Technology Science

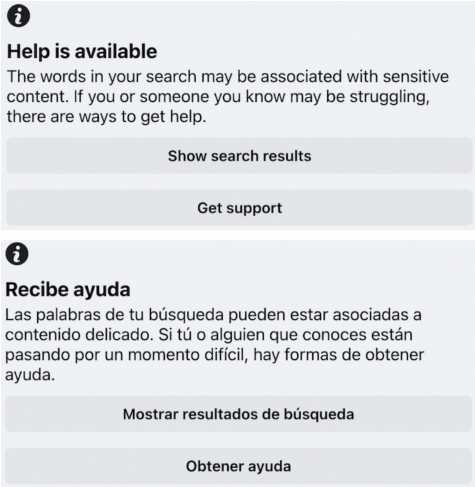
Facebook's Search Interventions: Bad in English, Peor en Español

Julia High

Abstract
Introduction
Background
Methods
Results
Discussion
References
Appendix

Download
Authors
Citation

Suggestions



Facebook's search feature displays interstitial message screens when a user searches using certain terms related to violence, nudity, sexual activity, or the sale of drugs. This study found that English-language searches were much more likely than searches in Spanish to trigger a pop-up intervention of the type shown here.

- This study explores the following question: How effective are Facebook's search interventions at responding to harmful queries in non-English languages?
- Facebook's search feature is instrumental in separating users from harmful content on the platform, and it is an essential part of the platform's content moderation ecosystem.
- The study found that 49.1% of English language words that should violate Meta's Community Standards triggered search interventions, whereas only 21.1% of similar Spanish words did.

- When searching in Spanish, it was far easier to access content that violates the Community Standards, including violent, sexually explicit, and harmful content. Furthermore, searches in Spanish rarely triggered the support or deterrence resources offered in response to comparable English searches.
- Facebook's search feature allows researchers a straightforward way of exploring the disparity between English and non-English languages without sowing harmful content onto the platform.

Abstract

Social media platforms such as Facebook have revolutionized the way we communicate online. However, while these platforms are meant to connect people across the globe, they are also leveraged by bad actors to post violent and harmful content. This study, conducted before Meta Platforms announced sweeping changes to its content moderation practices and standards in January 2025, used the Facebook search feature to examine one aspect of Facebook's content moderation practices. To do this, I recorded the occurrence of search interventions in response to Facebook search terms in English and Spanish. The interventions take the form of pop-up deterrence messages that appear immediately after searching provocative terms on Facebook and include comments such as "Are you sure you want to continue?" or "Help is available." For this study, I entered 57 words that violated Facebook's Community Standards in both English and Spanish—114 words in total—into the Facebook search dialog and recorded whether the system responded with an intervention.

Results summary:

In English, 49.1% of words triggered a pop-up deterrence message, which Meta calls an interstitial. In contrast, only 21.1% of Spanish words did. The fact that English words were more than twice as likely to trigger interstitials suggests that Facebook is significantly less effective at providing search interventions, deterrence measures, or helpful resources to potentially dangerous or inappropriate search queries in Spanish than in English. This has significant implications for online safety, demonstrating that speakers of other languages may endure more hostile digital environments owing to companies' insufficient protections against offensive and dangerous content and lack of interventions to prevent access to such content.

Introduction

Facebook's search feature acts as the first line of defense in its content moderation ecosystem. The search feature is instrumental in separating users from harmful content on the platform and identifying suspicious behavior. In a leaked internal document, Facebook employees have acknowledged that "'Search' plays the role of a connecting medium between actors and harmful content on Facebook" [1]. Another document includes a set of guidelines from Facebook employees that instruct raters (a term that is not explained in the document) to proactively assess potential search queries and determine whether they will lead to sensitive content including violence, hate speech, and sexual content [2].

Meta's written responses to questions for the record (QFRs) posed by the Senate Judiciary Committee during hearings on online child sexual exploitation in January 2024 yield further information on this search feature. In the QFRs, Meta outlines how many content moderation and digital safety decisions are made at the search level [3]. For example, Meta designates certain accounts as suspicious based on their frequent searches for inappropriate content. Or, when users search for terms related to suicide or self-harm, Meta wrote that it attempts to hide related results and instead direct the users to helpful resources using search interventions. Search interventions take the form of Facebook message screens, called interstitials, that pop up after certain searches containing sensitive terms. These interventions frequently contain "deterrence and prevention messaging" related to the sensitive topic [3]. During the hearings, Meta Chairman and Chief Executive Officer Mark Zuckerberg was asked about one of these interstitials, which warned users that their search may lead to images of child sexual abuse material. The warning screen offered users two options: "get resources" or "see results anyway" [3]. The option to "see results anyway" underscores how the search feature and these interstitials play a key role in

maintaining a safe—or unsafe—digital environment for users. While public information on Facebook's search feature is scarce, these statements from the company illustrate the importance of this feature in its broader content moderation efforts.

