, recall, and F1-score were applied to evaluate the model's performance. The training process involved preparing the labeled dataset and configuring the pre-trained model "PlanTL-GOB-ES/Roberta-base-bne", specialized in processing Spanish text

1. Prototype Architecture

The architecture of prototype (Figure 3) is based on the RoBERTa-b model, designed to process text sequences and generate accurate predictions in natural language processing tasks. This model operates in three main stages: dataset input, feature extraction, and classification.

Ilustración 3. Prototype Architecture.



Source: Author's elaboration

The model employs a Classifier with 7 neurons to assign values to class labels, followed Output Layer that generates final predictions. Using a Sigmoid function, it compresses values between 0 and 1, addressing a multilabel classification problem with independent predictions and a 0.5 threshold for determining outcomes. This process not only produces effective metrics but also provides information about the nature of analyzed content, thus contributing to accurate identification and a deeper understanding of hate speech in digital environments.

2. Training

The training of the hate speech classification model was meticulously designed to optimize its ability to identify and categorize offensive language across diverse social media contexts. The process encompassed phases of preprocessing, configuration, training, and optimization. A carefully prepared Final Dataset was used, split into 80/20 proportions for training and testing. The model was based on PlanTL-GOB-ES/Roberta-base-bne, specialized in Spanish, complemented with RobertaTokenizerFast and configured for seven-label classification. Hyperparameters were adjusted using Training Arguments, with emphasis on epochs and warmup steps. A prediction threshold of 0.5 was established for binary labels, aligned with the multilabel design. Evaluation metrics included precision, recall, F1-score, and accuracy, prioritizing the F1-score as an overall indicator.

The experimentation phase was crucial, conducting exhaustive tests with different parameter combinations. This systematic approach allowed for fine-tuning the learning process and improving the detection of subtle patterns in hate speech.

The ultimate goal was to develop a robust and sensitive model capable of effectively addressing the complex task of detecting hate speech in the dynamic environment of social media.

The most salient results of these experiments revealed significant patterns (Figure 4). It was observed that number of epochs showed a positive correlation with F1-Score for Hate Speech (HS) detection. In addition, training with four epochs generally produced better F1 scores for Aggressiveness (AG), a key metric due to its hierarchical relationship with the Threat (AM) and Vulgar Language (LV) labels.

Figure 4. Experiment Results

Evaluation		F1 Score							Epochs	warmup steps
Loss	Accuracy	HS	TG	AG	DF	AM	LV	AP		
0.166	0.765	0.877	0.821	0.841	0.699	0.882	0.806	0.627	4	700
0.163	0.766	0.871	0.821	0.840	0.678	0.908	0.806	0.562	4	700
0.167	0.758	0.869	0.820	0.839	0.728	0.870	0.802	0.640	4	700
0.163	0.788	0.858	0.832	0.815	0.705	0.787	0.819	0.659	5	700
0.173	0.756	0.855	0.812	0.787	0.675	0.756	0.795	0.551	4	600
0.175	0.769	0.850	0.828	0.807	0.726	0.874	0.808	0.641	5	600

Source: Author's elaboration
Note: HS: Presence of Hate Speech, TG: Target, AG: Aggressiveness, DF: Generalization or Slander, AM: Threat, LV: Vulgar Language, AP: Justification.

The results table showed interesting variations in performance metrics. For example, with 4 epochs and 700 warmup steps, a Loss of 0.166, Accuracy of 0.765, and F1Scores of 0.877 for HS, 0.841 for AG, 0.882 for AM, and 0.806 for LV were obtained. In contrast, increasing to 5 epochs while maintaining 700 warmup steps showed a slight improvement in Accuracy (0.788), but a decrease in the F1Score for HS (0.858) and AG (0.815).

After a thorough evaluation of these results, the selection of the final model was primarily based on the F1Score for Hate Speech, considered the most critical criterion. Although the second experiment showed generally superior metrics, the first experiment (4 epochs, 700 warmup steps) was selected as the final model due to its better overall performance, particularly in the apology category, and its consistency across other metrics.

This rigorous process of experimentation and selection ensured that the final model was not only highly accurate in detecting hate speech but also robust and balanced in its performance across all classification categories. The methodology employed, which combined careful data preparation, informed hyperparameter selection, and comprehensive evaluation of multiple configurations, resulted in a model capable of effectively addressing the complex task of identifying and classifying hate speech in diverse linguistic and social contexts.

