JQSS: (Journal Of Quranic and Social Studies)

ISSN: (e): 2790-5640 ISSN (p): 2790-5632

Volume: 5, Issue: 2, May-August 2025. P: 01-17

Open Access: https://jqss.org/index.php/JQSS/article/view/174

DOI: https://doi.org/10.5281/zenodo.15383008

HEC Recognized
Tri-Annual Research Journal
JQSS
Journal of Quranic and Social Studies
MAY-AUGUST 2025. VOL:5. ISSUE:2
Editor: Dr Aman Ullah
pISSN: 2790-5640
eISSN: 2790-5632
www.jqss.org
AL-HIDAYA RESEARCH ACADEMY(Rg)Balochistan,Quetta

The Role of Facebook and Instagram Hate Speech in Societal Polarization: Evidence from Pakistan in a Global Context

1. **Shanila** rj.maniha@gmail.com	M.Phil Scholar, Department of Media Studies, University of Balochistan, Quetta
2. **Dr. Muhammad Fahim Baloch** balochfahim@gmail.com	Associate Professor, Department of Media Studies, University of Balochistan, Quetta
3. **Dr. Nasreen Samar** Samarnasreen@yahoo.com	Assistant Professor, Department of Gender & Development Studies University of Balochistan, Quetta

How to Cite: Shanila and Dr. Muhammad Fahim Baloch and Nasreen Samar (2025). The Role of Facebook and Instagram Hate Speech in Societal Polarization: Evidence from Pakistan in a Global Context, *(JQSS) Journal of Quranic and Social Studies*,5(2), 01-17.

Abstract and

Publisher

HRA (AL-HIDAYA RESEARCH ACADEMY) (Rg)

Balochistan Quetta

The Role of Facebook and Instagram Hate Speech in Societal Polarization: Evidence from Pakistan in a Global Context

Journal of Quranic
and Social Studies
01-17
© The Author (s) 2025
Volume:5, Issue:2, 2025
DOI:10.5281/zenodo.15383008
www.jqss.org
ISSN: E/ **2790-5640**
ISSN: P/ **2790-5632**

OJS
OPEN
JOURNAL
SYSTEMS

PKP
PUBLIC
KNOWLEDGE
PROJECT

Shanila
M.Phil Scholar, Department of Media Studies,
University of Balochistan, Quetta

Dr. Muhammad Fahim Baloch
Associate Professor, Department of Media Studies,
University of Balochistan, Quetta

Dr. Nasreen Samar
Assistant Professor, Department of Gender & Development Studies
University of Balochistan, Quetta

Abstract

The research explores the relationship between hate speech on social media and polarization in Pakistan. It examines the relationship between the way hate speech varies in developing, underdeveloped, and developed countries, along with its contribution to the social, political, and cultural divisions. The research examines the role of social media platforms in either fuelling or controlling hate speech and reviews the effectiveness of institutional and government responses in different regions. Finally, the paper recommends that developing, underdeveloped, and developed countries improve policy regulations and digital governance to reduce the negative influence of hate and minimize polarization in society.

Keywords: Hate Speech, Social Media, Polarization, Digital Governance, Pakistan

Corresponding Author Email:
rj.maniha@gmail.com
balochfahim@gmail.com

Introduction

Social media has rapidly transformed modern communication, which allows individuals across the globe to connect and share their ideas. However, it also becomes a platform for hate speech, which means that there are expressions that give rise to hostility or violence based on religion, race, gender, or identity. This increase in hate speech has raised concerns about its role in enhancing polarization, affecting social, political, economic, religious, cultural, ideological, and identity-based dimensions. In developed countries, stronger regulatory mechanisms and institutions allow for a decrease in their effects, whereas developing and underdeveloped countries often lead to sufficient regulations that lead to more profound social impacts. The difference highlights the need for global and context-specific strategies to address hate speech issues on social media (Pérez-Escolar & Noguera-Vivo, 2022)

Research Aim:

The research aims to explore how hate speech on social media contributes to the polarization in different dimensions (social, political, economic, cultural, religious, ideological, and identity) in developing and underdeveloped countries and compare it to developed countries

Research Objectives:

- To identify the main types of hate speech on social media in Pakistan
- To analyze how hate speech develops polarization across different dimensions
- To compare polarization caused by the hate speech between developing, underdeveloped, and developed countries
- To assess the role of social media platforms in regulating or exacerbating hate speech across different regions.
- To analyze the cultural, institutional, and regulatory differences in response to hate speech.
- To provide policy recommendations to reduce hate speech's negative influence across this context.

