Detecting Obstetric Violence Tweets from Mexico: Annotation Guidelines and Classification with LLMs

Mónica Vázquez-Hernández, Helena Gómez-Adorno, Natalia Lerín-Hernández, Israel Islas Barajas, and Orlando Ramos-Flores

***Abstract*—This paper presents the construction and analysis of a manually annotated corpus of tweets related to Obstetric Violence (OV) shared on Twitter (now X). The study aims to identify different types of violence experienced by women during the perinatal period, as well as activism efforts that seek to raise awareness about OV. The methodology includes data collection through keyword filtering, manual annotation guided by typologies of OV, and a descriptive analysis using BERTopic to identify themes in the data. The tweets were classified into categories such as OV, Non-OV, and Activism, and further annotated based on narrator type and type of OV violence. The study also evaluates the performance of large language models (LLMs) — including ChatGPT, Copilot, and Meta's LLaMA — for zero-shot classification of tweets, highlighting their limitations in accurately identifying nuanced cases of OV. The research contributes a labeled dataset, a detailed annotation guide, and insights into the challenges of detecting OV in social media texts. It underscores the importance of addressing the invisibility and normalization of OV in both healthcare and NLP research.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10170

***Index Terms*—Obstetric violence, manual annotation, LLM, zero-shot, tweets, BERTopic.**

I. Introduction

WOMEN'S life experiences are determined by their social class, race, geographic location, age, sexual orientation, and educational level, among other relevant characteristics of the distribution of power; this framework is called intersectionality. The experience of a woman living her perinatal period is also connected to other categories of analysis that determine the conditions of exclusion and violence that are configured in a country like Mexico. Recognizing the differences between women, their privileges, and their oppression matrix allows us to determine if obstetric violence occurs to a small number of women or if it is the reality of the majority of women in Mexico [1]. Obstetric violence is understood as a set of practices occurring in healthcare institutions during childbirth that involve mistreatment, disrespect, or physical, emotional, and sexual abuse toward women, resulting in violations of their dignity, autonomy, and rights [2].

In Belgium, between 25-35% of women consider that their childbirth was traumatic; in the United Kingdom, 1 in 3 women have had a traumatic childbirth; in Australia, 1 in 10 women have been victims of obstetric violence [3]. While in Mexico, according to ENDIRE 2021, the prevalence of abuse in obstetric care against women between 15 and 49 years of age is 31.4% on average. The abuse is higher when the woman has a cesarean delivery (33.4%) than when she has a vaginal delivery (29.6%). In Mexico, 43% of women had a cesarean delivery, despite the WHO recommendation of 12%.

In Mexico, there are legislation and regulations to avoid obstetric violence (OV), such as the Official Mexican Standard NOM-007-SSA2-2016, for the care of women during pregnancy, childbirth, and puerperium, and the newborn[1], the General Health Law, the Law on Women's Access to a Life Free of Violence[2], and the General Recommendation 31/2017. According to all these, women have the right to decide freely, in a responsible and informed manner about their pregnancy, childbirth, and puerperium process, guaranteeing their physical integrity, the right not to suffer discrimination or coercion in all decisions related to sexual and reproductive life, as well as to receive a dignified and respectful medical attention.

Twitter (now X) is a social network that allows people who own a profile to share information, experiences, etc., in the form of text limited to 280 characters. Social networks are sometimes seen as spaces for dialogue where many women can complain about the abuses they have suffered[3]. Ideally, this space allows them to share their experiences and receive advice and support from other women. Some health personnel report cases of abuse done by colleagues, journalists share their investigative articles on OV, and feminist collectives make the issue visible. However, one of the most ethically accepted practices is to verify that the posts are public and are maintained for a period of time after being shared, in addition to belonging to identifiable profiles.

The associate editor coordinating the review of this manuscript and approving it for publication was Saul Zapotecas-Martinez (*Corresponding author: Mónica Vázquez-Hernández*).

Mónica Vázquez-Hernández, H. Gómez-Adorno, N. Lerín-Hernández, and O. Ramos-Flores are with Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Mexico City, 04510 Mexico (e-mails: monica.vazquez@iimas.unam.mx, helena.gomez@iimas.unam.mx, lerin.natalia@aries.iimas.unam.mx, and orlandxrf@gmail.com).

I. I. Barajas is with Facultad de Ciencias Políticas, Universidad Nacional Autónoma de México, Mexico City, 04510 Mexico (e-mail: israel.barajas.ir@politicas.unam.mx).

[1]Norma Oficial Mexicana-007-SSA2-2016. Para la atención de la mujer durante el embarazo, parto y puerperio y de la persona recién nacida.

[2]Ley General de Acceso de las Mujeres a una Vida Libre de Violencia

[3]Perez Barajas. Jovenes, Plataformas Digitales y Lenguajes. Página seis

In this work, we collected a set of Twitter posts (now X) related to keywords indicative of obstetric violence to analyze and classify the posts related to this topic. With this work, we aim to identify the types of profiles talking about the topic and make visible the most common types of OV reported by women on social networks. To achieve this objective in this work, we collected a set of posts between the years 2020 and 2022 to reach a total of 2156 tweets published by people and institutions [4].

In this work, we interpret women's participation on social media as a form of discursive engagement through which experiences of obstetric violence are shared, discussed, and made visible, including the search for support and collective awareness. Also, it is important for some midwives, gynecologists, and lawyers who are part of women's collectives, nonprofit civil associations, non-governmental organizations, nonprofit civil associations or non-governmental organizations, etc; Their activism is done by sharing their stories and testimonies of traumatic experiences, narrating the pain they have managed to position the term "obstetric violence" in social imaginary, and have helped to point out that OV is a structural problem, which goes from the practice of non-consensual cesarean to forced sterilizations. Their work helped thousands of women to learn about their rights and identify when they were being violated. International organizations such as the Committee on the Elimination of Discrimination against Women (CEDAW) have condemned and sentenced obstetric violence. This improved conditions for the care of mothers and children during pregnancy, childbirth, and postpartum[4,5].

As a non-judicial strategy, some women in the search for their cases are not invisibilized by institutions, carry out their activism to combat obstetric violence by encouraging other women to exercise, defend, and demand the fulfillment of their sexual and reproductive health rights, as an exercise of citizenship. This activism is carried out in person and virtually through socio-digital networks. In this work, we have analyzed tweets on activism to make OV visible. It should be noted that there is no specific descriptor to designate obstetric violence, which can be interpreted as an indicator by itself. This form of violence has not yet been constituted as an object of study in the field of NLP; therefore, this work is the first approximation.

This study is guided by the following research questions:

- RQ1. To what extent are explicit narratives of obstetric violence present in Spanish-language social media discourse from Mexico?
- RQ2. What types of experiences and forms of mistreatment are reflected in explicit obstetric violence narratives shared on social media?
- RQ3. How effectively can large language models identify obstetric violence narratives in a low-resource, zero-shot setting?