3. Results

The model achieves an overall accuracy of 77% and F1 scores above 60% for all labels, demonstrating its effectiveness in identifying various aspects of hate speech (Figure 5).

Figure 5: Best results obtained

Metric	Identifier	Description	Score
F1 score	accuracy	Accuracy	0.77
	HS	Presence of hate speech	0.88
	TG	Target	0.82
	AG	Aggressiveness	0.84
	DF	Defamation	0.70
	AM	Threats	0.88
	LV	Vulgar language	0.81
	AP	Apology	0.63

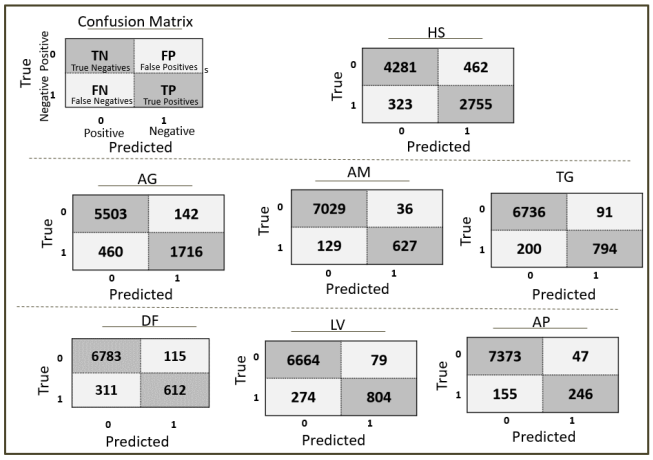
Source: Author's elaboration

The hate speech (HS) label shows one of the highest F1 scores at 0.88, indicating the model's strong ability to correctly identify and classify

texts related to hate speech. Other labels such as trolling (TG), aggressiveness (AG), threats (AM), and vulgar language (LV) also exhibit high scores, suggesting the model's competence in these categories. However, the F1 scores for defamation (DF) and apology for hate speech (AP) are lower, at 0.70 and 0.63 respectively, indicating greater difficulty in correctly classifying these labels. This may be due to insufficient or unrepresentative training data for these categories.

The confusion matrices for each label provide additional information on the model's performance (Figure 6), highlighting the model's effectiveness in terms of true positives and true negatives, while revealing a tendency to misclassify comments directed at individuals as if they were directed at generic groups. Despite this, both false positives and true positives indicate good classification capabilities

Figure 6. Confusion Matrix by Hate Speech category



Source: Author's elaboration

The label "Aggressiveness" (AG) demonstrates solid performance in both positive and negative predictions, attributable to a large number of instances and its hierarchical position under the label "HS." However, the number of false positives suggests the need for adjustments to reduce incorrect predictions of aggressiveness in non-aggressive contexts.

The confusion matrix for the label "Threats" (AM) shows a better capacity of the model to predict false positives than true positives, due to this label's dependence on "HS." Regarding the label "Vulgar Language" (LV), the model exhibits a high ability to discern the absence of such language, being more precise when this label is 0, due to data imbalance.

Despite efforts to balance the dataset by increasing positive "HS" instances, the "Apology of Hate Speech" (AP) label remains skewed. While the model has a low false positive rate, its high false negative rate indicates a need for further refinement in detecting apologies. Nevertheless, it is deemed a suitable classifier for identifying hate speech.

In addition to commonly used metrics such as precision and recall, the study includes specificity to provide a more comprehensive assessment of model performance (Figure 7). Specificity is particularly useful for complementing the information provided by other metrics in imbalanced datasets.

The results show that the model performs well in identifying true negatives, with specificity values ranging from 85.2% to 94.5% for different labels.

Figure 7. Predictions and associated probabilities

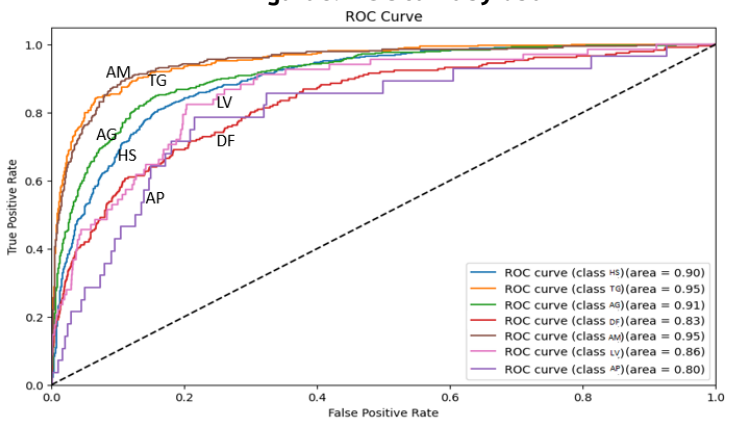
Label	Specificity	Accuracy	Recall
HS	85.20%	89.80%	93.50%
TG	87.30%	98.30%	96.80%
AG	91.80%	97.30%	91.90%
DF	89.90%	99.10%	94.50%
AM	94.50%	99.50%	98.20%
LV	89.20%	98.60%	95.90%
AP	86.60%	99.60%	97.30%

