unesco

THE "HATE SPEECH" POLICIES OF MAJOR PLATFORMS DURING THE COVID-19 PANDEMIC

Ana Laura Pérez

Published in 2021 by the United Nations Educational, Scientific and Cultural Organization, 7, place de Fontenoy, 75352 Paris 07 SP, France and UNESCO's Regional Bureau for Science in Latin America and the Caribbean, UNESCO Montevideo, Luis Piera 1992, 2nd floor, Montevideo, 11200, Uruguay.

MTD/CI/2021/PI/0/REV1

Managing Editor: Sandra Sharman
Graphic design: Trigeon.

This publication has received support from OBSERVACOM.

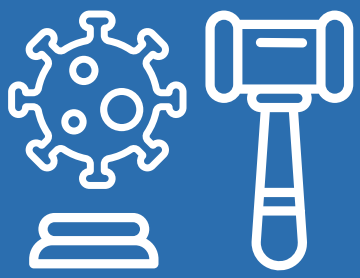
TABLE OF CONTENTS

EXECUTIVE SUMMARY

This document reports an increase in so-called “hate speech” posts on Facebook, Twitter, and YouTube after the onset of the COVID-19 pandemic. Although dissimilar, such an increase can be observed in the transparency reports of the different platforms and the surge in content moderation since March 2020.

During the same period—as a result of the lockdown measures adopted in most countries around the world—platforms increased the use of AI tools for content moderation. Therefore, we can’t fully say whether the interannual growth is linked to increased posts or changes in monitoring systems.

INTRODUCTION



The arrival of the COVID-19 pandemic has had significant impacts that go beyond public health services and the populations around the world. We are still unable to fully grasp its impact, and it will probably be a few years before we can do so.

Some studies show an increase in hate speech against certain groups across social media platforms due to the COVID-19 pandemic. There is also evidence of increased hate speech removal from social media.

Since 2020, platforms and social media have made substantial changes to their moderation criteria. New provisions have been added to their community guidelines, and—since many of their workers were sent home—they have been forced to boost the use of automated moderation. In addition, the impact of hate speech on social media platforms and its potential for inciting violence has become a subject of public debate.

This year, Twitter, Facebook, and YouTube—each to a different extent—shifted from remaining impartial over the public debate promoted on their platforms to, for example, blocking the account of a sitting President during his final days in office.

By looking at the transparency reports of social media platforms, it is clear that in 2020, so-called hate speech and the removal of such posts grew significantly on social media. There is not enough disaggregated data to understand what each platform classifies as hate speech, the decision-making processes, and error rates, making it more difficult to understand the root causes of this growth.

Platforms have publicly acknowledged having problems with their content moderation processes after sending workers in these areas home. This meant they were forced to increase the use of automated moderation and AI systems. They also acknowledged that this might have led to an increase in error rates due to machine learning software's difficulties in understanding the context in which these contents are created, as machine and human content moderators have different capacities in this regard.

In particular, Facebook and its sister platform, Instagram, had exponential growth in content classified as "hate speech" during the onset of the COVID-19 pandemic. After the second quarter of 2020, this growth spiked when most countries started implementing social distancing and lockdown measures.

However, there's not enough data to establish whether changes in the content review criteria and the shift towards a more aggressive content moderation model could explain this growth or if there was an actual increase of hate speech on social media in 2020.

This study explores the increase in online hate speech after the onset of the COVID-19 pandemic worldwide and the actions implemented by Facebook, Twitter, and YouTube, their reach, impact, causes, and potential consequences.

ONLINE HATE SPEECH

The term "hate speech" is complex, and it doesn't have a consistent definition. The different platforms, governments, laws, and regulations don't seem to agree on what it means. In the **United Nations Strategy and Plan of Action on Hate Speech launched by Secretary-General Antonio Guterres** hate speech is defined as "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language concerning a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor." It also states that "this is often rooted in, and generates intolerance and hatred and, in certain contexts, can be demeaning and divisive."

From a legal point of view, international law does not prohibit hate speech as such but rather "the incitement to discrimination, hostility, and violence." The United Nations defines the former as a "very dangerous form of speech because it explicitly and deliberately aims at triggering discrimination, hostility, and violence, which may also lead to or include terrorism or atrocity crimes." The document indicates that hate speech that does not reach the threshold of incitement is not something that international law requires States to prohibit. Nevertheless, the United Nations warns that "even when not prohibited, hate speech may be harmful."

In this document, the United Nations states that "Around the world, we are seeing a disturbing groundswell of xenophobia, racism, and intolerance, including rising anti-Semitism, anti-Muslim hatred and persecution of Christians. Social media and other forms of communication are being exploited as platforms for bigotry. Neo-Nazi and white supremacy movements are on the march. Public discourse is being weaponized for political gain with incendiary rhetoric that stigmatizes and dehumanizes minorities, migrants, refugees, women, and any so-called 'other.'" This is not an isolated phenomenon or the loud voices of a few people on the fringe of society. Hate is moving into the mainstream – in liberal democracies and authoritarian systems alike. And with each broken norm, the pillars of our common humanity are weakened. Hate speech is a menace to democratic values, social stability, and peace."

Likewise, **the General Policy Recommendation No. 15 on Combating Hate Speech and the Explanatory Memorandum of the European Commission against Racism and Intolerance (ECRI) of the Council of Europe**, defines hate speech as the "use of one or more particular forms of expression—namely, the advocacy, promotion or incitement of the denigration, hatred or vilification of a person or group of persons, as well any harassment, insult, negative stereotyping, stigmatization or threat of such person or persons and any justification of all these forms of expression—that is based on a non-exhaustive list of personal characteristics or status that includes "race," color, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity, and sexual orientation."

According to this definition, hate speech is "is not just intended to incite the commission of acts of violence, intimidation, hostility or discrimination but also such use that can reasonably be expected to have that effect" and "grounds that go beyond race, color, language, religion or belief, nationality, national or ethnic origin, and descent." The document also indicates that the term "'Expression' is understood in the Recommendation to cover speech and publications in any form, including through the use of electronic media, as well as their dissemination and storage." And that "Hate speech can take the form of written or spoken words, or other forms such as pictures, signs, symbols, paintings, music, plays or videos. It also embraces the use of particular conduct, such as gestures, to communicate an idea, message or opinion." In addition, this definition also includes "the public denial, trivialization, justification or condonation of crimes of genocide, crimes against humanity or war crime which have been found by courts to have occurred and the glorification of persons for having committed such crimes."

Several countries have banned hate speech by law, and such provisions tend to focus on incitement to hatred towards certain people based on their personal characteristics.

Marianne Díaz stated in her paper **"Hate Speech in Latin America: Regulation Trends, the Role of Intermediaries and the Risks to Freedom of Expression,"** the legal approaches most widely adopted in Latin America focus on direct criminal penalties, subsidiary criminal penalties (considered as an aggravating factor to the primary offense) and prohibition, which although it does not set forth criminal penalties, it does provide for reparations. Díaz Hernández adds that the criminal law in several countries in Latin America—Costa Rica, El Salvador, Peru, Argentina, Bolivia, and Uruguay, to name a few—"defines incitement to hatred as a criminal offense."

The criteria adopted to classify incitement to hatred as a crime is not consistent among the countries that have chosen the criminalization approach. Some require proof of actual or potential damage. The Inter-American Commission on Human Rights has stressed that "as a matter of principle, instead of restrictions, States should adopt preventive and educational mechanisms to promote deeper, broader debates to raise awareness on and fight against harmful stereotypes."

However, there is consensus that hate speech may play a role in promoting violence against specific social groups. Legal scholar Alexander Tsesis **argues that the very purpose of intimidating hate speech is to perpetuate and augment existing inequalities**:

"Although the spread of intimidating hate speech does not always lead to the commission of discriminatory violence, it establishes the rationale for attacking particular disfavored groups."

Brutality against the Rohingya people in Myanmar is evidence of the role Facebook content containing hate speech can play in this process. In 2018, an **investigation conducted by Reuters and the Human Rights Center at the University of California, Berkeley, School of Law de Reuters** found over 1000 posts calling the Rohingya and other Muslims dogs, maggots, and rapists.

These posts were created and disseminated at the beginning of Myanmar's army ethnic cleansing and crimes against humanity drive, which pushed 740,000 Rohingya to flee to Bangladesh.

2020: A "TSUNAMI OF HATE AND XENOPHOBIA."

In May 2020, Antonio Guterres, Secretary-General of the United Nations, stated that the COVID-19 pandemic had unleashed a "tsunami of hate and xenophobia, scapegoating and scare-mongering around the world" and called for action "to strengthen the immunity of our societies against the virus of hate."

"Migrants and refugees have been vilified as a source of the virus and then denied access to medical treatment. Contemptible memes about older persons have emerged, suggesting they are also the most expendable. And journalists, health professionals, aid workers, and human rights defenders are being targeted simply for doing their jobs", he added.

At the 13th session of the Forum on Minority Issues in November 2020, Michelle Bachelet, the United Nations High Commissioner for Human Rights, said that social media provide new "opportunities for exercising our fundamental freedoms of expression, association and participation have expanded in unparalleled ways. Yet, this expansion has brought with it new and significant threats to civic space and people's rights."

"One of them is hate speech, which is largely disseminated online through various social media platforms. Minorities have been disproportionately targeted with incitement to discrimination, hostility, and violence. This may lead to tensions, unrest, and attacks against individuals and groups. It may also be used to serve certain political interests, contributing to a climate of fear among minority communities."

The High Commissioner said "the same rights that people have offline must also be protected online" and added that social media companies "have a responsibility to prevent, mitigate and remedy human rights violations that they may cause or contribute to."

"Social media companies have alternatives to either taking down or leaving material online. They can also flag content, add countervailing material, warn the disseminator and suggest self-moderation. Take-downs would only be warranted in the most severe cases. Any solution proposed to tackle hate speech in social media should work towards closing an enormous gap in transparency and democratic accountability in the decision-making of the platforms. Not only should we expect them to follow human rights guidance, but we also need mechanisms to monitor and assess their acts."