Section II shows how other authors have approached OV and the different types of violence considered OV, some of which are understood from the medical habitus. Section III describes the process of compiling posts and refers to a labeling guide, which describes the rules to be followed for labeling the posts as VO and examples of data labeling. The descriptive analysis shows statistics about the post distribution according to the types of narrators and graphs of the classification of posts according to the type of violence they experienced the Section IV shows the characteristics of the corpus, and Section discusses the results obtained. Finally, the contributions of the work are summarized in Section VI.

II. Background

OV is a different type of violence from medical violence because it is specifically directed at women and involves objectification, reification, and pathologization of mostly healthy bodies. OV has been analyzed from various perspectives, Roberto Castro [5], and the European project on obstetric violence[6] argue that there are five analytical approaches found in studies on obstetric violence: a) typology approach; b) legal and public policy approach; c) social science, psychological and/or anthropological approach; d) women's perception approach; and e) health professionals' perception approach. Sara Cohen adds the philosophical analysis, particularly focused on how violence is lived and experienced by women in terms of gender violence.

The OV experienced by women during their perinatal period can take several forms, one of them is psychological violence, which occurs when violence is exercised on the emotional stability of women, it could happen by action and omission, and consists of "limiting women to freely express their emotions, feelings and concerns, opting for an attitude of submission, due to scolding, criticism and mockery by health personnel". Psychological violence by omission occurs when a) lack of information about her evolution and b) the impossibility of deciding autonomously, while psychological violence by action includes i) verbal aggressions, ii) ignoring the mother's questions about her process (power/knowledge relationship), iii) inducing guilt and iv) obtaining informed consent under circumstances in which the pregnant woman is not in a position to refuse.

On the other hand, physical violence [6] is that which is exercised on the woman's body, among the most common practices of this form of violence exercised by health personnel are: 1) performing invasive practices or providing medication that is not justified and that may affect the state of health of the person who is about to be born or the pregnant woman, 2) not respecting the times or the possibility of vaginal delivery, 3) performing unnecessary cesarean sections 4) practicing Kristeller maneuvers and 5) performing vaginal touch by more than one person as well as recurrent touches [7]. A traumatic birth experience can have impacts on the physical or mental health of women and their infants [8].

Institutional violence is violence carried out by public officials, physicians, health personnel, and agents belonging to a public hospital, whose purpose is to delay, hinder, or prevent women from having access to an adequate health service [9].

[4]El parto es nuestro. https://www.elpartoesnuestro.es/

[5]Grupo de Información en Reproducción. http://gire.org.mx/salud-reproductiva/

[6]Obstetric Violence Project. The New Goal for Research, Policies, and Human Rights on Childbirth https://www.obstetricviolence-project.com/

Obstetric violence is closely connected to misogyny through everyday gendered practices that devalue women's autonomy within healthcare settings. As discussed by [10], misogyny in medical contexts often appears in patronizing, dismissive, or infantilizing attitudes and language toward women, which undermine their agency during childbirth. These practices contribute to an institutional environment where disrespectful care and coercive interventions become normalized. Thus, obstetric violence can be understood as a structural manifestation of misogyny embedded in medical hierarchies and gendered power relations.

Finally, prior NLP research addressing gender-based violence in Spanish has focused on related but more broadly framed offensive language tasks. For example, [11] introduces a Spanish dataset for the detection of hate speech, racism, and misogyny in social networks, emphasizing the need for region-specific resources in the Colombian context. In addition, [12] analyzes zero-shot transfer scenarios for hate speech detection across European and Latin American Spanish variants, highlighting the challenges of adapting models across dialectal differences and offensive content categories. These studies have advanced the detection of overt abusive content in Spanish social media, but they primarily address misogyny and hate speech in a general sense rather than institutionally contextualized forms of gender-based harm. In contrast, our work focuses on obstetric violence, a specific and underreported form of gender-based violence that manifests indirectly and institutionally, and provides theoretically grounded annotation guidelines, a manually curated dataset of explicit OV narratives from Mexico, and zero-shot baseline evaluations suited to low-resource conditions.

III. Data Collection Methodology

This section provides a comprehensive description of the data collection methodological framework followed in this research, the manual tweet annotation process, and the quantification of the inter-annotators' agreement.

A. Data Collection

The prevalence of violence, abuse, and discrimination against women in health services is obscured by the persistent tendency to confine these occurrences to subjective, unreported experiences. The inadequate conceptualization of obstetric violence compounds this phenomenon, the consequent inhibition of its articulation, and the normalization of such violations, especially when juxtaposed with positive birth outcomes and healthy newborns[7].

To prioritize women's perspectives during the perinatal period, the social media platform X (formerly Twitter) was selected. This platform serves as a conduit for women to exchange experiential accounts and seek communal support. Furthermore, healthcare practitioners, journalists, and feminist organizations utilize X to publicize instances of professional abuse, document experiences from academic training, and disseminate research on obstetric violence.

This study utilized the Twitter API[8] to retrieve a corpus of public tweets spanning May 2020 to December 2022 [13] originating from Mexico. The Twitter API retrieves public tweets in a non-selective manner, and most of the collected content is expected to be written in Mexican Spanish, although tweets in other language varieties may also be present. From this corpus, a subset of 2,152 tweets was extracted using keyword[9] and contextual cues related to obstetric care in Mexico, focusing on content about the perinatal period (pregnancy, childbirth, and postpartum) and obstetric violence.

B. Data Labeling

Before the data labeling phase, a comprehensive preliminary analysis was conducted to characterize the textual content of the extracted tweets. This analysis aimed to discern the thematic focus across distinct user groups, including women within the perinatal period, healthcare practitioners, journalists, and advocacy collectives. Furthermore, it aimed to elucidate the specific facets of obstetric violence being addressed within these accounts, thereby providing a foundational understanding of the dataset's composition and the diverse perspectives it encompassed. This preliminary investigation served to inform the development of robust and nuanced annotation guidelines, ensuring that the subsequent labeling process accurately captured the complexities inherent in the data. This was followed by a comprehensive literature review to define the categories and typologies of obstetric violence described in [14].

Following the establishment of obstetric violence categories and typologies, comprehensive annotation guidelines were developed. These guidelines served to provide annotators with detailed descriptions of each category, enabling them to classify individual accounts and their corresponding types accurately. Furthermore, the guidelines included specific definitions and examples illustrating the nuances of each obstetric violence type, ensuring consistency and precision in the annotation process. The complete annotation guidelines are accessible via an online repository[10].

Table I details the classification categories for the tweets and provides illustrative examples for each. Following this initial categorization, our analysis focused specifically on tweets identified as relating to Obstetric Violence.

Following the classification, we identified profiles of individuals or institutions sharing experiences of Obstetric Violence. Table II presents the three distinct narrator types found within these posts, along with an illustrative example for each.

Finally, Table III provides a detailed description of each type of violence that can be perpetrated against women during the perinatal period, accompanied by an illustrative example.

