www.editada.org

*International Journal of Combinatorial Optimization Problems and Informatics, 17(5), 2026, 14-36. ISSN: 2007-1558.*
*https://doi.org/10.61467/2007.1558.2026.v17i5.1404*

# Multilevel Detection of Symbolic Violence Against Women in Reddit's Tech Communities Using BERT

Maricarmen Camacho-Pérez, Mireya Clavel-Maqueda, Eduardo Cornejo-Velazquez
Universidad Autónoma del Estado de Hidalgo, México.

**Abstract.** This study introduces an automatic system designed to identify and examine structural and symbolic violence against women within tech-focused Reddit subreddits. Rather than tracking overt hostility, this pipeline targets nuanced marginalization, such as the demeaning of technical proficiency, systemic gender profiling, and the normalization of exclusionary work cultures. The methodology relies on a four-tiered computational architecture: First, it uses a rule-based lexical module utilized for initial benchmarking; Then a trained BERT model identifies the type of violence; Third, a post-processing refinement layer calibrated to mitigate false positives; Finally a contextual classifier capable of isolating active verbal aggression (conversational stance) from first-hand personal narratives (testimonial stance).A dataset comprising 184,572 text entries spanning 17 distinct digital communities was evaluated over a ten-year timeframe (2016–2026). The optimized BERT model demonstrated robust performance, yielding an accuracy of 84.90% and a macro-F1 metric of 0.85, effectively uncovering 23.3% of violent instances in contrast to the restricted 9.4% retrieved by the baseline lexical rules. Empirical evidence indicates that while 22.3% of the total corpus exhibits markers of symbolic violence, an overwhelming 99.9% of these instances belong to testimonial discourse. This dynamic highlights that these forums operate primarily as critical environments for mutual support and whistleblowing rather than platforms for perpetrating harassment. The analysis over time shows a statistically significant downward trend ($r^2$=0.58, p=0.0067), with a reduction of about 0.36 percentage points per year. However, the presence of violence stays above 20% for most of the period. This suggests that despite social movements like #MeToo or the Google Walkout, symbolic violence is still a structural problem in the tech sector. Finally, subtle violence is more common in technical communities, while direct violence is more common in support communities.
**Keywords: S**ymbolic violence, Reddit, and technology.

Article Info
*Received April 21, 2026*
*Accepted Aug 23, 2026*

## 1 Introduction

The tech sector is currently one of the most dynamic areas of the global economy. Even so, women in this space constantly deal with quiet forms of discrimination that people rarely notice (World Economic Forum, 2024). This is not always about open insults or direct harassment. Instead, there is a deeper and more subtle type of hostility called symbolic violence. Bourdieu (2000), the French sociologist who created this concept, explained that these forms of domination become so normalized that both perpetrators and victims accept them as part of daily life.

Within tech environments, this subtle violence shows up in very specific ways. For example, people often question a woman's technical skills for no reason, or they practice mansplaining, which means explaining things to a woman in a condescending way when she already knows the topic. Another common issue is when people claim a woman's professional success is just a result of diversity policies instead of her own merit, all while normalizing a toxic work culture (Benítez-Eyzaguirre, 2021).

These actions do a lot of damage because, since coworkers do not see them as active aggression, they keep happening without any resistance inside companies.

Digital platforms, especially the spaces on Reddit (2026), offer a great window to study these social dynamics. Reddit uses specific forums called subreddits based on distinct topics, where users share their ideas, talk about their days, and help each other. Groups like r/TwoXChromosomes, r/WomenInTech, and r/girlsgonewired connect thousands of women in the tech world. Very often, these users post detailed stories about the bias and discrimination they face at work.

Despite the exponential growth of research on the automatic detection of hate speech, existing systems focus predominantly on explicit forms of aggression such as insults, threats, and obscene language, and have significant limitations when it comes to identifying symbolic violence. This is due to two fundamental reasons. First, the contextual nature of subtle violence requires deep semantic understanding, not merely the detection of keywords. Second, in support communities such as those mentioned, most texts addressing violence are testimonial accounts (people recounting what happened to them) rather than direct expressions of violence within the text itself.

Despite growing research on Reddit and gender-based violence in digital environments, large-scale longitudinal analysis of symbolic violence in specific tech communities remains an under-explored area. This study aims to address this gap in the literature.

This paper presents an automated analysis system that addresses these limitations through a multi-level architecture. The BERT-based model detects violence with greater accuracy (20.6% vs. 8.0% for the rule-based classifier), distinguishing between active violence and discourse about violence, across a decade-long corpus of Reddit communities in tech-related spaces.

### 1.1 Objective

To analyze symbolic violence against women in Reddit's tech communities using a system based on natural language processing and the BERT model, which allows us to describe how it manifests and how it has changed over time, in order to highlight its presence in spaces where issues of gender and technology intersect.

## 2 Theoretical Framework

### 2.1. Symbolic Violence and Its Manifestation in Digital Environments

The concept of symbolic violence, proposed by Bourdieu (2000), refers to a form of domination that is exercised imperceptibly through cultural and social structures. Symbolic violence does not depend on open force like verbal or physical abuse. Instead, it works because both the people who cause it and those who live through it see it as something normal, legitimate, or just a part of life (Bourdieu, 2000). The World Health Organization (WHO, 2021) shows that this type of bias affects women in many areas of their lives worldwide. At the same time, UN Women (2025) points out that around one in three women globally has faced some type of violence, which includes subtle forms that victims themselves might not recognize immediately. According to the World Report on Violence and Health (WHO, 2021), symbolic violence is actually one of the most invisible types of discrimination because it travels through everyday messages, stereotypes, and social images that keep inequality alive without using direct attacks.

In digital environments, these problems expand significantly. The Global Education Monitoring Report Team (2024) explains that digital violence against women is simply an extension of real-world structural discrimination moving into online platforms, showing up as anything from clear harassment to quiet ways of pushing women out. Along the same line, Sierra Caballero & Alberich Pascual (2021) study how digital tools create new spaces where people build their identities and repeat traditional power dynamics. Furthermore, Gómez Herrera & Cañedo (2025) shows that women on platforms like Twitch face sexist behavior that mixes explicit aggression with subtle ways of making them feel invalid.

This symbolic violence takes on a very specific shape in the tech sector. Benítez-Eyzaguirre (2021) studies the brogrammer phenomenon, a culture that combines the word bro with programmer to make a hostile, male centered tech environment seem normal. Ruiz (2023) looks deeper into this idea, explaining that brogrammer culture acts as a wall that keeps women away by spreading stereotypes about who belongs in the programming world. This culture leads to people constantly doubting the technical capacity of women or claiming that their professional success is due to diversity policies rather than their actual talent. Because of this, tech companies often accept environments where having a woman on the team is treated as an odd exception (Ruiz, 2021).

Lately, new studies confirm that this bias is not just a human problem. It also hides inside algorithmic hiring systems, which automatically copy and repeat historical gender biases (Delgado Bancalero, 2023). Even with all these known issues, using big data and computational tools to study these dynamics on social platforms on a large scale is still a field that needs much more research.

### 2.2. The Gender Gap in the Technology Sector

The low number of women in science, technology, engineering, and mathematics (STEM) is a well-known problem around the world. UNESCO (2024) points out that the obstacles keeping women away from STEM exist at many different stages of life, starting with childhood gender stereotypes and continuing up to company cultures that push women to leave these careers. Looking at the global numbers, the World Economic Forum (2024) states that women make up only 29.2% of the employees in the tech industry. This percentage drops even more when looking specifically at technical areas like software engineering and data science.

This difference does not happen because women lack talent or interest. Instead, it is caused by structural problems that affect their entire educational and professional path. UNESCO (2020) explains that social expectations and cultural bias are the main reasons that push girls away from STEM topics when they are very young. Looking at real experiences, Hernández (2023) studied stories from women who graduated from technical universities. Her work shows how these professionals face constant doubts about whether they belong in the tech world, forcing them to work twice as hard to prove their skills. Similarly, Perucha (2023) in Spain and Vargas (2023) in Chile found the exact same situations, which proves that this issue repeats itself across different countries.

The impact of this gap goes way beyond individual unfairness. Rivas (2024) explains that when few women participate in tech careers, the design of digital products and services ends up carrying deep male biases. This creates a bad cycle where most technology is built only from a male point of view. Both UNESCO (2024) and UN Women (2024) have made it clear that fixing the gender gap in STEM is not just about social justice. It is also a mandatory requirement to build technology that works for everyone. Finally, Pacheco and Rodriguez (2022) write that giving women more power in these areas requires changing both official company rules and everyday habits at work.

### 2.3. Automatic detection of hate speech and gender bias

Computers have become much better at finding violent or unfair content in text, mostly because of new tools in Natural Language Processing (NLP) and machine learning. Alarcón Becerra (2022) writes that these systems allow us to sort through huge amounts of data automatically, using methods that go from classic TF IDF calculations to complex deep learning setups. In an administrative comparison, ICHI.PRO (2020) analyzed three common ways to categorize text: TF IDF, Word2Vec, and BERT. Their test proved that models built around the Transformer architecture always beat older methods when a system needs to understand the actual context of a sentence.

These Transformer designs, especially Bidirectional Encoder Representations from Transformers (BERT), completely changed the field. Instead of looking at words as isolated pieces, BERT reads the whole sentence at once to grasp hidden meanings like irony or sutil bias. To make these systems work for specific jobs, developers use a step called fine tuning. Devlin et al. (2019) explains that this means taking a model that already knows basic language and training it a bit more on a smaller, targeted dataset. Hu et al. (2022) showed how well this works by creating ConfliBERT, a specialized version built only for studying political fights and violence, which easily beat standard BERT models. For practical setups, Fuentes (2024)

details how to build these tools using Python and Jupyter Notebooks, while Mora Andrade (2024) applies text mining methods to track aggression against women.