Along the same lines, in April 2020, the Special Rapporteur on freedom of religion or belief, Ahmed Shaheed, reported the spread of conspiracy theories on social media, claiming that "Jews or Israel are responsible for developing and spreading the COVID-19 virus."

"Countering online hate speech also will not succeed if the mainstream or social media do not take seriously the reports of cyber hate targeting Jews and other minorities." "They must remove any posts that incite to hatred or violence in addition to identifying and reporting fake news." "At this deeply challenging time, ensuring that all individuals are able to exercise their right to freedom of religion or belief without fear and to the greatest extent feasible while safeguarding public health is more essential than ever," said Shaheed.

There has been an increase in hate speech on social media during the pandemic. In February, the Chinese community was the first to be targeted as COVID-19 emerged in this country. Hate speech was then redirected against the use of masks and even against the LGBTIQ community, claiming they are to blame for the origin of the virus as it is regarded as divine punishment. A **study by L1ght** found a 900% increase in hate speech on Twitter directed towards China and the Chinese and a 200% increase in traffic to hate sites and specific posts against Asians.

In many cases, political leaders from different parts of the world were using these expressions on social media (with millions of followers) and offline. The use of the term **"Chinese virus"** on social media by the then President of the United States, Donald Trump, and the term "Wuhan virus" by the then also Secretary of State, Mike Pompeo, may be used to have fueled **hate speech in the US.**

In February 2020, Luca Zaia, governor of the Italian region of Veneto, one of the first pandemic hot spots, told journalists her country would fight the virus more efficiently than China thanks to **the personal hygiene standards of our people...the culture of Italian citizens, we are used to taking showers and washing our hands frequently (...), while we have all seen videos of Chinese people eating live mice."**

In April that year, the then Brazilian Minister of Education, **Abraham Weintraub** tweeted the pandemic was part of the Chinese government's "plan for world domination."

The increase in racist rhetoric on social and mainstream media is in line with the rise in violent acts against the same communities in several parts of the world. **In the UK, people of Asian descent have been attacked, targeted,** and blamed for spreading the coronavirus. In Australia, two women attacked Chinese female students, punching and kicking one of them and yelled, **"go back to China"** and **"damn immigrants."** In Spain, a young **Asian American was assaulted by two men** and was left in a coma for two days. In Texas, in the US, a man **attacked a Burmese family with a knife** accusing them of infecting people with coronavirus.

In Africa, there have been reports in **Kenya, Ethiopia,** and **South Africa** of discrimination and attacks against people of Asian descent– and foreigners in general—accusing them of carrying the coronavirus.

Cases have also been reported in Latin America. In Brazil, the media has reported **abuse and harassment** against people of

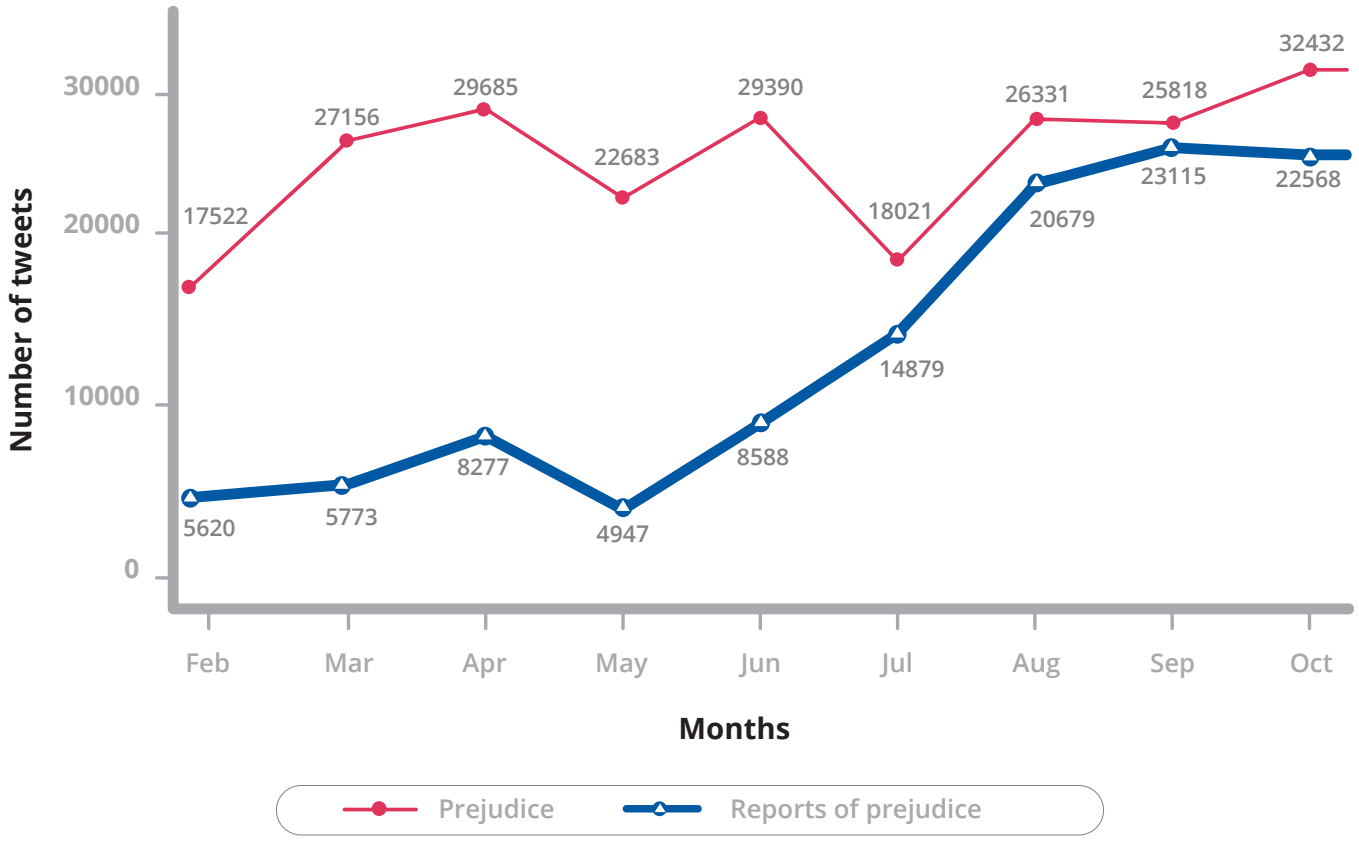
Asian descent. In one of the cases, a Law student said she was the victim of racism and xenophobia by a female passenger in the Rio de Janeiro subway. Marie Okabayashi tweeted, "the woman waited until I was close to the door and yelled 'look at the Chinese girl leaving, Chinese pig,' 'filthy' and 'she'll get us all sick if she stays,'" together with a video of the attacker.

Mexican historian Yuriko Valdez, of Chinese descent and author of the documentary "The Legacy of My Race. Chinese and Mixed-race in Mexicali", warns about the spread of xenophobia in this town and the myriad of racist comments on social media on posts about holidays like the Chinese New Year on January 25th. In addition to the usual "Chinese people eat rats and dogs" comments, we are now seeing comments like "Chinese people are pigs" or "they are going to infect us because China is the source of the infection." People who are "proud to truly be from Mexicali" reacted with similar comments to the launch of an exhibit of the Chinese Association at the City Zoo: "The Chinese don't deserve to be paid tribute" and "they are sick with coronavirus," are among the messages reported by Valdez.

"COVID-19-related expressions of racism and xenophobia online have included harassment, hate speech, the proliferation of discriminatory stereotypes, and conspiracy theories. Not surprisingly, leaders who are attempting to attribute COVID-19 to certain national or ethnic groups are the very same nationalist populist leaders who have made racist and xenophobic rhetoric central to their political platforms," stated E. Tendayi Achiume, Special Rapporteur on **contemporary forms of racism racial discrimination, xenophobia, and related intolerance.**