It is difficult to find meaningful and verifiable information on Meta's content moderation ecosystem, especially with regard to the number of users and human content moderators who speak a given language. Much of Facebook's content moderation work is outsourced. Even with some 15,000 outsourced content moderators working around the clock as of 2022, there were not enough workers to monitor every piece of hateful content on the platform [4]. The company has made claims about the efficacy of its automated systems at detecting hate speech, but their claims lack essential nuance. According to Meta, in 2024 the company's automated systems spotted 95.3% of hate speech that was taken down before it was recognized by a human [5]. Notably, Meta's account of the statistic only notes hate speech that was ultimately removed from the platform. The company's reports do not acknowledge harmful posts that may have been reported by users but never removed.

These oversights are amplified when it comes to content in languages other than English, given the lack of content moderation infrastructure tailored to non-English speakers. Meta offers no information on the language and region breakdown of users. It has stated that its hate speech detection tools can identify speech in over 100 languages [5]. However, internal documents indicate that the accuracy of this detection varies, and the detection tools can be wholly ineffective when identifying hate speech in non-included languages or dialects [6]. In order for Facebook's algorithms to learn how to identify certain speech, they must be trained on large datasets of content from a given language. Most of this training data is in English [7]. Other languages and dialects have far less training data, limiting the efficacy of the algorithms [6]. Unfortunately, many languages, often those spoken by vulnerable minorities, are left out of the picture entirely [7]. This absence has had deadly consequences in conflict regions like Myanmar and Ethiopia, where hate speech in Burmese and Amharic proliferated on Facebook, in part because of insufficient algorithmic capabilities [7, 8].

My study offers a simple way that researchers might gain firsthand insight into the efficacy of Facebook's search feature and directly observe content moderation outcomes. To that end, my research question was: How effective are Facebook's search interventions at responding to harmful queries in non-English languages?

My study tested the null hypothesis that there is no statistically significant difference between English and Spanish in the proportion of harmful queries that trigger a search intervention.

Background

Most content moderation research revolves around the efficacy of retrospective content moderation performed on content after it has been posted on a platform. I was unable to find research on flagging searches for harmful content at the source. Additionally, there are very few studies that focus specifically on Facebook's content moderation since there is little publicly available information on their strategies. In the QFRs for the January 2024 hearings, Meta stated that they do not disclose details about the mechanisms behind their search interventions or enforcement efforts so as not to provide a roadmap for bad actors who seek to evade detection [3]. Owing to the lack of comparable prior work, this study was partially informed by broader literature on linguistic disparities in natural language processing (NLP) models.

Classifiers are used in algorithmic moderation systems to categorize content within a dataset and recognize patterns based on its features [9]. Meta uses algorithmic classifier systems to identify hate speech and other Community Standards violations [10]. Meta's Community Standards are a set of guidelines that govern user conduct on Facebook, Instagram, Messenger, and Threads. The guidelines ensure that content on Facebook upholds values of authenticity, safety, privacy, and dignity. The Community Standards restrict or prohibit content that involves violence and criminal behavior, sexual activity, self-injury, hate speech, and a number of other sensitive topics.

Meta's classifiers draw on natural language processing (NLP) research, which explores how to give computers the ability to understand different languages and their nuances in the same way that humans do [10, 11]. Researchers at the Center for Democracy and Technology have reviewed evidence of the shortcomings of automated content analysis in non-English languages in general, not just on Facebook [12]. One commentary highlights the negative impact of the "resourcedness gap" between languages—the gap in quality training data between English and other languages [13]. Although Spanish, the second target language of this study, is one of the most widely spoken languages in the world, there is still exponentially less training data available in the language compared to English. A key issue is that many language models apply an Anglocentric framework to understanding less-resourced languages. This enables generalized language models to produce text in a

wider variety of languages, but it also means that they neglect cultural nuances that are key in identifying harmful terms [13].

Another study focuses specifically on the Indonesian language in NLP models [14]. In this study, researchers suggested that the majority of Indonesian NLP models were trained on formal or standard Indonesian, even though it is a language with many informalities in everyday speech. The researchers created a training dataset with phrases in formal Indonesian in parallel to phrases in informal Indonesian as a low-resource alternative to using machine translation models to transfer language style elements. According to the researchers, the model trained on this approach performed just as well as a "pre-trained GPT-2 fine-tuned to [the] task," but their model was far more cost-effective. This paper is fascinating because it highlights a cost-effective approach to analyzing the nuances in a low-resourced language, an approach that a well-resourced multinational company like Meta could potentially use to train classifiers to identify hate speech in different languages, addressing some of the shortcomings of multilingual LLMs, which draw on generalized speech patterns to interpret and generate text across multiple languages.

These studies have implications for content analysis more broadly. It is extremely difficult to find research that directly tests Facebook's content moderation, not surprising because such testing might require posting harmful content onto the platform. Studies conducted off the platform must serve as proxies for more in-depth research into Facebook's practices. Since Facebook is a platform with over two billion daily active users, and yet lacks transparency about some practices, it is crucial to conduct studies that can deepen public understanding of how their systems work [15]. This is especially important when it comes to content moderation, since opaque, automated decisions impact what content users have access to online and may have varying degrees of accuracy across languages.