Research Questions:

- What types of hate speech are most common on social media in developing, underdeveloped, and developed countries?
- How does hate speech contribute to polarization in different dimensions in Pakistan?
- How does the impact of hate speech-driven polarization differ between developing, underdeveloped, and developed countries?
- What role do social media platforms play in regulating hate speech across these regions?

- How do institutional and regulatory responses to head speech differ between developing, underdeveloped, and developed countries?
- What strategies can reduce the negative impact of hate speech in different contexts?

Problem Statement:

The increase in hate speech on social media has exacerbated polarization, mainly in social, economic, political, cultural, ideological, and religious identity-based dimensions. This problem is more present in developing and underdeveloped countries where weak institutional frameworks and Limited regulations fail to stop its spread. In contrast, developed nations with strong regulatory mechanisms experience less severe impact. The research aims to examine how hate speech drives polarization and the differences between regions in addressing it.

The rationale of the Research:

The increasing influence of social media in shaping the global narrative has intensified the need to understand its role in developing polarization, particularly through hate speech. This issue is more critical in developing and underdeveloped countries, where regulatory structures are usually inadequate. Additionally, it is necessary to understand how hate speech intensifies the political, social, and cultural divisions in these divisions to develop effective strategies to deal with it.

Theoretical Framework:

The research is based on the Social Identity Theory which expresses how hate speech amplifies the ingroup and outgroup dynamics by increasing polarization (Scheepers & Ellemers, 2019). Spiral of Silence Theory highlights how individuals refrain from expressing disobedient opinions among hate-driven narratives (Chaudhry & Gruzd, 2019). Lastly, Agenda Setting Theory discovers how social media platforms prioritize and shape public discussion which further amplifies hate narratives (Gilardi et al., 2021).

Conceptual Framework:

The conceptual Framework discovers the link between hate speech and polarization in different contexts. It will discover the relationship between hate speech on social media and polarization by focusing on the different variables. These elements can either increase or decrease the polarization based on their strength and effectiveness in different regional contexts.

Variables:

The independent variable in this research is measured by the frequency, content, and reach of hate speech. The dependent variables include forms of polarization (social, political, cultural, economic, religious, ideological, and identity-based). Moderating variables include government regulations, institutional capacity, media literacy, and civil service. Mediating variables focus on the perception of hate speech by the public and its online activism.

Literature Review

Hate Speech and Polarization on Social Media:

Schäfer et al., (2022), examine the impact of hate speech on public opinion and social attitude toward social groups like homosexuals and Muslims. Their experimental study is different in terms of the amount of heat speech and the targeted groups. The findings highlight that hate speech against homosexuals negatively influenced perceived social cohesion, while pre-existing attitudes influenced responses to discriminatory demands. However, the amount of hate speech did not affect the perceived public opinion (Schäfer et al., 2022).

Katsarou et al. (2021) highlight the sentiment polarization in online social networks by focusing on Twitter and examining the spread of hate speech. Using hashtag like #Coronavirus, #ClimateChange, #Immigrants, and #MeToo. The research categorized it into 5 classes which include hate speech, offensive, positive, sexist, and neutral. The research used pre-trained models like ULMFiT and AWD-LSTM for classification to achieve high accuracy. The research highlighted how sentiment-driven interactions contribute to the network evolution and emphasized the role of hate speech in increasing polarization online (Katsarou et al., 2021).

Akhtar et al., (2019), address the challenges in automatic hate speech detection because of the subjective nature of the manual annotations. The research highlights the growing issues of hate speech by targeting vulnerable groups on social media and presenting a method that uses fine-grained knowledge from individuals to improve the quality of the training data sets. The researcher presented a better performance in classifying sexist, racist, and homophobic hate speech in tweets by introducing a measure of polarization for individual instances, improving the effectiveness of supervised learning models (Akhtar et al., 2019).

Polarization in Developing and Underdeveloped Countries:

Tucker et al. (2018), Present a comprehensive review of the interaction between social media, political polarization, and disinformation. The analysis of this research categorized different types of politically related information which includes fake news and hyperpartisan content while analyzing their influence on public opinion and political discourse. The review identifies the significant gap in the existing literature about this relationship and focuses on the need for further research and data to better understand how social media contributes to polarization and spread of disinformation in current society (Tucker et al., 2018).

Udanor & Anyanwu, (2019), explores the complexities of defining hate speech within the diverse cultural and religious landscape of Nigeria. It highlights the challenges posed by social media's anonymity that develops the proliferation of hate speech. The implementation of POSA and R-studio shows that the research Quantatively examines hate speech problems in tweets and presents a significant percentage of hate content. The finding indicates a lack of automated monitoring systems on platforms such as

Facebook and Twitter which highlight the need for more effective mechanisms to address hate speech in Nigeria (Udanor & Anyanwu, 2019).