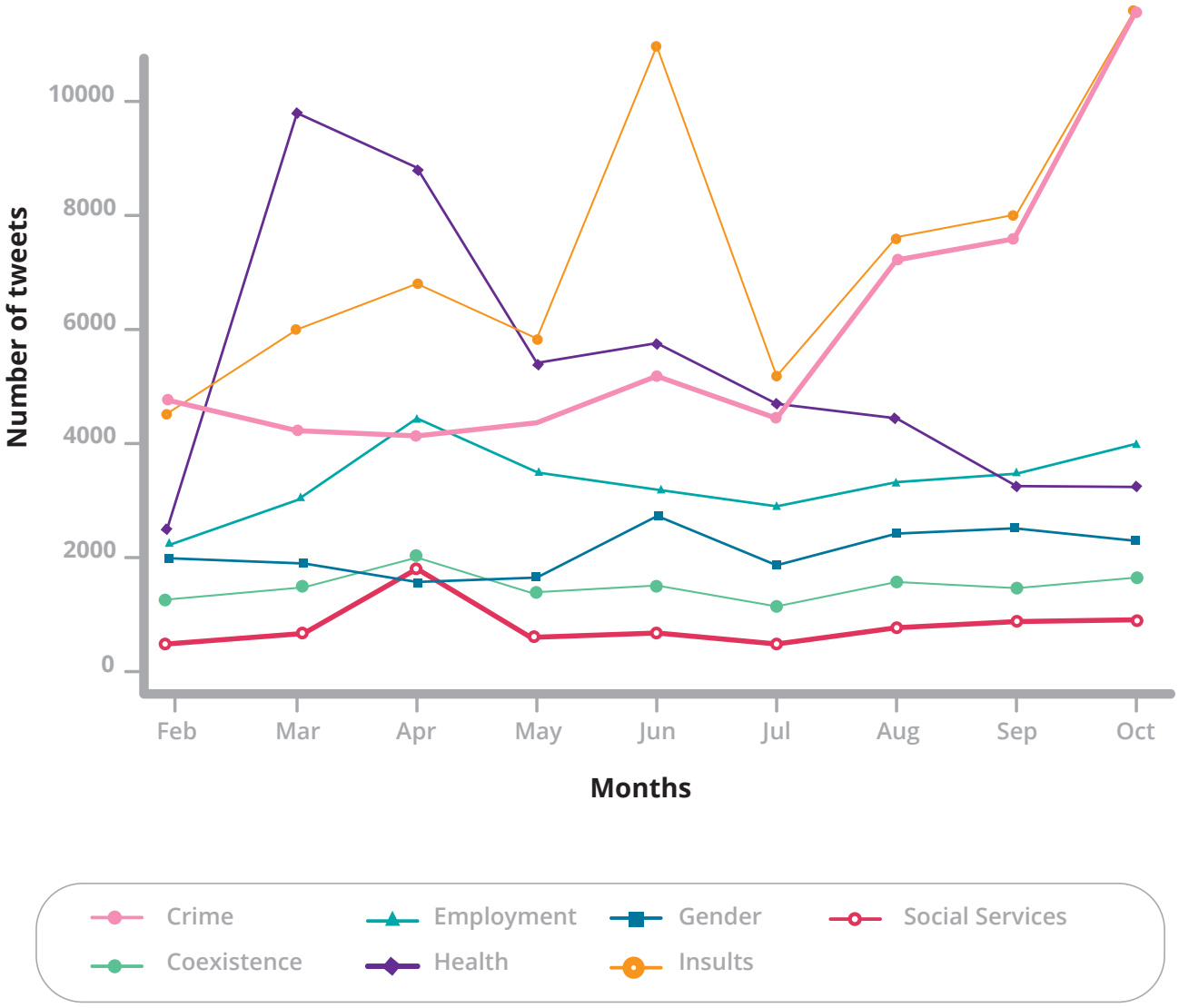
The Migration Unit of the Inter-American Development Bank (IDB) carried out a study between February and December 2020 monitoring threads on **Twitter about immigration**. The study monitored seven countries in the region considered significant receiving countries: Argentina, Chile, Colombia, Costa Rica, Ecuador, Panamá, and Peru. They monitored tweets mentioning the terms asylum, xenophobia, migrant, immigrant, refugee, and exile. Once identified, an algorithm classified them into eight mutually exclusive categories. The first seven categories grouped tweets expressing prejudice against migrants. These were: Crime, Employment, Gender, Social Services, Coexistence, Health, and General Insults. The eighth category comprises tweets reporting or condemning these prejudices.

The study used February tweets as a pre-pandemic baseline, and it found an increase in prejudice towards migrants of 70 % in two months, from 17,522 monthly tweets in February to 29,685 in April.



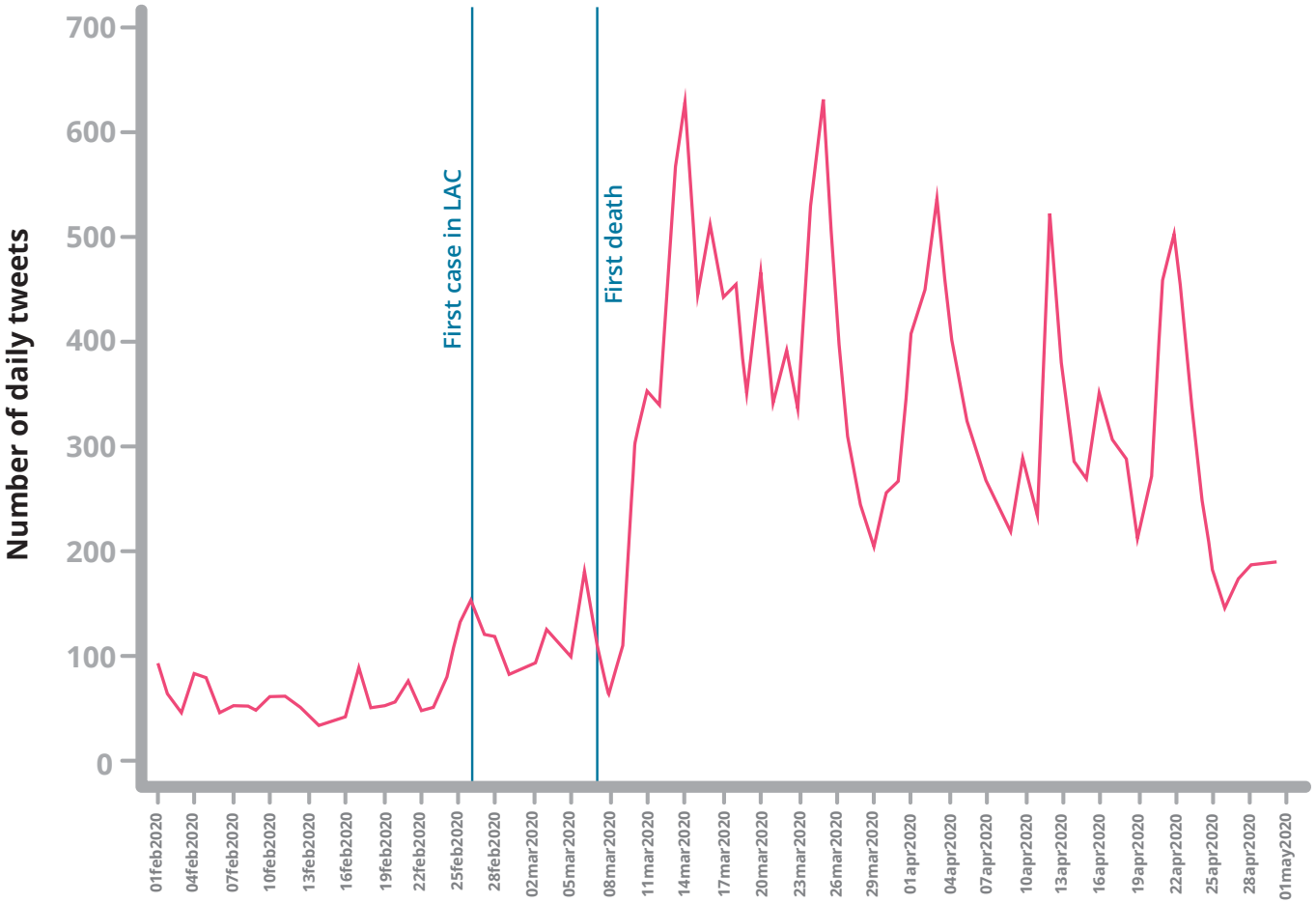
Source: The Migration Unit of the IDB based on Citibeats data.

According to the study, the increase observed between February and April can be explained by health-related biases "mainly fear of immigrants transmitting the disease or causing healthcare systems to collapse."



Source: The Migration Unit of the IDB based on Citibeats data.

The IDB study states that “prejudice spiked after the first COVID-19 death was announced in the region”— **In March 2020 in Argentina.”**



Source: The Migration Unit of the IDB based on Citibeats data.

The authors reported that they observed fluctuations in the levels of xenophobia or prejudice in the following months, although levels remained higher than those in February before the pandemic. There was an increase in October, which can be explained by other factors (as crime rate) that are not directly linked to the pandemic and a reduction in health-related tweets.

THE "HATE SPEECH" POLICIES OF MAJOR PLATFORMS DURING THE PANDEMIC

"I just believe strongly that Facebook shouldn't be the arbiter of truth of everything that people say online.".

Mark Zuckerberg repeated this statement over the years, and it is a good description of the attitude social media platforms had adopted on content moderation until 2020. Even after the 2016 elections in the US, when Facebook, Twitter, and YouTube were harshly criticized for their role in disseminating disinformation, hate speech, and conspiracy theories, they remained reluctant to take action in this regard.

However, this changed in 2020. Facebook, Twitter, and YouTube made changes to their community guidelines and terms of service—something they had fought back for years—and are now flagging content on the accounts of public figures as fake news and have even deleted the posts of a sitting US President and suspended his account.

In June 2020, the death of George Floyd, an African American man, after being arrested by four Minneapolis police officers, sparked protests around the world against racism and police brutality. The then President of the United States, Donald Trump, made **several posts on his social media and, in one, in particular, wrote: "When the looting starts, the shooting starts."** The African American community widely interpreted this as a violent threat against protesters. Twitter hid the post. Facebook didn't.

Amid criticism, Facebook's CEO, Mark Zuckerberg, explained his reasons for keeping Trump's post online: "I disagree strongly with how the President spoke about this, but I believe people should be able to see this for themselves because ultimately accountability for those in positions of power can only happen when their speech is scrutinized out in the open."

Weeks later, a group of companies, among them Unilever, Coca-Cola, Verizon, and Honda, announced the Stop Hate for Profit campaign's launch and paused all paid advertising on the platform for one month. Unilever's Media VP, Luis Di Como, said that "to continue advertising on these platforms now would not add value to the people and society." **"Given the existing polarization and the ongoing elections in the United States there is a need for more strict enforcement when it comes to hate speech," he claimed.**

"We respect any brand's decision and remain focused on the important work of removing hate speech and providing critical voting information," said Carolyn Everson, VP of Global Business Group at Facebook, on Monday. "Our conversations with marketers and civil rights organizations are about how, together, we can be a force for good."

However, in January 2021, Trump was permanently banned from Twitter and Facebook, and some of his videos were deleted from YouTube following comments about alleged election fraud in the US addressed at his supporters who stormed the Capitol building in Washington, sparking violence and fear among lawmakers y officials, and resulting in **several deaths.**

"The shocking events of the last 24 hours clearly demonstrate that President Donald Trump intends to use his remaining time in office to undermine the peaceful and lawful transition of power to his elected successor, Joe Biden," Zuckerberg said in his Facebook post explaining the ban.

FACEBOOK HATE SPEECH REMOVAL

Facebook's Community Standards define hate speech as "a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease."

Facebook states that "We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic."

And they add: "We recognize that people sometimes share content that includes someone else's hate speech to condemn it or raise awareness. In other cases, speech that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to indicate their intent clearly. If the intention is unclear, we may remove content."

