Facebook dithered in curbing divisive user content in India

October 24 2021, by Sheikh Saaliq and Krutika Pathi

In this Sept. 27, 2015, file photo, Facebook CEO Mark Zuckerberg, right, hugs Prime Minister of India Narendra Modi at Facebook in Menlo Park, Calif.. Facebook in India has been selective in curbing hate speech, misinformation and inflammatory posts, particularly anti-Muslim content, according to leaked documents obtained by The Associated Press, even as the internet giant's own employees cast doubt over the motivations and interests. Credit: AP Photo/Jeff Chiu, File

Facebook in India has been selective in curbing hate speech, misinformation and inflammatory posts, particularly anti-Muslim content, according to leaked documents obtained by The Associated Press, even as its own employees cast doubt over the company's motivations and interests.

From research as recent as March of this year to company memos that date back to 2019, the internal company documents on India highlight Facebook's constant struggles in quashing abusive content on its platforms in the world's biggest democracy and the company's largest growth market. Communal and religious tensions in India have a history of boiling over on social media and stoking violence.

The files show that Facebook has been aware of the problems for years, raising questions over whether it has done enough to address these issues. Many critics and digital experts say it has failed to do so, especially in cases where members of Prime Minister Narendra Modi's ruling Bharatiya Janata Party, the BJP, are involved.

Across the world, Facebook has become increasingly important in politics, and India is no different.

Modi has been credited for leveraging the platform to his party's advantage during elections, and reporting from The Wall Street Journal last year cast doubt over whether Facebook was selectively enforcing its policies on hate speech to avoid blowback from the BJP. Both Modi and Facebook chairman and CEO Mark Zuckerberg have exuded bonhomie, memorialized by a 2015 image of the two hugging at the Facebook headquarters.

The leaked documents include a trove of internal company reports on hate speech and misinformation in India. In some cases, much of it was intensified by its own "recommended" feature and algorithms. But they

also include the company staffers' concerns over the mishandling of these issues and their discontent expressed about the viral "malcontent" on the platform.

According to the documents, Facebook saw India as one of the most "at risk countries" in the world and identified both Hindi and Bengali languages as priorities for "automation on violating hostile speech." Yet, Facebook didn't have enough local language moderators or content-flagging in place to stop misinformation that at times led to real-world violence.

In this Sept. 27, 2015, file photo, India's Prime Minister Narendra Modi, left, speaks next to Facebook CEO Mark Zuckerberg at Facebook in Menlo Park, Calif.. Facebook in India has been selective in curbing hate speech,

misinformation and inflammatory posts, particularly anti-Muslim content, according to leaked documents obtained by The Associated Press, even as the internet giant's own employees cast doubt over the motivations and interests. Credit: AP Photo/Jeff Chiu, File

In a statement to the AP, Facebook said it has "invested significantly in technology to find hate speech in various languages, including Hindi and Bengali" which has resulted in "reduced the amount of hate speech that people see by half" in 2021.

"Hate speech against marginalized groups, including Muslims, is on the rise globally. So we are improving enforcement and are committed to updating our policies as hate speech evolves online," a company spokesperson said.

This AP story, along with others being published, is based on disclosures made to the Securities and Exchange Commission and provided to Congress in redacted form by former Facebook employee-turned-whistleblower Frances Haugen's legal counsel. The redacted versions were obtained by a consortium of news organizations, including the AP.

Back in February 2019 and ahead of a general election when concerns of misinformation were running high, a Facebook employee wanted to understand what a new user in the country saw on their news feed if all they did was follow pages and groups solely recommended by the platform itself.

The employee created a test user account and kept it live for three weeks, a period during which an extraordinary event shook India—a militant attack in disputed Kashmir had killed over 40 Indian soldiers, bringing the country to near war with rival Pakistan.

In the note, titled "An Indian Test User's Descent into a Sea of Polarizing, Nationalistic Messages," the employee whose name is redacted said they were "shocked" by the content flooding the news feed which "has become a near constant barrage of polarizing nationalist content, misinformation, and violence and gore."