Ali et al., (2019), highlight the double role of social media in the Arab Spring, mainly focusing on Egypt's 2011 revolution. It examines how social media develops social capital by allowing integration through bonding, bridging, and linking. However, the absence of contextual factors can lead to polarization instead of integration. The application of social capital theory in this research highlights the significant influence of social media on socio-political dynamics which provide key insight into the relationship between social media and political change in the region (Ali et al., 2019).

Polarization in Developed Countries:

Belcastro et al., (2020), explore the field of social media by emphasizing the development of the IOM-MN methodology for the identification of user polarization during electoral events. Research highlights the effectiveness of feed-forward neural networks in developing classification rules from initial hashtags linked to political factions. The research shows that this approach surpasses the traditional sentiment analysis method by comparing results from the 2018 Italian and 2016 US elections. It provided a more accurate representation of political polarization among social media users (Belcastro et al., 2020).

Urman (2019) examines the political polarization on Twitter and shows a major variation across different countries. The study analyzes data from 16 democratic nations by using a network analytical audience duplication approach and categorizing their political Twitter spheres into different polarization levels. The finding indicates that polarization peaks in a two-party system with plurality electoral rules, by contrast with lower levels in a multi-party system through promotional voting. The result challenges the previous single case study conclusions by focusing on the need for more comparative research to get a comprehensive understanding of polarization dynamics in social media (Urman, 2019).

Institutional and Regulatory Responses to Hate Speech:

Matamoros-Fernández & Farkas, (2021), a systematic review explores racism and hate speech within social media research. The research analyzed 104 articles to identify the geographical context, platforms, and methodologies used. It highlights a lack of diversity and critical engagement with systemic racism by highlighting the need for research to consider institutional and regulatory responses to these issues. The articles suggest that it is necessary to understand how user practices and platform politics shape racism. It needs a more detailed analysis of how policies and regulations of social media effectively reduce hate speech and develop an equitable digital environment (Matamoros-Fernández & Farkas, 2021).

MacAvaney et al., (2019), review the challenges linked with hate speech detection in online content by highlighting issues such as different definitions of hate speech, linguistic subtitles, and data limitations for training detection systems. It highlights the interpretability problem in current approaches which makes it hard to understand system decisions. The author proposes a multi-view support machine (SVM) method that achieves near state-of-the-art performance while enhancing interpretability as compared to the neural network methods.The review focuses on the ongoing technical and practical challenges by providing modified interventions to deal with hate speech issues (MacAvaney et al., 2019).

Research Methodology

The research used a secondary research approach by focusing on desk research to collect and analyze the existing literature and reports.The methodology facilitates the exploration of established knowledge and findings related to the research topic as it allows a comprehensive understanding of the subject matter without primary data collection (Sileyew, 2019).

Data Collection Methods:

Secondary data is collected from a variety of sources, which include reports, academic journal articles, social media studies, and government Publications.This range of material provides a comprehensive understanding of hate speech and polarization that allows the analysis of existing research findings and trends.The data collected will support the exploration of the relationship between social media, hate speech, and political polarization in current society (Sileyew, 2019).

Data Analysis Techniques:

The data analysis used primary techniques as described below

- **Comparative Analysis**: It focuses on examining the effects of speech across different regions, including developing, underdeveloped, and developed countries. It highlights the different impacts and responses in this context. It will allow for a clear understanding of the original factors that impact the prevalence and effects of hate speech (Ruggiano & Perry, 2017).
- **Thematic Analysis**: It includes the identification of main themes within the collected data such as diverse types of hate speech and their contribution to political polarization.The categorizing of data into themes allows for revealing the underlying patterns and dynamics that characterize hate speech in different settings. Together this method allows for a comprehensive framework to understand the complex relationship between polarizations and hate speech (Ruggiano & Perry, 2017).

Limitations of the Study:

The study faces several limitations, which include the availability and reliability of data from underdeveloped countries and may hinder comprehensive analysis. Moreover, cultural differences and variations in reporting practices may influence how hate speech is defined and interpreted across different backgrounds. This factor may limit the generalizability of findings and the accuracy of the conclusions drawn from research.