It is likely that the search feature uses a mechanism different from other content moderation systems on Facebook, and yet its accessibility offers another possibility for proxy study. By comparing the search function's response to potentially harmful content in different languages, I hope to offer a simple and feasible way that researchers might gain firsthand insight into the efficacy of Facebook's search feature at responding to harmful content and shed some light on the potential disparity in content responses and moderation in non-English languages.

Methods

Materials

To conduct this study, I created a list of words in English and Spanish that would violate Facebook's Community Standards. I chose to conduct the study in English and Spanish because I am proficient in both languages, allowing me to reliably identify words that violate Facebook's Community Standards. When I created my word lists, I had a few criteria. First, the words had to deal with sensitive topics or harmful content. That is, a user typing these words into the Facebook search bar would likely access content that violates the Community Standards. Second, I had to ensure that each English word could be directly translated into Spanish and vice versa (for example, "sex" in English translates to "sexo" in Spanish). This allowed me to make comparisons and observations about the disparity in search interventions between the two languages. The word list is provided as Supplemental Material. Please note that the list includes highly offensive and inappropriate terms. Finally, I wanted to use terms that were both colloquially used and that generally had a universal meaning across all Spanish-speaking cultures. To do this, I consulted with Harvard Romance Languages faculty members and teaching staff. Ultimately I ended up with 114 words that fit all of the aforementioned requirements. I created two new Facebook accounts to conduct this study.

Study Design

1. I generated two lists of words that would violate Facebook's Community Standards in accordance with the procedure above.

2. I organized the words into categories based on the FB Community Standards. Categories include: Violence and Criminal Behavior; Safety; and Objectionable Content, each further divided into subcategories.

3. I created two Gmail accounts with the same gender-neutral name and the same birthdate.

4. I used those Gmail accounts to create two new Facebook accounts. The Facebook accounts also used the same gender-neutral name and birthdate. I did not report a gender for either account.

5. For one Facebook account, I kept the app settings in English. For the other, I changed the settings to Spanish.

6. Next, I searched each English word on the English account, logged out, cleared browser cookies, and searched each Spanish word on the Spanish account. I then recorded the interstitials I was given by Facebook. The prompts included warnings about violence, drug use, and sexual activity.

Between searches, I cleared my search history to prevent any biased results.

7. Throughout the study, I used my computer's default virtual private network (VPN), which shows I am located within the United States. I chose to do this to simulate the experience of Spanish speakers in the United States, since the country has a wide variety of Spanish speakers with origins in different countries and varying cultural contexts.

8. Finally, to analyze my results, I recorded the percentage of words that triggered an interstitial in both languages. I also made observations about which subcategories triggered the responses.

Results

Overall, most of the Community Standards-violating terms that I used for searching in both languages did not prompt an intervention. I was easily and directly able to search for inappropriate, incendiary, and hateful content. It was far easier to access content that violates the Community Standards when I searched in Spanish. In both languages, search terms that mentioned drugs did not trigger any offers of assistance, terms about self-harm provoked assistance screens less than half of the time, and the majority of sexually explicit language did not trigger any warning about inappropriate content. In addition, many searches led to Facebook Groups that are entire communities dedicated to these objectionable search topics. In Spanish, it was far easier and more straightforward to access these unsavory groups by searching with a Community Standards-violating term.

I chose words that would, if typed into the search bar, lead to content that violates Facebook's Community Standards. Since my lists consisted of paired words that had nearly the same meaning and connotation in English and Spanish, the words in each pair should have been flagged in both languages had the algorithm been operating in a consistent manner. This was not the case. Of the 57 English words entered, 28, or 49.1%, triggered an interstitial. In contrast, only 12 of the 57 Spanish words, or 21.1%, triggered an interstitial. The results of a z-test for statistical significance show a z score of 3.1319, corresponding to a p-value of 0.00174, allowing me to establish this difference as significant and reject my null hypothesis.

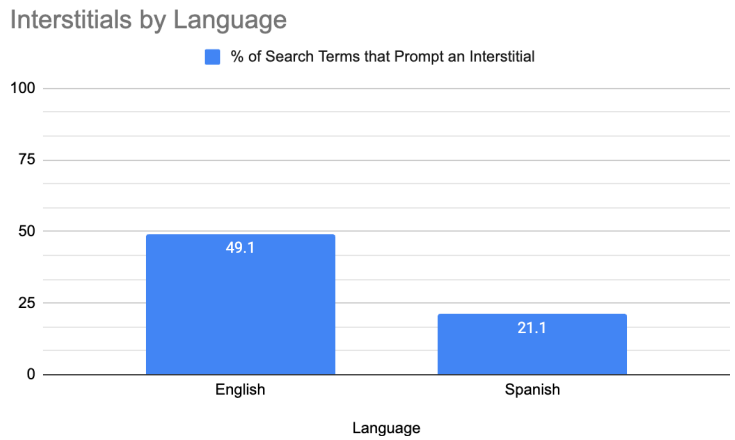
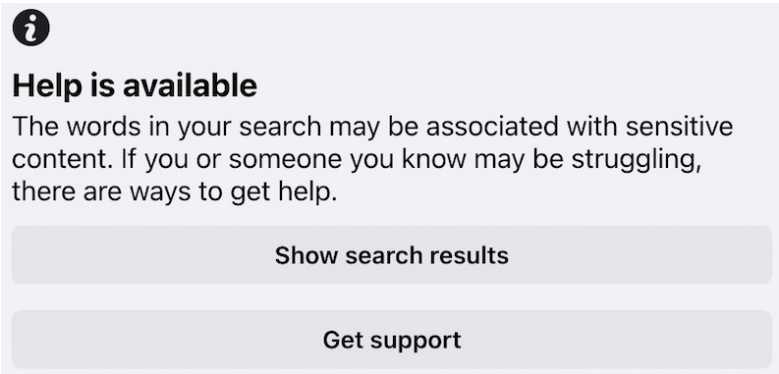