For labeling the tweets, a multidisciplinary team composed of two physicians and an activist, each bringing unique perspectives, classified the corpus into three predefined categories. This collaborative approach aimed to capture the nuanced nature of the data, particularly given the sensitive subject matter. The Argilla[11] labeling tool facilitated this process, providing

[7]https://www.obstetricviolence-project.com/
[8]https://developer.x.com/en/docs/x-api
[9]https://github.com/PLN-disca-iimas/obstetric-violence-tweets
[10]https://pln-disca-iimas.github.io/obstetric-violence-tweets/
[11]https://docs.argilla.io/latest/

TABLE I
TWEETS CATEGORIES

Category	Tweet
Obstetric Violence	Translated tweet: "Oh look, someone just does an episiotomy a 26-week 600g birthing". Original tweet: *Oh mira, alguien le acaba de hacer episiotomia a un parto de 26 semanas de 600g.*
Non-Obstetric Violence	Translated tweet:We undergo menstruation every month, we have to take hormone pumps as contraceptives, our body has multiple changes when we are pregnant, we are almost split in two to give birth to a baby, which we created in our beautiful body, and they tell me that we are the WEAK SEX?". Original tweet: Nos baja cada mes, tenemos que meternos bomba de hormonas como anticonceptivos, nuestro cuerpo tiene múltiples cambios al gestar un bebé, casi nos parten en dos para parir a un bebé, el cual nosotros creamos en nuestro bello cuerpo y me dicen que somos el SEXO DÉBIL?.
Information about Obstetric Violence	Translated tweet: Find out 31.4% of women between the ages of 15 and 49 who had a childbirth or cesarean section experienced some type of mistreatment #ObstetricViolence URL #16daysofactivism URL". Original tweet: *"Enterate 31.4% de las #Mujeres de 15 a 49 años que tuvieron parto o cesárea experimentó algún tipo de maltrato. #ViolenciaObstétrica URL #16DíasDeActivismo URL"*

TABLE II
OBSTETRIC VIOLENCE NARRATORS

Account	Tweet
First-person accounts	Translated tweet: "It's not even 12 yet and I've already shaved exactly 6 vaginas, if you're going to give birth, come shaved :(". Original tweet: *"Todavía no son ni las 12 y ya depile exactamente 6 vaginas, si van a parir venganse depiladitas :("*
Third-person accounts	Translated tweet: "Today I saw the scariest thing at my internship: A Pediatrics R1 doing Kristeller". Original tweet: *"Hoy vi la cosa más espantosa de mi internado: Una R1 de Pedia haciendo Kristeller"*
General information about Obstetric Violence	Translated tweet: "In Mexico, three out of ten women have suffered obstetric violence, and 4 percent were forcibly sterilized during their last birth, according to the National Survey on the Dynamics of Household Relationships conducted by INEGI in 2016". Original tweet: *En México tres de cada diez mujeres han sufrido violencia obstétrica y 4 por ciento fueron esterilizadas de manera forzada durante su último parto, según la Encuesta sobre la Dinámica de las Relaciones en el Hogar realizada por el INEGI en 2016.*

TABLE III
OBSTETRIC VIOLENCE TYPES

Violence type	Tweet
Psychological by Omission	Translated tweet: "Obstetric violence doesn't exist, but a doctor performed an ultrasound on me at 7 months pregnant without even looking at me and only talking to the trainee: 'Ugh! Breech, poor guy on call.' Me: 'No, poor me, not the doctor.' Then he did look at me." Original tweet: *La violencia obstétrica no existe, pero a mi me hizo una ecografía con 7 meses de embarazo un médico que ni me miró a la cara y solo hablaba con el de prácticas: "¡Buf! de nalgas, pobre del que le toque guardia". Yo: "No, pobre de mí, no del médico" Entonces sí me miró.*
Psychological by Action	Translated tweet: "When my mom was having a cesarean section to have me, she had a hemorrhage, and the doctor told her in the coldest way possible, 'In two hours you're going to bleed to death.' She just hugged me and cried. 21 years have passed and my mom is still very much alive." Original tweet: *Cuando a mi mamá le hacían la cesárea para tenerme a mi, tuvo una hemorragia, y el médico le dijo de la manera más fría posible "en dos horas se va a morir desangrada". Ella solo me abrazaba y lloraba. Han pasado 21 años y mi mamá sigue muy viva.*
Physical	Translated tweet: "When the Gynecology R2 performed the Kristeller maneuver on a patient, causing fetal distress and the death of the baby. On top of that, he said it was the fault of the Pediatric MIP who didn't know how to resuscitate it (I was the MIP) and my on-call R never came to support me." Original tweet: *Cuando el R2 de Gine le hizo Maniobra de Kristeller a una paciente, ocasionando sufrimiento fetal y la muerte del producto. Aunado a eso, dijo que fue culpa del MIP de Pediatría que no lo supo reanimar (yo era el MIP) y mi R de guardia nunca me fue a apoyar.*
Institutional (Ethical Violations)	Traslated tweet: "In my hospital, they're blaming an INTERN for not properly resuscitating a newborn with an undiagnosed diaphragmatic hernia in a SCHEDULED cesarean section. There was no pediatrician present. The joke tells itself." Original tweet: *En mi hospital están echándole la culpa a una PASANTE de no reanimar adecuadamente en una cesárea PROGRAMADA a un neonato con HERNIA DIAFRAGMATICA no diagnosticada prenatalmente. No había pediatra. El chiste se cuenta solo.*
Sexual	Translated tweet: "The case of the anesthesiologist who sexually abused patients during their labor is an example of how violent the world is against women, that not even during childbirth are we free from violence. This case should lead us to deep reflection. #NiUnaMás" Original tweet: *El caso del anestesiólogo que abusaba sexualmente de pacientes durante su labor de parto es un ejemplo de lo violento que es el mundo contra las mujeres, que ni durante el parto vivimos libres de violencia. Este caso debe llevarnos a una profunda reflexión. #NiUnaMás*

a structured environment for annotation and enabling efficient collaboration among the labelers.

Following the initial manual classification, we conducted a thorough inter-rater agreement analysis to ensure the labeled data's reliability and consistency. Specifically, we employed Cohen's Kappa metric to quantify the agreement between pairs of labelers. This metric accounts for the possibility of agreement occurring by chance, providing a more robust measure of actual agreement. Furthermore, we meticulously re-examined all tweets where disagreement was observed, especially those where each labeler assigned a different category. This comprehensive review allowed us to resolve discrepancies and refine the labeling, ensuring a more accurate and robust dataset. Table IV presents the inter-annotator agreement results, detailing the calculated Cohen's Kappa scores for each pair of labelers. A1, A2, and A3 denote Annotator 1, Annotator 2, and Annotator 3, respectively. This table provides a comprehensive overview of the inter-rater reliability achieved during the classification task for each classification category: tweet category, OV Narrator, and OV type. It can be observed a near-perfect agreement for all categories, being the OV type being the category with the lowest agreement rate.

TABLE IV
INTER-ANNOTATOR AGREEMENT (COHEN'S KAPPA SCORES) FOR EACH ANNOTATOR PAIR

Annotators	Tweet Category	OV Narrators	OV Types
A1 & A2	0.97	0.95	0.88
A2 & A3	0.95	0.89	0.87
A1 & A3	0.93	0.87	0.81

Annotators found that tweets describing physical violence and psychological violence through direct action were the easiest to classify. Conversely, tweets detailing psychological violence through omission or institutional violence generated the most discussion and debate among the team.