When searching for online sexism and misogyny, studies show that these behaviors leave clear linguistic footprints. Frenda et al. (2019) tested this by running support vector machines (SVMs) on three sets of tweets in English. They discovered that the exact same language traits, like vulgar terms or specific words about the female body, identify both problems, hitting accuracy scores between 76% and 89%. This proves that sexism and open hate are just different sides of the same coin. Along the same line, Farrell et al. (2019) did a massive review of Reddit by scanning 6 million posts across 7 subreddits from the manosphere between 2011 and 2018. Using a vocabulary checklist based on feminist theories, they found that anti women posts and violent talk are growing steadily. This is a vital background because it proves you can study Reddit on a huge scale across many years to see how digital conversations change.

Power dynamics in digital spaces heavily dictate which stories surface and which disappear. Foriest et al. (2024) illustrated this by analyzing Reddit conversations on gender violence, contrasting general forums with specific spaces like r/Trans and r/TwoXChromosomes. Applying the Theory of Silenced Groups alongside a linguistic classifier, they mapped out five distinct ways muting occurs online. Crucially, their research points out that standard moderation rules often end up suppressing non dominant voices by accident. We draw directly from this premise: since platform mechanics dictate how violence is validated, our framework explicitly separates active conversational aggression from personal testimonies.

On the technical side, model preparation strategies vary heavily depending on the input. Purnomo et al. (2025) found that while stripping non alphanumeric characters improves BERT performance for long texts, shorter inputs actually perform better when left completely untouched. Their work validated several configuration choices that guided our methodology, such as the superiority of domain specific architectures, extracting features from the last hidden layer, and setting the optimal learning rate at $2\times10^{-5}$. Adding to the case for these models, Shah et al. (2024) demonstrated that Distil BERT can handle sequential learning across multiple tasks without catastrophic forgetting, highlighting the core flexibility of transformer architectures.

As a practical precedent, Lamas Codesido (2024) successfully deployed classification models into a web application designed specifically for Reddit moderation. This proved the viability of using NLP to flag problematic threads in this exact environment. However, a major blind spot persists in much of the current literature. Most approaches fail to separate texts where a user is actively committing violence from those where a user is simply recounting an attack they survived. Building moderation tools or running analytical studies without drawing this line creates significant structural flaws.

### 2.4. Reddit as a Subject of Study

Social and computational sciences increasingly rely on Reddit as a primary digital laboratory. The scale alone makes it invaluable; (Backlinko Team, 2026) indicates the platform hosts over a billion registered users across countless niche interests. But size is only part of the appeal. As Proferes et al. (2021) pointed out in their systematic review, researchers from sociology to computer science gravitate toward Reddit because of its unique architecture. The subreddit based system creates distinct cultural silos, each governed by its own internal norms. This setup provides an ideal environment for comparative research. Following the methodological frameworks of Farrell et al. (2019) and Foriest et al. (2024), our study leverages these structural boundaries to contrast how discourse shifts across women only support networks, mixed professional spaces, and general workplace forums.

Beyond general discourse, these digital spaces actively function as crisis networks. Hui et al. (2026) demonstrated this by analyzing the r/domesticviolence board to see how survivors share web resources. They discovered that over 75 percent of the links circulated there directly addressed the specific needs of users asking for help. This metric strongly justifies using the platform to analyze how people facing violence exchange information, proving these boards operate not just as discussion areas, but as effective peer support ecosystems.

Extracting this vast amount of data relies heavily on web scraping techniques. Calvera-Isabal et al. (2023) along with Cocón et al. (2023) detail the technical automation and available platforms for web extraction. At the same time, GeeksforGeeks (2025) maps out the Python specific pipelines for Reddit. However, technical execution must always balance against ethical guardrails. Scraping digital platforms at large scale raises immediate questions about user consent and privacy expectations.

Brown et al. (2024) mapped out the legal and scientific friction points inherent to this process. Varela-Rodríguez & Vicente-Mariño (2021) echoed this same debate regarding mass content harvesting. Consequently, any computational analysis of this platform must carefully navigate the tension between public data accessibility and responsible research practices.

## 3 Methodology

This research relies on a quantitative Natural Language Processing framework to address its core objectives. Figure 1 outlines the complete architectural pipeline. The initial phase required extracting raw text across 17 distinct Reddit communities via the platform API. Once the information was secured, an exploratory analysis mapped out the foundational characteristics of the dataset. This preliminary stage tracked basic descriptive statistics, historical post distributions, individual subreddit activity, and baseline linguistic patterns. Following the data cleaning and formatting procedures, the primary analytical engine was constructed. The architecture operates across four sequential layers. A lexical rule-based classifier establishes the initial baseline. Subsequently, a custom trained BERT model separates the text into direct violence, subtle violence, or neutral content. A post processing module then filters out false positives, allowing the final context classifier to separate conversational aggression from personal testimonial accounts.

Model calibration relied on the foundational parameters established by Devlin et al. (2019), incorporating modern text classification validation techniques drawn from Shah et al. (2024) and Purnomo et al. (2025). Once the classification system reached full operation, a temporal analysis tracked how symbolic violence shifted throughout the last decade. The final phase synthesized these results for academic dissemination. The upcoming sections detail the specific technical executions for each component within this pipeline.

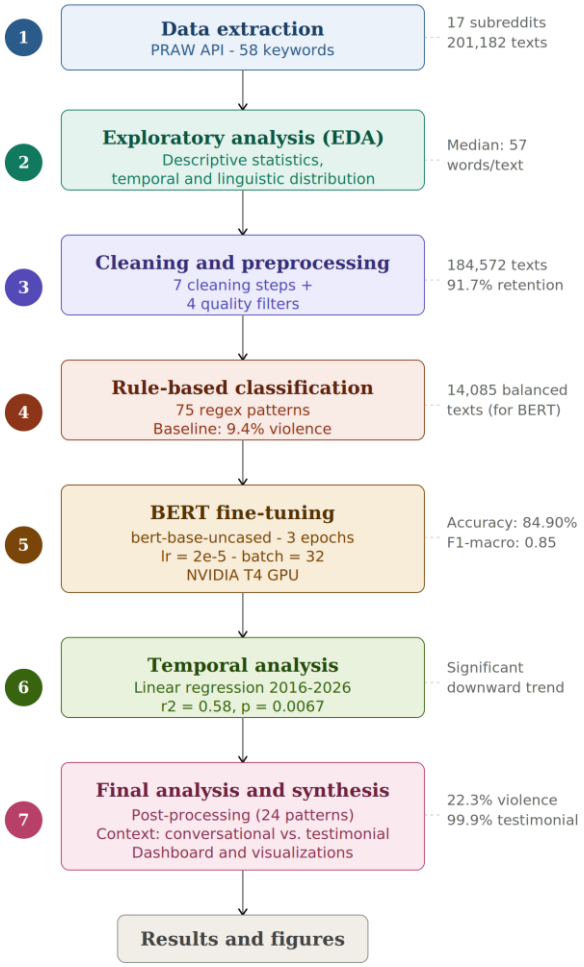


*Fig. 1. Data processing pipeline. Source: Own elaboration.*

### 3.1. Data Collection and Corpus Composition

Information extraction relied on the official Reddit API (2026) operated through a standard Python 3.10 environment. The underlying retrieval mechanism utilized the Python Reddit API Wrapper library at version 7.7.1. All resulting information was stored systematically using Comma Separated Values formatting to facilitate the upcoming analytical phases.

The collection architecture isolated specific metadata points for every recorded entry to ensure comprehensive tracking. Captured elements included a unique digital identifier, the source subreddit, a structural classification denoting either a main publication or a threaded response, the actual written content, the precise publication timestamp, and the final community voting score. The extraction logic depended entirely on a predetermined set of 58 distinct keywords. These terms were clustered into four specialized search modules targeting direct violence, subtle violence, testimonial experiences, and general neutral contexts.

Initial search terms functioned strictly as collection parameters to guarantee representative sampling rather than serving as definitive labels. The BERT architecture executed the final categorization into three primary states (direct violence, subtle violence, and neutral) by evaluating the complete semantic context of entire paragraphs instead of relying on isolated vocabulary matching. Consequently, texts acquired through aggressive keywords frequently received neutral classifications when the broader narrative revealed users actively condemning situations or conducting critical analysis. Section 3.3 details this entire analytical pipeline.

A total of 201,182 texts (posts and comments) were extracted from 17 communities selected for their relevance to the intersection of gender and technology. The selected communities fall into three distinct structural categories depending on their primary function.

Community selection adhered to highly specific parameters. The primary prerequisite mandated a definitive connection to technology, STEM disciplines, or the unique experiences women navigate within these fields, recognizing that evaluating unrelated environments would contribute no relevant data toward the core research objectives. The subsequent requirement focused on sustained engagement levels, specifically targeting spaces characterized by continuous publication frequency and active participation over extended periods to ensure adequate data density for reliable analytical execution.

A third imperative involved integrating diverse environmental archetypes. The evaluation encompassed dedicated female support networks, integrated professional forums welcoming diverse demographics, and generalized discussion platforms. This strategic inclusion proved vital, as restricting observation solely to women centric spaces would have severely truncated the analytical scope and hindered the identification of symbolic violence within ecosystems where gender is not the focal point. Incorporating this triad of community types facilitates a comparative assessment to determine whether the phenomenon fluctuates according to the environment or manifests uniformly across varied platform contexts.

Ultimately, spaces dedicated to entertainment, personal relationships, or topics entirely disconnected from technological or professional spheres were systematically discarded for failing to align with the targeted content parameters.