Figure 1. Percentage of Community Standards-violating terms entered in the Facebook search engine in English and Spanish that triggered interventions.

I received five types of search interventions in response to various searches. The first interstitial, depicted below, said that “Help is available” or “Recibe ayuda.” The text explains that the search triggered the response because “the words in your search may be associated with sensitive content.” It then offers the user the option to either continue to the search results page or to click a “Get support” link. If the user selects “Get support,” they are then presented with options to “Reach out to a friend,” “Contact a helpline,” or “See suggestions from professionals outside of Meta.” This prompt was triggered in response to violent words like “kill,” as well as words that referred to self-harm and eating disorders.



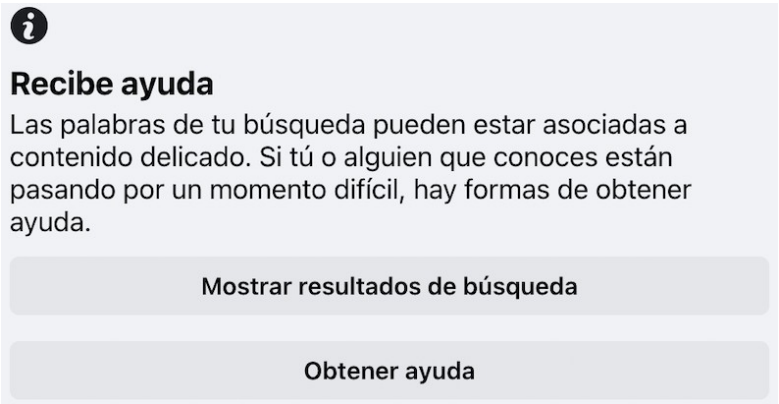


Figure 2. “Help is available” and “Recibe ayuda” Facebook search interstitials in English and Spanish.

A second interstitial asks the user “Are you sure you want to continue?” or “¿Confirmas que quieres continuar?” The pop-up explains that the search term “is sometimes associated with nudity or sexual activity, which isn’t allowed on Facebook.” In contrast to the above response, this wording suggests that Meta is attempting to warn the user to follow its guidelines. Users are then given the option to continue to the search results page or to return to the news feed. I received this prompt in response to searches about nudity or sexual activity.

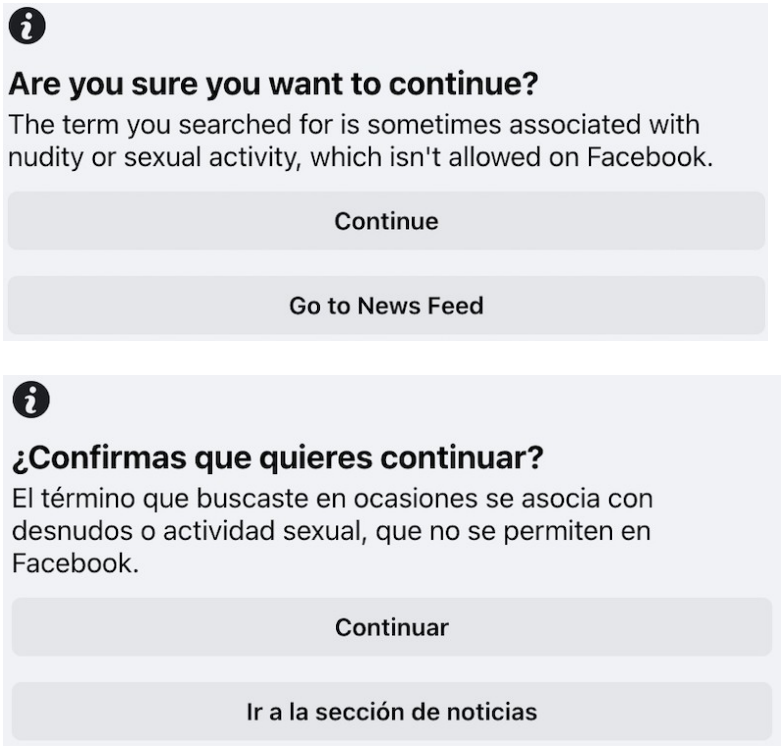


Figure 3. “Are you sure you want to continue?” and “¿Confirmas que quieres continuar?” Facebook search interstitials in English and Spanish for nudity or sexual activity.