Source: Author's elaboration

Precision, which evaluates the proportion of correct predictions within the predicted class, varies between 89.80% and 99.60%, highlighting the model's ability to accurately predict labels of interest. Recall, which indicates the proportion of actual label examples that are correctly predicted, ranges from 91.90% to 98.20%, underscoring the model's skill in correctly identifying positive cases.

The ROC curve, which illustrates the classification performance for each label, plots the true positive rate against the false positive rate as the classification threshold varies (Figure 7), revealing that the HS label exhibits an AUC of 0.90, indicating strong performance in overall hate speech detection.

Figure 8. ROC curve by label



Source: Author's elaboration

The TG and AM labels demonstrate exceptional performance with an AUC of 0.95, suggesting that the model effectively identifies the target of hate speech and the presence of threats. The AG and LV labels achieve AUCs of 0.91 and 0.86 respectively, implying moderate performance in detecting aggressive and vulgar language. The DF and AP labels exhibit the lowest AUC values, 0.83 and 0.80 respectively, indicating challenges in identifying defamatory content and justifications for hate speech.

During classification evaluation, the model's prediction for a given test text is assessed, with a probability threshold of 50% used as the decision boundary. A label is predicted to be present if its corresponding probability exceeds 50%; otherwise, it is considered absent. Figure 9 provides an example to verify the model's predictions against ground truth labels.

According to these predictions, the model indicates that text contains hate speech and a defamatory generalization towards the target with a high probability.

Figura 9. Predictions and associated probabilities

Label	Prediction	Probability	
HS	1	94.1%	Hate Speech: Indicates whether the text contains any form of hate speech according to the established theoretical framework.
TG	0	2.4%	Target: Specifies the target of hate speech: 0 if it is a collective group and 1 if it is a specific individual.
AG	0	3.3%	Aggression: Indicates whether a comment is aggressive (contains insults, contempt, or hostility.
DF	1	97.4%	Defamation: Indicates whether there is a negative generalization about the target (false statement or slander directed at a target.
AM	0	1.0%	Threat: Indicates whether the text suggests any kind of violence against the target, compromising their integrity.
LV	0	0.6%	Vulgar Language: Indicates whether the text contains any vulgar language (e.g., rude words.
AP	0	6.6%	Apology: Indicates whether the text presents any form of apology (justification of hate speech

Source: Author's elaboration

In addition, there is a low probability that the label target refers to a collective group. For the rest of labels, the probability of aggressiveness, threats, vulgar language, or hate speech justifications is low and predicted as non-existent. According to these predictions, the model indicates that text contains hate speech and defamatory generalization towards the target with high probability. In addition, it presents a low probability for labeling targets referring to a collective group. For the rest of the labels, the probability of aggressiveness, threats, vulgar language, or hate speech justifications is low and predicted as non-existent.

The predictions made by the model are correct according to the basis for labeling data since the text contains a racist-type hate message about immigrants, referring to Venezuelans in plural as a group of people, and defamation for spreading false statements about Venezuelans, the president of Venezuela and the UN. Regarding the other labels, the text does not present any aggressiveness or justifications to defend its position.

Discussion and Conclusions

The hate speech detection model developed in this study demonstrated highly significant results, outperforming benchmarks in the automated detection of such content. The model achieved an overall accuracy of 77% and an F1-score of 88% in hate speech (HS) detection, establishing it as a robust and effective solution. These results are particularly impressive in the categories of threats and aggression, where F1 scores reached 88% and 84%, respectively, showcasing superior capability in identifying potentially harmful content. This underscores the effectiveness of implementing the RoBERTa

architecture and leveraging the custom "HateGuard" dataset.