Seemingly benign and innocuous groups recommended by Facebook quickly morphed into something else altogether, where hate speech, unverified rumors and viral content ran rampant.

In this Tuesday, March 31, 2020, file photo, Indian paramedics after screening take down the names of Muslims wearing face masks before they are being to a bus that will take them to a quarantine facility, amid concerns over the spread of the new coronavirus at the Nizamuddin area of New Delhi, India. Facebook in India has been selective in curbing hate speech, misinformation and inflammatory posts, particularly anti-Muslim content, according to leaked documents obtained by The Associated Press, even as the internet giant's own

employees cast doubt over the motivations and interests. Credit: AP Photo/Manish Swarup, File

The recommended groups were inundated with fake news, anti-Pakistan rhetoric and Islamophobic content. Much of the content was extremely graphic.

One included a man holding the bloodied head of another man covered in a Pakistani flag, with an Indian flag in the place of his head. Its "Popular Across Facebook" feature showed a slew of unverified content related to the retaliatory Indian strikes into Pakistan after the bombings, including an image of a napalm bomb from a video game clip debunked by one of Facebook's fact-check partners.

"Following this test user's News Feed, I've seen more images of dead people in the past three weeks than I've seen in my entire life total," the researcher wrote.

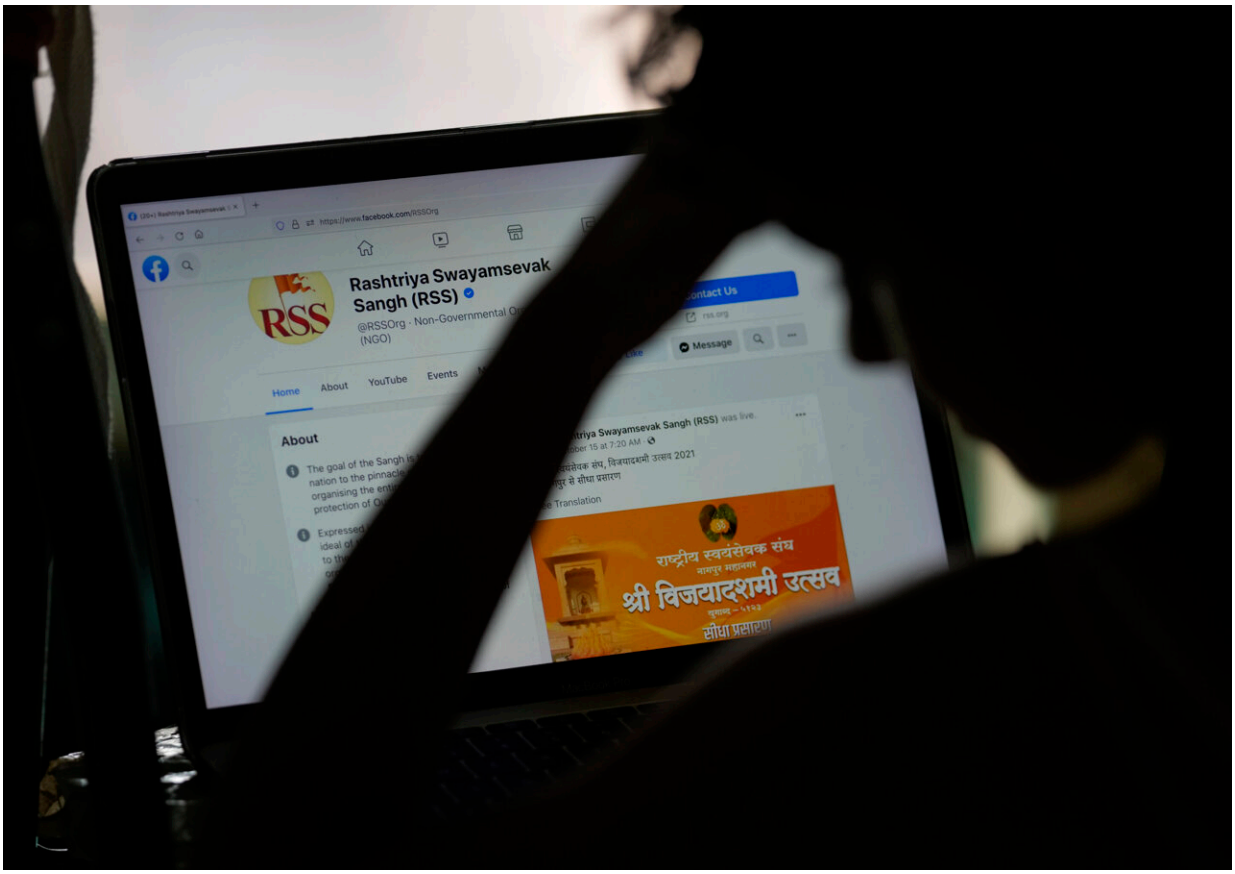
It sparked deep concerns over what such divisive content could lead to in the real world, where local news at the time were reporting on Kashmiris being attacked in the fallout.