A third interstitial also says “Are you sure you want to continue?” but focuses on drugs. The explanation given here is that the search term “is sometimes associated with the sale of drugs, which isn’t allowed on Facebook.” It then directs the user towards drug abuse resources. The user can either continue to the search results or select “Get Help.” If you select “Get Help,” you are directed to the Substance Abuse and Mental Health Services Administration (SAMHSA) website and their national helpline. This is a different pathway than the “Get support” option in the first prompt. This message did not pop up in response to any Spanish searches.

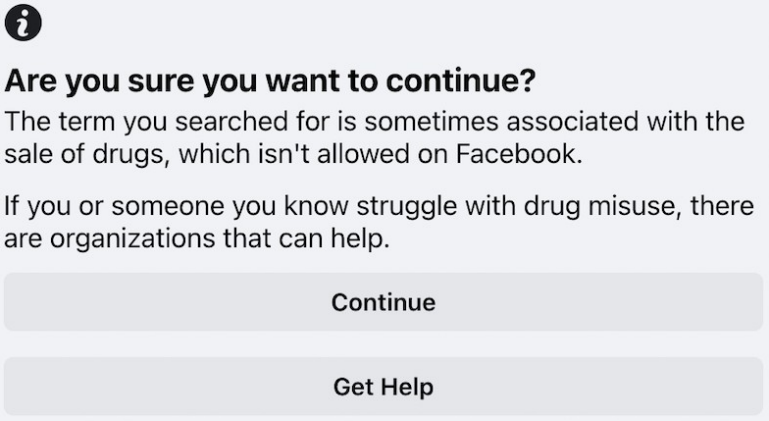


Figure 4. “Are you sure you want to continue?” Facebook search interstitial in English for drug-related content.

The fourth intervention is different from the others in that it offers the user more direct, action-oriented options. This interstitial also deals with drugs, telling the user “If you see the sale of drugs, please report it” or “Si detectas la venta de drogas, repórtalo.” However, in addition to offering a “Get Help” option, it first directs users to “See How to Report” the sale of drugs. This was the only interstitial I saw that urges the user toward an action.



Figure 5. “If you see the sale of drugs, please report it” and “Si detectas la venta de drogas, repórtalo” Facebook search interstitials in English and Spanish referring to the sale of drugs.

The fifth intervention also deals with drug-related search terms. After entering terms referring to certain drugs in English, I was allowed to continue my search. However, at the top of the search results, there appeared a window that asked, “Can we help?” followed by an explanation focused on “opioid misuse.” This screen offers the most proactive response, saying that Meta will

assist the user in finding “treatment referrals” and information on “substance misuse, prevention, and recovery.” Additionally, the “Get support” button directs users to the SAMHSA website. Interestingly, the most proactive response is also the page that fully allows the user to continue their search. This message was not presented in response to any Spanish searches.

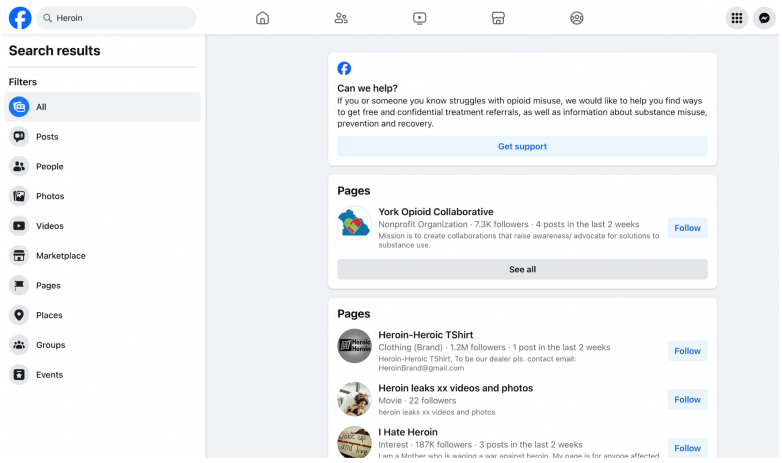


Figure 6. “Can we help?” and Facebook search interstitial in English for drug-related content.

Additional interesting trends organized by Facebook’s Community Standards subcategories are discussed below.

Violence and Incitement Subcategory

Of 10 words entered in the Violence and Incitement subcategory, only two —“kill” and “massacre”—were flagged in English, and only one— “masacre”— was flagged in Spanish. “Matar,” the Spanish equivalent for “kill,” was not flagged. Neither were several other words describing violent acts. Because of the small sample size, the difference between languages here is not statistically significant at a threshold of $p < 0.05$ with a z score of 0.6262, which corresponds to a p-value of 0.5287.