Figure 10 corroborates these findings by presenting a side-by-side comparison of the model with other state-of-the-art studies and datasets in the field. As illustrated, the proposed model outperforms previous approaches, such as the SVM by Basile et al. (2019) and the BETO model by Plaza del Arco et al. (2021), which achieved F1scores of 0.73 and 0.77, respectively. The methodology employed in this study, combining advanced preprocessing techniques and cutting-edge NLP architectures, offers a significant edge in Spanish hate speech detection, surpassing other approaches with F1-scores ranging from 0.76 to 0.87. This validates the model's robustness in complex and multilingual contexts.

Figure 10. Comparison of HS metrics for different models and Datasets

Dataset	Author	Model	F1 HS
Hateval	Basile et al, 2019	SVM	0.73
Haternet	Pereira et al, 2019	LSTM+MLP	0.61
Haternet	Valle-Cano, 2021	HaterBERT	0.99
Haternet Hateval	Aluru et al, 2020	mBERT	0.73
Haternet	Plaza del Arco et al, 2021	BETO	0.77
Haternet	Plaza del Arco et al, 2021	BETO	0.78
Hateval	Perez et al, 2021	RoBERTuito	0.76
HasCoVa-2022	Castillo-Lopez et al., 2022	BETO	0.85
HasCoVa-2022	Castillo-Lopez et al., 2022	mBERT	0.73
		RNN+LSTM	0.78
HateGuard	Current research	BETO	0.87
		RoBERTa	0.88

Source: Author's elaboration

It's crucial to note that alongside RoBERTa, we evaluated other architectures, including BETO and recurrent neural networks (RNNs) with LSTM layers, using the HatEval and HaterNet datasets. While these models showed promise, our RoBERTa-based model trained on HateGuard consistently outperformed them across all experiments, demonstrating the robustness and efficacy of our approach. However, the model exhibited some limitations, particularly in identifying defamation (F1 score of 0.70) and apologia for hate speech (F1 score of 0.63), indicating areas for refinement. In addition, it remains difficult to capture the nuances of discourse, such as sarcasm and irony, which require more sophisticated semantic analysis.

The potential of the model transcends the detection of hate speech. Its applications include content moderation on social media, online reputation protection, and the promotion of a safer and more respectful digital environment. Furthermore, this work establishes a solid foundation for future research in natural language processing, opening new opportunities for the development of tools such as sentiment analysis, fake news detection, and other related challenges.

In conclusion, this work represents a significant advancement in the automatic detection of hate speech. The developed model, in combination with the HateGuard dataset, offers an effective and scalable solution for identifying and mitigating hate speech on social media. The results obtained are encouraging and present new perspectives for future research. However, it is advisable, first and foremost, to improve the dataset and model by considering various strategies to enhance a detection model, especially for minority classes.

At the data level, it is suggested to increase diversity using techniques such as synonyms, paraphrasing, and generating synthetic data with models like GPT-3. Data balancing through sampling is also mentioned. At the model level, adjustments are proposed such as modifying thresholds, applying L2 regularization, assembling multiple models, and optimizing hyperparameters. The goal is to obtain a more robust and accurate model.

Additionally, exploring more advanced models that integrate figurative language analysis and enrich datasets with more complex examples is recommended. Investigating the combination of automated techniques and human supervision could also improve classification in ambiguous cases.

BIBLIOGRAPHY

Aluru, S; Mathew, B; Saha, P. and Mukherjee A (2020). *"Deep Learning Models for Multilingual Hate Speech Detection". arXiv*, 16 pages DOI:10.48550/arXiv.2004.06465. https://doi.org/10.48550/arXiv.2004.06465

Acheampong Francisca Adoma; Nunoo-Mensah Henry et al. (2020). *Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition.* 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). Publisher: IEEE , DOI: 10.1109/ICCWAMTIP51612.2020.9317379, IEEE Xplore: 15 January 2021.

Basile V., Bosco C., Fersini E., Nozza D., Patti V., et al(2019). "SemEval-2019 Task 5 : Multi lingual Detection of Hate Speech Against Immigrants and Women in Twitter". *In Proceedings of the 13th International Workshop on Semantic Evaluation*., Pages 54-63. DOI:10.18653/v1/S19-2007. https://doi.org/10.18653/v1/S19-2007

Camacho, D; Panizo-LLedot and Bello-Orgaz, G. (2020). *"The four dimensions of social network analysis: An overview of research methods, applications, and software tools".* Elsevier, Volume 63, November 2020, Pages 88-120. DOI: 10.1016/j.inffus.2020.05.009. https://doi.org/10.1016/j.inffus.2020.05.009