- Support communities for women in technology: r/TwoXChromosomes, r/WomenInTech, r/girlsgonewired, r/LadiesofScience, and r/WomenEngineers. These five digital environments function as dedicated spaces that encourage mutual assistance and the open exchange of personal lived experiences among female professionals and students within STEM disciplines.
- Mixed-gender professional tech communities: r/cscareerquestions, r/ITCareerQuestions, r/experienceddevs, r/AskComputerScience, r/programming, r/learnprogramming, r/webdev, r/datascience, and r/MachineLearning. These are nine forums where professionals of all genders discuss career topics, training, and professional development.
- General work communities and broad-perspective communities: r/technology, r/antiwork, and r/AskWomen. Three broader spaces where working conditions, tech culture, and women's experiences in different professional fields are addressed.

The corpus includes not only women's support communities but also encompasses the three types described: support communities (5 subreddits), mixed professional communities where people of all genders participate (9 subreddits), and workplace and general-interest communities (3 subreddits). In this way, mixed and general communities serve as a point of

comparison with support communities, allowing for an analysis of whether there are differences in the type and amount of violence depending on space.

Other studies analyzing gender issues on Reddit have employed similar strategies. Farrell et al. (2019) selected seven communities from Reddit's so-called "manosphere" to study how misogyny evolves, explaining that choosing thematic communities allows researchers to observe dynamics that would be lost if a random sample were taken from the entire platform. Similarly, Foriest et al. (2024) compared shared-identity communities (r/Trans and r/TwoXChromosomes) with general discussion spaces (r/AskReddit, r/CasualConversation) to see how the type of community influences the silencing of conversations about gender-based violence. Hui et al. (2026) used r/domesticviolence as a data source to analyze the resources shared among survivors of intimate partner violence, arguing that Reddit offers anonymity and has no character limit, which makes it easier for people to talk about difficult experiences.

### 3.2. Corpus Preprocessing

The initial corpus of 201,182 texts underwent a two-stage cleaning process using the Python libraries pandas (version 2.1.4) and re (regular expressions), as shown in Figure 2.

The first stage was text cleaning, which was applied to each record using a function with the following steps in order: (1) conversion of all characters to lowercase; (2) removal of web links (URLs) using the regular expression r'http\S+|www\.\S+'; (3) removal of user mentions using the pattern r'@\w+'; (4) removal of references to subreddits using the pattern r'r/\w+'; (5) normalization of line breaks and tabs to single spaces; (6) removal of special characters that do not contribute meaning, retaining letters, numbers, apostrophes, and basic punctuation marks; and (7) normalization of multiple spaces to a single space.

The second stage involved applying quality filters to remove records that would not be useful for the analysis: (1) 8,825 posts marked as "[deleted]" or "[removed]" were removed; these were posts deleted by the user or removed by Reddit moderators; (2) 6,902 texts with fewer than 5 words were discarded, as they lack sufficient context for the classifier to function correctly; (3) 883 texts that were not in English were filtered out using a heuristic based on the presence of common English function words (stopwords), since the BERT model used was trained on that language; and (4) content marked as NSFW (adult content) was excluded, as it could skew the analysis.

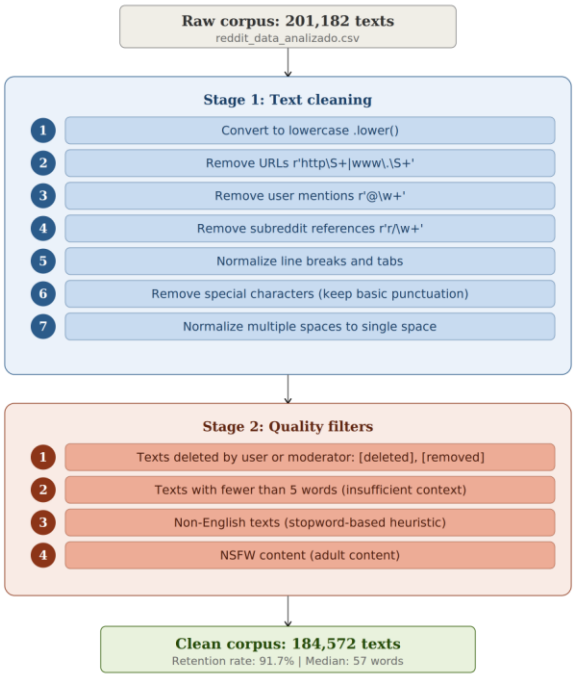


*Fig. 2. Stages of corpus preprocessing. Source: Own elaboration.*

Following this process, the final corpus consisted of 184,572 texts, representing a retention rate of 91.7%. Basic punctuation was retained because it is important for semantic analysis.

The decision not to apply additional techniques such as stopword removal or lemmatization was based on recent studies. Purnomo et al. (2025) tested 16 combinations of preprocessing techniques on three classification datasets using BERT and found that, for short texts, it is best not to remove anything because each word contributes important contextual information that BERT needs. For longer texts, cleaning non-alphanumeric characters was the most helpful. Since the texts in the corpus have a median of 57 words with highly varied lengths, we chose to apply only the cleaning described and leave the stopwords and the original form of the words intact.

### 3.3. Multilevel System Architecture

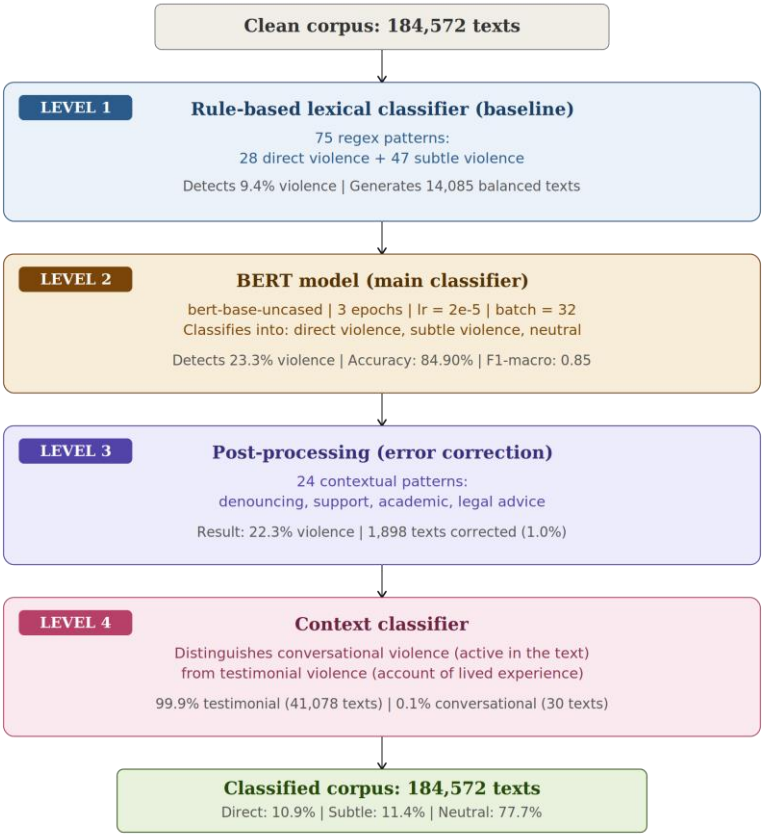


*Fig. 3. Multilevel architecture of the symbolic violence detection system. Source: Own elaboration.*

The analytical framework categorizes the extracted data into three distinct classifications:

- Direct violence: This segment encompasses expressions universally recognized as aggressive or discriminatory without requiring supplementary context. It captures manifestations such as sexual harassment, explicit discrimination, professional exclusion, gendered insults, and overtly hostile conduct. Recurring linguistic patterns include phrases like "sexual harassment", "hostile work environment", "women don't belong in tech", "diversity hire" (when weaponized derogatorily), "toxic culture", and broadly threatening behavior.
- Subtle violence: This grouping contains behaviors frequently overlooked as immediate hostility by the general public yet extensively documented as detrimental within academic literature. This covers dynamics like mansplaining, microaggressions, condescension, and the veiled questioning of professional competence. Typical textual patterns encompass "mansplained to me", "good for a woman", "didn't look like a programmer", "questioned my

competence", "explained my own code to me", "well actually", "only woman on the team", "called bossy for being assertive", and "double standard".

- Neutral category: This final division isolates discourse regarding women in technology completely devoid of symbolic violence. The textual data here consists of professional career advice, success narratives, mentorship discussions, coding bootcamp experiences, and purely technical exchanges lacking any discriminatory undertones.

Each classification tier performs a highly specific function within the broader analytical pipeline.

- Level 1: Lexical rule classifier (baseline). As a baseline of comparison, a classifier based on regular expressions was created. This system searches for predefined text patterns: 75 patterns organized into two categories. The 28 patterns of direct violence were drawn from the work of Farrall et al. (2019) and the dimensions of sexism detection proposed by Frenda et al. (2019). These patterns include expressions of explicit harassment, overt discrimination, workplace exclusion, and direct stereotypes. The 47 patterns of subtle violence were based on Sue's (2010) taxonomy of gender-based microaggressions and on the work on mansplaining and questioning of technical competence documented by Benítez-Eyzaguirre (2021), and they also incorporate patterns of assumed gender roles at work, comments on appearance, double standards, and exclusionary culture. Each pattern was written as a regular expression in Python, which allows for the detection of variations of the same expression. The main function of this level is to serve as a baseline for measuring how much the AI model improves, and to generate the balanced dataset used to train BERT.
- Level 2: BERT model (main classifier). Using the balanced dataset generated in Level 1 (14,085 texts, divided equally into 4,695 texts for each of the three categories: direct violence, subtle violence, and neutral), the BERT-base-uncased model was trained through fine-tuning. In contrast to the initial rule-based classifier, the BERT architecture possesses the capacity to process complete linguistic context. This advanced framework excels at identifying irony and isolating subtle discriminatory patterns. Furthermore, it effectively differentiates between texts that share identical surface structures but carry entirely distinct underlying meanings. Section 3.4 provides a comprehensive breakdown of the specific training methodologies applied to this component.
- Level 3: Post-processing (error correction). This module was responsible for correcting false positives by applying 24 contextual patterns organized into four categories: texts that denounce violence, texts that express support or solidarity, academic or analytical texts, and texts that offer legal or emotional advice. Its main purpose is to assign a neutral label to texts that feature violent vocabulary but actually serve to denounce situations, offer guidance, or provide critical analysis rather than perpetrate actual violence.
- Level 4: Context classifier (conversational vs. testimonial). A final classifier is applied to texts classified as violent to distinguish between two types of contexts. The first is conversational violence, where aggression is actively being perpetrated within the text itself. The second is testimonial violence, where the person recounts an experience of discrimination they experienced at another time or place. The detector also includes negation patterns that prevent texts that are actually defending, advising, or supporting the victim from being classified as conversational. This distinction is key to understanding the role of the communities analyzed.

