ISSN 1870-4069

A Methodological Framework for Detecting Benevolent Misogynistic Hate Speech in Mexican Political Discourse Using Transformer-Based Models

Monserrat Sánchez-Juárez, Eric Ramos-Aguilar, Daniel Sánchez-Ruiz, Ricardo Ramos-Aguilar

Instituto Politécnico Nacional, Unidad Profesional Interdisciplinaria de Ingeniería
Campus Tlaxcala, Tlaxcala, México
msanchezj1401@alumno.ipn.mx,{eramosa,dsanchezro,rramosa}@ipn.mx

Abstract. This study presents a methodological framework for developing a natural language processing (NLP) model specifically adapted to the Mexican political landscape, aimed at identifying and categorizing misogynistic and sexist hate speech on social media. Despite the widespread presence of such discourse in political environments, there is a marked absence of targeted strategies to address it within this specific domain. The proposed approach involves fine-tuning and training multiple large-scale pre-trained language models—such as MarIA, BETO, BERTIN, mBERT, RoBERTa, and RoBERTuito—using a custom binary-labeled dataset. This dataset is designed to differentiate between benevolent sexist hate speech and neutral or non-sexist language, with particular attention to implicit and coded forms that are often challenging to recognize. By addressing an under explored area in computational political discourse analysis, this work contributes both to academic research and to the development of practical tools that promote safer and more inclusive digital political communication.

Keywords: Sexist Hate Speech, Large Language Model, Natural Language Processing.

1 Introduction

The rise of social media the late decade has already transformed public and political communication by enabling direct and immediate interaction among users. However, this opening has also made easier the discriminatory expression increase among users pointing out misogynistic hate speech, These manifestations not only reinforce harmful gender stereotypes but also perpetuate dynamics of symbolic violence and exclusion, especially against women who actively participate in public life. The United Nations defines hate speech as any kind of communication that attacks or uses pejorative or discriminatory language against a person or group based on inherent identity traits such as gender, race, or religion [24].

pp. 73–85; rec. 2025-06-20; acc. 2025-07-16

Hate speech has historically existed as a tool of exclusion and symbolic violence, used to delegitimize and silence marginalized groups [11]. In political contexts, this phenomenon takes multiple forms, from explicit discriminatory statements to messages loaded with symbolic connotations that reinforce stereotypes related to gender, race, or social class [26].

Misogynistic hate, in particular, manifests in speech that does trivialize, sexualize, or discredit women participating in public life, reinforcing patriarchal power structures [18]. This discursive violence is not limited to formal spaces such as traditional media or parliamentary forums; it has intensified on social media platforms, where anonymity and algorithmic design facilitate its fast spread and normalization [14]. High-profile cases include media violence against figures such as Hillary Clinton in the United States [27], Dilma Rousseff in Brazil [16], and, in Mexico, women like Claudia Sheinbaum the first woman that became a president in Mexico, and Tatiana Clouthier, who have been targeted by systematic smear campaigns with clearly misogynistic overtones [13]. This reality highlights the urgency of analyzing such discourse from intersectional perspectives supported by natural language processing tools.

The reproduction of sexist hate speech has serious consequences at both the individual and societal levels. On a personal level, it can lead to negative psychological impacts on victims, such as anxiety, depression, and low self-esteem, affecting their well-being and limiting their participation in public and professional spheres [8]. On a societal level, sexist hate speech contributes to the perpetuation of gender stereotypes and reinforces structures of discrimination and exclusion, hindering equal opportunities between men and women [3]. Moreover, studies have shown that this type of discourse not only fosters the normalization of gender-based violence, but may also translate into physical assaults and other hate crimes against women [28].

The remainder of this paper is organized as follows: Section 2 presents the context and background of the study, along with a review of related work. Section 3 describes the proposed methodology, including data collection, database creation, text preprocessing, model tuning and training, and the analysis and classification process. Section 4 discusses the evaluation of the results and outlines the expected outcomes. Finally, Section 5 presents the conclusions and suggests directions for future research.

2 Context and Background

In Europe, female politicians are increasingly subject to discrediting tactics. In 2016, 41.8% of European parliamentarians reported having been targeted with humiliating or sexually suggestive images shared on social media [1]. By 2018, that figure had risen to 58.2%, and among respondents under the age of 40, it reached 76.2% [2]. In Spain, and specifically on X formerly Twitter, women from political and communication fields are the focus of 90% of gender-based insults and hate speech. Female politicians in particular receive an average of 15 negative mentions per day [20].

