TOM SIMONITE BUSINESS OCT 25, 2021 7:00 AM

Facebook Is Everywhere; Its Moderation Is Nowhere Close

Human reviewers and AI filters struggle to police the flood of content—or understand the nuances in different Arabic dialects.

Facebook users who speak languages such as Arabic, Pashto, or Armenian are effectively second class citizens of the world's largest social network. ILLUSTRATION: ELENA LACEY; GETTY IMAGES

The Facebook Papers

Thousands of internal documents show the world's biggest social media company often ignored warnings about its deepest problems.

What Badge Posts Reveal

Inside the Facebook Papers
How to Fix Facebook, According to Employees
Facebook Is Everywhere. Its Moderation Isn't
Now Reading

The AI Database →

APPLICATION: CONTENT MODERATION, SAFETY COMPANY: FACEBOOK TECHNOLOGY: NATURAL LANGUAGE PROCESSING

FACEBOOK LAUNCHED SUPPORT for Arabic in 2009 and scored a hit. Soon after, the service won plaudits for helping the mass protests known as the Arab Spring. By last year, Arabic was the third most common language on the platform, with people in the Middle East and North Africa spending more time each day with Facebook's services than users in any other region.

When it comes to understanding and policing Arabic content, Facebook has been less successful, according to two internal studies last year. One, a detailed account of Facebook's handling of Arabic, warns that the company's human and automated reviewers struggle to comprehend the varied dialects used across the Middle East and North Africa. The result: In a region wracked by political instability, the company wrongly censors benign posts for promoting terrorism while exposing Arabic speakers to hateful speech they shouldn't see.

"Arabic is not one language," the study says. "It is better to consider it a family of languages—many of which are mutually incomprehensible."

The documents on Facebook's foibles with Arabic are part of a tranche of internal material, known collectively as The Facebook Papers, that shows the company struggling—or neglecting—to manage its platform in places that are far from its headquarters in California, in regions where the vast majority of its users live. Many of these markets are in economically disadvantaged parts of the world, afflicted by the kinds of ethnic tensions and political violence that are often amplified by social media.

The documents were disclosed to the Securities and Exchange Commission and provided to Congress in redacted form by legal counsel for ex-Facebook employee Frances Haugen. The redacted versions were reviewed by a consortium of news organizations, including WIRED.

The collection offers a limited view inside the social network but reveals enough to illustrate the immense challenge created by Facebook's success. A site for rating the looks of women students at Harvard evolved into a global platform used by nearly 3 billion people in more than 100 languages. Perfectly curating such a service is impossible, but the company's protections for its users seem particularly uneven in poorer countries. Facebook users who speak languages such as Arabic, Pashto, or Armenian are effectively second class citizens of the world's largest social network.

Some of Facebook's failings detailed in the documents involve genuinely hard technical problems. The company uses artificial intelligence to help manage problematic content—at Facebook's scale humans cannot review every post. But computer scientists say machine learning algorithms don't yet understand the nuances of language. Other shortcomings appear to reflect choices by Facebook, which made more than $29 billion in profit last year, about where and how much to invest.

For example, Facebook says nearly two-thirds of the people who use the service do so in a language other than English and that it regulates content in the same way globally. A company spokesperson said it has 15,000 people reviewing

content in more than 70 languages and has published its Community Standards in 50. But Facebook offers its service in more than 110 languages; users post in still more.

> Nearly two-thirds of the people who use Facebook do so in a language other than English.

A December 2020 memo on combating hate speech in Afghanistan warns that users can't easily report problematic content because Facebook had not translated its community standards into Pashto or Dari, the country's two official languages. Online forms for reporting hate speech had been only partially translated into the two languages, with many words presented in English. In Pashto, also widely spoken in Pakistan, the memo says Facebook's translation of the term hate speech "does not seem to be accurate."

"When combating hate speech on Facebook, our goal is to reduce its prevalence, which is the amount of it that people actually see," a Facebook spokesperson said in a statement. The company recently released figures suggesting that on average, this has declined worldwide since mid-2020. "This is the most comprehensive effort to remove hate speech of any major consumer technology company, and while we have more work to do we remain committed to getting this right."

For Arabic, most of Facebook's content review takes place in Casablanca, Morocco, one document says, using locally recruited staff. That means errors when handling content from outside North Africa are "virtually guaranteed," the document says.

Even in North African dialects, errors are a problem. The document cites the case of Hosam El Sokkari, previously the BBC's head of Arabic, who in 2020 found himself unable to livestream on Facebook because the company said a 2017 post written in Egyptian Arabic that criticized a conservative Muslim cleric promoted terrorism. Algorithms flagged the post for breaking Facebook's rules and human reviewers concurred, according to the *Wall Street Journal*. El Sokkari's account was later locked after Facebook told him several other of his posts breached its policies. The document says an internal investigation found that staff who reviewed "a set" of El Sokkari's posts wrongly took action against them 90 percent of the time.

A Facebook spokesperson said the company reinstated El Sokkari's posts after it became aware they had been mistakenly removed; Facebook is reviewing options to address the challenges of handling Arabic dialects, including hiring more content reviewers with diverse language skills.