Canete, J; Chaperon, G; Fuentes, R; Ho, J; Kang, H. and Pérez, J. (2020). *"Spanish Pre-Trained BERT Model and Evaluation Data." In PML4DC at ICLR*. DOI:10.48550/arXiv.2308.02976. https://doi.org/10.48550/arXiv.2308.02976

Castaño-Pulgarín, SA. and Suárez-Betancur, N. (2021). *"Internet, social media and online hate speech. Systematic review".* Aggression and violent, Elsevier, Volume 58, May–June 2021, 101608. DOI: 10.1016/j.avb.2021.101608. https://doi.org/10.1016/j.avb.2021.101608

Castillo-López, G., Riabi, A. and Seddah, D. (2022). "Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection". *Association for Computational Linguistic*s, May 2023, pages 1-13.DOI: 10.18653/v1/2023.vardial-1.1. https://doi.org/10.18653/v1/2023.vardial-1.1.

Cem ÖZKURT (2024). Comparative Analysis of State-of-the-Art Q\&A Models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 Dataset. Research Square, February 2024, DOI:10.21203/rs.3.rs-3956898/v1 file:///C:/Users/osema/Downloads/Comparative_Analysis_of_State-of-the-Art_QA_Models.pdf

Chiril, P., Pamungkas, E.W., Benamara, F. et al(2022). Emotionally Informed Hate Speech Detection: A Multi-target Perspective. *Cogn Comput* 14, 322–352. DOI: 10.1007/s12559-021-09862-5. https://doi.org/10.1007/s12559-021-09862-5

Conneau A., Khandelwal K., Goyal N.(2019)"Unsupervised Cross-lingual Representation Learning at Scale". *arXiv preprint arXiv*. DOI:10.48550/arXiv.1911.02116. https://doi.org/10.48550/arXiv.1911.02116

Croce, Danilo, Castellucci, Giuseppe and Basili, Roberto (2020). "GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, 2114-2119. DOI: 10.18653/v1/2020.acl-main.191. https://aclanthology.org/2020.acl-main.191/.

Cunha, I. and Girona, R. (2021). "Detección Automática de Discurso de Odio en español en Redes Sociales". Investigan técnicas de aprendizaje automático para identificar y monitorear contenidos de odio en línea.

Davidson, T; Warmsley, D; Macy, M; and Weber, I. (2017) "Automated Hate Speech Detection and the Problem of Offensive Language". *In proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 512-515. DOI:10.1609/icwsm.v11i1.14955. https://doi.org/10.1609/icwsm.v11i1.14955

Devlin J., Chang M., Lee K. and Toutanova K.(2019). "BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding".*arXiv preprint arXiv*. DOI:10.48550/arXiv.1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Fortuna, P. and Nunes, S. (2018). "A survey on automatic detection of hate speech in text". ACM Computing SurveysVolume 51 Issue, 4 Article No: 85, pp 1–30. DOI: 10.1145/3232676. https://doi.org/10.1145/3232676

Gitari N.D.; Zuping Z.; Damien H. and Long J. (2015) "A lexicon-based approach for hate speech detection". *International Journal of Multimedia and Ubiquitous Engineering*, Vol.10, pp.215-230 DOI:10.14257/ijmue.2015.10.4.21 . http://dx.doi.org/10.14257/ijmue.2015.10.4.21

Gómez, M. and Fernández, C. (2020). "Impacto Psicológico del Discurso de Odio en Usuarios de Redes Sociales .Evalúan los efectos negativos del acoso y el discurso de odio en la salud mental de los usuarios.

Gomez, R; Gibert, J. and Gomez L. (2020). "Exploring hate speech detection in multimodal publications". Proceedings, openaccess.thecvf.com. DOI: 10.48550/arXiv.1910.03814 . https://doi.org/10.48550/arXiv.1910.03814

Guiora, Amos and Park, Elizabeth (2017). "Hate Speech on Social Media". Philosophia 45, 957–971 (2017). DOI:10.1007/s11406-017-9858-4. https://doi.org/10.1007/s11406-017-9858-4

Gutiérrez, A; Armengol, J; Pàmies, M; Llop, J. and Silveira, J. (2022). "MarIA: Spanish Language Models". Procesamiento del Lenguaje Natural, [S.l.], v. 68, p. 39-60, mar. 2022. ISSN 1989-7553. DOI:10.26342/2022-68-3. https://doi.org/10.26342/2022-68-3