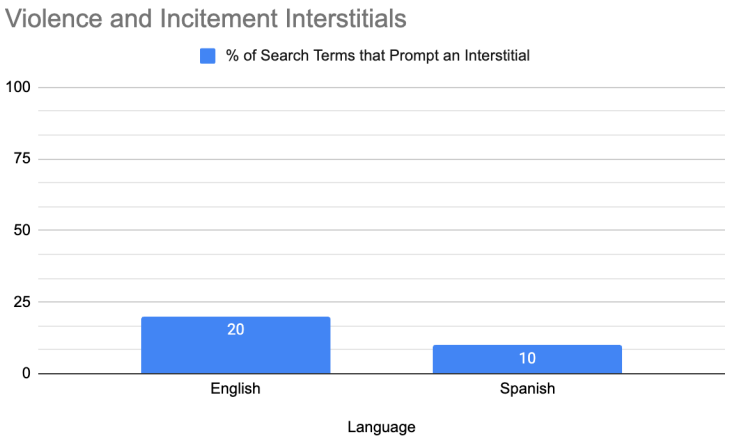


Figure 7. The percentage of interstitials prompted in English and Spanish for the Violence and Incitement Subcategory.

Restricted Goods and Services Subcategory

In the Restricted Goods and Services subcategory, 11 of the 17 English searches that mentioned drugs triggered one of the abovementioned interstitials, including slang terms like the use of “crack” to refer to cocaine. Four of the 17 Spanish words triggered an intervention. Just one, the word “crack,” (the English word “crack” is also used as slang for cocaine in Spanish) triggered an opioid misuse prompt. Searching for the word “opioid” or other types of opioids like “heroin” and “fentanyl,” triggered the sale of drugs prompt. There was no visible pattern between when the full-screen prompt depicted in Figure 4 occurred versus the small warning window depicted in Figure 5. The difference between English and Spanish here is statistically significant at a threshold of $p < 0.05$ with a z score of 2.4178, which corresponds to a p-value of 0.01552.

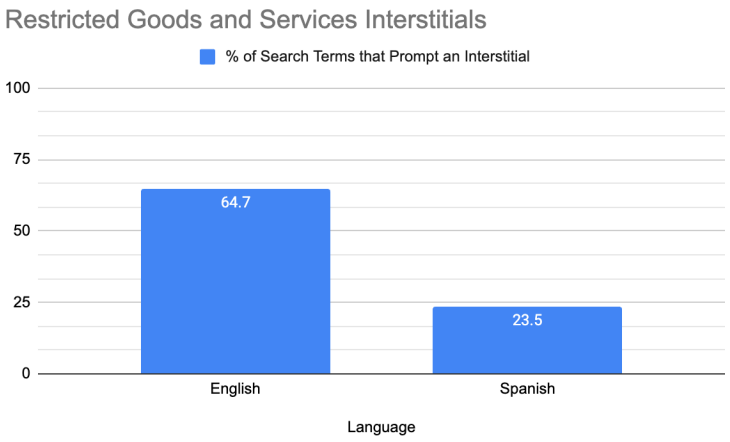


Figure 8. The percentage of interstitials prompted in English and Spanish for the Restricted Goods and Services Subcategory.

Suicide, Self-Injury, and Eating Disorders Subcategory

The Suicide, Self-Injury, and Eating Disorders subcategory had the highest search intervention rate across English and Spanish, with 100% of words in this category triggering an intervention in English and 42.8% in Spanish. The difference here is statistically significant at a threshold of $p < 0.05$ with a z score of 2.3664, which corresponds to a p-value of 0.01778.

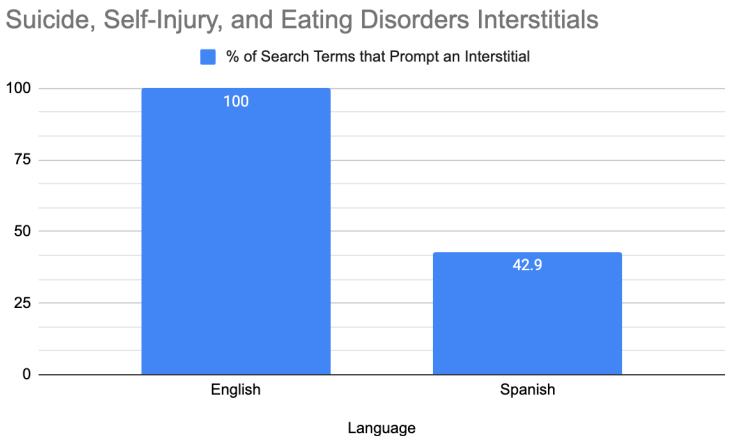


Figure 9. The percentage of searches in English and Spanish for the Suicide, Self-Injury, and Eating Disorders Subcategory that prompted interstitials.

Discussion

Although all of the words I searched should have violated Facebook's Community Standards, fewer than half of the words triggered aninterstitial even in English. In Spanish, the rate was around one-fifth.This suggests that Facebook's search interventions may be far less effective in non-English languages, particularly Spanish, refuting my null hypothesis. These results are particularly meaningful in light of the search interventions for drug-related content and self-harm.Although the system's English-language responses suggest that Facebook recognizes that searches related to these topics should direct a user to assistance, they simply do not appear to offer these resources in Spanish. This inequity has significant implications for digital safety and wellbeing on Facebook for Spanish speakers.

In this study, my goal was to compare the frequency of search interventions I received in response to Community Standards–violating searches and compare them between languages, not to look at the total number of terms flagged in each language separately. However, it is notable that the intervention rate was so low, even in English. This fact has numerous implications.