The company classifies hate speech into three tiers depending on the severity of the social media post. Tier 1 covers all "Content targeting a person or group of people on the basis of their aforementioned protected characteristics with violent speech or support in written or visual form and dehumanizing speech or imagery in the form of comparisons, generalizations, or unqualified behavioral statements (in written or visual form) to or about: insects, animals that are culturally perceived as intellectually or physically inferior, filth, bacteria, disease and feces, sexual predators, subhumanity, violent and sexual criminals, other criminals (including but not limited to "thieves," "bank robbers," and statements denying existence, mocking the concept, events or victims of hate crimes even if no real person is depicted in an image."

Also considered Tier 1: "Designated dehumanizing comparisons, generalizations, or behavioral statements (in written or visual form) that include black people and apes or ape-like creatures, black people and farm equipment, caricatures of Black people in the form of blackface, Jewish people and rats, Jewish people running the world or controlling major institutions such as media networks, the economy, or the government, denying or distorting information about the Holocaust, Muslim people and pigs, Muslim people and sexual relations with goats or pigs, Mexican people, and worm-like creatures, women as household objects or referring to women as property or "objects," Transgender or non-binary people referred to as "it," Dalits, scheduled caste or 'lower caste' people as menial laborers."

Facebook considers Tier 2 Hate Speech all "content targeting a person or group of people on the basis of their protected characteristics with generalizations that state inferiority (in written or visual form) such as 'physical deficiencies' about hygiene, including but not limited to: 'filthy,' 'dirty,' 'smelly'; statements about physical appearance, including but not limited to: 'ugly,' 'hideous'; about mental deficiencies, including but not limited to 'dumb,' 'stupid,' 'idiots'; about education, including but not limited to 'illiterate,' 'uneducated'; about mental health, including but not limited to 'mentally ill,' 'retarded', 'crazy,' 'insane'; and moral deficiencies, which are defined as those about character traits culturally perceived as negative, including but not limited to 'coward,' 'liar,' 'arrogant,' 'ignorant; and derogatory terms related to sexual activity, including but not limited to ' whore', 'slut', 'perverts.'" Also considered under Tier 2 are "other statements of inferiority, which we define as expressions about being less than adequate, including but not limited to: 'worthless,' 'useless.' Expressions about being better/worse than another protected characteristic. Expressions about deviating from the norm, including but not limited to: 'freaks,' 'abnormal.' Expressions of contempt, such as self-admission to intolerance on the basis of a protected characteristics,

including but not limited to: 'homophobic,' 'islamophobic,' 'racist.' Expressions that a protected characteristic shouldn't exist, and expressions of hate and dismissal, and disgust including but not limited to: 'don´t respect,' don't like, don´t care for, 'vomit,' 'throw up,' 'vile,' 'disgusting,' etc." This category also includes cursing: "Referring to the target as genitalia or anus; profane terms or phrases with the intent to insult; terms or phrases calling for engagement in sexual activity, or contact with genitalia, anus, feces or urine."

Finally, Facebook classifies as Tier 3 Hate Speech all content in written or visual form with any of the following: "Segregation in the form of calls for action, statements of intent, aspirational or conditional statements, or statements advocating or supporting segregation. Exclusion in the form of calls for action, statements of intent, aspirational or conditional statements, or statements advocating or supporting, defined as explicit exclusion, which means things like expelling certain groups or saying they are not allowed, political exclusion, which means denying the right to political participation, economic exclusion, which means denying access to economic entitlements and limiting participation in the labor market, social exclusion, which means things like denying access to spaces (physical and online) and social services," and "content that describes or negatively targets people with slurs, where slurs are defined as words that are inherently offensive and used as insulting labels."

In July 2020, the **Stop Hate for Profit campaign** brought together over 1200 companies from around the world and called for an ad boycott against major platforms demanding hate speech moderation and an ad pause on accounts that promote discrimination against certain groups. One of the main demands of the participating companies and organizations was the suspension of all of Trump's social media accounts.

The coalition called for platforms to remove "groups or accounts focused on white supremacy, militia, anti-Semitism, Islamophobia, and violent conspiracies" and to "increase resources focused on monitoring groups for hate speech and violence," "change platform policy to forbid any event page with a call to arms," and to "commit 5% of their annual revenue to an independently administered fund to support initiatives, academics and organizations doing the work to fight against racism, hate, and division caused by Facebook's inaction."

In a June 2020 post, Facebook **addressed some of the Stop Hate for Profit demands**. With regards to the organizers' request to "create a separate moderation pipeline staffed by experts on identity-based hate for users who express they have been targeted," Facebook stated that "hate speech reports on Facebook are already automatically funneled to a set of reviewers with specific training in our identity-based hate policies in 50 markets covering 30 languages. In addition, we consult with experts on identity-based hate in developing and evolving the policies that these trained reviewers enforce." They also announced they "intend to include the prevalence of hate speech in future Community Standards Enforcement Reports (CSER), pending no further complications from COVID-19."

That same month, Richard Allan, Facebook's VP of Public Policy, wrote a column addressing the differences in the definition of hate speech worldwide and the challenges the platform faces to **identify it and take action**. "There is no universally accepted answer for when something crosses the line. Although several countries have laws against hate speech, their definitions of it vary significantly." In Germany, for example, laws forbid incitement to hatred; you could find yourself the subject of a police raid if you post such content online. In the US, on the other hand, even the most vile kinds of speech are legally protected under the US Constitution", Allen stated. "People who live in the same country — or next door — often have different levels of tolerance for speech about protected characteristics. To some, crude humor about a religious leader can be considered both blasphemy and hate speech against all followers of that faith. To others, a battle of gender-based insults may be a mutually enjoyable way of sharing a laugh. Is it OK for a person to post negative things about a certain nationality as long as they share that same nationality? What if a young person who refers to an ethnic group using a racial slur is quoting from lyrics of a song?" the Facebook executive asked.

Allen also discusses the mistakes made in content moderation when content is misclassified as hate speech. He pointed out that "If we fail to remove content that you report because you think it is hate speech, it feels like we're not living up to the values in our Community Standards. When we remove something you posted and believe is a reasonable political view, it can feel like censorship. We know how strongly people feel when we make such mistakes, and we're constantly working to improve our processes and explain things more fully."

He added that Facebook's mistakes "have caused a great deal of concern in a number of communities, including among groups who feel we act—or fail to act—out of bias."

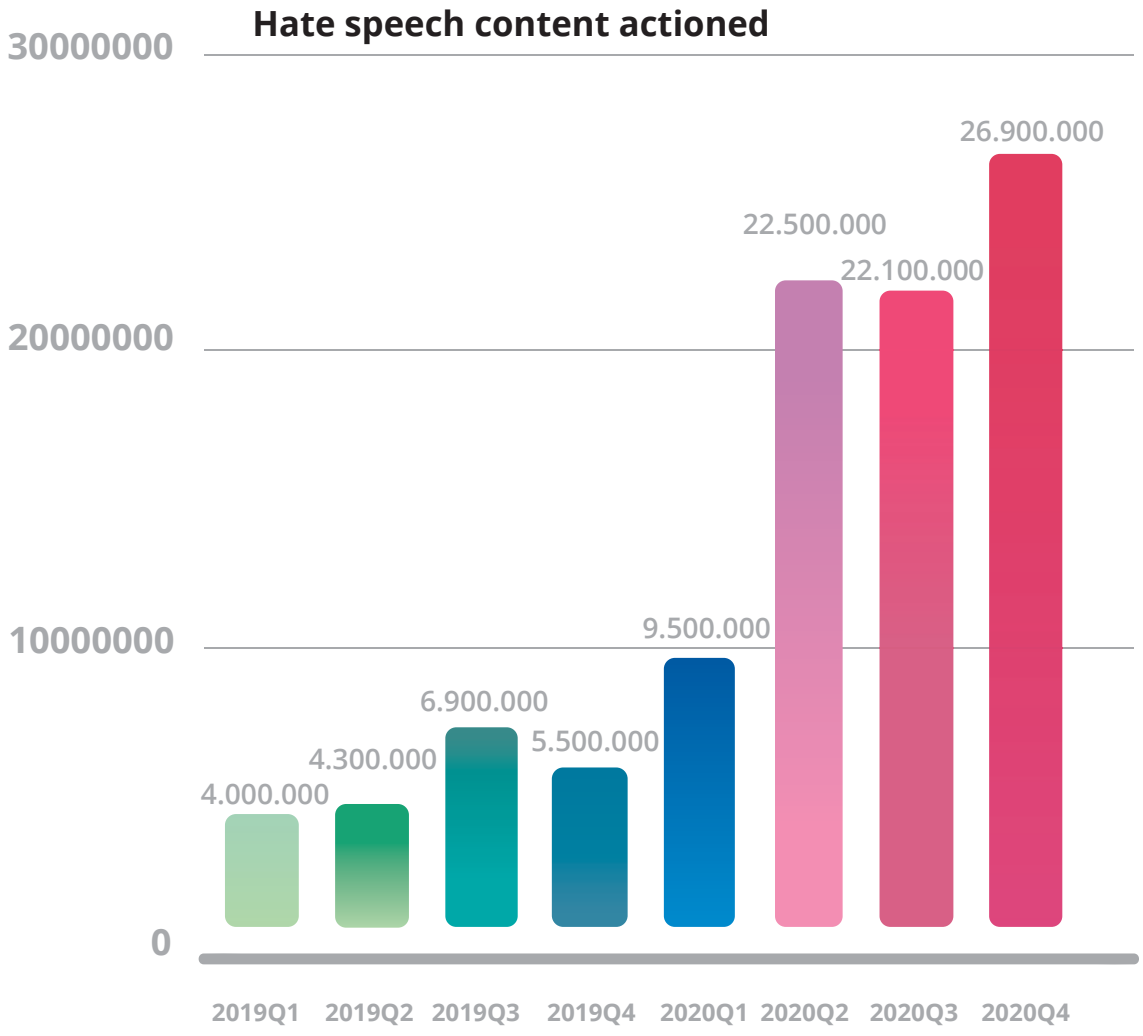
"Last year (2019), Shaun King, a prominent African-American activist, posted hate mail he had received that included vulgar slurs. We took down Mr. King's post in error, not recognizing at first that it was shared to condemn the attack," he said. In July, Nick Clegg, Facebook's VP of Global Affairs and Communications wrote an article **stating that the company had adopted several measures and made significant progress towards the elimination of hate speech on their platform**. Clegg wrote, "A recent European Commission report found that Facebook assessed 95.7% of hate speech reports in less than 24 hours, faster than YouTube and Twitter. Last month, we reported that we find nearly 90% of the hate speech we remove before someone reports it—up from 24% a little over two years ago. We took action against 9.6 million pieces of content in the first quarter of 2020—up from 5.7 million in the previous quarter. And 99% of the ISIS and Al Qaeda content we remove is taken down before anyone reports it to us."