"Should we as a company have an extra responsibility for preventing integrity harms that result from recommended content?" the researcher asked in their conclusion.

The memo, circulated with other employees, did not answer that question. But it did expose how the platform's own algorithms or default settings played a part in spurring such malcontent. The employee noted that there were clear "blind spots," particularly in "local language content." They said they hoped these findings would start conversations

on how to avoid such "integrity harms," especially for those who "differ significantly" from the typical U.S. user.

Even though the research was conducted during three weeks that weren't an average representation, they acknowledged that it did show how such "unmoderated" and problematic content "could totally take over" during "a major crisis event."



A girl looks at the face book page of Rashtriya Swayamevak Sangh or RSS, in New Delhi, India, Sunday, Oct. 24, 2021. Facebook in India has been selective in curbing hate speech, misinformation and inflammatory posts, particularly anti-Muslim content, according to leaked documents obtained by The Associated Press, even as the internet giant's own employees cast doubt over the motivations and interests. Credit: AP Photo/Manish Swarup

The Facebook spokesperson said the test study "inspired deeper, more rigorous analysis" of its recommendation systems and "contributed to product changes to improve them."

"Separately, our work on curbing hate speech continues and we have further strengthened our hate classifiers, to include four Indian languages," the spokesperson said.

Other research files on misinformation in India highlight just how massive a problem it is for the platform.

In January 2019, a month before the test user experiment, another assessment raised similar alarms about misleading content. In a presentation circulated to employees, the findings concluded that Facebook's misinformation tags weren't clear enough for users, underscoring that it needed to do more to stem hate speech and fake news. Users told researchers that "clearly labeling information would make their lives easier."

Again, it was noted that the platform didn't have enough local language fact-checkers, which meant a lot of content went unverified.

Alongside misinformation, the leaked documents reveal another problem plaguing Facebook in India: anti-Muslim propaganda, especially by Hindu-hardline groups.

India is Facebook's largest market with over 340 million users—nearly 400 million Indians also use the company's messaging service WhatsApp. But both have been accused of being vehicles to spread hate speech and fake news against minorities.

In this Thursday, Feb. 27, 2020, file photo, relatives and neighbors wail near the body of Mohammad Mudasir, 31, who was killed in communal violence in New Delhi, India. Facebook in India has been selective in curbing hate speech, misinformation and inflammatory posts, particularly anti-Muslim content, according to leaked documents obtained by The Associated Press, even as the internet giant's own employees cast doubt over the motivations and interests. Credit: AP Photo/Manish Swarup, File

In February 2020, these tensions came to life on Facebook when a politician from Modi's party uploaded a video on the platform in which he called on his supporters to remove mostly Muslim protesters from a road in New Delhi if the police didn't. Violent riots erupted within hours, killing 53 people. Most of them were Muslims. Only after

thousands of views and shares did Facebook remove the video.

In April, misinformation targeting Muslims again went viral on its platform as the hashtag "Coronajihad" flooded news feeds, blaming the community for a surge in COVID-19 cases. The hashtag was popular on Facebook for days but was later removed by the company.

For Mohammad Abbas, a 54-year-old Muslim preacher in New Delhi, those messages were alarming.