### 3.4. Technical Setup and Training of the BERT Model

The model used was BERT-base-uncased, obtained from the Hugging Face Model Hub repository. This model was selected for two main reasons supported by Purnomo et al. (2025): first, BERT-base is a reliable starting point when no domain-specific pre-trained model exists, as is the case with symbolic violence; second, its last hidden layer provides the best text representation for classification, which justifies using the [CLS] token, a special token placed at the beginning of each text that functions as a numerical summary of the entire sequence - as input to the classifier. The uncased version is particularly suited for Reddit texts, where informal writing rarely follows capitalization rules. Table 1 presents the model's features, and Figure 4 shows the classification process step by step.

**Table 1.** Features of the BERT model used.

| Hyperparameter | Value | Description |
|---|---|---|
| Base model | BERT-base-uncased | BERT version without case distinction, suitable for informal text |
| Source | Hugging Face Model Hub | Public repository from which the pretrained model was downloaded |
| Total parameters | 109M | Number of internal values the model adjusts during learning |
| Attention layers | 12 | Blocks that process text and capture relationships between words |
| Hidden dimension | 768 | Size of the vector used to represent each word internally |
| Attention heads | 12 per layer | Allow the model to focus on different parts of the text simultaneously |
| Vocabulary (WordPiece) | 30,522 subwords | Dictionary the model uses to split words into recognizable pieces |
| Maximum length | 128 tokens | Maximum number of text pieces the model can process at once |
| Output classes | 3 | Direct violence, subtle violence, and neutral |

Source: Own elaboration

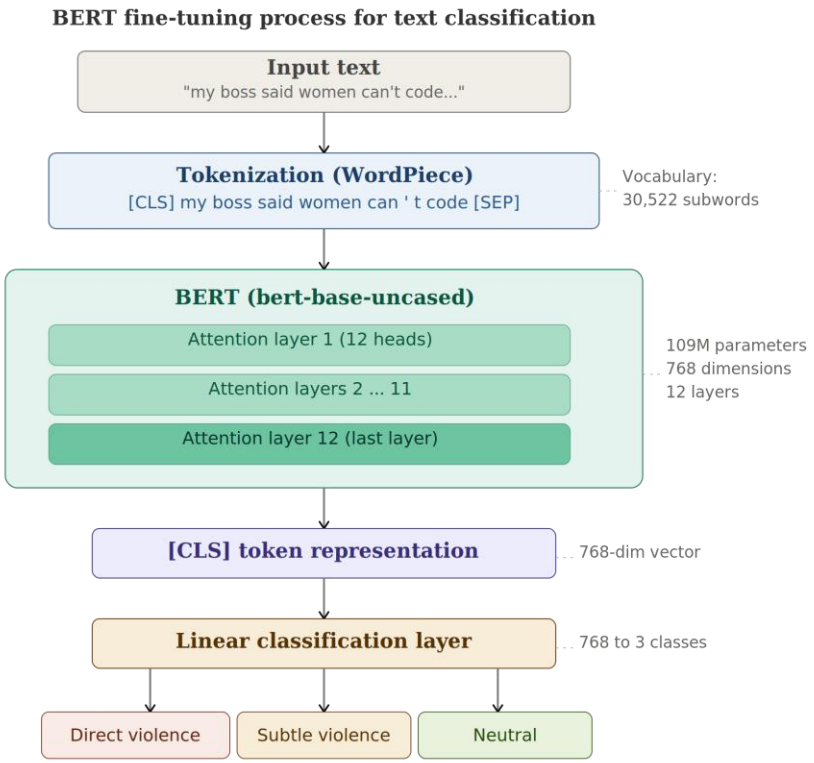


*Fig. 4. BERT fine-tuning process for text classification. Source: Own elaboration.*

Table 2 shows the hyperparameters used to train the model. The learning rate of $2\times10^{-5}$ was selected following the recommendation of Devlin et al. (2019) in the original BERT paper, and is further supported by Shah et al. (2024), who report that low rates with gradual reduction improve model convergence.

**Table 2.** Training hyperparameters.

| Hyperparameter | Value | Description |
|---|---|---|
| Learning rate | $2\times10^{-5}$ | How fast the model adjusts its internal values; lower = finer learning |
| Batch size | 32 | Number of texts the model processes at the same time in each step |
| Training epochs | 3 | Number of times the model goes through the entire training set |
| Optimizer | AdamW | Algorithm that decides how to update the model's values at each step |
| Regularization (weight decay) | 0.01 | Penalty to prevent the model from memorizing data instead of learning |
| Warm-up | 10% of steps | At the start the model learns slowly, then gradually increases the pace |
| Loss function | CrossEntropyLoss | Formula that measures how far predictions are from the correct answer |
| Framework | Transformers 4.40.0 | Software library used to program the training |
| Backend | PyTorch 2.1.2 | Computation engine that executes the model's operations |
| Hardware | GPU NVIDIA T4 (16 GB) | Graphics card that accelerates training computations |
| Platform | Google Colab | Cloud service where the entire process was executed |

Source: Own elaboration

The model parameters included three training epochs and a 128 token maximum limit. These seemingly conservative choices closely follow established guidelines. Devlin et al. (2019) advise using two to four epochs for fine tuning because exceeding this range on smaller datasets often triggers overfitting. Purnomo et al. (2025) similarly demonstrated that stabilization occurs precisely at three epochs with no benefit from extra cycles. The 128 token threshold is also well suited since the corpus median sits at 57 words, ensuring nearly all texts are processed completely. Shah et al. (2024) applied comparable constraints to DistilBERT across multiple datasets and achieved accuracies above 95 percent, confirming the viability of this setup for text classification.

The 14,085 labeled texts were divided into three sets using the `train_test_split` function from scikit-learn (version 1.3.2), as shown in Figure 5: 70% for training (9,859 texts), which the model uses to learn; 15% for validation (2,113 texts), which serve to ensure that the model does not overfit during training; and 15% for final evaluation (2,113 texts), used to calculate the metrics reported in the results section. The division was made with stratification by class; that is, in each of the three sets, the same proportion of direct violence, subtle violence, and neutral content (33.3% each) was maintained so that the model would not learn in an unbalanced manner.

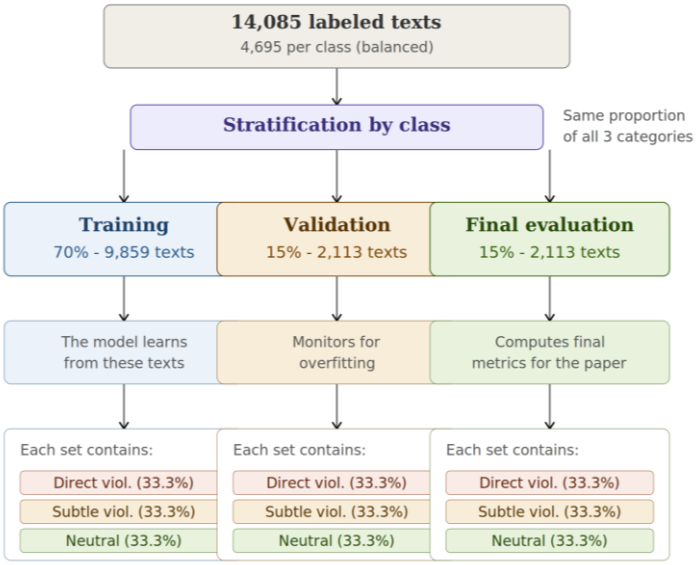


*Fig. 5. Splitting the dataset for BERT training. Source: Own elaboration.*

### 3.5. Temporal Analysis of Symbolic Violence

To assess whether the prevalence of symbolic violence has changed over time, the texts were grouped by year (2016 to 2026), and the percentage of violence detected in each period was calculated. To determine whether there is an upward or downward trend, a simple linear regression model was applied using the `linregress` function from the scipy.stats library. In this model, the independent variable is the year and the dependent variable is the percentage of violence detected.

The analysis reports two primary indicators. The coefficient of determination ($r^2$) evaluates the strength of the trend, while the p value determines if this pattern holds statistical significance rather than resulting from random chance. We consider any p value below 0.05 as statistically significant.

Major historical milestones were incorporated to provide context for year over year fluctuations. These events encompass the #MeToo movement in 2017, the 2018 Google Walkout, the COVID 19 pandemic of 2020, the tech sector layoffs spanning 2022 to 2023, and the 2023 surge in generative artificial intelligence. Instead of functioning as mathematical variables within the statistical model, these periods serve strictly as qualitative interpretive frameworks. Additionally, we examined how violence types and discourse contexts were distributed across this timeframe.

Detected violence represents the combined total of the direct and subtle categories outlined previously. Any texts receiving a neutral classification were completely excluded from this specific metric.

## 4 Results

This section outlines the empirical findings of the study, progressing from broad corpus characteristics to a granular community level analysis. We first examine the manifestation of symbolic violence across the 17 selected spaces and compare the performance of the BERT model directly against the traditional rule-based classifier. The final portion explores the temporal evolution of these behaviors between 2016 and 2026 while highlighting the contrasts observed among the various community types.