It's important to recognize that hospitals serve as training grounds for medical professionals, fostering an environment where students, teachers, and the practical field converge. Unfortunately, in this dynamic, women's bodies can be objectified, becoming the material for student learning.

Notably, two of the labelers were physicians, and their professional backgrounds sometimes presented challenges. Their academic training occasionally hindered their ability to recognize instances where they, or the institution, might be perpetrating violence. This highlights the potential for ingrained biases to influence perception, even among those dedicated to patient care.

The following examples illustrate instances where physicians initially failed to recognize institutional violence, potentially due to their ingrained medical perspectives. However, after collaborative discussion and reflection, it was determined that these accounts did indeed describe institutional obstetric violence:

- Translated tweet: "Something like this happened to me: a doctor was taking a GyO R0 to 'practice,' a birth comes up and she tells me no, leave this birth to X, you go see if the packages for the scheduled C-section are ready, I come back and the lady is in the delivery room, and the R0 is washing his face, she ruptured the membranes and bam!!! Face bathed." Original tweet: *Me paso algo así: una Dra. llevaba a un R0 de GyO a "practicar" sale un parto y me dice no, este parto dejaselo a X, tu ve a ver si ya están los paquetes de la cesárea programada, regreso y la doña en sala, y el R0 lavándose la cara, le reventó membranas y zas!!! Carita bañada*
- Translated tweet: "Enough with this lying on the bed with your legs open, having 3 people enter the room, and without saying 'hello' or who they are, they start touching you without telling you what they're going to do. And, of course, having previously kicked out your companion. #stopviolenciaobstetrica" Original tweet: *Basta ya eso de estar tumbada en la camilla abierta de patas, que entren 3 personas en la sala, y sin decir ni "hola" ni quiénes son, empiecen a tocarte sin decirte lo que van a hacer. Y, por supuesto, previamente habiendo echado a tu acompañante. #stopviolenciaobstetrica*

In the following account, we see a clear instance of psychological violence by omission, where the patient is denied timely and necessary attention:

- Translated tweet: "By the way, the baby was face up and with a nuchal cord. That's why she wasn't dilating, because she couldn't push. He ended up exhausted and bradycardic after 18 hours of labor. That's why they did the cesarean section. But she wasn't dilating 'because she was nervous.'" Original tweet: *Por cierto, el niño estaba de cara y con vuelta de cordón. Por eso no dilataba, porque no podía empujar. Terminó agotado y bradicárdico después de haber pasado 18h. de parto. Por eso hicieron la cesárea. Pero no dilataba "porque estaba nerviosa".*

The following narrative presented another point of contention, ultimately classified as institutional violence due to the denial of mother-newborn bonding:

- Translated tweet: "The victim and obstetric violence of a woman whose baby hadn't been brought to her because there was no staff!!" Original tweet: *La victima y la violencia obstetrica de una ñora que no le habían llevado a su bebé pq no había personal!!.*

In conclusion, our labeling process revealed a clear disparity in classification ease. Tweets detailing overt forms of violence, specifically physical violence and psychological violence enacted through direct actions, were readily categorized with minimal ambiguity. This can likely be attributed to the explicit nature of these violations, leaving little room for subjective interpretation. Conversely, tweets depicting more subtle forms of violence, such as psychological violence through omission or institutional violence, proved significantly more challenging. These instances required extensive discussion and debate among the labelers, highlighting the inherent complexity and nuanced nature of these categories. The difficulty in classifying these tweets underscores the importance of considering the broader context and systemic factors that contribute to obstetric violence. It also reinforces the need for rigorous training and collaborative discussion to ensure consistent and

accurate labeling, particularly when dealing with sensitive and potentially subjective data.

IV. Dataset Impact and Topic Analysis

Figs. 1 and 2 show characteristics of the corpus of tweets collected. There are 2157 tweets, of which 90.9% are from the category 'Non-obstetric violence', 4.5% from 'Activism to make obstetric violence visible', and 4.5% from the category 'Obstetric violence'. The majority of OV tweets are third-person narratives, followed by victim testimonials. Activist tweets mostly provide general information on obstetric violence. Although the number of OV-related tweets is limited, it reflects the intrinsic difficulty of identifying overt descriptions of obstetric violence in public social media discourse, where such experiences are often normalized, indirectly expressed, or underreported. Consequently, this dataset is not intended to be statistically representative of all linguistic manifestations of obstetric violence. Instead, its primary contribution lies in providing a high-quality, manually annotated gold standard of explicit OV narratives, which can support exploratory analyses and serve as a foundation for future corpus expansion and methodological development.

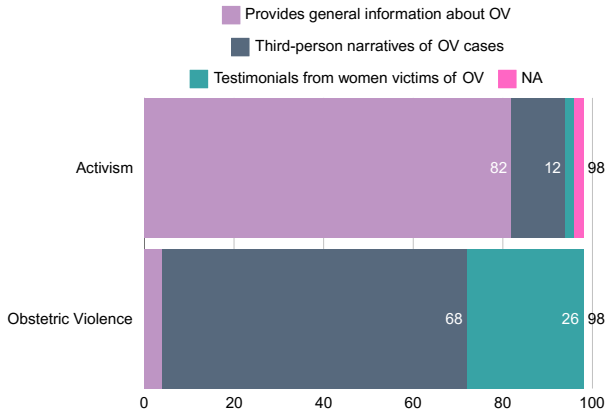


Fig. 1. Distribution of the type of account in tweets on obstetric violence and on activism to make it visible.

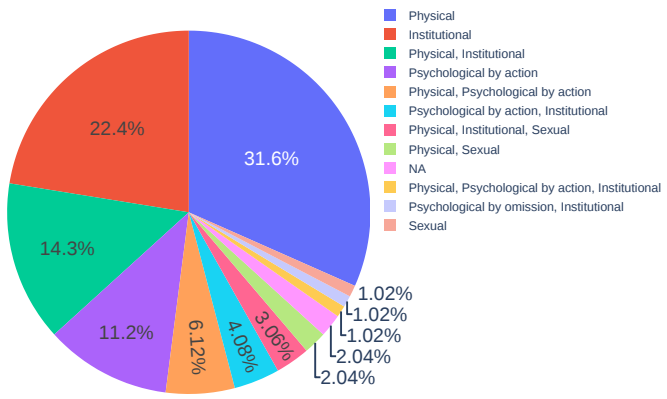


Fig. 2. Distribution of the type of violence in OV tweets.

Fig. 2 shows the distribution of tweets classified in the OV category. The most frequent form of violence in the accounts is physical violence, present in 31.6% of the total samples. It is followed by institutional violence, which appears in 22.4% of the stories, while 14.3% of the cases present both types of OV simultaneously. On the other hand, psychological violence by action is identified in 11.2% of the tweets, while psychological violence by omission is only observed in combination with institutional violence, representing 1.02% of the stories.