Data Analysis/Discussion

Types of Hate Speech across Regions:

The types of hate speech on social media vary significantly across regions that are shaped by political, cultural, and socio-economic factors.The underdeveloped countries like Myanmar, religious and ethnic hate speech is widespread, particularly targeting the Rohingya Muslim Minority. A study by the UN Human Rights Council shows that Facebook was a key platform for spreading anti-Rohingya sentiments, which contributed to real-world violence. Here, hate speech is often linked to ethnic conflict (Szurlej, 2016). In developing countries like India, hate speech is caused by religious and political divides. More than 900,000 posts were removed by Facebook for violating hate speech policies during the time of 2019, among which mostly involved anti-Muslim rhetoric. Facebook is mostly used for hate speech, as shown in Figure 1 in the Appendix. The European Journal of Communication highlights the way social media in India has become the reason for increasing caste-based discrimination, religious intolerance, and political partisanship (Al Jazeera, 2021). On the other hand, in developed countries like the U.S., racial and political hate speech dominates platforms. A Pew Research report in 2017 showed that nearly 41% of Americans have experienced online harassment in different ways. Some of it is largely linked to race and political polarization (Vogels, 2021).

Polarization across Dimensions:

Hate speech drives polarization across different dimensions which exacerbates the conditions in developing and developed countries (See Figure 2). Politically, hate speech has intensified the partisan divides in the U.S. The research shows that almost 55% of Americans viewed the opposing political party as a threat to the nation. It is also found that political polarization deepened during the 2020 elections and platforms such as Twitter, and Facebook were found to foster echo chambers (Yu et al., 2023). In developing countries such as Nigeria, religious hate speech resulted in wireless between Muslim and Christian communities. Figure 3 shows the distribution of religion in Nigeria (see Figure 3).The research on this issue shows that almost 20,000 deaths are linked to the religious conflict, which is caused by online hate speech. It economically affects minority groups and contributes to unequal access to resources (USCIRF, 2024). In India, the hate speech against Dalits marginalizes them socially and economically by restricting access to education and jobs. In terms of culture, identity-based polarization can be seen in countries like Brazil where social media platforms have raised racial and gender-based hate speech (Sharma, 2015). A UN Women report found that almost 60% of Brazilian women faced online Harassment because of race and gender which highlights the cultural and ideological polarization developed by hate speech in these regions (UN Women, 2024).

Role of Social Media in Spreading Hate Speech in Pakistan:

Social media has revolutionized communication by connecting people worldwide and sharing information. A diverse nation with many cultural, religious, and ethnic identities, Pakistan uses Twitter, Facebook, and WhatsApp for public discourse. These platforms have many benefits, but they have also promoted hate speech, which has increased political, social, economic, cultural, religious, ideological, and identity-based polarization. This phenomenon deepens social divisions, weakens social cohesion, and strains Pakistan's democracy. To understand how hate speech on social media can polarize people, one must study national history, sociopolitical dynamics, and how digital platforms amplify divisive narratives.
Pakistan's complex sociopolitical landscape exacerbates polarization. Punjabis, Sindhis, Pashtuns, and Baloch, each with their own culture and language, have shaped the country's history. Religious diversity—mostly Sunnis, a minority Shia, and smaller communities like Ahmadis, Christians, and Hindus—adds complexity.

The partition of India in 1947, Zia-ul-Haq's military regimes, and extremist groups have exacerbated political divisions, often based on religion and ethnicity. Social media platforms allow hate speech—language intended to incite hatred toward individuals or groups based on ethnicity, religion, or ideology—to spread quickly and unchecked. In contrast to traditional media, social media allows users to anonymously spread hateful content, reaching many people and escalating social tensions.
Pakistani political polarization has increased due to social media hate speech. On Twitter, political discourse often turns negative. PTI, PML-N, and PPP supporters name-call, accuse, and propagandize. These conversations reinforce "us versus them" and ideological divides (Hassan et al., 2020).

Social media was flooded with anti-candidate hashtags and campaigns during the 2018 general elections. Candidate corruption and disloyalty to the nation were common during campaigns. This rhetoric fits with affective polarization, in which people view political opponents as threats rather than competitors. In 2024, the Montreal AI Ethics Institute found a link between online hate speech and political upheavals like the 2022 PTI overthrow. These studies say social media data reflects offline political events in Pakistan at the time. This shows that hate speech deepens political divisions, reducing opportunities for compromise (montrealethics.ai, 2025).