### 4.1. General Statistics of the Corpus

Table 3 demonstrates that comments constitute 92.5 percent of the dataset, highlighting the conversational dynamics inherent to the platform. The analytical framework identified symbolic violence in 22.3 percent of the evaluated texts. This total divides into 10.9 percent direct manifestations and 11.4 percent subtle variations, leaving the remaining 77.7 percent categorized strictly as neutral.

**Table 3.** Text distribution and classification frequencies.

| Feature | Value |
|---|---|
| Total texts | 184,572 |
| Posts | 13,865 (7.5%) |
| Comments | 170,707 (92.5%) |
| Texts with direct violence | 20,080 (10.9%) |
| Texts with subtle violence | 21,028 (11.4%) |
| Neutral texts | 143,464 (77.7%) |
| Total violence detected | 41,108 (22.3%) |

Source: Compiled by the author

Figure 6 illustrates these proportions. The first panel maps the structural distribution of the content, while the second panel details the specific classification categories. The data reveals a near parity between direct and subtle aggression, recording 10.9 percent and 11.4 percent. This balance contrasts sharply with older literature that typically observed a heavy majority for overt hostility.

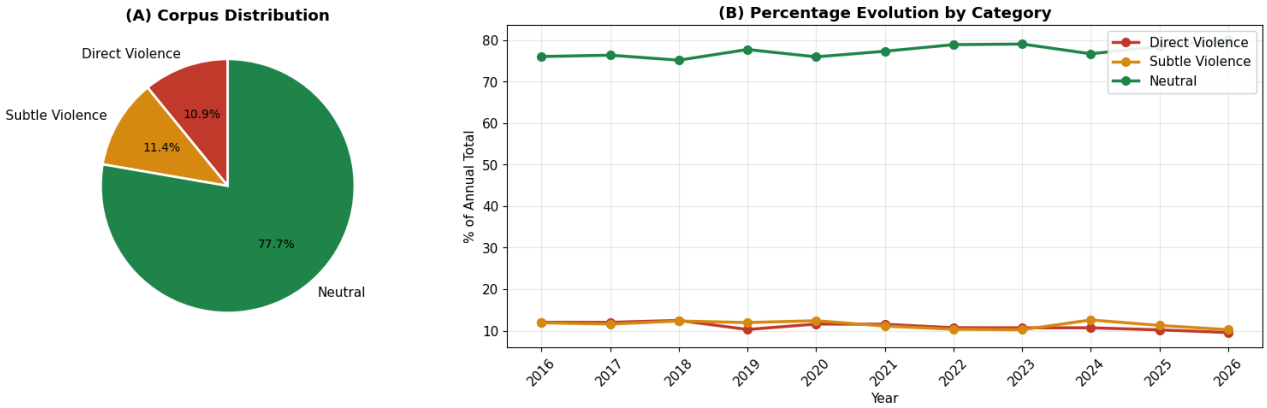


*Fig. 6. Overall system results (A and B).*

Source: Own elaboration.

The visualization in Figure 7 captures the most critical phenomenon observed during the analysis. Panel (C) shows the distribution of texts containing violence by the subreddit from which they originate. Support communities for women (r/TwoXChromosomes, r/WomenInTech, r/girlsgonewired, r/LadiesofScience, and r/WomenEngineers) account for the largest volume of texts containing detected violence, followed by mixed professional communities (r/careerquestions, r/experienceddevs, r/datascience) and general work-related communities (r/antiwork, r/AskWomen). The panel D section of the graphic demonstrates an overwhelming tendency where 99.9 percent of the detected aggression consists of testimonial narratives. Users systematically recount adverse situations originating from external environments instead of directing hostility toward each other. This dynamic remains stable across support, professional, and workplace forums. Participants utilize the platform to process outside experiences rather than initiate internal conflicts. As a result, the analysis identified a mere 30 individual posts throughout the entire dataset, representing 0.1 percent, that qualified as active interpersonal attacks. This indicates that, regardless of the type of community, these subreddits function as spaces for reporting and mutual support.

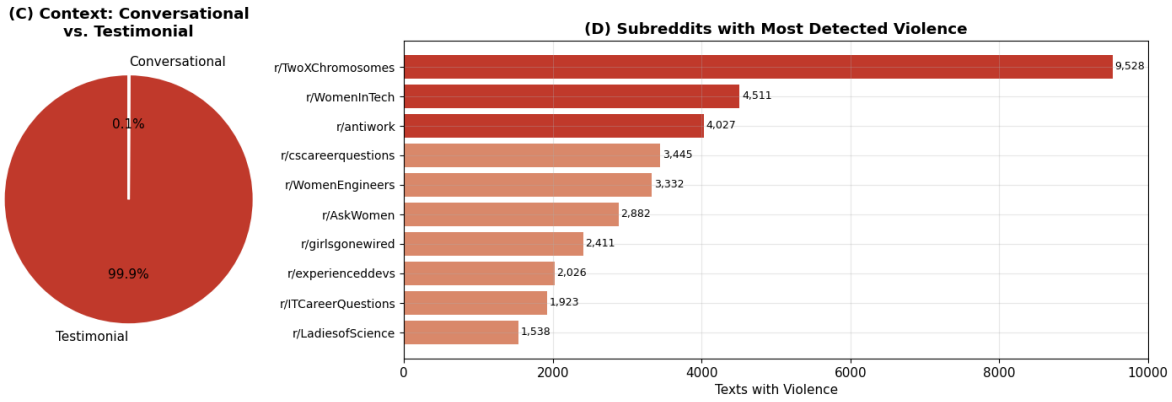


*Fig. 7. Overall system results (C and D).*

Source: Own elaboration.

### 4.2. Performance of the BERT Model

After training the BERT model, evaluation metrics were obtained on the test set, consisting of 2,113 texts, as shown in Table 4. The BERT model achieved an overall precision of 84.90%, with a recall of 0.85 and a macro-F1 score of 0.85, indicating balanced performance across all three categories. Precision measures what proportion of the texts classified in a category belong to it; recall measures what proportion of the texts that belong to a category were correctly identified; and F1 is the harmonic mean of both metrics.

**Table 4.** Evaluation metrics for the BERT model on the test set.

| Class | Precision | Recall | F1 | Support |
|---|---|---|---|---|
| Direct violence | 0.86 | 0.88 | 0.87 | 704 |
| Subtle violence | 0.82 | 0.83 | 0.83 | 705 |
| Neutral | 0.86 | 0.84 | 0.85 | 704 |
| Macro average | 0.85 | 0.85 | 0.85 | 2,113 |

Source: Own elaboration.

To understand the value added by each level of the system, Table 5 compares the amount of violence detected at each stage of the architecture. While the rule-based classifier detected 9.4% of the violence in the corpus, BERT identified 23.3%, that is, 2.5 times more.

**Table 5.** Comparison of detected violence by architectural level.

| **System Level** | **Violence Detected** | **% of corpus** |
|---|---|---|
| Level 1: Rule-based Classifier | 11,977 | 8.0% |
| Level 2: BERT | 30,730 | 20.6% |
| Level 3: BERT + Post-processing | 30,284 | 20.3% |
| Level 4: Conversational Context | 42 | 0.1% of violence |
| Level 4: Testimonial Context | 30,242 | 99.9% of violence |

Source: Own elaboration

The evidence confirms that basic lexical queries fail to capture the full scope of symbolic aggression. Consequently, an architecture capable of deep semantic comprehension becomes essential. A subsequent correction module resolved 1,898 false positives, equating to exactly 1.0 percent of the total dataset. Furthermore, the final contextual evaluation established that 99.9 percent of the flagged hostility originates from personal testimonies.

### 4.3. BERT Model Error Analysis

To properly gauge the predictive strength of the architecture, the methodology incorporated a strictly isolated testing cohort containing 2,113 distinct texts. This validation subset comprised 15% of the balanced dataset and was extracted via stratified sampling utilizing a random state of 42 to guarantee proportional distribution. Researchers purposefully quarantined these specific samples from the computational model during all foundational training cycles. A subsequent review of the confusion matrix maps the authentic categorical designations against the algorithmic output, yielding a generalized accuracy rate of 84.9%. Nevertheless, as Figure 8 explicitly demonstrates, accepting this aggregate figure at face value obscures critical operational nuances that demand a much more granular investigation.


*Fig. 8. Confusion matrix of the BERT model.*

Source: Compiled by the author.

*False Negatives: Undetected Violence within the Test Set*

Such omissions materialize when aggressive expressions lack the explicit vocabulary the architecture internalized during training. A query asking "is 22 too old to learn programming?" harbors a clear ageist microaggression, yet the model classifies it as neutral due to the absence of targeted keywords. Similarly, statements dismissing hostile environments, such as "I do not see how this is harassment", represent discriminatory behaviors that successfully bypass detection. The core limitation lies in the algorithmic necessity to comprehend underlying conversational intent rather than executing isolated semantic analysis.

*False Positives: Erroneously Detected Violence within the Test Set*

The evaluation matrix highlights 113 benign texts incorrectly flagged as violent out of 704 genuinely neutral samples (16.1%). This phenomenon surfaces when the architecture conflates academic or experiential discussions about aggression with actual perpetration. A statement declaring "women leave tech due to microaggressions" incorporates a specific trigger word but functions as an analytical observation rather than an attack. The system registers the term and subsequently mislabels the entire narrative. This specific systemic flaw appears frequently within communities like r/TwoXChromosomes, where participants routinely mention hostile terms while sharing personal testimonials or formulating structural critiques.

*Categorical Confusion within Violence Tiers*

An additional 119 instances (5.6% of the testing cohort) suffered misclassification between the designated severity tiers. Specifically, 66 overt violence texts received a subtle classification (9.4% of the respective class), while 53 subtle narratives were escalated to the direct category (7.5% of the class). This overlap highlights the blurred boundary separating overt aggression from veiled hostility. A comment stating "well, let me explain" directed at a female professional manifests clear

condescension, yet the exact phrasing could simultaneously represent standard technical guidance depending entirely on the surrounding dialogue.