According to the Community Standards Enforcement Report (CSER) published in February 2021, the number of pieces of content on which Facebook took action increased from 20,700,000 in 2019 to 81,000,000, which means there was a 300% increase in the content classified as hate speech in one year.

In November, Facebook started monitoring the hate speech prevalence on their platform and found that it ranged between 0.10% and 0.11% between July and September. This means that for every 10,000 posts viewed on the platform, around 10 or 11 would be classified as hate speech by Facebook. Prevalence went down to 0.07% and 0.08% between October and December 2020. Facebook's report doesn't state whether this fall results from a general increase in the number of posts, an actual reduction in the hate speech category, or a change in the monitoring processes and criteria.

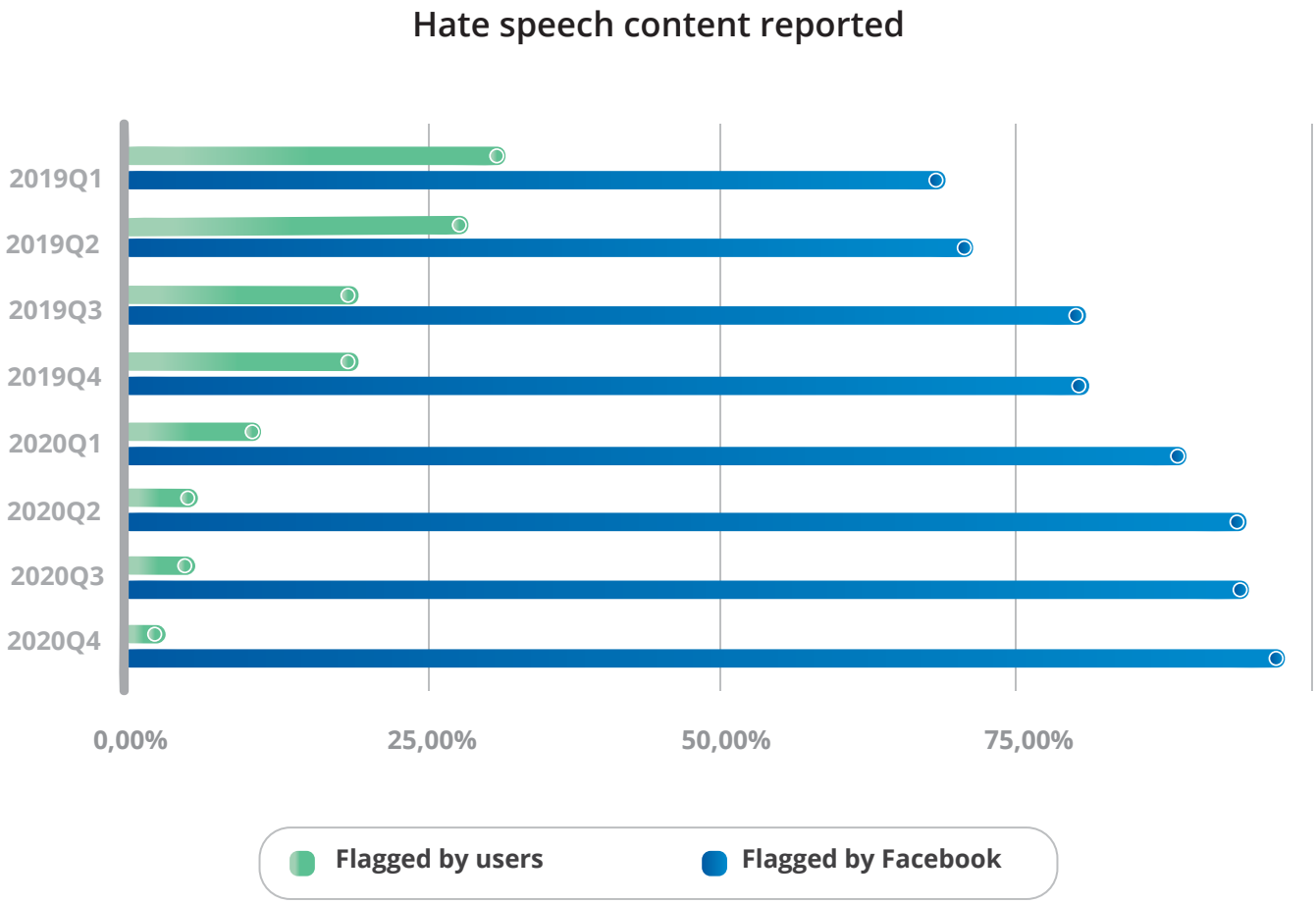
If we look at 2020 in depth, we can see a sharp increase in hate speech content actioned by Facebook during the second quarter. Between January and March, the company took action on 9,500,000 pieces of content, while in the following months, the figure doubled to 22,500,000 between April and June, 22,100,000 between July and September, and 26,900,000 between October and December.

The CSER shows that the number of pieces detected and the percentage of proactive detection are mainly the result of "enhanced monitoring systems in Arabic and Spanish" and "increased automated systems in Portuguese."



Another interesting aspect is the increase in the percentage of content spotted by Facebook compared to the content reported by users out of the total hate speech content actioned. The relevance of Facebook's internal systems on the overall hate speech content detection has been gradually increasing since 2018, accounting for almost the total in the last quarter of 2020.

In the last quarter of 2017, Facebook took action on 1,700,000 pieces of hate speech content. Out of these, 76.4% were flagged by users. In 2020, this percentage took a turn. Between January and March, 89.3% of the hate speech content was detected by Facebook's systems, and the same happened between April and June (94.7%), July and September (94.7%), and October and December (97.1%).



Something similar happened on Instagram—also owned by Facebook—where actions on hate speech content are monitored since the last quarter of 2019. Between January and March 2020, Instagram detected and took action on 578,000 pieces of content considered to fit the hate speech definition. That figure rose to 3,200,000 between April and June, 6,500,000 between July and September, and 6,600,000 between October and December 2020.

In the first quarter, 57.1% of the content was flagged by users, while in the second quarter the situation was reversed, and users accounted for only 15.1% of the hate speech actioned content. The same trend was observed during the following quarters: 5.2% between July and September, and 4.9% between October and December.

By mid-march 2020, Facebook sent over 15,000 content moderators in 20 locations home, following their requests amid the COVID-19 lockdown measures.

Facebook's CEO, Mark Zuckerberg, said that week that during the ongoing coronavirus pandemic, Facebook is "relying more heavily on AI software for content moderation decisions." The company also announced full-time training efforts to pay "extra attention" to "highly sensitive" content. Users should expect "more mistakes while the company expedites the process, in part because only a fraction of the humans will be involved and the software makes more naive decisions than humans, and there could be a rise in 'false positives,' including removal of content that should not be taken down." "It will create a trade-off against some other types of content that may not have as imminent physical risks for people," **Zuckerberg explained.**

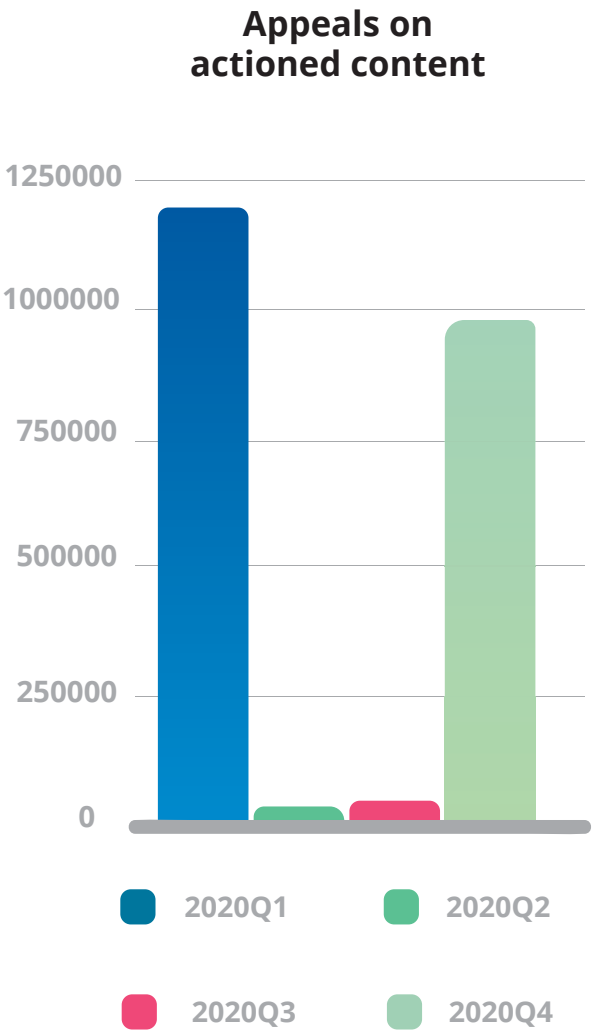
In November 2020, Facebook **announced changes in their moderation systems that implied** an increase in machine moderation during the initial stages of content review. Chris Palow, a software engineer in Facebook's interaction integrity team, admitted during a press conference that "AI is never going to be perfect" and that "AI has its limits" to sort between what should be flagged as hate speech and what should not be, for example, because it is meant as parody or satire. "The system is about marrying AI and human reviewers to make fewer total mistakes," he said. Facebook hasn't disclosed what percentage of posts are misclassified as content to be removed.