First, much of the contextual information on Facebook's search interventions for this study came from the U.S. Senate hearing on child safety on the platform. At the hearing, it was clear that child safety on social media is a politically salient issue, and Facebook took care to highlight its initiatives to keep children safe on the platform. Looking at Meta's political battles and the results of this study together suggests that Facebook may only implement search interventions on politically salient harmful search terms, not those that

violate their Community Standards. It raises questions about why Facebook does not offer interventions for searches that would lead to content including violent acts, clearly offensive slurs, and overtly sexual activity, given that they have the capability to do so.

Furthermore, Spanish is the second-most spoken language in the United States. In this study, I chose not to use a VPN to alter my location, instead logging into Facebook from my default, US-based location to simulate the digital experience of English and Spanish speakers who live in the US. There are millions of Spanish speakers who are as vulnerable as English speakers to child exploitation, drug addiction, and mental health struggles, and yet these results suggest they are far less likely to be offered search intervention resources.

Finally, this study may shed some light into Facebook's content moderation methods. Since fewer than half of words that directly violate Facebook's Community Standards in English triggered an intervention, this may suggest that Facebook does not filter searches for Community Standards–violating terms. If search entries were compared with a list of harmful terms, it seems reasonable to assume that more of my search terms would have been flagged, given that I used 57 very common words, some copied verbatim from the Community Standards. In the Senate testimony, Facebook claimed to have an algorithm that determines whether a search receives an intervention. This algorithm would appear to be ineffective, raising the question of whether Facebook chooses not to offer an intervention in response to certain search terms.

Meta is a business. Since social media companies are not held liable for inappropriate content posted on their platform under Section 230, they may choose to moderate their content only to the extent that doing so is profitable. The company presumably does not want Facebook to become a cesspool of hateful and inappropriate content because no one would want to use the site, and it would damage their reputation.However, controversial or inappropriate content may generate more engagement, which many critics note boosts Facebook's business [16]. Additionally, since many of the Community Standards-violating searches lead the user directly to entire communities dedicated to rule-breaking content, it seems as though it would be a good idea for Facebook to prevent people from searching certain unambiguously inappropriate terms at the source so that they cannot promote and engage with this content.It is possible that Facebook allows users to continue their searches even after receiving a search intervention to boost engagement and

help their business. Shedding light on these questions would require further study or disclosure.

Next Steps and Limitations

There are many ways in which this study might be extended beyond this initial experiment. First, researchers might expand the number of search terms in English and Spanish and test them multiple times. In this study, it was difficult to make judgments about certain subcategories because I had so few words in each. I developed my list of search terms directly from the Community Standards, and I believe they capture the essence of each category and subcategory. However, having a broader list of words would support stronger generalizations and allow examination of each Community Standards subcategory. Applying this methodology at a larger scale could be challenging, however, due to the fact that language is highly dependent on cultural context, and it is difficult to measure the harmfulness of a given term even with words that are near-direct translations. Part of the solution may be to introduce phrases into searches in addition to individual words.

Second, tests with other languages are needed to generalize the results beyond English and Spanish. In future versions of this study, researchers might examine languages that do not use the Latin alphabet and those with fewer speakers than English or Spanish. Facebook's search response algorithm may be even less effective in languages that use alphabets that are less similar to English. A difference in the search intervention rate between English and less common languages might shed light on whether a lack of training data or reliance on a multilingual model affect the system's ability to flag harmful terms. Given that English and Spanish are linguistically similar and that they are both widely spoken, researching these other factors may offer insight as to why there is such a large disparity in which words triggered a search intervention and provide insight on whether larger training datasets in other languages are needed to create more robust content moderation ecosystems. Translation issues may make extending the study difficult. However, a larger, multilingual research team might be able toaccomplish such a study. In the case of languages where one-to-one matches with English are less likely, it may be possible to create a list of words in a given language that fit into the Community Standards–violating categories and simply look at the overall proportion of words flagged rather than attempting a word-by-word comparison.

Finally, a study looking more closely at which harmfulEnglish-language search terms and phrases prompt search interventions could inform the supposition

raised in the discussion above about Facebook providing interventions only on politically relevant search terms. A study could hone in on words and phrases relating to child safety, given that is an issue where Facebook has come under fire from lawmakers and advocates. Instagram searches could be added to gain insight on Meta's platforms more broadly. Meta's internal documents revealed that as of 2020, 40% of hateful Arabic language content on Facebook is proactively detected by classifiers, while only 6% of such content is detected on Instagram [17]. This issue suggests that there may be further disparities in non-English language content moderation across Meta's platforms. Findings from such a study would be particularly significant in light of Meta's recent changes to their trust and safety policies that will loosen existing moderation and enforcement decisions [18].