In 2020, the organization “Sapari” and the Media Development Foundation, with support from the United Nations Development Programme, conducted media monitoring of 112 Facebook pages and gathered both quantitative and qualitative sociological data on female politicians in Georgia, USA, regarding sexist hate speech during the pre-election period [25]. The study reviewed prevalent forms of sexist hate speech in Europe and the United States based on European indicators of sexist hate speech, and sought to identify those forms and indicators—presented in Table 1—as adapted to the context of Georgia.

Table 1. Georgian Indicators of Sexist Hate Speech.

Hate Speech Forms	Indicators
Slut-shaming	Reference to a female politician as male politician’s property in any form – because of any male politician referring to women as a "prostitutes"
Spreading sexual photos	Spreading photos or videos of any kind of sexual content without the consent of the female politician
Body-shaming	Mentioning a female body / sexuality, in positive, negative, age-based, having explicitly feminine attire, being beautiful or in any other form
Positive sexism/false compliments	Ironical mocking, masked with political correctness, any kind of compliment that goes beyond work and refers to the visual and representation of a female politician.
Gender role of women	Any attempt to refer to a female politician, as a housewife, culinary specialist, or exhibiting behaviors positively reinforced for men.
Reprimand for family and motherhood	Mentioning a female politician’s childbearing, or her children’s behavior in any form, especially when she either has no children at all, or has only one child, or her child has deviant behavior.

One of the least explored forms is benevolent or positive sexism, usually expressed through "compliments" that involve implicit discriminatory undertones and reinforce gender stereotypes under the guise of politeness or affection. In the context of Mexican Spanish, this phenomenon becomes even more complex due to the use of colloquial language and culturally coded forms of expression, such as albur, which is characterized by double meanings and implicit sexual connotations. These expressions pose a particular challenge for automated hate speech detection systems, as they require a deep understanding of cultural context, figurative language, and the subtleties of everyday speech.

3 Related Work

3.1 General Hate Speech Detection

In recent years, artificial intelligence (AI) and natural language processing (NLP) techniques have enabled significant progress in the automated moderation of digital content. Tools such as CommentGuard [9], Hive Moderation [12], Respondology [22], Netino [17], and Sightengine [23] have been primarily developed to moderate offensive content in English, focusing on categories such as abusive language, racism, Islamophobia, and brand protection. However, these systems show substantial limitations in identifying sexist hate speech, especially in languages other than English.

Numerous efforts have been undertaken to combat online hate through the application of computational intelligence techniques. For instance, [6] utilized a dataset of 16,000 tweets annotated by Waseem and Hovy, which included 3,383 sexist tweets, 1,972 racist tweets, and the remainder labeled as non-offensive. They employed models based on Long Short-Term Memory (LSTM) networks and Gradient Boosted Decision Trees (GBDT), achieving an F1-score of 0.930. Similarly, [19] proposed an ensemble of Recurrent Neural Networks (RNNs) using the same dataset, slightly improving performance with an F1-score of 0.932 by incorporating user behavior data.

Furthermore, [6] compared various machine learning approaches for hate speech detection, including Logistic Regression (LR), Support Vector Machines (SVM), and GBDT, concluding that deep learning models such as CNNs and LSTMs outperformed traditional models by 13% to 20%. In a complementary study, [4] used SVM as a baseline and compared it with CNN, CNN + LSTM, GRU, and CNN + GRU architectures, reporting consistent improvements of at least 7% in accuracy when using CNN.

Recent studies have highlighted the superior performance of BERT in hate speech detection tasks. For instance, [21] found that BERT outperformed FastText, CNN, and LSTM significantly. Similarly, [5] reported better results with BERT compared to LSTM and BiLSTM. Additionally, BERT has shown strong results in multilingual tasks [10, 5].

3.2 Sexist Hate Speech in Political Contexts

In a broader political context, [7] applied NLP techniques for emotion detection and text mining on a corpus of over three million tweets. Their findings revealed that messages directed at female politicians tend to exhibit greater emotional polarity, while male politicians receive slightly higher volumes of hate speech.