Months later, in February 2021, Facebook's organic content policy manager, Varun Reddy acknowledged that the platform was having problems because of the lack of human reviewers in most content moderation processes. "AI learns from human moderators," he explained, adding that the reduction in human moderation has changed **"how effective the AI is over time."**

"We're working with partners to get as much capacity back online as we can. We're not there yet, but it has improved significantly since lockdown began on March 25th. In the coming weeks and months, we are hopeful the systems will come back to full efficacy," Reddy said in February this year.

The users' appeals process to report content allegedly wrongly removed was also affected by the lockdown measures and Facebook employees being home. Facebook's Report pointed out that "Due to a temporary reduction in our reviewer capacity as a result of COVID-19, we cannot always offer our users the option to appeal. We still gave users the option to tell us when they disagree with our decision, which has helped review many of these cases and restore content where appropriate." The report shows virtually no appeals between April and June 2020, only 70,000 worldwide in the six-month period. This is a minimal number if we consider there had been 1,200,000 the quarter before. In the following period, between October and December, the number rose to 984,200 cases.

In 2020, Facebook also reached record numbers of restored content compared to previous periods. They went from 483,400 pieces of content in 2019 to 703,200 in 2020. Out of the latter, Facebook restored 589,300 with no appeals.



HATE SPEECH CONTENT MODERATION ON TWITTER

In December 2020, Twitter announced they were updating the rules against hate speech on the platform and based the decision on the fact that "research shows dehumanizing language increases the risks of offline harm." In July 2019, Twitter expanded its rules against hate speech to include religion or caste as protected characteristics. In March 2020, they included age, disability, or disease, and in December 2020, **they announced a ban on language that dehumanizes people on the basis of race, ethnicity, or national origin.**

Several examples were included to show the kind of content that would not be allowed after the announcement:

> ***"All (national origin) are cockroaches who live off of welfare benefits and need to be taken away;" "People who are (race) are leeches and only good for one thing;" "There are too many (national origin, race, ethnicity) maggots in our country and they need to leave;" "All (age group) are leeches and don't deserve any support from us;" "People with (disease) are rats that contaminate everyone around them;" "(Religious group) should be punished. We are not doing enough to rid us of those filthy animals;"***

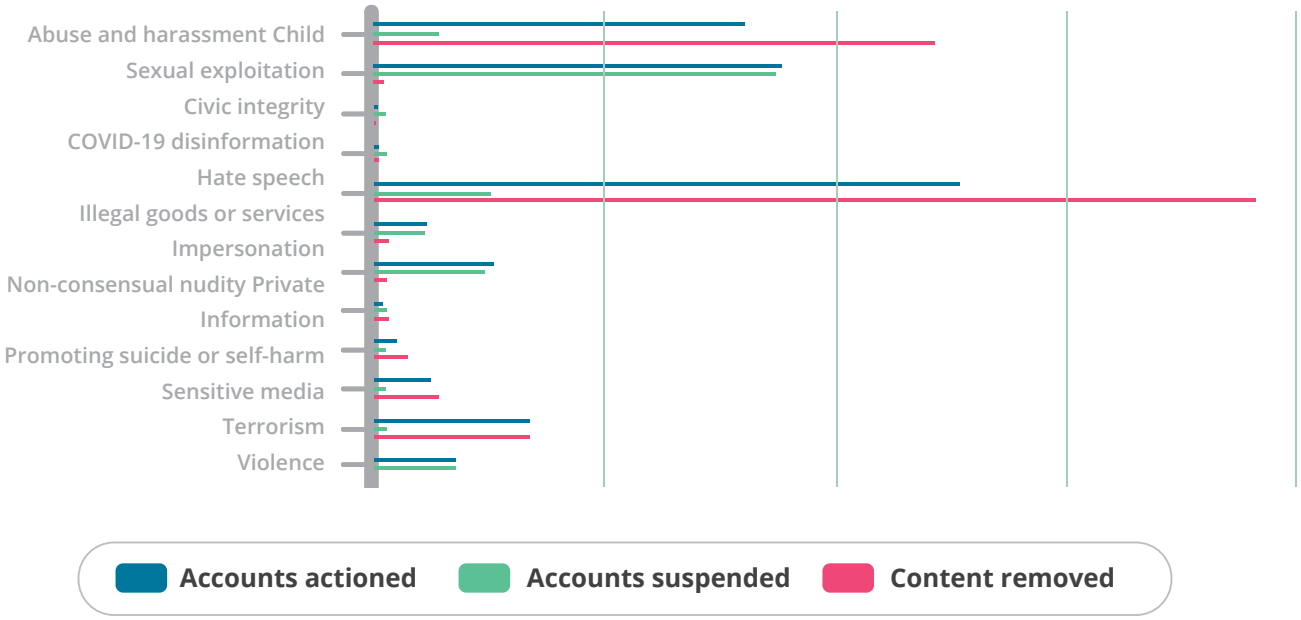
In October 2019, the now Vice President of the United States, Kamala Harris, published an **open letter** to Jack Dorsey, the CEO of Twitter, requesting some of the then President Donald Trump posts be removed for violating Twitter's community guidelines, among those, the provisions on hate speech. "No user, regardless of their job, wealth, or stature, should be exempt from abiding by Twitter's user agreement," pointed out Harris in her letter.

That same year, a study from **New York University (NYU)** found a link between the number of racist tweets and the number of hate crimes in 100 cities across the United States. Rumi Chunara, one of the study authors, said, "I think there is a sentiment in the targeted tweets that is likely related to fostering an environment for these crimes." And he added, **"Meanwhile, having a productive conversation might actually improve culture and outcomes."**

"Right now the system makes it super easy to harass and abuse others," said Dorsey in 2019, and he added that "one of the problems is that it places undue weight on followers and likes."

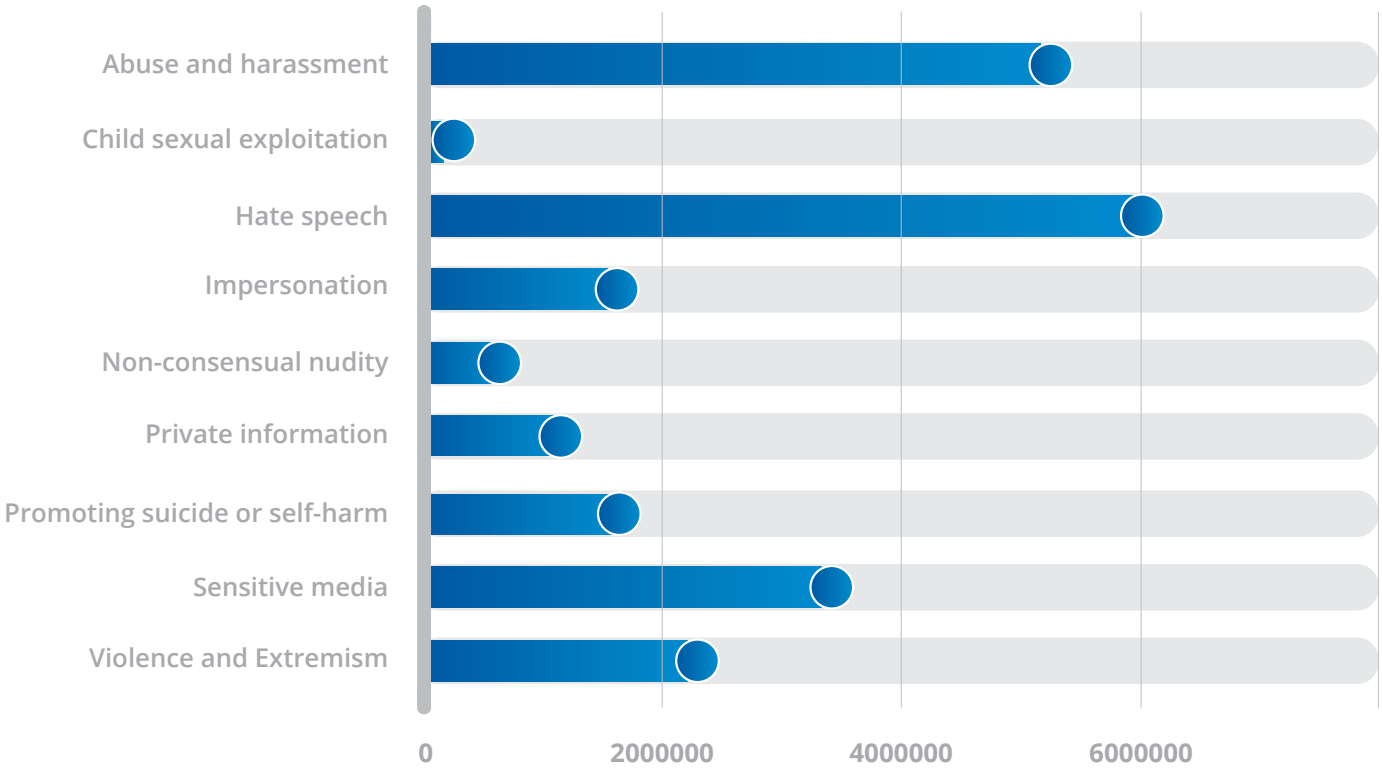
What happened in 2020 and during the COVID-19 pandemic? According to the **latest Twitter Transparency Report** available, between January and June that year, the company took action on 1,940,082 accounts, out of which 925,954 were suspended and 1,927,063 pieces of content removed. During the same period in 2019, a similar amount of content was removed (1,914,471), but fewer accounts were suspended (687,397).

Accounts actioned in January-June by removal reason



Actions were taken on 645,416 accounts on the basis of content flagged as hate speech, which accounts for 33.2% of all accounts actioned. A total of 12,400,000 accounts were reported in the January-June period, out of which almost half (6,055,642) were reported for hate speech content. The report states that 30% more accounts were reported than in the same period the year before.

Accounts reported January - June



Twitter reports a 35% reduction in the number of accounts actioned on the basis of hate speech compared to the previous period. However, they acknowledge that given the circumstances, the teams mainly focused on content that could lead to harm or the dissemination of COVID-19 misleading information, which caused a "significant setback in all the remaining areas."