> Users in Afghanistan can't easily report problematic content because Facebook had not translated its community standards into Pashto or Dari, the country's two official languages.

A document reviewing Facebook's moderation across the Middle East and North Africa, from December 2020, says algorithms used to detect terrorist content in Arabic wrongly flag posts 77 percent of the time—worse than a coin flip. A Facebook spokesperson said the figure is wrong, and that the company has not seen evidence of such poor performance.

That document also warns that flagging too many posts for terrorism may be harming Facebook's business prospects. The company's most recent earnings report said revenue per user grew fastest in its geographic category that includes the Middle East. The document says that when owners of advertiser accounts that had been disabled appealed

Facebook's decision, nearly half proved to have been shuttered incorrectly. It suggests that video views and growth in the region are constrained because accounts are being wrongly penalized.

Rasha Abdulla, a professor at the American University in Cairo who studies social media, says the findings of Facebook's research confirm suspicions by outsiders that the company quashes innocent or important content, such as jokes, news coverage, and political discussion. She believes the problem has worsened as the company has added more automation. "We really started seeing these problems arise in recent years, with increasing use of algorithms and AI," she says.

Increased reliance on algorithms is at the heart of Facebook's strategy for content moderation. The company recently said machine learning has reduced how often Facebook users encounter hate speech. But Facebook does not disclose data on how its technology performs in different countries or languages.

Internal Facebook documents show some staff expressing skepticism and include evidence that the company's moderation technology is less effective in emerging markets.

One reason for that is a shortage of human-labeled content needed to train machine learning algorithms to flag similar content by themselves. The 2020 document that discussed Arabic dialects says Facebook needs a pool of workers who understand the full diversity of Arabic to properly track problem content and train algorithms for the different dialects. It says a lead engineer on hate speech work considered building such systems impossible. "As it stands, they barely have enough content to train and maintain the Arabic classifier," the document says.

Earlier this month, Facebook agreed to commission an independent check on its content moderation for Arabic and Hebrew. The suggestion had come from Facebook's Oversight Board of outside experts funded by the company, after reviewers incorrectly removed an Egyptian user's post of a report by *Al Jazeera Arabic* on threats of violence by the military wing of Hamas. Facebook had already reinstated the post.

> "We really started seeing these problems arise in recent years, with increasing use of algorithms and AI."
>
> – RASHA ABDULLA, PROFESSOR, AMERICAN UNIVERSITY IN CAIRO

No one has ever had to manage a global network like Facebook's that reaches into nearly every country, language, and community on earth. Internal documents show staff functioning like an internet age diplomatic corps, attempting to apply data science to the world's thorniest conflicts. Documents show the company attempting to prioritize extra language and automated content moderation resources for a list of "at-risk countries" where violence or other harms are considered most likely. A version of the list for 2021 shows 10 countries on the top tier, including Pakistan, Ethiopia, and

Myanmar—where the UN said Facebook posts played a "determining role" in 2017 attacks on the country's Muslim Rohingya minority. A December 2020 document describes a push to hire staff with expertise in those countries and their languages. It says the company lacks such coverage for four of the 10 countries on the top tier.

Facebook says it has automated systems to find hate speech and terrorism content in more than 50 languages.

In internal posts, some Facebook engineers express blunt pessimism about the power of automation to solve the company's problems. A 2019 document estimates that properly training a classifier to detect hate speech in a market served by Facebook requires 4,000 manual content reviews a day. When one employee asks if that number might shrink as systems get better, a coworker says the answer is no because the company's algorithms are immature, like elementary school students: "They need teachers (human reviewers) to grow."

A Facebook data scientist who worked on "violence and incitement" before leaving the company last December estimated in a goodbye post included in Haugen's documents and previously reported by BuzzFeed News that the company removes less than 5 percent of hate speech on the platform—and claimed AI can't significantly improve that. "The problem of inferring the semantic meaning of speech with high precision is not remotely close to solved," the data scientist wrote.

Facebook says figures from June showed that on average across the world, the amount of hate speech users saw on Facebook fell by half in the previous nine months. The company doesn't disclose information on patterns for individual countries or languages.

The departing data scientist argued the company could do more, saying employees working on content problems were given impossible remits. Authors of the post described a deep sense of guilt over having to prioritize work on US English while violence flared in Armenia and Ethiopia and claimed Facebook has an easy way to improve its global moderation. "It's just not reasonable to have one person responsible for data science for all of violence and incitement for the entire world," the post said. "We can afford it. Hire more people."

Updated, 10-25-21, 3:35pm ET: This article has been updated to include additional information from Facebook about the number of languages in which it has automated systems to identify hate speech and the number of languages in which its community standards are available.

More Great WIRED Stories

Tom Simonite is a senior writer for WIRED in San Francisco covering artificial intelligence and its effects on the world. He once trained an artificial neural network to generate seascapes and is available for commissions. Simonite was previously San Francisco bureau chief at *MIT Technology Review*, and wrote and edited... Read more

SENIOR WRITER

TOPICS FACEBOOK THE FACEBOOK PAPERS - SERIES THE FACEBOOK PAPERS CONTENT MODERATION ALGORITHMS ARTIFICIAL INTELLIGENCE SOCIAL MEDIA