Social media hate speech amplifies political party divisions and polarizes society. Pakistani society is stratified by class, ethnicity, and region, so social media often targets marginalized groups. When targeted by hate speech, ethnic minorities like Baloch and Pashtuns are often labeled separatists or anti-state activists. Audience mistrust increases as stereotypes are strengthened. Online campaigns call Baloch activists traitors. Due to social isolation, state crackdowns are justified. Urbanites use social media to mock rural communities as backwards, alienating them. Because social media platforms provide anonymity, users can express prejudices they might otherwise repress, creating echo chambers that reinforce hateful narratives (Shafiq et al., 2024). Professors like Sunstein say the process causes group polarization. When like-minded people gather, radical views are emphasized, and the number of perspectives decreases (Mahmood et al., 2024).

Hate speech on social media worsens economic polarization and other issues. Pakistan has a large economic gap between urban elites and the rural poor, as well as between provinces. This inequality is widespread in Pakistan. Economic grievances are often weaponized by social media hate speech. Sindh and Balochistan users accuse Punjabis of resource monopolization. Punjab is the most populous and prosperous province. Resentment and economic inequality increase with these narratives. Urban elites blame rural or working-class communities for economic stagnation on social media. These exchanges widen global economic divides by portraying different groups as adversaries rather than stakeholders in a shared economy.

The 2023 World Economic Forum global risks report lists polarization as a long-term threat. Hate speech and misinformation on digital platforms erode social cohesion and hinder global economic cooperation and development, according to this report. Pakistan's ethnically and culturally diverse heritage and competing national identity narratives cause cultural polarization. Definition of "authentic". Social media discussions about Pakistani culture often involve modernist and secular voices and culturally conservative and religious voices. Hatred of music, dance, and traditional festivals stigmatizes them as un-Islamic or foreign, alienating communities that enjoy them. Extremists have called Pakistan's culturally significant Sufi shrines heretical online. This has increased the cultural gap between scripturalists and Sufis. These debates were shaped by General Zia-ul-Haq's Islamization policies. The policies institutionalized Sunni orthodoxy. The policies took effect. Social media promotes extremist voices and marginalizes moderate perspectives, creating a polarized cultural landscape that threatens pluralism and escalating tensions.

Social media hate speech has exacerbated Pakistan's biggest division, religious polarization. Internet rhetoric targets Shias, Ahmadis, Christians, and Hindus, undermining constitutional religious diversity protections. A 2021 Journal of Islamic Thought and Civilization study found a link between religion and Facebook hate speech. Particularly during sectarian events like Muharram, when Shia and Sunni users share inflammatory content. Ahmadis, non-Muslims, are persecuted online and called for exclusion or violence under Zia's laws. Blasphemous accusations, which can have serious legal and social consequences in Pakistan, are spread more easily on social media. Bytes for All's 2020 report showed how internet hate speech marginalizes religious minorities, reaching over half of young people. Religious polarization radicalizes majorities and isolates minorities. Tehreek-e-Labbaik Pakistan (TLP) has grown by using social media to rally supporters around blasphemy issues (Mahmood et al., 2024).

Social media hate speech polarizes ideologies, including political and religious views. Pakistan's ideological spectrum ranges from theocratic Islamists to modernizing secular liberals. Social media allows extremists to spread divisive narratives. Islamists are called terrorists or backwards, while secularists are called Western agents or atheists. Hate speech-inspired labels reduce complex ideological debates to binary oppositions, removing nuance. A 2023 study in the International Research Journal of Management and Social Sciences found that political speeches on social media often include hate speech to eliminate opponents, deepening ideological divides. Internet conflict between traditionalist Muslims and liberal women's rights advocates often escalates into threats and abuse. Internet conflicts between these groups demonstrate this. Social media hate speech polarizes gender, religion, and race identities (Ali Abid et al., 2021).

Pakistan's diverse identities are both a strength and a vulnerability because social media can bridge or widen these gaps. Hate speech includes portrayals of Pashtuns as terrorists or women as immoral for defying norms. X posted that the 2021 Pahalgam attack was followed by an increase in hate speech and unsubstantiated claims linking Pashtuns to terrorism. Gender-based hate speech and online harassment of female activists are common. This identity-based polarization creates a highly fragmented society where people value group affiliations over national unity. This undermines Pakistani unity (Ali Abid et al., 2021).