*Error Extrapolation Across the Comprehensive Corpus*

Projecting these localized error rates onto the complete 184,572 text corpus notably alters the foundational metrics. Approximations suggest 17,428 genuinely aggressive texts evaded detection (9.4%), while roughly 9,967 benign entries received incorrect hostility markers (5.4%). Consequently, the authentic prevalence rate likely rests between 11.4% and 21.2%, diverging from the initial 23.3% output reported by the system. Despite these statistical deviations, the architectural implementation remains highly valuable: it identified 2.5 times more targeted content compared to standard lexical rules (23.3% versus 9.4%), empirically validating its methodological deployment.

Comparative literature evaluating online toxicity detection systems (Shah et al., 2024) documents identical misclassification patterns, confirming these error margins align perfectly with standard expectations for complex natural language processing tasks.

### 4.4. Temporal Analysis (2016–2026)

Figure 9 tracks the expansion of community engagement throughout the analyzed decade. Monthly publication volumes display a consistent upward trajectory starting in 2016. A notable acceleration occurs at the onset of 2020, eventually reaching an absolute maximum in 2025. The incorporated dotted boundaries represent the historical milestones previously outlined in section 3.5. These visual markers indicate a clear alignment between specific external events and sudden surges in forum participation.

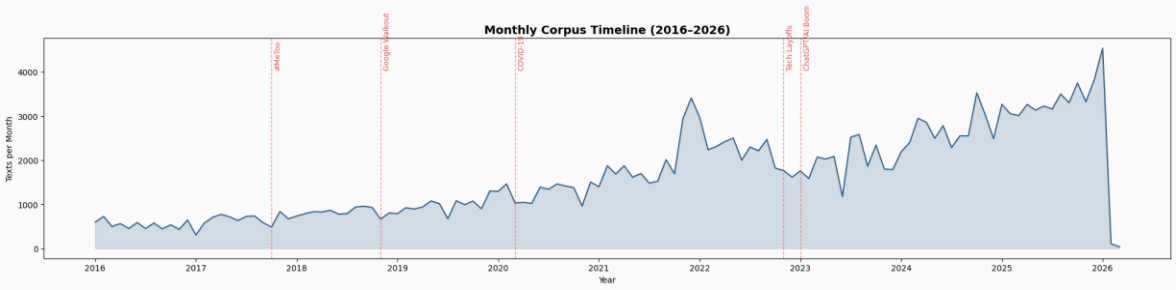


*Fig. 9. Monthly dataset progression from 2016 to 2026 highlighting major historical intersections.*

Source: Own elaboration.

Moving to the chronological evaluation, Figure 10 details the temporal dynamics. The initial panel captures the absolute expansion of recorded texts. This metric scales from an initial 5,842 entries in 2016 to an apex of 37,227 in 2025, reflecting the organic adoption rate of these specific platform spaces. Most importantly, the data confirms that neutral interactions multiply at a substantially higher velocity compared to hostile manifestations. Panel (B) is the most revealing: the percentage of total violence ranges from 20.9% to 25.9%, with a peak in 2018 coinciding with the Google Walkout and a spike in 2020 associated with the impact of the COVID-19 pandemic on women's working conditions in the tech industry. The regression analysis yielded an $r^2$ of 0.58 and a p value of 0.0067. This annual decrease of 0.36 percentage points confirms a statistically valid downward trajectory. This means that the proportion of symbolic violence has been gradually decreasing over the decade, falling from approximately 25% in 2016 to 21% in 2026.

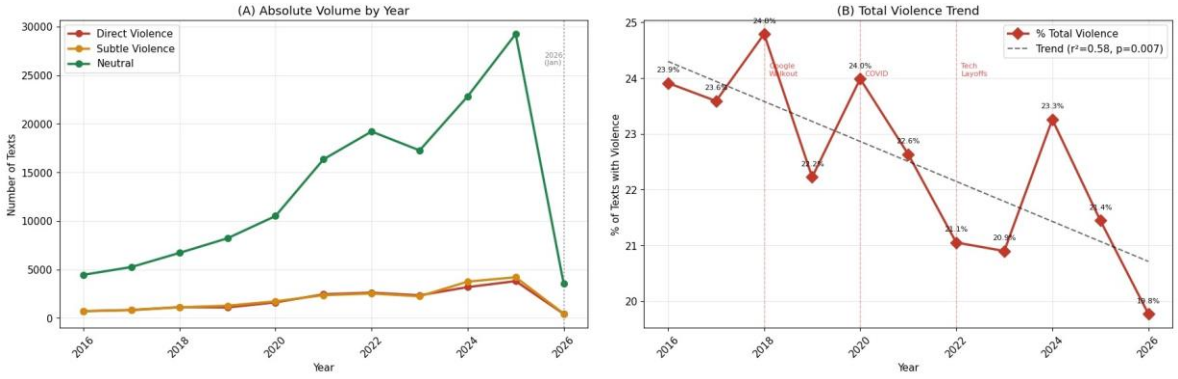


*Fig. 10. Temporal analysis of symbolic violence A-B (2016–2026). Source: Own elaboration.*

Figure 11, panel (C), confirms that the percentage breakdown among the three categories shows gradual changes over the course of the decade. Panel (D) compares conversational and testimonial violence, confirming the overwhelming dominance of the testimonial context throughout the entire period.

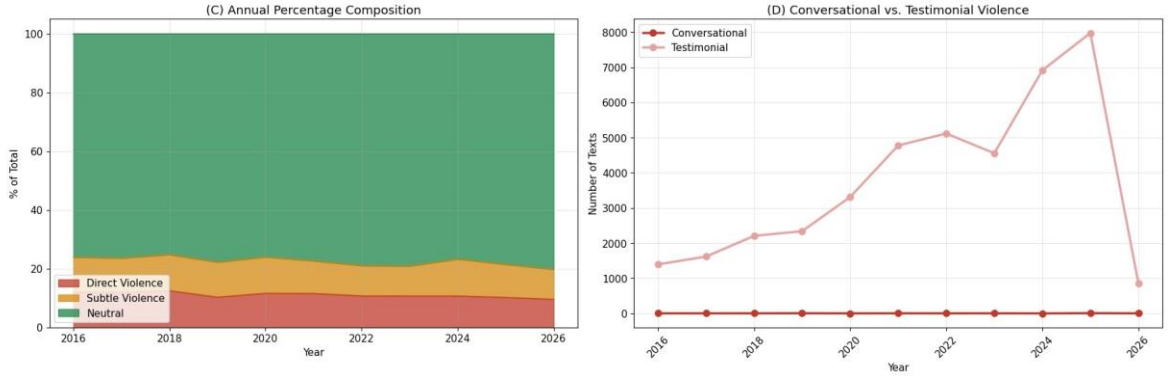


*Fig. 11. Temporal analysis of symbolic violence C-D (2016–2026). Source: Own elaboration.*

Despite this meaningful decline, overall prevalence stays above 20% throughout most of the timeframe. Prominent media events like #MeToo triggered sudden bursts of engagement, yet the overall reduction stems from a gradual shift rather than isolated interventions.

### 4.5. Analysis by community

Figure 12 outlines clear behavioral distinctions depending on the forum category. Technology focused groups display a higher proportion of subtle violence. Discrimination in these specific spaces frequently translates into implicit skepticism regarding technical competence, mansplaining, and microaggressions. Conversely, r/TwoXChromosomes and r/antiwork exhibit the exact opposite configuration. Direct violence predominates in these boards, aligning perfectly with their primary role as platforms for explicitly denouncing severe incidents.

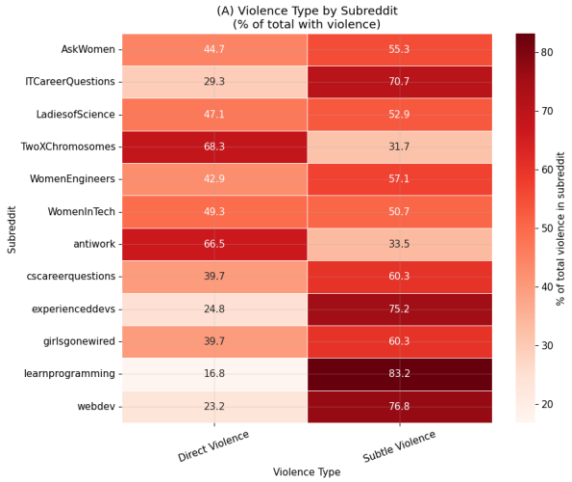


*Fig. 12. Distribution of violence by community and type.*

Source: Own elaboration.

Activity fluctuations across each subreddit over the years highlight distinct evolutionary paths (Figure 13). Groups like r/WomenInTech and r/WomenEngineers concentrate the vast majority of their engagement between 2024 and 2025, mirroring their recent rapid expansion. Meanwhile, r/TwoXChromosomes and r/cscareerquestions sustain remarkably consistent engagement metrics throughout the decade. Furthermore, r/AskWomen records an evenly distributed volume of posts across the entire observed period.

Regarding absolute volume, r/TwoXChromosomes holds the highest overall count of violent posts. The forums r/antiwork, r/cscareerquestions, and r/WomenInTech follow sequentially in total numbers.