Topic modeling is a machine learning and natural language processing (NLP) technique for discovering topics in large volumes of text [15]. It allows organizing, understanding, and summarizing text collections by inferring hidden topics [16]. It is based on the identification of co-occurrences of words within a corpus, grouping them into sets representing different topics, so that documents are associated with a number of underlying topics and an interpretable representation of the documents is provided [17]. Among the most widely used methods for topic modeling is Latent Dirichlet Allocation (LDA), which assumes that documents contain a mixture of topics, each with a probabilistic distribution over words. However, LDA has limitations when working with short texts, such as social media posts [15].

Therefore, for the descriptive and exploratory analysis of the OV corpus, we considered the BERTopic neural approach, which uses embeddings of the text using BERT (Bidirectional Encoder Representations from Transformers) [18]. The latter is a language model based on Transformer networks that learns different representations for polysemous words [16]. In particular, BERTopic is a model based on the embedding clustering approach, incorporating a class-based variant of TF-IDF (c-TF-IDF) to create interpretable thematic clusters [19]. While BERTopic does not require extensive preprocessing of the documents [15], in the specific case of the OV tweets in the corpus, a data cleaning process was applied with the library *pysent* [20]. Especially, usernames and links were removed, hashtags were converted to text after removing the symbol #, and emojis were transformed into their respective textual descriptions to preserve their semantic content. Within the modularity framework of BERTopic, the vectorization model was configured to incorporate bi-gram generation in the topic extraction process. In addition, in the c-TF-IDF weighting scheme, a reduction of the importance assigned to high-frequency words was applied.

Thus, three topics were found in the tweets labeled as obstetric violence. Figure 3 shows the 10 words with the highest c-TF-IDF score for each of the three topics found. Topic 1 presents words related to visibilization ("obstetric violence does exist") and to OV of the physical type. Topic 2 contains the terms "medical malpractice" and "scheduled cesarean-section", as well as "hospital" and "pediatrician", so they can be related to institutional violence. Finally, Topic 3 is associated with physical violence, as it mainly includes the terms "Kristeller maneuver" (KM) and "fetal distress".

KM is discouraged by the WHO[12], since it has not been proven to be effective or safe, and its application leads to serious physical consequences for both the mother and the newborn. This explains the close relationship between the terms mentioned in Topic 3. Such a maneuver, also called fundal pressure, has been linked to adverse effects, such as neonatal fractures and brain damage [21], as well as an increased likelihood of lacerations to the mother's perineum, damage to the levator ani muscle and the cervix, and long-term consequences, such as rib pain and dyspareunia [22], [23]. Also, it has been documented that senior physicians often request their subordinates to apply KM [24], which is reflected in the representative words in Topic 3, "request" and "orders". Because of the hierarchical structure of the hospital setting, which favors higher-ranking physicians, lower-ranking practitioners are often forced to abide by these directions [25]. Finally, Table V shows the tweets taken as representative by BERTopic, which coincide with the previous interpretation. This can be seen in the usage of the words with the highest c-TF-IDF score on each tweet shown. The distribution of accounts in the topics is shown in Fig. 4, where it can be seen that the tweets that provide general information about O.V. are all located in Topic 1, which is the topic with the most tweets. Topic 2 contains the greatest proportion of third-person narratives.

Given the small size of the OV subset, the topic modeling analysis presented in this section should be interpreted as exploratory and descriptive. Rather than aiming to identify stable or generalizable thematic structures, this analysis illustrates recurring themes within explicit OV narratives and provides qualitative insight into the types of experiences reported. These results are therefore not intended for quantitative generalization, but to complement the annotation analysis and highlight challenges associated with studying obstetric violence in low-resource settings.

V. Clasification Models

Text classification is the assignment of predefined categories to a given text. We focus our experimental evaluation on zero-shot large language models due to the limited availability of labeled OV data. We also evaluated a traditional supervised text classification approach using BOW and TF–IDF combined with a linear SVM classifier. While these models typically require substantially larger training sets to avoid overfitting and to yield meaningful performance estimates, they are incorporated for completeness in the experiments. The automation of this task on a large scale has generated enormous interest due to the recent advancements of LLMs such as ChatGPT, Gemini, or Llama, which have shown excellent performance in text classification tasks [26].

There are various ways to obtain results with different LLMs, and although it is possible to make adjustments to LLMs for text classification, in this research responses are obtained through a *zero-shot* approach, this allows to not make any additional adjustment to the LLM, therefore this removes the need for extra steps to carry out the classification task, and for this reason it has gained attention in the research on evaluating LLM responses. However, it requires distinguishing the category descriptions in the *prompts* [27]. In some cases, the *prompt* design tries to improve *zero-shot* classification by segmenting texts and creating specific *prompts* for each category or adding label names with an adaptation for *zero-shot* classification [27].

From the LLM evaluation perspective, generating the desired response is the predominant approach [28]. There are different evaluation criteria, such as *Performance* (how well LLMs produce the desired results), *Ethicality* (privacy protection, reduction of misinformation, and transparency), *Fairness* (mitigation of biases to prevent discriminatory decisions), *Generalization* (LLMs' ability to adapt to unseen data), *Robustness* (resilience to errors, manipulation, or adversarial attacks, providing consistent responses), *Reasoning* (the ability of LLMs to logically infer or deduce information) [28].

Three basic measurement methodologies are considered: *Multiple-Classification* (MC, responses that classify texts based on predefined categories), *Token-Similarity* (TS, how similar different tokens are), and *Question-Answering* (QA, whether the responses LLMs provide to different questions are correct).

This research focuses on the *Performance* criterion through evaluation with MC using metrics such as accuracy, precision, recall, and F1-score. It is primarily used for classification tasks and is characterized by offering simplicity in LLM evaluations, although it depends on a set of pre-labeled responses, assuming perfect labeling, and involves a significant amount of resources in the labeling task [28].

A. Zero-shot Classification with Large Language Model

In order to evaluate the performance of different LLMs, such as ChatGPT, Copilot, and Meta, for the classification of the collected tweets, a prompt was designed that contains the same definitions of OV, Activism to highlight OV, and Non-OV used in human classification. This way, the information to classify the same tweets is the same for both human classification and the LLMs considered. The designed prompt is as follows:

```
Considera las siguientes categorías: 1.
Activismo para visibilizar violencia obstétrica:
Para que el tweet cumpla con el criterio de
activismo para visibilizar violencia obstétrica
dentro del tweet se debe dar información sobre
normas, legislación nacional o internacional
sobre qué prácticas clínicas son consideradas
violencia obstétrica, así como recomendaciones
de organizaciones internacionales de derechos
humanos y/u organizaciones no gubernamentales,
así como referencias a notas periodísticas y/o
reportajes gráficos sobre Violencia Obstétrica.
2. Violencia obstétrica: Para que el tweet cumpla
con el criterio de Violencia Obstétrica tiene que
describir una experiencia en primera o tercera
persona de una paciente o de un(a) médico(a)
que sufrió, ejerció o vió de cerca como alguien
del personal de salud ejerció algún tipo de
violencia contra alguna mujer o persona gestante
en cualquier etapa del periodo perinatal. 3. No
Violencia Obstétrica: El tweet no cumple con el
criterio de violencia obstétrica cuando no se
comparte ningún testimonio, ni hace referencia a
```