Leracitano, F; Balenzano, C. and Girard, S. (2023). "Online hate speech as a moral issue: Exploring moral reasoning of young italian users on social network sites." Social Science Computer Review, 42(1), 25-47. DOI:10.1177/08944393231161124. https://doi.org/10.1177/08944393231161124

Jahan, MS. and Oussalah, M. (2023). "A systematic review of Hate Speech automatic detection using Natural Language Processing". Neurocomputing, Volume 546, 14 August 2023, 126232 . DOI: 10.1016/j.neucom.2023.126232. https://doi.org/10.1016/j.neucom.2023.126232

Kennedy, CJ; Bacon, G; Sahn, A. and von Vacano, C. (2020). "Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application." *arXiv preprint arXiv*,35 pages.DOI:10.48550/arXiv.2009.10277 https://doi.org/10.48550/arXiv.2009.10277

MacAvaney, A; Yao, HR; Yang, E; Russell, K. and Goharian, N. (2019). "Hate speech detection: Challenges and solutions". PloS one. DOI: 10.1371/journal.pone.0221152. https://doi.org/10.1371/journal.pone.0221152

Martínez, A. and Pérez, J. (2019). "Regulación del Discurso de Odio en Plataformas Digitales". Analizan el marco legal y las políticas de moderación de contenidos empleadas por las redes sociales.

Matamoros-Fernández, A. and Farkas, J. (2021). "Racism, hate speech, and social media: A systematic review and critique". Television & new media, journals.sagepub.com, Volume 22, Issue 2. DOI: 10.1177/1527476420982230. https://doi.org/10.1177/1527476420982230

Mathew, B; Dutt, R; Goyal, P. and Mukherjee, A.(2019). "Spread of hate speech in online social media". Proceedings of the 10th ACM on Web Science, Pages 173–182. DOI: 10.1145/3292522.3326034. https://doi.org/10.1145/3292522.3326034

Min X, Lin H, Li X, Zhao H, Lu J, Yang L and Xu B (2023). "Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective". Information Fusion,Volume 96, Pages 214-223. DOI: 10.1016/j.inffus.2023.03.015. https://doi.org/10.1016/j.inffus.2023.03.015

Mondal, M; Silva, LA. and Benevenuto, F.(2017). "A measurement study of hate speech in social media". Conference on hypertext and social, Pages 85–94. DOI: 10.1145/3078714.3078723. https://doi.org/10.1145/3078714.3078723

Mozafari, M; Farahbakhsh, R. and Crespi, N.(2019)."A BERT-based transfer learning approach for hate speech detection in online social media". COMPLEX NETWORKS, 2019, 8. Conference paper, pp 928–940. DOI: 10.1007/978-3-030-36687-2_77. https://doi.org/10.1007/978-3-030-36687-2_77

Nockleby, JT. (2000). Hate Speech in Context: The Case of Verbal Threats. *Buffalo Law Review*, Volumen 42 Número 3, Artículo 2, 10-1-1994. https://digitalcommons.law.buffalo.edu/buffalolawreview/vol42/iss3/2/

Pereira-Kohatsu JC; Quijano-Sánchez, L; Liberatore, F. and Camacho-Collados, M. (2019). "Detecting and Monitoring Hate Speech in Twitter". *Sensors* 19, no. 21: 4654. DOI: 10.3390/s19214654. https://doi.org/10.3390/s19214654

Pérez, J; Furman, A. Alemany, L. and Luque, F. (2022)."RoBERTuito: a pre-trained language model for social media text in Spanish".*In Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, DOI:10.48550/arXiv.2111.09453. https://doi.org/10.48550/arXiv.2111.09453

Philippy, F; Guo, S. and Haddada, S. (2023). "Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review". Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 5877–5891 July 9-14, 2023. DOI: 10.48550/arXiv.2305.16768. https://doi.org/10.48550/arXiv.2305.16768

Plaza-del-Arco, F; M., Molina-González, MD; Ureña-López, LA. and Martín-Valdivia, MT. (2021). "Comparing pre-trained language models for Spanish hate speech detection." Expert Systems with Applications, Volume 166, 15 March 2021, 114120. DOI: 10.1016/j.eswa.2020.114120. https://doi.org/10.1016/j.eswa.2020.114120

Pyingkodi M, Thenmozhi K, Chitra K, and Karthikeyan M, et al(2023) "Hate Speech Analysis using Supervised Machine Learning Techniques". *In International Conference on Computer Communication and Informatics (ICCCI), Coimbator*e, India, 2023, pp. 1-6, DOI: 10.1109/ICCCI56745.2023.10128591. https://doi.org/10.1109/ICCCI56745.2023.10128591