Specifically targeting sexist content in Spanish, [15] employed the EXIST 2023 corpus, which includes tweets annotated by gender and age group, to identify sexist content, author intent, and the type of sexism. Using transformer-based models and NLP techniques, they achieved an F1-score of 0.854.

Although these studies mark significant progress, the detection of benevolent sexism—a subtle and often implicit form of gender-based discrimination—within Mexican political discourse remains under explored.

3.3 Challenges in Spanish-Language NLP

In the case of Spanish—particularly Mexican Spanish—the development of culturally and linguistically adapted NLP models is still in early stages. Sexist hate speech in Spanish social media often incorporates local expressions, irony, and euphemistic language, which generalist models fail to detect accurately.

Although some studies have addressed hate speech detection in Spanish (e.g., [15]), the literature on benevolent misogynistic discourse in the Mexican political context is notably scarce, highlighting the need for domain-specific approaches. This lack of specialized tools hinders effective content moderation and timely responses to digital gender-based violence, which disproportionately affects women in political and media spheres.

4 Methodological Framework

The methodology proposed is structured into three consecutive phases, detailed below and illustrated in Fig.1

The proposed methodology for hate speech detection consists of the following sequential phases:

1. Data Collecting: Collecting relevant text data from social media.
2. Database creation: Organizing and annotating the collected data to build a labeled custom dataset.
3. Text preprocessing: Cleaning and normalizing the text to prepare it for analysis.
4. Model tuning and training: Adjusting model parameters to detect hate speech.
5. Analysis and classification: Applying the trained models to classify text as hateful or non-hateful.
6. Results evaluation: Evaluating the model's performance using standard metrics like precision, recall, and F1-score.

4.1 Data Collecting

The first phase involves collecting representative textual data of both misogynistic and non-misogynistic speech. Data sources include social media platforms such as X (formerly Twitter) and Facebook, as well as public corpora from existing databases focused on hate speech in Spanish. The data collection process employs the following methods:

– **Web scraping.** Automated extraction of textual content from relevant websites and online forums where political discourse occurs.
– **API usage.** Structured access to large volumes of public posts using the X API, focusing on content generated in Mexico.
– **Preexisting databases.** Integration of publicly available corpora such as the *EXIST 2023* dataset and *HateEval*, which contain manually annotated examples of hate speech in Spanish.

Text selection is guided by specific inclusion and exclusion criteria. Only messages authored in Mexican Spanish that are public, non-redundant, and directed at or referring to Mexican female political figures (e.g., senators, deputies, governors) are considered.