Many factors cause social media hate speech polarization. The algorithms that power Facebook and Twitter prioritize engaging content, which often includes sensationalist or inflammatory posts. A feedback loop makes hate speech more noticeable and increases reactions and shares. Echo chambers, where users only see similar views, amplify divisive rhetoric (Shafiq et al., 2024). In Pakistan, false narratives about political opponents or minorities spread quickly, according to a 2023 World Economic Forum report. Misinformation and hate speech erode social cohesion, the report shows. The lack of robust content moderation exacerbates the issue. Twitter has policies against hate speech, but enforcement is inconsistent, especially in non-English languages like Urdu and Pashto, which allow hate speech to flourish. Polarization will have serious consequences. Politicians prioritizing partisan gains over national interests cause legislative gridlock and undermine democratic institutions. A political issue. Socially, it promotes distrust and hostility as communities view each other as competitors. It hinders economic cooperation, as provinces disagree on resource distribution. It promotes extremist narratives, threatening Pakistan's pluralistic culture. Religion causes sectarian violence and marginalizes minorities (Farooq et al., 2024).

From an ideological standpoint, it stifles debate and promotes dogma. Identity polarization undermines social cohesion and makes national identity formation harder. The 2024 WIREs Computational Statistics study found no definitive solution to hate speech detection despite advances. This shows the ongoing challenge of solving this issue. To combat hate speech and its polarizing effects, a multifaceted approach is needed.

Legal reforms like blasphemy law reform and anti-hate speech legislation are essential. Enforcement must balance protecting vulnerable groups and free speech.

Education and media literacy programs can help users critically evaluate online content, reducing their risk of misinformation. Social media platforms must improve content moderation, especially in local languages, and collaborate with local organizations to understand Pakistan's sociocultural context. Civil society can help promote tolerance and diversity counter-narratives. In conclusion, political leaders should model inclusive discourse rather than using hate speech to win elections. The conclusion is that social media hate speech has deepened political, social, economic, cultural, religious, ideological, and identity polarization in Pakistan. Divisive narratives diminish trust, foster hostility, and undermine Pakistani pluralism. Social media connects, but unchecked hatred is dangerous. Citizens, platforms, and policymakers must work together to create a digital environment that promotes communication and unity to solve this problem. Pakistan alone can overcome polarization and build a united future with such measures (Akbar & Safdar, 2024).

Comparative Analysis of Polarization:

Hate speech develops polarization differently, which is based on the regulatory strengths of the region. Countries with strong regulations such as Germany impose hate speech laws under the Network Enforcement Act (NetzDG) to pay fines on social media platforms for selling to removing illegal content which decreases the spread of inflammatory speech. The law applies only to social media networks having 2 million or more registered users and it is illegal if they fail to remove the content after a complaint within 24 hours (see Figure 4) (LOC, 2024). It has contributed to the lower level of online polarization for instance; the 2021 Digital report of the Reuters Institute noted that Germany's polarized political climate on social media platforms is less severe to the US. In contrast, countries with weak regulations face more polarization (Reuters, 2024). The hate speech on Facebook in Myanmar played an important role in increasing violence against the Rohingya minority due to the lack of regulatory oversight on online platforms (Szurlej, 2016). In countries like India, it shows a mixed impact with moderate regulations. Online polarization remains high despite recent legislative efforts like the Information Technology Rules, particularly regarding religious and political issues (Meity, 2024).

The Role of Social Media Platforms:

Social media platforms play an important role in either increasing or decreasing hate speech issues. The algorithms often intensify the content as observed in Facebook's role in Myanmar, where the UN found that the Rohingya were not adequately removed and contributed to the violence (Szurlej, 2016). Same as in India, platforms like WhatsApp have been used to spread disinformation and hate speech mainly in the lead-up to elections, causing religious and political polarization (Meity, 2024). In contrast, the platform in regions with stronger regulation, such as the European Union, has applied strict moderation policies. For example under NetzDG laws in Germany, social media companies face fines for not actively removing illegal hate speed which led to quicker takedowns and reduced online hostility (LOC, 2024). Figure 5 shows the cases of online hate speech removed by the Police (see figure 5). Kenya the 2017 elections, institutions like the National Cohesion and Integration Commission (NCIC) monitored hate speech, resulting in several arrests and deterrence of inflammatory rhetoric (Cohesion, 2024). In the US, platforms such as Facebook and Twitter have presented more aggressive content moderation policies but these actions have raised debates around free speech that demonstrate the complexity of the platform's responsibility in different regions (Reuters, 2024).

Conclusion:

In conclusion, research shows that hate speech significantly increases polarization across social political and cultural dimensions mainly in regions with weak regulatory frameworks. Insufficient government interventions in developing and underdeveloped countries increase polarization, whereas developed nations with strong regulations see moderated effects. Social media platforms play are double role in both spreading and reducing hair speech by highlighting the need for strong institutional oversight to reduce polarization caused by online hate.