Qasim R, Bangyal WH, Alqarni MA and Ali Almazroi A. (2022). "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification". *Journal of Healthcare Engineering*, Volume 2022, 17 pages. DOI:10.1155/2022/3498123. https://doi.org/10.1155/2022/3498123

Rajendran, MP; Ramaswamy, K. and Sekar, R. (2024)." Systematic detection and analysis of hate speech in social networking. AIP Conference." AIP Conf. Proc. 3042, 020042 (2024). DOI: 10.1063/5.0194201. https://doi.org/10.1063/5.0194201

Ruiz, A. and Sánchez, D. (2018). "Dinámicas de Propagación del Discurso de Odio en Twitter". Estudian cómo se viralizan y amplifican los mensajes de odio en esta red social.

Schmidt A. and Wiegand, M. (2017). "A survey on Hate Speech Detection using Language Processing". In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pages 1–10, DOI: 10.18653/v1/W17-1101. https://doi.org/10.18653/v1/W17-1101

Serrano, J. and Díaz, JA. (2022). "Desinformación y Discurso de Odio en Redes Sociales". Este estudio analiza cómo se propaga el discurso de odio y la desinformación a través de plataformas como Facebook, Twitter e Instagram.

Su, X; Li, Y; Branco, P and Inkpen, D. (2023)"SSL-GAN-RoBERTa: A robust semi-supervised model for detecting Anti-Asian COVID-19 hate speech on social media". Natural Language Engineering, Published online 2023:1-20. DOI:10.1017/S1351324923000396. https://doi.org/10.1017/S1351324923000396

Valle Cano, Gloria (2021). "Detección de mensajes de odio en Twitter: un estudio basado en perfiles dentro de la red social". *Universidad Autónoma de Madrid*. http://hdl.handle.net/10486/697828

Vaswani, A., Shazeer, N., Parmar, N. and Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, N., Polosukhin, I. (2017)" Attention Is All You Need". *Advances in neural*, 15 Pages. DOI: 10.48550/arXiv.1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Watanabe H., Bouazizi M. and Ohtsuki T. (2018). "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *in IEEE Access*, vol. 6, pp. 13825-13835, 2018, doi: 10.1109/ACCESS.2018.2806394. https://doi.org/10.1109/ACCESS.2018.2806394

Waseem, Zeerak and Hovy, Dirk (2016). "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter". *In Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics. DOI: 10.18653/v1/N16-2013. https://aclanthology.org/N16-2013/.

Yi Le Chan, Jireh; Thye Bea, Khean, Mun Hong Leow, Steven, Wai Phoong, Seuk and Khuen Cheng, Wai (2023). "State of the art: a review of sentiment analysis based on sequential transfer learning". *Artificial Intelligence Review* (2023), 56:749–780, DOI: 10.1007/s10462-022-10183-8. State of https://link.springer.com/content/pdf/10.1007/s10462-022-10183-8.pdf.

Zampieri M.; Malmasi S.; Nakov P.; Rosenthal S.; Farra N.; Kumar R. (2019). "Predicting the type and target of offensive posts in social media". *arXiv preprint arXiv*. DOI:10.48550/arXiv.1902.09666. https://doi.org/10.48550/arXiv.1902.09666

Zhang, Z. and Luo, L. (2019). "Hate speech detection: A solved problem? the challenging case of long tail on twitter". Semantic Web, vol. 10, no. 5, pp. 925-945, 2019. DOI: 10.3233/SW-180338. https://doi.org/10.3233/SW-180338

Zhang Z; Robinson D. and Tepper J. (2018) "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network". *In the Semantic Web*. ESWC 2018. Lecture Notes in Computer Science, vol 10843, June 2018. DOI:10.1007/978-3-319-93417-4_48. https://doi.org/10.1007/978-3-319-93417-4_48.

FdDRS (2023). Fuentes de Datos – Redes Sociales
a. Antecedentes:
1. YouTube:
Los videos presentados provienen de fuentes confiables de noticias en YouTube, las pertenecen a Mega Noticias, Biobío, 24 Horas - TVN Chile y Tele 13. Estas cuentas son conocidas por su integridad periodística y sin sesgo.