[12]World Health. Organization and Regional Office for Europe, Hospital Care for Mothers and Newborns: Quality Assessment and Improvement Hospital care for mothers and newborn babies: quality assessment and improvement toolhttps://www.who.int/europe/publications/i/item/WHO-EURO-2014-6059-45824-65972

TABLE V
REPRESENTATIVE TWEETS PER TOPIC

Topic	Tweet
1	Translated tweet: "Doctor accused of performing forced sterilization will no longer treat migrants in the US: Dr. Mahendra Amin faces accusations of performing hysterectomies and other procedures on women at the Irwin County Detention Center; URL (via ElUniversal) URL" Original tweet: *Médico acusado de practicar esterilización forzada ya no atenderá a migrantes en EU: El doctor Mahendra Amin enfrenta acusaciones de realizar histerectomías y otros procedimientos a mujeres en el Centro de Detención del Condado Irwin; URL (vía ElUniversal) URL*
2	Translated tweet: "The anesthesiologist arrived but there was no anesthesia!!!! Another woman about 38 years old and first timer had even lost her baby the day before and they could not do a cesarean section because there was NO ANESTHESIA, they induced labor and the poor woman had to deliver her baby dead and in pieces."Original tweet: *Llegó el anestesiólogo pero no había anestésia!!! Otra mujer como de 38 años y primeriza incluso había perdido a su bebé un día antes y no le pudieron hacer cesárea porque NO HABIA ANESTESIA, le inducieron el parto y la pobre mujer tuvo que parir a su bebé muerto y en pedazos*
3	Translated tweet: "They performed the Kristeller maneuver on me without even asking for my consent (suddenly I saw a guy on top of me squeezing me and I felt like I was losing my life inside). At no time did they inform me or my partner of the risks involved #stopkristeller" Original tweet: *A mi me practicaron la maniobra de Kristeller sin ni siquiera pedirme el consentimiento (de repente me vi un maromo encima apretando y yo notando que se me iba la vida dentro). En ningún momento nos informaron a mí o a mi pareja los riesgos que suponía #stopkristeller*

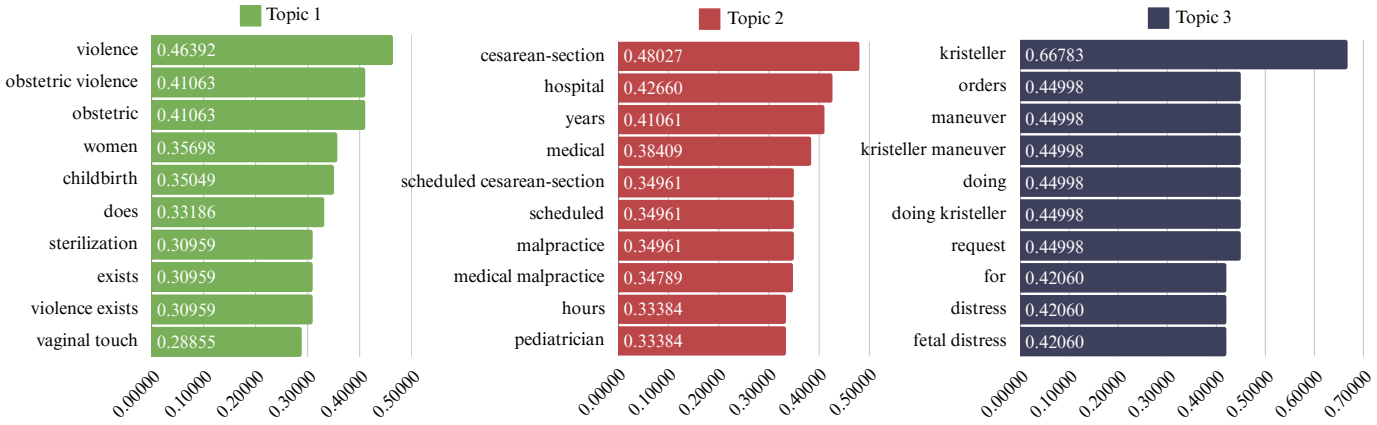


Fig. 3. Words with the highest c-TF-IDF score by topic of tweets classified as OV.

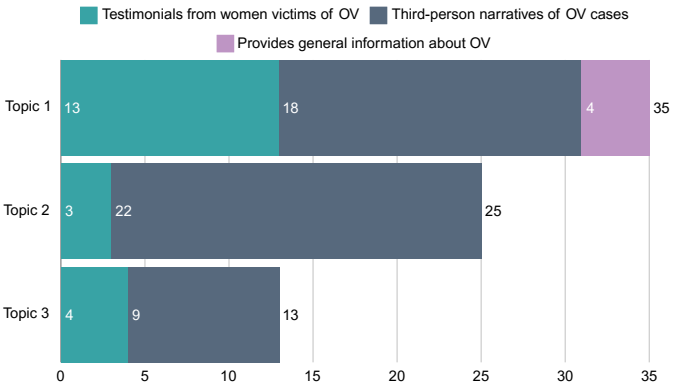


Fig. 4. Number of obstetric violence tweets by type of account and by topic.

```
ninguna norma, ley o recomendación de organismos
nacionales o internacionales de derechos humanos.
Clasifica los siguientes textos en alguna de las
tres definiciones. Únicamente responde utilizando
el ID con el nombre de la clasificación. Evita
comentarios adicionales.
```

Finally, the results of ChatGPT (GPT-4-turbo), Copilot, and Meta (Llama3) were compared with those of human classification using metrics to evaluate the classification performance. The results can be observed in the Table. VI.

As the results of each metric approach 1, the performance is better, and as they approach 0, the performance is worse. According to the results in Table VI, all LLMs show a low level of performance in classifying the categories of OV and Activism to highlight OV, compared to Non-OV. This is likely due to the number of tweets collected: 98 for OV, 98 for Activism to highlight of OV, and 1961 for Non-OV. Considering that, for the case of OV, it is important to avoid false negatives; recall gains greater importance than the Precision metric, where we avoid false positives. With this particular consideration, the LLM that shows the best performance is Copilot for OV; while for detecting Activism to raise awareness of OV, Meta has the best performance.

The results in Table VI also show that the BOW+SVM baseline clearly outperforms the zero-shot LLM approaches across most evaluation metrics. In particular, BOW+SVM achieves the highest macro F1-score (0.605) and overall accuracy (0.925), substantially surpassing ChatGPT, Copilot, and Meta models. A closer examination of the class-level metrics shows that BOW+SVM performs especially well on the majority class (Non OV), achieving a very high recall (0.985) and F1-score (0.963), which contributes significantly to the overall accuracy of the model. However, recall and F1-score for the OV class remain relatively low (0.306), reflecting the difficulty of detecting rare and linguistically diverse narratives of obstetric violence.

Regarding the number of correct predictions of the LLMs, ChatGPT has the highest average accuracy score of 0.790. However, when we consider the Macro Average F1-score,

TABLE VI
RESULTS OF PERFORMANCE METRICS FOR CLASSIFICATION OF VO, WITH CHATGPT, COPILOT, META AND A BOW BASELINE.