Recommendations:

Strategies for Developing and Underdeveloped Countries:

The government must develop a legal Framework for hate speech by focusing on awareness and education to counteract its effect. Civil society can play a role by promoting digital literacy, which helps the user to identify and report hate speech. Social media platforms must collaborate with local authorities and NGOs to apply region-specific moderation strategies and offer tools for reporting abusive content. Moreover, cross-border cooperation can help in applying policies more effectively in regions where there are weak regulations.

Strategies for Developed Countries:

Developed countries must improve existing regulations by implementing stricter content moderation policies on social media platforms to ensure transparency. The government can improve platforms' ability by using AI tools to detect hate speech proactively while safeguarding free speech rights. Strengthening International collaboration in hate speech regulation can also address challenges caused by cross-border online platforms.

Bibliography

1. Akhtar, S., Basile, V., & Patti, V. (2019). A new measure of polarization in the annotation of hate speech. Lecture Notes in Computer Science, 588–603. https://doi.org/10.1007/978-3-030-35166-3_41
2. Akbar, M. and Safdar, A. (2024) 'Exploring ethnic discrimination and hate speech in online political discourses: A comprehensive analysis from

the Pakistani context', Annals of Human and Social Sciences, 5(I). doi:10.35484/ahss.2024(5-i)25.

3. Ali Abid, A., Shami, S. and Ashfaq, A. (2021) 'Facebook and hate speech: Analyzing relationship between consumers' attributes and Islamic sectarian content on social media in Pakistan', Journal of Islamic Thought and Civilization, 11(1), pp. 453–462. doi:10.32350/jitc.111.2.
4. Ali, M., Azab, N., Sorour, M. K., & Dora, M. (2019). Integration v. Polarisation among social media users: Perspectives through social capital theory on the recent Egyptian political landscape. Technological Forecasting and Social Change, 145, 461–473. https://doi.org/10.1016/j.techfore.2019.01.001
5. Al Jazeera. (2021, October 25). Facebook failing to check hate speech, fake news in India: Report. https://www.aljazeera.com/news/2021/10/25/facebook-india-hate-speech-misinformation-muslims-social-media
6. Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2020). Learning political polarization on social media using Neural Networks. IEEE Access, 8, 47177–47187. https://doi.org/10.1109/access.2020.2978950
7. Chaudhry, I., & Gruzd, A. (2019). Expressing and challenging racist discourse on Facebook: How social media weaken the "Spiral of silence" theory. Policy & Internet, 12(1), 88–108. https://doi.org/10.1002/poi3.197
8. Cohesion. (2024). Kenya's National Action Plan Against Hate Speech. https://cohesion.go.ke/images/docs/downloads/Kenyas_National_Action_Plan_Against_Hate_Speech.pdf
9. Dixon, S. J. (2024, September 12). Facebook hate speech removal per quarter 2024. Statista. https://www.statista.com/statistics/1013804/facebook-hate-speech-content-deletion-quarter/
10. Davies, K. (2022, November 10). Cases of hate speech recorded by the police in Germany 2009-2020. Statista. https://www.statista.com/statistics/961603/cases-of-sedition-recorded-by-the-police-in-germany/
11. Fleck, A., & Richter, F. (2024, October 21). Infographic: 2 in 3 people often encounter hate speech online. Statista Daily Data. https://www.statista.com/chart/33299/online-hate-speech-encounters/
12. Farooq, S., Zain, A. and Sartaj, S. (2024) Hate and polarization in society: A case study of imran khan and Maryam Nawaz Speeches, ResearchGate. Available at: https://www.researchgate.net/publication/379929809_Hate_and_pola

rization_in_society_A_case_study_of_Imran_khan_and_Maryam_Nawaz_Speeches (Accessed: 10 May 2025).

13. Gilardi, F., Gessler, T., Kubli, M., & Müller, S. (2021). Social Media and political agenda setting. Political Communication, 39(1), 39–60. https://doi.org/10.1080/10584609.2021.1910390

14. Hassan, A.A.U., Fazal, H. and Khalid, T. (2020) Political hate speech in political processions: A comparative analysis of PMLN, PPP and PTI processions for election 2018, Pakistan Journal of Social Sciences. Available at: https://pjss.bzu.edu.pk/index.php/pjss/article/view/920 (Accessed: 10 May 2025).

15. Katsarou, K., Sunder, S., Woloszyn, V., & Semertzidis, K. (2021). Sentiment polarization in online social networks: The flow of hate speech. 2021 Eighth International Conference on Social Network Analysis, Management and Security (SNAMS), 99, 01–08. https://doi.org/10.1109/snams53716.2021.9732077

16. LOC. (2024). Germany: Network Enforcement Act amended to better fight online hate speech. The Library of Congress. https://www.loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/#:~:text=Background%20on%20the%20Network%20Enforcement%20Act&text=The%20Network%20Enforcement%20Act%20is, after%20receiving%20a%20user%20complaint.