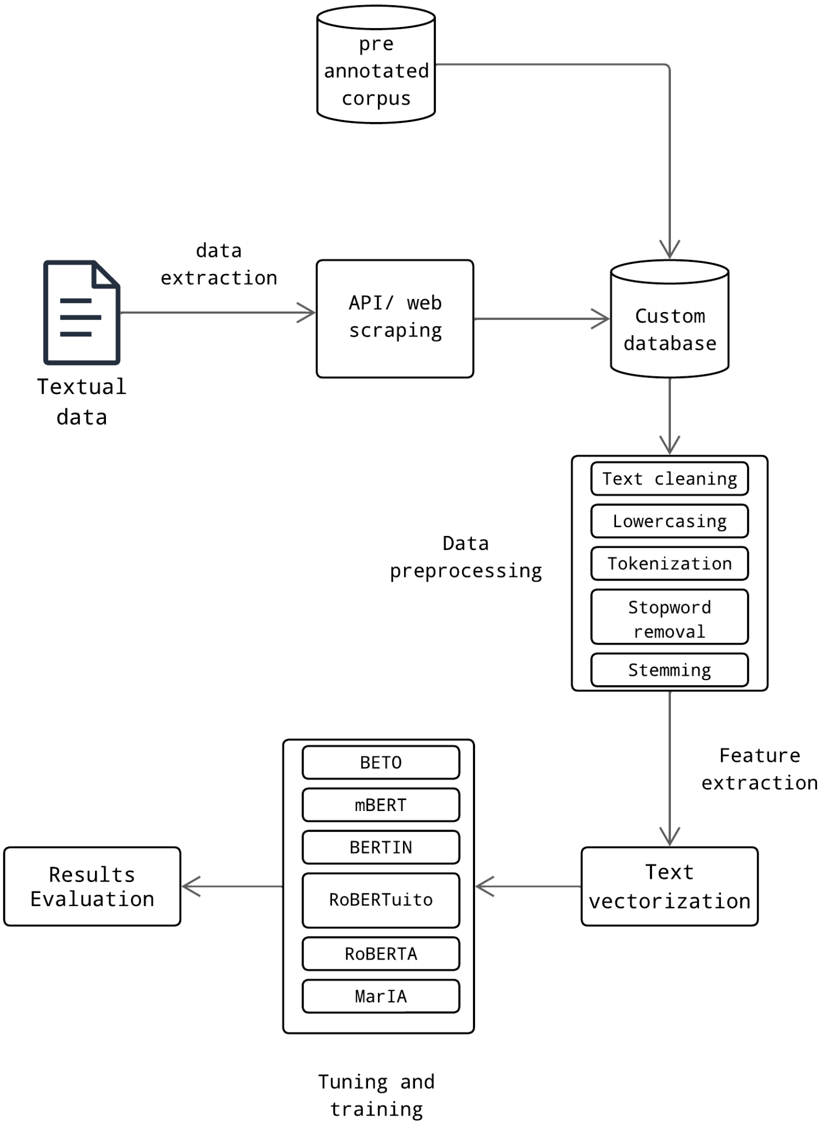

Fig. 1. Proposed methodology for the detection of hate speech in Mexican political discourse.

Examples of inclusion keywords are: *senadora, diputada, candidata, gobernadora, feminista, igualdad, paridad, INE, Congreso, morena, PRI, PAN, oposición, política mexicana.*

The initial dataset is expected to include approximately 2,000 posts from X and Facebook, spanning the years 2023–2024. Manual annotation will be carried out by experts in gender studies and computational linguistics to ensure accurate labeling of each instance as either *benevolent sexist*, or *non-misogynistic* content.

To ensure representativeness and balance, the dataset will be stratified to maintain a proportional distribution between the three classes mentioned above. Sampling strategies will also account for temporal diversity (e.g., during electoral periods) and thematic variety (e.g., policy debates, campaign events, controversies).

This curated dataset will serve as the foundation for the subsequent preprocessing and modeling stages.

4.2 Database Creation

Once the information has been collected, it is structured into a machine-readable format, typically in a CSV file. In this database, each row corresponds to an individual text entry, and the columns contain the text and a binary label: Benevolent Misogynistic Sexist (SBM) or Non-SBM (NSBM).

The binary labeling simplifies the classification task in the initial stage, deferring the application of multi-labels or subcategories to subsequent phases. An example of this dataset is presented in Table 2.

Table 2. Initial sample of the training corpus. SBM: Benevolent Misogynistic Sexist. NSBM: Non-Benevolent Misogynistic Sexist or neutral.

Text	SBM	NSBM
Qué suerte que tenemos una Presirvienta científicA ... @Claudiashein	X	
Vamos a tener un ama de casa 6 años	X	
No se les olvide que López Obrador y Claudia Sheinbaum robaron en el segundo piso		X
La Presirvienta de todos los mexicanos siempre tan cariñosa con la gente	X	

Although hostile language is not employed, benevolent sexist expressions subtly undermine the president by evaluating her through traditional gender roles rather than her professional competence or political actions.

To mitigate subjective bias in the labeling process, multiple annotators will participate in the manual classification of the dataset. Labels will be assigned according to predefined categories: *benevolent sexist*, and *non-misogynistic.* All annotators will do detailed annotation guideline developed in collaboration with a psychologist specialized in gender perspective.

To ensure annotation consistency, a subset of 20% of the corpus will be independently labeled by three annotators. Inter-annotator agreement will be evaluated using Cohen's Kappa coefficient, which quantifies the degree of agreement beyond chance. A Kappa score above 0.75 will be considered indicative of substantial agreement.

In cases of disagreement, a resolution protocol will be followed: conflicting instances will be reviewed and discussed in consensus meetings moderated by the gender expert. Final labels will be assigned based on agreement reached during these sessions.

This multi-annotator strategy, combined with quantitative agreement metrics and expert oversight, aims to enhance the reliability and validity of the labeled dataset.