Some video clips and posts purportedly showed Muslims spitting on authorities and hospital staff. They were quickly proven to be fake, but by then India's communal fault lines, still stressed by deadly riots a month earlier, were again split wide open.

The misinformation triggered a wave of violence, business boycotts and hate speech toward Muslims. Thousands from the community, including Abbas, were confined to institutional quarantine for weeks across the country. Some were even sent to jails, only to be later exonerated by courts.

"People shared fake videos on Facebook claiming Muslims spread the virus. What started as lies on Facebook became truth for millions of people," Abbas said.

Criticisms of Facebook's handling of such content were amplified in August of last year when The Wall Street Journal published a series of stories detailing how the company had internally debated whether to classify a Hindu hard-line lawmaker close to Modi's party as a "dangerous individual"—a classification that would ban him from the platform—after a series of anti-Muslim posts from his account.

In this Thursday, Feb. 27, 2020, file photo, an Indian woman walks past as Indian paramilitary soldiers patrol a street vandalized in Tuesday's violence in New Delhi, India. Facebook in India has been selective in curbing hate speech, misinformation and inflammatory posts, particularly anti-Muslim content, according to leaked documents obtained by The Associated Press, even as the internet giant's own employees cast doubt over the motivations and interests.Credit: AP Photo/Altaf Qadri,File

The documents reveal the leadership dithered on the decision, prompting concerns by some employees, of whom one wrote that Facebook was only designating non-Hindu extremist organizations as "dangerous."

The documents also show how the company's South Asia policy head herself had shared what many felt were Islamophobic posts on her

personal Facebook profile. At the time, she had also argued that classifying the politician as dangerous would hurt Facebook's prospects in India.

The author of a December 2020 internal document on the influence of powerful political actors on Facebook policy decisions notes that "Facebook routinely makes exceptions for powerful actors when enforcing content policy." The document also cites a former Facebook chief security officer saying that outside of the U.S., "local policy heads are generally pulled from the ruling political party and are rarely drawn from disadvantaged ethnic groups, religious creeds or casts" which "naturally bends decision-making towards the powerful."

Months later the India official quit Facebook. The company also removed the politician from the platform, but documents show many company employees felt the platform had mishandled the situation, accusing it of selective bias to avoid being in the crosshairs of the Indian government.

"Several Muslim colleagues have been deeply disturbed/hurt by some of the language used in posts from the Indian policy leadership on their personal FB profile," an employee wrote.

Another wrote that "barbarism" was being allowed to "flourish on our network."

It's a problem that has continued for Facebook, according to the leaked files.

This May 16, 2012, file photo, shows the Facebook logo displayed on an iPad. Facebook in India dithered in curbing hate speech and anti-Muslim content on its platform and lacked enough local language moderators to stop misinformation that at times led to real-world violence, according to leaked documents obtained by The Associated Press. Credit: AP Photo/Matt Rourke, File

As recently as March this year, the company was internally debating whether it could control the "fear mongering, anti-Muslim narratives" pushed by Rashtriya Swayamsevak Sangh, a far-right Hindu nationalist group which Modi is also a part of, on its platform.

In one document titled "Lotus Mahal," the company noted that members with links to the BJP had created multiple Facebook accounts to amplify anti-Muslim content, ranging from "calls to oust Muslim populations

from India" and "Love Jihad," an unproven conspiracy theory by Hindu hard-liners who accuse Muslim men of using interfaith marriages to coerce Hindu women to change their religion.

The research found that much of this content was "never flagged or actioned" since Facebook lacked "classifiers" and "moderators" in Hindi and Bengali languages. Facebook said it added hate speech classifiers in Hindi starting in 2018 and introduced Bengali in 2020.

The employees also wrote that Facebook hadn't yet "put forth a nomination for designation of this group given political sensitivities."

The company said its designations process includes a review of each case by relevant teams across the company and are agnostic to region, ideology or religion and focus instead on indicators of violence and hate. It did not, however, reveal whether the Hindu nationalist group had since been designated as "dangerous."