17. Meity. (2024). The Information Technology (intermediary guidelines and ... https://www.meity.gov.in/writereaddata/files/Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (updated 06.04.2023)-.pdf

18. MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PLOS ONE, 14(8). https://doi.org/10.1371/journal.pone.0221152

19. Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and Social Media: A systematic review and Critique. Television & New Media, 22(2), 205–224. https://doi.org/10.1177/1527476420982230

20. Mahmood, F., Zahra, Ms.S.M. and Mehdi, Dr.A. (2024) The role of social media in amplifying hate speech: A qualitative analysis of Imran Khan and Nawaz Sharif's rhetoric on YouTube, Policy Research Journal. Available at: https://theprj.org/index.php/1/article/view/226 (Accessed: 10 May 2025).

21. montrealethics.ai (2025) Democratizing AI Ethics Literacy, Montreal AI Ethics Institute. Available at: https://montrealethics.ai/ (Accessed: 10 May 2025).

22. Pérez-Escolar, M., & Noguera-Vivo, J. M. (2022). Hate speech and polarization in participatory society (p. 278). Taylor & Francis.

23. Ruggiano, N., & Perry, T. E. (2017). Conducting secondary analysis of qualitative data: Should we, can we, and how? Qualitative Social Work, 18(1), 81–97. https://doi.org/10.1177/1473325017700701

24. Reuters. (2024). Ox. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf

25. Sasu, D. D. (2023, August 3). Nigeria: Distribution of religions. Statista. https://www.statista.com/statistics/1203455/distribution-of-religions-in-nigeria/

26. Sharma, S. (2015). Caste-based crimes and economic status: Evidence from India. Journal of Comparative Economics, 43(1), 204–226. https://doi.org/10.1016/j.jce.2014.10.005

27. Szurlej, C. (2016). (PDF) preventing genocide against the Rohingya Muslim minority in Myanmar. https://www.researchgate.net/publication/312277661_Preventing_Genocide_against_the_Rohingya_Muslim_Minority_in_Myanmar

28. Shafiq, S., Rehman, Dr.S. ur and Khanum, K. (2024) The role of social media echo chambers in promoting divisive opinions and hate speech, PAKISTAN JOURNAL OF LAW, ANALYSIS AND WISDOM. Available at: https://pjlaw.com.pk/index.php/Journal/article/view/v3i9-96-104?articlesBySimilarityPage=4 (Accessed: 10 May 2025).

29. Scheepers, D., & Ellemers, N. (2019). Social Identity Theory. Social Psychology in Action, 129–143. https://doi.org/10.1007/978-3-030-13788-5_9

30. Sileyew, K. J. (2019). Research design and methodology (Vol. 7). Cyberspace.

31. Schäfer, S., Sülflow, M., & Reiners, L. (2022). Hate speech as an indicator for the state of the Society. Journal of Media Psychology, 34(1), 3–15. https://doi.org/10.1027/1864-1105/a000294

32. Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3144139

33. Udanor, C., & Anyanwu, C. C. (2019). Combating the challenges of social media hate speech in a polarized society. Data Technologies and Applications, 53(4), 501–527. https://doi.org/10.1108/dta-01-2019-0007

34. Urman, A. (2019). Context matters: Political polarization on Twitter from a comparative perspective. Media, Culture &amp; Society, 42(6), 857–879. https://doi.org/10.1177/0163443719876541

35. UN Women. (2024). Frequently asked questions: Tech-facilitated gender-based violence. UN Women – Headquarters.

https://www.unwomen.org/en/what-we-do/ending-violence-against-women/faqs/tech-facilitated-gender-based-violence

36. USCIRF. (2024, October 23). Central Nigeria: Overcoming dangerous speech and endemic religious divides. https://www.uscirf.gov/publications/central-nigeria-overcoming-dangerous-speech-and-endemic-religious-divides
37. Vogels, E. A. (2021, January 13). The state of online harassment. Pew Research Center. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/
38. Yu, X., Wojcieszak, M., & Casas, A. (2023). Partisanship on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users. Political Behavior, 46(2), 799–824. https://doi.org/10.1007/s11109-022-09850-x
39. Zandt, F., & Richter, F. (2024, June 18). Infographic: Meta's hate speech problem. Statista Daily Data. https://www.statista.com/chart/21704/hate-speech-content-removed-by-facebook/