In April 2020, Twitter posted an announcement on their blog reporting some changes resulting from their decision to send most employees home to support the social-distancing measures adopted by governments worldwide.

One of the measures adopted was to "increase the use of machine learning and automation to take a wide range of actions on potentially abusive and manipulative content." The post stated: "We want to be clear: while we work to ensure our systems are consistent, they can sometimes lack the context may result in us making mistakes. As a result, we will not permanently suspend any accounts based solely on our automated enforcement systems. Instead, we will continue to look for opportunities to build in human review checks where they will be most impactful."

The company announced that automated technology would be used during the COVID-19 pandemic to "surface content that's most likely to cause harm and should be reviewed first," and to "proactively identify rule-breaking content before it's reported (our systems learn from past decisions, so over time, the technology can help us rank content or challenge accounts automatically)." Twitter pointed out that "For content that requires additional context, such as misleading information around COVID-19, our teams will continue to review those reports manually."

The social media platform announced their response times would be "longer than normal" and acknowledge that since "automated systems don't have all of the context and insight our team has, we'll make mistakes."

YOUTUBE HATE SPEECH CONTENT MODERATION DURING THE PANDEMIC

YouTube's last Community Guidelines update on hate speech dates back to 2019. The Google-owned company currently defines hate speech as "content that promotes violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity, national origin, race, immigration status, religion, sex or gender, sexual orientation, victims of a major violent event and their kin and veteran status."

The Community Guidelines also state that YouTube won't allow content "dehumanizing individuals or groups based on the attributes noted above, that states they are physically or mentally inferior, or that praises or glorifies violence against them." They don't allow content "that uses stereotypes that incite or promote hatred based on any of the attributes noted above or racial, religious or other slurs intended to promote hate," or "allege the superiority of a group

over those with any of the attributes noted above to justify violence, discrimination, segregation, or exclusion," or "denies that a well-documented, violent event took place."

In March 2021, there was a heated debate over YouTube's hate speech policies when they removed a video from commentator Steven Crowder citing violations of the platform's COVID-19 misinformation policy. In the video, Crowder made several comments about the Republican administration's decision to give a subsidy to racial minority farmers on the grounds that they had been historically excluded from policies to help that sector. **The comments mocked the ways African Americans speak, move, and think.**

After complaints from several racial minority organizations, YouTube issued a statement in which it assured that its "policies prohibit content that promotes hate towards groups based on their race" but "while offensive, this video from the Steven Crowder does not violate this policy."

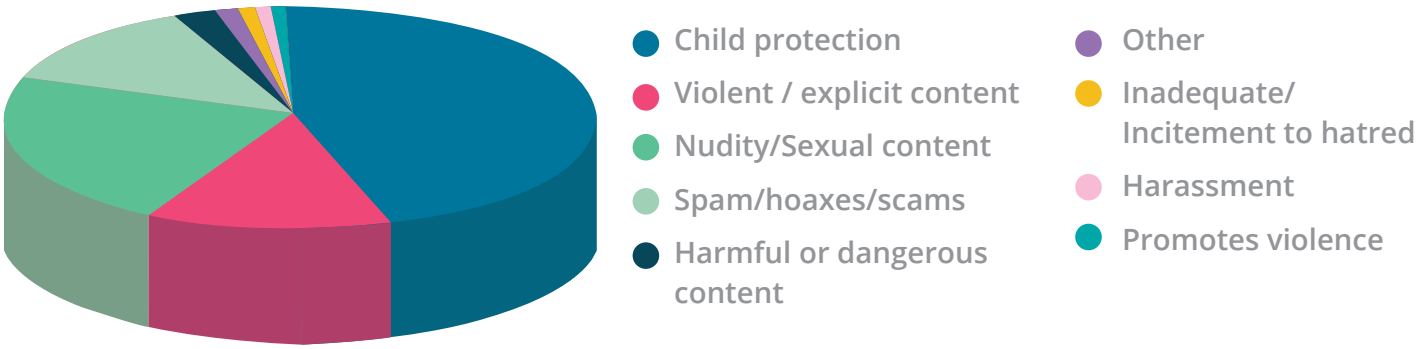
In April 2021, YouTube released information claiming that it had enhanced its hate speech detection systems on the platform. "We don't want YouTube to be a platform that can cause egregious real-world harm," **said Chief Product Officer Neal Mohan.**

Hate speech seems to be harder to detect on YouTube. It is not clear from the available data that there has been a significant increase in hate speech on this platform. However, there were isolated episodes that were highlighted in the media and public opinion.

Between April and June 2020, YouTube took down 11,401,696 videos, in addition to the more than 30,000,000 videos that were removed after 1,998,635 channels were deleted in the same period. Of these more than eleven million videos, only 552,062 videos were deleted without using automatic detection systems. Between July and September, 7,872,684 videos were removed and only 481,721 without automatic detection. Between October and December, 9,321,948 videos were removed and only 521,866 without the use of automated detection systems to spot violations to YouTube's Community Guidelines.

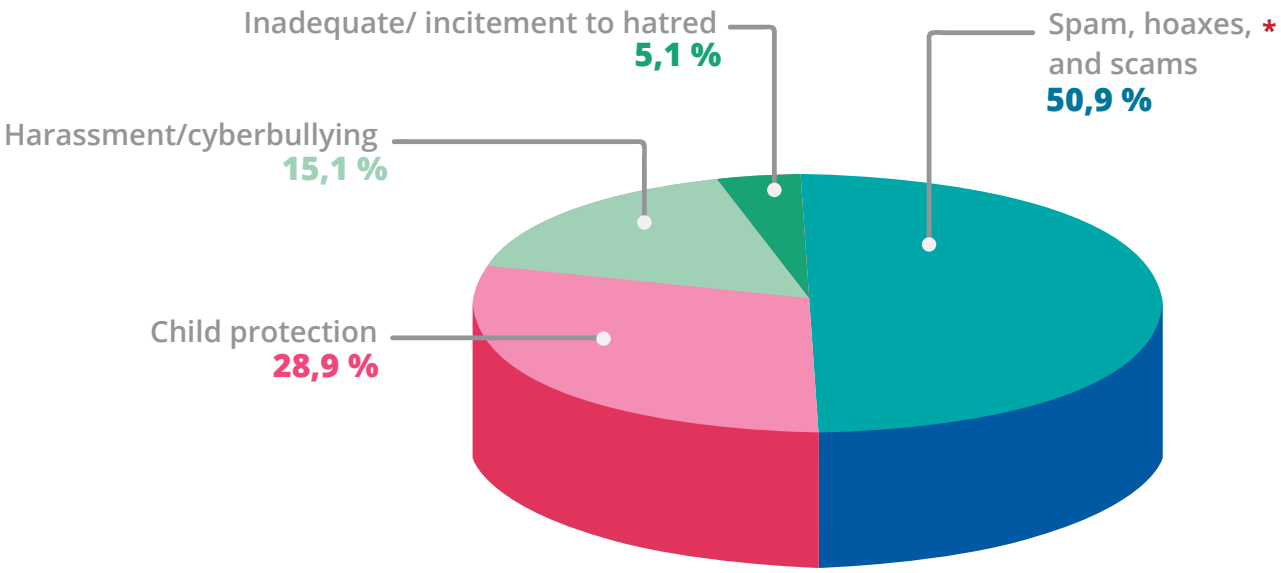
Regarding the reasons for flagging content, hate speech did not account for a significant number since only 97,362 videos were removed for this reason in the last quarter of 2020. However, it did see a slight increase between the April-June quarter and the following two, going from 0.7% of the videos removed to more than 1%.

Videos removed by removal reason - October - December



If we look at the video comments removed, some 906,196,160 were taken down in the last quarter of 2020, and "incitement to hatred" climbs to 5% among the reasons specified for removing such content. This means that in the last quarter of 2020, more than 46 million comments were removed from YouTube because the automated moderation systems considered that they violated the platform's hate speech rules.

Comments removed by removal reason - Oct-Dec



After YouTube sent its content moderators home in March due to the COVID-19 pandemic and significantly extended the use of automated filters, the number of videos removed doubled in the second quarter of 2020. This growth opened a debate at the Alphabet-owned social network about the accuracy of automated moderation processes.

"As the coronavirus response evolves, we are taking steps to prioritize the well-being of our employees and reducing in-office staffing. As a result, we will temporarily start relying more on technology to help with some of the work normally done by human reviewers, which means we are removing more content that may not be violative of our policies. This impacts some of the metrics in this report and will likely continue to impact metrics going forward," the company wrote in a blog post accompanying its transparency **report for the latest quarter.** "Since accountability is our top priority, we chose the latter: to use technology to help with some of the work normally performed by human reviewers," Google explained.

In the second-quarter report, YouTube admitted the increase in content removal resulted from the company accepting "lower efficiency levels to make sure we are removing as many pieces of content as possible."

"One of the decisions we made at the beginning of the pandemic when it came to machines which couldn't be as precise as humans, we were going to err on the side of making sure that our users were protected, even though that might have resulted in a slightly higher number of videos coming down," Neal Mohan, YouTube's chief product officer told **the specialized American site Mashable.**

In September, YouTube announced human moderators are were getting back to the office, and they would work on their moderation systems to try to get back to early 2020 figures.

As mentioned before, YouTube executives admitted that automatic detection systems led to the company removing a lot of content that did not violate its Community Guidelines. As a result, the number of appeals doubled from 166,000 in the first quarter to 325,000 in the second quarter of 2020.

Unlike Facebook, YouTube did not put the appeals process on the back burner and maintained the pre-COVID-19 process timelines. As a result, the number of videos restored after the appeal went from 41,000 to 161,000 during that period. **This meant an increase in the number of successful appeals as YouTube usually reverses its rulings on less than 25%, and it now went up to almost half.**

In its Transparency Report, YouTube explains its hate speech moderation process and discusses some of the challenges of this type of content compared to others that violate their Community Guidelines.