2. Reddit:
Los comentarios provenientes de Reddit son extraídos de comentarios a publicaciones más populares respecto al ámbito buscado. Estas publicaciones son publicadas en el apartado oficial del subreddit del país y pueden presentar sesgo en alguna de estas.

3. Hateval:
Se Accede a HateEval mediante el sitio web Github con comentarios de la red social Twitter (X), el cual contiene los elementos que componen el modelo, durante el año 2023.

b. Fuentes de Datos
1. YFDRS_Xv (2023). YouTube. Fuentes de Datos – Redes Sociales de Xenofobia (venezolanos).
- Jóvenes, venezolanos y con antecedentes: ¿Quiénes son los sospechosos del crimen de Carabinero? (2023), https://www.youtube.com/watch?v=7FO91r8tXXM
- ¿Podría haber consecuencias en Chile? Gobierno de Venezuela da golpe a cúpula del Tren de Aragua (2023) , https://www.youtube.com/watch?v=IRJustxxOzA
- Viuda de venezolano abatido por carabinero cuestiona procedimiento (2023) , https://www.youtube.com/watch?v=vyjoOofOrgk
- Largas filas en Embajada de Venezuela por masiva entrega de pasaportes (2023), https://www.youtube.com/watch?v=_32Eoj2QwUA

2. YFDRS_Xh (2023). YouTube. Fuentes de Datos – Redes Sociales de Xenofobia (haitianos). Haitianos argumentan que racismo los motiva a salir de Chile (2023), https://www.youtube.com/watch?v=2Yv3wq9lXfo

3. YFDRS_Xp (2023). YouTube. Fuentes de Datos – Redes Sociales de Xenofobia (peruanos).
- Inmigrantes: peruanos y colombianos los nuevos habitantes del centro (2017), https://www.youtube.com/watch?v=r12OZ5GzfRE
- Perú impide ingreso de migrantes desde Chile: Llevan días en la frontera (2023), https://www.youtube.com/watch?v=jVDUbO30AxE

4. YFDRS_Mf (2023). YouTube. Fuentes de Datos – Redes Sociales de Misoginia (feminismo).
- Marcha feminista del 8M se toma las calles de Santiago (2024) https://www.youtube.com/watch?v=HZsG4JzGhoI

5. YFDRS_ML (2023). YouTube. Fuentes de Datos – Redes Sociales de Grupo Minoritario (LGBT).
- Miles de PERSONAS participan de la MARCHA del ORGULLO en BUENOS AIRES (2023), https://www.youtube.com/watch?v=cSNv5jW-mCQ
- Historias del Orgullo LGBTQ+ en Argentina (2023), https://www.youtube.com/watch?v=BYavB_eWfkU
- Multitudinaria marcha por el Orgullo LGBTIQ+ (2023), https://www.youtube.com/watch?v=UI23ZqvEigY

RDRS_ML (2023). Reddit. Fuentes de Datos – Redes Sociales de Grupo Minoritario (LGBT).
- Militantes de Massa LGBT agreden y acosan sexualmente a militante de Milei (2023), https://www.reddit.com/r/argentina/comments/17tb1bi/militantes_de_massa_lgbt_agreden_y_acosan/
- Proyecto de ley prohíbe discriminación de género (pub, bares, discos)(2017), https://www.reddit.com/r/chile/comments/703ndj/proyecto_de_ley_prohibe_discriminacion_de_genero/
- El feminismo se movilizó después del paso (2023), https://www.reddit.com/r/argentina/comments/1608uew/comment/jxmjwr0/
- Feministas en Tetas bailando VS Libertarios capítulo 1 (2023) , https://www.reddit.com/r/argentina/comments/16ksm29/feministas_en_tetas_bailando_vs_libertarios/
- En la marcha del 25 ¿se soro roraron con la votante de Milei que le rompieron la napia? (2023), https://www.reddit.com/r/argentina/comments/18590n3/en_la_marcha_del_25_se_sorororaron_con_la_votante/

RFDRS_Mf (2023). Hateval. Fuentes de Datos – sitio web GitHub.

- Elementos que componen el modelo, durante el año 2023. https://github.com/cicl2018/HateEvalTeam/blob/master/Data%20Files/Data%20Files/%232%20Development-Spanish-A/train_es.tsv

UNIVERSIDAD TECNOLÓGICA METROPOLITANA
del Estado de Chile
UTEM

TRILOGÍA

CIENCIA · TECNOLOGÍA · SOCIEDAD

EDICIONES UNIVERSIDAD
TECNOLÓGICA METROPOLITANA