Labels	**Precision**	**Recall**	**F1-Score**	**Support**	**Model**
Activism to highlight OV	0.136	0.122	0.129	98	ChatGpt
	0.189	0.776	0.304	98	Copilot
	0.129	0.867	0.224	98	Meta
	0.636	0.367	0.468	98	BOW+SVM
Non Obstetric Violence (Non OV)	0.973	0.83	0.895	1961	ChatGpt
	0.992	0.749	0.853	1961	Copilot
	0.994	0.671	0.801	1961	Meta
	0.943	0.985	0.963	1961	BOW+SVM
Obstetric Violence (OV)	0.162	0.653	0.259	98	ChatGpt
	0.284	0.796	0.418	98	Copilot
	0.345	0.612	0.441	98	Meta
	0.556	0.306	0.395	98	BOW+SVM
Macro avg	0.423	0.535	0.428	2157	ChatGpt
	0.488	0.773	0.525	2157	Copilot
	0.489	0.717	0.489	2157	Meta
	0.712	0.549	**0.605**	2157	BOW+SVM
Accuracy			0.790	2157	ChatGpt
			0.752	2157	Copilot
			0.677	2157	Meta
			0.925	2157	BOW+SVM

that is, the average scores unweighted by class size, Copilot achieves the best score, followed by Meta, and finally ChatGPT.

The performed error analysis shows consistent confusions between *Obstetric Violence* and *Activism to highlight OV* across all models, indicating that predictions are largely driven by salient violence-related lexical cues rather than by narrative perspective. As a result, advocacy and awareness-raising tweets are often interpreted as direct victimization narratives. This pattern suggests a bias toward over-attributing violence in gender- and healthcare-related content, reflecting both model training data and broader societal discourse. These findings highlight the need for cautious interpretation of LLM outputs and motivate future work on interpretability and bias-aware modeling in sensitive healthcare contexts.

VI. CONCLUSIONS

This paper presents the development of a manually labeled corpus of social media posts containing narratives of obstetric violence (OV), with the aim of supporting the creation of automatic classification models. Such models are crucial for identifying and analyzing women's experiences during the perinatal period, particularly given the widespread normalization and invisibilization of OV practices.

Our main contributions include: (a) a comprehensive labeling guide that defines the criteria for identifying OV, the narrator's perspective, and the specific types of violence; (b) a curated corpus of tweets annotated in accordance with this guide; and (c) an initial evaluation of large language models (LLMs) for the automatic classification of these posts.

Beyond its methodological contributions, this work advances academic understanding of obstetric violence by providing empirical evidence that supports feminist and sociological accounts of this phenomenon as normalized, institutional, and often underreported in public discourse. From a practical perspective, the annotated corpus and baseline analyses can support policy-oriented research aimed at monitoring obstetric violence in digital spaces, contribute to digital activism efforts by increasing the visibility of experiential narratives, and inform the development of healthcare monitoring tools that analyze public perceptions and reports of childbirth care. Moreover, the dataset and annotation guidelines are designed to be reusable by other researchers and institutions, enabling replication studies, cross-regional comparisons, model adaptation, and fine-tuning as larger datasets become available, and interdisciplinary collaborations between NLP researchers, social scientists, and healthcare stakeholders.

Experimental results show that current LLMs, when used in a zero-shot setting, struggle to accurately detect OV-related content, with a maximum F1-score of 0.525. These findings highlight the limitations of generic LLMs for sensitive and domain-specific classification tasks, particularly when dealing with nuanced forms of violence.

A key limitation of this study is the relatively small number of tweets explicitly labeled as Obstetric Violence (98 instances). This constraint does not stem from limited data access or sampling bias, but rather from the intrinsic characteristics of obstetric violence as a social phenomenon. The corpus was extracted from a previously collected Twitter (X) dataset gathered using the research Twitter API prior to the introduction of current pricing and access constraints, covering a multi-year period (2020–2022). Despite applying targeted keyword-based filtering, only a small fraction of tweets contained explicit experiential narratives that met the strict criteria defined in our annotation guidelines. This outcome is consistent with the normalization, invisibilization, and underreporting of obstetric violence described in sociological and feminist research, and highlights the challenges of identifying such experiences in public social media discourse [5]. Consequently, the findings presented in this work should be

interpreted as exploratory and not statistically generalizable. Nevertheless, the curated corpus and annotation guidelines provide a reliable foundation for future work on corpus expansion, data augmentation, and supervised model training.

In future work, we propose to enhance classification performance through fine-tuning transformer-based models using the labeled corpus. This will include expanding the label set to cover the full range of violence types and narrative perspectives, to improve both detection accuracy and the interpretability of model outputs.

ACKNOWLEDGMENTS

This work was supported by UNAM Posdoctoral Program (POSDOC) and PAPIIT project IN104424. This manuscript was revised with the assistance of artificial intelligence tools (ChatGPT, OpenAI) to enhance clarity and correct grammatical issues.

REFERENCES

[1] K. Crenshaw, "Mapping the margins: Intersectionality, identity politics, and violence against women of color," *Stanford Law Review*, vol. 43, no. 6, pp. 1241–1299, 1991, DOI: 10.2307/1229039.

[2] S. M. Frías and R. Castro, "Mistreatment, abuse, and gender-based violence during childbirth: A longitudinal analysis of obstetric violence in méxico (2011–2021)," *Violence Against Women*, vol. 31, no. 14, pp. 3496–3522, 2025, DOI: 10.1177/10778012241289426.

[3] I. O. Fernández, "Ptsd and obstetric violence," *Midwifery today with international midwife*, no. 105, pp. 48–9, 2013. [Online]. Available: https://www.midwiferytoday.com/mt-articles/ptsd-and-obstetric-violence/

[4] I. Soldevilla and N. Flores, "Natural language processing through bert for identifying gender-based violence messages on social media," in *2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE)*, 2021, pp. 204–208, DOI: 10.1109/ICICSE52190.2021.9404127.

[5] R. Castro and S. M. Frías, *Violencia obstétrica y ciencias sociales : estudios críticos en América Latina*. UNAM, 2022, DOI: 10.22201/crim.9786073058698p.2022.

[6] L. I. Díaz García and Y. Fernández M., "Situación legislativa de la violencia obstétrica en américa latina: el caso de venezuela, argentina, méxico y chile," *Revista de derecho e la Pontificia Universidad Católica de Valparaíso*, pp. 123 – 143, 12 2018, DOI: 10.4067/S0718-68512018005000301.

[7] E. Jojoa-Tobar, Y. D. Cuchumbe-Sánchez, J. B. Ledesma-Rengifo, M. C. Muñoz-Mosquera, J. P. Suarez Bravo, and A. M. Paja Campo, "Violencia obstétrica: haciendo visible lo invisible," *Salud UIS*, vol. 51, no. 2, p. 136–147, abr. 2019, DOI: 10.18273/revsal.v51n2-2019006.