4.3 Text Preprocessing

The collected texts, characterized by informal language, typographical errors, emojis, symbols, user tags (@), hashtags, and other elements, undergo a series of preprocessing steps to normalize the data and prepare it for analysis. These stages are shown in Fig. 2.

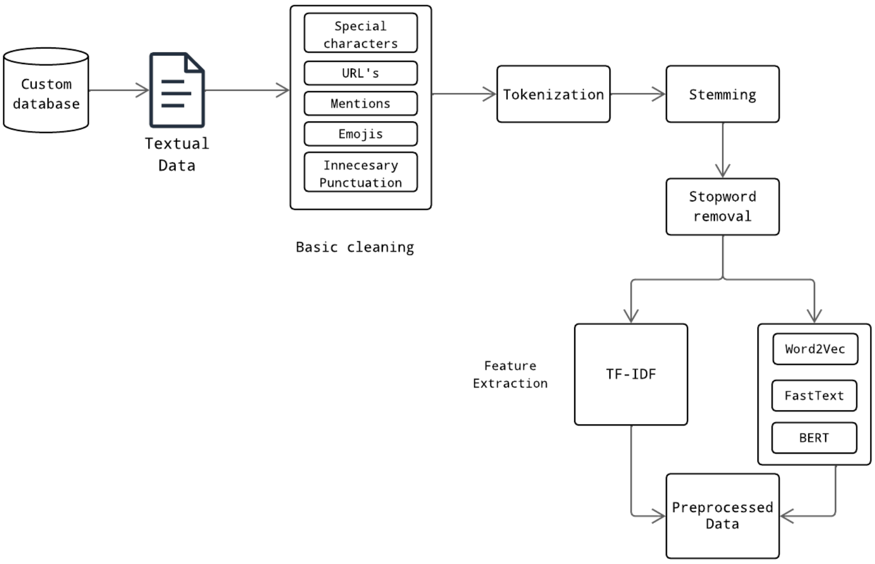

Fig. 2. Text preprocessing stages for textual data.

The preprocessing steps outlined above enable large language models (LLMs) to capture semantic relationships between words, thereby enhancing the predictive performance of the system.

4.4 Model Training and Fine Tuning

Once the texts have been preprocessed and converted into numerical representations, several language models will be trained. The selected models include pretrained Spanish or multilingual transformers that have shown strong performance in similar tasks, such as: BETO, mBERT, BERTIN, RoBERTuito, RoBERTa, MarIA.

These models will be fine-tuned for the binary classification task by adapting their internal parameters to the specific corpus of this project. To avoid overfitting, cross-validation will be applied, and key hyperparameters such as learning rate, number of epochs, batch size, and loss function will be optimized.

4.5 Text Analysis and Classification

The trained model is employed to classify new texts as either *benevolent misogynistic* or *non-misogynistic*. Once predictions are obtained, the outputs are analyzed to interpret the results, identify frequent misclassification patterns, and assess the model's generalization capabilities. To further enhance interpretability, SHAP (SHapley Additive exPlanations) values will be computed to determine the relative contribution of specific terms or phrases to the model's decisions. These explainability techniques provide transparency and support a deeper understanding of the underlying decision-making process.

The analysis of SHAP values will focus on extracting key linguistic patterns that are strongly associated with benevolent misogyny. For instance, words or expressions such as *"ama de casa"*, *"cariñosa"*, or *"débil"*, *"presirvienta"* may emerge as recurrent indicators in positively labeled predictions. These findings will be compiled and synthesized in the form of accessible reports and visual summaries tailored to non-technical audiences. The findings aim to provide policy stakeholders and civil society organizations with actionable linguistic evidence on the presence of benevolent sexism in political communication, facilitating more effective interventions against online gender violence.

However, one of the primary challenges in modeling this task lies in the class imbalance inherent to the dataset, especially given the relatively subtle and less frequent nature of benevolent sexist expressions. To mitigate this issue, multiple strategies will be implemented. First, the Synthetic Minority Over-sampling Technique (SMOTE) will be applied during training to generate synthetic samples of the minority class and balance the dataset distribution. Additionally, experiments with random undersampling of the majority class will be conducted to evaluate potential trade-offs in model performance.