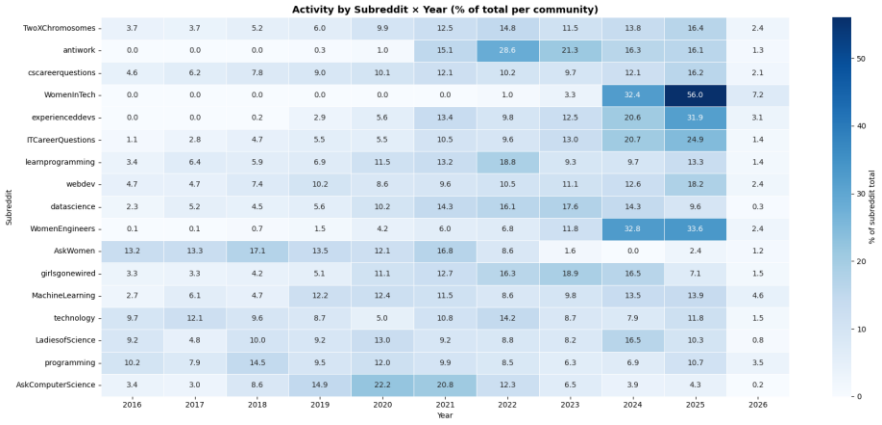


*Fig. 13. Activity by subreddit and year (percentage of the total per community).*

Source: Own elaboration.

### 4.6. Validation Using Real Examples

A qualitative assessment of random excerpts from the classified corpus ensured semantic consistency across the model. This manual review verified whether the algorithmic labels accurately captured the core textual content. Representative instances detailing direct violence are documented in Table 6, while Table 7 provides corresponding examples of subtle violence. All texts come from the real corpus and were selected using fixed-seed random sampling (random_state=42) to ensure reproducibility.

**Table 6.** Real-life examples of direct violence.

| Context | Subreddit | Year | Text (fragment) |
|---|---|---|---|
| Conversational | r/TwoXChromosomes | 2022 | "incompetent women being labelled as diversity hires...while incompetent men are just labelled incompetent. i'm so sick of this double standard..." |
| Conversational | r/TwoXChromosomes | 2016 | "im very sorry youres going through this but i really dont see what is harassment about this. usually when people say they hate me for no reason its never no reason." |
| Testimonial | r/LadiesofScience | 2021 | "comments about having earned or achieved something just because of being a woman are so annoying and undermining! i scored a great score on a grant and one of my labmates had the audacity to ask if i earned the grant by fulfilling a diversity quota." |
| Testimonial | r/TwoXChromosomes | 2020 | "nope, get out. why waste any time or energy for such a small-minded bunch. when you get a job and leave make sure you are really damn clear on how their sexism and gender bias played into it..." |

Source: Own elaboration

**Table 7.** Real-life examples of subtle violence.

| Context | Subreddit | Year | Text (fragment) |
|---|---|---|---|
| Conversational | r/cscareerquestions | 2021 | "i am the first female on my team. usually i don't feel any different at all, other than an irrational anxiety about oh my god if i don't perform well my team will think girls can't do programming." |
| Conversational | r/TwoXChromosomes | 2025 | "virginia tech, 1967,69-1977...engineering prof. women don't belong in this field and i will fail you (paraphrasing as it's been a long time) both switched to bio." |
| Testimonial | r/WomenInTech | 2025 | "even after being in the tech field, from help desk to software development, i still feel like a fraud from time to time. that kind of never goes away. i'm just better at dealing with it." |
| Testimonial | r/TwoXChromosomes | 2025 | "told a date what i do in my line of work. he corrected me and said that is not at all what people in that profession do. he knew because he had spoken to some people at a party. there was no convincing him of anything else." |
| Conversational | r/cscareerquestions | 2021 | "i am the first female on my team. usually, i don't feel any different at all, other than an irrational anxiety about oh my god if i don't perform well my team will think girls can't do programming." |

Source: Own elaboration

Instances detailed in Table 6 feature sexual harassment, explicit workplace discrimination, and insults targeting gender. These elements precisely match the taxonomy boundaries established for the classification system. Conversely, Table 7 details subtle microaggressions that typically evade standard lexical filters. These include being dismissed as "argumentative" for sharing ideas, externally triggered imposter syndrome, male peers correcting female professionals on their own specialized work, technical condescension, and stereotypes regarding female mathematical proficiency. Such behaviors mirror the microaggression framework targeting gender documented by Sue (2010).

Conversational violence accounts for a negligible fraction of the findings, comprising exactly 30 texts or 0.1 percent of the total detected. A close review of this specific corpus demonstrates that users in these instances are typically quoting hostile phrases strictly to criticize or expose them, rather than committing active aggression. One sample labeled under this category includes the quote "women can't be engineers," yet the surrounding paragraph actively condemns double standards.

## 5 Discussion

An analysis of 184,572 text posts from REDDIT, published from 2016 through 2026 in 17 technology-related communities, provided substantial evidence that symbolic violence has a decreasing tendency at a rate of 0.36% per year ($r^2$=0.58, p=0.0067), dropping from 25% in 2016 to 21% in 2026.

However, the percentage of symbolic violence in student communities is rampant while the rate of decline is non-existent. Global phenomenons related to manifestations against violence targeted towards women (MeToo and Google Walkout) corresponded with the analyzed period of symbolic violence. The identified consequences of symbolic violence indicate that this form of aggression gradually concedes into cultural shifts, but is still entrenched in the culture and persists across time.

Online discourse or the way people discuss these issues may be related to the diminishment of violent remarks. Moreover, it was evidenced that 99% of texts classified as violent correspond to testimonials, which proves that the analyzed online communities function as support groups where women can speak out against abuse and violence.

The results reveal disparate patterns: in technical subreddits, subtle violence dominates the discourse such as questioning women's technical capacity, mansplaining, or criticizing the hiring policies related to gender parity disguised as a jab at women. On the other hand, in online support groups, such as r/TwoXChromosomes, violence is not masked and instead it is explicitly addressed by community members through their experiences and reports with harassment and discrimination.

## 6 Conclusions

The objective of this work was to analyze symbolic violence against women in Reddit's technology communities using a system based on natural language processing and the BERT model, which also describes how it manifests and changes over time.

This objective was achieved through the implementation of a BERT model that was capable of distinguishing between texts of direct violence, unobtrusive masked violence and neutral stances. Furthermore, a classifier was employed to categorize texts as either anecdotes of personal experiences with violence or discourse that perpetuates violence.

Regarding temporal evolution, a decreasing tendency was identified with a reduction of 0.36% per year. Social campaigns and diversity policies have had a moderate effect of ease over this structural problem that requires even more profound efforts to overturn.

The BERT model detected 23.3% violence compared to a 9.4% found by the rule-based method, with an accuracy of 84.90% and an F1-macro of 0.85. BERT found 2.5 times more violence instances because symbolic violence is expressed in ways or patterns that cannot be captured by simply searching for key words; BERT captures and understands the full context. The post-processing module and context-classification module both demonstrated their significance in the implementation of the system: the first one corrected 1,898 false positives and the latter one added an interpretation layer that a binary classifier could not have offered.

The ability to identify and discriminate between texts that this work offers is very important, since due to the inconspicuous nature of subtle violence it makes it difficult to detect or expose, and thus is the most prevalent and unnoticed type of violence in professional environments. But above all, having the possibility to perform the identification and even the classification of such violence through automated methods, that is susceptible to betterment and replication, is in itself a remarkable contribution of this work, both in the technical and social scope.

These findings can be of great benefit for the design of public policies and preventative and corrective programs such as retention strategies and the development of female talent in tech companies. Interventions should not only target instances of evident acts of discrimination and violence but also commonplace dynamics of power.

Limitations

Interpretations of these metrics remain bound by specific methodological constraints. The corpus relies entirely on English language data, preventing direct generalization to Spanish speaking networks or other linguistic demographics. Furthermore, utilizing the initial lexical rules to label training data introduces a circular bias, potentially restricting the BERT architecture from recognizing violent patterns omitted from the baseline vocabulary. Securing independent human annotation and calculating Cohen kappa coefficients are essential next steps to validate labeling integrity. Finally, the context classifier generated residual false positives when users quoted hostile phrases for criticism. Integrating negation patterns successfully lowered these specific errors from 41 to 30 instances, though complete eradication remains pending.

It is crucial to acknowledge that Reddit does not serve as a universal representative sample for all women navigating the technology sector. The platform demographic predominantly consists of individuals fluent in English, located in developed nations, and possessing advanced educational backgrounds. Female professionals who avoid this specific site, whether due to personal preference, limited digital access, or linguistic barriers, remain completely absent from the current analytical framework. Alternative online environments, including Twitter, LinkedIn, or private digital forums, likely harbor significantly different manifestations of symbolic violence. Consequently, these empirical findings exclusively characterize technological communities housed within Reddit between 2016 and 2026. These overarching conclusions cannot be directly extrapolated to other social networking platforms, distinct geographical territories, or physical workplace environments.

Future work

Subsequent research phases will prioritize adapting the infrastructure for the BETO model to evaluate Spanish speaking environments and contrast behavioral trends. Validating the entire dataset through human annotators to establish robust Cohen kappa metrics remains a primary objective. Expanding the analytical scope to encompass highly technical environments like Stack Overflow or GitHub will provide further comparative depth.

Regarding the model configuration, a maximum token length of 128 and 3 training epochs was used, which has been shown to be sufficient for text classification tasks according to the literature reviewed (Devlin et al., 2019; Shah et al., 2024; Purnomo et al., 2025), as detailed in Section 3.4. Researchers should also investigate whether expanding the maximum sequence length to 256 or 512 tokens enhances detection capabilities within longer publications. Ultimately, comparing these computational findings against public metrics concerning gender diversity within technology companies could expose tangible links between digital discourse and physical workplace environments.

**Data and code availability**. The complete analytical pipeline, underlying source code, and public dataset identifiers remain fully accessible via: https://github.com/mariimanguito/ViolenciaTechReddit

**Declaration of generative AI in the writing process.** During the preparation of this work, Generative AI was used to support minor tasks like language refinement, text structuring and readability improvements in this manuscript. After that the authors review and edit the content and take full responsibility for the final version of this paper.

## References

Alarcón Becerra, A. E. (2022). *Clasificación de texto con técnicas de procesamiento del lenguaje natural y machine learning para analizar la relación entre noticias publicadas en la web y la variación de los índices de la bolsa de comercio en Chile* [Master's thesis, Universidad del Desarrollo]. Repositorio Universidad del Desarrollo.