References

1. FBarchive Image 117693_w32 in Document odoc4847272106w32 (Green Edition). Frances Haugen [producer]. Cambridge, MA: FBarchive [distributor], 2022. Web. 15 May 2023. URL: fbarchive.org/user/doc/odoc4847272106w32

2. FBarchive Document odoc7242350415w32 (Green Edition). Frances Haugen [producer]. Cambridge, MA: FBarchive [distributor], 2022. Web. 15 May 2023. URL: fbarchive.org/user/doc/odoc7242350415w32

3. U.S. Senate. 118th Congress, 2nd Session. (2024, Jan. 31). Big Tech and the Online Child Sexual Exploitation Crisis: Hearing Before the U.S. Senate Committee on the Judiciary. [Online]. Available: https://www.judiciary.senate.gov/committee-activity/hearings/big-tech-and-the-online-child-sexual-exploitation-crisis

4. "The people behind Meta's review teams." Meta Transparency Center. Last Updated Jan. 19, 2022. Accessed: Feb. 12, 2025. [Online]. Available: https://transparency.meta.com/en-gb/enforcement/detecting-violations/people-behind-our-review-teams/

5. "Hate Speech." Meta Transparency Center. Last Updated Sep. 2024. Accessed: Feb. 12, 2025. [Online]. Available: https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/

6. FBarchive Document odoc817183w32 (Green Edition). Frances Haugen [producer]. Cambridge, MA: FBarchive [distributor], 2022. Web. 15 May 2023. URL: fbarchive.org/user/doc/odoc817183w32

7. G. Nicholas and A. Bhatia. “Languages Left Behind: Automated Content Analysis in Non-English Languages.” Center for Democracy and Technology. Aug. 18, 2022. Accessed: Feb. 12, 2025. [Online]. Available: https://cdt.org/insights/languages-left-behind-automated-content-analysis-in-non-english-languages/

8. FBarchive Image 110749_w32 in Document odoc2985054037w32 (Green Edition). Frances Haugen [producer]. Cambridge, MA: FBarchive [distributor], 2022. Web. 15 May 2023. URL: fbarchive.org/user/doc/odoc2985054037w32

9. R. Gorwa, R. Binns, and C. Katzenbach, “Algorithmic content moderation: Technical and political challenges in the automation of platform governance,” Big Data & Society, vol. 7, no. 1, Feb. 2020. Accessed Feb. 12, 2025, doi: 10.1177/2053951719897945. [Online]. Available: https://journals.sagepub.com/doi/10.1177/2053951719897945

10. R. Dansby et al. “AI advances to better detect hate speech.” Meta ML Applications | Integrity. May 12, 2020. Accessed: Feb. 12, 2025. [Online]. Available: https://ai.meta.com/blog/ai-advances-to-better-detect-hate-speech/

11. C. Stryker and J. Holdsworth. “What is NLP (natural language processing)?” IBM Think. Accessed: Aug. 11, 2024. Feb. 12, 2025. [Online]. Available: https://www.ibm.com/think/topics/natural-language-processing

12. G. Nicholas and A. Bhatia, “Toward Better Automated Content Moderation in Low-Resource Languages”, JOTS, vol. 2, no. 1, Sep. 2023. Accessed: Feb. 12, 2025, doi: 10.54501/jots.v2i1.150. [Online]. Available: https://doi.org/10.54501/jots.v2i1.150

13. P. Joshi et al., “The State and Fate of Linguistic Diversity and Inclusion in the NLP World,” in Proceedings of the 58tyh Annual Meeting of the Association for Computational Linguistics., July 2020, pp. 6282–93, doi: [https://doi.org/10.18653/v1/2020.acl-main.560] (https://doi.org/10.18653/v1/2020.acl-main.560)

14. H. A. Wibowo, T. A. Prawiro, M. Ihsan, A. F. Aji, R. E. Prasojo, R. Mahendra, and S. Fitriany, “Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation,” Nov. 2020, arXiv:2011.03286.

15. “Meta Reports Third Quarter 2023 Results.” Meta Investor Relations. Oct. 25, 2023. Accessed: Feb. 12, 2025. [Online]. Available:

https://investor.atmeta.com/investor-news/press-release-details/2023/Meta-Reports-Third-Quarter-2023-Results/default.aspx

16. K. Hao. “The Facebook Whistleblower Says Its Algorithms Are Dangerous. Here’s Why.” MIT Technology Review, Oct. 5, 2021. Accessed: Feb. 12, 2025 . [Online]. Available: www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/

17. FBarchive Image 105069_w32 in Document odoc817183w32 (Green Edition). Frances Haugen [producer]. Cambridge, MA: FBarchive [distributor], 2022. Web. 15 May 2023. URL: fbarchive.org/user/doc/odoc817183w32

18. J. Kaplan. “More Speech and Fewer Mistakes.” Meta Newsroom. Jan. 7, 2025. Accessed: Feb. 12, 2025. [Online]. Available: https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Appendix

Supplemental Material

Authors

Julia High is a junior at Harvard College (class of 2026) studying Government with a focus in Technology Science and a language citation in Spanish. Julia has studied technology and international security as a Visiting Student at Oxford University and served as the Director of Ethical Technology at Harvard Tech for Social Good 501(c)(3). She aims to pursue a career at the intersection of law, policy, and innovation.

Author email

juliahigh@college.harvard.edu

Citation

High, J. Facebook’s Search Interventions: Bad in English, Peor en Español. Technology Science. 2025022503. February 25, 2025. https://techscience.org/a/2025022503/

Copyright © 2025. President and Fellows Harvard University.