To further support balanced learning, the loss function will incorporate class weighting, assigning greater penalization to errors involving the underrepresented class. Model evaluation will rely on weighted performance metrics, including the weighted F1-score, precision, and recall, to ensure that results reflect both the accuracy and fairness of the classifier. Together, these methods aim to improve the model's ability to detect nuanced patterns of benevolent misogyny while maintaining general robustness.

To evaluate the model's performance, the following standard evaluation metrics are show in Table 3.

Table 3. Evaluation metrics to measure the model's performance.

Metric	**Description**
Precision	The proportion of correctly predicted misogynistic instances among all instances predicted as misogynistic.
Recall	The proportion of actual misogynistic instances that were correctly identified by the model.
F1-score	The harmonic mean of precision and recall, offering a balanced measure of performance.
Confusion Matrix	A representation that displays true positives, false positives, true negatives, and false negatives to visualize the model's overall performance.

5 Anticipated Outcomes and Impact

The proposed methodology is expected to enable the effective classification of political discourse through the use of fine-tuned transformer-based models. These models aim to accurately distinguish between *benevolent misogynistic* and *non-misogynistic* speech within the specific context of Mexican political communication. Leveraging pre-trained Spanish-language models such as BETO, mBERT, and RoBERTuito, combined with rigorous preprocessing and interpretability techniques (e.g., SHAP), is anticipated to yield strong results across standard evaluation metrics such as precision, recall, and F1-score.

Beyond performance metrics, the project seeks to uncover linguistic patterns and implicit expressions of gender bias, addressing a significant gap in the computational analysis of non-hostile forms of sexist speech. The findings are intended to contribute to the development of safer and more inclusive digital environments, particularly for women engaging in political discourse online.

To assess the generalization capacity of the models across Spanish dialects, cross-linguistic transfer experiments will be conducted. Specifically, the model will be evaluated on a subset of 1,000 tweets written in Peninsular Spanish from the EXIST 2023 corpus. Fine-tuning of BETO and RoBERTuito will be performed as needed, allowing for the identification of dialectal variations and necessary linguistic adaptations.

In terms of practical applications, the trained models will be integrated into a prototype content moderation tool for social media platforms such as

X. This tool will enable detection of benevolent misogynistic discourse, offering a foundation for digital interventions. Furthermore, collaboration with public institutions such as the Instituto Nacional Electoral (INE) or INMUJERES is envisioned, particularly for supporting awareness campaigns and policy-making initiatives aimed at combating political gender-based violence.

Ultimately, this work bridges the gap between computational modeling and actionable policy, highlighting the potential of NLP technologies to foster more equitable and respectful political communication online.

6 Conclusions

This study presents a structured methodology for the detection of *benevolent misogynistic discourse* in Mexican political speech using advanced natural language processing techniques. By targeting a form of hate speech that often goes unrecognized—subtle, non-aggressive language that reinforces gender stereotypes—the research addresses a critical and underexplored dimension of online misogyny. The proposed pipeline, which encompasses data collection, preprocessing, model fine-tuning, and interpretability analysis, enables not only accurate classification but also the extraction of valuable insights into the persistence of gender bias within digital narratives.

The outcomes of this work are intended to inform the development of analytical and moderation tools that promote equitable political communication and support the creation of safer digital environments for women's participation in the public sphere. Furthermore, the methodological framework lays the groundwork for several future extensions. These include the integration of multi-label classification schemes that could capture overlapping categories of sexist discourse (e.g., combining benevolent and hostile traits), as well as the adaptation of the pipeline to other linguistic and cultural contexts. For instance, cross-lingual transfer to other languages may be explored by leveraging multilingual transformer models and domain-specific corpora.

By expanding the scope and granularity of hate speech detection, future work may contribute to a more nuanced and globally adaptable approach to combating digital gender-based violence.