Backlinko Team. (2026, January 16). *Reddit user and growth stats*. Backlinko. https://backlinko.com/reddit-users

Benítez-Eyzaguirre, L. (2021). La cultura brogrammer y el sexismo digital. In F. Sierra Caballero & J. Alberich Pascual (Eds.), *Cultura digital, nuevas mediaciones sociales e identidades culturales* (pp. 71–96). Comunicación Social Ediciones y Publicaciones.

Bourdieu, P. (2000). *La dominación masculina*. Anagrama.

Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., & Zimmer, M. (2025). Web scraping for research: Legal, ethical, institutional, and scientific considerations. *Big Data & Society, 12*(4), 1–17. https://doi.org/10.1177/20539517251381686

Calvera-Isabal, M., Santos, P., Hoppe, H.-U., & Schulten, C. (2023). Cómo automatizar la extracción y análisis de información sobre ciencia ciudadana con propósitos educativos. *Comunicar, 31*(74), 23–35. https://doi.org/10.3916/C74-2023-02

Cocón, F., Pérez-Cruz, D., Pérez-Rejón, J. Á., Zavaleta-Carrillo, P., Barradas-Arenas, U., Gómez-Ramón, R., & Pérez Cruz, J. A. (2023). Web scraping: Uso de plataformas de extracción de datos aplicadas a un sitio web sobre profesiones en México. *RISTI – Revista Ibérica de Sistemas e Tecnologias de Informação, 52*, 61–73. https://doi.org/10.17013/risti.52.61-73

Delgado Bancalero, R. (2023). *La discriminación de la mujer en la selección algorítmica de profesionales de la comunicación* [Master's thesis, Universidad de Cádiz]. Repositorio UCA. http://hdl.handle.net/10498/29442

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

*Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Di Pietro, M. (2020, July 18). *Text classification with NLP: Tf-Idf vs Word2Vec vs BERT*. Towards Data Science. https://medium.com/data-science/text-classification-with-nlp-tf-idf-vs-word2vec-vs-bert-41ff868d1794

Farrell, T., Fernandez, M., Novotny, J., & Alani, H. (2019). Exploring misogyny across the manosphere in Reddit. In *Proceedings of the 11th ACM Conference on Web Science* (pp. 87–96). Association for Computing Machinery. https://doi.org/10.1145/3292522.3326045

Foriest, J. C., Mittal, S., Bray, K., Tran, A.-T., & De Choudhury, M. (2024). A cross community comparison of muting in conversations of gendered violence on Reddit. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW2), Article 401. https://doi.org/10.1145/3686940

Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. *Journal of Intelligent & Fuzzy Systems, 36*(5), 4743–4752. https://doi.org/10.3233/JIFS-179023

Fuentes, F. (2020, October 16). *Clasificación de textos con Python y Jupyter Notebooks*. Arsys. https://www.arsys.es/blog/clasificaciontextos-python-jupyternotebooks

GeeksforGeeks. (2025, July 23). *Scraping Reddit using Python*. https://www.geeksforgeeks.org/python/scraping-reddit-using-python/

Global Education Monitoring Report Team. (2024). *Global education monitoring report 2024, gender report: Technology on her terms*. UNESCO. https://doi.org/10.54676/WVCF2762

Global Education Monitoring Report Team. (2024). *Support girls and women to pursue STEM subjects and careers: Advocacy brief*. UNESCO. https://doi.org/10.54676/BPSL3344

Gómez Herrera, M. J., & Cañedo, A. (2025). Mujeres streamers en Twitch: Ciberfeminismo frente a las violencias machistas. *Teknokultura. Revista de Cultura Digital y Movimientos Sociales, 22*(2), 209–217. https://doi.org/10.5209/tekn.97131

Hernández Herrera, C. A., & Hernández Herrera, M. C. (2023). Revelando la brecha de género en STEM: Experiencias de mujeres egresadas de un Instituto Tecnológico Federal. *Acta Universitaria, 33*, 1–14. https://doi.org/10.15174/au.2023.3862

Hu, Y., Hosseini, M. S., Skorupa Parolin, E., Osorio, J., Khan, L., Brandt, P. T., & D'Orazio, V. J. (2022). ConfliBERT: A pre-trained language model for political conflict and violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5469–5482). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.400

Hui, V., Tian, L., Eby, M., Zhang, B., & Constantino, R. E. (2026). Evaluating website resources shared online amongst women with intimate partner violence experiences: Analysis of an online health community. *Frontiers in Psychology, 17*, 1750050. https://doi.org/10.3389/fpsyg.2026.1750050

Lamas Codesido, M. (2024). *Aplicación web para la ayuda a la moderación de foros de Reddit usando modelos de clasificación* [Bachelor's thesis, Universidade da Coruña]. Repositorio UDC. http://hdl.handle.net/2183/39848

Mora Andrade, S. E. (2024). *Aplicación de técnicas de minería de textos al estudio de la violencia contra la mujer* [Doctoral dissertation, Universidad de Granada]. DIGIBUG.

ONU Mujeres. (2024, March). *Iniciativas transformadoras: Educación y empoderamiento para la igualdad en campos de ciencia, tecnología, ingeniería y matemáticas*. https://lac.unwomen.org/es/stories/noticia/2024/03/iniciativas-transformadoras-educacion-y-empoderamiento-para-la-igualdad-en-campos-de-ciencia-tecnologia-ingenieria-y-matematicas

Pacheco Montaño, L. C., & Díaz-Rincón, S. V. (2022). Participación política y democrática femenina en Colombia: Machismo, empoderamiento e igualdad de género. *Tejidos Sociales, 4*(1), 1–13.

Perucha Calvo, B. (2023). *Análisis de la brecha de género en las carreras STEM de la Universidad de Castilla-La Mancha, situación actual e influencia en el desarrollo socioeconómico de la región* [Report]. Instituto de la Mujer de Castilla-La Mancha.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society, 7*(2). https://doi.org/10.1177/20563051211019004

Purnomo, R. D., Fatyanosa, T. N., & Data, M. (2025). A comparative analysis of the impact of data preprocessing and fine-tuning methods on BERT's performance in text classification. In *2025 7th International Conference on Natural Language Processing (ICNLP)* (pp. 654–658). IEEE. https://doi.org/10.1109/ICNLP65360.2025.11108690

Reddit. (n.d.). *r/[subreddit name]* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/[subreddit]/

Reddit. (n.d.). *r/antiwork* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/antiwork/

Reddit. (n.d.). *r/cscareerquestions* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/cscareerquestions/

Reddit. (n.d.). *r/datascience* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/datascience/

Reddit. (n.d.). *r/ExperiencedDevs* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/ExperiencedDevs/

Reddit. (n.d.). *r/girlsgonewired* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/girlsgonewired/

Reddit. (n.d.). *r/ITCareerQuestions* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/ITCareerQuestions/

Reddit. (n.d.). *r/LadiesofScience* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/LadiesofScience/

Reddit. (n.d.). *r/learnprogramming* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/learnprogramming/

Reddit. (n.d.). *r/MachineLearning* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/MachineLearning/

Reddit. (n.d.). *r/programming* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/programming/

Reddit. (n.d.). *r/technology* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/technology/

Reddit. (n.d.). *r/TwoXChromosomes* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/TwoXChromosomes/

Reddit. (n.d.). *r/webdev* [Online forum]. Retrieved May 29, 2026, from https://www.reddit.com/r/webdev/

Reddit. (n.d.). *Reddit* [Social media platform]. Retrieved May 29, 2026, from https://www.reddit.com/

Rivas Sepúlveda, E., & Lamas Huerta, P. A. (2024). Más allá de los estereotipos: Participación de mujeres en carreras tecnológicas y STEM. *Revista Educación Superior y Sociedad, 36*(2), 247–271. https://doi.org/10.54674/ess.v36i2.903

Ruiz, L. L. (2021, November 12). *Frente al sesgo tecnológico, más mujeres programadoras*. Concilia2. https://www.concilia2.es/blog/sesgo-tecnologico-mujeres-programadoras/

Ruiz, L. L. (2023, March 20). *'Brogrammer' o cómo el machismo no quiere programadoras*. Concilia2. https://www.concilia2.es/brogrammer-machismo-tecnologia/

Shah, S., Manzoni, S. L., Zaman, F., Es Sabery, F., Epifania, F., & Zoppis, I. F. (2024). Fine-tuning of Distil-BERT for continual learning in text classification: An experimental analysis. *IEEE Access, 12*, 104964–104982. https://doi.org/10.1109/ACCESS.2024.3435537

Sierra Caballero, F., & Alberich Pascual, J. (Eds.). (2021). *Cultura digital, nuevas mediaciones sociales e identidades culturales*. Comunicación Social Ediciones y Publicaciones.

Sue, D. W. (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. Wiley.

That yields:

UN Women. (2025, November 19). *Facts and figures: Ending violence against women*. https://knowledge.unwomen.org/en/articles/facts-and-figures/facts-and-figures-ending-violence-against-women

UNESCO Bangkok. (2020). *STEM education for girls and women: Breaking barriers and exploring gender inequality in Asia*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000375106

Varela-Rodríguez, M., & Vicente-Mariño, M. (2021). Imágenes desgarradas: El uso de scrapers en investigación social en Instagram sobre cáncer. *Cuadernos.info, 49*, 72–97. https://doi.org/10.7764/cdi.49.27809

Vargas Poblete, R. B. (2023). *Rezagadas: Brecha de género en carreras STEM* [Undergraduate thesis, Universidad de Chile]. Repositorio Académico de la Universidad de Chile. https://doi.org/10.58011/1se5-r109

World Economic Forum. (2024). *Global gender gap report 2024*. https://www.weforum.org/publications/global-gender-gap-report-2024/

World Health Organization. (2021). *Violence against women prevalence estimates, 2018: Global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women*. https://www.who.int/publications/i/item/9789240022256