"Hate speech is a complex policy area to enforce at scale, as decisions require a nuanced understanding of local languages and contexts. To help us consistently enforce our policy, we have expanded our review team's linguistic and subject matter expertise. We're also deploying machine learning to detect potentially hateful content better to send for human review, applying lessons from our enforcement against other types of content, like violent extremism. Sometimes we make mistakes, and we have an appeals process for creators who believe their content was incorrectly removed. We constantly evaluate our policies and enforcement guidelines and will continue to consult with experts and the community and make changes as needed," they stated.

YouTube also points out that "In addition to removing content that violates our policies, we work to reduce recommendations of content that comes close to violating our guidelines. **We also have long-standing advertiser-friendly guidelines that prohibit ads from running on videos that include hateful content."**

CONCLUSIONS

Following intense political, social, and media backlash, Facebook, YouTube and Twitter have in recent months made changes to their hate speech Community Guidelines and have adopted measures they had been deeply reluctant to adopt in the past, and which imply a substantial increase in their role as regulators of what can and cannot be said in these new public spaces.

It is hard to know how successful these changes have been and even define success, as the platforms admit they are not sure whether these measures are working amid the COVID-19 pandemic. In cases like Facebook and Instagram, these measures include highly restrictive actions, and in some cases, the removal of safeguards as appeals processes were brought to a halt for several months. For millions of users in Latin America, this meant removing public interest content and the inability to demand an appeal.

Although we don't have enough elements to thoroughly examine each and every one of the reasons for this change in criteria, the fact is that in 2020 social media platforms made decisions and made changes in their content moderation processes. Those changes in the processes and the Community Guidelines that regulate them meant a dramatic shift from how Facebook, Twitter, and YouTube had treated user-created content so far.

There were two significant developments this year. The first one is a very substantial increase in hate speech posts on social media due to COVID-19. Based on the data provided by the platforms themselves, Facebook seems to be the social media outlet with the most significant increase in such posts. Between 2019 and 2020, the number of hate speech posts actioned by this platform grew by almost 300%. When looking at 2020 figures, it is striking that this growth was more dramatic during the second quarter of the year. As described in the previous chapter, in March—as the COVID-19 pandemic gathered pace around the world—the number of hate speech posts actioned by the company doubled and remained high throughout the year. Twitter and YouTube also saw an uptick in posts, but it was not as significant.

The second development was that due to the increase in hate speech and the demands of civil society, Facebook, Twitter, and YouTube tighten their oversight and enforcement processes and expanded the type of content considered non-compliant with their Community Guidelines. However, analysts worldwide question whether these measures are effective or adequate, and there are major problems in the way they were implemented, as they have affected fundamental rights.

Despite popular belief, social media outlets were never intended as fully open or "unregulated" spaces for interaction. For many years, platforms have been moderating "illegal" content and content that falls under more vague definitions and—although not legally prohibited—is considered indecent, obscene, and not in line with the morals of their countries of origin.

The onset of COVID-19, a global pandemic that forced millions of people to lockdown at home, reduce social contact, and work remotely, had other types of impacts. One of them was an increase in hate speech on social media, and another one—probably less noticeable at first sight—was a change in user-created content moderation. Crowdtangle—a search platform that allows tracking hashtags or keywords on Facebook, I nstagram, and Twitter—conducted a study that shows that between February 2020 and March 2021, there were 43,779 posts on Facebook using the expression "Chinese virus" and a total of 3,535,409 interactions. The two major peaks were recorded in March and April 2020.

Governments worldwide called their citizens to sustained social distancing forcing platforms to send thousands of human reviewers home. This led to a significant increase in automated tools and artificial intelligence to review the millions of posts uploaded to social media every minute. Although automated systems are constantly improving, they cannot yet understand the nuances and differences in the language, quirks, and culture of millions of users worldwide and the use of context to define concepts as complex as hate speech.

According to the UNESCO Countering Online Hate Speech study, there are at least five possible non-regulatory mechanisms to counter online hate speech. They all directly involve platforms as a substantial part of the solution to the problem. In this document, UNESCO suggests oversight and monitoring initiatives by members of the civil society, coordinated actions by NGOs to report hate speech cases to authorities, campaigns to promote Internet Service Providers hosting specific content to become more involved, and user empowerment through media literacy, education, and capacity building on the ethical aspects of freedom of expression online.

It is also clear that any mistakes that social media platforms make in flagging hate speech may remove content that does not meet the criteria. It would therefore be a violation of freedom of expression as a human right.

Platforms have grown exponentially worldwide and have become forums for exchange; thus, what happens in these environments directly affects public debate (or could potentially affect it). Allowing both governments and platforms to become content moderators may silence dissident voices, especially in authoritarian societies.

As Díaz Hernández points out, the problem is not just about bans resulting in wrongful or disproportionate restrictions to freedom of expression, but the fact that they also usually end up being inefficient to address and tackle the underlying problem because they fail to counter hate speech, and all too often exacerbate the climate of violence and social divide that gave rise to such content.

It is also essential to bear in mind that the problem around the content moderation actions taken by platforms involves regulating the content itself and the architecture of the Internet and their characteristics as—theoretically—extra-spatial and extra-territorial environments. Based on this structure and the role that platforms and social media outlets play in this ecosystem, each domain has its own rules and definitions of what is or is not prohibited and allowed. In this sense, part of the problem lies in that it is not only a matter of what the legislation of each State understands by hate speech, but also what that term means for Facebook, Twitter, or YouTube, but they are not subject to democratic controls and do not provide due process guarantees or transparency, among other things.

The global pandemic has brought along significant changes in people's lives. Perhaps one of these could also be the beginning of a new discussion about platforms' role as content moderators and the problems that result from allowing or encouraging them to play the role of gatekeepers on the Internet.

ANA LAURA PÉREZ

Uruguay

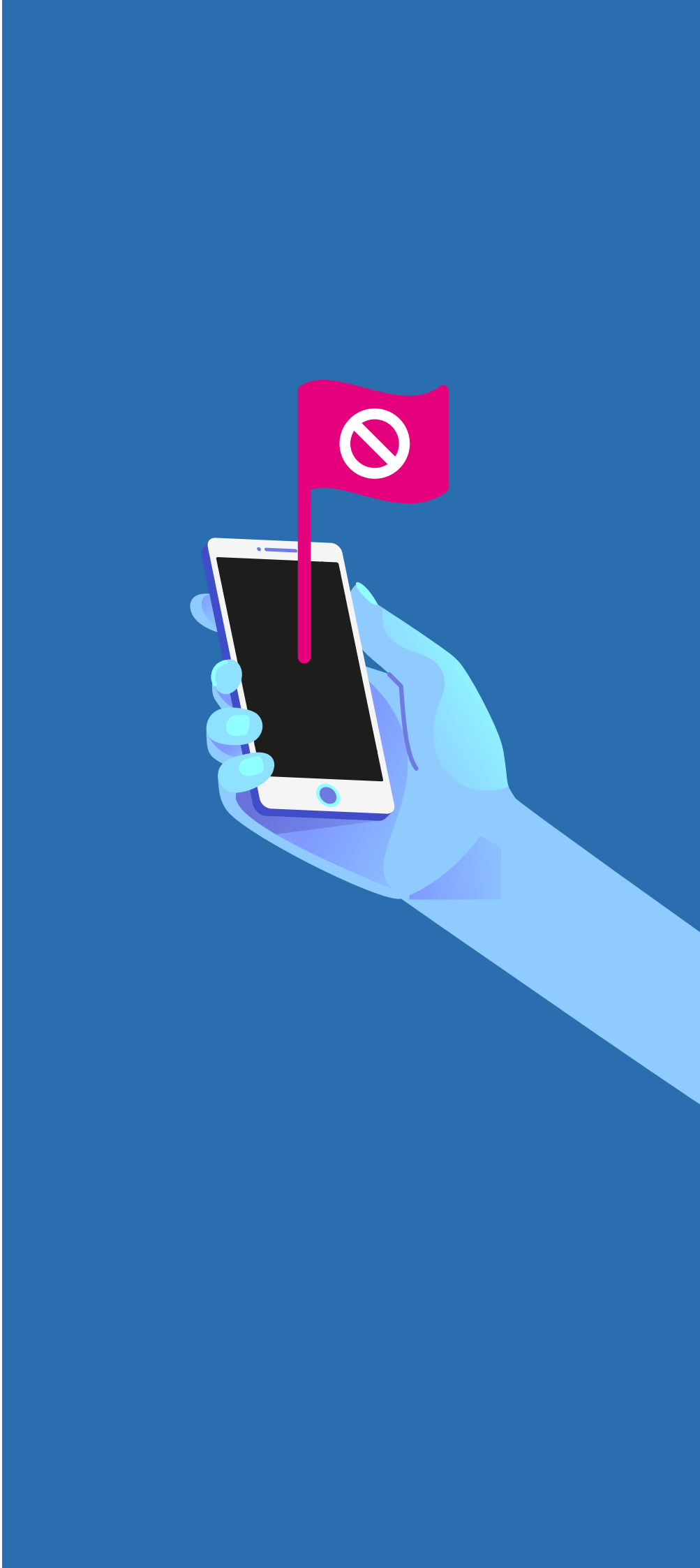
ABOUT THE AUTHOR

She holds a B.A. in Communication and Journalism from ORT University, a diploma in Latin American Studies from the University of Montevideo, and a Master in Business Administration from Montevideo's Institute of Business Studies.

She has worked for 20 years as a journalist and editor in some of the most influential media outlets in her country: El Observador and El País newspapers and the weekly publication Búsqueda. She also hosts and is featured in programs on TV Ciudad, the public channel of the Municipality of Montevideo. She is also currently Digital Product Manager at the El País newspaper.

She was the coordinator of the Journalism and Digital Content studies of the B.A. in Communication program at ORT University, where she has taught for almost ten years.

She has participated as a lecturer, speaker, and panelist in several events on journalism, particularly on disinformation and digital platforms, topics in which she has specialized in recent years and on which she has given training courses to journalists in Uruguay and several Latin American countries.



Funded by the
European Union