[8] C. T. Beck and S. Watson, "Impact of birth trauma on breast-feeding: a tale of two pathways," *Nursing research*, pp. 228–236, July-August 2008, DOI: 10.1097/01.NNR.0000313494.87282.90.

[9] C. Bellamy and R. Castro, "Formas de violencia institucional en la sala de espera de urgencias en un hospital público de méxico," *Revista Ciencias de la Salud*, vol. 17, no. 1, p. 120–137, february 2019, DOI: 10.12804/revistas.urosario.edu.co/revsalud/a.7621.

[10] N. von Benzon, J. Hickman-Dunne, and R. Whittle, "'my doctor just called me a good girl and i died a bit inside': From everyday misogyny to obstetric violence in uk fertility and maternity services," *Social Science & Medicine*, vol. 344, p. 116614, 2024, DOI: 10.1016/j.socscimed.2024.116614.

[11] L. G. Moreno-Sandoval, A. Pomares-Quimbaya, S. A. Barbosa-Sierra, and L. M. Pantoja-Rojas, "Detection of hate speech, racism and misogyny in digital social networks: Colombian case study," *Big Data and Cognitive Computing*, vol. 8, no. 9, p. 113, 2024, DOI: 10.3390/bdcc8090113.

[12] G. Castillo-López, A. Riabi, and D. Seddah, "Analyzing zero-shot transfer scenarios across spanish variants for hate speech detection," in *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, 2023, pp. 1–13.

[13] H. Gomez-Adorno, G. Bel-Enguix, G. Sierra, J. Barajas, and W. Álvarez, "Machine learning and deep learning sentiment analysis models: Case study on the sent-covid corpus of tweets in mexican spanish," *Informatics*, vol. 11, 2024, DOI: 10.3390/informatics11020024.

[14] C. Roberto, "Génesis y práctica del habitus médico autoritario en méxico," *Revista Mexicana de Sociología*, vol. 76, no. 2, pp. 167–197, May 2014, DOI: 10.22201/iis.01882503p.2014.2.46428.

[15] R. Egger, "Topic modelling," *Tourism on the Verge*, vol. Part F1051, pp. 375–403, 2022, DOI: 10.1007/978-3-030-88389-8_18.

[16] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Information Systems*, vol. 112, p. 102131, 2 2023, DOI: 10.1016/J.IS.2022.102131.

[17] S. P. Crain, K. Zhou, S.-H. Yang, and H. Zha, *Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond*. Springer US, 2012, pp. 129–161, DOI: 10.1007/978-1-4614-3223-4_5.

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, 10 2018, DOI: 10.48550/arXiv.1810.04805.

[19] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 3 2022, DOI: 10.48550/arXiv.2203.05794.

[20] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, and M. V. Martínez, "pysentimiento: A python toolkit for opinion mining and social nlp tasks," 2024, DOI: 10.48550/arXiv.2106.09462.

[21] Z. O. Merhi and A. O. Awonuga, "The role of uterine fundal pressure in the management of the second stage of labor: a reappraisal," *Obstetrical & gynecological survey*, vol. 60, pp. 599–603, 9 2005, DOI: 10.1097/01.OGX.0000175804.68946.AC. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16121114/

[22] G. Acmaz, E. Albayrak, G. Oner, M. Baser, G. Aykut, G. Tekin, G. Zararsiz, and I. Muderris, "The effect of kristeller maneuver on maternal and neonatal outcome," *Archives of Clinical and Experimental Surgery (ACES)*, vol. 4, p. 29, 2015, DOI: 10.5455/ACES.20140328024258.

[23] A. Youssef, E. Brunelli, L. Bianchini, M. G. Dodaro, F. Bellussi, and G. Salsi, "Fundal pressure in the second stage of labor: time to face the invisible enemy," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 34, no. 18, p. 3094–3095, 2019, DOI: 10.1080/14767058.2019.1677600.

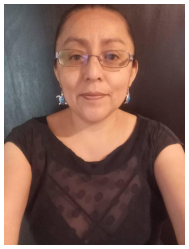
[24] S. Tongate and J. D. Gibbs, "Nurses, physicians and disagreements about fundal pressure: how we used evidence to change practice," *Nursing for women's health*, vol. 14, pp. 137–142, 2010, DOI: 10.1111/j.1751-486X.2010.01527.x. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20409137/

[25] A. Malvasi, S. Zaami, A. Tinelli, G. Trojano, G. Montanari-Vergallo, and E. Marinelli, "Kristeller maneuvers or fundal pressure and maternal/neonatal morbidity: obstetric and judicial literature review." *Matern Fetal Neonatal Med.*, vol. 32, no. 15, pp. 2598–2607, august 2019, DOI: 10.1080/14767058.2018.1441278.

[26] Y. Zhang, M. Wang, C. Ren, Q. Li, P. Tiwari, B. Wang, and J. Qin, "Pushing the limits of llm capacity for text classification," *arXiv*, 2024, DOI: 10.48550/arXiv.2402.07470.

[27] P. Lepagnol, T. Gerald, S. Ghannay, C. Servan, and S. Rosset, "Small language models are good too: An empirical study of zero-shot classification," *arXiv*, 2024, DOI: 10.48550/arXiv.2404.11122.

[28] T. Hu and X.-H. Zhou, "Unveiling llm evaluation focused on metrics: Challenges and solutions," *arXiv*, 2024, DOI: 10.48550/arXiv.2404.09135.



M. Vazquez Hernandez received Ph.D. degree in Electrical Engineering with a specialty in Bioelectronics at CINVESTAV-IPN in 2006. She received her Bachelor's degree in electronic engineering at the Technological Institute of Puebla in 2000. In 2007, she joined the Department of Computer System Engineering and Automation at IIMAS-UNAM, where she worked on processing signals and incorporating gender perspectives into her research.

Helena M. Gómez-Adorno received Ph.D. in Computer Science at the Center for Computing Reseach, IPN. Currently, she is a researcher at the Institute of Research in Applied Mathematics and Systems (IIMAS), National Autonomous University of Mexico (UNAM). Her research interests are in natural language processing and text mining. She has worked on question-answering systems, semantic similarity, authorship attribution, author profiling, and text classification problems. She is a current Mexican National System of Researchers of SECI-HTI Level 1 member.

Natalia Lerín-Hernández received Bachelor degree in Physics at UNAM, where she supported research in the area of Statistical Physics and Complex Systems. She also has a Bachelor degree in Data Science at UNAM, supporting research in Natural Language Processing.

Israel Islas Barajas is a master's student in Government and Public Affairs, UNAM. Specialist in Public Security and a graduate in Criminology. His research interests are the digital transformation of public administration, management of government institutions, public security, justice, and violence.

Orlando Ramos-Flores received a Ph.D. in Language & Knowledge Engineering from Benemérita Universidad Autónoma de Puebla (BUAP). Subsequently, he completed a postdoctoral research stay at IIMAS-UNAM, where he applied Named Entity Recognition to extract information from Electronic Health Records. His research interests span Natural Language Processing, Information Extraction, Information Retrieval, Machine Learning, Deep Learning, and Web Development.