References

1. Sexism, harassment and violence against women parliamentarians. Tech. rep., Inter-Parliamentary Union (2016), https://www.ipu.org/resources/publications/issue-briefs/2016-10/sexism-harassment-and-violence-against-women-parliamentarians
2. Sexism, harassment and violence against women in parliaments in europe. Tech. rep., Inter-Parliamentary Union (IPU) and Parliamentary Assembly of the Council of Europe (PACE) (2018), https://www.ipu.org/resources/publications/reports/2018-10/sexism-harassment-and-violence-against-women-in-parliaments-in-europe

3. Al-Hassan, A., Al-Dossari, H.: Detection of hate speech in social networks: a survey on multilingual corpus. In: 6th international conference on computer science and information technology. vol. 10, pp. 10–5121. ACM (2019)
4. Al-Hassan, A., Al-Dossari, H.: Detection of hate speech in arabic tweets using deep learning. Multimedia systems **28**(6), 1963–1974 (2022)
5. Alatawi, H.S., Alhothali, A.M., Moria, K.M.: Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. IEEE Access **9**, 106363–106374 (2021)
6. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 759–760. International World Wide Web Conferences Steering Committee (2017)
7. Blanco-Alfonso, I., Rodríguez-Fernández, L., Arce-García, S.: Polarización y discurso de odio con sesgo de género asociado a la política: análisis de las interacciones en twitter. Revista de Comunicación **21**(2), 33–50 (2022)
8. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing. pp. 71–80. IEEE (2012)
9. CommentGuard: The #1 comment moderation tool for facebook & instagram (2025), `https://commentguard.io/`
10. Dowlagar, S., Mamidi, R.: Hasocone@ fire-hasoc2020: Using bert and multilingual bert models for hate speech detection. In: Forum for Information Retrieval Evaluation. pp. 1–8 (2021)
11. Gagliardone, I., Gal, D., Alves, T., Martinez, G.: Countering Online Hate Speech, UNESCO Series on Internet Freedom, vol. 16. UNESCO Publishing, Paris (2015)
12. Hive Moderation: Hive moderation (2025), `https://hivemoderation.com/`
13. Instituto Nacional Electoral (INE): Violencia política contra las mujeres (2023), `https://igualdad.ine.mx/mujeres-en-la-politica/violencia-politica/`
14. Jane, E.: Misogyny Online: A Short (and Brutish) History. SAGE Publications Ltd (2016)
15. Martínez, M.P.J., López-Nava, I.H., y Gómez, M.M.: Identificación de sexismo en redes sociales. Tesis de maestría en ciencias, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Ensenada, Baja California, México (2025)
16. Meneguelli, G., Ferré Pavia, C.: El discurso de odio contra dilma rousseff desde la perspectiva semiolingüística. Estudos Feministas **32**(1), 187865 (2024)
17. Netino by Concentrix: Content & marketing services (2024), `https://netino.fr/en/`
18. ONU Mujeres: Violencia contra las mujeres en política (2023), `https://lac.unwomen.org/sites/default/files/2023-06/5eeb7511-c851-4b46-a15d-0089190e14a6.pdf`
19. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Effective hate-speech detection in twitter data using recurrent neural networks. Applied Intelligence **48**(12), 4730–4742 (2018)
20. Piñeiro-Otero, T., Martínez-Rolán, X.: Eso no me lo dices en la calle. análisis del discurso del odio contra las mujeres en twitter. Profesional de la Información **30**(5), 1–17 (2021)
21. Ranasinghe, T., Zampieri, M., Hettiarachchi, H.: Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. In: FIRE Working Notes. vol. 2517, pp. 199–207 (2019)

22. Respondology: Social media comment activation platform (2025), https://respondology.com/
23. Sightengine: Detect nudity, porn, suggestive and explicit adult content in images and videos (2025), https://sightengine.com/nudity-detection-api
24. United Nations: ¿Qué es el discurso de odio? (2022), "https://www.un.org/es/hate-speech/understanding-hate-speech/what-is-hate-speech
25. Urchukhishvili, G.: Indicators of sexist hate speech (2020), undp.org/sites/g/files/zskgke326/files/migration/ge/undp_ge_sexist_language_indicators_sapari_eng.pdf
26. Van Dijk, T.A.: Discourse as Social Interaction. SAGE Publications (1997)
27. Weaving, M., Alshaabi, T., Arnold, M.V., Blake, K., Danforth, C.M., Dodds, P.S., Haslam, N., Fine, C.: Twitter misogyny associated with hillary clinton increased throughout the 2016 u.s. election campaign. Scientific Reports **13**(1), 5266 (2023)
28. Wigand, C., Voin, M.: Speech by commissioner jourová–10 years of the eu fundamental rights agency: A call to action in defence of fundamental rights, democracy and the rule of law (2017), https://ec.europa.eu/